ANNI JÄRVELIN

Applications of S-grams in Natural Language Information Retrieval

■

ACADEMIC DISSERTATION
To be presented, with the permission of
the Board of the School of Information Sciences of the University of Tampere,
for public discussion in the Auditorium Pinni B 1097,
Kanslerinrinne 1, Tampere, on December 18th, 2014, at 12 o'clock.

UNIVERSITY OF TAMPERE

ANNI JÄRVELIN

Applications of S-grams in Natural Language Information Retrieval

Acta Universitatis Tamperensis 2010
Tampere University Press
Tampere 2014

ACADEMIC DISSERTATION
University of Tampere
School of Information Sciences
Finland

The originality of this thesis has been checked using the Turnitin OriginalityCheck service in accordance with the quality management system of the University of Tampere.

Cover design by
Mikko Reinikka

Distributor:
kirjamyynti@juvenes.fi
http://granum.uta.fi

Acta Universitatis Tamperensis 2010
ISBN 978-951-44-9674-5 (print)
ISSN-L 1455-1616
ISSN 1455-1616

Acta Electronica Universitatis Tamperensis 1499
ISBN 978-951-44-9675-2 (pdf)
ISSN 1456-954X
http://tampub.uta.fi

Suomen Yliopistopaino Oy – Juvenes Print
Tampere 2014

Acknowledgements

This has been a long road, with many detours from my side and much patience required from my supervisors Professor Eero Sormunen, Dr. Heikki Keskustalo and Dr. Ari Pirkola. Thank you for always being there whenever I returned from one of my detours to continue the thesis work. Thank you, Eero, for guidance in how to think, work and write science and for fixing my reference lists, over and over again. Thank you, Heikki, for your energy and commitment. Your ability to solve problems and your will to really dive in and understand the practical work kept me going when facing seemingly insurmountable obstacles. Thank you, Ari, for your expertise and inspiration and for always giving me all the practical advice and support I needed.

I'd like to thank my co-authors: Sanna Kumpulainen, Antti Järvelin, Kalervo Järvelin, Peter Wilkins, Eija Airio, Kimmo Kettunen, Miamaria Saastamoinen, and others. It has been a privilege working with you. Sanna, we started this work together and many of the things I now know I first learned with you. Thanks for making approximate string matching fun. Antti, thank you for patiently teaching me math and for all the programming work you did. It made my work much easier.

My friends and colleagues from my time at SICS: without your kind support, I would definitely have finished this thesis a year or two earlier, but knowing less and having had less fun. I appreciate every minute spent on other things with you. I would not have been able to finish this work without the funding from the Tampere Graduate Programme for Information Science and Engineering (TISE).

Finally, my sincere thanks to my family and friends. Thank you for being there and not too often asking about it. Thank you, Lotta and Daniel, for lending your home when I was desperate to find a peaceful place to work at. Mom and dad, for always (and uncritically) believing in me. Aino and Antti, for showing the way. Having seen you struggle with your theses, I was prepared and never felt alone. Alvar and Elsa, for giving me other things to think about, and Anders, for your patience, support and love. And for always reading all my texts. Thank you!

Stockholm, November 13th 2014
Anni Järvelin

Abstract

Several phenomena cause words to occur in different surface forms in natural language texts including variation and fluctuations in the orthographical rules of languages, morphological variation, historical variation, and variation due to errors in optical character recognition (OCR) of digitized documents. This string level variation interferes with the word matching and weighting mechanisms of information retrieval.

This thesis examines the use of approximate string matching methods for handling the string level variation occurring in alphabetic languages. An approximate string matching technique called classified *s*-gram matching is used for query translation and expansion in cross-lingual and historical retrieval scenarios. Test collection based evaluation methodology is adopted and three different test collections are used for analyzing the variation occurring in different types of collections, in different text domains and between different language pairs: (1) between the closely related languages Norwegian and Swedish in the news text domain, (2) between several European languages in a collection of lightly annotated photographic images, and (3) between contemporary and historical Finnish in a digitized historical document collection.

The findings show that classified *s*-gram matching is a useful approach to handling the variety of string level variation that occurs in natural language document collections. It is well suited for modelling cross-lingual, historical and morphological variation in strings, as well as the most common OCR errors. The linguistic relations between the source and target languages, the text domain, and the morphological complexity affect the performance of the classified *s*-gram matching technique. It is best suited for applications where the target index can be expected to contain many relevant variants of the query words, and where query expansion through adding several of them to the queries is desirable. The optimal number of *s*-gram variants to be added to the target language queries depends on the target language and the collection. Only a few variants are needed for query translation in clean, lemmatized, bilingual collections, while dozens of variants may beneficially be added to queries in collections where combinations of many productive string level variation types occur.

Table of contents

Original publications

This dissertation consists of a summary and the following original research publications:

I. Järvelin, A., Järvelin, A. & Järvelin, K. (2007). *S*-grams: Defining Generalized *n*-grams for Information Retrieval. Information Processing and Management, 43(4), 1005-1019.

II. Järvelin, A. & Järvelin, A. (2008). Comparison of *s*-gram Proximity Measures in Out-of-Vocabulary Word Translation. In Amir, A. & al. (Eds.), Proceedings of the 15th String Processing and Information Retrieval Symposium (SPIRE 2008), Melbourne, Australia, November 2008. Springer Berlin Heidelberg, pp. 75-86.

III. Järvelin, A., Kumpulainen, S., Pirkola, A. & Sormunen, E. (2006). Dictionary-independent translation in CLIR between closely related languages. In de Jong, F.M.G. & Kraaij, W. (Eds.), Proceedings of the 6th Dutch-Belgian Information Retrieval workshop (DIR 2006), Delft, The Netherlands; March 2006, pp. 25-32.

IV. Järvelin, A., Wilkins, P., Adamek, T., Airio, E., Jones, G., Smeaton, A. & Sormunen, E. (2008). DCU and UTA at ImageCLEFPhoto 2007. In Peters, C. & al. (Eds.), Advances in Multilingual and Multimodal Information Retrieval: Proceedings of the eighth Workshop of the Cross-Language Evaluation Forum, CLEF 2007 (Budapest, Hungary), 530-537. Springer Berlin Heidelberg.

V. Järvelin, A., Keskustalo, H., Sormunen, E., Kettunen, K. & Saastamoinen, M. Information retrieval from historical newspaper collections in highly inflectional languages: a query expansion approach. Accepted for publication in Journal of the Association for Information Science and Technology (JASIST).

The publications will be referred to as Studies I-V in the dissertation summary. They are reprinted by permission of the publishers.

The author's research contributions and the roles of the studies

The studies included in this thesis form two themes. Studies I and II build the background: they present definitions for the classified *s*-gram matching and measures for *s*-gram based string proximity (Study I), evaluate different proximity measures and combinations of *s*-gram lengths and other *s*-gram settings (Study II), and present a literature review of approximate string matching applications in information retrieval (Study I). Studies III-V then describe experiments in three application domains: retrieval between closely related languages (Study III), cross-language image retrieval (Study IV), and historic document retrieval (Study V).

I have had collaborators and co-authors in all studies included in this thesis. In Study I, Dr. Antti Järvelin was responsible for the definitions for classified *s*-grams and their proximity measures, while I was more perfunctorily involved in that discussion. I was responsible for the introduction, literature review and the motivating example. Therefore, the *s*-gram definitions are not a part of the contributions of this thesis, and they are not discussed in the introduction. Professor Kalervo Järvelin wrote the discussion and contributed with valuable comments.

In Study II, I was responsible for the study design and preparation of the publication. Dr. Antti Järvelin built the test environment for *s*-gram matching and evaluation, while I was responsible for the test data, analysis of cross-lingual variation types for defining useful gram classes, runs and analysis of the results. The article was written collaboratively with me as the lead author.

In Study III, I was responsible for the preparation of the publication. All experimental work was conducted in collaboration with Dr. Sanna Kumpulainen. Dr. Ari Pirkola and Professor Eero Sormunen acted as the supervisors of the work.

Study IV was a result of collaboration between University of Tampere (UTA) and Dublin City University (DCU). Peter Wilkins (DCU) was responsible for visual retrieval component and data fusion with support from Dr. Tomasz Adamek, Dr. Gareth Jones and Professor Alan Smeaton. I was responsible for UTA's work on the text retrieval and query translation component. Eija Airio provided technical

support and was closely involved in the dictionary-based translation set-up. Professor Eero Sormunen acted as the supervisor of the work.

In Study V, I collected and analyzed the data and acted as the lead author. Dr. Heikki Keskustalo provided valuable technical support and together with Professor Eero Sormunen had a major impact on the final form of the article. Dr. Kimmo Kettunen contributed with the work on FCG. Miamaria Saastamoinen contributed to the creation of the test collection and thus contributed with her knowledge of the collection in various phases of the study and helped describe the test collection. Dr. Heikki Keskustalo, Professor Eero Sormunen, and Dr. Ari Pirkola acted as my academic supervisors.

1 Introduction

There is a long tradition of word-based text representation in natural language information retrieval (Kettunen, 2009). The contents of text documents are usually represented in indexes by the words that occur in them and the frequency-based weights of the words. Users' information needs are represented to the system by query words, and information retrieval is typically based on matching query words with document words. There are several phenomena that cause words to occur in different surface forms in natural language texts, including morphological variation, spelling errors and alternative spellings, historical variation, cross-lingual spelling variation, and variation due to errors in optical character recognition (OCR) of digitized documents. This string level variation interferes with the simple word matching and weighting mechanisms of information retrieval. The different forms of a word represent the same concept and are therefore equal from the standpoint of users' information needs (Pirkola, Keskustalo, Leppänen *et al.,* 2002). Therefore, recognizing all surface form variants of a word as occurrences of the same word is essential for information retrieval performance.

Different solutions have been suggested for handling string level variation in natural language information retrieval. **Linguistically motivated approaches** extensively rely on linguistic models, e.g., on morphology and vocabulary. Implementations of these models are used for handling morphological variation (lemmatizers, stemmers), for translation (dictionary-based translation, machine translation), and for handling historical variation in texts (dictionaries) (see, e.g., Airio, 2006; Hedlund, Airio, Keskustalo *et al.,* 2004; Dolamic & Savoy, 2010; Gotscharek, Reffle, Ringsletter *et al.,* 2011). **Rule-based approaches**, such as transliteration, typically rely on rules learned from comparable corpora or from pairs of string variants. They are also typically language dependent techniques and require a notable initial effort of creating the rules: even if they rely on statistical approaches, language specific knowledge is regularly utilized (see, e.g., Pirkola, Toivonen, Keskustalo *et al.,* 2003b; Larkey, AbdulJaleel & Connell, 2003; Karimi, Scholer & Turpin, 2011). **Approximate string matching** techniques are language independent approaches to handling string level variation. Approximate string matching is used for measuring the similarity between strings, and identifying the

strings that are most similar to a source string. Its application in natural language information retrieval is based on the expectation that words that consist of similar strings of characters also have similar meanings (Robertson & Willett, 1998).

This thesis examines the use of approximate string matching methods for handling the string level variation occurring in alphabetic languages. The focus is on natural language information retrieval, particularly cross-language information retrieval (CLIR) and historical document retrieval (HDR); two fields where many string level variation phenomena need to be handled simultaneously. The general goal of this thesis is developing low cost, language independent query translation and expansion approaches for related languages and historical language variants. The focus is on applications of an approximate string matching technique called **classified *s*-gram matching** developed at the FIRE research group at the University of Tampere (Pirkola *et al.*, 2002; Keskustalo, Pirkola, Visala *et al.*, 2003). The classified *s*-gram matching technique has been experimentally shown to perform well in out-of-vocabulary (OOV) word translation, but its performance as a stand-alone query translation and expansion technique has not been studied. There have not been extensive studies on whether the *s*-gram settings should be adjusted for different languages and for handling different types of surface form variation: what *s*-gram classes, what kind of padding and which proximity measures should be used. These issues are considered in this thesis.

Generally, to reach high retrieval accuracy in CLIR and HDR, many resources for processing texts in the source and target languages and for translation are required. Stop word lists, stemmers, lemmatizers, or case form generators for handling morphological variation, compound splitters, and translation resources, and resources for handling the OCR errors created in the process of digitizing historical documents are needed. Obtaining and incorporating such resources is often time consuming and expensive, and therefore difficult particularly for smaller languages and historical language variants. As McNamee and Mayfield (2004) pointed out, incorporating language-specific resources for handling all string level variation that occurs in a cross-lingual setting may lead to very complex systems.

Against this background, approximate string matching techniques have the major benefits of being simple, available, and cheap solutions to the problems caused by string level variation. Because no language specific linguistic knowledge is usually utilized, approximate string matching techniques are generally easy to apply to new languages and to new string level variation types. Approximate string matching has previously been suggested as a solution to handling various string level variation types: handling morphological variation (McNamee & Mayfield,

2004; McNamee, Nicolas & Mayfield, 2009), spelling errors (Kukich, 1992), OCR errors (Harding, Croft & Weir, 1997; Amati, Celi, Di Nicola *et al.*, 2011), out-of-vocabulary query words in CLIR (Pirkola *et al.*, 2003b; Keskustalo *et al.,* 2003) and historical variants of query words in HDR (Robertson & Willett, 1993; Braun, Weisman & Sprinkhuizen-Kuyper, 2002). Therefore, one approximate string matching technique could potentially simultaneously handle all the different types of variation offering remarkable simplifications to CLIR and HDR systems.

Approximate string matching techniques are fuzzy techniques and inevitably introduce noise to the IR process. The most similar strings are not always variants of each other. For example, Loponen, Pirkola and Järvelin (2008) found that 2-6 closest matches needed to be inspected to find the correct English variants for French, German and Spanish query words. However, information retrieval is in general an application area that allows quite a lot of noise, and where even lower quality language processing can be very useful (Kettunen, 2013). Therefore, the relations of source and target language similarity, query translation and query expansion precision, and retrieval performance are studied in this thesis. What level of language similarity or query translation and expansion precision is required for approximate string matching to be a plausible stand-alone approach to information retrieval?

The goal of this thesis is to contribute with new empirical knowledge on where and how approximate string matching techniques may be applied to query translation and expansion. The application area of approximate string matching is broadened by introducing new areas where *s*-gram matching can function as a stand-alone approach for handling string level variation. The limits of the approach are also studied, through examining different language pairs in different domains, and how the domain affects the similarity of the language pairs and the translation and retrieval performance through it. McNamee and Mayfield (2004) have previously measured the similarity of several European languages based on parallel texts, and studied how the similarity affects *n*-gram based query translation. However, we hypothesize that the similarity of a language pair is not stable, but varies between different domains depending on typical search topics and domain vocabulary: the level of vocabulary shared by a language pair might be notably different for example in news text and in photo retrieval domains. This thesis also analyses how the classified *s*-gram matching technique can be adjusted and optimized to handling different string level variation types in natural language IR and how the different choices made when setting up the *s*-gram matching affect its performance on different language pairs.

Three different application areas are considered in this thesis: information retrieval between closely related languages (Study III), cross-language photo retrieval (Study IV), and historical document retrieval (Study V). The access to high quality translation resources is limited in all these domains. Extensive translation resources are often not prioritized between closely related languages where the speakers of one of the languages usually can understand the other. The user population of historical documents was until recently quite limited and digitized historical documents have only recently become accessible for the wider public. High quality translation resources do not exist for most historical languages. The extensive spelling variation occurring in historical texts makes the creation of comprehensive translation resources difficult, and the OCR errors littering the collections hinder effective normalization of morphological variation. Photographic image retrieval is another example of a domain where the coverage of the translation resources is a problem: it is a proper name dense domain, where both image annotations and user search queries very often contain proper names which usually are poorly covered in translation resources. Therefore, even if translation resources for the targeted language pairs exist, major portions of the central vocabulary of the domain will be untranslatable.

To summarize, this thesis aims to define the limits of approximate string matching as stand-alone approach to handling string level variation in CLIR and HDR. The focus is on classified *s*-gram matching, but also other string matching techniques were used. The question of which approximate string matching technique might finally be the best one for the different application areas is not answered in this thesis. Neither is the *s*-gram matching technique extensively compared to other fuzzy translation techniques including other approximate string matching techniques or rule-based translation, even if such comparison would be interesting. The issues addressed include: how the use of approximate string matching instead of standard query translation (Studies III and IV) and expansion (Study V) techniques affects retrieval performance; how similar do the source and target languages need to be for approximate string matching to be a sufficient translation approach (Studies II-IV); how domain affects language similarity (Studies III-V); and what kind and how extensive variation can be modeled by approximate string matching (Studies III-V).

The rest of this thesis summary is organized as follows: the central concepts related to cross-lingual spelling variation and approximate string matching (in IR) are first defined in chapter 2. Chapter 3 then discusses previous research related to approximate string matching in information retrieval covering proper name

matching, OOV word translation, handling morphological variation, OCR errors, historical variation, query translation in CLIR, and cross-language image retrieval. Chapter 4 presents the laboratory research methodology adopted in the studies included in the thesis. The individual studies are summarized in chapter 5. The reliability and validity of the studies are discussed and the results of the individual studies are brought together and discussed in chapter 6.

2 Defining the central concepts of the thesis

2.1 String level variation in natural language information retrieval

Natural languages are constantly evolving, complex and flexible, but however structured systems, with systematically recurring phenomena and regularities. Karlsson (2004) divides natural languages into five sub-systems: semantics, lexicon, syntax, morphology and phonology. Semantics is the sub-system related to meaning and lexicon the sub-system of lexemes; the inventory of words in a natural language. Syntax deals with the structure of sentences and morphology with the internal structure of words. Phonology studies the sound systems of languages. (Karlsson, 2004.) Written languages are regulated by orthographical rules defining how sounds are represented as graphemes (characters). This thesis focuses on string level phenomena in alphabetic languages, including word internal variation depending on the orthographical rules of the languages considered, morphological variation, historical (diachronic) variation, and variation due to errors in optical character recognition of digitized documents. We begin by discussing language change as a source of historical variation within languages and of cross-lingual spelling variation between related languages.

2.1.1 Related languages and language change

Two languages are said to be related, if they share a common language of origin, called a proto-language (Dahl, 2007). Language change within language families is often described in terms of vertical transmission of language, where a language divided in two, e.g., geographically or socially separated language communities gradually evolves towards two separate languages when passed on from generation to generation. According to Dahl (2007), the mechanisms of language change are however complex and the relations between languages cannot be explained by a hierarchical model of language families and inheritance only. Even language

external factors affect language change: linguistic features tend to spread in contacts between geographically or culturally close languages (*ibid.*).

In historical linguistics, it is often considered that changes in languages typically spread from cultural and economic centers towards periphery, both within a language and between languages (Dahl 2007): new features – most commonly vocabulary, but even grammatical features and pronunciation – are adopted from high prestige official or international languages to languages with lower status (*ibid.*). Consequently, European languages belonging to different language families share many words with common origin: borrowed from one language to another or to many languages from a common donor language, such as Latin, Greek or French, or more recently English. For example Finnish has through centuries borrowed many features from Latin and Swedish, including words and grammatical constructions especially in the written language (Häkkinen, 1994).

Loanwords form cross-lingual spelling variants between the languages, when they are adapted to the orthographical conventions of the recipient languages. The cross-lingual pairs of words that share the same etymological origin are often called *cognates*. Both the surface form and the meaning of cognates may differ between languages, due to changes occurring independently in the languages after the cognates entered the languages. Cross-lingual spelling variants in this thesis refer to the subset of cognates which share the same (or very close) meaning and can thus be usefully translated into each other in cross-language information retrieval context. "Hands-free", "roll-on", and "franchise" are examples of recent foreign loan words in Finnish that have retained the original spelling. "Viski" (*whisky*) and "parfee" (*parfait*) are examples of cases where the spelling has been more clearly adapted to fit Finnish pronunciation, inflection and orthography[1].

According to Dahl (2007) language change occurs in varying pace, with periods of relative stability and periods of rapid change, often depending on changes in the cultural or geographical surroundings. The process of language change involves both *invention* of new forms and *propagation* of those forms into the language during longer periods of time. During the gradual propagation of new features into a language, the change is often visible as variation in the language use: some people have adopted the new form, while others still use the old one (*ibid.*). The literary Finnish of the early 19th century is an example of a language in a period of rapid change. Häkkinen (1994) described how Finnish during this period of time became a broadly used written language and was consciously developed by adopting

[1] Examples from www.kotus.fi (The Institute for the Languages of Finland)

features from different dialects and by replacing foreign (mostly Swedish and Latin) vocabulary and constructions by native Finnish ones. The variation in vocabulary, inflection and spelling was wide and visible during this period (Häkkinen, 1994).

2.1.2 Orthographical variation

Orthographies are systems that define how sounds are represented in writing in a particular language. They include a set of graphic symbols (*graphemes*), such as signs, characters, letters and punctuation marks (depending on the type of the writing system, i.e., logographic, syllabic, alphabetic), and a set of rules for spelling and pronunciation, for spelling word boundaries, capitalization and punctuation (Seifart, 2006). All languages have their own orthographic conventions and there is wide variation even between related alphabetic languages in how sounds are coded into graphemes. Consequently, cross-lingual spelling variants may look rather different in different languages. Examples of cross-lingual spelling variants for a few European languages are given in Table 1. The orthographical variation in cross-lingual spelling variants is usually rather regular, and follow the spelling conventions of the languages. At string level, the variation can often be described in terms of single character substitutions, insertions and deletions, or combinations of them (Keskustalo *et al.*, 2003).

Table 1. Examples of cross-lingual spelling variants in five European languages.

	Finnish	**English**	**French**	**German**	**Swedish**
Person names	Dmitri Medvedev	Dmitry Medvedev	Dmitri Medvedev	Dmitri Medwedew	Dmitrij Medvedev
Place names	Bryssel Alpit Kiina Kreikka	Brussels Alps China Greece	Bruxelles Alpes Chine Grèce	Brüssel Alpen Kina Griechenland	Bryssel Alperna Kina Grekland
Technical terms	resessiivinen morfologia	recessive morphology	récessif morphologie	rezessiv Morphologie	recessiv morfologi
Popular words	chatata/tsätätä tomaatti	chat tomato	tchatcher tomate	chatten Tomate	chatta tomat

Historically, standardized spelling is a relatively new invention, introduced after the invention of printing, mass production of books and newspapers, and the introduction of grammars. English spelling became more or less fixed by the late 18th century (Robertson & Willett, 1993), while German spelling was standardized by 1902 (Ernst-Gerlach & Fuhr, 2007) and Finnish spelling was stabilized by the end of 1800s (Häkkinen, 1994). Previously, spelling was based on pronunciation and wide local (dialectal) and even personal variation occurred in spelling. The rate of graphical variants can therefore be high in historical texts.

Variation in orthography also occurs within languages as alternative spellings, i.e., there is not necessarily always one stable, correct way of spelling words. Alternative spellings may co-occur due to ongoing changes in spelling conventions or regional (national, dialectal) differences in a language (American vs. British English). There may be periods of active orthographical reforms and periods of relative stability, but the regulated written language is continuously changing, reflecting the changes in language use. For example, Finnish orthography has been relatively stable since the major reforms of the 1800s. However, some changes and instability constantly occur, recent examples in Finnish including the acceptance of the spelling "haltia" as equal with "haltija" in the meaning *elf* (but not in the meaning *posessor*); acceptance of both capitalized and lowercase spelling of the word "internet" ~ "Internet"; and acceptance of both spelling together and spelling separately the words "päinvastoin" ~ "päin vastoin" (*contrary*) and "itsestäänselvä" ~ "itsestään selvä" (*self-evident*)[2].

The spelling of multi-word expressions (regulated by orthography) is particularly interesting from an information retrieval perspective. Fluctuations in spelling compounds together or with a blank separating the words, as in the examples "päinvastoin" and "itsestäänselvä" above, occur in many languages (Karlsson, 2004). The orthographical conventions vary between languages: English and French are languages where spelling multi-word expressions as phrases is often preferred, while in German, Swedish and Finnish multi-word expressions are typically spelled together as compounds (Hedlund, 2002). In the compounding languages, compound formation is often productive in the sense that new compound expressions can be generated on demand by combining words. The share of compounds of the vocabulary of these languages is therefore usually high, and non-lexicalized or occasional compounds frequently occur in texts (*ibid.*). The problem from an information retrieval perspective is that it is not easy (for a user)

[2] Suomen kielen lautakunta, http://www.kotus.fi/index.phtml?s=280. Retrieved 31/3/2014

to guess what compounds are used in relevant documents. Therefore compound splitting is usually recommended for compounding languages (Alkula, 2001; Airio, 2006). From a cross-language information retrieval perspective, both phrases and compounds require special attention: phrases should be identified and translated as units to avoid losing their specific meanings. Occasional compounds in turn are rarely included in any translation resources, and should therefore be split and the constituents translated separately (Levow, Oard & Resnik, 2005).

2.1.3 Morphological variation

Morphology is concerned with the internal structure of words. *Inflectional morphology* refers to the use of morphological means to form inflected word forms of lexemes and affixes; *derivational morphology* to formation of new words from other words or root forms (Kettunen, 2009.) From an information retrieval perspective, inflected word forms are string level variants of the same lexeme should therefore be mapped together in indexing and retrieval. The degree of morphological variation varies between languages depending, e.g., on the role of morphology in expressing the grammatical relations between words: the morphological variation occurring in English is very limited (4 forms for a noun), while Finnish nouns can theoretically be inflected in over 2000 grammatical forms (Karlsson, 1983), even if most of them are too rare to be of practical importance in information retrieval applications (Kettunen & Airio, 2006). Table 2 illustrates the morphological variation in Finnish through a few examples of possible inflectional forms of the word "kissa" (*cat*).

Some technique for managing morphological variation is virtually always used in information retrieval, though the need for morphological processing depends on the morphological complexity of the language. The gap between results with no morphological processing and the best achieved results correlates with the (intuitive) morphological complexity of the language (Kettunen, 2009). Generally, morphological processing has quite a low impact on information retrieval effectiveness for languages such as English (Kettunen, 2009; Harman, 1991), but brings about major improvements in effectiveness for morphologically more complex languages such as Finnish, German and Turkish (Kettunen, 2009; Airio, 2006; Braschler & Ripplinger 2004; Can, Kocberber, Balcik *et al.*, 2008).

Table 2. Examples of inflectional forms for a Finnish noun. The singular and plural forms are given for the 11 most common case forms for the word "kissa" (*cat*), together with the first person possessive forms both for the singular and for the plural forms (e.g. nominative, singular possessive in first person, "kissani", means *my cat*.).

Case	Singular	Plural	SG + poss. 1. pers.	PL + poss. 1. pers.
Nominative	kissa	kissat	kissani	kissani
Genetive	kissan	kissojen	kissani	kissojeni
Partitive	kissaa	kissoja	kissaani	kissojani
Essive	kissana	kissoina	kissanani	kissoinani
Translative	kissaksi	kissoiksi	kissakseni	kissoikseni
Inessive	kissassa	kissoissa	kissassani	kissoissani
Elative	kissasta	kissoista	kissastani	kissoistani
Illative	kissaan	kissoihin	kissaani	kissoihini
Allative	kissalle	kissoille	kissalleni	kissoilleni
Adessive	kissalla	kissoilla	kissallani	kissoillani
Ablative	kissalta	kissoilta	kissaltani	kissoiltani

Morphology changes both due to causes internal and external to the language (Joseph, 1998; Dahl, 2007). The external causes refer to the transfer of morphological features from one language to another in a language contact. Virtually any morphological element can be transferred this way (Joseph, 1998). The internal changes include those triggered by regular sound changes, which may lead to once distinct case endings falling together, leading to additional changes to maintain the distinctions between the cases (Häkkinen, 1994; Joseph, 1998), and those triggered by analogic influence of one form over another. Morphological change in general is unpredictable. According to Joseph (1998), much of morphological change involves lexically particular developments, and is sporadic in its propagation - it does not necessarily spread to all candidates of the change i.e., the same change does not necessarily occur in all words with the same morphological feature. Older morphological forms may continue to occur in some lexemes or in particular semantic functions, such as the older plural form "brethren" of *brother*, which now is restricted to the meaning "fellow members of the church" (ibid.).

2.1.4 OCR errors

Digitization is a good way to preserve cultural heritage documents and make them widely accessible for researchers and the general public. Transforming the printed

cultural heritage collections into digital resources accessible and searchable through modern information and communication technologies requires that the digitized document images are transformed into digital text through OCR. OCR can currently reach over 99 percent accuracy in recognition of characters from high quality images of original documents with a simple book layout. However, OCR quality is dependent on environment and the condition of the original documents, and the accuracy is typically lower for historical documents. Generally, the older a document is, the lower the accuracy rate is likely to be. Holley (2009) reported raw character recognition accuracy rates varying from 71 percent to 98 percent in a sample of digitized newspapers from 1803-1954, the lowest rate indicating almost every third character being erroneously recognized and virtually all words containing errors. With an average word length of around eight characters, even 98 percent accuracy rate results in an error, on average, in every sixth word in Finnish texts. Digitized cultural heritage collections are therefore often riddled with OCR errors that hamper the performance of information retrieval systems.

OCR errors make it difficult to apply the conflation approaches developed for modern languages in indexing of historical texts: dictionary-based approaches (most lemmatization approaches) cannot handle words including OCR errors, because they are not included in their dictionaries; stemming algorithms will not recognize (and strip) word endings altered by OCR errors. Handling words containing OCR errors is consequently a major problem for information retrieval from historic document collections.

2.2 Approximate string matching in information retrieval

Navarro (2001) defined the approximate string matching problem to be "*to find the positions of a text where a given pattern occurs, allowing a limited number of "errors" in the matches*". Different application areas of approximate string matching have different error models for estimating how likely one string is to be an erroneous variant of another string, depending on the kind of variation typically occurring in the application area. Early application areas of approximate string matching included recovering original signals after transmission over noisy channels, finding DNA sequences after possible mutations or finding misspelled variants of words. (*Ibid.*)

Recognizing and measuring similarity in strings is useful in several situations in natural language information retrieval due to the natural morphological and diachronic word form variation and variation arising from, e.g., typing or OCR

errors occurring in databases. In cross-language information retrieval, approximate string matching can be used in translation of cross-lingual spelling variants. Especially related languages often share a number of cross-lingual spelling variants, proper names and technical terms being typical examples of words where such variation occurs. Hall and Dowling (1980) point out that the approximate string matching techniques need to be adjusted to the source and magnitude of the surface level variation. The precision of approximate string matching decreases, when the magnitude of variation increases (e.g., the similarity of the source and target languages decreases).

2.2.1 Phonetic matching

String similarity can be measured with a variety of different methods. **Phonetic matching** is used to identify strings that are pronounced similarly, regardless of their actual spelling (Zobel & Dart, 1996). They are generally language dependent methods, where the language specific orthographical rules for coding sounds into graphemes need to be accounted for. The first phonetic matching approaches were developed for English to cope with the spelling errors occurring when proper names are spelled based on their pronunciation. Soundex, developed in early 1900s by Odell & Russell (Hall & Dowling, 1980), is the best known phonetic matching algorithm and a commonly used baseline in approximate string matching studies in the field of information retrieval. In Soundex, strings are transformed into phonetic codes, where letters are replaced by numbers representing categories of similarly sounding letters (letters that might be mixed, based on how they sound) and all vowels are eliminated. The length of the phonetic codes is restricted to four characters (Hall & Dowling, 1980). Several variants of Soundex have been developed and tested, including Phonix developed by Gadd in 1990 (here: Zobel & Dart, 1996) and it's variant Phonix+ where the Phonix code was not truncated at four characters, but a minimal edit distance was used for finding approximate matches of the codes (Zobel & Dart, 1996); and a variant by McNamee *et al.* (2009) applied for 11 languages. Generally, neither the original Soundex algorithm nor its newer variants have particularly performed well, not in name matching (Zobel & Dart, 1996; Pfeifer, Poersch & Fuhr, 1996), and not in handling morphological variation in general (McNamee *et al.,* 2009).

2.2.2 Edit distance

The group of proximity measures called **edit distances** (or Damerau-Levenshtein metrics) are based on the idea that the distance between two strings can be measured as the minimum number of single character insertions, deletions, or substitutions (even transpositions) required for transforming one of the strings into the other (Navarro, 2001). The name Damerau-Levenshtein comes from the two pioneering authors who (separately) were the first to suggest the distance function (Levenshtein, 1966; Damerau, 1964). Dynamic programming algorithms, such as the Wagner-Fischer algorithm, were among the early efficient solutions for computing the minimum distances (Hall & Dowling, 1980). At the simplest, all edit operations are given an equal cost (e.g., 1). However, even different costs for the operations or specific character changes may be used, depending on the application area. For example, Zobel and Dart (1996) developed a phonetic distance measure Editex that combined edit distance with the letter groupings used in Soundex and Phonix: different edit costs were attached to different letter replacements, depending on their phonetic similarity. Kempken, Luther and Pilz (2006) used a machine learning algorithm for learning edit costs adjusted for the specific historical phenomena occurring in a historical document collection. Edit distances are widely used in a variety of applications as they are both efficient and deliver good results (Kempken *et al.,* 2006).

Longest Common Subsequence distance (LCS) measures the length of the longest pairing of characters that can be made between two strings, while retaining the order of letters in the strings (Navarro, 2001). It is then an edit distance variant, where only additions and deletions of characters are allowed. For example for the English–German word pair "motivation"–"motivierung" the longest common subsequence "motivin" has length 7. Similarity can then be measured, e.g., as the Longest Common Subsequence Ratio (LCSR), by dividing the length of the longest common subsequence by the maximum length of the strings compared (e.g. Kondrak, Marcu & Knight, 2003). For example, the LCSR(*motivation, motivierung*) = $7/11 \approx 0.64$. The Jaro algorithm is similar to edit distances and widely used in the field of record linkage (Christen, 2006; Bilenko, Mooney, Cohen *et al.*, 2003). It compares the number and order of characters shared between two strings, allowing the position of the characters in the strings to vary by half of the length of the shorter string ($\min(|s|, |t|)/2$). The Jaro-Winkler algorithm also uses the length of the longest common prefix between the strings to maximum length of 4 characters and increases the similarity score for strings that agree on initial characters. This

may be useful, because fewer errors typically occur at the beginning of words (Christen, 2006).

2.2.3 *N*-grams

***N*-gram** (or "*q*-gram") matching is another approximate string matching method that is often used in information retrieval applications (e.g., McNamee *et al.*, 2009; Pearce & Nicholas, 1996; Hedlund *et al.*, 2004; Pfeifer *et al.*, 1996). In the *n*-gram method, the strings that are compared are split into substrings of length *n*, i.e., *n*-grams, and the string proximity is measured based on the share of the strings' overlapping substrings of all of their unique substrings. The characters in the *n*-grams retain the order of the characters in the source strings. The substring length *n* usually varies between 2 to 7 in natural language applications. The higher values on *n* are particularly useful when word spanning *n*-grams are used (e.g., McNamee *et al.,* 2009). It is common to add extra padding-characters around the compared strings when forming the *n*-grams (Robertson & Willett, 1998). The padding helps getting the beginnings and the endings of the strings properly presented in their *n*-gram sets, and thus gives more weight to them in matching. Commonly, a padding of *n*-1 characters is used (*ibid.*). The *n*-gram based proximity comparisons have (in natural language applications) often been based on binary *n*-gram profiles, which only record the presence and not the number of occurrences of each distinct *n*-gram in a string (*ibid.*). The *n*-gram matching method can be further specified, by weighting the *n*-grams depending on their frequency (weight in inverse proportion to *n*-gram frequency), or by accounting for the position of the *n*-grams in the strings compared, e.g., by down-weighting *n*-gram matches from different parts of the strings (Brew & McKelvie, 1996; Robertson & Willett, 1998). Word beginnings may be simply weighted by using padding only at the beginnings of the strings compared.

2.2.4 Classified *s*-gram matching

N-grams are usually formed of adjacent characters, but even using non-adjacent characters is possible. For example Ullman (1977) used 4-grams of non-adjacent characters for spelling error correction and Brew and McKelvie (1996) used di-grams with skipping one character for finding verb cognates. Adjacent *n*-grams are in fact a special case of the more general *s-grams*, where characters may or may not

be skipped when forming the substrings of length *n* (*s* then refers to the word "skip"). **Classified *s*-gram matching** was originally developed by Pirkola *et al.* (2002). It is a further development of the *s*-gram matching technique, where a varying number of characters may be skipped when forming the substrings, and the *s*-grams are grouped into *classes* depending on the number of skipped characters. In the string proximity calculations, only the *s*-grams belonging to the same class are compared to each other. This way, different types of string level variation can be modelled. A gram class where one character is skipped, when forming the *s*-grams, is denoted with "{1}" (the set of *s*-grams formed with skip *k*=1), a gram class including *s*-grams formed by skipping both one and two characters is denoted "{1, 2}", and the traditional adjacent *n*-grams are denoted "{0}" (i.e., they are a special case of the *s*-grams, where zero characters are skipped). Examples of these gram classes are given in Figure 1.

Example string:	**Pariisi**	**Paris**
s{0}-grams with no skips	{pa ar ri ii is si}	{pa ar ri is}
s{1}-grams with skip length 1	{pr ai ri is ii}	{pr ai rs}
s{2}-grams with skip length 2	{pi ai rs ii}	{pi as}
s{1, 2}-grams with 1 and 2 skips	{pr ai ri is ii pi rs}	{pr ai rs pi as}

Figure 1. Examples of how *s*-grams are formed in different gram classes for strings "Pariisi" and "Paris". (The *s*-gram length *n*=2)

The gram classes can be further combined into more general gram classes. *Character Combination Index* (CCI) then indicates the set of all the *s*-gram classes to be formed from a string, and which *s*-gram sets should be compared to each other when calculating the *s*-gram proximity for strings. For example, CCI={{0},{1,2}} means that two *s*-gram classes are formed from a string: the gram class {0} with conventional *n*-grams and the gram class {1, 2} with *s*-grams formed by skipping both one and two characters. In the string proximity calculations then, the s{0}-grams of the compared strings are compared to each other and the s{1, 2}-grams to each other separately, and the final proximity value is based on a combination of the results of the two comparisons. This way the strings can be at once compared using several *s*-gram classes that model different types of variation. Figure 2 illustrates the comparison of strings based on gram classes.

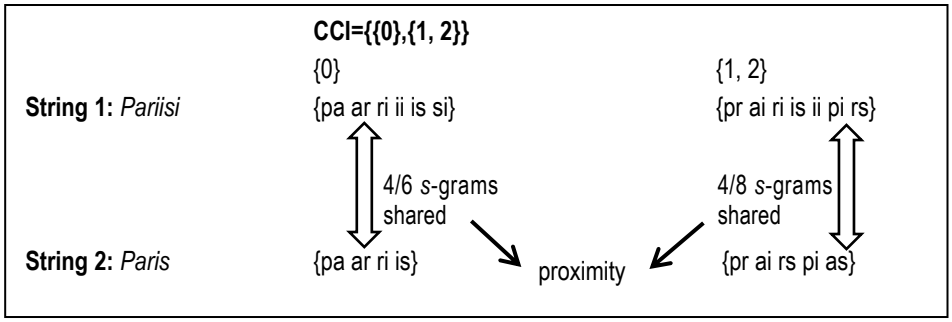

Figure 2. Example of how two strings can be compared using the classified s-grams, with CCI={{0},{1, 2}}

Different *s*-gram classes carry forward different evidence from their host string and *s*-grams can therefore be tuned to handle different phenomena by adjusting the *CCI*. Different lengths of substrings and skips are possible: Ullman (1977) and McNamee *et al.* (2009) used 4-grams with skips in one position. However, in all studies where multiple, classified *s*-grams have been used, the *s*-gram length has been set to two (n=2) (Pirkola *et al.*, 2002; Keskustalo *et al.*, 2003). Gram classes where the skip lengths vary between 0-2 have been found to model cross-language spelling variation well, because the variation often includes combinations of single character insertions, deletions and substitutions. Gram classes corresponding to frequently occurring cross-lingual variation patterns include: gram class {1}, which allows one substitution (as the substitution *z* → *c* between German and English "rezessiv" and "recessive"); gram class {0, 1}, which allows one insertion between adjacent characters, or deletion of characters separating two characters (as the insertion of *e* in French "Alpes" between the *ps* of the English "Alps") and gram class {1, 2}, which allows two character combinations of substitution and insertion (as the *f* → *ph* correspondence between "morfologia" and "morphology"), or single character insertions between characters separated by one character (corresponding e.g., to long vowels in Finnish, as *i* → *ii* between "recessive" and "resessiivinen"). For more examples of the correspondences between string level variation, skip lengths and *s*-gram classes, see Keskustalo *et al.* (2003).

Spelling errors and many common OCR errors can also be at string level described in terms of single character substitutions, insertions and deletions, or two character combinations of them (e.g., Zamora, Pollock & Zamora, 1981; Rice, Kanai & Nartker, 1993). Especially OCR errors may however introduce even more complex variation, such as substitutions, deletions or insertions of longer strings of

characters (e.g., substitution of *m* with *iii*) and erroneous wordbreaks due to insertion of punctuation and white spaces (Taghva & Stofsky, 2001). Historical variation in orthography and morphology also seems to include similar variation: substitutions of characters (using "w" instead of "v") and variation in marking the long vowels (e.g., "pispa", "piispa", "pīspa") being examples of differences between 19th century and contemporary Finnish (Häkkinen, 1994).

Using extra-padding characters around the compared strings is again useful for getting the beginnings and the endings of the strings properly represented in the *s*-gram sets. The suitable number of padding characters depends on the length of the substring (*n*) and the number of the skipped characters (*k*), and is typically defined as (*n*-1)(*k*+1). For example, when forming s{2}-grams (digrams skipping two characters) three padding characters are used to ensure that the characters at the beginnings and the endings of strings will be included in as many *s*-grams as the characters in the middle of the strings. Figure 3 shows the *s*-gram sets of Figure 1, but this time formed using padding on both sides of the strings.

Example string:	**Pariisi**	**Paris**
Padded s{0}-grams	{_p pa ar ri ii is si i_}	{_p pa ar ri is s_}
Padded s{1}-grams	{_p _a pr ai ri is ii s_ i_}	{_p _a pr ai rs i_ s_}
Padded s{2}-grams	{_p _a _r pi ai rs ii i_ s_}	{_p _a _r pi as r_ i_ s_}
Padded s{1, 2}-grams	{_p _a pr ai ri is ii s_ i _ _r pi rs}	{_p _a pr ai rs i_ s_ pi as r_}

Figure 3. Examples of *s*-gram sets formed using different gram classes for the padded strings "Pariisi" and "Paris". (The *s*-gram length *n*=2)

Padding may be used only at one side of the strings to give more weight to that side, while down-weighting the other. This allows adjusting the *s*-gram matching to different application areas and languages: for example, for handling morphological variation in suffixing languages such as Finnish, where mainly the word endings change in inflected forms, the use of left padding only might be useful to give more weight to the stable word beginnings.

Different proximity measures can be used for calculating string similarity based on the classified *s*-grams. Pirkola *et al.*, (2002) and Keskustalo *et al.*, (2003) used similarity measure based on the Jaccard's coefficient, where the number of intersecting *s*-grams in the *s*-gram sets of the compared strings is compared to the number of the unique *s*-grams in the *s*-gram sets of the compared strings. However, other proximity measures used in *n*-gram matching are also possible: e.g.,

the Hamming distance used by Zobel & Dart (1996), Dice's coefficient suggested by Robertson & Willett (1998), or L1 distance used by Ukkonen, (1992). Jaccard's coefficient, Dice's Coefficient and Hamming distance are all binary proximity measures, where only the presence and not the number of occurrences of each distinct *n*-gram in a string. This has been motivated with repetitions of *n*-grams being relatively rare in natural language word. (Robertson & Willett, 1998; Zobel & Dart, 1996). Non-binary measures such as the L1 distance are however more accurate where repetitions of *s*-grams occur. The number of *s*-gram repetitions occurring in gram classes combining *s*-grams with different skip lengths might be higher than in traditional *n*-grams.

N-gram matching is a commonly used string matching techniques, and the adjacent and non-adjacent *s*-grams have been used for many applications, including: spelling error correction (Hall & Dowling, 1980), retrieval from OCR error degraded texts (Harding et al., 1997), proper name matching (Zobel & Dart, 1996), language identification (Cavnar & Trenkle, 1994), cognate matching (Pirkola *et al.,* 2002; Keskustalo *et al.*, 2003), and normalization of morphological variation (McNamee & Mayfield 2004; McNamee *et al.*, 2009).

3 Previous research on approximate string matching in information retrieval

Approximate string matching and phonetic matching techniques have been suggested for many information retrieval related tasks including handling (monolingual) variation in names, handling morphological variation in mono- and cross-lingual IR, handling OOV words in CLIR or historical spelling variants of words in historic document retrieval, and handling errors in OCR scanned documents. Finally, a few studies have evaluated the use of approximate string matching in query translation between related languages.

3.1 Spelling variants in mono- and cross-lingual IR

3.1.1 Proper names

Proper names, including personal names, geographical names, and names of companies and products, are problematic both in monolingual and cross-lingual information retrieval. They are common and often central query words, and their successful processing is crucial for query performance (Pirkola, 1998; Christen, 2006). In monolingual context, the variation arising from typing and spelling errors, difficulties in translation and transliteration of foreign names, and cultural and historical traditions in how names are reported makes exact matching approaches inadequate (Borgman & Sigfried, 1992). Transliteration of foreign names to a host language using a different script is a creative process that allows creation of multiple valid variants of source terms: any phonetically reasonable representation of the source language word can be acceptable, and the acceptable transliterations vary between languages using the same script (e.g., Karimi *et al.*, 2011). For example, when foreign words are transliterated to Finnish, both Finnish and international transliteration systems might be used. In cross-language information retrieval, proper names are typically OOV words, which are not included in the translation resources and cannot therefore be translated to the target language.

Therefore, some other way of handling the cross-lingual spelling variation in proper names is needed.

Many phonetic and string similarity measures, and combinations of them have been suggested for name matching. The results from the studies are however somewhat indecisive: different measures have been included in different studies, and different techniques seem to have performed best depending on the implementation of the measures and the test data. For example, Zobel & Dart (1996) found that Editext performed better than edit distance, and edit distance better than di-grams in matching surname queries from Melbourne white pages, while Pfeifer *et al.* (1996) found that di-grams and even a Phonix variant outperformed Edit distance in matching surnames collected from both English and German language sources. Christen (2006) compared 27 variants of approximate string matching techniques using four datasets, of which one was the same dataset as in Pfeifer *et al.*, (1996). In their implementation, Edit distance outperformed both Phonix and di-grams, and different techniques performed best in each dataset. *S*-grams were included in the comparison, and performed worse than adjacent di-grams in all datasets, and worse than edit distances in the two datasets containing only surnames. However, the *s*-gram implementation is unclear. Different approaches performed best in the given name, surname and full name datasets, the best performing approaches being a LCS variant, the Jaro metric and a Jaro-Winkler version adjusted for handing name phrases where the order of the components could be swapped (Christen, 2006).

3.1.2 Morphological variants

Word form variation caused by inflectional morphology is one of the basic problems of text retrieval. Many approaches for handling inflectional variation have been suggested, including stemming algorithms and lemmatizers that return all inflected word forms into their base forms. The choice of the approach depends usually on the language: most work on English information retrieval has favored simple stemming algorithms, because they can sufficiently well handle the simple inflectional variation occurring in English (Kettunen, 2013). For languages with a more complex morphology, lemmatization based on full morphological analysis was long considered the best alternative (Alkula, 2001; Kettunen, 2004; Airio, 2006). However, even stemmers can improve the IR performance notably for languages that are morphologically rather complex such as Finnish (Airio, 2006;

Hollink, Kamps, Monz & De Rijke, 2004; McNamee & Mayfield, 2004) German (Braschler & Ripplinger, 2004; McNamee & Mayfield, 2004) and Turkish (Can *et al.*, 2008). Kettunen and Airio (2006), and Kettunen (2009) suggested that for practical information retrieval purposes, covering the most common inflectional forms is enough even for morphologically complex languages, because usually only a small share of the theoretically possible inflections occur frequently. Therefore, Kettunen and Airio (2006) developed a technique called Frequent Case form Generation (FCG), where instead of reducing all variants occurring in free text documents to their base forms, query words are expanded by their most frequent case forms.

McNamee and Mayfield (2004), McNamee *et al.*, (2009), Kettunen, McNamee and Baskaya (2010) and Kettunen (2013) compared simple, language independent and "non-linguistic" approaches to handling morphological variation, particularly in alphabetic languages. Their results have shown particularly *n*-gram indexing as an effective method for handling morphological variation. For example, McNamee *et al.* (2009) tested *n*-gram and *s*-gram indexing, keyword truncation and phonetic matching among other techniques. They showed that 4-5 character *n*-gram indexing performed better than stemming for most of the 18 languages considered. *S*-grams with single skips and gram length $n=4$ performed only slightly worse than the *n*-grams. Even automatic truncation performed almost as well as stemming for many languages (*ibid.*). Mustafa (2005) found hybrid di-grams, consisting of a combination of adjacent di-grams and *s*{1}-grams (one character skipped) to perform well for finding inflectional variants from Arabic texts. The use of *s*{1}-grams was motivated by the common use of one character prefixes in inflection of Arabic words. Kettunen (2013) recommended that heavier normalization methods, such as lemmatization or stemming, should only be used for handling morphologically complex languages. For other languages, simple methods such as five-character truncation and *n*-gram indexing work sufficiently well.

3.1.3 OCR errors

The effect of OCR errors in information retrieval has been studied since the early 1990. OCR accuracy from high quality source documents is generally high enough to not seriously affect information retrieval performance (e.g. Mitra & Chaudhuri, 2000; Taghva, Borsack & Condit, 1994). OCR quality is however strongly environment dependent: poor condition of the source documents, poor print or

paper quality, typefaces used and complexity of layout may all notably increase the OCR error rate. Therefore, particularly digitized historical text documents still suffer from poor OCR quality. High OCR corruption levels may lead to unstable retrieval performance, as the corruption interferes both with term weight calculations and with matching search keys to index terms (Beitzel, Jensen & Grossman, 2003; Mittendorf & Schäuble, 1996).

The two main approaches for handling OCR errors in information retrieval have been to either handle the errors in the collection during indexing through OCR error correction or *n*-gram indexing (Liu, Babad, Sun & Chan, 1991; Taghva *et al.*, 1994; Harding *et al.*, 1997; Amati *et al.*, 2011; Savoy & Naji, 2011), or at query time using approximate string matching for query expansion (Harding *et al.*, 1997; Amati *et al.*, 2011). The previous has often performed better of the two (e.g. Harding *et al.*, 1997). OCR error correction has the additional advantage of making the standard tokenization and conflations approaches viable. However, handling the errors at query time has the advantage that the processing of the database index can be avoided, when it is not possible or desirable.

N-grams have been one of the most frequently and successfully used solutions to retrieval of text corrupted by OCR errors. For example Harding *et al.* (1997) and Amati *et al.* (2011) used a combination of full words and their *n*-gram representations as indexing features; Savoy and Naji (2011) used 4-gram indexing only, while Liu *et al.* (1991) used frequency distributions of *di*-grams in OCR error recognition and correction. Harding *et al.* (1997) compared *n*-gram indexing to *n*-gram based query expansion. *N*-gram indexing performed better than *n*-gram query expansion, but also required much more storage space and computational resources as both the full words and a combination of several length *n*-grams were indexed. The *n*-gram query expansion results improved when the number of *n*-gram variants added to the queries was increased suggesting that using a looser *n*-gram distance threshold and thus adding more words to the expanded queries might further improve the approach. (Harding *et al.*, 1997.)

3.1.4 Historical variants

In addition to the OCR errors resulting from the digitization process, historical document retrieval is complicated by historical vocabulary change and historical variation in spelling. Studies on historical information retrieval have most often only focused on the historical spelling variation, while handling OCR errors and

vocabulary change have more rarely been considered. However, Pilz, Luther, Fuhr and Ammon (2006) included rules for handling OCR errors in their historic information retrieval system, and Hauser, Heller, Leiss *et al.* (2007) and Gotscharek *et al.* (2011) have presented approaches for supporting the construction of historical lexica.

The two main approaches to handling historical spelling variation have been approximate string matching and rule-based transliteration approaches. Robertson and Willet (1993) studying 17th century English, O'Rourke, Robertson and Willett (1997) studying 12th century French, and Braun *et al.* (2002) studying 16th and 17th century Dutch all tested *n*-gram matching, and the Wagner-Fischer algorithm for finding historical spelling variants of modern query words from the target document collections. Kempken *et al.* (2006) used an edit distance variant where the edit costs were automatically learned from a set of modern–historical word pairs. They concluded that algorithms that are trained on the specific historic phenomena of the collection can reach a better translation recall and precision than simpler string matching methods such as edit distance and *n*-grams (Kempken *et al.,* 2006).

Transliteration rules for rewriting modern spellings as their historical variants (or vice versa) have been constructed manually by Gotscharek *et al.* (2011) and automatically by Koolen, Adriaans, Kamps and De Rijke (2006) and Ernst-Gerlach and Fuhr (2007). Koolen *et al.* (2006) used a combination of a phonetic matching, and relative frequencies of consonant and vowel sequences and of *n*-gram sequences in a historic and a modern corpus for generating transliteration rules for historical Dutch. Ernst-Gerlach and Fuhr (2007) constructed transliteration rules for German from a modern–historical word pair list by recording the transformations needed for rewriting the modern word as the historical variant. Pilz, Ernst-Gerlach, Kempken, *et al.* (2008) showed that automatically generated rules for historic variant generation can rather well reproduce manually generated gold standard rules and also capture variation that is not discovered by the manual rules. They also suggested that a transliteration model trained on historical German might perform rather well on transliteration of old English.

3.1.5 Out-of-vocabulary words

The most common translation resources used in CLIR are machine readable dictionaries, (Hedlund, 2002; Hull & Grefenstette, 1996), machine translation

systems (McCarley, 1999; Dolamic & Savoy, 2010) and statistical translation resources induced from parallel or comparable corpora (Darwish & Oard, 2003; McNamee & Mayfield, 2004; Talvensaari, Laurikkala, Järvelin & Juhola, 2007; Sheridan & Ballerini, 1996). The coverage of these translation resources varies, but is always limited as new terms, such as technical terms, non-lexicalized compound words, proper names, acronyms, and abbreviations are continuously introduced to languages. These out-of-vocabulary terms can severely degrade the retrieval effectiveness of a CLIR engine, especially when the source queries being translated are very short. (Pirkola, Keskustalo & Järvelin, 2001; Zhou, Truran, Brailsford *et al.*, 2012)

Proper names and technical terms are often spelling variants in different languages, and therefore approaches based on string similarity can often be successfully used for handling their translation. The simplest approach to handling OOV words in CLIR has been to pass them to the target language as they are, in the hope that they will match the target language cognates (Pirkola *et al.*, 2001; Kishida, 2005). This can often work especially for proper names between languages that use the same script. An approach with a better coverage of the different types of OOV words and language pairs is to use approximate string matching to find similar, but not identical target language spelling variants of the source words. Pirkola *et al.* (2002) and Keskustalo *et al.* (2003) developed the classified *s*-gram matching technique for handling cross-language spelling variants between several European languages. They showed that the cross-lingual spelling variation could be modeled using single character insertions, deletions, substitutions and combinations of them and therefore considered the problem similar to e.g., spelling correction and name matching. The classified *s*-grams were therefore compared to edit distance, LCS and traditional adjacent *n*-grams in cross-language spelling variant translation, and outperformed them all. Recently, Montalvo, Pardo, Martínez and Fresno (2012) studied a combination of 3-grams (called TRI-SIM, including partially matching tri-grams), positional di-grams formed of adjacent characters and by skipping one character (called XXDICE, by Brew & McKelvie, 1996) and Longest Common Subsequence Ratio (LCSR) in cognate matching between Spanish and English and French and English cognates. They reported that XXDICE and adjacent di-grams (using Dice for proximity measure) were the best performing individual techniques, while the combined approach further improved the results achieved by any of the individual techniques, or other baseline classifiers for combining the metrics. (Montalvo *et al.*, 2012.)

Another approach to OOV word translation in CLIR is machine transliteration. Pirkola *et al.* (2003b), Toivonen, Pirkola, Keskustalo *et al.* (2005) and Loponen *et al.* (2008) studied transliteration of typical OOV words between several European languages using transformation rule based translation (TRT). Udupa, Saravanan, Bakalov and Bhole (2009) studied transliteration of OOV words in CLIR between English and Hindi, and English and Tamil. Klementiev and Roth (2006) studied English-Russian transliteration; and Al-Onaizan and Knight (2002), AbdulJaleel and Larkey (2003), and Sherif and Kondrak (2007) Arabic-English transliteration. Machine transliteration is particularly useful between languages that use different scripts (or writing systems), where direct approximate string matching is not possible. The goal is to adapt words from one script to another, so that the target language spelling of the words represents the sounds of the source language words correctly. Unlike approximate string matching, transliteration techniques are typically language dependent: even if they rely on statistical approaches, language specific knowledge is regularly utilized in the form of training data or bilingual pronunciation dictionaries (Karimi *et al.*., 2011). Character (or substring) correspondences are usually learned from a set of bilingual variant pairs or from aligned corpora.

Transliteration algorithms generate usually many theoretically correct transliteration candidates for each query word. The usefulness of these transliterations in CLIR mainly depends on their frequency in the target document collection. Therefore, different approaches for selecting the best transliterations from a set of generated transliteration candidates have been suggested, based on temporal alignment (Klementiev and Roth, 2006); finding approximate matches of the candidates in the target document collections (Pirkola *et al.,* 2003b; Toivonen et al., 2005); and relative frequencies of the source language word and the transliteration candidates (Pirkola, Toivonen, Keskustalo & Järvelin, 2006). Udupa *et al.* (2009) mined correct transliterations for OOV query words from among the initial retrieval results of the query.

Loponen *et al.* (2008) compared a transformation rule based translation approach using frequency based identification of translation equivalents (FITE-TRT, Pirkola *et al.*, 2006) to *s*-gram matching in OOV word translation, in a scenario where only one correct translation existed for each query word. FITE-TRT was shown to reach a clearly higher recall and precision when only the first translation candidate was considered. On average, 2-6 closest *s*-gram matches needed to be considered to reach the recall of FITE-TRT, i.e., a correct English translation found for ≈70 % of French, German and Spanish query words.

However, Larkey *et al.* (2003) showed that sometimes more than one transliteration (cross-lingual spelling variant) is needed. In their study up to 20 transliterations for each name in a translated query was needed to reach the best results (Larkey *et al.*, 2003). This was the effect of the several possible transliterations of Arabic names in English: for best retrieval results all the possible transliterations needed to be covered, instead of the one best. (Larkey *et al.*, 2003.) Similar effects can be expected in other situations, where many cross-lingual spelling variants, morphological variants, OCR variants and historical variants may occur.

3.2 Fuzzy translation in CLIR

The goal of cross-language information retrieval is to help searchers find documents that are written in languages that are different from the language in which their query is represented (Levow *et al.*, 2005). Cross-language information retrieval is based on translation: the queries, the documents or both of them need to be translated to make word matching based information retrieval techniques applicable. Query translation has received more attention in CLIR literature, mainly because document translation is computationally and in terms of storage space much more demanding than query translation, especially if many languages are to be covered (Kishida 2005; McCarley 1999). A recent survey of translation techniques in CLIR is given by Zhou *et al.* (2012).

Buckley, Mitra, Waltz and Cardie (1998) were possibly the first to rely entirely on cognate matching in CLIR: they treated English queries as misspelled French and used a French spelling correction program for transforming the English cognates to their French spelling allowing for deletions and insertions of characters and a few intellectually developed rules for character transformations. Their results were promising, and it was estimated that around 40 % of the vocabulary were identical matches or close enough cognates to be translated using the simple cognate matching approach. (*ibid.*).

McNamee and Mayfield (2004) and (2005) extended the use of *n*-gram indexing to corpus based translation, and showed how translation knowledge can be derived from parallel corpora using *n*-grams instead of words. Vilares, Oakes and Vilares (2007), and Chew, Bader and Abdelali (2008) applied this approach to English-Spanish and English-Arabic CLIR, respectively. While McNamee and Mayfield (2004, 2005) used word spanning *n*-grams of single *n*-gram length (4 or 6), Chew *et al.* (2008) used word internal and non-overlapping *n*-grams of varying lengths up to

a maximum length: morphologically insignificant *n*-grams were discarded and a single tokenization consisting of one, two or more *n*-grams of different lengths, was selected for each word (e.g., "comingle" might be tokenized to *co*+*mingle*). For most languages the optimal maximum length was $n \leq 9$, while for Arabic (where written words tend to be shorter), the optimal maximum length was $n \leq 6$ (Chew *et al.*, 2008). This approach worked even for translation between unrelated languages, as long as comparable corpora were available (*ibid.*).

McNamee and Mayfield (2002, 2004) also examined a language independent "no translation" approach to CLIR between several European languages which is particularly interesting from the point of view of this thesis. They used word overlapping 4-grams and 6-grams as indexing features, and simply matched target and query language *n*-grams to find approximate cognate matches. They reported results doubling the performance as compared to raw word cognate matching, 4-grams performing best for most language pairs, except for Finnish where 6-grams performed better. McNamee & Mayfield (2002) reported a performance level of around 33-60 % of the monolingual runs for the 6-gram translation, depending on the language pair. McNamee (2008) further tested using various *s*-grams for "no translation" CLIR. Results comparable to using *n*-grams were achieved, but with a much higher computational cost due to the higher redundancy of *s*-gram text representation (McNamee, 2008). Makin, Pandey, Pingali, and Varma (2007) experimented with Levenshtein edit distance, Longest Common Subsequence Ratio (LCSR) and Jaro-Winkler's similarity score for cognate matching between the Indian languages Telugu and Hindi. These Indian languages usually share many words with Sanskrit, Persian or English origin, adjusted to the script and morphology of the specific languages. Query translation relying solely on the cognate matching (particularly Jaro-Winkler's) performed better than dictionary-based query translation alone; combining cognate matching and dictionary-based translation further improved the results, reaching 50 % of monolingual performance. (Makin *et al.*, 2007.)

In McNamee and Mayfield's (2004) study the intuitive relatedness or similarity of the language pairs seemed to have a major effect on the *n*-gram retrieval performance. They calculated the relatedness of the language pairs based on the Bendetto language pair distance (Bendetto, Caglioti & Loreto, 2002) using the United Nation's Universal Declaration of Human Rights in each language as data. They found that the language pair distance was clearly predictive of the *n*-gram translation retrieval performance. This seems natural, as success of approximate

string matching entirely depends on the existence of spelling variants between the source and target languages.

Fuzzy translation is a particularly promising approach to query translation between closely related languages. For example, the Scandinavian languages Swedish, Norwegian and Danish form a language group where the share of identical words and cross-lingual spelling variants of the vocabulary is high. Some 50% of the Swedish and Norwegian (Bokmål) vocabulary is identical (when inflected word forms and orthographical differences of using æ/ø instead of ä/ö are not considered) and another 40% similar (Barðdal, Jörgensen, Larsen & Martinussen, 1997). The Scandinavian languages are comprehensible for speakers of any one of the languages, especially in written form – to such an extent that e.g., the Nordic council has chosen to interpret the three languages as a single Scandinavian language and does not provide translations of its resources in one language in the others (Dalianis, Rimka & Kann, 2009). However, major difficulties have been reported for employees and visitors to find information from the Nordic council's website (ibid.). This demonstrates that active language skills are needed for query formulation in CLIR.

3.3 Cross-language image retrieval

While image retrieval research focuses more and more on visual retrieval and the use of metadata (e.g. in recent editions of ImageCLEF, http://www.imageclef.org), text-based retrieval remains the most common way users search for photographic images: text-based retrieval still yields better performance levels than content-based image retrieval, and more importantly, users seem yet unconvinced of the benefits of initiating their queries with an example image, drawing, or the like. Keyword queries are preferred to using example images and browsing categories for image retrieval. (Yoon, 2011, Menard & Khashman, 2014).

Image retrieval has often been considered a realistic scenario for cross-language information retrieval, due to the language independent nature of images: they can be relevant to users even when the texts related to them are written in a language that users understand only poorly (or not at all). Image collections are also often multilingual in their nature (e.g. Flickr[3], or the Web as a collection), and particularly non-English-speaking users of image retrieval systems seem to regularly search for

[3] https://www.flickr.com

images using keywords in a language other than their first language (Menard & Khashman, 2014). Supporting text-based image retrieval in multiple languages seems therefore important.

Photographic image retrieval somewhat differs from text document retrieval in the types of search topics, queries and the texts associated with images that can be used in text-based retrieval. Photo annotations and queries contain names of people, places and things more frequently than typical text documents and corresponding queries. Image queries often have unique targets such as specific persons, places or constructions of which there only is one ("London Bridge") (Pu, 2005; Tsikrika, Popescu & Kludas, 2011; Yoon, 2011). For example, Pu (2005) found that 78 % of frequent Web image queries[4] had specific targets, as compared to only 34 % of Web text queries. In addition to known persons, queries related to plants, animals and nature, and medical information are more common than in text retrieval ("rhododendron" or "chickenpox") (Pu, 2005; Yoon, 2011). This means that a large part of the central query words are out-of-vocabulary in the common translation resources, but also that a larger share of that vocabulary is likely to be translatable using approximate string matching.

Translation of textual queries has not been a major focus for image retrieval research recently. The best retrieval results are usually achieved by systems combining text retrieval and visual features. For example ImageCLEF 2012 and 2013 did not even include tasks where text-based multilingual retrieval was in focus. In ImageCLEF 2011 Wikipedia task, typical Web search was simulated using a heterogeneous collection of images, associated Wikipedia articles and user-provided image annotations, when available. Three query languages were offered, but no results for translated textual queries were reported (Tsikrika *et al.*, 2011). Previously, standard CLIR approaches have been used for query and document translation in image retrieval (e.g., O'Hare, Wilkins, Gurrin *et al.*, 2009 and Ruiz, Chen, Pasupathy *et al.*, 2010 used Google translate), and results comparable to monolingual multimodal image retrieval have been achieved. The combination of text and visual retrieval reduces the problems caused by the "language barrier". Arni, Clough, Sanderson & Grubinger (2009) even concluded that "… the language barrier is no longer a critical factor in achieving good retrieval results" when discussing the results of the CLEF2009 photo retrieval task where the annotations were randomly either in English or in German.

[4] from the VisionNext search engine search log, www.visionNEXT.com

3.4 Summary of the previous research results

Approximate string matching techniques have been shown to be reasonable solutions to handling morphological variation and monolingual, historical and cross-lingual spelling variation. There's no conclusive evidence of one approach being crucially better than the others. Rather, the results have varied from one application area and test setting to another. In recent information retrieval studies, McNamee and colleagues (2004, 2009) have shown that using longer character *n*-grams as indexing features is a particularly useful approach to handling morphological variation, when extensive linguistic resources are not available. This thesis however focuses on query processing. The focus is on cross-lingual information retrieval and handling cross-lingual variants, rather than morphological variation per se.

There are several approaches to fuzzy translation of query words that could be adopted. Studies focusing on cross-language information retrieval have shown that classified *s*-gram matching outperforms many of the standard approaches, such as edit distances and *n*-grams in finding cross-lingual spelling variants of query words (Pirkola *et al.*, 2002; Keskustalo *et al.*, 2003). Transliteration techniques, such as the FITE-TRT, have been shown to have clearly higher translation precision than the classified *s*-grams and are probably the recommended approach to OOV word translation, where precise one-word translation is required (Loponen *et al.* 2008). However, in many applications finding *many* spelling variants may be required: the cross-lingual spelling variants of OOV words may occur in inflected word forms, there may be several possible transliterations of foreign names and digitized historic documents typically contain dozens of surface form variants of words, due to OCR errors, non-standard spellings, and historical changes in languages. It seems likely that the performance differences between approximate string matching techniques and transliteration would decrease under such circumstances. Such comparisons have however not been made, to our knowledge. We do one comparison in Study III, but using the old version of the transformation rule based translation program (Toivonen *et al.*, 2005) that is still relying on *n*-grams.

Approximate string matching techniques (including classified *s*-gram matching) have the benefit of being entirely language independent. Unlike for the transliteration approaches, no language specific training data or linguistic knowledge is needed, and *s*-grams are therefore readily applicable to new languages. Therefore, *s*-gram matching seems like a promising approach to query translation

in situations where resource-lean translation approaches are required; and where query expansion with several surface form variants is desirable.

A few studies have suggested that approximate string matching might be a sufficient approach to translation in cross-lingual information retrieval between related languages. It has been identified that the relatedness of the source and target languages is crucial for the performance of such fuzzy translation. Loponen *et al.* (2008) showed that the performance of *s*-gram matching degraded fast, when the similarity between the source and target words decreased. Their results suggested that *s*-gram matching can perform reasonably well when the source and target word similarity is above the LCS ratio of 0.8. In McNamee and Mayfield's (2002) study, the CLIR performance varied depending on the relatedness of the source and target languages: between the closely related Portuguese to Spanish 64 % of monolingual performance was reached, while only 33 % of the monolingual performance was reached when Finnish *n*-gram queries were used for retrieval from a Spanish collection (*ibid.*).

There is however a difference between the general relatedness of two languages and the similarity of those two languages in specific domains, as measured by the share of cross-lingual spelling variants of the vocabulary or the orthographic similarity of the cross-lingual spelling variants. The (linguistic) relatedness of two languages is somewhat stable, but their similarity varies depending on the language use in the different domains. Therefore, language similarity as measured based on some unrelated general text corpus is not perfectly informative of the language similarity in any specific document collection or domain. From an information retrieval perspective, two languages are likely to be more similar to each other on a proper name dense domain such as photographic image retrieval than in general news domain, for example. In the domain of historical document retrieval, the similarity varies depending on the age of the documents. For example, Gotscharek *et al.* (2011) estimated that 19th and 18th century German was similar enough to modern German for transliteration rule based translation to lead to an acceptable translation precision and recall, while for older German a combination of approximate string matching and historical lexicon was needed (due to vocabulary change).

We suggest that fuzzy query translation may work well for even less closely related languages on specific domains. This thesis aims to increase the understanding of the domain-to-domain variation in language similarity and identify domains of relatively high language similarity. Approaches to estimating the potential usefulness of the *s*-gram query translation are also considered.

Suggesting a single threshold value for language pair similarity seems undesirable, as the acceptable performance level probably depends on the availability of alternative techniques for handling variation. It would however be useful to be able to estimate, based on the language pair similarity, the potential of the *s*-gram query translation/expansion to improve results over no-translation, or match the results of the available state-of-the art linguistic tools.

4 Data and methods: test collection based evaluation of information retrieval systems

Table 3 summarizes the test collections and the methods used in the studies of this thesis. The data, resources and methods are then discussed briefly in the following chapters.

4.1 Test collections and language processing

Three different test collections were used in the studies included in this thesis, as shown in Table 3. The test collections used in Studies I-IV were acquired from the Cross Language Evaluation Forum, CLEF. In Study V, a test collection created at the University of Tampere for evaluation of information retrieval from historical document collections was used. The creation of the historical test collection was not a part of this thesis.

In Studies I and III, we utilized the Swedish test collection from CLEF 2003, containing 142 819 newswire articles (352 MB) from the Swedish news agency Tidningarnas Telegrambyrå and the 54 search topics (out of the 60 created for CLEF 2003) which had relevant documents in the Swedish collection. A set of Norwegian topics was created by translating the English search topics into Norwegian by a native Norwegian speaker. The relevance assessments in the collection are on a binary scale. The CLEF 2003 newswire document collections were also used in Study II. Target word lists (TWLs) were created from the indexes of the English, Finnish, German and Swedish collections. The German collection was much larger than the other collections and therefore only a part of it was used in the TWL. The final TWL sizes varied between 257 000 (English) and 535 000 (Finnish) unique word forms. A set of 271 cross-lingual spelling variants in seven languages (English, Finnish, French, German, Italian, Spanish and Swedish) was used as source words. Each source word had exactly one correct spelling variant in each of the four target language word lists.

Table 3. Resources & languages used in the studies

Resources	Study I	Study II	Study III	Study IV	Study V
Test collections	CLEF 2003 Swedish newswire collection	Test words; CLEF 2003 document collections (word lists)	CLEF 2003 Swedish newswire collection	ImageCLEF Photo 2007 collection.	Historical Finnish Newspaper test collection
Source languages	Norwegian, Swedish	English, Finnish, French, German, Italian, Spanish, Swedish	Norwegian, Swedish	Danish, English, French, German, Norwegian, Swedish	Modern Finnish
Target languages	Norwegian, Swedish	English, Finnish, German, Swedish	Norwegian, Swedish	English, German	Early modern Finnish
Morphology	TWOL lemmatizers	TWOL lemmatizers	TWOL lemmatizers	TWOL lemmatizers	FCG
Fuzzy translation methods	*n*-grams, *s*-grams	*n*-grams (di & tri), *s*-grams	*n*-grams, *s*-grams, TRT	*s*-grams	*s*-grams
Translation dictionaries	GlobalDix	—	GlobalDix	GlobalDix	—
Query structure	#SUM, #SYN	—	#SUM, #SYN, #UW7	#SYN	#SYN, (#combine)
Retrieval system	INQUERY	—	INQUERY	Lemur Indri	Lemur Indri
Evaluation measures	MAP	MRR	MAP, P@ Recall levels	MAP, P@K	MAP, P@K nDCG

The IAPR TC-12 photo collection (Grubinger, Clough, Müller & Deselaers, 2006; Grubinger, Clough, Hanbury & Müller, 2007) from ImageCLEF Photo 2007 was used in Study IV. The collection contains 20 000 photographic images including pictures of people, animals, cities, landscapes, sports and other actions, and 60 search topics. The collection was provided by a travel company Viventura. The photos are taken by travel guides working for Viventura, who regularly photograph the tours organized and provide customers with access to the photos from the tours they participated. The topics were mainly developed based on a log of previous customer searches, but even topics for testing various aspects of text and visual retrieval were included (e.g., "Black and white photos of Russia"). The topics contain a clearly lower share of proper names than has been reported common for Web photo search, which makes the topics more difficult from *s*-gram translation perspective. Each image is associated with a light annotation containing an image title, date, the location at which the photograph was taken and additional notes concerning the image. We used English and German annotations, and Danish, English, French, German, Norwegian and Swedish topics. Only the topic titles were used. In addition, each topic included three example images that were used for visual retrieval. The collection offers relevance assessments on a three point scale, but only binary relevance was used in Study IV. The collection is described in detail in Grubinger (2007).

The historical Finnish newspaper test collection used in Study V consists of a subset of the historical newspaper archive of the Finnish National library[5] and contains digitized Finnish newspapers from the years 1829-1890. The collection contains 180 468 documents (84 512 newspaper pages, 772 MB) (Raitanen, 2012), and 56 search topics related to 19th century history, together with relevance assessments on a four point scale. The topics are written from a contemporary perspective using contemporary language. The topics model relatively broad, topical information needs and are typically considered with historical events, institutions or persons or with the attitudes or wider developments in the society during the 19th century. The assessment pools for relevance assessments were created as a part of the topic creation process, independently of the experimental runs. The topics were created based on extensive test searches. The assessment pools for each topic were created based on an extensive conceptual analysis of the topics and inclusive query plans, following Sormunen (2000). The final queries

[5] http://digi.lib.helsinki.fi/sanomalehti/secure/main.html.

used for pool creation covered various letter, word and term variants as well as all the facets of the topic at hand.

4.1.1 Morphology

In all studies where dictionary based translation was used the TWOL lemmatizers by the Lingsoft Ltd. were used for transforming word forms in documents and queries into their basic (dictionary) forms. The only exception is the French queries in Study IV, which were lemmatized manually because a lemmatizer was not available. This might lead to somewhat higher lemmatization quality for the French queries. Compounds were always split before the dictionary based translation, and the constituents were translated separately (Studies I and III), or the constituents were separately translated only when there was no translation available for the complete compound (Study IV). In Studies II and III lemmatized indexes and queries were used even for the fuzzy translation. However, in Study III compounds were not split in the *s*-gram queries as Swedish and Norwegian compounds were expected to be similar. In later studies (I, IV and V), the document and query words were not normalized prior to *s*-gram translation to better adhere to the goal of developing low cost translation resources not relying on extensive linguistic analysis. In study V, where handling even morphological variation using the classified *s*-gram matching was an explicit goal, *s*-gram query expansion was compared to and combined with the Frequent Case form Generator (FCG) by Kettunen and Airio (2006). The goal was to improve the coverage of the inflected word forms in the expanded *s*-gram queries.

4.1.2 Translation methods

The main approach to translation in this thesis is to use the classified *s*-gram matching for translating query words to a set of their closest matches in the target language. Different *s*-gram settings have been tested and evaluated in the different studies; the settings are summarized in Table 4. Study II contained the most extensive comparisons between the different CCI's, paddings and proximity measures: translation candidates were generated using twelve different CCIs, three padding alternatives, and six proximity measures.

Table 4. The *s*-gram settings tested in the different studies.

	Study I	Study II	Study III	Study IV	Study V
CCI	{{0},{0,1},{1,2}} {{0},{1,2}}	{{0}} {{0},{0,1}} {{0},{0,1},{1},{1,2}} {{0},{0,1},{1,2}} {{0},{1}} {{0},{0,1},{1}} {{0},{1},{1,2}} {{0},{1,2}} {{0,1}} {{0,1,2}} {{1}} {{1,2}}	{{0},{1}} {{0},{1,2}}	{{0},{1,2}}	{{0},{0,1},{1,2}} {{0},{1},{1,2}} {{0},{0,1},{1}, {1,2}}
Padding	Left	no, left, left & right	Left	Left	Left
Proximity measure	Jaccard	L1, Tanimoto, Cosine similarity, Hamming distance, Jaccard, Dice	Jaccard	Jaccard	Dice
No. *s*-gram variants in queries	3	N/A	4	3	2-60

The classified *s*-gram translation was compared to a few other translation techniques in the different studies: Dictionary-based translation was used as a baseline in Studies I, III and IV and *n*-gram translation was used in Studies I, II and III. In Study III, the transformation rule based translation approach TRT was also tested, both alone and combined with *n*-gram matching. In the latter case, source language words were first transformed closer to their target language variants using the statistical transformation rules. Then, the transformations were matched against the target language index to find the closest variants actually occurring in the document collection. In Study V, query expansion using *s*-grams and FCG was compared to non-expanded queries and to the pseudo relevance feedback feature of the Lemur Indri search engine.

4.2 Query formulation

Much of the work in this thesis winds around the questions of where query words are attained from and of how queries are formulated, in situations where the original query words need to be translated (or transformed) to better match the words representing the relevant documents in the collection. Translation in CLIR is associated with problems related to ambiguity and loss of meaning. The correct senses of phrases are often lost if the phrases are not identified and kept intact in translation and in target language queries; alternative translations in target language increase ambiguity; occasional compounds are rarely covered in translation resources and cause serious coverage problems in compounding languages. Consequently, compounds are often split and the constituents translated separately. As each constituent may be ambiguous and get several possible translations, problems with loss of meaning similar to those in phrase translation are frequently encountered. (Ballesteros & Croft, 1997; Pirkola, 1998; Hedlund, 2002; Hedlund *et al.*, 2004.)

Similar problems occur when translating or expanding queries using approximate string matching: the fuzziness of the approach necessarily adds ambiguity and noise to the translated queries. The suitable number of closest matches to use is likely to depend on the target phenomena, the languages involved and the precision of the approximate string matching technique used. The correct match is not always the most similar string, and therefore adding a few closest matches increases the chance of including the correct match (e.g., 2-6 closest matches in Loponen *et al.*, 2008). Hedlund *et al.* (2004) reported that adding just two closest *n*-gram matches of Finnish query words to the translated English queries yielded the best balance between translation recall and precision in OOV word translation in cross-language information retrieval.

Often however, more than one correct variant exists: if the target language index is not conflated it might contain more than one inflectional variant of the target language spelling variant. Mutually interchangeable spelling variants or alternative historical variants might occur, as well as different OCR errors and combinations of these phenomena. In such situations, adding a higher number of approximate matches to the target language queries might be useful to cover a larger share of the relevant variants. For example, Harding *et al.* (1997) found that in retrieval from OCR error degraded text the performance of query expansion using *n*-gram matches improved when the similarity threshold was relaxed and more expansion terms were allowed. Even more generally, query expansion has a

notable positive effect on CLIR performance (c.f. the seminal work by Ballesteros & Croft , 1997).

In the studies included in this thesis, the number of closest *s*-gram matches included in the queries varied depending on the goals of the studies. Only a few closest *s*-gram matches were included in the queries in studies I, III and IV, as shown in Table 4. The goal of these studies was to find one or few correct translations of the source language query words, and the number of closest matches was chosen based on previous CLIR studies by Pirkola, Puolamäki and Järvelin (2003a) and Hedlund *et al.* (2004). In Study V, the goal was to study the optimal number of *s*-gram matches to use for expanding the target language queries in a very noisy collection, where each query word potentially had tens of correct variants. Therefore, a wide range of expansion levels were tested, with the number of the closest *s*-gram matches added to the expanded queries varying between 2 and 60.

Pirkola (1998) suggested the use of structured queries to alleviate the problems related to ambiguity and loss of meaning in query translation. Grouping the alternative translations using the synonym operator of the INQUERY search engine nearly doubled the performance of the translated queries in his study, showing that it was an effective solution to reduce the query drift caused by irrelevant translations with high document frequencies (*ibid.*). The use of simple synonym based query structuring in CLIR was also recommended by Hedlund *et al.* (2004), who showed that additional proximity based structures for combining compound and phrase constituents only decreased retrieval effectiveness. However, Pirkola *et al.* (2003a) reported that structured queries combining a Boolean and-structure (#band) with a synonym structure performed better than the synonym structure alone for queries where dictionary-based translation and *n*-gram translation of OOV words were combined. Harding *et al.* (1997) used similar synonym structures in the context of OCR degraded text for combining the *n*-gram query expansion terms.

In this thesis, INQUERY search engine (Callan, Croft & Harding, 1992) was used in Studies I and III. In Studies IV and V the Lemur Indri search engine (Strohman, Metzler, Turtle & Croft, 2005; The Lemur Project[6]) was used. The query language of Indri was developed based on INQUERY's query language and supports its query structures (even if instead of tf.idf estimates, estimates from the language model are used in Indri). This allowed for the use of corresponding

[6] http://www.lemurproject.org

synonym based query structuring in all studies included in this thesis, where retrieval experiments were run (i.e., all but Study II). Using the plain synonym structure instead of the Boolean and (#band) structure seemed motivated in the context of this thesis: Pirkola's results are from a study where dictionary based translation was complemented with *n*-gram translations of OOV query words (Pirkola *et al.*, 2003a). It was pointed out in their study that combining the #band-structure with a synonym structure was necessary for relaxing the Boolean conjunction in cases where the *n*-gram translation introduced bad query keys to the translated queries (*ibid.*). In our studies, where the entire query translations were based on *s*-gram matching, the #band structure seemed too strict because of the noisy translations generated by *s*-gram matching for most query words.

4.3 Evaluation measures

A variety of evaluation measures exists for quantifying the effectiveness of information retrieval systems. The evaluation measures used in the studies included in this thesis are listed in Table 3 and described below. Most effectiveness measures are based on the concepts of recall and precision. *Recall* describes the ability of a retrieval system to retrieve relevant documents. It is defined as the share of the (known) relevant documents in the collection that are retrieved by a retrieval system for a query. As the real number of relevant documents is not known, recall is calculated based on the relevant documents in the relevance corpus.

$$Recall = \frac{Relevant\ documents\ retrieved}{All\ (known)\ relevant\ documents}$$

Precision describes the ability of a retrieval system to differentiate between relevant and irrelevant documents. It is measured as the share of the relevant documents of all of the retrieved documents.

$$Precision = \frac{Relevant\ documents\ retrieved}{All\ documents\ retrieved}$$

Recall and precision can be calculated at each position of a ranked result list. The position where recall and precision are calculated then models the stopping point where the user ceases to examine the result (Tague-Sutcliffe, 1992). Different stopping points and evaluation measures may be used to model the different goals

that users might have. Precision measured at various stopping points in the ranked result list (P@K, precision at rank *K*) is an intuitive user oriented measure, where the stopping points can be selected to reflect the number of documents a user is likely to examine. It is often used for evaluation of Web search tasks, where users are known to focus on results ranked early in the result list. Commonly used cutoff points include 1, 10 and 20 retrieved documents, i.e. P@1, P@10, P@20. Buckley and Voorhees (2000) noted that P@K does not average well over topics, if the number of relevant documents varies much between the topics. Therefore, a higher number of topics is needed for the evaluation based on P@K to be stable, than is needed for example for evaluation based on Mean Average Precision (MAP), which is a more stable measure (Buckley & Voorhees, 2000). MAP has long been the standard evaluation measure for comparing retrieval methods. It combines recall and precision in a single value by measuring precision after each retrieved relevant document on a result list (P), averaging the precision for the topic over the result list (AP), and finally calculating the mean of average precisions over all topics (MAP) (see e.g., Buckley & Voorhees, 2004).

Reciprocal Rank (RR) score for a query is measured as the reciprocal of the rank of the first relevant result, or zero if no relevant document is found among the documents inspected (Voorhees, 2001). For example, if the first relevant result for a query is found at rank 3, the RR score of the query is 1/3. Mean Reciprocal Rank (MRR) is then calculated as the mean of the individual queries reciprocal ranks (*ibid.*). MRR has been widely used in question answering and known item search tasks, where only the first relevant result is of interest (Voorhees, 1999; Soboroff, 2004).

Cumulated Gain (CG), developed by Järvelin and Kekäläinen (2002), is a measure for computing the gain a user obtains by examining the ranked result list up to a given rank,. Different relevance scores can be assigned to documents depending on the level of their relevance, and the scores are simply summed from rank position 1 to *n*. Discounted Cumulated Gain (DCG) adds a rank-based discounting factor to decrease the gain acquired from documents ranked further down in the result list. The discount models the stopping point as a decreasing probability of a user encountering a document when the document's rank increases. The relevance score of a document is divided by the logarithm of its rank, and the discounted scores are summed from rank 1 to *n*, as in the plain CG. Different logarithm bases can be used for modelling varying user behavior: e.g., logarithm base 2 for a steeper discounting to model impatient users and logarithm base 10 for more lenient discounting and modelling patient users (Järvelin &

Kekäläinen, 2002). Finally, normalized Discounted Cumulated Gain (nDCG) calculates the relative-to-ideal performance based on DCG by comparing the acquired discounted cumulated gain vector to the ideal vector corresponding to the best possible ranking of the documents. Measuring the relative-to-ideal performance makes it possible to compare retrieval methods based on their discounted cumulated gain. The average nDCG up to a given rank position k gives a single-value performance summary and can be used in statistical significance testing (similarly to average precision over specific document cut-off value points). (Järvelin & Kekäläinen, 2002.)

MAP was used as an evaluation measure in all retrieval experiments (Studies I, III, IV and V), due to its position as a standard evaluation measure frequently used in information retrieval evaluations. It balances precision and recall in a single, stable evaluation measure. In addition to MAP, precision at the standard recall levels of 10 % and 50 % was reported in Study III. In later retrieval studies, reporting precision at recall levels was replaced by P@K and DCG measures, which have clearer and more intuitive usage scenarios. P@K was used in Studies IV and V. It is a simple and intuitive measure with a clear connection to user satisfaction, especially when measured at lower depths, such as among the 10 top-ranked results. In Study V, the normalized discounted cumulated gain (nDCG) was the main evaluation measure to allow for the use of graded relevance assessments and discounting the value of relevant documents found at low ranks.

In Study II the goal was to measure at which rank the correct translation was found on average. Therefore, the mean reciprocal rank was used as an evaluation measure. The translation performance for a single query word was measured as the reciprocal of the rank at which the correct translation occurred. The evaluation focused on the top-5 ranks of the result lists, because only the top matches are usually used in s-gram based translation of OOV query words. Therefore, if the correct translation did not occur among the top-5 matches for a query, the query was given the score 0.

4.4 Statistical testing

Non-parametric significance tests suitable for paired samples or multiple related samples are often used in information retrieval experiments (Hull, 1993; Van Rijsbergen, 1979). Non-parametric significance tests are preferred as they do not make as stringent assumptions concerning the distribution of the measurements as

the more powerful[7] parametric tests do: recall and precision are discrete variables and the values they can take on are not normally distributed (Hull, 1993).

The Friedman test is a non-parametric version of the two-way analysis of variance (ANOVA) for several related samples (Conover, 1999). It is suitable for test collection based information retrieval evaluations, where several retrieval methods are evaluated using the same set of topics. The alternative hypothesis (H_1) of the Friedman test states that at least one of the compared retrieval methods tends to perform better than at least one of the other methods. The test uses a randomized complete block design, where retrieval methods are compared topic-wise, and the measurements (for each method and topic) are replaced with the ranks of each method's performance on the topic. Using ranks instead of measurement values avoids the problems with the non-normal distribution of the values and normalizes differences between topics (Hull, 1993). The sums of the ranks for each method are compared and the probability of the rank totals being observed by chance is calculated. An initial test is performed to determine whether there is any difference between the evaluated retrieval methods. If an initial difference is detected, the methods are compared to each other to determine between which methods there are significant differences. (Hull, 1993; Conover, 1999).

The Wilcoxon signed-rank test is the non-parametric alternative for the t-test for comparison of two related samples (Hull, 1993). The Wilcoxon test replaces each difference with the rank of its absolute value. The ranks are multiplied by the sign of the difference, and the sums of the ranks are compared to their expected values. (*ibid.*) The Wilcoxon test has been frequently used in information retrieval evaluations. However, it has been criticized for both lack of power and potential to lead to false detections of significance (Smucker, Allan & Carterette, 2007).

Pearson's chi-square is a test statistic used for goodness-of-fit tests applied to nominally scaled data (Lind, Marchal & Wathen, 2014, pp. 543-547). It compares the observed distribution of events to an expected theoretical distribution calculated based on the number of observations in each category. The null hypothesis (for equal expected frequencies) is that there is no difference between observed and expected frequency distributions, and that any differences are due to a chance (*ibid.*). A limitation of the chi-square test is that it might lead to an erroneous conclusion if the expected frequency in a cell is unusually small.

[7] The more powerful tests make fewer Type II errors of accepting the null hypothesis, when it actually could be rejected.

Therefore, the test statistic may not be used if more than 20 % of the expected value cells have an expected frequency less than 5 (*ibid.*, p. 551).

The presence of statistically significant performance differences does not guarantee that the performance differences are noticeable in practice. Spärck Jones (1974) suggested that in addition to statistical significance, an absolute performance difference of at least 5-10 %-units is required for the differences to be considered noticeable in practice, and above 10 % to be considered material (*ibid.*).

Statistical significance testing was applied to the results of all of the studies included in this thesis, with the exception of Study IV were statistical significance was not measured. The Friedman test was used in Studies II, III and V, while the Wilcoxon signed-ranks test was used in Study I. In addition to the Friedman test, Pearson's chi-squared test was used in the topic level analysis of the *s*-gram query expansion's performance in Study V, to control if the differences observed between different topic categories were statistically significant. We also applied Spärck Jones' (1974) categorization for estimating the practical importance of the results.

4.5 Strengths and weaknesses of test collection based evaluation

Empirical evaluation is a critical component of information retrieval research, because the performance of an IR system cannot be reliably predicted prior to a search being conducted; the current theoretical models of information retrieval simply do not allow for such predictions (Sanderson, 2010). Therefore, most information retrieval research has followed the test collection based evaluation methodology. The strength of the methodology is that it creates a controlled test setting and stable evaluation measures, and produces repeatable experimental results as a basis for comparisons of the different information retrieval methods (Sanderson, 2010; Kekäläinen & Järvelin, 2002a).

However, the test collection based evaluation methodology has been criticized for its lack of realism. For example, Ingwersen and Järvelin (2005) argued that the applicability of the evaluation results from test collection based evaluations is limited by the exclusion of real users and their tasks. Indeed, transferring results from laboratory studies to operational environments is a complex task and requires that the assumptions made and the variables used in the experiment are reasonable and relevant in the operational environment. Much of the criticism has targeted the

simplistic models of users and interaction that are reflected in the test collections and evaluation measures: the focus on single query interactions (Keskustalo, Järvelin, Pirkola *et al.*, 2009; Kanoulas, Carterette, Clough & Sanderson, 2011), exclusion of interface features other than ranked result lists (such as summaries, Turpin, Scholer, Järvelin *et al.*, 2009), and the user models of patient users who gain equal value from all relevant documents despite their rank in the result list (Järvelin & Kekäläinen, 2002; Moffat & Zobel, 2008). The operationalization of relevance as binary topicality has been much discussed (Sormunen, 2002; Kekäläinen & Järvelin, 2002b).

It has been questioned whether there is a correlation between retrieval performance as measured in test collection-based evaluations and user performance when using the evaluated systems (Hersh, Turpin, Pierce *et al.*, 2000; Turpin & Hersh, 2001; Turpin & Scholer, 2006; Smith & Kantor, 2008). However, in later studies clear correlations have been identified (Al-Maskari, Sanderson, Clough & Airio, 2008; Turpin & Scholer, 2006). Al-Maskari *et al.* (2008) argued that the difficulty to find correlations between system and user performance at least partly is explained by the subjectivity of the information retrieval process: even in controlled studies, users tend to disagree on issues such as what a topic is about, which topics are easy and difficult, and which documents are relevant for a given topic (*ibid.*). Voorhees (2002) noted that the relevance assessments used in test collections are not necessarily representative, given the inherent subjectivity of relevance assessment, and the limitations of pooling. However, she also showed that - with adequate pool depth and diversity – the bias caused by the missing relevance assessments only had a minor effect on the evaluation results (*ibid.*). Scholer and Turpin (2009) actually showed that users could be categorized based on their relevance assessment profiles. When users with relevance assessment profiles similar to those of the TREC assessors were used, stronger correlations between the system and the user performances were observed (*ibid.*).

Graded relevance scales are today widely used in major test collections, and evaluation measures that account for graded relevance and users' effort and willingness to examine results (including nDCG and RBP) are available and used in the major evaluation conferences. Smucker (2009) recently suggested an evaluation approach where more complex user and interaction models are integrated, and can be tuned to model different usage scenarios based on results of user studies. The single query-result interactions are a core function of search engines and therefore a motivated target of evaluation. For the task of evaluation over multiple query sessions, a session track is running in TREC since 2010. Several authors

(Keskustalo *et al.,* 2009, Kanoulas *et al.*, 2011; Azzopardi, 2011; Baskaya, Keskustalo & Järvelin, 2012) have shown that evaluation based on sessions of queries is possible within the scope of test collection-based evaluation methodology.

The test collection based evaluation methodology was adopted in this thesis because it is well-suited for the goals of the thesis: evaluating how *s*-gram query translation and expansion affect retrieval effectiveness, and analysing how the source word types and source and target language similarity affect *s*-gram translation precision. Following the test collection based evaluation methodology made it possible to use existing test collections, which provided us with enough data for stable measurements of the differences in our translation approaches. This made it possible to experiment in three different text domains and compare results between the domains. We could make use of standard evaluation measures, which are widely used and the behavior of which is well understood.

We attempted to account for the limitations of the model related to topic types, relevance assessments and evaluation measures: different evaluation measures were used for modeling different user behavior (recall-oriented vs. precision-oriented) and graded relevance assessments were used in Study V to better model user preferences. The types of topics used may limit the applicability of the results. The focus of this thesis is on different text domains, not on specific search tasks. However, different search tasks may have different typical vocabulary usage and present different challenges to *s*-gram translation (e.g., long vs. short queries; share of proper name queries).

More generally, the limitations of the test collection based evaluation methodology are quite well understood and do not take away the value of the methodology for the task of evaluating information retrieval algorithms' effectiveness in ranking topically relevant documents.

5 Summary of the individual studies

This section presents a summary of the studies I-V included in this thesis (see list of original publications, page 7).

5.1 Study I: Defining *s*-grams

Study I focused on developing definitions for classified *s*-grams and two proximity measures for them: L1 distance and Jaccard coefficient. There were no previous stringent definitions for classified *s*-gram matching and proximity measures for calculating string proximity based on classified *s*-grams. The goal was to develop definitions applicable for *s*-grams of different lengths and for different character combinations indexes. Another goal was to make a literature review on *s*-grams and *n*-grams and their applications in natural language information retrieval. A case study building on the experiments run in (the chronologically earlier) Study III was also included, making three notable changes: (1) inflected (non-lemmatized) document collection and query words were used, (2) the number of *s*-gram variants added to the translated queries was reduced to 3 (from 4 in Study III), and (3) the dictionary-based baseline query structure was changed to a simpler synonym structure, instead of using proximity operators as in Study III.[8]

The research questions of the Study I that were related to this thesis included:

- What are the efficiency limitations of gram profile definitions for classified *s*-grams in cross-language information retrieval? What is distinct for cross-language information retrieval as an application area of *s*-gram matching? (Literature review)
- Is classified *s*-gram matching a competitive and effective approach to query translation between closely related languages in a "non-linguistic" setting, i.e., when no linguistically motivated tools are used for language analysis or translation?

[8] The case study also corrected the errors in the Swedish monolingual baseline that occurred in Study III.

Results. Definitions for classified *s*-grams and their proximity measures were developed in a bottom-up manner, and the efficiency of the classified *s*-gram approach was discussed in the context of natural language information retrieval, based on the *s*-gram profiles, alphabet size and length of the *s*-grams. The *s*-gram profile definitions are mathematically elegant and solid, but not necessarily a computationally efficient solution, because each gram class of a CCI requires its own *s*-gram profile vector, and each vector contains all *s*-grams of length *n* that can be formed of a given alphabet. However, the *s*-gram vectors of natural language words are very sparse, because the average word length is well below ten characters and because di-grams are regularly used. Similarity comparisons based on sparse vectors are usually computationally relatively efficient. We concluded that for the application area of natural language query translation (and expansion) the efficiency issues are not a serious problem.

The absolute performance differences between the *s*-gram queries and dictionary translated queries were similar to the differences measured in Study III: the absolute difference in MAP between the approaches was 0.056 %-units. Here, the difference was statistically significant, suggesting that when no tools for handling morphological variation are used, dictionary-based query translation performs noticeably better (following Spärck Jones' (1974) definition of practically noticeable differences) than *s*-gram query translation. Reaching 80 % of the dictionary-based baseline's performance and improving by 70 % over the monolingual Norwegian baseline are however promising results: *s*-gram query translation may definitely be a competitive approach to query translation between closely related languages when linguistically motivated tools for translation and morphological processing are not available.

5.2 Study II: Comparison of *s*-gram proximity measures

Study II concentrated on comparing different *s*-gram proximity measures, CCIs and padding options in a CLIR context. In previous *n*-gram and *s*-gram studies, various proximity measures had been used for calculating the proximity between strings, but no evaluations had been made considering the mutual superiority of the measures. In Study II, seven proximity measures for classified *s*-grams were compared, including the most commonly used non-binary proximity measures L1, Tanimoto coefficient and cosine similarity, and their binary counterparts Hamming distance, Jaccard's coefficient and binary cosine similarity. In addition the binary

Dice coefficient was tested. Twelve different CCIs were tested. Eleven languages pairs were used to see if the extent of variation between source and target strings affected the performance of the proximity measures.

The research questions in Study II were:

- Does using non-binary proximity measures instead of binary measures improve *s*-gram translation precision in CLIR?
- Does CCI affect which proximity performs best?
- Does the relatedness of the language pairs affect which proximity measures perform best?

Our hypothesis was that the non-binary proximity measures would be more effective than their binary counterparts, due to their better sensitivity: they account for the number of the unique *s*-grams in strings, while the binary measures only count each unique *s*-gram once.

Results. Contradicting to our hypothesis, the binary proximity measures generally performed better than the non-binary ones. The differences between the binary measures were negligible. The CCIs that included several gram classes and where several *s*-gram types were combined in the gram classes performed somewhat better than the simpler CCIs. The two gram class CCI {{0},{1, 2}}, or combinations of any three of the gram classes {0}, {1}, {0, 1} and {1, 2} performed the best. Adding a fourth gram class into a CCI did not lead to additional performance improvements. {1}, {0, 1} and {1, 2} were the gram classes that corresponded to the most common character changes in our analysis of English–French and English–German cross-lingual spelling variants.

The differences between the binary and non-binary measures were found to depend on the combination of the CCIs used and the use of padding: where several types of *s*-grams were combined in the gram classes, the same *s*-grams including padding characters were produced multiple times. The non-binary proximity measures then heavily over-weighted the padding *s*-grams in the proximity calculations. In the test runs with only left padding and without padding, the differences between the binary and non-binary measures were reduced and disappeared, respectively. Simultaneously, the general performance however deteriorated showing that a minimum of one-sided padding is needed. Character changes were found especially common at the ends of the cross-lingual spelling variants, which suggested that (at least) the more stable beginnings of the variants

should be padded. Consequently, any of the binary proximity measures Dice, Jaccard and Binary Cosine were recommended for string matching in natural language information retrieval applications, where the alphabet is relatively large and strings relatively short and the repetition of *s*-grams is not extensive. Non-binary proximity measures might be better suited for applications such as gene matching, where extensive *s*-gram repetition occurs due to much longer strings and much smaller alphabet.

The language pair did not affect the performance of the proximity measures in any major way, even if it did have a notable effect on the general performance level of the *s*-gram translation. The highest performance levels of MRR@5≈0.76 were measured for Swedish to German and German to Swedish translations. The lowest translation performance of MRR@5≈0.45 was measured for Finnish to English and English to Finnish translations. The results for the language pairs followed the intuitive order of similarity of the languages at large. However, Spanish and Italian to English translation performance diverged somewhat from the expected. According to the language pair similarity measurements by McNamee and Mayfield (2004), these language pairs are more similar than the pair Swedish-German. However, in Study II the MRR for these language pairs was clearly lower (MRR@5≈0.52 for Italian to English and MRR@5≈0.57 for Spanish to English).

5.3 Study III: Query translation between closely related languages

In Study III, we introduced closely related languages as a separate line of research for cross-language information retrieval. Closely related languages typically share a high number of cross-lingual spelling variants, opening for translation approaches based on string similarity only. Instead of the common query translation approaches based on linguistic knowledge, such as translation dictionaries, we suggested using fuzzy translation techniques based on the similarities in spelling between the source and target languages. We studied the use of fuzzy translation techniques based on *s*-gram matching, transformation rule-based translation (TRT) and the combination of TRT and *n*-grams for query translation between Swedish and Norwegian. We also measured the similarity of the Norwegian source and Swedish target words using the LCSR measure. The goal was to study how language pair similarity affects *s*-gram query translation performance. We also measured the language pair similarity for another, less closely related European

language pair, German and English, to get a comparison point for the relatedness of the languages.

The research questions in Study III were:

- Can query translation between closely related languages be handled using fuzzy translation approaches?
- Are there clear performance differences between the tested approximate string matching and rule-based translation approaches?
- How closely related should the source and target language be, for the fuzzy query translation to be a viable alternative?

Results. The fuzzy translation techniques were shown effective and applicable translation techniques in CLIR between closely related languages. The *s*-gram query translation approach reached results comparable to those of the dictionary-based query translation. The absolute performance differences between the translation approaches were small, around 0.055 %-units. The performance differences were not statistically significant. Given the lack of statistically significant differences, and given that a sufficient number of topics were used for the results to be reliable (cf. Voorhees & Buckley, 2002), we can conclude – following the definition of Spärck Jones (1974) – that the observed performance differences are hardly noticeable in practice. The absolute performance differences were higher at lower recall levels, but still not statistically significant. These results were encouraging and supported our hypothesis that dictionary-based translation could be replaced by fuzzy string matching techniques in CLIR between closely related languages.

Performance differences between the different fuzzy translation approaches were small: all the four approaches including *s*-grams (*n*-grams) performed similarly. There were minor differences in their performance on the different recall levels, but none of the differences were statistically significant. The average similarities measured for Swedish and Norwegian and for English and German source and target words were 0.815 and 0.556, respectively. Language similarity values approaching LCSR=0.8 could then be seen as an indication of the *s*-gram based techniques applicability for a language pair.

5.4 Study IV: Cross-language photographic image retrieval

Study IV focused on *s*-gram based query translation in retrieval of photographic images. Our aim was to develop a language-independent photo retrieval approach combining visual retrieval with *s*-gram translation. We studied six language pairs with different levels of relatedness using English and German as target languages and Danish, English, French, German, Norwegian, and Swedish as source languages. None of the language pairs were as closely related as Swedish and Norwegian studied earlier. The photo annotations and queries were however expected to contain a high share of proper names, which would increase the level of language similarity within the language pairs. The images were lightly annotated, i.e. contained relatively little textual content, increasing the importance of the successful fusion of text and visual features.

The research questions targeted in study IV were:

- Is the classified *s*-gram matching technique a suitable and sufficient query translation technique in multimodal photographic image retrieval?
- How does it perform compared to dictionary-based translation?
- Is the classified *s*-gram matching technique a sufficient approach to handling morphological variation in monolingual text-based and multimodal photo retrieval?

Results. The *s*-gram based query translation performed nearly as well as the dictionary-based translation in this study. The absolute differences in MAP between *s*-gram and dictionary translation were below 0.05 %-units for all six language pairs (avg. difference 0.03 %-units). Similar absolute differences were measured in P@10 and P@20 (avg. difference 0.03 %-units even here).

The results varied depending on the language pair following rather well the intuitive similarity of the languages (and McNamee and Mayfield's (2004) measurements): the *s*-gram translation reached its best results with Norwegian and Danish topics and German annotations - over 90 % of the dictionary translation's MAP (or a below 0.02 %-units absolute difference), and the worst results between German and English - less than 80 % of the dictionary translation's MAP (corresponding to a 0.04 %-units absolute difference). Swedish to German queries were an exception, where the performance differences are higher than would be expected for the language pair, in fact the highest for any of the language pairs when P@10 and P@20 are considered.

Statistical significance was unfortunately not measured in Study IV. However, the number of topics used (60) ought to be sufficient for the results to be relatively reliable and repeatable. Therefore (even without statistical testing), we can carefully conclude that *s*-gram translation seems to be a sufficient approach to query translation in multimodal image retrieval of lightly annotated images. Generally, the *s*-gram query translation's performance relative to the baseline was high for all language pairs: as high (or higher) as in Study II for the closely related languages Swedish and Norwegian. This suggests that when combined to visual retrieval, *s*-gram query translation is a sufficient translation approach for a wider selection of languages than in pure text-based retrieval.

The monolingual results were typical of the languages tested. The monolingual text-only queries using *s*-grams instead of lemmatization performed well for English reaching 95 % of the performance of the lemmatized queries. For German, which has a notably more complex inflectional morphology than English, 84 % of the performance of the lemmatized queries was achieved.

5.5 Study V: Information retrieval from historical document collections

In study V, *s*-gram matching was applied to historical document retrieval (a.k.a. historic document retrieval). *S*-gram matching was used for identifying variants of modern Finnish query words from a digitized Finnish newspaper collection from the 1800s. The *s*-gram variants were then used for expanding the modern queries. The task was slightly different from the query translation tasks in the previous cross-language information retrieval studies due to the amount of variation that occurs in automatically digitized historical collections: while query words may be expected to have a single correct spelling variant (+ a few inflections) in the cross-language retrieval scenarios, tens or even hundreds of historical, inflectional and OCR variants typically exist for modern query words in OCR scanned historical collections. Study V therefore focused on finding a query expansion approach that balances a good coverage of the different variation types and reasonable precision in generating the expansion keys.

The research questions targeted in study V were:

- What kind of variation does *s*-gram matching capture? Does the captured variation reflect the variation actually occurring in the collection?

- Is query expansion using *s*-gram variants of query words useful in historic document retrieval?
- How do topic and query word properties affect the performance of *s*-gram query expansion in historical collections?

Results. The categorization scheme for *s*-gram variant candidates included three categories: 1. variants of the query words, 2. words semantically related to the query words, and 3. noise. Categories 1 and 2 contained the potentially useful query expansion keys. 78 % of the generated *s*-gram variant candidates were categorized into one of these two categories. The largest relevant variant subcategory was that of inflectional variants. OCR variants were much more common among the top *s*-gram matches than historical variants, which partly depended on common OCR errors reversing historical changes. In over 700 of the ca 1600 relevant variants analyzed, variation of more than one type occurred.

Query expansion based on approximate string matching was superior to using the clean inflectional forms of the query words, showing that coverage of the different types of variation is more important than precision in handling one type of variation. The combined approach did not improve performance over the *s*-gram query expansion performance. Extensive expansion of around 30 variants for each query word was required to achieve the highest performance improvements. This level of query expansion is notably higher than that found useful in our previous cross-language information retrieval studies. The *s*-gram query expansion especially improved the results of poorly performing short baseline queries. The average performance improvement was around 100 % for the best *s*-gram query expansion approaches, reducing the share of low performing queries from 42 % to 16 % of all queries.

6 Discussion and conclusions

This thesis studied the use of classified *s*-gram matching in query translation between related languages, and in query expansion in historical document retrieval. The main contributions are related to the use of *s*-gram matching in natural language information retrieval, and to in what kind of document collections and between what kinds of language pairs *s*-gram matching can be applied. The thesis also addresses questions related to how the features of natural languages determine the best approaches to classified *s*-gram matching and how well they perform. In the following the main findings will be summarized and discussed.

6.1 Classified *s*-grams and natural language information retrieval

In Studies I and II different proximity measures for classified *s*-gram matching were compared, from the point of view of natural language information retrieval. The main point of comparison was the use of binary vs. non-binary proximity measures. We found that binary proximity measures were better suited for classified *s*-gram matching in natural language information retrieval applications than the non-binary ones. Natural languages contain relatively little repetition (as compared to e.g. gene sequences), and therefore the higher sensitivity of the non-binary measures is not needed, just as Robertson and Willett (1998) previously suggested. The use of padding together with gram classes that combine several *s*-gram types leads to generating the same "padding *s*-grams" repeatedly in the gram classes. The beginnings and endings of strings get therefore over-weighted when using non-binary proximity measures, leading to lower translation precision. On the other hand, it was found in Study II that the CCIs that contain gram classes where several *s*-gram types are combined generally perform better than the "simpler" gram classes, and that using padding particularly at the beginnings of strings improves performance.

The size of an *s*-gram profile for a string depends on the number of *s*-gram classes included in a CCI and the number of *s*-grams of length *n* that can be

formed of the alphabet. Increasing the alphabet size and *s*-gram length rapidly leads to efficiency issues in *s*-gram matching. In natural language IR applications, the best performing CCIs contain 2-3 *s*-gram classes, and the alphabet has usually less than 30 symbols. Focusing on natural language query translation and expansion, the *s*-gram length is typically 2. With such values, the *s*-gram profile size remains reasonable and *s*-gram matching can be implemented efficiently using the *s*-gram profiles. Using higher *s*-gram lengths, the *s*-gram profile size easily grows impracticable.

6.2 Language pair similarity

The potential of *s*-gram based query translation and expansion depends on the similarity of the source and target languages. It has been previously shown that the translation precision varies on same type of data between different language pairs (Loponen et al., 2008), and that the retrieval performance of *n*-gram queries correlates with the similarity of the source and target languages (McNamee & Mayfield, 2004). The results of this thesis supported similar conclusions: *s*-gram translation precision and retrieval performance of *s*-gram translated queries generally depended on the similarity of the source and target languages. However, language pair similarity as measured by average LCSR did not (alone) reliably predict the retrieval performance of *s*-gram queries. An analysis of the topic words in Studies III and IV suggested that particularly the frequency of compound words in the source language affected the relation of language pair similarity as measured by LCSR and retrieval performance (see Table 5). High share of compounds in the source language led to low LCSR scores, because the source language compounds frequently did not have exact translations in the target language document collection (particularly when the target language was not a compounding language). These differences are however not reflected in the MAP scores: e.g., in Study IV German to English LCSR calculated for the German query words and their correct translations in the English collection was clearly lower than the LCSR for English to German in the same study. The absolute performance of the German to English *s*-gram queries was however slightly higher than the performance of the corresponding English to German queries. This is likely to depend on the different problems related to translation of compound words and phrasal expressions in CLIR (Pirkola, 1998; Hedlund, 2002; Hedlund *et al.,* 2004), but also on the behavior of *s*-gram matching on long source words. *S*-gram translation precision

generally increases when the source word length increases, because fewer random matches occur. For compounds lacking exact translations, constituents of the compounds are often identified among the top *s*-gram matches. This phenomenon was observed and more closely analyzed in Study V, where *s*-gram matching proved a useful approach to decompounding compound words that did not occur in the target document collection. Therefore, we conclude that it might be useful to consider the language pair similarity score together with measures of the linguistic complexity of the source and target languages, e.g., based on the average word length as suggested by McNamee *et al.* (2009).

Table 5. The language pair similarities measured by LCSR and retrieval performances measured by MAP in Studies III and IV. N=number of topic words analyzed. % dict. refers to the relative performance of *s*-gram queries compared to the baseline of dictionary translated baseline queries. Compounds indicate the percentage of the source language topic words that were compounds.

	Language pair (N)	**LCSR**	**Compounds**	**MAP (% dict.)**
Study III	NO-SW (365)	0.79	24 %	0.29 (84%)
	NO-DE (365)	*0.62*	*24 %*	*N/A*
	SW-DE (337)	*0.62*	*26 %*	*N/A*
Study IV	FR-EN (143)	0.56	2 %	0,15 (81 %)
	NO-DE (131)	*0.50*	*27 %*	*0.15 (92 %)*
	DA-DE (132)	0.49	30 %	0.16 (91 %)
	SW-DE (138)	*0.49*	*25 %*	*0.15 (82 %)*
	EN-DE (142)	0,48	13 %	0,14 (79 %)
	DE-EN (125)	0.38	38 %	0.15 (77 %)

Previous research suggests that the similarity of languages is not constant the way their relatedness is. It varies between collections and domains, and between different types of queries. For example, up to 78 % of Web image queries are proper name queries (Pu, 2005) leading potentially to a high level of similarity between the vocabulary of the queries and the target language, even for distantly related languages as long as they share the same writing system. The language pair similarity scores varied between the test collections used in this thesis, even though to the opposite of the expected direction: the average LCSR scores measured for

Swedish and Norwegian to German were clearly higher in the newswire collection of Study III than in the photo collection of Study IV (Table 5, the rows marked in italic). The frequency of compounds, or untranslatable topic words did not explain this difference leaving topic vocabulary and collection vocabulary size as possible explanations. The IAPR TC-12 collection used in Study IV is not focused on proper name retrieval and actually contains a lower share of search topics that contain proper names than the CLEF 2003 newswire collection does.

How similar should then two languages be for the *s*-gram translation to be a useful query translation and expansion approach? In Studies I and III it was found that in text retrieval applications *s*-gram query translation performed comparably with dictionary-based translation in Norwegian to Swedish retrieval. The average LCSR≈0.8 measured between the Norwegian source and Swedish target words could then be seen as an indication of the *s*-gram based techniques applicability for a language pair. In Study IV where the queries also included a visual retrieval component, *s*-gram query performance was competitive with the dictionary based translation despite clearly lower average LCSR scores (≈0.5). Therefore, the applicability of the *s*-gram query translation depends on the application domain, document collection and the other resources available for the search system. Translation is but one component of a retrieval system, and the performance differences between *s*-gram and dictionary-based translation may decrease when the techniques are combined with other evidence of relevance. Language similarity measurements can serve as an estimate of the potential of the *s*-gram matching to improve the results as compared to no translation or no query expansion.

The similarity of languages in different document collections and for different types of queries can be measured by using rather simple approaches. McNamee and Mayfield (2004) suggested using the Bendetto language pair distance for measuring the relatedness based on parallel documents. We suggested measuring average LCSR between a set of query words and their translations in the target document collection, because it allows accounting for the vocabulary typical of the application domain. This approach was used by Loponen *et al.* (2008). When parallel documents are available in the test collection, the Bendetto language pair distance can be used for measuring collection or domain specific similarity of the languages.

6.3 Number of variants

Loponen *et al.* (2008) showed that *s*-gram matching is sensitive to the decreasing similarity between the source and target languages. Translation precision decreases fast as more variant candidates need to be included to capture the correct translation. Thus, *s*-gram matching might not be the first choice as a fuzzy translation approach for applications where high translation precision is required and only one or few correct translations are likely to exist – transliteration approaches such as the FITE-TRT advocated by Loponen *et al.* (2008) can offer more precise translations. However, where a query expansion effect is desirable and where the tools for handling morphological variation, compounding, spelling variation, and OCR errors are not available, *s*-gram matching can be very useful as it can simultaneously handle all these different types of variation. In Study V, concerning *s*-gram query expansion in a noisy collection where query words typically had tens of relevant variants in the collection, *s*-gram query expansion reached a 78 % precision in generating query expansion keys. It was found that extensive query expansion of 30 (and above) *s*-gram variants for each query word gave the best performance improvements reaching over 100% relative improvements over the baseline. Classified *s*-gram matching is then rather a technique for generating query expansion terms than a precise translation tool: it is best suited for applications where the target index contains many relevant variants of the query words (and words related to them), and where adding several of them to the queries is desirable. The optimal number of *s*-gram variants to be added to the target language queries depends on the target language, and on the collection: only few for CLIR in clean, lemmatized collections (Hedlund *et al.*, 2004), but many in collections where a combination of many productive variation types occur. Web photo retrieval is an example of an application area where *s*-gram matching might be a useful cross-lingual query expansion tool: the Web is a large multilingual and largely unedited document collection, where many different types of variants occur for most query words.

The query expansion behavior of *s*-gram matching in noisy collections reminds of the massive query expansion behavior discussed by Buckley *et al.* (1995). The suitable *s*-gram query expansion level is limited by the number of variants that words typically have in a specific collection. However, as long as the expansion level is not exceeded, the inevitably generated noise does not damage ranking because good expansion terms co-occur non-randomly in relevant documents, while poor expansion terms co-occur randomly (cf. Buckley *et al.*, 1995). Harding *et*

al. (1997) showed similar behavior for *n*-gram-based query expansion in an OCR scanned collection where much variation occurred: relaxing the similarity threshold and adding more terms to the expanded queries improved retrieval performance.

6.4 Validity, reliability and limitations

The studies of this thesis followed the test collection-based evaluation methodology, with its strengths and limitations. A minimum of 50 topics was used in all studies, which together with statistical significance testing and careful consideration of the relative and absolute differences in the performance scores provides enough data for reliable results replicable in different test settings (Spärck Jones, 1974; Voorhees & Buckley, 2002). The topics used are typical for test collection base evaluation: they describe topical, informational information needs that require finding several relevant documents. Relevance scales used in the collections vary from binary to a four point scale. The binary scale used in Studies I, III and IV means that rather liberal relevance criteria are applied, which might have a negative effect on the correlation of the results with user performance (cf. Scholer & Turpin, 2009). In most studies, several evaluation measures were used to reflect differences in search strategies and user persistence. Studies I and III had a clear focus on recall-oriented search. MAP and the PR-graph were reported in Study I, and MAP and precision at recall levels 10 and 50 in Study III. In Studies IV and V, MAP was complemented with precision at early ranks (P@10 and P@20) and nDCG (in Study V) to reflect more precision-oriented user models. The results from the different evaluation measures were similar and supported the same conclusions.

The experimental methodology strengthens the reliability of the results, but also limits the scope of the experiments. The studies included in this thesis are system oriented and do not account for interaction (cf. Keskustalo *et al.,* 2009), or different support offered by the interface for the users' relevance decisions (cf. Turpin *et al.*, 2009). The studies evaluate the effect of *s*-gram translation and query expansion on result ranking for single queries, from the viewpoint of users who a) want to find several relevant documents and b) either examine the complete result list or only consult the top results. The results are valid for comparisons between the rankings generated using *s*-gram query translation and expansion with rankings generated using other query translation or expansion approaches. Studies by Kelly, Fu and Shah (2007), and Al-Maskari *et al.* (2008) have shown that P@10 and average

precision correlate well with user satisfaction and efficiency on informational search tasks. However, it is difficult to conclude how the differences in rankings between *s*-gram based and dictionary based translation approaches affect user performance, i.e., are the differences small enough for the *s*-gram based approaches to be acceptable alternatives to the baseline methods for any specific users, tasks, or contexts.

A few potential validity problems can be identified from the studies included in this thesis. In Study III, the query structures used for dictionary-based baseline queries and the *s*-gram queries differed due to the compound word handling. Translated compound constituents in the dictionary-based baseline were combined using a proximity operator. Otherwise, the synonym structure was used for combining the alternative translations of the query words. In the *s*-gram queries, only the synonym structure was used. The aim was an intuitive query structure, but the proximity operator might have had a negative effect on the baseline query performance (cf. Hedlund *et al.,* 2004). Therefore, in the subsequent Study I, a new experiment was run using the same data, where only the synonym structure was used for combining the dictionary translated query words. The relative performance of the queries with the synonym structure was around 4 % better than the performance of the queries with synonym and proximity structure in Study I. This was enough for making the performance difference between *s*-gram and dictionary queries in Study I statistically significant.

In Study IV, pseudo relevance feedback (PRF) and a visual retrieval component were used in combination with the text-based retrieval approaches. The effects of these additional system variables to the results were not separately studied. Therefore, isolating the exact effect of the *s*-gram matching as compared to dictionary-based translation is impossible. It however does not seriously undermine the conclusion that *s*-gram queries can perform nearly as well as dictionary-based translation in multimodal photo retrieval, as a component in a complex retrieval system.

In Studies I and IV different indexes were used for the dictionary-based queries and the *s*-gram queries: lemmatized indexes with the dictionary-based queries and inflected indexes with *s*-gram queries. This difference is motivated: the goal of using *s*-gram matching instead of dictionary-based translation is to avoid the use of linguistic resources with a limited availability. Using lemmatizers in such a linguistic-resource-free scenario seems unmotivated. This is likely to affect the performance of the *s*-gram queries negatively, as handling the morphological variation has been shown to have a major effect on retrieval performance in many

of the languages used in the studies (Kettunen, 2013). The effect of the translation quality on query performance is then not separable from the (negative) effect of morphological variation, or from the *s*-gram matching's ability to handle morphological variation. The results of these experiments show the performance of *s*-grams given the dual problems of morphological variation and translation. It is a realistic and valid comparison between two different approaches.

After the publication of Study III, we noticed an error in the monolingual Swedish baseline queries concerning the handling of the words not recognized by the Swedish lemmatizer: the unrecognized words were not marked in the queries and were thus matched against the index of recognized lemmatized words, instead of the unrecognized words, where e.g., many proper names get indexed. This had a clear negative effect on the Swedish baseline performance. The error was corrected in Study I, where the Swedish baseline then performed better than the dictionary translated baseline and statistically significantly better than the *s*-grams. The Swedish baseline was never used as a major point of comparison for the *s*-gram query translation: its role was to function as a control for a reasonable cross-language retrieval performance. Therefore, this mistake does not affect the conclusions related to the performance of the *s*-gram query translation.

6.5 Future

S-gram variant generation and selection could be further studied, as well as combining the *s*-gram query expansion with different approaches to handling string level variation during indexing. In this thesis, the *s*-gram variants added to the expanded queries have been selected based on string similarity only. Identifying the most frequent correct variants could be useful. It could also be studied how the number of *s*-gram expansion terms could be varied depending on the query word, instead of using as many expansion terms for all query words. Both the precision and coverage of the *s*-gram variants could potentially be improved by combining *s*-grams to different rule-based approaches, for handling morphological, historical or cross-lingual variation. This should be further explored, despite the insignificant improvements gained in Studies III and V using TRT and FCG.

In the studies included in this thesis, indexes have either being lemmatized or tokenized as-is without morphological processing. There are several light-weight, non-linguistic approaches to handling (morphological) variation in index words, including *n*-gramming, index word truncation and syllabication (McNamee et al.,

2009; Kettunen et al., 2010). Combinations of such indexing approaches and *s*-gram query expansion would be interesting to study.

6.6 Conclusions

Classified *s*-gram matching is a useful tool for handling cross-lingual, morphological, historical, and OCR variation of query words in document collections. It is best suited for query expansion (and translation) in noisy document collections for relatively close language pairs. In such collections, adding tens of the closest *s*-gram matches of the query words to the expanded queries delivers major performance improvements, as compared to unexpanded queries, standard relevance feedback and to query expansion using morphological variants of the query words.

Classified *s*-gram matching also offers a low-cost alternative to query translation between related languages. It performs comparably to dictionary-based query translation between closely related-languages and in cross-language image retrieval where visual retrieval and text retrieval are combined. Linguistically motivated approaches for handling morphological variation or compound words are not needed, because the *s*-gram based query translation can handle morphological variation through its query expansion effect, and because it effectively splits compounds that do not have exact matches in the collection.

7 References

AbdulJaleel, N. & Larkey, L. (2003). Statistical transliteration for English-Arabic cross language information retrieval. In Proceedings of the 12th International ACM CIKM Conference on Information and Knowledge Management. New York: ACM, 139-146.

Airio, E. (2006). Word normalization and decompounding in mono- and bilingual IR. Information Retrieval, 9(3), 249-271.

Al-Maskari, A., Sanderson, M., Clough, P. & Airio, E. (2008). The good and the bad system: does the test collection predicts users' effectiveness? In Proceedings of the 31st International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 59-66.

Al-Onaizan, Y. & Knigth, K. (2002). Named entity translation: extended abstract. In Proceedings of the second International Conference on Human Language Technology. Stroudsburg: ACL, 122-124.

Alkula, R. (2001). From plain character strings to meaningful words: Producing better full text databases for inflectional and compounding languages with morphological analysis software. Information Retrieval, 4(3-4), 195-208.

Amati, G., Celi, A., Di Nicola, C., Flammini, M. & Pavone, D. (2011). Improved stable retrieval in noisy collections. In Amati, G. & Crestani, F. (eds.), Advances in Information Retrieval Theory. Berlin-Heidelberg: Springer, 342-345.

Arni, T., Clough, P., Sanderson, M. & Grubinger, M. (2009) Overview of the ImageCLEFphoto 2008 photographic retrieval task. In Peters, C. et al. (Eds.), Evaluating Systems for Multilingual and Multimodal Information Access: Proceedings of the 9th Workshop of the Cross-Language Evaluation Forum. Berlin-Heidelberg: Springer, 500-511.

Azzopardi, L. (2011). The economics in interactive information retrieval. In Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 15-24.

Ballesteros, L. & Croft, B. (1997). Phrasal translation and query expansion techniques for cross-language information retrieval. In Proceedings of the 20th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 84-91.

Barðdal, J., Jörgensen, N., Larsen, G. & Martinussen, B. (1997). Nordiska: Våra språk förr och nu. Lund: Studentlitteratur.

Baskaya, F., Keskustalo, H. & Järvelin, K. (2012). Time drives interaction: Simulating sessions in diverse searching environments. In Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 105-114.

Beitzel, S., Jensen, E. & Grossman, D. (2003). A survey of retrieval strategies for OCR text collections. In Proceedings of the fifth Symposium on Document Image Understanding Technology. College Park: University of Maryland, 145-151.

Benedetto, D., Caglioti, E. & Loreto, V. (2002). Language trees and zipping. Physical Review Letters, 88(4).

Bilenko, M., Mooney, R., Cohen, W., Ravikumar, P. & Fienberg, S. (2003). Adaptive name matching in information integration. IEEE Intelligent Systems, 18(5), 16-23.

Braschler, M. & Ripplinger, B. (2004). How effective is stemming and decompounding for German text retrieval? Information Retrieval, 7(3-4), 291-316.

Braun, L., Wiesman, F. & Sprinkhuizen-Kuyper, I. (2002). Information retrieval from historical corpora. In Proceedings of the 3rd Dutch-Belgian Information Retrieval Workshop, DIR. Leuven: KU Leuven, 106–112.

Brew, C. & McKelvie, D. (1996). Word-pairs extraction for lexicography. In Proceedings of the second International Conference on New Methods in Language Processing (Ankara, Turkey), 45-55. Retrieved from http://www.ling.ohio-state.edu/~cbrew/papers/nemlap96.ps.

Borgman, C. & Sigfried, S. (1992). Getty's SynonameTM and its cousins: A survey of applications of personal name-matching algorithms. Journal of the American Society for Information Science, 43(7), 459-476.

Buckley, C., Mitra, M., Walz, J. & Cardie, C. (1998). Using clustering and super concepts within SMART: TREC-6. In Voorhees E. & Harman, D. (Eds.) Proceedings of the Sixth Text REtrieval Conference (TREC-6). NIST Special Publication 500-240. Gaithersburg: NIST, 107–124.

Buckley, C., Salton, G., Allan, J. & Singhal, A. (1995). Automatic query expansion using SMART: TREC-3. In Proceedings of the Third Text REtrieval Conference (TREC-3), NIST Special Publication 500-225. Gaithersburg: NIST, 69-79.

Buckley, C. & Voorhees, E. (2000). Evaluating evaluation measure stability. In Proceedings of the 23th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 33-40.

Buckley, C. & Voorhees, E. (2004). Retrieval evaluation with incomplete information. In Proceedings of the 27th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 25-32.

Callan, J., Croft, B. & Harding, S.M. (1992). The INQUERY retrieval system. In Proceedings of the Third International Conference on Database and Expert Systems Applications. Berlin-Heidelberg: Springer, 78-83.

Can, F., Kocberber, S., Balcik, E., Kaynak, C., Ocalan, H. C. & Vursavas, O. N. (2008). Information retrieval on Turkish texts. Journal of the American Society for Information Science and Technology, 59(3), 41-53.

Cavnar, W. B. & J. M. Trenkle. (1994). N-gram-based text categorization. In Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval. Las Vegas: University of Nevada, 161–175.

Chew, P., Bader, B. & Abdelali, A. (2008). Latent morpho-semantic analysis: Multilingual information retrieval with character n-grams and mutual information. In Proceedings of the 22nd International Conference on Computational Linguistics COLING 2008. Stroudsburg: ACL, 129-136.

Christen, P. (2006). A Comparison of Personal Name Matching: Techniques and Practical Issues. Technical report TR-CS-06-02, Joint Computer Science Report Series. Canberra: Australian National University.

Conover, W., J. (1999). Practical Nonparametric Statistics, 3rd edition. New York: John Wiley and Sons.

Dahl, Ö. (2007). Språkets enhet och mångfald [The unity and diversity of language]. Lund: Studentlitteratur.

Dalianis, H., Rimka, M. & Kann, V. (2009). Using Uplug and SiteSeeker to construct a cross language search engine for Scandinavian languages. In Proceedings of the Workshop on Automatic Treatment of Multilinguality in Retrieval, Search and Lexicography (Copenhagen, Denmark).

Damerau, F. (1964). A technique for computer detection and correction of spelling errors. Communications of the ACM, 7(3), 171-176.

Darwish, K. & Oard, D. (2003). Probabilistic structured query methods. In Proceedings of the 26th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 338-344.

Dolamic, L. & Savoy, J. (2010). Retrieval effectiveness of machine translated queries. Journal of the American Society for Information Science and Technology, 61(11), 2266-2273.

Ernst-Gerlach, A. & Fuhr, N. (2007). Retrieval in text collections with historic spelling using linguistic and spelling variants. In Proceedings of the 7th Joint Conference on Digital Libraries JCDL'07. New York: ACM, 333-341.

Gotscharek, A., Reffle, U., Ringsletter, C., Schulz, K. & Neumann, A. (2011). Towards information retrieval on historical document collections: the role of matching procedures and special lexica. International Journal on Document Analysis and Recognition, 14(2), 159-171.

Grubinger, M. (2007). Analysis and Evaluation of Visual Information Systems Performance. PhD Thesis. Melbourne: Victoria University.

Grubinger, M., Clough, P., Hanbury, A. & Müller, H. (2007). Overview of the ImageCLEFphoto 2007 photographic retrieval task. In Peters, C. et al. (eds.), Advances in Multilingual and Multimodal Information Retrieval: Proceedings of the eighth Workshop of the Cross-Language Evaluation Forum, CLEF 2007. Lecture Notes in Computer Science, Vol. 5152. Berlin-Heidelberg: Springer, 433-444.

Grubinger, M., Clough, P., Müller, H. & Deselaers, T. (2006). The IAPR benchmark: A new evaluation resource for visual information systems. In Proceedings of OntoImage 2006 - Workshop on Language Resources for Content-based Image Retrieval during LREC 2006 (Genoa, Italy), 13-23.

Hall, P. & Dowling, G. (1980). Approximate string matching. Computing Surveys 12(4), 381-402.

Harding, S. M., Croft, W. B. & Weir, C. (1997). Probabilistic retrieval of OCR degraded text using n-grams. In Research and advanced technology for digital libraries. Berlin-Heidelberg: Springer, 345-359.

Harman, D. (1991). How effective is suffixing?" Journal of the American Society of Information Science, 41(1), 7-15.

Hauser, A., Heller, M., Leiss, E., Schulz, K. U. & Wanzeck, C. (2007). Information access to historical documents from the early new high German period. In Proceedings of IJCAI-07 Workshop on Analytics for Noisy Unstructured Text Data (Hyderabad, India), 147-154.

Hedlund, T. (2002). Compounds in dictionary-based cross-language information retrieval. Information Research, 7(2). Retrieved from http://InformationR.net/ir/7-2/paper128.html.

Hedlund, T., Airio, E., Keskustalo, H., Lehtokangas, R., Pirkola, A. & Järvelin, K. (2004). Dictionary-based cross-language information retrieval: Learning experiences from CLEF 2000-2002. Information Retrieval, 7(1-2), 99-119.

Hersh, W., Turpin, A., Price, S., Chan, B., Kraemer, D., Sacherek, L. & Olson, D. (2000) Do batch mode and user evaluations give the same results? In Proceedings of the 23th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 17-24.

Holley, R. (2009). How good can it get? Analysing and improving OCR accuracy in large scale historic newspaper digitisation programs. D-Lib Magazine, 15(3-4). Retrieved from http://www.dlib.org/dlib/march09/holley/03holley.html.

Hollink, V., Kamps, J., Monz, C. & De Rijke. M. (2004). Monolingual document retrieval for European languages. Information Retrieval, 7(1-2), 33.52.

Hull, D. (1993). Using statistical testing in the evaluation of retrieval experiments. In Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM , 329-338.

Hull, D. & Grefenstette, G. (1996). Querying across languages: A dictionary-based approach to multilingual information retrieval. In Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 49-57.

Häkkinen, K. (1994). Agricolasta nykykieleen. Suomen kirjakielen historia [From Agricola to modern language. The history of standard Finnish]. Helsinki: WSOY.

Ingwersen, P. & Järvelin, K. (2005). The Turn: Integration of Information Seeking and Retrieval in Context.: Dortrecht: Springer.

Joseph, B. (1998). Historical morphology. In Zwicky, A. & Spencer, A. (eds.), The Handbook of Morphology. Malden: Blackwell.

Järvelin, K. & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. ACM Transactions on Information Systems, 20(4), 442-446.

Kanoulas, E., Carterette, B., Clough, P., and Sanderson, M. (2011). Evaluating multi-query sessions. In Proceedings of the 34th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 1053-1062.

Karimi, S., Scholer, F. & Turpin, A. (2011). Machine transliteration survey. ACM Computing Surveys (CSUR), 43(3).

Karlsson, F. (2004). Yleinen kielitiede [General linguistics]. Helsinki: Yliopistopaino.

Karlsson, F. (1983). Suomen kielen äänne- ja muotorakenne [Phonology and morphology of Finnish]. Helsinki: WSOY.

Kekäläinen, J. & Järvelin, K. (2002a). Evaluating information retrieval systems under the challenges of interaction and multidimensional dynamic relevance. In Proceedings of the 4th CoLIS Conference. Greenwood Village: Libraries Unlimited, 253-270.

Kekäläinen, J. & Järvelin, K. (2002b). Using graded relevance assessments in IR evaluation. Journal of the American Society for Information Science and Technology, 53(13), 1120-1129.

Kelly, D., Fu, X. & Shah, C. (2007). Effects of rank and precision of search results on users' evaluations of system performance. Technical Report TR-2007-02. Chapel Hill: University of North Carolina.

Kempken, S., Luther, W. & Pilz, T. (2006). Comparison of distance measures for historical spelling variants. In Bramer, M. (eds.), Artificial Intelligence in Theory and Practice, IFIP Series 219. New York: Springer, 295-304.

Keskustalo, H., Järvelin, K., Pirkola, A., Sharma, T. & Lykke Nielsen, M. (2009). Test collection-based IR evaluation needs extension toward sessions - A case of extremely short queries. In Proceedings of AIRS 2009, the 5th Asia Information Retrieval Symposium. Lecture notes in computer science, 5839. Berlin-New York: Springer, 63-74.

Keskustalo, H., Pirkola, A., Visala, K., Leppänen, E. & Järvelin, K. (2003). Non-adjacent digrams improve matching of cross-lingual spelling variants. In Proceedings of SPIRE 2003, the 10th International Symposium on String Processing and Information Retrieval. String Processing and Information Retrieval, vol. 10. Berlin/New York: Springer, 252-265.

Kettunen, K. (2004). Covering the morphological variation of Finnish query nouns in a probabilistic best-match system. In Proceedings of The First Baltic Conference, Human Language technologies - The Baltic Perspective. Frontiers in Artificial Intelligence and Applications, vol. 219. Fairfax: IOS Press, 73 - 80.

Kettunen, K. (2009). Reductive and generative approaches to management of morphological variation of keywords in monolingual information retrieval - an overview. Journal of Documentation, 65(2), 267-290.

Kettunen, K. (2013). Managing word form variation of text retrieval in practice - Why language technology is not the only cure for better IR performance? Trends in information management, 9(1).

Kettunen, K. & Airio, E. (2006). Is a morphologically complex language really that complex in full-text retrieval? In Salakoski et al. (eds.), Advances in Natural Language Processing, LNAI 4139. Berlin-Heidelberg: Springer, 411-22..

Kettunen, K., McNamee, P. & Baskaya, F. (2010). Using syllables as indexing terms in full-text information retrieval. In Proceedings of The First Baltic Conference, Human Language technologies - The Baltic Perspective. Frontiers in Artificial Intelligence and Applications, vol. 219. Fairfax: IOS Press, 225-232.

Kishida, K. (2005). Technical issues of cross-language information retrieval: a review. Information Processing and Management. 41(3), 433-455.

Klementiev, A. & Roth, D. (2006). Weakly supervised named entity transliteration and discovery from multilingual comparable corpora. In Proceedings of the 21st International Conference on Computational Linguistics and 44th annual meeting of the Association for Computational Linguistics. New York: ACL, 817-824.

Kondrak, G., Marcu, D. & Knight, K. (2003). Cognates can improve statistical translation models. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (HLT-NAACL 2003): companion volume of the Proceedings of – short papers, 46-48.

Koolen, M., Adriaans, F., Kamps, J. & De Rijke, M. (2006). A cross-language approach to historic document retrieval. In: Lalmas, M., et al., (eds.), Proceedings of 28th European Conference on Information Retrieval Research. Lecture Notes in Computer Science, vol. 3936. Berlin-Heidelberg: Springer, 407-419.

Kukich, K. (1992). Techniques for automatically correcting words in text. ACM Computing Surveys, 24(4), 377-439.

Larkey, L. S., AbdulJaleel, N. & Connell, M. (2003). What's in a name? Proper names in Arabic cross language information retrieval. CIIR Technical Report No. IR-278. Amherst: University of Massachusetts. Retrieved from http://ciir.cs.umass.edu/pubfiles/ir-278.pdf.

Levenshtein, V.I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics - Doklady, 10(8), 707-710.

Levow, G.-A., Oard, D. & Resnik, P. (2005). Dictionary-based techniques for cross-language information retrieval. Information Processing and Management, 41 (3), 523-547.

Lind, D. A., Marchal, W. G. & Wathen, S. A. (2014). Statistical techniques in business & economics. 16th edition. Irwin: McGraw-Hill.

Liu, L-M., Babad, Y., Sun, W. & Chan, K-K. (1991). Adaptive post-processing of OCR text via knowledge acquisition. In Proceedings of the 19th annual conference on computer science (CSC'91). New York: ACM, 558-569.

Loponen, A., Pirkola, A. & Järvelin, K. (2008). An Effective Implementation of the FITE-TRT Method for OOV Word Translation. In McDonald et al. (eds.), Advances in Information Retrieval: Proceedings of the 30th European Conference on IR Research. Berlin-Heidelberg: Springer , 138-149.

Makin, R., Pandey, N., Pingali, P. & Varma, V. (2007). Approximate string matching techniques for effective CLIR among Indian languages. In Masulli, F. et al. (eds.), Applications of Fuzzy Sets Theory: Proceedings of the 7th International Workshop on Fuzzy Logic and Applications. Berlin-Heidelberg: Springer, Italy), 430-437.

McCarley, J. (1999). Should we translate the documents or the queries in cross-language information retrieval? In Proceedings of the 37th annual meeting of the Association for Computational Linguistics. New York: ACL, 208-214.

McNamee, P. (2008). Textual representations for corpus-based bilingual retrieval. Ann Arbor: ProQuest.

McNamee, P. & Mayfield J. (2002). Scalable multilingual information access. In Peters, C. et al. (eds.) Advances in Cross-Language Information Retrieval: Proceedings of the Third Workshop of the Cross-Language Evaluation Forum, CLEF 2002. LNCS 2785, Berlin-Heidelberg: Springer, 207-218.

McNamee, P. & Mayfield J. (2004). Character n-gram tokenization for European Language Text Retrieval. Information Retrieval, 7(1-2), 73-97.

McNamee, P. & Mayfield, J. (2005). Translating pieces of words. In Proceedings of the 28h Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 643-644.

McNamee, P., Nicolas, C. & Mayfield, J. (2009). Addressing morphological variation in alphabetic languages. In Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 75-82.

Menard, E. & Khashman, N. (2014). Image retrieval behaviours: users are leading the way to a new bilingual search interface. Library Hi Tech, 32(1), 50-68.

Mitra, M. & Chaudhuri, B. (2000). Information retrieval from documents: A survey. Information retrieval, 2(2-3), 141-163.

Mittendorf, E. & Schäuble, P. (1996). Measuring the effects of data corruption on information retrieval. In Proceedings of the SDAIR'96 Conference (Las Vegas, Nevada), 179 –189.

Moffat, A. & Zobel, J. (2008). Rank-biased precision for measurement of retrieval effectiveness. ACM Transactions Information Systems, 27(1):2,1-27.

Montalvo, S., Pardo, E., Martínez, R., and Fresno, V. (2012). Automatic cognate identification based on a fuzzy combination of string similarity measures. In

Proceedings of Fuzzy systems (FUZZ-IEEE), IEEE World Congress on Computational Intelligence. New York: IEEE, 1-8.
Mustafa, S. (2005). Character contiguity in N-gram-based word matching: the case for Arabic text searching. Information Processing and Management, 41(4), 819-827.
Navarro, G. (2001). A guided tour to approximate string matching. ACM Computing Surveys, 33(1), 31-88.
O'Hare, N., Wilkins, P., Gurrin, C., Newman, E., Jones & G., Smeaton, A. (2009). Diversity in image retrieval: DCU at ImageCLEFPhoto 2008. In Peters, C. et al. (Eds.), Evaluating Systems for Multilingual and Multimodal Information Access: Proceedings of the 9th Workshop of the Cross-Language Evaluation Forum. Berlin-Heidelberg: Springer, 500-511.
O'Rourke, A., Robertson, A. & Willett, P. (1997). Word variant identification in old French. Information research, 2(4), paper 22. Retrieved from http://informationr.net/ir/2-4/paper22.html.
Pearce, C. & Nicholas, C. K. (1996). TELLTALE: Experiments in a dynamic hypertext environment for degraded and multilingual data. Journal of the American Society for Information Science and Technology, 47(4), 263-275.
Pfeifer, U., Poersch, T. & Fuhr, N. (1996). Retrieval effectiveness of proper name search methods. Information Processing & Management, 32(6), 667-679.
Pilz, T., Luther, W., Fuhr, N. & Ammon, U. (2006). Rule-based search in text databases with nonstandard orthography. Literary and Linguistic Computing, 21(2), 179-186.
Pilz, T., Ernst-Gerlach, A., Kempken, S., Rayson, P. & Archer, D. (2008). The identification of spelling variants in English and German historical texts: Manual or automatic? Literary and Linguistic Computing, 23(1), 65-72.
Pirkola, A. (1998). The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 55-63.
Pirkola, A., Hedlund, T., Keskustalo, H. & Järvelin, K. (2001). Dictionary-based cross-language information retrieval: problems, methods, and research findings. Information Retrieval, 4(3/4), 209-230.
Pirkola, A., Keskustalo H., Leppänen, E., Känsälä, A.P. & Järvelin, K. (2002). Targeted s-gram matching: a novel n-gram matching technique for cross- and monolingual word form variants. Information research, 7(2), paper 126. Retrieved from http://InformationR.net/ir/7-2/paper126.html.
Pirkola, A., Puolamäki, D. & Järvelin, K. (2003a). Applying query structuring in cross-language retrieval. Information Processing & Management, 39(3), 391-402.
Pirkola, A., Toivonen, J., Keskustalo, H., Visala, K. & Järvelin, K. (2003b). Fuzzy translation of cross-lingual spelling variants. In Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 345 - 352.
Pirkola, A., Toivonen, J., Keskustalo, H. & Järvelin, K. (2006). FITE-TRT: A high quality translation technique for OOV words. In Proceedings of the 2006 ACM symposium on Applied computing. New York: ACM, 1043-1049.
Pu, H-T. (2005). A comparative analysis of web image and textual queries. Online Information Review, 29(5), 457-467.
Raitanen, I. (2012). "Etsikäät hywää ja älläät pahaa". Tiedonhakumenetelmien tuloksellisuuden vertailu merkkivirheitä sisältävässä historiallisessa

sanomalehtikokoelmassa [Comparison of the effectiveness of information retrieval methods in historical newspaper collections containing character errors]. Master's thesis. Tampere: University of Tampere. Retrieved from http://urn.fi/urn:nbn:fi:uta-1-22596.

Rice, S., Kanai, J. & Nartker, T. (1993). An evaluation of OCR accuracy. Information Science Research Institute, Annual Research Report 1993, Las Vegas: University of Nevada.

Rijsbergen, C. V. (1979). Information Retrieval. 1979. London: Butterworths.

Robertson, A. M. & Willett, P. (1993). A comparison of spelling correction methods for the identification of word forms in historical text databases. Literary and linguistic computing, 8(3), 143-152.

Robertson, A. M. & Willett, P. (1998). Applications of n-grams in textual information systems. Journal of Documentation, 54(1), 48–69.

Ruiz, M., Chen, J., Pasupathy, K., Chin, P. & Knudson, R. (2010). UNT at ImageCLEF 2010: CLIR for Wikipedia images. CLEF 2010 working notes. Retrieved from http://www.clef-initiative.eu/edition/clef2010/working-notes.

Sanderson, M. (2010). Test collection based evaluation of information retrieval systems. Foundations and Trends in Information Retrieval, 4(4), 247-375.

Savoy, J. & Naji, N. (2011). Comparative information retrieval evaluation for scanned documents. In Proceedings of the 15th WSEAS international conference on Computers. Athens: WSEAS Press, 527-534.

Scholer, F. & Turpin, A. (2009). Metric and relevance mismatch in retrieval evaluation. In proceedings of AIRS 2009, the 5th Asia Information Retrieval Symposium. Berlin-Heidelberg: Springer, 50-62.

Seifart, F. (2006). Orthography development. In Gippert, J. et al. (Eds.), Essentials of language documentation. Berlin-New York: Mouton de Gruyter, 275-299.

Sheridan, P. & Ballerini, J. P. (1996). Experiments in multilingual information retrieval using the SPIDER system. In Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 58-65.

Sherif, T. & Kondrak, G. (2007). Substring-based transliteration. Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. New York: ACL, 944–951.

Smith, C. L., and Kantor, P. B. (2008). User adaptation: Good results from poor systems. In Proceedings of the 31st International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 147-154.

Smucker, M. (2009) Towards timed predictions of human performance for interactive information retrieval evaluation. In Proceedings of the third Workshop on Human-Computer Interaction and Information Retrieval (HCIR'09) (Washington DC, USA). Retrieved from http://www.mansci.uwaterloo.ca /~msmucker/publications/smucker-hcir-2009.pdf

Smucker, M., Allan, J. & Carterette, B. (2007). A comparison of statistical significance tests for information retrieval evaluation. In Proceedings of the 16th ACM conference on Conference on information and knowledge management. New York: ACM, 623-632.

Soboroff, I. (2004) On evaluating web search with very few relevant documents. In Proceedings of the 27th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 530-531.

Sormunen, E. (2000). A Method for Measuring Wide Range Performance of Boolean Queries in Full-Text Databases. Doctoral Thesis. Acta Electronica Universitatis Tamperensis 34. Tampere: University of Tampere. Retrieved from http://acta.uta.fi/teos.phtml?3786.

Sormunen, E. (2002). Liberal relevance criteria of TREC: counting on negligible documents? In Proceedings of the 25th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 324-330.

Spärck Jones, K. (1974). Automatic indexing, Journal of Documentation, 30(4), 393-432.

Strohman, T., Metzler, D., Turtle, H. & Croft, W. B. (2005). Indri: A language model-based search engine for complex queries. In Proceedings of the International Conference on Intelligent Analysis, 2(6).

Taghva, K., Borsack, J. & Condit, A. (1994). Results of applying probabilistic IR to OCR text. In Proceedings of the 17th annual international ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 202-211.

Taghva, K. & Stofsky, E. (2001). OCRSpell: an interactive spelling correction system for OCR errors in text. International Journal on Document Analysis and Recognition 3(3), 125-137.

Tague-Sutcliffe, J. (1992). The pragmatics of information retrieval experimentation, revisited. Information Processing and Management, 28(4), 467-490.

Talvensaari, T., Laurikkala, J., Järvelin, K. & Juhola M. (2007). Creating and exploiting a comparable corpus in cross-language information retrieval. ACM Transactions on Information Systems, 25 (1), article 4.

Toivonen, J., Pirkola, A., Keskustalo, H., Visala, K. & Järvelin, K. (2005). Translating cross-lingual spelling variants using transformation rules. Information Processing & Management, 41(4), 859-872.

Tsikrika, T., Popescu, A. & Kludas, J. (2011). Overview of the Wikipedia image retrieval task at ImageCLEF 2011. In CLEF (Notebook Papers/Labs/Workshop). Retrieved from http://ims-sites.dei.unipd.it/documents/ 71612/86377/CLEF2011wn-ImageCLEF-TsikrikaEt2011.pdf.

Turpin, A. & Scholer, F. (2006). User performance versus precision measures for simple search tasks. In Proceedings of the 29th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 11-18.

Turpin, A. & Hersh, W. (2001). Why batch and user evaluations do not give the same results. In proceedings of the 24th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 225-231.

Turpin, A., Scholer, F., Järvelin, K., Wu, M. & Culpepper, J. S. (2009). Including summaries in system evaluation. In Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval. New York: ACM, 508-515.

Udupa, R., Saravanan, K., Bakalov, A. & Bhole, A. (2009). They are out there, if you know where to look: Mining transliterations of OOV query terms for cross-language information retrieval. In Advances in Information Retrieval: Proceedings of the 31th European Conference on IR Research, ECIR 2009. Berlin-Heidelberg: Springer, 437-448.

Ukkonen, E. (1992). Approximate string-matching with q-grams and maximal matches. Theoretical Computer Science, 92(1), 191–211.

Ullman, J. R. (1977). A binary n-gram technique for automatic correction of substitution, deletion, insertion and reversal errors in words. The Computer Journal, 20(2), 141-147.

Vilares, J., Oakes, M. P. & Vilares, M. (2007). Character n-grams translation in cross-language information retrieval. In Kedad, Z. et al. (eds.), Natural Language Processing and Information Systems: Proceedings of the 12th International Conference on Applications of Natural Language to Information Systems. Berlin-Heidelberg: Springer, 217-228.

Voorhees, E. (1999). The TREC-8 question answering track report. In Proceedings of the Eighth Text REtrieval Conference (TREC-8), NIST Special Publication 500-246. Gaithersburg: NIST, 77-82.

Voorhees, E. (2001). Evaluation by highly relevant documents. In Proceedings of the 24th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 74-82.

Voorhees, E. (2002). The philosophy of information retrieval evaluation. In Peters, C. et al. (Eds.), Evaluation of Cross-Language Information Retrieval Systems: Proceedings of the Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001. Berlin-Heidelberg: Springer, 355-370.

Voorhees, E., and Buckley, C. (2002). The effect of topic set size on retrieval experiment error. In Proceedings of the 25th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 316-323.

Yoon, J. (2011). Searching images in daily life. Library and Information Science Research, 33(), 269-275.

Zamora, E. M., Pollock, J. J. & Zamora, A. (1981). The use of trigram analysis for spelling error detection. Information Processing and Management, 17(8), 305–316.

Zhou, D., Truran, M., Brailsford, T., Wade, V. & Ashman, H. (2012). Translation techniques in cross-language information retrieval. ACM Computing Surveys, 45(1), article 1.

Zobel, J. & Dart, P. (1996). Phonetic string matching: Lessons from information retrieval. In Proceedings of the 19th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 166-172.

Available online at www.sciencedirect.com
ScienceDirect
Information Processing and Management 43 (2007) 1005–1019
INFORMATION PROCESSING & MANAGEMENT
www.elsevier.com/locate/infoproman

s-grams: Defining generalized *n*-grams for information retrieval

Anni Järvelin [a,*], Antti Järvelin [b], Kalervo Järvelin [a]

[a] *University of Tampere, Department of Information Studies, FIN-33014 University of Tampere, Finland*
[b] *University of Tampere, Department of Computer Sciences, FIN-33014 University of Tampere, Finland*

Received 1 June 2006; received in revised form 24 September 2006; accepted 26 September 2006
Available online 22 November 2006

Abstract

n-grams have been used widely and successfully for approximate string matching in many areas. *s*-grams have been introduced recently as an *n*-gram based matching technique, where di-grams are formed of both adjacent and non-adjacent characters. *s*-grams have proved successful in approximate string matching across language boundaries in Information Retrieval (IR). *s*-grams however lack precise definitions. Also their similarity comparison lacks precise definition. In this paper, we give precise definitions for both. Our definitions are developed in a bottom-up manner, only assuming character strings and elementary mathematical concepts. Extending established practices, we provide novel definitions of *s*-gram profiles and the L_1 distance metric for them. This is a stronger string proximity measure than the popular Jaccard similarity measure because Jaccard is insensitive to the counts of each *n*-gram in the strings to be compared. However, due to the popularity of Jaccard in IR experiments, we define the reduction of *s*-gram profiles to binary profiles in order to precisely define the (extended) Jaccard similarity function for *s*-grams. We also show that *n*-gram similarity/distance computations are special cases of our generalized definitions.

Keywords: Information retrieval; Approximate string matching; *n*-grams; *s*-grams

1. Introduction

s-gram matching is an approximate string matching technique, where the text strings compared are decomposed into *s*-grams, i.e., into fixed length substrings. The degree of similarity between the strings can be computed by comparing their *s*-gram sets. The idea dates back to Shannon's Mathematical Theory of Communication (1948) and it has in earlier literature also been referred to as *n*-grams (e.g. Pfeiffer, Poersch, & Fuhr, 1996; Robertson & Willett, 1998) or *q*-grams (e.g. Ukkonen, 1992; Zobel & Dart, 1996). Typically *n*-grams consist of adjacent character pairs, triples etc. of the original strings. The name *s*-gram originates from a study by Pirkola, Keskustalo, Leppänen, Känsälä, and Järvelin (2002), who devised a novel classified *s*-gram matching technique. In this technique the *s*-grams are formed of both adjacent and non-adjacent characters of

* Corresponding author.
E-mail addresses: anni.jarvelin@uta.fi (A. Järvelin), antti.jarvelin@uta.fi (A. Järvelin), kalervo.jarvelin@uta.fi (K. Järvelin).

0306-4573/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2006.09.016

the text strings and classified into sets for computing the similarity. The name *s*-gram comes from the word skip and points to the idea that a number of characters are skipped when the substrings (*s*-grams) are formed. In the original study, Pirkola et al. (2002) found classified *s*-grams better than the adjacent *n*-grams in matching cross-lingual spelling and monolingual morphological variants.

In the literature, the terminology used for referring to *n*-grams and *s*-grams has varied. In this paper, the term *n-gram* will be used to refer to the adjacent *n*-grams only. The term *s-gram* is used when referring to both the adjacent and non-adjacent *s*-grams. Thus *n*-grams are a special case of *s*-grams, where zero characters are skipped when the substrings are formed. Di-grams are conventional *n*-grams where the substrings' length is two ($n = 2$), for tri-grams n is three. When referring to different length *s*-grams expressions such as *s*-di-grams and *s*-tri-grams will be used. Furthermore, if also the skip length of an *s*-gram needs to be specified, *s*-grams of length n and skip-length k will be referred as $s_{n,k}$-grams.

n-gram matching is a widely used technique both within IR and outside (e.g. Downie & Nelson, 2000; Grossman & Frieder, 2004; Keskustalo, Pirkola, Visala, Leppänen, & Järvelin, 2003; McNamee & Mayfield, 2003; O'Rourke, Robertson, & Willettt, 1997; Pirkola et al., 2002; Pfeiffer et al., 1996; Robertson & Willett, 1998; Toivonen, Pirkola, Keskustalo, Visala, & Järvelin, 2005; Uitdenbogerd & Zobel, 2002; Ukkonen, 1992; Ullman, 1977; Zobel & Dart, 1996). Ukkonen (1992) gives a formal definition to *n*-grams. However, the classified *s*-grams found superior to *n*-grams in information retrieval applications lack such defendable definitions. Also their similarity comparison (Pirkola et al., 2002) lacks a stringent definition. This paper provides stringent definitions both for *s*-grams and their classified similarity comparison. We do this firstly by formalizing the *similarity* comparison proposed in (Pirkola et al., 2002). Secondly, we propose the use of *distance* measures used earlier in the literature and extend them for classified *s*-grams. As these have not been tested empirically so far, our work suggests that new empirical work be done.

The rest of the paper is organized as follows. Approximate string matching techniques and the use of *s*-grams in IR are discussed in Section 2. In Section 3 *s*-grams are presented in more detail and an example application of the *s*-gram matching technique is given. Section 4 discusses prior formalizations of *n*-grams and Section 5 presents our formalization of *s*-grams with *n*-grams as a special case. Finally, Section 6 provides discussion and conclusions.

2. Approximate string matching techniques in IR

Approximate string matching techniques are based on the expectation that two strings that have similar strings of characters will have similar meaning and should therefore be regarded as being equivalent (Robertson & Willett, 1998). Recognizing and measuring similarity in strings is useful in several situations in IR as both natural morphological word form variation and variation due to e.g. typing errors or OCR (optical character recognition) errors occur in databases. Recognizing such word form variants as occurrences of the same string is essential as different forms of a word represent the same concept and are therefore equal from the standpoint of users' requests (Pirkola et al., 2002).

Cross-lingual spelling variation is a type of word form variation occurring between languages. Especially related languages often share a number of words that have the same origin (e.g. Latin based words) and only differ due to the orthographical differences between the languages. Proper names and technical terms are typical examples of words where cross-lingual spelling variation occurs (e.g. Brussels in Finnish is Bryssel). Approximate string matching can be used in cross-language information retrieval (CLIR) to recognize cross-lingual spelling variants as equivalents. In CLIR, a source language query is typically translated into the target language with machine-readable dictionaries. The general translation dictionaries often do not cover most proper names and technical terms, which therefore remain untranslatable in queries. They can nevertheless often be recognized as similar by approximate string matching, and therefore translated. This has a positive effect on query translation as proper names and technical terms often are important query keys (Pirkola et al., 2002).

The approximate similarity between strings can be measured with different methods. Here, Soundex and its variant Phonix, Edit distance (ED), Longest common subsequence (LCS) and the *s*-gram technique will be discussed. Soundex is an early phonetic matching scheme of Odell and Russell from 1918 (Hall & Dowling, 1980) and Phonix its newer variant developed by Gadd in late 1980's (Zobel & Dart, 1996). Matching schemes

based on comparing the phonetic similarity of strings are language dependent techniques. Soundex and Phonix were developed for the English language but can be modified for other languages (Pfeiffer et al., 1996). They use phonetic codes based on grouping similar sounding letters together. Strings sharing the same code are assumed to be similar (Zobel & Dart, 1996). Phonix also uses rules for letter-group transformations to provide context for the phonetic codes. Especially Soundex makes quite commonly the error of transforming dissimilar-sounding strings into the same code, and Pfeiffer et al. (1996) found both Soundex and Phonix clearly inferior to *n*-grams in proper name matching. Zobel and Dart (1996) tested various string similarity techniques for phonetic matching and found that an edit distance variant Editex that uses the letter-groupings of Soundex outperformed both Soundex and Phonix.

Edit distances (ED) are distance measures, which count the minimal number of single character instertions, deletions and replacements needed to transform one string to another. Different operations can be assigned different costs, depending on their likelihood (Navarro, 2001). ED is sometimes referred to as Damerau–Levenstein metric as Damerau and Levenstein developed the metric separately during the 1960's. Damerau developed an early edit distance measure for handling spelling errors, which accepts a difference of one insertion, deletion, replacement or transposition of a character in the strings compared (Damerau, 1964). Pfeiffer et al. (1996) and Zobel and Dart (1996) studied different approximate string matching techniques in name matching and ED was tested in both studies. While both found that combining evidence from different string matching techniques was the best solution, the results concerning the individual techniques and specially ED and *n*-grams diverged: Pfeiffer et al. (1996) found *n*-grams clearly a better technique than the ED, whereas Zobel and Dart (1996) reported that ED outperformed *n*-grams.

LCS is a string matching technique that measures the length of the longest pairing of characters that can be made between two strings, so that the pairings respect the order of the letters (Navarro, 2001). O'Rourke et al. (1997) found LCS to be the best method for matching historical word form variants (compared to di- and tri-grams). They nevertheless concluded that di-grams would be the method of choice in an operational system due to LCS's high demand on computational time. Keskustalo et al. (2003) compared *s*-grams to different length *n*-grams, ED and LCS in matching cross-lingual spelling variants and found that, where ED often outperforms the adjacent *n*-grams, the classified *s*-grams performed better than ED for all the six language pairs studied. LCS was always inferior to *s*-grams and ED.

The modern applications of the *n*-gram matching in IR are discussed in, e.g., Grossman and Frieder (2004) and in Robertson and Willett (1998). The *n*-gram matching, and its generalization *s*-gram matching, are language independent techniques and can therefore be easily applied to all languages in which the strings consist of space- or punctuation-delimited sequences of characters (Robertson & Willett, 1998). Grossman and Frieder (2004) point to *n*-gram applications in handling OCR errors, spelling error correction, text compression, authorship detection, and discuss applications in traditional text retrieval at some length. The present article has its focus on *s*-gram matching in information retrieval (IR) and above all in cross-language information retrieval (CLIR) and therefore mainly IR research is discussed. Results from several studies (Pfeiffer et al., 1996; Pirkola et al., 2002; Keskustalo et al., 2003) support the proposition that *s*-gram matching is the choice approximate string matching technique in IR and CLIR and therefore an extensive discussion of other approximate matching techniques is not provided here.

Biological and genetic IR are application areas of *s*-grams, where the strings compared can be thousands of characters long and therefore higher value of n may be used (Altschul, Gish, Miller, Myers, & Lipman, 1990; Bishop & Thompson, 1984; Miller, Gurd, & Brass, 1999). Califano and Rigoutsos (1993) proposed a $s_{n,k}$-gram method for matching molecule biological sequences. For large values of n and k (e.g., $n > 10$) the proposed method led to serious efficiency problems due to the size of the index needed for retrieval. Similar problems are faced when word-spanning *n*-grams are used for indexing document collections instead of word-based indexing (McNamee & Mayfield, 2004): When the strings decomposed to *n*-grams are long (queries, documents) and the number of unique *n*-grams in any collection is bounded by $|\Sigma|^n$, where $|\Sigma|$ denotes the size of an alphabet Σ, the index size grows rapidly with n. The value of n appropriate for *n*-gram based indexing varies with the language considered. The optimal value of n for European languages is between 4 and 5 (McNamee & Mayfield, 2004), as for Chinese, where word based indexing faces difficulties because of the short comings of the programs used for recognizing word boundaries, di-gram based indexing has been popular (Chen, He, Xu, Gey, & Meggs, 1997) and index size is not a problem.

The definitions given for n-grams in the present article hold for these applications. The approaches are nevertheless different: in n-gram based indexing the value of n needs to be set high enough to ensure the discriminating power of the indexing features (n-grams). In the present article the n-grams are seen as a word level string similarity measure used before a query is matched against a document collection. The problems with long text passages and high values of n are not addressed with length, as the average word length is well under ten characters and as the focus is on (CL)IR, where n can be limited to low values. The effect that the classified s-grams special features have for the index size is discussed in more detail in Section 6.

3. *s*-grams

3.1. *s*-gram basics

The classified s-gram technique was introduced as a solution for monolingual morphological and cross-lingual spelling variation problems in IR (Pirkola et al., 2002) and its performance has been tested with several language pairs (Pirkola et al., 2002; Keskustalo et al., 2003). In the classified s-gram matching technique s-grams are divided into categories (classes) on the basis of the number of the skipped characters and only the s-grams belonging to the same class are compared to each other. *Skip-gram class* indicates the skip length used when generating a set of s-grams belonging to a class. Two or more skip-gram classes may also be combined into more general skip-gram classes (Pirkola et al., 2002; Keskustalo et al., 2003). The *character combination index* (CCI) then indicates the set of all the skip-gram classes to be formed from a string. Different combinations of skipped characters can be used. For example the CCI $\{\{0\},\{1,2\}\}$ means that two skip-gram classes are formed from a string: one with conventional n-grams formed of adjacent characters and one with s-grams formed both by skipping one and two characters. An example of forming the skip-gram classes is given in Table 1. The largest value in a skip-gram class is called the spanning length of the skip-gram class (Keskustalo et al., 2003), e.g., for skip-gram class $\{0,1\}$, the spanning length is one.

Different skip-gram classes carry forward different evidence from their host string and s-grams can therefore be tuned to handle different phenomena by adjusting the skip-gram classes. Keskustalo et al. (2003) gives a good presentation on how the skip-gram classes relate to different variation in strings from the cross-lingual spelling variation point of view. Cross-lingual spelling variation typically involves single character insertions, deletions and substitutions, or their two-character combinations. For example transforming Swedish variant *heksaklorid* into the English *hexachloride* involves a single insertion (e) and combinations of deletion and substitution ($ks \rightarrow x$) and substitution and insertion ($k \rightarrow ch$) Keskustalo et al. (2003). Therefore it is reasonable to use only skip-gram classes with spanning length of two or less when matching cross-lingual spelling variants. Also spelling errors typically involve character substitution, insertion, deletion and reversal errors and their combinations (Ullman, 1977). The spanning length of s-grams can be restricted to two or less again, as Zamora, Pollock, and Zamora (1981) have reported that most of the misspelled strings in text databases only contain a single error.

It is common to use padding spaces in the beginning and in the end of the strings when forming s-grams. The padding helps to get the characters at the beginning and at the end of a string properly presented in its s-gram set. For conventional n-grams it is common to use a padding of $n-1$ characters (Robertson & Willett, 1998). For s-grams a padding that varies together with the length of the substring (n) and the number of the skipped characters can be used. In accordance with these rules, the set of padded $s_{2,0}$-grams for the string *abra-*

Table 1
The skip-gram classes for forming the s-di-grams with different CCIs for the string *abradacabra*

Type	CCI	s-gram classes
$s_{2,0}$	$\{0\}$	$\{ab, br, ra, ad, da, ac, ca, ab, br, ra\}$
$s_{2,1}$	$\{1\}$	$\{ar, ba, rd, aa, dc, aa, cb, ar, ba\}$
$s_{2,2}$	$\{2\}$	$\{aa, bd, ra, ac, da, ab, cr, aa\}$
$s_{2,\{1,2\}}$	$\{1,2\}$	$\{ar, ba, rd, aa, dc, cb, ar, bd, ra, ac, da, ab, cr\}$
$s_{2,\{\{0\},\{1,2\}\}}$	$\{\{0\},\{1,2\}\}$	$\{\{ab, br, ra, ad, da, ac, ca, ab, br, ra\}, \{ar, ba, rd, aa, dc, cb, ar, bd, ra, ac, da, ab, cr\}\}$

dacabra is: {_*a*, *ab*, *br*, *ra*, *ad*, *da*, *ac*, *ca*, *ab*, *br*, *ra*, *a*_}. Keskustalo et al. (2003) tested with several languages different types of padding spaces for conventional di-grams, tri-grams and *s*-di-grams, and found that using padding spaces both in the beginning and the end of the words gave the best results. However, leaving out the padding spaces can help down-weighting the derivational suffixes and prefixes, when handling morphological variation or cross-lingual spelling variants in inflected forms. For example, the use of the beginning padding only has been found beneficial for Finnish, which is an inflectionally complex suffix language (Pirkola et al., 2002).

3.2. Motivating example

The typical dictionary-based query translation approach to CLIR has a downside in the constant need for updating the dictionaries and in that different dictionaries are needed for each language pair. These features can make the approach costly and replacing it by cheaper, language independent techniques would be desirable. Between closely related languages that share a high number of cross-lingual spelling variants, approximate string matching techniques can be used for a simpler, fuzzy query translation technique. A fuzzy *s*-gram query translation between the Scandinavian languages Norwegian and Swedish is explored here to give an example of a *s*-gram matching application.[1] Norwegian and Swedish are closely related languages that share a high number of cross-lingual spelling variants: around 90% of the vocabularies of the languages are similar having only some orthographical and inflectional differences (Baròdal, Jörgensen, Larsen, & Martinussen, 1997). This provides a good basis for fuzzy translation.

Our fuzzy query translation experiment was done using adjacent di-grams and classified *s*-di-grams for translating Norwegian search topics to Swedish. The goal was to, with fuzzy techniques, reach translation quality sufficient to enable effective searching of a Swedish document collection and competitive with the dictionary translation. *s*-grams were formed with two different character combination indices: *s*-gram1's with CCI $\{\{0\},\{0,1\},\{1,2\}\}$ and *s*-gram2's with CCI $\{\{0\},\{1,2\}\}$. Only the better performing *s*-gram1's are discussed in the following. The fuzzy translation was compared to dictionary-based query translation and to monolingual Swedish IR. Also a Norwegian baseline query was formed to see how much the fuzzy translation improves the results compared to no translation at all.

A typical CLIR test setting with the search topics and test collection from Cross-Language Evaluation Forum (CLEF) 2003 was used. The CLEF'03 environment includes document collections and a set of 60 search topics in several languages, including Swedish (Peters, 2003). The Swedish document collection contains 142819 news wire articles from the Swedish news agency TT published in 1994–1995. As the Norwegian search topics were not included in the CLEF test environment, the English test topics were translated to Norwegian by a native Norwegian-speaker. The document collection and the search topics were not morphologically preprocessed for the fuzzy translation—the words were used in the inflected word forms in which they appeared in text. The collection and the topics were nevertheless normalized for the dictionary-based translation and monolingual IR, to ensure hard baselines. The stop words, as well as the duplicates of words appearing in identical word forms, were removed from the search topics. Finally, six of the search topics did not have any relevant documents in the Swedish document collection and were therefore removed from the topics. The tests were run with the remaining 54 topics.

Test queries were formed from the words of the title and description fields of the test topics (on average 7.5 words after duplicates were removed). The fuzzy translation was done by translating the Norwegian topics into Swedish by matching the *n*- or *s*-grams of the topic words against the Swedish document collection's index words' *n*- or *s*-grams. The three best matches were selected as translations for each word. For the dictionary translation, the GlobalDix dictionary by Kielikone Ltd. was used. All the translations for each topic word were selected to the query and the queries were structured according to the Pirkola method (Pirkola, 1998). The Norwegian and Swedish baseline queries were formed directly from the Norwegian and Swedish topics' words. The performance of the translation techniques was measured by interpolated mean average precision

[1] Fuzzy translation in CLIR between related languages has received some attention before: McNamee and Mayfield (2004) studied a fuzzy query translation technique based on comparing the source language query keys' *n*-grams directly to document collection's *n*-gram index, the results being promising.

over recall levels 10–100 averaged over all queries and as average precision at the recall level 10. A precision-recall graph was created using ten recall levels (10–100). The statistical significance of the results was tested using the Wilcoxon signed ranks test.

The results are presented in Table 2 and in Fig. 1. The statistical significance levels are given in the table. The *s*-grams achieved on average 80% of the dictionary baseline's performance and 62% of the monolingual Swedish baselines performance. The differences in the techniques' average precisions are statistically significant (Table 2). Still, 80% of dictionary baselines performance is a result that shows that *s*-gram translation is a promising and interesting query translation technique in CLIR between related languages. Especially so, when the language dependent dictionary translation setting requires two morphological analyzers (Norwegian and Swedish) and a translation dictionary to function, while the *s*-gram translation only needs a (language independent) program for producing the *s*-grams. When comparing the precision at the higher ranks, i.e., among the 20 and the 50 first retrieved documents (The document cut-off value (DCV) at 20 and 50 retrieved documents) and at the 10% recall level, the differences between the translation techniques are not statistically significant. These documents placed at the top of the result list are the most important ones from the practical user perspective.

The *s*-grams clearly outperformed the Norwegian baseline, with an average precision statistically significantly better than that of the Norwegian baseline (Table 2). The high percentual difference in performance suggests that the difference also has practical significance. It can be seen from Fig. 1 that the *s*-gram technique's precision-recall curve settles clearly above the *n*-gram technique's curve at recall levels 10–70 and always clearly above the Norwegian baseline (nobase). The difference in the *s*- and *n*-gram techniques' average precision, 16.3%, is not statistically significant. The fuzzy *s*-gram translation can be further improved by combining it to other fuzzy techniques (see e.g. Toivonen et al. (2005) for a description of an efficient fuzzy translation technique for words missing from dictionaries that combines *n*-grams to statistical rewriting rules). Therefore, it can be concluded that the *s*-gram translation is a promising technique in query translation between closely related languages.

Table 2
Interpolated average precision values (over recall levels 10–100) and the differences between the techniques (%)

Technique	Nobase	Swebase	Dicbase	*n*-gram	*s*-gram1
Average precision	13.38	35.6	28.4	19.6	22.8
Difference to nobase (%)			+112.3**	+46.5*	+70.4**
Difference to swebase (%)			−20.2	−44.9**	−36**
Difference to dicbase (%)				−31**	−19.7*
Difference to *n*-gram (%)					+16.3

Statistical significance levels are indicated in the table: * = statistically significant difference at level 0.01, ** = statistically highly significant difference at level 0.001.

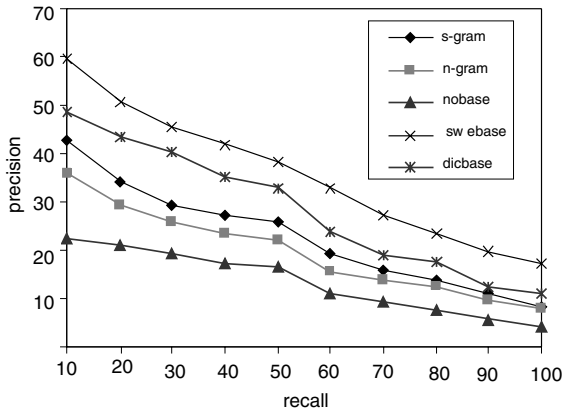

Fig. 1. Precision-recall graph over ten recall levels 10–100.

4. Formalizations of *n*-grams and their proximity functions

We begin by briefly reviewing the basic concepts of formal languages. We follow the conventions of Hopcroft, Motwani, and Ullman (2001). Then we define *n*-grams by following the conventions of Ukkonen (1992). Finally we give two proximity functions for strings based on their *n*-gram overlap. The first one is the L_1 metric originally defined for *n*-grams by Ukkonen (1992) and the second one is Jaccard's similarity function, which is very often used in IR experiments involving *n*-grams.

4.1. Basic concepts of formal languages

We define an *alphabet* Σ as a finite, non-empty set of symbols. A *string* is a finite sequence of symbols from given alphabet. For example, if we have an alphabet $\Sigma = \{a, b\}$, then a, b, ab and bba are strings over Σ. There is also the *empty string* ϵ which does not contain any symbols. Note that an empty string can be chosen from any alphabet. The *length* $|w|$ of a string w is the number of positions for symbols in the string. For instance $|abba| = 4$ and $|\epsilon| = 0$. A string $v = b_1 b_2 \ldots b_m$ is a *substring* of string $w = a_1 a_2 \ldots a_n$, $m \leqslant n$, if $b_1 b_2 \ldots b_m = a_i a_{i+1} \ldots a_{i+m-1}$ for some i, $1 \leqslant i \leqslant n$. If $w = abb$, then ϵ, a, ab, abb, bb and b are substrings of w.

If Σ is an alphabet we denote Σ^k a set of strings over Σ whose length is k. For example, if $\Sigma = \{a, b\}$, then $\Sigma^0 = \{\epsilon\}$, $\Sigma^1 = \{a, b\}$, $\Sigma^2 = \{aa, ab, ba, bb\}$ and $\Sigma^3 = \{aaa, aab, aba, baa, abb, bab, bba, bbb\}$. The set of all strings over an alphabet Σ is denoted Σ^*. In other words, $\Sigma^* = \Sigma^0 \cup \Sigma^1 \cup \Sigma^2 \cup \Sigma^3 \cup \ldots$.

For strings v and w, their *concatenation* vw is a string, which first has all the symbols of v in order of their appearance in v and then all the symbols in w in order of their appearance in w. Thus, the concatenation of strings $v = abba$ and $w = babba$ is $vw = abbababba$.

4.2. Definitions for n-grams

As we saw in previous sections, comparing two strings by calculating the overlap of their common substrings of certain length has a wide range of applications in IR. Now we are ready to give formal definitions for *n*-grams and their selected proximity functions.

Let Σ be a finite alphabet. An *n*-gram is any string $w \in \Sigma^n$. For example, if $\Sigma = \{a, b\}$ its di-grams are aa, ab, ba and bb. To be able to derive proximity functions between strings based on their *n*-gram overlap we need the following definition of string's *n*-gram profile (Ukkonen, 1992).

Definition 1. Let $w = a_1 a_2 \ldots a_m \in \Sigma^*$ and let $x \in \Sigma^n$ be a *n*-gram. If $a_i a_{i+1} \ldots a_{i+n-1} = x$ for some i, then w has an occurrence of x. Let $G(w)[x]$ denote the total number of occurrences of x in w. The *n*-gram profile of w is the vector $G_n(w) = (G(w)[x]), x \in \Sigma^n$.

Now, the distance of the strings can be defined as the L_1 norm (a.k.a Manhattan distance or city block distance) of the difference of their *n*-gram profiles (Ukkonen, 1992).

Definition 2. Let $v, w \in \Sigma^*$ and $n \in \mathbb{N}^+$. The *n*-gram distance between v and w is

$$D_n(v, w) = \sum_{x \in \Sigma^n} |G(v)[x] - G(w)[x]|. \tag{1}$$

Example 1. Let $\Sigma = \{a, b\}$ and $v = abba$, $w = babba \in \Sigma^*$. Their di-gram profiles, listed in lexicographical order of the di-grams, are $(0, 1, 1, 1)$ and $(0, 1, 2, 1)$. Thus, their di-gram distance $D_2(v, w) = 1$. Table 3 lists the di-grams of strings v and w and all di-grams over alphabet $\Sigma = \{a, b\}$.

The *n*-gram distance is pseudo metric (Ukkonen, 1992), i.e., for all $v, w, x \in \Sigma^*$:

(1) $D_n(v, w) \geqslant 0$ (non-negativity),
(2) $D_n(v, w) = D_n(w, v)$ (symmetry) and
(3) $D_n(v, w) \leqslant D_n(v, x) + D_n(x, w)$ (triangle inequality).

Table 3
The di-grams of strings $v = abba$, $w = babba$ over the alphabet $\Sigma = \{a,b\}^2$

	v	w	Σ^2
1	*ab*	*ba*	*aa*
2	*bb*	*ab*	*ab*
3	*ba*	*bb*	*ba*
4		*ba*	*bb*

Di-grams of v and w are listed in the order of their appearance in the strings and all occurrences of the di-grams are included in the list. Di-grams over Σ are listed in the lexicographical order. The di-gram profiles of v and w over the alphabet Σ are $G_2(v) = (0,1,1,1)$ and $G_2(w) = (0,1,2,1)$, when the di-grams are counted in lexicographical order.

It is not a metric since $D_n(v,w)$ can be 0 although $v \neq w$ (for example for di-grams of strings $v = abaa$ and $w = aaba$). Ukkonen (1992) lists also other properties of the n-gram distance which are here omitted.

Another approach to n-gram based string proximity is to calculate *similarity* between two strings instead of distance as in Eq. (1). Indeed, Keskustalo et al. (2003), Pirkola et al. (2002) and Toivonen et al. (2005) used Jaccard's formula based similarity function for s-grams. In order to formalize Jaccard's formula for s-gram similarity we first define it for ordinary n-grams—but based on n-gram profiles rather than n-gram sets.

In its basic form, Jaccard's formula for measuring the similarity between two sets A and B is written as

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}, \tag{2}$$

where $|A|$ denotes the size of the set A. If $A = B$, then $A \cap B = A \cup B$, and thus $J(A,B) = 1$. On the other hand if A and B do not share common elements, then $A \cap B = \emptyset$ and $J(A,B) = 0$.

Since in set theory the duplicate occurrences of any element in the set are discarded, the Jaccard's formula based n-gram similarity function for strings ignores the multiple occurrences of n-grams. Therefore, instead of n-gram profile of Definition 1 we need a *binary n-gram profile*:

Definition 3. Let $w \in \Sigma^*$ and let $x \in \Sigma^n$ be a n-gram. Let $B(w)[x]$ denote the occurrences of string x in w as follows:

$$B(w)[x] = \begin{cases} 1 & \text{if } G(w)[x] > 0, \\ 0 & \text{otherwise.} \end{cases}$$

The binary n-gram profile of w is the binary vector $B_n(w) = (B(w)[x])$, $x \in \Sigma^n$.

With the binary n-gram profile of w the Jaccard's formula based n-gram similarity function takes the following form.

Definition 4. Let v and w be non-empty strings from Σ^* and $n \in \mathbb{N}^+$. The Jaccard's n-gram similarity between v and w is

$$J_n(v,w) = \frac{\sum_{x \in \Sigma^n} B(v)[x]B(w)[x]}{\sum_{x \in \Sigma^n} (B(v)[x] + B(w)[x] - B(v)[x]B(w)[x])}. \tag{3}$$

Example 2. Let $\Sigma = \{a,b\}$ and $v = abba$, $w = babba \in \Sigma^*$. Now binary di-gram profiles of v and w, listed in lexicographical order of the di-grams, are $(0,1,1,1)$ and $(0,1,1,1)$. Thus their Jaccard's di-gram similarity $J_2(v,w) = 1$ and therefore the strings v and w are treated as equal (cf. Example 1).

As we mentioned in the previous section, it is common in IR applications to use padding spaces around the strings to get the symbols in the beginning and the end of the strings properly represented in the string comparison. It was also mentioned that it is common to use $n - 1$ padding spaces around the strings. Next we show how the padded n-gram comparison of the strings can be performed. For this purpose we assume a special padding symbol ♯. For example, in text retrieval applications ♯ could be thought as a regular space character.

Let Σ be a finite alphabet including the padding symbol $\sharp$, n an integer > 0 and $v, w \in (\Sigma \setminus \{\sharp\})^*$. Comparing strings v and w based on their n-gram overlap with, say, $n-1$ padding spaces in both the beginning and the end of the strings is performed as follows. Let p be a special string consisting only $n-1$ padding symbols, i.e., $p \in \{\sharp\}^{n-1}$ (clearly p is also in Σ^*). The padded comparison of the strings v and w is now trivial since operations $D_n(pvp, pwp)$ and $J_n(pvp, pwp)$ do the job. This follows from the fact that Σ^n contains (trivially) all n-grams of both v and w plus those n-grams which begin or end $n-1$ or less padding spaces (because $\sharp \in \Sigma$). If padding is used only in the beginning of the strings, the comparison is performed with $D_n(pv, pw)$ and $J_n(pv, pw)$. The situation where padding is used only in the end of the strings is handled correspondingly.

5. Formalizations of *s*-grams and their proximity functions

We shall now generalize the definition of n-grams by allowing skips between the symbols of the string w from which the n-gram is formed, i.e., the definition is generalized for s-grams. While Keskustalo et al. (2003) only consider s-grams of length of two, our definitions are for s-grams of any length. We also show how we get n-grams as special a case of our s-gram definition.

5.1. Definitions for s-grams

For simplicity we require that the skips in the s-grams are *systematical*, i.e., (1) each skip has equal length and (2) the skips are performed in each character position. With gram length of 2 and skip of 1 the $s_{2,1}$-grams of $w = abbababba$ are ab, ba, bb, aa, bb, ab, and ba.

Definition 5. Let $w = a_1 a_2 \ldots a_m \in \Sigma^*$, $n \in \mathbb{N}^+$ be a gram length, $k \in \mathbb{N}$ a skip length and let $x \in \Sigma^n$ be an n-gram. If $a_i a_{i+k+1} \ldots a_{i+(k+1)(n-1)} = x$ for some i, then w has a $s_{n,k}$-gram occurrence of x. Let $G_k(w)[x]$ denote the total number of $s_{n,k}$-gram occurrences of x in w. The $s_{n,k}$-gram profile of w is the vector $G_{n,k}(w) = (G_k(w)[x])$, $x \in \Sigma^n$.

Now, the $s_{n,k}$-gram based L_1 norm is defined as for n-grams:

Definition 6. Let v, w be strings from Σ^*, $n \in \mathbb{N}^+$ be a gram length and $k \in \mathbb{N}$ a skip length. The $s_{n,k}$-gram distance between v and w is

$$D_{n,k}(v, w) = \sum_{x \in \Sigma^n} |G_k(v)[x] - G_k(w)[x]|. \tag{4}$$

Example 3. Let $\Sigma = \{a, b\}$ be an alphabet and $v = aabab$, $w = babab$ strings from Σ^*. Their $s_{2,1}$-gram profiles, listed in lexicographical order of the $s_{2,1}$-grams, are $(1, 1, 0, 1)$ and $(1, 0, 0, 2)$. Thus, their $s_{2,1}$-gram distance $D_{2,1}(v, w) = 2$. Note that their di-gram distance defined in Eq. (1) would be 1. Table 4 lists the $s_{2,1}$-grams of strings v and w and all $s_{2,1}$-grams over alphabet $\Sigma = \{a, b\}$.

Note that Eq. (1) is a special case of Eq. (4), because $D_n(v, w) = D_{n,0}(v, w)$ for all strings $v, w \in \Sigma^*$. Furthermore, the following theorem shows that distance $D_{n,k}$ is a pseudo metric. The theorem is easy to prove and the proof is thus omitted.

Table 4
The $s_{2,1}$-grams of strings $v = aabab$ and $w = babab$ from the alphabet $\Sigma = \{a, b\}$

	v	w	Σ^2
1	*ab*	*bb*	*aa*
2	*aa*	*aa*	*ab*
3	*bb*	*bb*	*ba*
4			*bb*

$s_{2,1}$-grams of v and w are listed in the order of their appearance in the strings and all occurrences of the $s_{2,1}$-grams are included in the list. Di-grams over Σ are listed in the lexicographical order. The $s_{2,1}$-gram profiles of v and w over the alphabet Σ are $G_{2,1}(v) = (1, 1, 0, 1)$ and $G_{2,1}(w) = (1, 0, 0, 2)$, when the $s_{2,1}$-grams are counted in lexicographical order.

Theorem 1. *For all* $v, w, x \in \Sigma^*$,

(1) $D_{n,k}(v,w) \geqslant 0$ (*non-negativity*),
(2) $D_{n,k}(v,w) = D_{n,k}(w,v)$ (*symmetry*) *and*
(3) $D_{n,k}(v,w) \leqslant D_{n,k}(v,x) + D_{n,k}(x,w)$ (*triangle inequality*).

Thus also distance $D_{n,k}$ is pseudo metric. It is not metric, because $D_n(v,w)$ can be 0 although $v \neq w$ (for example for $s_{2,1}$-grams of strings $v = aaba$ and $w = aaab$).

Because, Keskustalo et al. (2003), Pirkola et al. (2002) and Toivonen et al. (2005) used Jaccard's similarity function for *s*-grams it is reasonable to formalize Jaccard's formula also for *s*-grams. This gives us also another point of view to *s*-gram based string proximity using the concept of similarity instead of distance.

The definition of Jaccard's similarity function for *s*-gram based string matching is analogous to that of *n*-grams. First, we need to define the *binary s-gram profile*:

Definition 7. Let $w \in \Sigma^*$, $n \in \mathbb{N}^+$ be a gram length and $k \in \mathbb{N}$ a skip length of the *s*-grams. Let $B_k(w)[x]$ denote the occurrences of string x in w as follows:

$$B_k(w)[x] = \begin{cases} 1 & \text{if } G_k(w)[x] > 0, \\ 0 & \text{otherwise.} \end{cases}$$

The binary $s_{n,k}$-gram profile of w is the binary vector $B_{n,k}(w) = (B_k(w)[x])$, $x \in \Sigma^n$.

With the binary $s_{n,k}$-gram profiles of strings v and w, the *s*-gram similarity function based on Jaccard's formula takes the following form.

Definition 8. Let v and w be non-empty strings from Σ^*, $n \in \mathbb{N}^+$ a gram length and $k \in \mathbb{N}$ a skip length. The Jaccard's *s*-gram similarity between v and w is

$$J_{n,k}(v,w) = \frac{\sum_{x \in \Sigma^n} B_k(v)[x] B_k(w)[x]}{\sum_{x \in \Sigma^n} (B_k(v)[x] + B_k(w)[x] - B_k(v)[x] B_k(w)[x])}. \tag{5}$$

Example 4. Let $\Sigma = \{a, b\}$ be an alphabet and $v = aabab$, $w = babab$ strings from Σ^*. Their binary $s_{2,1}$-gram profiles, listed in lexicographical order of the $s_{2,1}$-grams, are $(1,1,0,1)$ and $(1,0,0,1)$. Thus, their Jaccard's $s_{2,1}$-gram similarity $J_{2,1}(v,w) = 2/3$ (cf. Example 3). Note that their Jaccard's di-gram distance defined in Eq. (3) would also be 2/3.

Note that Eq. (3) is a special case of Eq. (5), because $J_n(v,w) = J_{n,0}(v,w)$ for all strings $v, w \in \Sigma^*$.

As with *n*-grams, it is also common with *s*-grams to use padding spaces around the strings to get the symbols in the beginning and the end properly represented in string comparison. The approach illustrated in the previous section for *n*-grams also works with *s*-grams, and therefore it is not repeated here. However, it should be noted that with *s*-grams it is common to use $(k+1)(n-1)$ padding spaces around the strings instead of $n-1$ (i.e, the length of the padding string p is $(k+1)(n-1)$).

5.2. Skip-gram classes and their proximity functions

Pirkola et al. (2002) and Keskustalo et al. (2003) found that the *s*-gram matching performance is improved when the *s*-gram sets of two strings are produced in several different ways and classified so that the similarity in each skip-gram class is computed separately. The evidence from the different skip-gram classes is then combined for the comparison of the strings. Therefore, we will now begin the formulation skip-gram classes and their proximity functions.

Definition 9. The skip-gram class of skip lengths, or shortly skip-gram class, is a set $C \in \mathscr{P}(\mathbb{N})$. Character Combination Index (CCI) is a set of skip-gram classes, i.e., a set $\mathscr{C} \in \mathscr{P}(\mathscr{P}(\mathbb{N}))$. If we want to refer to *s*-grams of length n in certain skip-gram class C we will simply write $s_{n,C}$-gram.

Definition 10. Let $w \in \Sigma^*$, $C \in \mathscr{P}(\mathbb{N})$ a skip-gram class and $x \in \Sigma^n$. Let $G_C(w)[x] = \sum_{k \in C} G_k(w)[x]$. The skip-gram class profile of w is the vector $G_C(w) = (G_C(w)[x])$, $x \in \Sigma^n$. In other words, $G_{n,C}(w) = \sum_{k \in C} G_{n,k}(w)$.

The next definition gives the L_1 norm for skip-gram classes:

Definition 11. Let v, w be strings in Σ^*, $n \in \mathbb{N}^+$ be a gram length and $C \in \mathscr{P}(\mathbb{N})$ a skip-gram class. The skip-gram class distance between v and w is

$$D_{n,C}(v,w) = \sum_{x \in \Sigma^n} |G_C(v)[x] - G_C(w)[x]|. \tag{6}$$

Example 5. Let $\Sigma = \{a,b\}$ be an alphabet and $v = aabab$, $w = babab$ strings from Σ^*. Their $s_{2,\{0,1\}}$-skip-gram class profiles, listed in lexicographical order of the $s_{2,0}$-grams and $s_{2,1}$-grams, are $(2,3,1,1)$ and $(1,2,2,2)$. Thus, their $s_{2,\{0,1\}}$-skip-gram class distance $D_{2,\{0,1\}}(v,w) = 4$.

Note that we can calculate Eq. (4) with Eq. (6), because $D_{n,k}(v,w) = D_{n,\{k\}}(v,w)$ for all strings $v, w \in \Sigma^*$. Especially n-gram distance of Eq. (1) between strings v and w is given by $D_{n,\{0\}}(v,w)$.

According to the following theorem also skip-gram class distance is a pseudo metric. Again, the proof is obvious and thus omitted.

Theorem 2. *For all* $v, w, x \in \Sigma^*$,

(1) $D_{n,C}(v,w) \geqslant 0$ (*non-negativity*),
(2) $D_{n,C}(v,w) = D_{n,C}(w,v)$ (*symmetry*) *and*
(3) $D_{n,C}(v,w) \leqslant D_{n,C}(v,x) + D_{n,C}(x,w)$ (*triangle inequality*).

Next we want to derive the Jaccard's similarity function for gram class based string matching analogous to previous Jaccard's formulas. Again, this requires us to define a *binary skip-gram class profile*.

Definition 12. Let $w \in \Sigma^*$, and $C \in \mathscr{P}(\mathbb{N})$ a skip-gram class and $x \in \Sigma^n$. Let

$$B_C(w)[x] = \begin{cases} 1 & \text{if } G_C(w)[x] \geqslant 0, \\ 0 & \text{otherwise.} \end{cases}$$

The binary skip-gram class profile of w is the binary vector $B_{n,C}(w) = (B_C(w)[x])$, $x \in \Sigma^n$.

The Jaccard's formula based s-gram similarity using binary gram-class profiles of strings can now be defined like Jaccard's similarities earlier.

Definition 13. Let v and w be non-empty strings from Σ^*, $n \in \mathbb{N}^+$ be a gram length and $C \in \mathscr{P}(\mathbb{N})$ a skip-gram class. The Jaccard's skip-gram class similarity between v and w is

$$J_{n,C}(v,w) = \frac{\sum_{x \in \Sigma^n} B_C(v)[x] B_C(w)[x]}{\sum_{x \in \Sigma^n} (B_C(v)[x] + B_C(w)[x] - B_C(v)[x] B_C(w)[x])}. \tag{7}$$

Example 6. Let $\Sigma = \{a,b\}$ be an alphabet and $v = aabab$, $w = babab$ strings from Σ^*. Their binary $s_{2,\{0,1\}}$-skip-gram class profiles, listed in lexicographical order of the $s_{2,0}$-grams and $s_{2,1}$-grams, are $(1,1,1,1)$ and $(1,1,1,1)$. Thus, their $s_{2,\{0,1\}}$-skip-gram class similarity is $J_{2,\{0,1\}}(v,w) = 1$. Therefore, according to the Jaccard's skip-gram class similarity, strings v and w are equal. This is a notable difference to the L_1 norm between v and w calculated in Example 5.

Note again that we can calculate Eq. (5) with Eq. (7), because $J_{n,k}(v,w) = J_{n,\{k\}}(v,w)$ for all strings $v, w \in \Sigma^*$. Especially n-gram similarity of Eq. (3) between strings v and w is given by $J_{n,\{0\}}(v,w)$.

Finally, we define a proximity functions for strings based on their total s-gram overlap in given set of skip-gram classes specified by a CCI.

5.3. CCI-based string proximity

In Pirkola et al. (2002) and Keskustalo et al. (2003), the final similarity function for *s*-grams was based on the Character Combination Index (CCI) and an extension of Jaccard's formula. The idea was to calculate the Jaccard similarity in each skip-gram class and then to combine the similaries in a novel way to an overall similarity function. In this section, we shall continue our approach from preceeding sections by, firstly, defining a *distance* function for two strings based on their *s*-gram profiles and the given CCI, and secondly, binarizing these profiles and defining (precisely) the Jaccard style similarity function for the same situation. As we shall see though some examples, the two functions may give quite different results. This is due to the distance function counting each occurrence of any *s*-gram and the similarity function only counting one, because it is set oriented. Only the latter has so far been experimentally tested in the case of *s*-grams.

Definition 14. Let v and w be strings in Σ^* and $\mathscr{C} \in \mathscr{P}(\mathscr{P}(\mathbb{N}))$ a character combination index. The distance $D_{n,\mathscr{C}}(v,w)$ of v and w with regard to $\mathscr{C}$ is defined as the average distance of their skip-gram class distances, i.e.,

$$D_{n,\mathscr{C}}(v,w) = \frac{1}{|\mathscr{C}|}\sum_{C\in\mathscr{C}} D_{n,C}(v,w). \tag{8}$$

Example 7. Let $\Sigma = \{a,b\}$ be an alphabet, $v = abbababba$, $w = baabaaba$ strings from Σ^* and $\mathscr{C} = \{\{0,1\},\{2\}\}$ a CCI. To calculate the distance $D_{2,\mathscr{C}}(v,w)$, we first need to calculate the skip-gram class distances $D_{2,\{0,1\}}(v,w)$ and $D_{2,\{2\}}(v,w)$. The skip-gram class profiles listed in lexicographical order of the *s*-grams are for string v $G_{2,\{0,1\}}(v) = (1,5,5,4)$ and $G_{2,\{2\}}(v) = (2,1,1,2)$, and for string w $G_{2,\{0,1\}}(w) = (4,4,5,0)$ and $G_{2,\{2\}}(w) = (3,0,0,2)$. Thus,

$$D_{2,\mathscr{C}}(v,w) = \frac{1}{|\mathscr{C}|}\sum_{C\in\mathscr{C}} D_{2,C}(v,w) = \frac{D_{2,\{0,1\}}(v,w) + D_{2,\{2\}}(v,w)}{|\{\{0,1\},\{2\}\}|} = \frac{8+3}{2} = 5.5.$$

The advantage of using average distance in the above definition is that the CCI distance remains as a pseudo metric. However, the CCI distance $D_{n,\mathscr{C}}(v,w)$ has never been empirically tested. Instead, the following definition gives a proximity function loosely on based Jaccard's formula that is used by Pirkola et al. (2002) and Keskustalo et al. (2003) for skip-gram class based string comparison.

Definition 15. Let v and w be non-empty strings over an alphabet Σ, $n \in \mathbb{N}$ and $\mathscr{C} \in \mathscr{P}(\mathscr{P}(\mathbb{N}))$ a CCI. The similarity $S_{n,\mathscr{C}}(v,w)$ of v and w with regard to CCI $\mathscr{C}$ is

$$S_{n,\mathscr{C}}(v,w) = \frac{\sum_{C\in\mathscr{C}}\sum_{x\in\Sigma^n} B_C(v)[x]B_C(w)[x]}{\sum_{C\in\mathscr{C}}\sum_{x\in\Sigma^n}(B_C(v)[x] + B_C(w)[x] - B_C(v)[x]B_C(w)[x])}. \tag{9}$$

Example 8. Let $\Sigma = \{a,b\}$ be an alphabet, $v = abbababba$, $w = baabaaba$ strings from Σ^* and $\mathscr{C} = \{\{0,1\},\{2\}\}$ a CCI. To calculate the similarity $S_{2,\mathscr{C}}(v,w)$, we need to calculate the binary skip-gram class profiles. The profiles, listed in lexicographical order of the *s*-grams, for string v are $B_{2,\{0,1\}}(v) = (1,1,1,1)$, $B_{2,\{2\}}(v) = (1,1,1,1)$, and for string w are $B_{2,\{0,1\}}(w) = (1,1,1,0)$, $B_{2,\{2\}}(w) = (1,0,0,1)$ Thus,

$$S_{2,\mathscr{C}}(v,w) = \frac{\sum_{C\in\mathscr{C}}\sum_{x\in\Sigma^2} B_C(v)[x]B_C(w)[x]}{\sum_{C\in\mathscr{C}}\sum_{x\in\Sigma^2}(B_C(v)[x] + B_C(w)[x] - B_C(v)[x]B_C(w)[x])} = \frac{(1+1+1+0)+(1+0+0+1)}{(1+1+1+1)+(1+1+1+1)} = \frac{5}{8} = 0.625.$$

We gave the similarity function of Eq. (9) by convention, following Pirkola et al. (2002) and Keskustalo et al. (2003). However, this choice for similarity function for CCI based string matching is not the most intuitive one. Therefore, we propose a new, simpler similarity function calculated as an average of the skip-gram classes of the CCI, in the spirit of Eq. (8).

Definition 16. Let v and w be non-empty strings from Σ^* and $\mathscr{C} \in \mathscr{P}(\mathscr{P}(\mathbb{N}))$ a character combination index. The similarity $S'_{n,\mathscr{C}}(v,w)$ of v and w with regard to CCI $\mathscr{C}$ is

$$S'_{n,\mathscr{C}}(v,w) = \frac{1}{|\mathscr{C}|}\sum_{C\in\mathscr{C}} J_{n,C}(v,w). \tag{10}$$

The similarities given by the Eqs. (9) and (10) are not far apart but nevertheless different as the reader may find through the simple example constructed with strings *aabba* and *bbab* and a CCI $\mathscr{C} = \{\{0\},\{1\}\}$.

We end this section by noting that following equalities hold between the CCI based string proximity functions and s-gram proximity functions. Let v and w be strings over an alphabet Σ, $n \in \mathbb{N}^+$ a gram length and $k \in \mathbb{N}$ a skip length. Then

(1) $D_{n,k}(v,w) = D_{n,\{\{k\}\}}(v,w)$ and
(2) $J_{n,k}(v,w) = S_{n,\{\{k\}\}}(v,w)$,

where in (2) $S_{n,\{\{k\}\}}(v,w)$ denotes either Eq. (9) or Eq. (10). Thus proximities between n- and s-grams can be evaluated by using Eqs. (8)–(10) only.

6. Discussion and conclusions

n-grams have been used widely and successfully for approximate string matching in many areas. Recently, Pirkola et al. (2002) devised a novel classified s-gram matching technique, where di-grams are formed of both adjacent and non-adjacent characters. s-grams have proved successful in approximate string matching across language boundaries—especially in matching proper names and technical terminology. While n-grams have been precisely defined, s-grams so far lack such precise definitions. Also their similarity comparison in (Keskustalo et al., 2003) lacks a stringent definition. In this paper, we have given precise definitions both for s-grams and their distance/similarity comparison.

Following established practices in the literature (Ukkonen, 1992) we first presented the n-gram profiles of strings and then their *distance* measure, the L_1 metric. This is a well established distance measure, and takes into account both the kinds of n-grams two strings contain, and their number. Based on this we also defined binary n-gram profiles and the Jaccard *similarity* measure, which has been popular in IR experiments. As pointed out, this similarity measure is weaker than the L_1 distance metric because Jaccard is insensitive to the counts of each n-gram in the strings to be compared.

Turning to s-grams, we provided novel definitions of s-gram profiles and extended the L_1 distance metric for them. We gave definitions for simple s-gram profiles and distances and progressively extended them to skip-gram classes and collections of skip-gram classes (as specified by the CCI). The s-gram profiles were also reduced in each case to binary profiles in order to precisely define the (extended) Jaccard similarity functions for s-grams in each case. Again, as pointed out, the extended L_1 distance metrics may be claimed stronger than the corresponding Jaccard similarity measures because the Jaccard measures are insensitive to the counts of each n-gram in the strings to be compared. Interestingly, the L_1 distance metrics have never been employed in IR experiments based on s-grams. Their greater strength in measuring string proximity suggests their potential for IR experiments, which are planned for in our forthcoming research.

Pirkola et al. (2002) only considered s-di-grams, which entails one skipping possibility between the two characters. Our definitions were for general s-grams with multiple skipping positions and constant skip length. Unfortunately, at the sub-word character string level, we do not know of text retrieval applications of $s_{n,k}$-grams for $n > 2$.

Finally, we showed that n-gram similarity/distance computations are special cases of our generalized definitions. In fact, all s-gram and n-gram distance computations may be carried out by Eq. (8) and their Jaccard similarities by Eq. (9) or Eq. (10).

The size of the n-gram index has been pointed out as a critical factor in their application (Califano & Rigoutsos, 1993; McNamee & Mayfield, 2004), where the determining factors are the size of the symbol set, and n-gram length, which determine the length of the profile vectors per object (i.e. word, text, biological sequence). In our case, a further complication arises from the skip-gram classes as each class requires its own profile. However, in the present problem area, focussing on s-grams for mono-lingual and cross-language word matching, the symbol set Σ typically has less than 30 symbols, the gram length typically is 2 and the CCI $\mathscr{C}$ contains 2–3 classes. With such values the s-gram profile size per keyword ($= |\mathscr{C}||\Sigma^n|$) remains under 2 KB per keyword which is very reasonable. Recall that the keyword index takes, for each keyword, the keyword length and the address list length, which is governed by address specificity, address length, and the database size (word count).

At a more general level, our strong belief is that popular proximity measures need precise definitions. This fosters an understanding of their relative strengths (e.g. the L_1 distance metric being stronger than the Jaccard measure) and their consistent implementation and application. Our definitions were developed in a bottom-up manner, only assuming character strings (based on some alphabet) and elementary mathematical concepts.

Acknowledgements

This study was funded in part by Department of Information Studies, University of Tampere, TISE (Tampere Graduate School for Information Science and Engineering), and Academy of Finland under the Grant # 1209960.

References

Altschul, S., Gish, W., Miller, W., Myers, E., & Lipman, D. (1990). A basic local alignment search tool. *Journal of Molecular Biology, 215*, 403–410.

Baròdal, J., Jörgensen, N., Larsen, G., & Martinussen, B. (1997). Nordiska: Våra språk förr och nu. Studentlitteratur.

Bishop, M., & Thompson, E. (1984). Fast computer search for similar DNA sequences. *Nucleic Acid Research, 12*(13), 5471–5474.

Califano, A., & Rigoutsos, I. (1993). FLASH: A fast look-up algorithm for string homology. In *Proceedings of the 1st international conference on intelligent systems for molecular biology* (pp. 56–64). AAAI Press.

Chen, A., He, J., Xu, L., Gey, F., & Meggs, J. (1997). Chinese text retrieval without using a dictionary. In: Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 42–49.

Damerau, F. J. (1964). A technique for computer detection and correction of spelling errors. *Communications of the ACM, 7*(3), 171–176.

Downie, S., & Nelson, M. (2000). Evaluation of a simple and effective music information retrieval method. In *SIGIR '00: Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval* (pp. 73–80). New York, NY USA: ACM Press.

Grossman, D. A., & Frieder, O. (2004). *Information retrieval: Algorithms and heuristics* (2nd ed.). Springer.

Hall, P. A. V., & Dowling, G. R. (1980). Approximate string matching. *ACM Computing Surveys, 12*(4), 381–402.

Hopcroft, J. E., Motwani, R., & Ullman, J. D. (2001). *Introduction to automata theory languages and computation* (2nd ed.). Addison Wesley.

Keskustalo, H., Pirkola, A., Visala, K., Leppänen, E., & Järvelin, K. (2003). Non-adjacent digrams improve matching of cross-lingual spelling variants. In M. A. Nascimento, E. de Moura, & A. Oliveira (Eds.), *Proceedings of the 10th international symposium on string processing and information retrieval SPIRE. Lecture Notes in Computer Science* (2857, pp. 252–265). Berlin: Springer.

McNamee, P., & Mayfield, J. (2003). JHU/APL experiments in tokenization and non-words translation. In: CLEF 2003 working notes. Available from http://clef.iei.pi.cnr.it/.

McNamee, P., & Mayfield, J. (2004). Character n-gram tokenization for European language text retrieval. *Information Retrieval, 7*, 73–97.

Miller, C., Gurd, J., & Brass, A. (1999). A rapid algorithm for sequence database comparisons: application to the identification of vector contamination in the EMBL databases. *Bioinformatics, 15*(2), 111–121.

Navarro, G. (2001). A guided tour to approximate string matching. *ACM Computing Surveys, 33*(1), 31–88.

O'Rourke, A., Robertson, A., & Willettt, P. (1997). Word variant identification in old french. *Information Research, 2*(4).

Peters, C. (2003). Introduction to the CLEF 2003 working notes. Available from http://clef.iei.pi.cnr.it/.

Pfeiffer, U., Poersch, T., & Fuhr, N. (1996). Retrieval effectiveness of proper name search methods. *Information Processing and Management, 32*(6), 667–679.

Pirkola, A. (1998). The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. In *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval* (pp. 55–63). New York, NY, USA: ACM Press.

Pirkola, A., Keskustalo, H., Leppänen, E., Känsälä, A. P., & Järvelin, K. (2002). Targeted *s*-gram matching: a novel *n*-gram maching technique for cross- and monolingual word form variants. *Information Research, 7*(2). Available from http://InformationR.net/ir/7-2/paper126.html.

Robertson, A. M., & Willett, P. (1998). Applications of *n*-grams in textual information systems. *Journal of Documentation, 54*(1), 48–69.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal, 27*(8), 379–423.

Toivonen, J., Pirkola, A., Keskustalo, H., Visala, K., & Järvelin, K. (2005). Translating cross-lingual spelling variants using transformation rules. *Information Processing and Management, 41*, 859–872.

Uitdenbogerd, A. L., & Zobel, J. (2002). Music ranking techniques evaluated. In *ACSC '02 Proceedings of the twenty-fifth Australasian conference on computer science* (pp. 275–283). Darlinghurst, Australia: Australian Computer Society, Inc.

Ukkonen, E. (1992). Approximate string-matching with *q*-grams and maximal matches. *Theoretical Computer Science, 92*(1), 191–211.

Ullman, J. R. (1977). A binary *n*-gram technique for automatic correction of substitution, deletion, insertion and reversal errors in words. *The Computer Journal, 20*(2), 141–147.

Zamora, E. M., Pollock, J. J., & Zamora, A. (1981). The use of trigram analysis for spelling error detection. *Information Processing and Management, 17*(8), 305–316.

Zobel, J., & Dart, P. (1996). Phonetic string matching: lessons from information retrieval. In *SIGIR '96: Proceedings of the 19th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 166–172). New York, NY, USA: ACM Press.

Comparison of *s*-gram Proximity Measures in Out-of-Vocabulary Word Translation

Anni Järvelin[1] and Antti Järvelin[2]

[1] University of Tampere, Department of Information Studies,
FIN-33014 University of Tampere, Finland
anni.jarvelin@uta.fi
[2] University of Tampere, Department of Computer Sciences,
FIN-33014 University of Tampere, Finland
antti.jarvelin@cs.uta.fi

Abstract. Classified *s*-grams have been successfully used in cross-language information retrieval (CLIR) as an approximate string matching technique for translating out-of-vocabulary (OOV) words. For example, *s*-grams have consistently outperformed other approximate string matching techniques, like edit distance or *n*-grams. The Jaccard coefficient has traditionally been used as an *s*-gram based string proximity measure. However, other proximity measures for *s*-gram matching have not been tested. In the current study the performance of seven proximity measures for classified *s*-grams in CLIR context was evaluated using eleven language pairs. The binary proximity measures performed generally better than their non-binary counterparts, but the difference depended mainly on the padding used with *s*-grams. When no padding was used, the binary and non-binary proximity measures were nearly equal, though the performance at large deteriorated.

1 Introduction

Cross-Language Information Retrieval (CLIR) refers to retrieval of documents written in a language other than that of the user's request. The document collection's language is called the *target language* and the query language the *source language*. A typical approach to CLIR is automatically translating the query into the target language. For an overview of CLIR, see [1]. Out-of-vocabulary (OOV) words constitute a major problem in query translation in CLIR. Due to the terminology missing from dictionaries, untranslatable keys appear in queries. Many typical OOV words, like proper names and technical terms, are often important query keys [2]. Therefore their translation is essential for query performance. In European languages, technical terms often share a common Greek or Latin root but are rendered with different spelling. This provides a good basis for the use of approximate string matching in translating the OOV words, as the words similar to a query's OOV words can be found from the target document collection and recognized as the translations of the query words.

The classified *s*-gram matching technique is a generalization of the well-known *n*-gram matching technique developed as a solution to the OOV word problem

A. Amir, A. Turpin, and A. Moffat (Eds.): SPIRE 2008, LNCS 5280, pp. 75–86, 2008.

in dictionary-based CLIR [3]. In s-gram matching the strings compared are decomposed into shorter substrings, called s-grams. Skipping characters is allowed when forming the s-grams and the degree of similarity between the strings is computed by comparing their s-gram sets. s-grams, or gapped q-grams, have also been described e.g. in [4] where they were applied for fast and efficient filtering for approximate string matching. The classified s-grams differ from the other gapped q-grams in that several different s-grams are grouped together into sets of s-grams prior to calculating the similarity. The classified s-grams have been developed with CLIR and natural language processing in mind, i.e., for relatively short strings including relatively little repetition of s-grams. In CLIR applications, the technique has outperformed several other established approximate string matching techniques, such as the edit distance, the longest common subsequence and n-grams [3,5].

There are several ways of calculating the s-gram proximity between two strings. In the context of n-gram matching the L_1 distance [6], its binary version Hamming distance [7], the Dice coefficient [8], and the Jaccard coefficient [9] among others have been used. Robertson and Willett [10] mention that any proximity measure could be used, while Zobel and Dart [7] propose that measures used in IR, such as the cosine measure [8], should not be appropriate for phonetic n-gram matching as they factor out the document length.

Only similarity measures based on the Jaccard coefficient have previously been tested with classified s-grams [3,5]. Clearly, other proximity measures could also be applied, but it is not obvious which might be the best suited ones. Järvelin et al. [11] formalized a few proximity measures for s-gram matching, e.g., the L_1 distance. They argued that, theoretically, the Jaccard coefficient may not be the choice proximity measure to be used in the s-gram matching, as it is binary and thus insensitive to the counts of each s-gram in the strings to be compared. The non-binary L_1 distance should be a more sensitive proximity measure, as it takes both the types of s-grams and their number in the strings compared into account. Järvelin et al. [11] did not test their claim empirically, but their definitions enable the comparison of different s-gram proximity measures.

As the choice of the proximity measure used with the s-grams may affect the performance of the technique, testing the different proximity measures is needed. This article contributes to the issue by reporting the results of an evaluation of several proximity measures for s-gram matching of cross-lingual spelling variants. Especially the differences between the binary and the non-binary proximity measures are considered. The binary proximity measures Jaccard coefficient, binary cosine similarity and Hamming distance were compared to their non-binary counterparts Tanimoto coefficient, cosine similarity and L_1 distance respectively. Also, the binary Dice coefficient was tested. Cross-lingual spelling variants in seven languages (English, Finnish, French, German, Italian, Spanish and Swedish) were used as source words that were translated into four target languages, English, German, Swedish and Finnish, using classified s-gram matching. In total eleven language pairs were used. The proximity measures' performance was evaluated as average translation precision.

Next, Section 2 provides an introduction to the s-grams and their proximity measures. Section 3 presents the materials and methods and Section 4 the results. Finally, section 5 contains a brief discussion and the conclusions.

2 s-gram Definitions

2.1 Introduction to s-grams

Word variation, where a language pair shares words written differently but having the same origin, is called cross-lingual spelling variation. Pirkola et al. [3] and Keskustalo et al. [5] showed that this kind of variation can advantageously be modeled with the s-grams. In s-gram matching the text strings to be compared are decomposed into substrings and the similarity between the strings is calculated as the overlap of their common substrings. Unlike in n-gram matching, skipping some characters is allowed when forming the s-grams. In CLIR applications substring length of two has been used. It has been found beneficial in IR applications to use padding spaces around the strings when forming s-grams [5,10]. This helps to get the characters at the beginning and at the end of a string properly presented in string comparison.

In *classified* s-gram matching technique [3] the s-grams originating from the same string are classified into sets based on the number of characters skipped prior to calculating the similarity. Only the s-grams belonging to the same set are compared to each other. *Gram class* indicates the skip length(s) used when generating a set of s-grams. The largest value in a gram class is called the *spanning length* of the gram class [5], e.g., for gram class $\{0, 1\}$, the spanning length is one. Two or more gram classes may also be combined into more general gram classes. The *character combination index (CCI)* then indicates the set of all the gram classes to be formed from a string, e.g. CCI $\{\{0\}, \{1, 2\}\}$ means that two gram classes are formed from a string: one with conventional n-grams formed of adjacent characters ($\{0\}$) and one with s-grams formed both by skipping one and two characters ($\{1, 2\}$). For the string "abracadabra", the s-gram set produced by the CCI $\{\{1, 2\}\}$ is $\{ar, ba, rc, aa, cd, db, bc, ra, ad, ca, ab, dr\}$, when duplicate s-grams are not listed.

2.2 s-gram Profiles and Their Proximities

s-gram-based string proximity measures are based on strings' *s-gram profiles*. The s-gram profile definitions given in this paper are extended from Ukkonen's [6] n-gram profile definition. Next strings' s-gram profiles are defined, which are then generalized for gram classes. Then various gram class based proximity measures are given, because the strings' CCI based proximity measures are calculated as the average gram class distance of the CCI's gram classes.

Definition 1. *Let $w = a_1 a_2 \dots a_m$ be a string over a finite alphabet Σ, $n \in \mathbb{N}^+$ be a gram length, $k \in \mathbb{N}$ a skip length and let $x \in \Sigma^n$ be an s-gram. If $a_i a_{i+k+1} \dots a_{i+(k+1)(n-1)} = x$ for some i, then w has a $s_{n,k}$-gram occurrence of*

x. Let $G_k(w)[x]$ denote the total number of $s_{n,k}$-gram occurrences of x in w. The $s_{n,k}$-gram profile of w is the vector $G_{n,k}(w) = (G_k(w)[x]), x \in \Sigma^n$.

s-gram profile can easily be generalized for gram classes. The gram class profiles are formed by summing up the s-gram profiles in a given gram class.

Definition 2. *Let $w \in \Sigma^*$, $C \in \mathcal{P}(\mathbb{N})$ a gram class and $x \in \Sigma^n$. Let $G_C(w)[x] = \sum_{k \in C} G_k(w)[x]$. The gram class profile of w is the vector $G_{n,C}(w) = (G_C(w)[x])$, $x \in \Sigma^n$. In other words, $G_{n,C}(w) = \sum_{k \in C} G_{n,k}(w)$.*

Sometimes the exact number of the occurrences of s-grams in the string is irrelevant, but merely the information if a specific s-gram occurs at all in the string is needed. This leads to the notion of binary gram class profile.

Definition 3. *Let $w \in \Sigma^*$, and $C \in \mathcal{P}(\mathbb{N})$ a gram class and $x \in \Sigma^n$. Let*

$$B_C(w)[x] = \begin{cases} 1 \text{ if } G_C(w)[x] > 0 \\ 0 \text{ otherwise} \end{cases}.$$

The binary gram class profile of w is the vector $B_{n,C}(w) = (B_C(w)[x]), x \in \Sigma^n$.

Various proximity measures can now be used to calculate string proximities based on the general and binary gram class profiles. Next, only the proximity measures using the general gram class profile of Definition 2 are given, because the corresponding proximity measures using binary profiles are defined by substituting the general s-gram profiles with binary ones in the following equations.

Let v and w be strings in Σ^*, $n \in \mathbb{N}^+$ be a gram length and $C \in \mathcal{P}(\mathbb{N})$ a gram class. L_1 distance for gram classes of strings v and w is

$$L1_{n,C}(v, w) = \sum_{x \in \Sigma^n} |G_C(v)[x] - G_C(w)[x]|. \tag{1}$$

The L_1 distance has been used with n-grams by Ukkonen [6] and its binary version, the Hamming distance, was proposed by Zobel and Dart [7]. Therefore its performance was investigated in s-gram based OOV word translation.

The cosine gram class similarity between v and w is defined as

$$Cos_{n,C}(v, w) = \frac{G_C(v)^T G_C(w)}{\|G_C(v)\| \|G_C(w)\|}, \tag{2}$$

where $\|\cdot\|$ denotes the Euclidean norm and T the transpose of a vector. Cosine similarity (or normalized dot product) is a widely utilized proximity measure in text retrieval applications [12] and therefore its performance in s-gram matching was also investigated along its binarized counterpart.

The Tanimoto coefficient [13] between gram classes of v and w is given by

$$T_{n,C}(v, w) = \frac{G_C(v)^T G_C(w)}{\|G_C(v)\|^2 - G_C(v)^T G_C(w) + \|G_C(w)\|^2}. \tag{3}$$

The Tanimoto coefficient was tested, because its binary counter part, the Jaccard coefficient, has traditionally been used in s-gram matching [3,5,11].

Turning to the binary profile based proximity measures, the Hamming distance $H_{n,C}(v,w)$ between v and w is derived by substituting the general gram class profile with binary profile in (1), binary cosine similarity $BinCos_{n,C}(v,w)$ by substituting with binary profiles in (2), and Jaccard coefficient $J_{n,C}(v,w)$ by doing the same substitution in (3).

Lastly, the Dice's coefficient was investigated, because it has been used in n-gram matching [10]. It is closely related to the Jaccard coefficient, but weights more the matching profile positions between the gram class profiles than the mismatching ones [12]. The Dice coefficient between v and w is given by

$$D_{n,C}(v,w) = \frac{2B_C(v)^T B_C(w)}{\|B_C(v)\|^2 + B_C(v)^T B_C(w) + \|B_C(w)\|^2}. \tag{4}$$

The character combination index based string proximity measures tested in this paper are defined as the average of strings' gram class proximities. For example, for a CCI $\mathscr{C} \in \mathcal{P}(\mathcal{P}(\mathbb{N}))$, and a gram length n, the CCI-distance corresponding to L_1 distance is

$$L1_{n,\mathscr{C}}(v,w) = \frac{1}{|\mathscr{C}|} \sum_{C \in \mathscr{C}} L1_{n,C}(v,w). \tag{5}$$

All CCI-based proximity measures tested below were defined analogously.

One problem that might arise when using the s-gram profiles in approximate string matching is the length of the profiles. With $s_{n,k}$-grams, the profile length is $|\Sigma|^n$ where Σ is the specified alphabet. For example, the standard English alphabet consists of 26 letters, and thus even the di-gram profiles are quite long. However, with natural languages, the s-gram and the gram class profiles are typically very sparse, and well suited for sparse vector implementations. Therefore, the proximities between the s-gram profiles can be evaluated efficiently.

3 Materials and Methods

3.1 Materials

The test data consisted of three parts: the search keys, the target words and the set of correct translations, i.e., the relevance judgments. 271 search keys were expressed in seven languages, which were English, Finnish, French, German, Italian, Spanish and Swedish. The search keys were mostly technical terms from the domains of biology, medicine, economics and technology, but also a list of geographical names obtained from [5] was included. These are typical cases of cross-lingual spelling variants that tend to be OOV words and thus problematic in CLIR. In total, 11 language pairs were used in the study, with four target languages: English, German, Finnish and Swedish. English was combined with all of the other languages as a target language and was also used as a source

language with Finnish, German and Swedish. Translation was also done both ways between Swedish and German.

Target word lists (TWLs) consisted of CLEF 2003 [14] document collection's indices for the target languages. The collections are full-text newspaper document collections from 1994–1995. The size of the collections, and thus the TWLs, varies between languages. The English TWL consisted of ca 257,000, the Swedish TWL of ca 388,000 and the Finnish TWL of ca 535,000 unique word forms. The German CLEF03 collection was considerably bigger and thus only a part of it was used for creating a TWL including ca 391,000 unique word forms.

All the TWLs were lemmatized (i.e. the index words were returned into their basic forms) with the TWOL morphological analyzer by Lingsoft Ltd. The words not recognized by the morphological analyzer were indexed in the word forms they appeared in the text. Compounds were split and both the original compounds and their constituents were indexed. The missing translation equivalents of the search keys were added to the TWLs, and there was only one correct translation for each search key in the TWLs.

3.2 Methods

The performance of the proximity measures was tested as follows. The *s*-gram length was set to two, because earlier research [3,5] suggests it to be the most appropriate gram length for CLIR. Padding was used at both ends of the strings and the length of the padding string was $(n-1)(k+1)$, where n is the gram and k the skip length. Also, *s*-grams with no padding and padding only at the beginning of strings were tested for two language pairs (English-German and German-English) with CCI $\{\{0\},\{1,2\}\}$ to see how the padding affects the results. For each search key 100 best translations were produced, with exception of ties in the last place when all translations within the cohort of equal proximity values were included into the result set. Translations found later were not taken into consideration, as taking more than 2-4 *s*-gram translation candidates into a query deteriorates its performance [15].

This study concentrates on comparing the proximity measures. Exhaustive testing of all possible CCIs, proximity measures and language pairs was not sensible or even possible within this study. If skip lengths 0 – 4 were considered, there would be $2^5 - 1 = 31$ possible gram classes, and thus about $2^{31} - 1$, about two billion, combinations as possible CCIs. To be able to restrict the scope of the study to some evidently useful CCIs, statistics on typical cross-lingual spelling variation between French and English and German and English were used. Pirkola et al. [16] generated statistical transformation rules that model typical character changes and correspondences between several language pairs. The rules were generated from over 10,000 term pairs of medical words. They model the same cross-lingual spelling variation phenomenon as the *s*-grams, but are based on an independent method and character correspondence statistics from an independent large dataset. We mapped a subset of ca 200 most frequent transformation rules to the corresponding gram classes for both language pairs and calculated the frequency of each gram class in the data.

Table 1. The number of transformation rules corresponding to each gram class for French to English and German to English cross-lingual spelling variants

Gram class	$\{1\}$	$\{0,1\}$	$\{1,2\}$	$\{0,2\}$	$\{2\}$	$\{1,3\}$	$\{0,3\}$	$\{2,3\}$	$\{3\}$
Fr-En	56	65	44	11	12	7	3	5	3
Ge-En	117	37	36	20	7	3	4	1	0
Total	173	102	80	31	19	10	7	6	3

Table 2. The twelve CCIs used for the comparison of the proximity measures. Note that CCI_0 corresponds to the traditional n-grams. For CCI_0 gram length of two and three was experimented, for the remaining CCIs only gram length of two was used.

CCI_0	$\{\{0\}\}$	CCI_4	$\{\{0\},\{1\}\}$	CCI_8	$\{\{0,1\}\}$
CCI_1	$\{\{0\},\{0,1\}\}$	CCI_5	$\{\{0\},\{0,1\},\{1\}\}$	CCI_9	$\{\{0,1,2\}\}$
CCI_2	$\{\{0\},\{0,1\},\{1\},\{1,2\}\}$	CCI_6	$\{\{0\},\{1\},\{1,2\}\}$	CCI_{10}	$\{\{1\}\}$
CCI_3	$\{\{0\},\{0,1\},\{1,2\}\}$	CCI_7	$\{\{0\},\{1,2\}\}$	CCI_{11}	$\{\{1,2\}\}$

There were some differences between the language pairs, but the transformation rules that model character changes corresponding to the gram classes $\{1\}$, $\{0,1\}$ and $\{1,2\}$ were clearly the most common ones in the data. Table 1 summarizes the results for both languages. Based on this it seemed reasonable to use only gram classes with spanning length of two or less when matching cross-lingual spelling variants. Changes corresponding to the remaining clearly less frequent gram classes were thus discarded. Keskustalo et al. [5] reached an equal conclusion, when deciding which gram classes they should use.

Based on the results of Table 1, the gram classes $\{1\}$, $\{0,1\}$, and $\{1,2\}$ and gram class $\{0\}$ corresponding to the n-grams were selected as the base gram classes for the tested CCIs. In total, the twelve CCIs of Table 2 were used in the tests. For CCI_0, in addition to the gram length of two (di-grams), also gram length of three (tri-grams) was used. This set of CCI_0 - CCI_{11} is a representative set on effective s-grams, and by using this set a reliable picture of various s-gram proximity measures in s-gram matching can be obtained.

To compare the performances of the proximity measures, the average precision (AP, or reciprocal rank - as there is only one correct translation, these are the same) was calculated for each proximity measure at three different levels: among top 2, top 5 and top 100 highest ranked translation candidates. If the correct translation was in a cohort of words sharing the same proximity value with the target word, the average rank of the cohort was used. The top 2 and top 5 levels were the most interesting ones, as more translation candidates would deteriorate the query performance. The Friedman test [17] was used to test the statistical significance of the differences between the proximity measures. Below, statistically significant difference corresponds to α-level $\alpha = 0.01$, statistically highly significant difference to α-level $\alpha = 0.001$, and statistically almost significant α-level $\alpha = 0.05$.

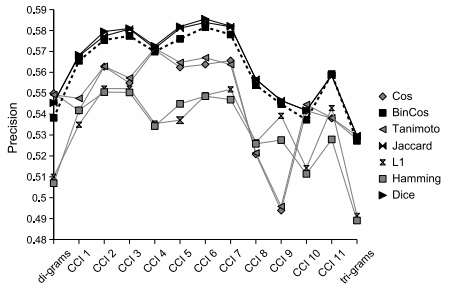

Fig. 1. The medians of APs of the proximity measures at top 5 translation candidates for all CCIs over all language pairs, zoomed in for clarity

4 Results

4.1 CCIs and Proximity Measures over All Languages

The results for all proximity measures and CCIs over all language pairs are presented in Fig. 1 and in Table 3 as the medians of AP when the top 5 translation candidates are considered. The results in top 2 and top 100 followed the same trends and are not presented due to the lack of space. The results divide the *s*-grams into two groups: the *s*-grams with CCIs that combine several *s*-gram types into a gram class and the *s*-grams where only one *s*-gram type is present in each gram class. In the former group (CCIs 1, 2, 3, 5, 6, 7, 8, 9, 11), the binary proximity measures performed clearly better than their non-binary counterparts, i.e., Jaccard performed better than Tanimoto, binary cosine better than cosine and Hamming better than L_1. The differences between Jaccard and Tanimoto and binary cosine and cosine were statistically significant for 7 of these 9 CCIs for 8 language pairs out of 11. For CCI_5 the differences were statistically significant only for five language pairs (EN-FI, EN-GE, FI-EN, FR-EN, IT-EN) of which two (EN-GE, FR-EN) were only almost significant. For CCI_6 the differences were statistically significant for seven language pairs (EN-FI, EN-GE, FR-EN, GE-EN, IT-EN, SP-EN, SW-EN), two of these (IT-EN, SW-EN) being almost significant. For the two closest related language pairs (GE-SW and SW-GE) the differences were not statistically significant. Also, for EN-SW the differences were statistically significant only for CCIs 1, 2, 8, and 9. The differences between Hamming and L_1 were typically not statistically significant. The three best measures Dice, Jaccard and binary cosine performed similarly and clearly better than the rest of the proximity measures. L_1 and Hamming were the worst proximity measures. The performance difference between them and the other proximity measures was statistically significant for all language pairs and CCIs.

In the latter group, including the adjacent di-grams and tri-grams (CCI_0) and the *s*-grams with CCI_4 and CCI_{10}, the difference between the binary and non-binary proximity measures was smaller and always to the advantage of the non-binary measures. These differences were nevertheless never statistically significant. Tanimoto was the best proximity measure, while L_1 and Hamming were

Table 3. The medians of the APs of the proximity measures among top 5 translation candidates for all CCIs over all language pairs. The best proximity measures for each CCI are in bold. Tanimoto coefficient performs best for n-grams and s-grams with CCI_{10} and Dice coefficient performs best for the s-grams with CCI's that combine several s-gram types into more general gram classes.

	Proximity measure						
CCI	Cos	BinCos	Tanimoto	Jaccard	L_1	Hamming	Dice
di-grams	0.5490	0.5382	**0.5493**	0.5454	0.5100	0.5070	0.5454
CCI_1	0.5417	0.5655	0.5475	0.5678	0.5349	0.5418	**0.5683**
CCI_2	0.5627	0.5755	0.5629	0.5774	0.5522	0.5506	**0.5795**
CCI_3	0.5549	0.5774	0.5573	0.5807	0.5522	0.5504	**0.5810**
CCI_4	0.5708	0.5699	0.5715	0.5715	0.5355	0.5343	**0.5726**
CCI_5	0.5624	0.5760	0.5647	0.5811	0.5371	0.5449	**0.5819**
CCI_6	0.5638	0.5816	0.5670	0.5839	0.5490	0.5486	**0.5855**
CCI_7	0.5656	0.5781	0.5637	0.5818	0.5518	0.5469	**0.5821**
CCI_8	0.5208	0.5539	0.5214	**0.5567**	0.5265	0.5258	**0.5567**
CCI_9	0.4939	0.5448	0.4958	**0.5465**	0.5392	0.5276	**0.5465**
CCI_{10}	0.5417	0.5373	**0.5446**	0.5418	0.5142	0.5114	0.5418
CCI_{11}	0.5380	**0.5592**	0.5382	0.5585	0.5429	0.5279	0.5585
tri-grams	0.5280	0.5272	**0.5296**	**0.5296**	0.4913	0.4891	**0.5296**
MEDIAN	0.5490	0.5655	0.5493	0.5678	0.5371	0.5343	**0.5683**

the worst ones the difference being always statistically significant. n-grams performed clearly worse than the s-grams with CCIs combining s-gram types into more general gram classes. The s-grams with CCI_4, combining two gram classes of a single s-gram type, performed better. This suggests that the s-gram technique benefits from combining gram classes into one CCI. It also seems that the more s-gram types were combined into a gram class, the more the performance of Tanimoto and cosine suffered. The CCI_9 is an example of this, showing a notable fall in the performance of Tanimoto and cosine in Fig. 1.

4.2 Results for Each Language Pair

The results in Fig. 1 and in Table 3 are medians over all the language pairs tested. To give a better picture of the results for the different language pairs, a typical CCI was selected to represent each group. CCI_6 represents the s-grams that combine several s-gram types in the gram classes. The results are presented for all language pairs in Fig. 2 (a) as AP among the top 5 translation candidates. The binary proximity measures performed better than their non-binary counterparts. Differences between Jaccard and Tanimoto and binary cosine and cosine were statistically significant for 7 language pairs, as mentioned above (not for FI-EN, SW-GE, GE-SW, EN-SW). Dice, Jaccard and binary cosine were the best proximity measures, while L_1 and Hamming were the worst ones. The differences between these were consistently statistically significant. Fig. 2 (b) presents the results for CCI_0 di-grams representing the other class of s-grams as AP among

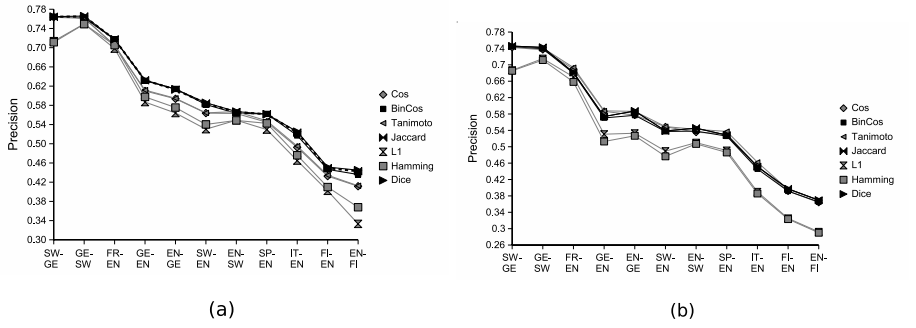

Fig. 2. (a) The AP of the proximity measures at top 5 for all language pairs for the s-grams with CCI_6. (b) The AP of the proximity measures at top 5 for all language pairs for traditional di-grams (CCI_0). Both figures are zoomed in for clarity.

the top 5 translation candidates. The results were scattered depending on the language pair, though still in line with the median results presented in Table 3 and Fig. 1. The non-binary proximity measures (Tanimoto and cosine) performed on average better than their binary counterparts, but the differences were not statistically significant. Hamming distance and L_1 were the worst measures, with statistically significant difference to the other proximity measures. Tri-grams performed generally worse than di-grams.

4.3 Padding

The differences between the binary and non-binary proximity measures were clearly reduced when no padding or padding only at the beginning of the strings were used. When no padding at all was used, the results deteriorated for all proximity measures and more for the binary than the non-binary proximity measures. For cosine and Tanimoto, the results even improved slightly for one of the two language pairs (GE-EN). Thus the differences between corresponding binary and non-binary proximity measures were reduced and were not statistically significant. When only the left-side padding was used, the overall effect on results was a little unclear: for English to German matching the best results deteriorated slightly, but for German to English the top results improved slightly. The non-binary proximity measures improved in comparison to their binary counterparts and the differences between them were not statistically significant. L_1 and Hamming suffered both from not using padding and also from using padding only at the beginning of the strings. They were always clearly the worst proximity measures with a statistically highly significant difference between them and the other proximity measures.

5 Discussion

To sum up the results, the binary proximity measures performed better than their non-binary counterparts in s-gram based matching of OOV words. Dice,

Jaccard and binary cosine performed best and any of these measures could be beneficially used. The difference between the binary and non-binary proximity measures seems to depend on the CCI used: when a number of different s-gram types were combined into a more general gram class (such as $\{\{1, 2\}\}$), the binary proximity measures clearly outperformed their non-binary counterparts. For the CCIs where only one s-gram type was present in each gram class (the traditional n-grams, CCI_4, and CCI_{10}), the differences between the binary and non-binary proximity measures vanished. Also, the more s-gram types were combined into a gram class, the more the performance of Tanimoto and cosine suffered.

This seems to be linked to the padding used with s-grams: When several s-gram types are combined into one gram class and padding was used, identical s-grams from both ends of strings are formed repeatedly and become overweighted when using non-binary proximity measures. As character changes are especially common at the ends of cross-lingual spelling variants (e.g. *antiseptic* - *antiseptique*), this damages the performance of the non-binary proximity measures. Removing the padding is nevertheless not a guarantee of success as it may affect the overall performance of the s-gram matching negatively. Keskustalo et al. [5] have found earlier that whether the padding on both sides of strings or only at the beginning performs best depends on the language pair at hand. For s-gram matching implementations using non-binary s-gram profiles, the repetitive occurrences of s-grams including padding characters should be ignored.

L_1 and its binary counterpart Hamming distance did not perform well and they do not seem suitable proximity measures for this application area. With these proximity measures the distance between two strings is calculated as the mean value of the different s-grams in the gram classes. This causes the measures to favor short words as no s-grams can be formed of one letter words (without padding) and none or very few of two or three letter words. Therefore, L_1 and Hamming give more non-relevant short words at the top ranks in the result lists than the other proximity measures. This is also reflected in the fact that the results for L_1 and Hamming deteriorated when the padding was removed.

Non-binary proximity measures are suitable for applications where a lot of repetition of s-grams occur (e.g. gene matching). In cross-lingual OOV word matching the alphabet used is rather large and the strings processed quite short. Consequently the repetition of s-grams is not extensive and therefore the binary and non-binary s-gram profiles approach each other. Therefore, no advantage is achieved with the use of the non-binary proximity measures.

Acknowledgments

The authors wish to thank Academy Professor Kalervo Järvelin, Ph.D., Docent Ari Pirkola, Ph.D., and Mr. Heikki Keskustalo, M.Sc. from University of Tampere for their support and comments on the paper. The first author was funded by Tampere Graduate School in Information Science and Engineering (TISE) and Academy of Finland under grant # 1209960.

References

1. Kishida, K.: Technical issues of cross-language infromation retrieval: a review. Inf. Process. Manage 41(3), 433–455 (2005)
2. Pirkola, A., Järvelin, K.: Employing the resolution power of search keys. JASIST 52(7), 575–583 (2001)
3. Pirkola, A., Keskustalo, H., Leppänen, E., Känsälä, A.P., Järvelin, K.: Targeted s-gram matching: a novel n-gram maching technique for cross- and monolingual word form variants. Information Research 7(2) (2002), http://InformationR.net/ir/7-2/paper126.html
4. Burkhardt, S., Kärkkäinen, J.: Better filtering with gapped q-grams. Fundamenta Informaticae 56(1–2), 51–70 (2003)
5. Keskustalo, H., Pirkola, A., Visala, K., Leppänen, E., Järvelin, K.: Non-adjacent digrams improve matching of cross-lingual spelling variants. In: Nascimento, M.A., de Moura, E.S., Oliveira, A.L. (eds.) SPIRE 2003. LNCS, vol. 2857, pp. 252–265. Springer, Heidelberg (2003)
6. Ukkonen, E.: Approximate string-matching with q-grams and maximal matches. Theor. Comput. Sci. 92(1), 191–211 (1992)
7. Zobel, J., Dart, P.: Phonetic string matching: lessons from information retrieval. In: SIGIR 1996: Proceedings of the 19th ACM SIGIR Conference, pp. 166–172. ACM Press, New York (1996)
8. Salton, G., McGill, M.J.: Introduction to Modern Information Retrieval, 1st edn. McGraw-Hill, New York (1983)
9. Pfeiffer, U., Poersch, T., Fuhr, N.: Retrieval effectiveness of proper name search methods. Inf. Process. Manage 32(6), 667–679 (1996)
10. Robertson, A.M., Willett, P.: Applications of n-grams in textual information systems. J. Doc. 54(1), 48–69 (1998)
11. Järvelin, A., Järvelin, A., Järvelin, K.: s-grams: defining generalized n-grams for information retrieval. Inf. Process. Manage. 43(4), 1005–1019 (2007)
12. Hand, D.J., Mannila, H., Smyth, P.: Principles of Data Mining, 1st edn. MIT Press, Cambridge (2001)
13. Theodoridis, S., Koutroumbas, K.: Pattern Recognition, 2nd edn. Academic Press, London (2003)
14. Gonzalo, J., Peters, C.: Introduction. In: Peters, C., Gonzalo, J., Braschler, M., Kluck, M. (eds.) CLEF 2003. LNCS, vol. 3237, pp. 1–6. Springer, Heidelberg (2004), http://clef.iei.pi.cnr.it/
15. Hedlund, T., Airio, E., Keskustalo, H., Lehtokangas, R., Pirkola, A., Järvelin, K.: Dictionary-based cross-language information retrieval: Learning experiences from CLEF 2000 – 2002. Information Retrieval - Special Issue on CLEF Cross-Language IR 7(1–2), 99–119 (2004)
16. Pirkola, A., Toivonen, J., Keskustalo, H., Visala, K., Järvelin, K.: Fuzzy translation of cross-lingual spelling variants. In: SIGIR 2003: Proceedings of the 26th ACM SIGIR Conference, pp. 345–352. ACM Press, New York (2003)
17. Conover, W.J.: Practical Nonparametric Statistics, 3rd edn. Wiley, New York (1999)

Dictionary-independent translation in CLIR between closely related languages

Anni Järvelin
+46-480-411662
anni.jarvelin@uta.fi

Sanna Kumpulainen
+358-505298901
sanna.kumpulainen@uta.fi

Ari Pirkola
+358-14-762278
pirkola@cc.jyu.fi

Eero Sormunen
+358-3-35516972
eero.sormunen@uta.fi

Department of Information Studies, FIN 33014, University of Tampere, Finland

ABSTRACT
This paper presents results from a study, where fuzzy string matching techniques were used as the sole query translation technique in Cross Language Information Retrieval (CLIR) between the closely related languages Swedish and Norwegian. It is a novel research idea to apply only fuzzy string matching techniques in query translation. Closely related languages share a number of words that are cross-lingual spelling variants of each other. These spelling variants can be translated by means of fuzzy matching. When cross-lingual spelling variants form a high enough share of the vocabulary of related languages, the fuzzy matching techniques can perform well enough to replace the conventional dictionary-based query translation. Different fuzzy matching techniques were tested in CLIR between Norwegian and Swedish and it was found that queries translated using skipgram matching and a combined technique of transformation rule based translation (TRT) and n-grams perform well. For the best fuzzy matching query types performance difference with respect to dictionary translation queries was not statistically significant.

Categories and Subject Descriptors
H.3.3 [**Information Systems**]: Information Search and Retrieval

General Terms
Performance, Experimentation

Keywords
Cross-language retrieval, Fuzzy matching

1. INTRODUCTION
Information retrieval methods are based on comparing the words in requests with the words in documents. Cross Language Information Retrieval (CLIR) refers to the retrieval of documents in other languages than the language of the request. For an overview of different approaches to CLIR, see [6]. Fuzzy string matching methods are used for finding matches between words that are similar but not identical. In CLIR fuzzy string matching has been used for handling proper names and technical terms, as well as other cross-lingual spelling variants not found in translation dictionaries [5, 12, 17]. McNamee and Mayfield [7] have used n-grams in corpus-based query translation.

Closely related languages have not been considered as a separate line of research in CLIR. The dominating approach, dictionary-based translation of queries, is a fairly effective technique, but has its problems in the limited coverage of dictionaries and the constant need for updating, which can make it an expensive technique. Closely related languages typically share a high number of spelling variants, i.e., equivalent words that share the same origin and are similar (but not identical). If the number of the shared cross-lingual variants is high enough, query translation can be handled by much cheaper and simpler fuzzy techniques.

Among fuzzy techniques n-gram and skipgram matching have been found to be effective in monolingual proper name [10] and cross-lingual spelling variant matching [5, 12] and transformation rule based translation technique (TRT) has been found to be an effective method for translating cross-lingual spelling variants [17]. N-grams and skipgrams are language independent techniques and the TRT technique can be easily adjusted for new language pairs. The methods are therefore easily applicable for new languages and thus ideal translation methods in CLIR. They are not dependent on expensive linguistic resources. In this study we used these dictionary-independent fuzzy string matching techniques as a query translation technique between closely related languages. The techniques were tested with the Scandinavian language pair Norwegian and Swedish, with Norwegian as the source language and Swedish as the target language.

Scandinavian languages have not been studied extensively from the information retrieval point of view. Hedlund et al. [3] is an exception. In their study characteristics of Swedish in information retrieval were analyzed. Swedish and Norwegian together with Danish form a language group where the speakers of one language can quite easily understand the other languages, especially in written form. Both the grammar and vocabulary of the languages are similar as they have developed in a close historical and cultural relation to one another. Some 50% of the Swedish and Norwegian (Bokmål) vocabulary is identical and around 40% similar, when inflected word forms and orthographical differences of using æ/ø instead of ä/ö are not considered [1]. There are also consistent and frequently occurring differences in the orthographies of Swedish and Norwegian. For

Proceedings of the sixth Dutch-Belgian Information Retrieval Workshop (DIR 2006)

example, Norwegian avoids the use of letters c, z, and x (*center (Swe) – senter (No)*) and the letter d is often left out of words where Swedish has it (*kunde (Swe) – kunne (No)*), the Danish letters æ/ø are used in Norwegian instead of Swedish ä/ö and the Swedish word endings -sion, -ssion and –tion are written –sjon in Norwegian. These similar features suggest that the use of fuzzy string matching techniques and the statistical transformation rules might be efficient in query translation from Norwegian to Swedish.

The research problems investigated in this paper are as follows:

1. Are fuzzy string matching methods as effective as the dictionary-based translation techniques in CLIR between closely related languages like Norwegian and Swedish?

2. Which of the fuzzy string matching methods tested is the most suitable translation technique for CLIR between closely related languages?

To the best of our knowledge, attempting to solve the query translation problem in CLIR between closely related languages with fuzzy string matching techniques without dictionary translation is a novel research idea not tried before.

The rest of the paper is organized as follows. The fuzzy string matching techniques used in this study are introduced in Section 2. Section 3 presents the test environment, methods and data. The similarity between Norwegian and Swedish is discussed in Section 4. Section 5 presents the findings and Section 6 discussion and conclusions for the study.

2. TRANSLATION TECHNIQUES

2.1 N-grams and Skipgrams

N-gram matching is a language independent method for matching words whose character strings are similar [13, 14]. Query keys and words in documents are decomposed into n-grams, i.e. into substrings of length *n*. The degree of similarity between the query keys and index terms can then be computed by comparing their n-gram sets. For a description of the applications of the technique, see [14]. N-gram matching has been reported to be an effective technique among fuzzy string matching techniques in name searching [10] and in cross-lingual spelling variant matching [5]. McNamee and Mayfield [7] used a direct corpus-based n-gram query translation technique, where the source language n-grams were directly translated to the target language n-grams using aligned corpora. The translation technique using 4- and 5-grams was found feasible. They also found n-grams an effective technique in tokenization, as it outperformed the stemmer used. Also Adafre et al. [1] have used 4-grams combined to a parallel corpus in query translation.

N-grams can consist both of adjacent characters or non-adjacent characters of the original words. Pirkola et al. [12] devised a novel matching technique for n-grams formed of non-adjacent characters, called the classified skipgram matching technique. In this technique digrams are divided into categories (classes) on the basis of the number of the skipped characters and only the digrams belonging to the same class are compared with each other. *Gram class* indicates the number of skipped characters when digrams are formed from a string *S*. *Character combination index* (CCI) then indicates a set of gram classes enumerating all the digram sets to be produced from the string *S*. For example *CCI* = {{0},{1,2}} means that two gram classes are formed from the string: one with conventional digrams formed of adjacent characters and one with skip-digrams formed both by skipping one and two characters [5]. The classified skipgrams have performed better than the traditional n-grams in the empirical tests examining the matching of cross-lingual spelling variants [5, 12].

It is common to use padding spaces in the beginning and in the end of the strings when forming n- and skipgrams. If the padding spaces are not used, the characters at the front and at the end of the strings will be under-represented in the gram set that is generated. Keskustalo et al. [5] tested different types of padding spaces for conventional digrams, trigrams and skipgrams, and found that using padding spaces both in the beginning and the end of the words gave the best results. However, the use of end padding spaces has been found unsuitable for inflectionally complex suffix languages, such as Finnish, where the use of the beginning padding only has been found beneficial [12]. This way of down-weighting the word ends – the inflectional suffixes – was assumed to be useful also when handling Swedish and Norwegian. For n-grams it is common to use a padding of n-1 characters [14]. For skipgrams a padding that varies according to the number of the skipped characters can be used.

The similarity values for n-grams are computed with a string similarity scheme [10]:

$$SIM(w_1, w_2) = \frac{|N_1 \cap N_2|}{|N_1 \cup N_2|}, \text{ where}$$

N_i is a digram set of a string w_i, $|N_1 \cap N_2|$ denotes the number of intersecting n-grams in strings w_1 and w_2, i.e. n-grams that the strings have in common, and $|N_1 \cup N_2|$ denotes the number of unique n-grams in the union of N_1 and N_2. The similarity measure for skipgrams is then defined between two strings *S* and *T* with respect to the given CCI as follows [5]:

$$SIM_{\mathrm{CCI}}(S,T) = \frac{\Sigma i_{\mathrm{CCI}} |DS_{\mathrm{i}}(S) \cap DS_{\mathrm{i}}(T)|}{\Sigma i_{\mathrm{CCI}} |DS_{\mathrm{i}}(S) \cup DS_{\mathrm{i}}(T)|}, \text{ where}$$

DS_i is the digram set of a string, *i* denoting the gram class, $|DS_i(S) \cap DS_i(T)|$ denotes the number of intersecting n-grams and $|DS_i(S) \cup DS_i(T)|$ the number of unique n-grams in the union of the strings *S* and *T*.

2.2 The TRT Technique

Transformation rule based translation (TRT) is a fuzzy translation technique based on the use of statistically generated rules of regular character correspondences in cross-lingual spelling variants within a language pair. The technique resembles transliteration, phonetic translation across languages with different writing systems, but no phonetic elements are included and the technique is meant for processing languages sharing the same writing system. It is applied in two-steps: the transformation rules are combined to n-gram matching. The idea of the TRT and the generation of the transformation rules are described in more detail in [17].

A *transformation rule* contains source and target language characters and their context characters [17]. In addition the frequency and the confidence factor of the rule are recorded. *Frequency* refers to the number of the occurrences of the rule in the data used for generating the rules. *Confidence factor* is the frequency of a rule divided by the number of source words where the source substring of the rule occurs. They are important threshold factors that can be used for selecting the most reliable rules for the translation. An example of a Norwegian to Swedish rule is:

for för beginning 132 147 89.80

The rule can be read: the letter o, prior to r and after f, is transformed into the letter ö in the beginning of words, with the confidence factor being 89.80. The confidence factor is calculated from the frequency of the rule (132) and the number of source words where the string occurs (147).

In this study we used the thresholds of confidence factor = 50% and frequency = 2.

3. METHODS AND DATA

3.1 Test Topics and Collection

The performance of the fuzzy translation methods was tested by running CLIR tests with a set of 60 topics used in the CLEF evaluation forum in the year 2003 [9]. Norwegian and Swedish topics were used, of which Swedish topics were included in the collection of the CLEF topics. To get the Norwegian topics, English topics were translated by a native Norwegian speaker. Of the two official Norwegian languages the more common Bokmål was used. In ten of the topics, queries failed in preliminary test runs for technical reasons. These topics were removed from all of the queries and the final tests were run with the remaining 50 topics. The target document collection was the Swedish CLEF document collection containing 142819 newspaper articles obtained from the Swedish news agency TT (Tidningarnas Telegrambyrå) published in 1994-1995 [9]. The document collection was lemmatized using Swetwol morphological analyzer by Lingsoft Inc. Compounds were split into their constituents and both the original word and the constituents were lemmatized and indexed. Words not recognized by the morphological analyzer were indexed as such to a separate index of unrecognized words. We used the InQuery Retrieval System as the search engine. InQuery is a probabilistic information retrieval system based on the Bayesian inference net model, where queries can be presented as unstructured bag-of-words queries or they can be structured with a variety of operators [2].

3.2 Creating TRT Rules

To create the word-pair list used for generating the Norwegian to Swedish transformation rules a part of the Swedish document collection's index was translated to Norwegian with the Global Dix dictionary by Kielikone plc. Words not recognized by the morphological analyzer were removed and, as the index was too large to use as a whole, every sixth word of the index was chosen. This list contained 6714 word-pairs. Word-pairs with an edit distance value bigger than half of the length of the longer word in the word-pair or including a word shorter than four characters were removed. The final word-pair list included 3058 unique word-pairs. This list seemed to be insufficient for generating enough high frequency transformation rules. This lack of high quality rules may have affected negatively the TRT technique's translation results.

3.3 N- and Skipgram Matching

The n- and skipgram translations were done by matching the n- or skipgrams of the topic words against the normalized index words of the Swedish test collection. The index was divided into two: the index of the words recognized by the morphological analyzer and the index of unrecognized words. Dividing the index is helpful when matching proper names [4]. For n-digram translation we used beginning weighted n-digrams with the padding of 1. Leaving out the padding at ends of words gives more weight to the beginnings of words, which can be useful when the words are inflected. For skipgram translation, a padding of the number of the skipped characters + 1 was used. For example for gram class 1, the skipgrams were formed with two padding spaces.

3.4 Queries

We used five sets of test queries, which were compared to three sets of baseline queries. The five translation methods tested were n-digrams, classified skipgrams with CCI = {{0}{1}} (*Skip1*) and CCI = {{0}{1,2}} (*Skip2*), plain TRT translation and the combined TRT and n-digram technique. The set of baseline queries consisted of a monolingual Swedish query set (*Swebase*), a monolingual Norwegian query set (*Nobase*) and a dictionary translated Norwegian to Swedish query set (*Dicbase*). The Global Dix dictionary was used for the translations. The Swebase and Dicbase gave high performing baselines, while the Nobase was used for testing how high performance is achieved without any translation and how much the fuzzy methods can improve this result.

The test query types were as follows. The query operators used in a query are presented in parentheses and examples of the queries are presented in Appendix 1.

1) Swedish monolingual baseline (#sum)

2) Norwegian monolingual baseline (#sum)

3) Dictionary baseline (#sum, #syn, #uw7)

4) N-digram query (#sum, #syn)

5) Skip1 query (#sum, #syn)

6) Skip2 query (#sum, #syn)

7) Plain TRT query (#sum, #syn)

8) Combined TRT and n-digram query (#sum, #syn)

The queries were formed from the title- and description fields of the CLEF topics. The topic words were lemmatized with the morphological analyzer Twol. For the dictionary translation, compound words were split into constituents that were then translated separately. This is because compound components are more often found in dictionaries than the whole compounds. For other query types, no compound splitting was done, as we assumed the compounds in Norwegian and Swedish to be similar. The lemmatized source words were translated and stop words were removed both before and after the translation.

The queries were formulated by grouping the query keys with InQuery's operators *sum*, *syn* and *uwn*. The sum-operator computes an average of query key weights for keys grouped by the operator. It is used for grouping the whole query and can include either the query keys without any structure or query key sets structured with the other operators. The syn-synonym operator treats its operand query keys as synonyms. The unordered proximity operator with a window size n (uwn) allows free word-order and combines the translations equivalents of the constituents of a source language compound [13].

The Swedish and Norwegian monolingual baseline queries were formed directly from the Swedish and Norwegian topic words as bag-of-words queries without any structure. The rest of the queries were structured with the syn-structure (*Pirkola's method*), which has been found effective in CLIR [11, 13, 16]. For the Dicbase queries all the translation equivalents of a source word were selected to the query and were grouped together with the syn-operator. When the translation was a noun phrase, its words were combined with a proximity operator of uwn, where we set the value of *n* to seven. Words not found in the dictionary were added to the query as such.

All the five test query types were structured queries, where the translation equivalents selected for a source word were grouped together with the syn-operator. For the n-gram and skipgram queries we selected for each source word the four highest ranked keys from the result list of n-gram matching. This selection was based on the findings by Hedlund et al. (2004), who showed that the best retrieval performance is achieved using just a few n-gram keys in queries [4]. These keys included two keys from the index of words recognized by the morphological analyzer and two from the index of unrecognized words.

For plain TRT-queries all the translated keys from the TRT result list were selected for each of the source word for the final queries. The combined TRT and n-digram queries were formed by selecting the first word form of each of the original source words from the TRT result list, which was then matched to the Swedish database index using n-digram matching. The word forms created with a rule combination with the highest confidence factor and frequency values get the highest position in the TRT result list. The four highest ranked keys from the result list of n-gram matching were then selected for the final queries like in other n-gram techniques.

3.5 Performance Measures

The effectiveness of the test queries was measured by Mean Average Precision (MAP) i.e., the average non-interpolated precision calculated over all relevant documents, and by interpolated recall precision averages at standard recall levels of 10 and 50, averaged over all queries. The test queries' precision-recall graphs were created using the eleven standard recall levels and the test queries' graphs were compared. The statistical significance of the results was tested using the Friedman two-way analysis of variance by ranks. The statistical significance levels are indicated in the tables.

4. SIMILARITY BETWEEN NORWEGIAN AND SWEDISH

To get an insight to how close two languages should be for the fuzzy matching to be practicable, the similarity of Swedish and Norwegian language was measured. A measure based on the Longest Common Subsequence (LCS) [8] was used, and German and English were used as a baseline language pair. They belong to the same language group but are not so closely related to make fuzzy matching alone a sufficient translation technique. The average similarity values measured for Swedish and Norwegian and for English and German were 0,815 and 0,556 respectively.

LCS is a measure that counts the maximum amount of letters that two words share and have in the same order, for example for a English - German word pair *motivation - motivierung* the longest common subsequence *motivin* has length 7. The data used for measuring the similarities between the languages included 167 word pairs for both language pairs. The vocabulary was selected from two sources: 71 words were chosen from the CLEF'03 topics and 96 words from a word list containing work environment vocabulary in all four languages (from the TNC-termbank by the Swedish national centre for terminology, TNC). The similarities were measured by first measuring the LCS values pair wise for all the words. Then each of these LCS values was divided by the length of the longer word of the word pair. Finally a mean value was calculated of these pair wise word similarity values for both language pairs. The similarity values range between 0-1. For example for the Swedish-Norwegian word pair *brevbomb - brevbombe* the LCS value is 8 and the similarity is counted by dividing it with the length of the longer of the words (here 9), with the similarity value being $8/9 \approx 0{,}889$.

Swedish, Norwegian and German are *compound languages* [4], i.e. languages where the components of multi-word expressions are written together, whereas English is a *non-compound language* where multi-word expressions are written as phrases (*fackförening, fagforening, gewerkschaft,* but *trade union*). The way the multi-word expressions are written is an important feature when measuring the orthographical similarity of languages. Therefore the test data included multi-word expressions. Phrases were written together by using a '_' to mark the space between the components (*trade_union*).

The similarity value of 0,815 measured for Swedish and Norwegian can be illustrated with examples: For a pair of short words such as *skola – skole* one character substitution results in a similarity value of 0,8. A longer word pair with a similarity value of 0,818 is *ioniserende – joniserande*, where two character substitutions happen. The orthographical differences in Swedish and Norwegian words are typically at this level. The mean similarity value of 0,556 measured for English and German corresponds to changes like *north_sea – nordsee*, which share five out of nine letters and get the similarity value of 0,556. The short word pairs *night – nacht* and *level - pegel,* where three letters out of five are common, get a similarity value of 0,6. The source of the vocabulary affected the similarity values slightly: the Swedish-Norwegian values for CLEF and TNC vocabularies were 0,829 and 0,805, respectively. English-German values were 0,582 for CLEF words and 0,536 for TNC words.

5. FINDINGS

5.1 The Performance of Fuzzy String Matching in Comparison to Baselines

Table 1 summarizes the Mean Average Precision values for all query types, and the performance differences between the test queries and the baseline queries. As the performances of the n-digram, skipgram and the combined TRT-n-gram queries were quite close to each other, they are referred together as the *n-gram queries* in the following. The performance differences between these queries are considered in Section 5.2.

The MAP is a measure that rewards techniques that retrieve relevant documents quickly [18]. When comparing the MAP values, the dictionary translation gives the best results, the monolingual Swedish baseline being second. The n-gram queries perform well: differences to the Dicbase and Swebase results are not statistically significant for any of the queries. The practical differences to the Dicbase are nevertheless noticeable (according to [15]) being over 5% for all fuzzy queries. All these techniques performed both statistically significantly and practically noticeably better than the Norwegian monolingual baseline. The plain TRT query's performance was better than the Nobase's, the difference not being statistically significant. The TRT query's performance was statistically significantly weaker than the Dicbase and Swebase baselines' performance.

Tables 2 and 3 present the recall precision averages at standard recall levels of 10 and 50. The Precision-Recall curves for all query types are shown in Figure 1. As can be seen from the P-R curves, the dictionary baseline and the Swedish monolingual baseline perform best on the high precision levels (0-20) and middle recall levels (20-80). For the high recall levels (80-100) the differences even up and the two skipgram queries perform as well as the Swebase baseline. Nobase and plain TRT queries still perform worse than the other queries.

At the recall level of 10 (Table 2), the dictionary baseline gets the highest precision average. The Swedish baseline is again the second best query type. The n-gram queries perform well, the differences to Dicbase and Swebase not being statistically significant. All the n-gram queries perform markedly better than the Nobase. Plain TRT query type is clearly worse than the Dicbase and Swebase baselines.

At the recall level of 50 (Table 3), the differences between different techniques diminish but the trend is still clear: Dictionary translation gives the best result followed by the monolingual Swedish query. The n-gram queries perform also well, the difference to Dicbase and Swebase not having statistical significance, although the practical differences between n-gram queries and Dicbase are noticeable. The plain TRT queries and Nobase are clearly the two weakest query types; their differences to the other query types are statistically significant.

5.2 Best Fuzzy String Matching Technique

The fuzzy queries were also compared to each other to determine the most suitable technique for CLIR between closely related languages. As can be seen from Figure 1, the plain TRT queries' P-R curve is consistently clearly below the other curves. The difference to the other fuzzy queries is most of the time statistically significant or highly significant, and the practical difference is always noticeable. Therefore it can be concluded that, when used alone, it is not an adequate translation technique in CLIR between closely related languages. Earlier research results from [17] support this conclusion. In this research, the TRT queries' performance may have been negatively affected by the lack of high frequency transformation rules. This may also have affected the results of the combined TRT-n-gram queries.

The findings do not suggest one fuzzy string matching technique as being the best translation method in CLIR between closely related languages. The differences between the different n-gram queries were small and statistically insignificant. The combined TRT-n-gram queries performed best on the high precision levels and the practical difference to the plain n-gram queries was noticeable at the recall level of 10. On the middle recall levels all the n-gram queries were quite even and their differences had no statistical or practical significance at the recall level of 50. Here the skipgram queries gave the best results, the Skip2 -grams with CCI={{0}{1,2}} being the best query type. From the Figure 1 it can be seen that the P-R curves of skipgram queries are above the others fuzzy queries' curves at the high recall levels.

Even if the differences are small, the Skip2 queries and the combined TRT-n-gram queries performed slightly better than the other queries. At the same time, the combined TRT-n-gram queries outperformed the plain n-gram queries indicating that the transformation rules do improve n-gram results in CLIR between closely related languages.

Table 1. The MAP values (%) for the test queries and their difference to the baselines (%) (* statistically significant difference, ** statistically highly significant difference)

	Baseline queries			**Test queries**				
	Nobase	Swebase	Dicbase	Skip1	Skip2	N-gram	Plain TRT	TRT-n-gram
Precision	12,64	31,76	34,13	28,34	28,63	26,53	16,88	27,74
Difference to Nobase	0	19,12	21,49	15,7*	*15,99**	13,89**	4,24	15,1**
Difference to Swebase		0	2,37	-3,42	-3,13	-5,23	*-14,88***	-4,02
Difference to Dicbase			0	-5,79	*-5,5*	-7,6	*-17,25***	-6,39

Table 2. The interpolated recall precision averages (%) at standard recall level 10 for the test queries, and their difference to the baselines. (* statistically significant difference, ** statistically highly significant difference)

	Baseline queries			Test queries				
	Nobase	Swebase	Dicbase	Skip1	Skip2	N-gram	Plain TRT	TRT-n-gram
Precision	21,85	50,65	54,91	44,39	43,95	41,44	28,17	46,54
Difference to Nobase	0	28,8	33,06	22,54**	22,1*	19,59**	6,32	24,69**
Difference to Swebase		0	4,26	-6,26	-6,7	-9,21	-22,48**	-4,11
Difference to Dicbase			0	-10,52	-10,96	-13,47	-26,74**	-8,37

Table 3. The interpolated recall precision averages (%) at standard recall level 50 for the test queries, and their difference to the baselines. (* statistically significant difference, ** statistically highly significant difference)

	Baseline queries			Test queries				
	Nobase	Swebase	Dicbase	Skip1	Skip2	N-gram	Plain TRT	TRT-n-gram
Precision	13,1	31,03	35,64	28,81	29,58	27,02	15,78	28,77
Difference to Nobase	0	17,93	22,54	15,71	16,48	13,92*	2,68	15,67**
Difference to Swebase		0	4,61	-2,22	-1,45	-4,01	-15,25**	-2,26
Difference to Dicbase			0	-6,83	-6,06	-8,62	-19,86**	-6,87

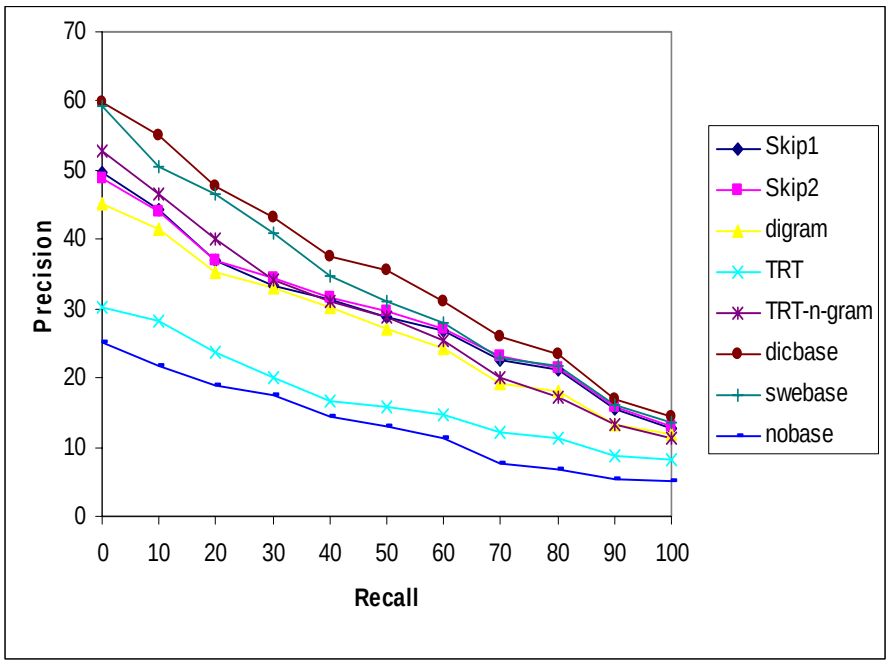

Figure 1. Recall-precision curves for all queries.

6. DISCUSSION AND CONCLUSIONS

The aim of this research was to find out (1) if fuzzy matching techniques are as effective as the dictionary-based translation techniques in CLIR between closely related languages like Norwegian and Swedish, and (2) the most suitable fuzzy string matching technique for query translation in CLIR between closely related languages. The effectiveness of five fuzzy string matching techniques was tested for Norwegian to Swedish query translation with CLEF search topics from the year 2003. The fuzzy techniques were compared to three baseline techniques, which

were a dictionary translation baseline, a monolingual Swedish baseline and a monolingual Norwegian baseline.

Our main findings were:

- The fuzzy (n-gram) matching techniques are effective and applicable translation techniques in CLIR between closely related languages. For the best fuzzy matching query types performance difference with respect to dictionary translation queries was not statistically significant.
- The results do not suggest one best fuzzy matching technique for CLIR between closely related languages.
- The TRT technique alone is not a good approach (however, see below for the generation of transformation rules).

The results were encouraging giving support to our hypothesis that dictionary-based translation could be replaced by fuzzy string matching techniques in CLIR between closely related languages. The n-gram based techniques performed well, skipgrams being slightly better than conventional n-grams. This is in line with earlier research, where skipgrams have been found to be better than n-grams in matching cross-lingual spelling variants [5, 12]. Combining n-grams to the TRT techniques' statistical transformation rules improved results, the practical difference being of noticeable (5,1%) at the recall level 10. The TRT-n-grams also outperformed the best skipgrams at low recall levels. This suggests that the combined technique is useful in CLIR, as also found in earlier research [17]. The results also give reason to assume that combining the transformation rules to skipgram matching would be a good approach. This combination can be assumed to perform well, as the skipgrams have been shown to outperform the conventional n-grams in cross-lingual spelling variant matching [5, 12].

The results suggests that the transformation rules should be formed on a basis of a larger term pair list than was done in this study, or the list should be formed from technical terms instead of general vocabulary. The performance of the TRT queries might improve if the transformation rules were thereby improved. Better transformation rules might also further improve the performance of the combined TRT and n-gram queries.

In the present research, all the query words were lemmatized because the transformation rules in their current state can only handle base forms. Creating transformation rule collection capable of handling inflected word forms will be one of the next steps in our research. Our future research will also include testing the combination of TRT and skipgrams, and extending the research to concern Danish language.

7. REFERENCES

[1] Adafre, S., van Hage, W., Kamps, J., de Melo, G. & de Rijke, M. 2004. The University of Amsterdam at CLEF 2004. CLEF 2004 Working Notes. Available at: Available at: http://clef.iei.pi.cnr.it/

[2] Barðdal, J., Jörgensen, N., Larsen, G., & Martinussen B. 1997. Nordiska: Våra språk förr och nu. Lund, Studentlitteratur.

[3] Broglio, J., Callan, J. & Croft B. 1993. Inquery system overwiev. In Proceedings of the TIPSTER text program, 47-67. Available: http://acl.ldc.upenn.edu/X/X93/X93-1008.pdf

[4] Hedlund, T., Pirkola, A. & Järvelin, K. 2001. Aspects of Swedish morphology and semantics from the perspective of mono- and cross-language information retrieval. Information Processing & Management, 37, 147-161.

[5] Hedlund, T., Airio, E., Keskustalo, H., Lehtokangas, R., Pirkola, A. & Järvelin, K. 2004. Dictionary-based Cross-Language Information Retrieval: Learning Experiences from CLEF 2000-2002. Information Retrieval – Special Issue on CLEF Cross-Language IR, 7, 99-119.

[6] Keskustalo, H. & Pirkola, A. & Visala, K. & Leppänen, Erkka & Järvelin, K. 2003. Non-adjacent Digrams Improve Matching of Cross-Lingual Spelling Variants. In: Nascimento, M.A., de Moura, E.S., Oliveira, A.L, (Eds.). Proceedings of the 10th International Symposium, SPIRE 2003. Manaus, Brazil, October 2003. Berlin: Springer, Lecture Notes in Computer Science 2857, pp. 252 - 265. ISSN 0302-9743, ISBN 3-540-20177-7.

[7] Kraaij, W. 2004. Variations on language modeling for information retrieval. PhD thesis, University of Twente.

[8] McNamee, P. & Mayfield, J. 2003. JHU/APL Experiments in Tokenization and Non-Words Translation. CLEF 2003 Working Notes. Available at: http://clef.iei.pi.cnr.it/

[9] Navarro, G. 2001. A Guided tour to approximate string matching. ACM Computing surveys (CSUR) (33)1.

[10] Peters, C. 2003. Introduction to the CLEF 2003 Working Notes. Available at: http://clef.iei.pi.cnr.it/

[11] Pfeiffer, U., Poersch, T. & Fuhr, N. 1996. Retrieval effectiveness of proper name search methods. Information Processing & Management, 32(6), 667-679.

[12] Pirkola, A. 1998. The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. In: Proceedings of the 21st Annual International ACM Sigir Conference on Research and Development in Information Retrieval, Melbourne, August 24-28. New York: ACM, 55-63.

[13] Pirkola, A., Keskustalo H., Leppänen, E., Känsälä, A.P. & Järvelin, K. 2002. Targeted s-gram matching: a novel n-gram matching technique for cross- and monolingual word form variants. Information research, 7(2) [http://InformationR.net/ir/7-2/paper126.html]

[14] Pirkola, A., Puolamäki, D. & Järvelin, K. 2003. Applying query structuring in Cross-Language Retrieval. Information Processing & Management 39(3), 391-402.

[15] Robertson, A.M. & Willet, P. 1998. Applications of n-grams in textual information systems. Journal of Documentation, 54(1), 48-69.

[16] Spark Jones, K. 1974. Automatic indexing. Journal of Documentation (30) 4, 393-432.

[17] Sperer, R. & Oard, D. W. 2000. Structured Translation for Cross-Langiage Information Retrieval. In Belkin, N. & Ingwersen, P. & Leong, M-K. (Eds.), Proceedings of the 23rd Annual International SIGIR Conference on Research and Development in Information Retrieval, 120-127. Athens, Greece.

[18] Toivonen, J., Pirkola, A., Keskustalo, H., Visala, K., & Järvelin, K. 2005. Translating cross-lingual spelling variants

using transformation rules. Information Processing & Management, 41, 859-872.

[19] Voorhees, E. M. 2002. Overview of TREC 2002, Appendix 1. Common Evaluation Measures. The Proceedings of the eleventh Text REtrieval Conference. Gaithersburg, Maryland. National Institute of Standards and Technology. [http://trec.nist.gov/pubs.htm

Appendix 1. Examples for query types

Swebase

#sum(christo packeterar tyska riksdagshus konstnär christo inslagning tyska riksdagshus)

Nobase

#sum(christo pakke tysk riksdagsbygning innpakking tysk riksdag berlin kunstner christo)

Dicbase

#sum(christo #syn(paket packe bunt ask packa) #syn(tysk tyska) #syn(regerings stats stat statlig) dag #syn(byggnadsverk byggnad konstruktion hus) #syn(packning) #syn(tysk tyska) #syn(regerings stats stat statlig) dag berlin konstnär christo)

N-digram query

#sum(#syn(mchistori chefshistorik @christo @christos) #syn(paket pakets @paker @pak) #syn(tysk tysktysk @tyskl @tysklan) #syn(riksdagsbyggnad riksdagsbevakning @riksdagsoch @landsbygsriksdagen) #syn(skinnpaj inpassning @pakkinen @iakkinen) tysk #syn(tysk tysktysk @tyskl @tysklan) #syn(riksdag riksdagsdag @riksdagsoch @riksdagsrupp) #syn(berliner berlinsk @berlin @berlins) #syn(kungstiger kungakonst @kunst @kunstler) #syn(mchistori chefshistorik @christo @christos))

Skip1 query (CCI = {{0},{1}})

#sum(#syn(chefjurist charterturistort @christo @christos) #syn(packe paket @takke @pakue) #syn(tysk tysktysk @tyskl @otysk) #syn(riksdagsbevakning riksdagsordning @riksdagsoch @riksdagsrupp) #syn(inpackning inpassning @ing @king) #syn(tysk tysktysk @tyskl @otysk) #syn(riksdag riksdagsdag @riksdagsoch @riksdagsrupp) #syn(berglin merlin @berlin @berlins) #syn(konstnär konstnummer @kunstler @köstner) #syn(chefjurist charterturistort @christo @christos))

Skip2 query (CCI = {{0},{1,2}})

#sum(#syn(tyristor mchistori @christo @christos) #syn(paket packe @pakue @takke) #syn(tysk tysktysk @tyskl @otysk) #syn(riksdagsbyggnad riksdagsbevakning @riksdagsebatten @riksdagsrupp) #syn(inpackning inpassning @king @parking) #syn (tysk tysktysk @tyskl @otysk) #syn(riksdag riksdagsdag @riksdagsoch @riksdagsrupp) #syn(berglin merlin @berlin @berlins) #syn(konstnär konstcenter @kunstler @kunstlers) #syn(tyristor mchistori @christo @christos))

TRT query

#sum(#syn(christo) #syn(packa pakka packe pakke) #syn(tysk) #syn(riksdagsbygning) #syn(innpacking innpakking) #syn(tysk) #syn(riksdag) #syn(berlin) #syn(kunstner) #syn(christo))

Combined TRT and n-digram

#sum(#syn(mchistori chefshistorik @christo @christos) #syn(packa packad @packard @packalén) #syn(tysk tysktysk @tyskl @tysklan) #syn(riksdagsbyggnad riksdagsbevakning @riksdagsoch @landsbygsriksdagen) #syn(inpackning inpacka @inpac @racking) #syn(tysk tysktysk @tyskl @tysklan) #syn(riksdag riksdagsdag @riksdagsoch @riksdagsrupp) #syn(berliner berlinsk @berlin @berlins) #syn(kungstiger kungakonst @kunst @kunstler) #syn(mchistori chefshistorik @christo @christos))

DCU and UTA at ImageCLEFPhoto 2007

Anni Järvelin[1], Peter Wilkins[2], Tomasz Adamek[2], Eija Airio[1], Gareth J.F. Jones[2], Alan F. Smeaton[2], and Eero Sormunen[1]

[1] University of Tampere (UTA), Finland
Anni.Jarvelin@uta.fi
[2] Dublin City University (DCU), Ireland
pwilkins@computing.dcu.ie

Abstract. Dublin City University (DCU) and University of Tampere (UTA) participated in ImageCLEF 2007 photographic retrieval task with several monolingual and bilingual runs. The approach was language independent with text retrieval utilizing fuzzy *s*-gram query translation and combined with visual retrieval. Data fusion was achieved through unsupervised query-time weight generation approaches. The baseline was a combination of dictionary-based query translation and visual retrieval, which achieved the best result. The best mixed modality runs using fuzzy *s*-gram translation reached on average around 83% of the baselines' performance. This approach was much closer at the early precision levels of P@10 and P@20. This suggests that our language independent approach could be a cheap alternative for cross-lingual image retrieval. Both sets of results further emphasize the merit in our query-time weight generation schemes for data fusion, with the fused runs exhibiting marked performance increases over single modalities without the use of prior training data.

1 Introduction

Retrieving images by their associated text is a common approach in image retrieval [1]. When cross-language image retrieval is considered, this approach requires language dependent linguistic resources for query translation. Machine-readable dictionaries, machine translation tools or corpus-based translation tools are expensive and not available for all the language pairs. However, there are alternative approaches which may be used to compensate linguistic tools, for example the fuzzy string matching technique *n*-gram matching and its generalization *s*-gram matching. These techniques have previously been used for translation of query words missing from dictionaries [2][3], but only rarely for the whole query translation [4][5].

In earlier ImageCLEF campaigns combined text and visual retrieval approaches have performed better than text or image retrieval alone. In this year's campaign, text retrieval faced a new challenge of retrieval of lightly annotated photographs [1]. A negative impact on the performance of the text retrieval techniques was to be expected and therefore successful fusion of text and visual

C. Peters et al. (Eds.): CLEF 2007, LNCS 5152, pp. 530–537, 2008.

features became even more important. We tested a language independent image retrieval approach, where s-gram-based fuzzy query translations were fused with visual retrieval. We explored data fusion techniques with query-time coefficient generation for retrieval expert combination. We experimented primarily with altering the stages at which we fuse various experts together. For instance we experimented with fusing all the visual experts into a single expert, then fusing with text, as opposed to treating all experts equally. To have a strong baseline, the performance of the language independent approach was compared to a combination of dictionary-based query translation and visual retrieval.

To study the effect of the source and target languages on the quality of the fuzzy translation we selected six language pairs where source/target languages were related to each other at different levels. The Scandinavian languages Danish, Norwegian and Swedish were translated into German, to which they are quite closely related to. French was the source language that was closest to English. German and English are not very closely related and translation between them was done both ways. (See [4] for a matrix for similarities between English, French, German and Swedish) A total of 138 runs were submitted. Reporting the results for all of these would be impractical and therefore only the results for the most interesting runs are presented here.

2 Background

s-gram-based query translation. The s-gram matching is a fuzzy string matching technique that measures similarity between text strings. The text strings to be compared are decomposed into substrings (s-grams) and the similarity is calculated as the overlap of their common substrings. s-gram matching is a generalization of n-gram technique, commonly used for matching cognates in Cross Language Information Retrieval (CLIR). While the n-gram substrings consist of adjacent characters of the original strings, skipping some characters is allowed when forming the s-grams. In classified s-gram matching [6], different types of s-grams are formed by skipping different number of characters. The s-grams are then classified into sets based on the number of characters skipped and only the s-grams belonging to the same set are compared to each other when calculating the similarity. Character Combination Index (CCI) indicates the set of all the s-gram types to be formed from a string. CCI $\{\{0\},\{1,2\}\}$ for example means that three types of s-grams are formed and classified into two sets: one set of conventional n-grams formed of adjacent characters ($\{0\}$) and one of s-grams formed both by skipping one and two characters ($\{1,2\}$). For an extensive description of the s-gram matching technique, see [6][7].

Proper names are very common query terms when searching from image databases [8] and are typically not covered by translation dictionaries and thus remain untranslatable in queries. Proper names in related languages are often spelling variants of each other and can thus be translated using approximate string matching.

Visual Retrieval. To facilitate visual retrieval we made use of six 'low-level' global visual experts. Our visual features are MPEG7 features and were extracted using the aceToolBox, developed as part DCU's collaboration in the aceMedia project [9]. These six features included: Colour Layout, Colour Structure, Colour Moments, Scalable Colour, Edge Histogram and Homogenous Texture. Further details on these descriptors can be found in [10] and [11]. Distance metrics for each of these features were implementations of those specified in the MPEG7 specification [11].

Query-Time Fusion. The combination of retrieval experts for a given information need can be expressed as a *data fusion* problem [12]. Given that for any given information need different retrieval experts perform differently, we require some form of weighting scheme in order to combine experts. Typical approaches to weight generation include the use of global query-independent weights or query-class expert weights learnt through experimentation on a training collection to name a few.

Our approach to weight generation differs in that it is a query-dependant weighting scheme for expert combination which requires no training data [13]. This work was based upon an observation, that if we were to plot the normalized scores of an expert, against that of scores of other experts used for a particular query, that the expert who's scores exhibited the greatest initial change in scores correlated with that expert being the best performer for that query. Examples of this observation can be seen in [13] and our ImageCLEF workshop paper [14]. This approach is not giving us any absolute indication of expert performance, which other approaches to examining score distributions attempt to provide, an excellent review of which is given by Robertson [15]. We would note that this observation is not universal, and we can identify failure cases where this observation will not occur. If we assume though that in a majority of queries this observation will hold, then we can employ techniques that leverage this approach to create query-time expert coefficients for data fusion. Our main technique involves measuring the change in scores for a retrieval expert over a top subset of its results, versus the change in scores over a larger sample of that experts scores. The expert which undergoes the greatest initial change in score is assigned a greater weight. A complete explanation of this process can be found in [13].

3 Resources, Methods and Runs

Text Retrieval and Indexing. We utilized the Lemur toolkit (Indri engine) [16] for indexing and retrieval. Indri combines language modeling to inference nets and allows structured queries to be evaluated using language modeling estimates. The word tokenization rules used in indexing included converting punctuation marks into spaces and capitals were converted into lower case. Strings separated by the space character were tokenized into individual words. For the dictionary-based translation, lemmatized English and German indices were created. The image annotation text was lemmatized, words not recognized by the lemmatizers were indexed as such. Compound words were split and both the

original compound and the constituents were indexed. For the *s*-gram-based translation we used inflected indices, where the words were stored in the inflected word forms in which they appeared in the image annotations. Stop words were removed.

Topics were processed as the indices. For the *s*-gram-based translation stop words were removed from the queries and the remaining words were translated into the target language with *s*-gram matching. The CCI was set to be {{0},{1,2}}, and the Jaccard coefficient [7] was used for calculating the *s*-gram proximity. Three best matching index words were selected as translations for each topic word. A similarity threshold value of 0.3 was set to discard bad translations, only the index words exceeding this threshold with respect to a source word were accepted as translations. As a consequence some of the query words could not be translated. Queries were structured utilizing a synonym operator where target words derived from the same source word were grouped into the same synonym group (*the Pirkola method*, [17]). Indris Pseudo Relevance Feedback (PRF) was used with 10 keys from the 10 highest ranked documents in the original result list.

For the dictionary-based query translation, the UTACLIR query translation tool was used. UTACLIR was originally developed for the CLEF 2000 and 2001 campaigns [2]. It utilizes external language resources, such as translation dictionaries, stemmers and lemmatizers. Topic words were lemmatized, stop words removed and finally the non-stop words translated. Next, untranslatable compound words were split and the constituents were translated. Translation equivalents were normalized utilizing a target language lemmatizer. Untranslatable words were matched against the database index using the *s*-gram matching. Queries were structured with the synonym operator. A morphological analyzer for French was not available and therefore the French topics were analyzed manually. This might have resulted in a slightly better quality of lemmatization than automatic analysis, even though we strived for comparable quality. We used PRF with 20 expansion keys from the 15 top ranked documents.

Data Fusion. The query-time data fusion approach specified in Section 2 describes our basic approach to expert combination. However, one set of parameters that was not specified was the order in which experts will be combined. This is the focus of our experimental work in this section.

One commonality between all the combination approaches we try in this work is the fusion of the low-level visual experts. For each query image we fuse the six low-level visual experts into a single result for each image, where the combination of these is using the aforementioned technique. Therefore for each query, the visual component was then represented by three result sets, one for each query image. Additionally for a subset of our runs we introduce a seventh visual expert, the FIRE baseline [18]. In cases where FIRE was used, because it was a single result for the three visual query images, we first combined our MPEG7 visual features into a single result for each image, then these combined into a overall

image result, which was then combined with the FIRE baseline. There are four variants that we tried in our combination work, which are:

- dyn-equal: Query-time weighting method used, text and individual image results combined at the same level (i.e. we have three image results and one text result which is to be combined).
- dyn-top: As above, except the results for each query image were fused into a single image result, which was then combined with the text result (i.e. image results combined into a single result, which is then combined with the single text result).
- stat-eventop: Query-time weighting to produce single image result list, image and text fused together with equal static weighting (0.5 coefficient).
- stat-imgHigh: As above, except with the image result assigned a static weight of 0.8 and text a static weight of 0.2.

Additionally, any of our runs which ended in 'fire' incorporated the FIRE baseline into the set of visual experts used for combination.

4 Results

Our tables of results are organized as follows. Table 1 presents our baseline runs, including monolingual text-only, visual-only and baseline fusion results mixing these two types. Table 2 presents our central cross-lingual results with mixed modalities. In all tables where data fusion is utilized, we present only the best performing data fusion approach. Except where noted, all visual results used in data fusion presented here incorporated the FIRE baseline as visual data which included the FIRE baseline with our global MPEG7 features consistently outperformed global MPEG7 by themselves.

For the monolingual runs in Table 1, the runs where morphological analysis (dict) was used performed slightly better than the *s*-gram runs. The difference is small for the English runs. For German runs the difference is greater, which is understandable as German has a more complex inflectional morphology than

Table 1. ImageCLEF Baseline Results

Language Pair	Modality	Text	Fusion	FB	MAP	P@10	P@20
EN-EN	Text	dict	n/a	no	0.1305	0.1550	0.1408
EN-EN	Text	*s*-gram	n/a	yes	0.1245	0.1133	0.1242
DE-DE	Text	dict	n/a	yes	0.1269	0.1717	0.1533
DE-DE	Text	*s*-gram	n/a	yes	0.1067	0.1233	0.1125
MPEG7 With FIRE	Visual	na	dyn-equal	no	0.1340	0.3600	0.2658
MPEG7 Without FIRE	Visual	na	dyn-equal	no	0.1000	0.2700	0.1958
EN-EN	Mixed	dict	dyn-equal	yes	0.1951	0.3967	0.3150
EN-EN	Mixed	*s*-gram	dyn-equal	yes	0.1833	0.3833	0.3092
DE-DE	Mixed	dict	dyn-equal	yes	0.1940	0.4033	0.3300
DE-DE	Mixed	*s*-gram	dyn-equal	yes	0.1628	0.3350	0.2792

English. Our text and visual retrieval techniques were almost equal, which is notable in the context of earlier years' ImageCLEF results. Our best visual-only run performed well being the second best visual approach in terms of Mean Average Precision (MAP). Its MAP value 0.1340 is comparable to our best monolingual English text run scoring 0.1305. We believe that the comparative low performance of the text expert (when compared to the dominance of text in previous years) was due to the reduced length of the annotations for 2007. Table 1 also presents our fused monolingual text and visual retrieval runs, which performed clearly better than any of the text or visual runs alone. Fusion of these modalities produced consistent improvements in MAP of between 65% and 67%.

From a data fusion perspective, no single approach of the four we tried consistently performed the best. Whilst our results presented here show the "dyn-equal" fusion as being superior, this is because it was the only fusion type attempted with visual data which incorporated the FIRE baseline. For runs where FIRE was not used, there best performing fusion type varied depending on the text type (dictionary or sgram) or language pair used. In a majority of cases all fusion types performed similarly, as such deeper investigation with significance testing will be required in order to infer any meaningful interpretations. However, we can conclude that as all four fusion types made use of our query-time weight generation method at some level, that this technique is capable of producing weights which lead to performance improvements when combining results. What is unknown is how far from the optimal query-dependant combination performance we achieved, and that will be one of the ultimate measures of the success of this approach.

Table 2 presents our central cross-lingual results. Dictionary-based query translation was the best query translation approach. The best mixed modality runs using the *s*-gram-based query translation nevertheless reached on average around 84% of the MAP of the best mixed modality runs using dictionary-based translation. The difference between the approaches further decreased when the early precision values of P@10 and P@20 were considered. The best *s*-gram runs

Table 2. ImageCLEF CLIR Fusion Results

Language Pair	Modality	Text	Fusion	FB	MAP	P@10	P@20
FR-EN	Mixed	dict	dyn-equal	yes	0.1819	0.3583	0.2967
FR-EN	Mixed	*s*-gram	dyn-equal	no	0.1468	0.3483	0.2667
DE-EN	Mixed	dict	dyn-equal	yes	0.1910	0.3483	0.3042
DE-EN	Mixed	*s*-gram	dyn-equal	yes	0.1468	0.3233	0.2533
DA-DE	Mixed	dict	dyn-equal	yes	0.1730	0.3467	0.2942
DA-DE	Mixed	*s*-gram	dyn-equal	yes	0.1572	0.3350	0.2717
NO-DE	Mixed	dict	dyn-equal	yes	0.1667	0.3517	0.2700
NO-DE	Mixed	*s*-gram	dyn-equal	yes	0.1536	0.3167	0.2667
SV-DE	Mixed	dict	dyn-equal	yes	0.1788	0.3817	0.2942
SV-DE	Mixed	*s*-gram	dyn-equal	yes	0.1472	0.3050	0.2500
EN-DE	Mixed	dict	dyn-equal	yes	0.1828	0.3633	0.3008
EN-DE	Mixed	*s*-gram	dyn-equal	yes	0.1446	0.3350	0.2667

reached on average around 91% of the best dictionary-based runs performance at P@10 and around 89% at P@20. If the high ranks of the result list are considered to be important from the user perspective, the *s*-gram translation could be seen as almost equal with the dictionary-based translation in mixed modality runs. These results varied depending on the language pair. *s*-gram-based and dictionary-based translation performed similarly for the closely related language pairs, while the differences were greater for the more distant language pairs. The *s*-gram translation reached its best results with Norwegian and Danish topics and German annotations - over 90% of the dictionary translation's MAP, and the worst ones between German and English - less than 80% of the dictionary translation's MAP.

5 Conclusions

In this paper we have presented the joint DCU and UTA ImageCLEF 2007 Photo results. In our work we experimented with two major variables, that of cross-lingual text retrieval utilizing minimal translation resources, and query-time weight generation for expert combination. Our results are encouraging and support further investigation into both approaches. Further work is now required to conduct a more thorough analysis of contributing factors to performance emphasized by each approach. Of particular interest will be the degree to which each of these approaches introduced new information, or re-ordered existing information presented by the systems. For instance, we do not know yet if the *s*-gram retrieval found relevant documents that were missed by the dictionary based approach. Likewise for data fusion, we do not know yet if we promoted into the final result set relevant results which were only present in some and not all of the experts used.

Acknowledgments

We are grateful to the AceMedia project (FP6-001765) which provided the image analysis tooklit. Research leading to this paper was supported by the European Commission under contract FP6-027026 (K-Space). The work of the first author is funded by Tampere Graduate School of Information Science and Engineering (TISE).

References

1. Grubinger, M., Clough, P., Hanbury, A., Müller, H.: Overview of the ImageCLEF-photo 2007 photographic retrieval task. In: Working Notes of the 2007 CLEF Workshop, Budapest, Hungary (September 2007)
2. Hedlund, T., Keskustalo, H., Pirkola, A., Airio, E., Järvelin, K.: Utaclir @ CLEF 2001 - effects of compound splitting and n-gram techniques. In: Peters, C., Braschler, M., Gonzalo, J., Kluck, M. (eds.) CLEF 2001. LNCS, vol. 2406, pp. 118–136. Springer, Heidelberg (2002)

3. Hiemstra, D., Kraaij, W.: Twenty-One at TREC7: Ad-hoc and cross-language track. In: TREC, pp. 174–185 (1998)
4. Mcnamee, P., Mayfield, J.: Character n-gram tokenization for european language text retrieval. Information Retrieval 7(1-2), 73–97 (2004)
5. Järvelin, A., Järvelin, A., Järvelin, K.: s-grams: Defining generalized n-grams for information retrieval. Information Processing and Management 43(4), 1005–1019 (2007)
6. Pirkola, A., Keskustalo, H., Leppänen, E., Känsälä, A.P., Järvelin, K.: Targeted s-gram matching: a novel n-gram matching technique for cross- and mono-lingual word form variants. Information Research 7(2) (2002)
7. Keskustalo, H., Pirkola, A., Visala, K., Leppänen, E., Järvelin, K.: Non-adjacent digrams improve matching of cross-lingual spelling variants. In: SPIRE, pp. 252–265 (2003)
8. Markkula, M., Sormunen, E.: End-user searching challenges indexing practices inthe digital newspaper photo archive. Information Retrieval 1(4), 259–285 (2000)
9. AceMedia: The AceMedia Project, `http://www.acemedia.org`
10. O'Connor, N., Cooke, E., le Borgne, H., Blighe, M., Adamek., T.: The AceToolbox: Low-Level Audiovisual Feature Extraction for Retrieval and Classification. In: 2nd IEE European Workshop on the Integration of Knowledge, Semantic and Digital Media Technologies (2005)
11. Manjunath, B., Salembier, P., Sikora, T. (eds.): Introduction to MPEG-7: Multimedia Content Description Language. Wiley, Chichester (2002)
12. Belkin, N.J., Kantor, P., Fox, E.A., Shaw, J.A.: Combining the evidence of multiple query representations for information retrieval. Information Processing and Management 31(3), 431–448 (1995)
13. Wilkins, P., Ferguson, P., Smeaton., A.F.: Using score distributions for querytime fusion in multimedia retrieval. In: MIR 2006 - 8th ACM SIGMM International Workshop on Multimedia Information Retrieval (2006)
14. Jarvelin, A., Wilkins, P., Adamek, T., Airio, E., Jones, G., Smeaton, A.F., Sormunen, E.: DCU and UTA at ImageCLEFPhoto 2007. In: ImageCLEF 2007 - The CLEF Cross Language Image Retrieval Track Workshop (2007)
15. Robertson, S.: On score distributions and relevance. In: Amati, G., Carpineto, C., Romano, G. (eds.) ECiR 2007. LNCS, vol. 4425, pp. 40–51. Springer, Heidelberg (2007)
16. Strohman, T., Metzler, D., Turtle, H., Croft, W.B.: Indri: A language-model based search engine for complex queries (extended version) (2005-02-14) (2005)
17. Pirkola, A.: The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. In: SIGIR 1998: Proceedings of the 21st Annual ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 55–63 (1998)
18. Deselaers, T., Weyand, T., Keysers, D., Macherey, W., Ney, H.: FIRE in ImageCLEF 2005. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, pp. 652–661. Springer, Heidelberg (2006)

Information retrieval from historical newspaper collections in highly inflectional languages: a query expansion approach

Anni Järvelin*)

School of Information Sciences, FIN-33014 University of Tampere, Finland

anni.jarvelin@uta.fi

Heikki Keskustalo

School of Information Sciences, FIN-33014 University of Tampere, Finland

heikki.keskustalo@uta.fi

Eero Sormunen

School of Information Sciences, FIN-33014 University of Tampere, Finland

eero.sormunen@uta.fi

Kimmo Kettunen

Centre for Preservation and Digitisation, National Library of Finland

Saimaankatu 6, FI-50100 Mikkeli, Finland

kimmo.kettunen@helsinki.fi

Miamaria Saastamoinen

School of Information Sciences, FIN-33014 University of Tampere, Finland

miamaria.saastamoinen@uta.fi

*) corresponding author

Abstract

PURPOSE: The aim of the study was to test whether query expansion by approximate string matching methods is beneficial in retrieval from historical newspaper collections in a language rich of compounds and inflectional forms (Finnish).

APPROACH: First, approximate string matching methods were used to generate lists of index words most similar to contemporary query terms in a digitized newspaper collection from the 1800s. Top index word variants were categorized to estimate the appropriate query expansion ranges in the retrieval test. Second, the effectiveness of approximate string matching methods, automatically generated inflectional forms and their combinations was measured in a Cranfield-style test. Finally, a detailed topic-level analysis of test results was conducted.

FINDINGS: In the index of historical newspaper collection the occurrences of a word typically spread to many linguistic and historical variants along with optical character recognition (OCR) errors. All query expansion methods improved the baseline results. Extensive expansion of around 30 variants for each query word was required to achieve the highest performance improvement. Query expansion based on approximate string matching was superior to using the inflectional forms of the query words, showing that coverage of the different types of variation is more important than precision in handling one type of variation.

VALUE: There are very few test collections for, and rigorous evaluations on, historical document retrieval. The specific viewpoint adopted in this study is to evaluate an approach for matching modern query words with their historical variants both in terms of translation precision and retrieval performance. Our results show that striving for high translation precision is not necessarily primary: coverage of the variants might be more important.

Introduction

Digitization is a good way to preserve cultural heritage documents and make them widely accessible for researchers and the general public. Cultural institutions are aware of this potential and often consider digitization of their cultural heritage collections as an obligation. Consequently, the quantity of digitized historical documents available is constantly growing. Transforming the printed cultural heritage collections into digital resources accessible and searchable through modern information and communication technologies requires that the digitized document images are transformed into digital text through Optical Character Recognition (OCR). While OCR can currently reach over 99 percent accuracy in recognition of characters from high quality images of original documents with a simple book layout, the accuracy for historical newspapers is lower than that. OCR quality is dependent on environment and the condition of the original

documents: print and paper quality, typefaces and layout complexity affect the accuracy of the result. Generally, the older the newspaper is, the lower the accuracy rate is likely to be. Holley (2008) reported raw character recognition accuracy rates varying from 71 percent to 98 percent in a sample of digitized newspapers from 1803-1954, the lowest rate indicating almost every third character being erroneously recognized and virtually all words containing errors. Even 98 percent accuracy rate results in an error in, on average, every sixth word in Finnish text (with an average word length of around 8 characters), if the errors are evenly distributed. Such error rates may lead to quadrupling the number of unique index words and notably increasing the size of the collection (cf. Taghva, Borsack & Condit, 1994).

Digitized cultural heritage collections are therefore often riddled with OCR errors that hamper the performance of information retrieval systems. Handling OCR errors is one of the two major problems for information retrieval from collections of historical documents. The second problem relates to historical change in languages: digitized texts are written in the language of the time of their origin. Natural languages continuously change reflecting the changes in the surrounding society. New words are added to the vocabulary and old words are given new meanings, or fall out of use. Even grammar and spelling change, often towards simpler forms, reflecting language internal (e.g. new sounds) and external (languages, dialects, language users) changes (Dahl, 2000). Standardized spelling is a modern invention, introduced in varying pace in different countries after the invention of printing and the introduction of dictionaries (Robertson & Willett, 1993). For example, English spelling became more or less fixed by the late 18th century (Robertson & Willett, 1993), while German spelling was not standardized until 1901/1902 (Ernst-Gerlach & Fuhr, 2007) and Finnish spelling was stabilized by the end of 1800s (Häkkinen, 1994). Before standardization, spelling was often based on the pronunciation of words leading to wide local variations reflecting the local dialects. Spelling could even be adjusted to help with the justification of lines. The rate of graphical variants can therefore be high in historical texts. (Rayson, Archer, Baron & Smith, 2007.)

Historical document retrieval (or "historic document retrieval", HDR) can be understood as a cross-language information retrieval (CLIR) problem (Koolen, Adriaans, Kamps & De Rijke, 2006): for retrieving documents written in a historical language given a modern query, either the queries or the documents need to be translated. Document translation aims at mapping all historical variants of a word into a single modern index term. This has the benefit of making the different linguistic tools developed for modern languages available in indexing. Document translation can therefore greatly reduce index size and improve retrieval results. Moreover, documents from different time periods, regions and sources can be treated differently, to adjust for the temporal and regional differences in spelling and differences in typography and layout in different publications. Most studies on HDR have however focused on query translation, i.e., generating query word variants at retrieval time (Robertson & Willet, 1993; O'Rourke, Robertson &

Willett, 1997; Braun, Wiesman & Sprinkhuizen-Kuyper, 2002; Ernst-Gerlach & Fuhr, 2007; Kempken, Luther & Pilz, 2006). Query translation has similar benefits in historical document retrieval as in CLIR. Control over the text digitization and indexing processes is not needed and the problems with existing collections can be handled post-hoc, without re-processing large document collections. Tools developed for modern languages can be used for generating morphological variants and splitting compounds in the modern queries. Retaining the original historical word forms is also desirable in some applications. On the other hand, the need to process the queries at query time reduces the capacity of the query processing component. For languages with rich inflectional morphology, such as Finnish or German, inflectional variation increases the number of variants that need to be processed at query time: in addition to the historical variants of query words, even the inflectional variants need to be generated at query time (Ernst-Gerlach and Fuhr, 2007).

Most HDR studies (Robertson & Willet, 1993; O'Rourke et al., 1997; Braun et al., 2002; Ernst-Gerlach & Fuhr, 2007; Koolen et al., 2006; Kempken et al., 2006) have focused on handling the graphical variants of modern words occurring in historical documents. The focus has been on recognizing correct pairs of modern and historical spelling variants, even if the need to handle OCR errors has been acknowledged. Words altered by changes in grammar and spelling (or OCR errors) are typically quite similar to their modern equivalents and can therefore often be reliably recognized using simple rule-based and string similarity approaches. Handling the changes in vocabulary requires more specific translation resources, such as dictionaries that are not readily available and are expensive to construct.

This article reports on a study on the use of fuzzy query expansion for handling the variation occurring in historical documents at query time. The study was motivated by the observed difficulties in accessing digitized cultural heritage documents due to the poor performance of the dedicated retrieval systems. We explored the use of approximate string matching in finding of historical, inflectional and OCR variants of modern query words from a collection of digitized Finnish newspapers from the 1800s. Our goal was to study the extent and kind of variation that occurs in digitized historical texts written on a morphologically complex language, and the effectiveness of approximate string matching in capturing that variation. We studied whether the approximate string variants could successfully be used as query expansion keys and the effect of the level of query expansion (the number of variants used for expanding the query words) on query performance. Furthermore, modern inflectional variants of the query words were generated to improve the coverage of morphological variants in the queries. The queries were then expanded with the inflectional variants, and with approximate matches of the inflectional variants found in the document collection. The study focused on string level variation, but our detailed analysis of search topics also revealed issues related to historical changes in vocabulary and concepts.

The rest of the paper is organized as follows. First, we present the background of the study including a description of inflectional morphology and compounding, as well as historical language change from the perspective of Finnish language, and a description of the methods we adopted for query expansion. The background is followed by a review of the related research. We then present our research questions and describe our data and methods. Finally, we report our findings and conclude with a discussion and conclusions.

Background

Finnish language in historical document retrieval

Finnish is a highly inflectional language, where suffixes are commonly used for inflection and derivation. The relations of words and expressions can be indicated in detail by morphological means and therefore the syntactic rules for, e.g., word order are rather loose. Due to the rich inflectional morphology, Finnish nouns can theoretically be inflected in over 2000 grammatical forms (Karlsson, 1983). Kettunen and Airio (2006) have showed that only a few of the theoretical noun forms are frequent enough in Finnish texts to be of practical importance in IR applications. Handling the morphological variation is however a major issue for Finnish information retrieval and necessary for good retrieval performance.

Finnish is a compounding language, where multi-word expressions are written together as closed compounds and new compound expressions can be readily generated by combining other words. Some two-thirds of the entries in the Dictionary of Modern Standard Finnish are compounds (Koskenniemi, 1983) and non-lexicalized or occasional compounds are frequent in Finnish texts. The word “kamarisuhteet” (*the relations between the houses of the parliament*) in our test data is a good example of a nifty occasional compound: it is a very efficient expression. The problem from information retrieval perspective is that the occasional compounds in documents and queries rarely match. It is not easy for the user to guess what compounds are used in relevant documents. The meanings of occasional compounds are often directly derived from the meanings of their constituents and therefore splitting compound words both in queries and indexes into their constituents is often beneficial in Finnish information retrieval (Alkula, 2001).

Finnish became established as a written language and went through some major reforms during the 19th century. The 1880s can be seen as a turning point during which the Finnish language more or less reached its current form (modern Finnish). Table 1 (based on the seminal work by Häkkinen (1994)) summarizes the most common differences between 19th century and contemporary Finnish. In the early 19th century Finnish, remains of the Old Finnish orthography can still be seen. Many alternative inflections and spellings exist concurrently due to the dialectal differences and the on-

going effort to develop and standardize Finnish. Different newspapers had different practices and there were clear geographical variations in the language. The orthographical differences include e.g. the use of the letter *w* instead of *v* (line 1, Table 1), and variations in marking long vowel sounds (line 2, Table 1). Common options for marking the long vowels included using one letter, one letter with a macron diacritic or two letters.

Line	Type of change	Description (modern – 19th century)	examples (modern – 19th century)
1	The character "v"	*v – w*	"vaimo" (*wife*) – "waimo"
2	Long vowel	VV – VV̄, V, V̄	"piispa" (*bishop*) – "piispa", "pispa" or "pīspa"
3	Marking the lost consonant	modern *Ø* – archaic *Ø*, apostrophe, *t*, *h*, *k*	"arvata" (*guess*) – "arwata' " "puhekieli" (*spoken language*) – "puhek'kieli" "sade" (*rain*) – "sadet"
4	Consonant gradation	*Ø – g*, or apostrophe	"pian" (*soon*) – "pi'an" "luvun" (*number's*) – "lugun", "luwun"
5	Hard and soft plosives	*p*, *t*, *k – b*, *d*, *g*	"hampaat" (*teeth*) – "hambat" "henki" (*life*) – "hengi"
6	Compound construction	spelled together – spelled together, hyphen, whitespace	"diakonissalaitos" (*Deaconess order*) – "diakonissalaitos", "diakonialaitos", "diakonissa-laitos", "diakonissa laitos"
7	Plural markers	*ei – i*; *loi*/*löi*	"tyttöjä" ((*some) girls*) – "tyttölöitä"
8	Plural genitive formation	*i+en – en*	"polkupyör-i-en" (*bicycles'*) – "polkupyörä-in"
9	Essive stem	vowel stem – vowel or consonant stem	"vuonna" (*In the year of*) – "vuotena"
10	Partitive suffixes	distribution of suffixes *-ta/tä and – a/ä* has changed	"omenaa" ((*an unspecified part of) apple*) – "omenata"
11	Illative suffixes	*-seen*/*siin* (and many others) – *- sen*/*sin* (and many others)	*vieraisiin –vieraisin*

Table 1. Differences between 19th century and contemporary Finnish. Capital "V" in describing the changes refers to a vowel, double "VV" to a double vowel. Based on Häkkinen (1994).

Some Finnish words that had previously ended in a consonant had lost the final consonant at some point of the development of the language. The final consonant sound has however not just disappeared, but can still be heard in the pronunciation of the words as the first consonant of the following word becomes geminated (lengthened). If the next word begins with a vowel, there is a double glottal stop between the words (e.g. *anna minulle* is pronounced [*annam minulle]* and *ota itse* [*ota' 'itse*][1]. Even word boundaries in compound words are geminated. While the missing consonant is not marked in contemporary Finnish text, it was still sometimes marked in the 19th century Finnish with an apostrophe ', *t, k* or *h* (line 3, Table 1).

There are some differences in consonant gradation between contemporary and 19th century Finnish (line 4, Table 1). It was common in the 19th century Finnish to use an apostrophe in the middle of words to mark the lost consonant in connection to the consonant gradation of *k*.[2] Also *g* was sometimes used as *k*'s weak variant between two *u*'s, even if the predominant spelling was already "*w*". The final syllable could also be spelled with a vowel and an apostrophe, as in *su'un, ky'yn*. Soft plosives were sometimes used instead of the hard plosives used in modern Finnish (line 5, Table 1).

The rules for compound construction were also less stable: compounds could be spelled together, with a hyphen or as a fixed phrase with the compound constituents spelled separately (line 6, Table 1).

Due to its central role in Finnish, the inflectional system changes slowly and only a few changes have occurred since the early 19th century. Most of the changes concern the appearance of the suffixes. One suffix could have different appearances in different dialects. Dialectal differences were quite visible in the 19th century written Finnish, and several types of variants could occur depending on the dialect used by the author. The ways how inflectional suffixes were added to stems also varied. Due to different kinds of reductions, weakening and disappearance of sounds, suffixes assimilated to stems or to each other. These changes sometimes led to situations where two separate word forms became too similar and other changes (not related to changes in pronunciation of the word forms) were made to maintain the differences between forms.

In the following, we give examples of historical changes in the most common case forms of nouns due to the central role of nouns in information retrieval (cf. Baeza-Yates & Ribeiro-Neto, 1999) and the fact that few common case forms cover in practice most of the inflectional variation (Kettunen & Airio, 2006). Especially during the early 19th century there was some variation in the use of plural markers (line 7, Table 1). The *ei*-diphthong has gradually replaced *i* as a plural marker since 1840s. The plural markers *loi*/*löi* typical for south-eastern and Savonian dialects were replaced by *oja/öjä* markers. There were also two mutually interchangeable ways of forming the plural genitive form (line 8, Table 1): the case suffix (originally *-ten/δen)* was either added to the plural stem formed with *i*-marker (the so called 1st genitive, e.g., **talo-i-en* > *talojen*) or directly to the singular stem (2nd genitive, e.g, **talo-en* > *taloen* > *taloin*). In the 19th century, the 2nd genitive was usually the primary choice. In modern Finnish, the 1st genitive has become a standard and the 2nd genitive is regarded as archaic. There was also variation in preferred essive stems, and in partitive and illative suffixes, inter alia (lines 9-11, Table 1).

S-gram matching

Recognizing and measuring similarity in strings is useful in information retrieval because words that have similar strings of characters often also have similar meanings (Robertson and Willet 1998). Different morphological and orthographical variants of a word represent the same concept and are therefore equal from the standpoint of users' requests (Pirkola, Keskustalo, Leppänen, Känsälä & Järvelin, 2002). Measures of string similarity can also capture cross-lingual orthographical variation in the spelling of related words in different languages. They have therefore been used in cross-language information retrieval, especially for the translation of out-of-vocabulary words such as proper

names and technical terms (Hedlund et al., 2004; Keskustalo, Pirkola, Visala, Leppänen & Järvelin, 2003). There are many different techniques for measuring the similarity between strings. For an introduction, see e.g. Navarro (2001).

S-gram matching is an approximate string matching technique that is based on the well-known *n*-gram matching technique. The strings that are compared are split into substrings (*n*-grams) of length *n*, and the proximity of the strings is then defined as the share of the strings' overlapping substrings of all of their unique substrings. (See e.g. McNamee and Mayfield (2004) for an overview of *n*-grams.) While *n*-grams are formed of adjacent characters, skipping some characters is allowed when forming *s*-grams (the "*s"* refers to skipping). The suitable lengths of the substrings and the skips vary depending on the application area. In natural language information retrieval applications the substring length *n* tends to vary between 2-6 characters depending on the application (Pirkola et al., 2002; Keskustalo et al., 2003; McNamee, Nicholas & Mayfield, 2009; Ullman, 1977). In query translation and query expansion applications *di*-grams (*n*=2) have been found to perform well, when using word internal sets of *n*-grams to identify spelling variants of search keys from among index words (Pirkola et al., 2002, Keskustalo et al., 2003). Skip lengths of 0-2 have been found to model cross- language spelling variation well (Keskustalo et al., 2003).

S-grams formed with different skip lengths can be combined into *s*-gram classes to better model the character changes typically occurring in natural language spelling variants. *Character Combination Index* (CCI) then indicates the set of all the *s*-gram classes to be formed from a string. For example, $CCI_{\{\{0\},\{1,2\}\}}$, denotes that two *s*-gram classes are formed from a string: the gram class {0} with conventional *n*-grams formed of adjacent characters (no skipping) and the gram class {1,2} with *s*-grams formed by skipping both one and two characters. Examples of the *s*-grams formed in the different gram classes are given in Table 2 for the string "Pariisi" (*Paris*). The proximity between strings is again measured based on the numbers of their overlapping and unique *s*-grams, but only the *s*-grams belonging to the same s-gram class are compared to each other. For detailed definitions of proximity measures based on *s*-gram profiles, see Järvelin, Järvelin and Järvelin (2007).

Using extra "padding-characters" at the beginnings and the ends of the compared strings is common in *n*-gram and *s*-gram matching. The padding helps getting the beginnings and the ends of the strings properly presented in their *s*-gram sets, and thus gives more weight to them in matching. For *n*-grams a padding of *n*-1 characters has been common (Robertson & Willett, 1998). For *s*-grams the padding depends on the length of the substring (*n*) and the number of the skipped characters (*k*). Padding is typically set as (*n*-1)(*k*+1) (Järvelin et al. 2007). Padding can also be used only at one side of the strings to give more weight to that side. Table 2 gives examples of *s*-grams where only left padding is used, to give more weight to string beginnings.

Gram class	{0}	{1}	{2}	{0,1}	{1,2}
s-grams	pa ar ri ii is si	pr ai ri is ii	pi ai rs ii	pa ar ri ii is si pr ai	pr ai ri is ii pi rs
s-grams with padding	_p pa ar ri ii is si	_p _a pr ai ri is ii	_p _a _r pi ai rs ii	_p _a pa ar ri ii is si pr ai s	_p _a _r pr ai ri is ii pi rs

Table 2. Examples of *s*-grams generated with and without padding for the word "pariisi" (*Paris*) using gram length *n*=2 and gram classes with varying skip length (*k*=0, 1, 2, or their combination).

Frequent case generation

The information retrieval performance of morphologically complex languages is usually notably enhanced by morphological processing. Usually reductive techniques, such as lemmatization or stemming, are used for mapping all morphological forms of a word in a document collection into a single index term. Kettunen and Airio (2006) and Kettunen (2009) utilized the skewed distributions of Finnish word forms in text corpora and developed a generative method, the Frequent Case Generation (FCG), for handling morphological variation at query time. The goal is to generate only the subset of the most significant inflected forms (from the standpoint of IR) for the input query words. No prior reductive processing of the database indexes is then needed, which is practical in large and frequently updated databases or where the reductive techniques cannot be utilized, such as digitized historical collections containing many OCR errors and historical variants.

For determining which inflected forms should be generated for the query keywords, the frequency distributions of case forms are studied in a representative text corpus. After the most frequent (case) forms for the language have been identified with corpus statistics, the suitable combination of case forms is established experimentally based on information retrieval experiments. The number of frequent case form combinations to test depends on the morphological complexity of the language. After the evaluation, the best FCG process with respect to lemmatization, stemming or other word form variation management methods is usually identified. The FCG method has been used so far successfully both in mono- and cross-lingual IR of Finnish, German, Swedish and English (Kettunen, Airio & Järvelin, 2007; Airio & Kettunen, 2009). Recently three Indian languages, Bengali, Gujarati and Marathi were evaluated for the application of the FCG method (Paik, Kettunen, Pal & Järvelin, 2013).

Related research

The main problems generally associated with historical document retrieval have to do with the (i) changes that occur in languages over the centuries and (ii) with the often poor OCR quality of the digitized collections. The effect of OCR errors in information retrieval has been studied since early 1990 and has been surveyed by e.g. Mitra and Chaudhuri (2000) and Beitzel, Jensen and Grossman (2003). Generally, the OCR accuracy from high quality source documents is

high enough to not seriously affect information retrieval performance (e.g. Mitra & Chaudhuri, 2000; Taghva et al., 1994). However, higher corruption levels lead to unstable retrieval performance and interfere both with term weight calculations and with matching search keys to index terms (Beizel et al., 2003; Mittendorf & Schäuble, 1996). Short documents are particularly sensitive to OCR errors, because there is not enough repetition of the central words to recover from the errors.

The two main approaches to handling OCR errors in information retrieval have been either handling the errors during indexing through OCR error correction or *n*-gram indexing (Liu, Babad, Sun & Chan, 1991; Taghva et al., 1994; Harding, Croft & Weir, 1997; Amati, Celi, Di Nicola, Flammini & Pavone, 2011; Savoy & Naji 2011), or handling them at query time using approximate string matching for query expansion (Harding et al., 1997; Amati et al., 2011). The previous has often performed better of the two (e.g. Harding et al., 1997). OCR error correction has the additional advantage of making the standard tokenization and conflations approaches viable. However, handling the errors at query time has the advantage that the processing of the database index entries can be avoided.

N-grams have been one of the most frequently and successfully used solutions to retrieval of text corrupted by OCR errors. For example Harding et al. (1997) and Amati et al. (2011) used a combination of full words and their *n*-gram representations as indexing features; Savoy and Naji (2011) used 4-gram indexing only, while Liu et al. (1991) used frequency distributions of *di*-grams in OCR error recognition and correction. Harding et al. (1997) compared *n*-gram indexing to *n*-gram based query expansion. They found that *n*-gram indexing performed better than *n*-gram query expansion, but also required much more storage space and computational resources. The *n*-gram query expansion results improved when the number of *n*-gram variants added to the queries was increased, suggesting that using a looser *n*-gram distance threshold and thus adding more words to the expanded queries might further improve the approach. (Harding et al., 1997.)

Studies on historical document retrieval have generally focused on the differences between historical and modern languages. OCR errors have been omitted from the experimental settings by using manually created or manually corrected test data (e.g. Braun et al., 2002; Gotscharek, Reffle, Ringsletter, Schulz & Neumann (2011); Hauser, Heller, Leiss, Schulz & Wanzeck, 2007; Kempken et al., 2006, Koolen et al., 2006; O'Rourke et al., 1997). An exception is Pilz, Luther, Fuhr and Ammon (2006), who created rules for handling OCR errors both manually and automatically based on edit costs between character replacements.

Historical document retrieval suffers from several problems that need to be solved for high retrieval performance. Braun et al. (2002) identified the general problems to be (i) historical changes in vocabulary, (ii) historical changes in spelling

and (iii) historical inconsistencies in spelling. Hauser et al. (2007) added a more detailed analysis of the sources of variation in German early prints and noted e.g. historical variation in spelling of compound words and replacements of sub-words as specific problems. It is generally assumed that for effective handling of both the vocabulary change and the spelling variation, a combination of dictionary-based and fuzzy translation approaches is necessary. Reaching a full coverage of the historical variation in a dictionary is difficult (Pilz et al., 2006) and while approximate string matching and rule-based translation approaches may reach a better coverage of variation they cannot handle vocabulary change (Braun et al., 2002). Hauser et al. (2007) suggested a complex translation tool combining a manually constructed dictionary, rule based variant generation and an approximate string matching approach based on character sequence edit distance. Gotscharek et al. (2011) described a corpus-based approach to efficient construction of historical lexica with focus on reducing the manual workload of lexicon construction. However, most studies have focused on the string level variation, ignoring the more complex issues related to conceptual, vocabulary and syntactic change.

Approximate string matching techniques (and *n*-grams in particular) have been used in many studies concerning historical document retrieval. For example, Robertson and Willet (1993) studying 17th century English, O'Rourke et al. (1997) studying 12th century French, and Braun et al. (2002) studying 16th and 17th century Dutch approached historical document retrieval as a spelling correction problem and tested a variety of approximate string matching techniques for identifying historical variants of modern query keys occurring in historical texts. Kempken et al. (2006) used an edit distance variant where the edit costs were automatically learned from the German historical document collection. They concluded that algorithms that are adapted to the specific historical phenomena of the collection can reach a better translation recall and precision than standard edit distance and *n*-grams (Kempken et al., 2006).

Koolen et al. (2006) constructed data-driven rules for modernizing historical Dutch documents based on a phonetic similarity measure, relative frequencies of consonant and vowel sequences in a historical and a modern corpus, and relative *n*-gram frequencies. Gotscharek et al. (2011) manually constructed a list of 140 rewrite patterns for modifying historical German spellings into words occurring in a modern German lexicon. They estimated that for 19th and 18th century German the rewrite patterns alone led to acceptable precision and recall of matching historical variants to their modern variants. For processing older German texts, they suggested that a combination of the rewrite patterns and a historical lexicon was needed (Gotscharek et al., 2011).

Ernst-Gerlach and Fuhr (2007) constructed rewrite rules for generating historical variants of modern query words. The rules were generated from a manually collected list of pairs of historical spellings and their modern variants. For each rule, the context where the rule was applicable and the reliability of the rule were recorded. Query words were first

expanded with their modern inflectional and derivational forms and historical variants were then generated for all of these word forms. This approach outperformed different combinations of stemming, edit distance and word form generation showing that handling both inflection and historical variation is important for highly inflectional languages. Pilz, Ernst-Gerlach, Kempken, Rayson and Archer (2008) found that automatic approaches to historical variant generation can reproduce manually generated gold standard rules quite well and may also capture variation that is not discovered manually. They argued for generic letter-replacement heuristics for Germanic languages and showed that an edit distance variant where the edit costs were learned from German historical corpus outperformed the standard edit distance of English historical data.

S-gram matching has not been tested previously in historical document retrieval. However, it has previously been successfully used in translation of out-of-vocabulary words and cross-lingual spelling variants in CLIR and it has outperformed many of the string similarity measures previously suggested in historical document retrieval studies (e.g. *n*-grams, LCS, edit distance) in cross-language query translation tasks (Pirkola et al., 2002; Keskustalo et al., 2003).

Research design

Research questions

We formulated the following three research questions:

RQ1 What kind of variation does the *s*-gram matching capture? Does the captured variation reflect the variation actually occurring in the historical collection?

RQ2 Is query expansion using fuzzy query word variants useful in historical document retrieval?

RQ3 How do the properties of topics and query words affect the performance of the fuzzy query expansion in historical collections?

RQ1 aims to create an understanding of the kind of variation that occurs in the collection, and how the properties of the query words and the selected *s*-gram matching approach affect the types of variants that are generated. Our goal is to measure the level of *s*-gram translation precision and identify potential biases or poor coverage of variation types in the translation results. RQ2 aims to reveal the most promising approach to fuzzy query expansion in historical document retrieval. The main focus is on the questions of coverage and precision. What is the suitable level of query expansion, and what is the role of high precision handling of one variation type (inflectional variation) as compared to a noisier approach with a better coverage of all variation types? RQ3 focuses on analyzing how query words and their

combinations affect retrieval performance. It requires that we identify the characteristics of topics that typically lead to good and poor retrieval performance.

Historical newspaper test collection

Our test collection consists of a subset of the historical newspaper archive of the National Library of Finland[3] and contains digitized Finnish newspapers published between 1829 and 1890. The collection contains 180 468 documents (84512 newspaper pages, 772 MB) (Raitanen, 2012). The original document images are available through the Web service of the National Library. The OCR quality in the collection is varied which makes the usual approaches to conflation of morphological variants into a common form infeasible. A stemming test conducted showed that the Snowball stemmer for Finnish reduced the number of unique index words in the collection from 7.03 million to 4.87 million (to 69.3 % of the original). The corresponding conflation-rate for Snowball stemmer on the OCR error-free Kotus corpus[4] on 19th century literary Finnish was clearly higher, 52.2 %. These rates can be compared to how Snowball performs on modern Finnish newspaper text, where the number of unique words forms can be reduced to 49.9 % of the original unprocessed word forms (Kettunen & Baskaya, 2011). This suggests that it is mainly the OCR errors that reduce the stemming performance in the historical collection, while historical variation has a smaller effect. On error-free 19th century Finnish data, stemmers developed for modern Finnish might perform well. However, due to the OCR noise, stemming was not used in this study during indexing.

Raitanen (2012) estimated based on a small text sample of 2105 words that roughly one fifth of the words occurring in the historical newspaper collection differ from contemporary Finnish and another fifth was incorrectly OCR scanned. This is quite well in line with the 95-97 % character recognition accuracy rates previously reported by Bremer-Laamanen (2001), which would correspond to errors occurring approximately in every fourth word if the OCR errors were evenly distributed. The most common historical change occurring in the text sample was using “w” instead of “v” used in contemporary Finnish. The sample contained 127 different OCR errors, the replacement of “w” with “m” being the most common one. (Raitanen, 2012.)

The test collection contains 56 search topics related to 19th century history, together with relevance assessments. The topics are written from the contemporary perspective using contemporary language. The historical background and alternative vocabulary may be introduced in the topic narratives to support the relevance assessments. The topics model relatively broad, topical information needs and typically require collecting and synthesizing information from several documents. They are typically considered with historical events, institutions or persons, or with the attitudes or wider developments in the society during the 19th century. English translation of a sample topic is shown in Figure 1.

Topic number: 1

Title: Australian aboriginals

Narrative: Australia was a British colony during the 1800s. Before the arrival of Europeans, Australia was inhabited by aboriginal peoples. What was written about the Australian aboriginals in the Finnish newspapers of the 19th century?

Relevant documents discuss the languages, looks, habits or culture of the aboriginal peoples. Documents only discussing the colonization of Australia without discussing the original population are not relevant.

Figure 1. An example topic (English translation).

Creating suitable topics was an iterative process. We made several test searches to ensure that a topic had enough but not too many (less than 500) documents in the assessment pools. The final assessment pools for each topic were created by searching with one complex query that covered extensively various letter, word and term variants as well as all the facets of the topic at hand. Each document in this set was then assessed intellectually by the relevance assessors. The relevance assessments were made on a four point scale: non relevant (level 0), marginally relevant (level 1), relevant (level 2) and highly relevant (level 3). The relevance levels were similar to the definitions in Sormunen (2002), but the exact criteria for the content depended on the topic. The news articles in the collection are typically very short. Therefore, the information needs described in the topics could rarely be answered by a single relevant document. Documents on the relevance level 0 contain no information about the topic or otherwise do not meet the requirements of the topic description. Documents on level 1 typically contain a single, information-poor reference to the topic. Documents on level 2 contain a few facts about the topic, or information about only some aspects of a multifaceted topic. Documents on level 3 discuss the topic more comprehensively meeting the requirements of the topic description. In this study, only the documents assessed to be relevant on the levels 2 and 3 were included in the recall-base. Marginally relevant documents were regarded as non-relevant. The average recall-base size for a topic was 52 relevant documents. On average, there were 36 documents on recall level 2 and 16 on recall level 3. Topic 48 (*Children's smoking*) had the smallest recall-base with 5 relevant documents, and topic 12 (*The construction project of The House of the Estates*) the largest with 253 relevant documents. The creation of the recall-bases for the topics was a challenging task in presence of the OCR errors, historical spelling variants and changes in the use of concepts and vocabulary. During the present study, new relevant documents were identified for two topics. These were added to the relevance corpus for future use, but not accounted for in this study. We used the title fields of 50 topics which produced altogether 100 unique search words after the removal of the stop words. The remaining six topics out of 56 were used for pre-testing *s*-gram settings and query structures, and thus were left out here.

Variant generation and query expansion

This study is concerned with a novel text domain, i.e., historical Finnish containing OCR errors. Because there are no established baseline approaches for handling word form variation occurring in the texts in this domain, we used a simple unprocessed baseline in which the queries words were left as-is. In other words, we only included the modern query words in the forms they occur in the topic titles. This simple baseline facilitates observing the effects of various query expansion approaches.

Query expansion has generally been found to improve performance on informational search tasks (e.g., Buckley, Salton, Allan & Singhal, 1995), considered also in this study. Therefore, another baseline based on using a simple query expansion approach could also be motivated in order to examine the effectiveness of *s*-grams and FCG for generation of useful terms for query expansion. We used the pseudo-relevance feedback (PRF) feature of the Lemur Indri search engine[5] as a simple query expansion baseline approach. We experimented with several pseudo relevance feedback settings: 2-20 top-ranked documents as sources for new terms, adding 10-50 terms to the expanded queries, and varying the weight given to the original respectively the expanded query (between 0.3-0.7 weight given to the original query). We saw modest performance improvements, as compared to the unprocessed baseline but none of the differences were statistically significant. Therefore, we only report the unprocessed baseline results.

The *s*-gram settings were chosen based on previous studies on Finnish monolingual and cross-language information retrieval. The gram length was set to two (n=2) and skip lengths of 0-2 characters were used, because they have been found useful for finding cross-lingual spelling variants for e.g., proper names and technical terms (Keskustalo et al., 2003); and for translation between closely related languages (Järvelin, Kumpulainen, Pirkola & Sormunen, 2006). Moreover, the inflectional, historical and OCR variants present in a Finnish 19th century text corpus follow similar patterns of single character substitutions, insertions and deletions, and their two character combinations. Three different CCI's combining the gram classes {0}, {1}, {0,1} and {1,2} were tested, namely, CCI1={{0},{0,1},{1,2}}, CCI2={{0},{1},{1,2}}, and CCI3={{0},{0,1},{1},{1,2}}. We used left padding only when forming the *s*-grams to give more weight to word beginnings: Finnish is a suffixing language, where mainly the word endings change in inflected forms. The beginnings of words are more stable and thus more important for determining the string similarity. Strings shorter than three characters and numerals were not expanded using the *s*-gram matching, because the *s*-gram variants produced for very short strings and numerals tend to be very noisy. In our test data only one string, "ii" (*the second* - as in *Tsar Alexander the Second*), matched this length criterion and it was added to the query as is. Finally, we used the Dice coefficient for calculating the string similarity (Järvelin & Järvelin, 2008).

In CLIR query translation applications using only a few (2-3) *n*-gram or *s*-gram variants for a search key has been found useful (Hedlund et al., 2004; Järvelin et al., 2006). Given the noisiness of digitized historical document collections, we expected that a wider coverage of the variants may be necessary. In order to find a balance between the coverage of the query expansion and avoiding adding an excess of noise, we experimented with nine different query expansion levels, namely 2, 5, 10, 15, 20, 30, 40, 50 and 60 *s*-gram variants generated for each query word and each of the three CCI's. All query words were expanded by the same number of *s*-gram variants. For the sake of the clarity of presentation, the results are only reported for four of these query expansion levels: 5, 20, 30 and 50.

The noisiness of the historical newspaper collection makes acquiring reliable corpus statistics over the case form distributions difficult. Therefore, the case forms used in query expansion by Frequent Case Generation were based on case form statistics of contemporary Finnish as shown in Kettunen and Airio (2006). The most frequent cases for contemporary Finnish are nominative, genitive and partitive, followed by the three inner locative cases, inessive, elative and illative. Some differences in the case distributions of contemporary and 19th century Finnish might occur, and it is difficult to foresee how historical variation and OCR errors affect the application of the method. We expect however the differences in the case distribution and inflectional system to be small due to the central role of inflection in Finnish, and not to have a considerable effect on the case form distributions.

Three query expansion levels were tested for FCG. First, in FCG6 approach queries the singular and plural forms of the three most common cases in modern Finnish, i.e., nominative, genitive and partitive were generated for each modern query word (altogether 6 forms). Secondly, in FCG12 queries, the singular and plural forms of the inner locatives inessive, elative and illative were generated in addition to the cases included in FCG6 (12 forms). Last, in FCG22 queries also the singular and plural forms of allative, adessive, ablative, essive and transitive cases were included in addition to the cases included in FCG12 (22 forms).

Frequent case generation is a nearly noise-free approach to handling morphological variation in inflected word-based indexes. It does not however identify the historical and OCR variants which are a major problem when searching from digitized historical text collections. *S*-grams can handle all kind of string level variation, but are prone to introduce noise into the queries. Therefore, we combined FCG and *s*-grams in order to ensure a good coverage of inflectional variation and reduce the amount of noise involved in *s*-gramming while still covering the different types of variation. We generated *s*-gram variants for each case form produced by the FCG6 and the FCG12 approaches, using both CCI1 and CCI2. Several expansion levels were tested for the combined FCG+*s*-gram queries. We generated 5, 10, 15, 20, 30, 40 and 50 *s*-gram variants for each case form generated by FCG6 and FCG12, resulting in up to 300 and 600 variants

for each search key, respectively. However, the different case forms of a word are usually similar to each other and produce many overlapping *s*-gram variants. We only added each variant once into a query, and therefore, the actual number of variants used for expanding the search keys is generally much lower than the maximum number of variants. For the FCG6+*s*-gram queries, around 45 % of the generated variants were unique. When 30 CCI2 *s*-gram variants were generated for each of the six the FCG6 case forms, in total 180 variants were generated for each query word. Out of them, on average only 81 were unique, with a minimum of 38 and maximum of 163 unique variants for any query word. When 50 *s*-gram variants were generated, on average 129 (of 300 variants) were unique, with a minimum of 57 and maximum of 260 unique variants. Some queries however became very long, e.g., the query "ida aalbergin ura ulkomailla" (*Ida Aalberg's career abroad*) was expanded with nearly 1300 words when using FCG12 and generating 50 *s*-gram variants for each query word and case form. For the combined FCG-*s*-gram method, the results for only the best two combinations for FCG6 and FCG12 are reported: using CCI2 and generating 20 variants for each case form (CCI2_20); and using CCI2 and generating 30 variants for each case form (CCI2_30). Adding more *s*-gram variants to the queries causes the query performance to deteriorate. The results for the marginally weaker CCI1 are omitted.

The variants of a single query word were combined in the queries by using the synonym operator of Indri. This way a common weight was computed for the variants and the queries were balanced against query drift due to expanding one/some of the query words with high frequency variant candidates. This so called "Pirkola's method" (Pirkola, 1998) is known to improve the performance of translated queries in cross-language information retrieval.

Variant categorization

For each of the 100 title words from the topic descriptions (i.e. search keys), twenty most similar *s*-gram variant candidates (database index entries) were intellectually analyzed for their semantic similarity with the search keys and for the kind of surface level variation that occurred in the strings. The occurrences of variant candidates were analyzed both in the PDF page images of the original documents and in the OCR generated texts, because the referent of a variant candidate was not always clear in the context given by the OCR scanned text only. Analyzing the variant candidates was manual work and for practical reasons it was not possible to analyze all occurrences of all variant candidates in the test collection. Therefore, we limited the analysis to the context of the 10 highest ranked documents retrieved with each variant candidate. Most variant candidates occurred in fewer than ten documents in the collection. The number of documents to analyze intellectually was 8 431.

A categorization scheme was developed for systematically recording the variation occurring in the variant candidates. The scheme divides the variant candidates into three main categories: (1) exact matches and surface level variants of the

query keys, (2) words which were semantically related to the query words, and (3) noise. The surface level variants in category 1 contain both historical and erroneous spelling variants, inflected word forms and OCR variants of the query words as well as different combinations of these variations. The semantically related strings in category 2 are not variants of the query word, but refer to different lemmas. However, their referents are semantically related to the query word, for example synonyms, cross-lingual spelling variants, derivatives, broader or narrower terms, or parts of compound query words. In many of the category 2 strings even surface level (spelling, inflection or OCR) variation occurred. These words were categorized as related words with multiple types of variation. Expanding queries exhaustively with words semantically related to the query words has been found useful in studies focusing on concept-based query expansion (Kekäläinen & Järvelin, 2000; Sormunen, 2000). Therefore, variant candidates in the main category 2 are potentially useful query expansion terms. Category 3 contains the variant candidates that are not semantically related to the query words, and the variant candidates for which the referent could not be identified (i.e. for which the intended original meaning was unclear). The categorization scheme is depicted in Table 3, with examples of each category given for a modern query word "vaimo" (*wife*).

Category	Explanation	Example variants
1	**VARIANT**	
1a	Exact match – the search key	"vaimo" (*wife*; married woman)
1b	Spelling variant	
1b-i	Historical/alternative	"waimo" (*wife*, historical)
1b-ii	Abbreviation	"w." (abbreviation, in compounds, e.g. "työmiehenw." (*worker's w.*)
1b-iii	Spelling/typesetting error	"waimonmon" (accidental repetition of the final syllable)
1c	Inflectional variant	"vaimon" (*wife's*, inflection);
1d	OCR variant	"vaiino" (*wife*, OCR error *m--in*)
1e	MULTIPLE: surface level only	"waimoin" (wives', historical inflection "in" and spelling "w"
2	**RELATED**	
2a	Synonym	"aviovaimo" (*wife*, synonym);
2b	Cross-lingual (foreign words)	Foreign translations of "vaimo" are not similar enough to the query word to be considered as variant candidates
2c	Derivative	"vaimoke" (*wifelet*)
2d	Related meaning	"leskivaimo" (*widow*, narrower term/compound);
2e	MULTIPLE: related	"leskiwaimo" (*widow*, narrower term + historical spelling)
3	**NOISE**	
3a	Noise	"vaim" (*only*: "vain" + OCR error); "vaimo" (*troubles,* coincidental overlap: "vaiwoja" + OCR errors *w--m* & insertion of a space → "vaimo ja")
3b	Unclear	-

Table 3. The categorization developed for analyzing variant candidates. Examples are given for possible analyses of variant candidates, given the query word "vaimo" (*wife*).

Evaluation

The *s*-gram translation quality was measured as precision among the 20 most similar variant candidates produced for each query word, i.e. as the share of relevant variants of all the variant candidates generated. *Relevant* was here defined

as *related referents*: all variant candidates assigned in the categories 1 and 2 were considered relevant. The effects of the query word length and the frequencies on the *s*-gram variant candidates generated for the query word were analyzed. Pearson's chi-squared was used for statistical testing.

We adopted two evaluation scenarios for the retrieval experiment to model two types of potential system usage. For *recall oriented search* (Scenario 1), we measured the normalized discounted cumulated gain (nDCG) with cut-off points at result list ranks 10 and 50 (nDCG@10, nDCG@50) and mean average precision (MAP). nDCG@10 is motivated as the quality of the top results is important even in recall oriented search. Users are unlikely to continue browsing through results, if the performance at the top ranks is not satisfactory. We used the logarithm base 10 for a lenient discounting of the cumulated gain in nDCG to model patient users, as suggested by Järvelin and Kekäläinen (2002). For *high precision oriented search* (Scenario 2), we measured nDCG at rank 10 (nDCG@10), with logarithm base 2 used for discounting the cumulated gain, to model impatient users who are likely to examine only few of the top results (Järvelin & Kekäläinen, 2002) and average precision at rank 10 (P@10).

MAP and P@10 were calculated based on binary relevance using trec_eval 9.0. Marginally relevant documents (relevance level 1) in the relevance corpus were considered non-relevant, following Scholer and Turpin (2009), who found that marginally relevant documents were not strongly differentiable from non-relevant documents, from the point of view of users' search effectiveness. Marginally relevant documents are hardly useful for the users of the historical newspaper library who often have topical information needs. Relevant and highly relevant documents (relevance levels 2 and 3) were counted as relevant. The nDCG@K values were calculated using Vectora[6] because it, unlike trec_eval 9.0, allowed changing the logarithm base used in discounting the document scores. Here, documents at relevance levels 0 and 1 were given a zero score, documents at relevance level 2 were given the score 10 and documents at the relevance level 3 were given the score 100.

The statistical significance of the results was tested using the Friedman's test and Pearson's chi-squared test. In addition to the standard effectiveness measures, we analyzed the performance of the *s*-gram query expansion on individual topic level to better understand how the differences between topics, query keys and combinations of query keys affect the usefulness of the approach. The results for *s*-gram query expansion using CCI1 and 30 expansion terms for each search key were analyzed, based on the P@10 results of both the baseline and the expanded *s*-gram queries. Coming up with good query terms is often a challenge for the users engaging in historical retrieval. Even retrieving just one relevant document at the top ranks, instead of retrieving nothing, can then be a very important from user perspective because it can help the user in selecting query words. Therefore, we consider the topics with low baseline performance in top

ranks as particularly important in the analysis, especially if no relevant documents were retrieved. We computed the Pearson's chi-squared test to control whether the differences observed between different topic categories were statistically significant.

Findings

Variant categorization and *s*-gram translation precision

The average translation precision for the query words among the top-20 *s*-gram variant candidates was 78. In other words, 78 % of the analyzed variant candidates were assigned into one of the two categories of (potentially) relevant query words: 62 % were surface form variants or exact matches of the query words and were categorized into the category 1. 16 % were semantically related to the query words and categorized into the category 2. The remaining 22 % of the variant candidates were semantically unrelated to the query words and categorized as noise in category 3. For most query words, variant candidates from several different sub-categories were found in the top-20. On average two out of the three main categories and five different subcategories were used in the analysis of the top-20 variant candidates of a single query word. 22 query words had variants from one main category only: 19 with only category 1 variants, and 3 with only category 2 variants. The distribution of the variant candidates into the sub-categories is illustrated in Table 4.

Category	Explanation	No. occurrences	Percentages
1	**VARIANT**	**1239**	**62 %**
1a	Exact match – the search key	82	4.1 %
1b	Spelling variant	77	3.9 %
1c	Inflectional variant	372	18.6 %
1d	OCR variant	191	9.6 %
1e	MULTIPLE: surface level only	517	25.9 %
2	**RELATED**	**328**	**16 %**
2a	Synonym	5	0.3 %
2b	Cross-lingual (foreign words)	8	0.4 %
2c	Derivative	50	2.5 %
2d	Related meaning	28	1.4 %
2e	MULTIPLE: related	*237*	*11.9 %*
3	**NOISE**	**433**	**22 %**
3a	Noise	425	21.3 %
3b	Unclear	8	0.4 %

Table 4. Translation precision and the distribution of the *s*-gram variants into the different categories.

Most of the modern Finnish query words had exact matches in the collection: 82 variants were categorized as exact matches in category 1. Six more query words had an exact match in the collection, but they were results of OCR scanning errors of historical variants and were categorized into category 1e. Only 36 of the surface form variants were

historical variants. Another 39 variants were spelling errors, 372 variants were inflectional forms of the modern query word and 191 variants were created through OCR errors. This does not necessarily reflect the variation occurring in the collection. In fact in categories 1e and 2e (multiple types of variation) historical variation was more common: 42 % (316 out of 754) of the variants containing multiple types of variation contained historical variation. Of these, 216 even contained OCR errors and were often cases where the OCR errors rendered the historical variants closer to the modern query word. This was particularly often observed in variants of query words containing the letter "v", because historical variation and OCR errors tend to have opposite effects on the letter: modern words containing "v" were often historically spelled with "w"; "w" is often erroneously OCR scanned as "v" in this collection. For example, the erroneously scanned variants of the historical "tulowero" (*income tax*), "tulovero" (corresponding the modern spelling) or "tulolvero" (non-word) are both more similar to the modern query word ("tulovero") than the correct historical spelling. The use of left padding benefits inflectional variants as compared to historical or OCR variants, because inflectional changes occur almost only in the down-weighted word endings, while historical and OCR changes may occur in any part of words. In Table 5 examples of five query words containing the letter "v" are shown. The relative frequencies of the modern and historical spelling variants vary. For some words several historical spelling variants co-exist, while for others historical variation only occurs in inflectional endings; some OCR errors are so common that the erroneously scanned words may have higher collection frequencies than the correct historical (or modern) variants. The modern variants of "kanava" and "vaimo" are the same as their most frequent OCR variants. In both cases, the "w" in the historical variants was often erroneously scanned as "v" and the OCR and modern variants thus coinciding. The other OCR variants are non-words.

Modern variant	Frequency	Historical variant	Frequency	OCR variant	Frequency
"työväenyhdistys" (*workers' union*)	64	"työwäenyhdistys"	86	"työmäenyhdistys"	**210**
"kansanvalistusseuran" (*lifelong learning foundation's*)	155	"kansanwalistusseuran"	329	"kansanmalistusseuran"	**634**
"vapaaehtoiset" (*volunteer*)	**159**	"wapaaehtoiset"	23	"vapaehtoiset" "rvapaaehtoiset"	50 1
"kanava" (*canal*)	43	"kanawa"	**455**	"kanava" (w→v) "kanaiva"	43 14
"vaimo" (*wife*)	2054	"waimo"	**3376**	"vaimo" (w→v) "vvaimo"	2054 1

Table 5. Examples of how the different types of surface form variants may have different frequency distributions depending on the word. The modern variant is the exact match of the query word in the document collection.

Word length

To analyze the relation of string length to the string variation categories, the 100 search keys were divided into three groups based on their length: short, medium and long words, as follows:

- **Short words:** 2-6 characters (27 words)
- **Medium-long words:** 7-11 characters (44words)
- **Long words:** 12-20 characters (29 words)

The groups were loosely based on the normal Finnish word length distribution, where the average word length is around 8 characters (Niemikorpi 1991), while ensuring that each group had enough occurrences for statistical analysis. Table 6 shows a cross-tabulation of the word length groups and variant occurrences in the different string variation categories. The length of a query word has a clear effect on the types of variant candidates generated for it (in top-20) and on the precision of the *s*-gram translation. Pearson's chi squared test confirms this result: the distribution of the variant candidates into the variant categories is not independent of word length (p=0.000).

WORD LENGTH	**VARIANT CATEGORIES**				**TOTAL**
	1a Exact match	**1b-1e Surface forms**	**2 Related terms**	**3 Noise**	
Short words (n=540)	4.4 %	36.9 %	11.3 %	47.4 %	100 %
Medium words (n=880)	4.8 %	72.3 %	8.5 %	14.4 %	100 %
Long words (n=580)	2.8 %	55.5 %	33.1 %	8.6 %	100 %
TOTALS (n=2000)	4.1 %	57.9 %	16.4 %	21.7 %	100 %

Table 6. Word length and string variation categories. Pearson's chi-squared $X^2(6) = 449.7$, $p = 0.000$.

The general trend is that as the length of the query words increase, the share of noise (category 3 variants) decreases to give way for more category 1b-1e surface form variants at first and, as the query words get longer, even for category 2 (related) variants. Particularly the query words with a length of four or less characters generate mainly noise. This seems reasonable given that a change of a single character can radically alter the meaning of a four character word, while long strings rarely resemble semantically unrelated strings as closely. On the other hand, related variants (category 2) are much more common for long words than expected based on the null hypothesis of independence of variant categories of the query word length. Long query words are often (occasional) compounds with rather specific meanings and do not match the expressions used in the document collection; the most common and central words in all languages tend to be relatively short. Due to their low frequency long words tend to have fewer surface form variants, and consequently many derivatives, compound parts and related words are found among the top 20 matches. Though not directly visible from Table 6, the share of variant candidates where multiple types of variation occurs (categories 1e and 2e) increases with query word length. Clearly more category 1e and 2e variants are observed among the long words than expected: 312 variants as compared to the 220 expected based on the null hypothesis. Naturally, there are differences between the variant candidates generated for query words of the same length: e.g. most long query words have virtually no noise among their variant candidates, while "huumausaineiden" (*narcotic substances'*) only has 2 relevant variants among the 20 variant candidates, because the exact expression is not used in the collection, and because it matches a variety of other compounds with the constituent "aineiden" (*substances'*).

Query expansion

Tables 7 and 8 summarize the average effectiveness results for the different query expansion methods in the high recall scenario (Scenario 1), and in the high precision scenario (Scenario 2).

The high recall scenario was evaluated using nDCG@10 and nDCG@50 with the lenient log 10 discounting, and MAP (Table 7). Each effectiveness measure ranks the best query expansion methods differently. However, the general trends are clear: all query expansion methods clearly improve performance as compared to the baseline. The improvement varies between 17-118 % depending on the query expansion method and effectiveness measure used. The ten best query expansion methods are always the same and the performance differences between them are minor. The plain FCG was the weakest query expansion method. It improved the results over the baseline by 17-47 % depending on how many case forms were added to the expanded queries and which effectiveness measure was used. The differences to the baseline were typically not statistically significant, despite the relatively high percentage differences, except for MAP (FCG12 and FCG22). The differences between FCG expansion using 6, 12 or 22 case forms were not statistically significant.

Run name	nDCG@10 log10	rank	Diff-% to baseline	nDCG@50 log10	rank	Diff-% to baseline	MAP	rank	Diff-% to baseline
Baseline	0.221	20	+0 %	0.248	20	+0 %	0.128	20	+0 %
FCG6	0.279	19	+26 %	0.290	19	+17 %	0.171	19	+34 %
FCG12	0.286	17	+29 %	0.301	17	+21 %	0.180	18	+41 % *
FCG22	0.281	18	+27 %	0.300	18	+21 %	0.188	15	+47 % *
SG_CCI1_5	0.309	14	+40 %	0.333	14	+34 %	0.195	14	+52 % *
SG_CCI2_5	0.303	15	+37 %	0.317	15	+28 %	0.183	17	+43 %
SG_CCI3_5	0.302	16	+37 %	0.316	16	+27 %	0.184	16	+44 %
SG_CCI1_20	0.371	13	+68 % *	0.393	13	+58 % *	0.246	13	+92 % *
SG_CCI2_20	0.398	11	+80 % *	0.401	11	+62 % *	0.253	11	+98 % *
SG_CCI3_20	0.385	12	+74 % *	0.394	12	+59 % *	0.251	12	+96 % *
SG_CCI1_30	**0.426**	**1**	**+93 % ***	0.430	7	+73 % *	0.276	3	+116 % *
SG_CCI2_30	0.417	6	+89 % *	0.421	10	+70 % *	0.267	10	+109 % *
SG_CCI3_30	0.418	5	+89 % *	0.429	9	+73 % *	0.275	4	+115 % *
SG_CCI1_50	0.411	9	+86 % *	0.436	4	+76 % *	0.277	2	+116 % *
SG_CCI2_50	0.422	2	+91 % *	0.440	2	+77 % *	**0.279**	**1**	**+118 % ***
SG_CCI3_50	0.411	9	+86 % *	0.436	4	+76 % *	0.275	4	+115 % *
FCG6_CCI2_20	0.420	3	+90 % *	0.430	7	+73 % *	0.269	7	+110 % *
FCG6_CCI2_30	0.420	3	+90 % *	0.438	3	+77 % *	0.270	6	+111 % *
FCG12_CCI2_20	0.414	7	+87 % *	0.435	6	+75 % *	0.268	9	+109 % *
FCG12_CCI2_30	0.414	7	+87 % *	**0.443**	**1**	**+79 % ***	0.269	7	+110 % *

Table 7. The results for Scenario 1: high recall. * means that the difference to the baseline is statistically significant.

All *s*-gram query expansion methods improved the results as compared to the baseline: using 5 variants (low expansion level) by 27-52 %, using 20 variants by 58-98 %, and using 30 variants by 70-116 %, depending on the effectiveness

measure used. Adding more variants no longer resulted in notable improvements. The difference to the baseline was statistically significant for all *s*-grams at the higher expansion levels of 20-50 variants. The differences between the different *s*-grams with higher expansion levels were not statistically significant. The performance difference between the low expansion level *s*-grams and the baseline were not statistically significant. Also, the CCI used did not notably affect the results. Finally, the combinations of FCG and *s*-grams improved the results over the baseline by 73-111 % depending on the effectiveness measures used. The different combinations were quite even. The combinations did improve results over the plain *s*-gram query expansion methods, but no differences measured were of statistical significance.

The high precision scenario was evaluated using nDCG@10 with the strict log 2 discounting, and P@10 (Table 8). The general trends for this scenario are the same as for scenario one. Measured performance was above the baseline for all query expansion methods. However, the measured differences passed statistical significance tests only in *s*-gram settings when the expansion level was 20 or higher. High enough expansion level is more important for performance than the specific *s*-gram settings used. There were no notable differences between the CCIs when the expansion level was kept constant. Expanding queries with the inflectional variants only was not enough. Including *s*-gram matching in the query expansion method was clearly more effective than the use of FCGs. These findings further underline the benefit of query expansion at level of at least 30 added *s*-gram variants and above. Performance improvement levels out around 50 *s*-gram variants.

Run name	nDCG@10 log2	rank	Diff-% to baseline	P@10	rank	Diff-% to baseline
Baseline	0.243	20	+0 %	0.294	20	+0 %
FCG6	0.305	19	+26 %	0.336	19	+14 %
FCG12	0.317	17	+30 %	0.358	18	+22 %
FCG22	0.308	18	+27 %	0.376	17	+28 %
SG_CCI1_5	0.331	14	+36 %	0.436	14	+48 %
SG_CCI2_5	0.328	15	+35 %	0.418	16	+42 %
SG_CCI3_5	0.325	16	+34 %	0.426	15	+45 %
SG_CCI1_20	0.395	13	+63 % *	0.488	11	+66 % *
SG_CCI2_20	0.418	11	+72 % *	0.498	5	+69 % *
SG_CCI3_20	0.406	12	+67 % *	0.494	9	+68 % *
SG_CCI1_30	0.438	7	+80 % *	0.506	2	+72 % *
SG_CCI2_30	0.437	8	+80 % *	0.498	5	+69 % *
SG_CCI3_30	0.440	6	+81 % *	0.504	3	+71 % *
SG_CCI1_50	0.435	9	+79 % *	0.484	13	+65 % *
SG_CCI2_50	0.442	4	+82 % *	0.490	10	+67 % *
SG_CCI3_50	0.432	10	+78 % *	0.486	12	+65 % *
FCG6_CCI2_20	**0.451**	**1**	**+86 % ***	**0.508**	**1**	**+73 % ***
FCG6_CCI2_30	0.445	3	+83 % *	0.498	5	+69 % *
FCG12_CCI2_20	0.449	2	+85 % *	0.504	3	+71 % *
FCG12_CCI2_30	0.442	4	+82 % *	0.496	8	+69 % *

Table 8. The results for Scenario 2: high precision. * means that the difference to baseline is statistically significant

Topic-level analysis

The average results show that *s*-gram query expansion is useful in Finnish historical document retrieval. However, not all topics and query keys benefit from *s*-gram expansion equally. Differences between the topics, query keys and combinations of query keys are likely to affect the usefulness of the approach. To gain more insight into the strengths and limitations of the *s*-gram query expansion, we analyzed the CCI1_30 results topic by topic.

Table 9 depicts the impact of query length on the baseline performance and on *s*-gram query expansion. Most queries in this experiment were short, one or two-word queries (37 of 50 queries). It is clear from Table 9 that especially one-word queries benefit from the *s*-gram query expansion: 15 out of the 17 one-word queries were improved by the query expansion and not one was damaged by it. Only 6 queries out of 50 were damaged by the *s*-gram query expansion. They were all queries with two or more query words and a relatively good baseline performance. Especially long queries with high baseline performance were damaged. The differences between the different query lengths were statistically significant when considering the categories of improved and not-improved (i.e., unaffected or damaged) queries (p=0.047): Pearson's chi-squared is calculated for the categories in the *TOTAL* column of Table 9 so that categories unaffected ("/") and damaged ("-") were combined into a single category.

Query length	**P@10≤ 0.1 N=21**			**0.2 ≤ P@10 ≤ 0.4 N=17**			**P@10 ≥ 0.5 N=12**			***TOTAL N=50***		
	+	/	-	+	/	-	+	/	-	+	/	-
1 word queries (n=17)	6	1	0	7	1	0	2	0	0	15	2	0
2 word queries (n=20)	8	3	0	3	0	1	2	2	1	13	5	2
3-5 word queries (n=13)	1	2	0	5	0	0	0	1	4	6	3	4
TOTAL	*15*	*6*	*0*	*15*	*1*	*1*	*4*	*3*	*5*	*34*	*10*	*6*

Table 9. Query length, baseline performance and *s*-gram QE performance. Columns marked with +, /, and - show the number of topics that were improved, unaffected and damaged by *s*-gram query expansion, respectively.

To better understand how query length, recall-base size and baseline performance impact the potential usefulness of *s*-gram query expansion, we examined a subset of the retrieved and missed relevant documents for set of topics with a focus on understanding why one-word queries were so uniformly improved by the *s*-gram query expansion and why long queries were more often damaged by the query expansion.

One-word queries

The performance of one-word queries depends heavily on the selection of a good query word. The major reasons for the low baseline performance of the one-word queries are related to the length and the low frequency of the query words used – the topics were not necessarily simpler or contained fewer facets than the topics expressed with longer queries; they were simply expressed more compactly. This is quite natural for Finnish, where new compound expressions can be readily formed. The query words were then generally long and many (11 out of 17) of them compounds. The average length of compound query words was 12.7 characters compared to the overall average of 9.7 characters. Long words tend to have specific meanings and consequently often low collection frequencies. Five out of the seventeen query words in the one-word queries had a zero collection frequency, eleven were low-frequency words (with less than 100 occurrences in the collection) and only one had a medium frequency (with 232 occurrences).

Typical reasons for the low collection frequencies in one-word queries included historical changes in compound formation, and in the use of vocabulary. Some vocabulary mismatches were due to the use of specific occasional compounds as query words. For example "Imatrankoski" (*Imatra rapid*) was simply called "Imatra" until founding the town of Imatra in 1948 led to a need of specifying the name; "diakonissalaitos" (*the deaconess order*) was varyingly spelled as "diakonissalaitos", "diakonissa laitos", "diakonissa-laitos", "diakonialaitos", etc. The targeted phenomena of the topic "seksuaaliradikalismi" (*sexual radicalism)* were very differently understood and discussed during the 19th century than today; and more often than using the expression "sokeainopetus" (*education of the blind*), words related to *schools for the blind* were used in relevant documents. The *s*-gram query expansion improved the results of many of these queries because the variants added to the queries included relevant parts of the compounds ("Imatra", "diakonissa", etc.). All these issues also occurred in longer queries, but they were not necessarily equally detrimental for the baseline performance because other query words could balance the situation.

Other mismatch phenomena – that even occurred in longer queries – included historical spelling variation, OCR errors and inflectional variation. All these had an overall negative effect on scoring due to scattering of the word frequencies to many variants; vocabulary mismatch due to use of synonymous or related expressions, anaphoras, or ellipses; or due to implicit references to a topic or one facet of a topic in relevant documents. For example "keksinnöt", (*inventions*) in topic 20, *Edison's inventions*, were often referred to by the names of the specific inventions; "diakonissalaitos", (*the deaconess order*), topic 2, was sometimes implicitly referred to in articles considering e.g. the role of deaconesses in missionary and charitable work. The high variant rate (of OCR errors and historical and inflectional variants) is typical for digitized historical collections.

Two-word and longer queries

Translation and expansion of fixed phrases is problematic in two-word queries and especially longer queries: while the fixed phrases are often unambiguous as phrases, their constituent words may be general and frequent. Separate occurrences of the constituent words may be too general or simply irrelevant, as observed in the case of first names in proper name phrases (in the newspapers the public figures are not referred to by first name only); the second constituents in phrases such as "Suezin kanava" (*Suez Canal)*, "Eiffel-torni" *(Eiffel tower)* and "Vepsän kieli" (*Veps language)*, of both constituents of the name of the play "Työmiehen vaimo" (*Worker's wife)* (this phrase is very ambiguous even as-is, because it was a common salutation for women during 1800s). Importantly, the inflectional patterns of (especially the first) constituent words in phrases are limited because the phrases are inflected as a single unit and therefore expanding the constituents with inflectional variants is not necessarily desirable. Inflected forms of first names most often refer to fictional persons in short stories etc.; derivatives generated for "vepsän" referred to the people (not the language) and any other inflection of "työmiehen" than the requested singular genitive form (*worker's*) is certainly noise.

The damaged queries more often contained query words that did not match the vocabulary or the concepts in the historical collection (e.g., "seksuaaliradikalismi" (*sexual radicalism*)*;* "sään ennustaminen" (*weather forecasting*)), or that were sensitive to query drift due to the combination of query words. For example, the *s*-gram query expansion decreased the performance of topic 52, "Maanalaiset rautatiet" (*Underground railways*), because the low frequency word *underground* that efficiently increased precision of the baseline query was expanded with non-relevant high frequency words related to peasants and countryside (e.g. "maalaiset") that co-occur with *railway* more often than *underground* does. The query words in "sään ennustaminen" (*weather forecasting*, topic 44) did not match the words used in relevant documents due to a historical shift in the preferred words and word forms in the context of weather forecasting. The words "ilman" (*weather's*, a synonym to "sään") and "ennustus" (*forecast*, *prophecy*) were commonly used in the collection. Of these, especially "ennustus" is not used in present day (scientific) weather forecasting contexts, due to the connotation *prophecy*.

Discussion and conclusions

In this article, we have described fuzzy approaches to generating query expansion terms in the context of historical document retrieval on a highly inflectional compounding language, Finnish. We combined *s*-gram matching and frequent case generation (FCG) to find inflectional, historical and OCR variants of query words from the historical document collection, and words semantically related to the modern query words. All the tested query expansion methods clearly improved the results over the baseline. *S*-gram query expansion seemed more promising than the more

precise FCG alone. The combination of FCG and *s*-grams slightly improved performance over the level achieved by the corresponding *s*-gram query expansion alone, especially in the precision-oriented evaluation scenario. These differences were not statistically significant however. Therefore, we conclude that capturing the spectrum of different sources and combinations of variation that occur in digitized historical documents is more important than covering the linguistically correct inflectional forms of query words. Variant recall seems more important than precision.

The choice of the exact *s*-gram query expansion approach was less important than the level of expansion used. All CCI's performed well, as long as the expansion level was high enough. The query expansion results clearly continued to improve until a query expansion level of 30 *s*-gram variants for each query word was reached. After that the performance improvement levelled out. This is in line with the findings of Harding et al. (1997), who found that in retrieval from OCR degraded texts using a higher *n*-gram distance threshold, and thus allowing more expansion terms to be used, consistently improved results as compared to lower distance thresholds. The level of query expansion required for best performance is notably higher than that in comparable cross-language information retrieval studies, where using 2-3 *s*-gram (or *n*-gram) variants has performed well (Hedlund et al., 2004; Järvelin et al., 2006). Digitized historical document collections and historical document retrieval clearly differ from the cross-language scenario by the amount of variation that occurs in the collections. The number of variants for words in our collection tended to be very high due to the several layers of variation: the historical instability in spelling, inflection and OCR errors added up to tens, even hundreds of variants for each word. The task of historical document retrieval from OCR-scanned collections is therefore rather that of query *expansion* than that of query *translation*.

Reaching a high precision in identifying the correct historical variants of modern query words seems less important for historical document retrieval performance than previously expected in studies focusing on translation precision (Kempken et al., 2006; Gotscharek et al., 2011). The average *s*-gram translation precision of around 78 % among the top-20 *s*-gram variant candidates was enough for reaching almost (and over) 100 % improvement over the baseline in average effectiveness with the best *s*-gram query expansion methods. High translation precision alone was no guarantee for improvements in query performance, and low translation precision did not always lead to deteriorating performance. In short queries query drift was kept reasonably low despite the noisiness of the expanded queries, probably because the good expansion terms co-occur non-randomly in relevant documents, while poor expansion terms co-occur randomly as shown by Buckley et al. (1995). However, the longer queries were more often damaged by the *s*-gram query expansion. The query structuring following the Pirkola's method (Pirkola, 1998) did not solve the problems of translation ambiguity occurring with the multi-word queries, when the noise in the *s*-gram expansion created irrelevant combinations of the query word variants that co-occurred more often than the correct translations of the query words.

Our topic level analysis showed that there was no single topic or query word feature that explained why the query expansion damaged or benefitted query performance in general. The potential of the *s*-gram query expansion – or its sensitivity to noise – was related to a combination of several factors. These entail the combination of query words, how precisely and comprehensively they defined a topic, their frequencies and variation and their likelihood to co-occur in relevant or irrelevant contexts, and the *s*-gram translation precision. The longer queries containing original query words with low frequency were then more sensitive to noise, because the likelihood of detrimental query term combinations increases with the number of query words.

In this article, we focused on one highly inflectional compounding language, Finnish. We believe that our results are applicable to other inflectional and compounding languages: the occurrence of many inflectional variants and historical instability in compound spelling may radically increase the number of word form variants occurring in historical document collections and thus lead to a need for extensive query expansion. The inflectional systems of languages do not usually go through fast, radical changes; it is reasonable to expect that the morphological complexity of historical language variants (in the age span relevant to historical document retrieval) is not radically different from the morphological complexity of the present day languages. The need of handling morphological variation varies between languages, even in monolingual information retrieval. The benefits achievable by morphological processing in, e.g., monolingual English retrieval are limited due to the limited extent of morphological variation occurring in English. We have no reason to believe that inflectional variation would be a major factor for English historical retrieval either. However, for morphologically complex languages major improvements in retrieval performance are achievable with adequate morphological processing (McNamee et al., 2009; Kettunen, 2013). The need of handling morphological variation has been acknowledged in previous studies on historical retrieval. For example, Ernst-Gerlach and Fuhr (2006; 2007) showed that a high coverage of German word form variants in a historical collection was achieved by first generating all modern inflectional variants of the query words, and then generating the historical variants of the inflectional forms. Even a combination of a modern stemmer and the historical variant generation performed well (Ernst-Gerlach & Fuhr, 2007).

Similarly to cross-language information retrieval, compounds were clearly problematic query words. Identifying and expanding queries with the compound parts was one of the most common reasons for why *s*-gram query expansion improved performance over the baseline. We therefore argue that splitting compound query words is highly recommended for historical document retrieval in compounding languages regardless of the approach chosen for handling historical, OCR and inflectional variation. We also suggest that extensive query expansion by *s*-gram variants of the query words is especially suitable for compounding languages and handling compound words in queries.

Historical instability in compound spelling has been reported as a problem for German historical document retrieval (Hauser et al., 2007). Similar historical instability in compound spelling is likely to occur even in other compounding languages such as Dutch, Swedish, and Russian given the lack of stable spelling rules. For languages such as English where phrases are more common than compounds written as single words, average queries might include more and shorter words and the optimal level of query expansion might be different.

Some types of variation are captured better than other by *s*-gram matching. For example, high frequency historical variants are sometimes missed in the *s*-gram query expansion, because strings where OCR errors have reversed the historical variation are more similar to the query words than the correct historical variants. The use of left padding probably increased the share of inflectional variants among the variant candidates to the detriment of historical and OCR variants. It is not obvious which type of variation is the most important to capture from information retrieval point of view. Therefore, it is important to be aware that the way the *s*-gram matching (or other approximate string matching approaches) is set up affects which variants will be retrieved.

S-gram query expansion improved especially the results of the short baseline queries that performed poorly in the top-10 ranks (P@10). This is important because in practice even small improvements in poorly performing queries can mean major benefits for the users by providing them with the first access points to the topics in the collection. This is more important than finding one more relevant document for an already high performing query, or losing one relevant document out of many in the top ranks. Table 10 summarizes the *s*-gram query expansion results for CCI1 with query expansion level 30. The category of poorly performing queries has been split into two categories in this table to highlight the change in the share of queries for which the baseline query retrieved no relevant documents in top-10: P@10=0 (no relevant documents retrieved in top-10), and P@10=0.1 (one relevant document retrieved in top-10). The baseline run had 10 topics (20 %), for which nothing relevant was found in top-10, the corresponding number for the s-gram run being 4 (8 %). The share of users who would find at most one relevant document among the top-10 results decreases from 42 % using the baseline queries to 16 % with the expanded queries. Likewise, the share of users who would find at least five relevant documents among the top-10 results increases from 24 % to 54 % when using the *s*-gram query expansion. *S*-gram matching is clearly a promising approach to managing variation in information retrieval from digitized historical document collections.

Performance level	**Baseline**	**Baseline %**	***s*-gram**	***s*-gram %**
P@10=0	10	20 %	4	8%
P@10=0.1	11	22 %	4	8 %
0.2 ≤ P@10 ≤ 0.4	17	34 %	15	30 %

P@10 ≥ 0.5	12	24 %	27	54 %
TOTAL	50	100 %	50	100 %

Table 10. Summary of the practical implications of the query expansion: comparing baseline and *s*-gram query performance.

While query expansion is usually considered to improve retrieval performance on informational search tasks, it has not proved fully as useful for more specific search tasks such as known-item search in previous studies (e.g., White & Marchionini, 2007). This study focused on informational search tasks, and the applicability of the *s*-gram query expansion approach to known-item search is therefore uncertain. *S*-gram query expansion broadens the scope of the queries especially given the number of related words (category 2 variants) added to the expanded queries. We have no knowledge of previous studies on query expansion in known-item historical document retrieval, but the results from e.g. TREC confusion track (Kantor & Voorhees, 2000) suggest that query expansion is not useful for known-item search from document collections corrupted by OCR errors. Koolen et al. (2006) who studied known-item historical document retrieval used a document translation approach, where the historical variants in the collection were modernized into a single modern spelling. They reached good results, so perhaps such more precise translation approaches are to prefer for known-item search tasks.

In this article, we adopted a simple approach to *s*-gram query expansion: the *N* most similar *s*-gram variant candidates were added to the expanded queries. Frequent case forms of the query words were used for improving the coverage of the inflectional variation. In future studies it would be interesting to combine *s*-gram matching and frequent case generation with other non-linguistic or resource lean approaches: frequency-based approaches to *s*-gram variant selection, rule-based approaches to handling regular historical, morphological and OCR variation (Ernst-Gerlach & Fuhr, 2006; Pirkola, Toivonen, Keskustalo & Järvelin, 2007), statistical approaches to lemmatization (Loponen & Järvelin, 2010) or other simple approaches to handling morphological variation in indexing, such as index word truncation (McNamee et al., 2009; Kettunen et al., 2010).

Footnotes

1. Examples from http://www.finnlectura.fi/verkkosuomi/Fonologia/sivu191.htm.

2. Consonant gradation means the weakening of the consonants (P, T and K in Finnish) in oblique stems compared to the nominative stems.

3. Http://digi.lib.helsinki.fi/sanomalehti/secure/main.html.

4 The corpus of early modern Finnish is available through the home page of the material service of the Institute for the Languages of Finland (KOTUS): www.kaino.kotus.fi (pages in Finnish).

5. Http://www.lemurproject.org.

6. Vectora is an application for computing cumulated gain-based evaluation metrics developed and administered at the School of Information Sciences at the University of Tampere.

7. Direct comparisons of the precision values reported in the different studies are difficult though: the test data are different and the definition of precision varies from study to study. Reporting e.g. recall values would not have been possible in our study due to the extent of variation in the collection.

Acknowledgements

This work was partially funded by Tampere Graduate School for Information Science and Engineering (TISE). We are grateful to Docent Ari Pirkola for his valuable comments and support.

List of references

Airio, E., & Kettunen, K. (2009). Does dictionary based bilingual retrieval work in a non-normalized index? *Information Processing and Management*, 45(6), 703-713.

Alkula, R. (2001). From Plain Character Strings to Meaningful Words: Producing Better Full Text Databases for Inflectional and Compounding Languages with Morphological Analysis Software. *Information Retrieval*, 4(3-4), 195-208.

Amati, G., Celi, A., Di Nicola, C., Flammini, M., & Pavone, D. (2011). Improved stable retrieval in noisy collections. In *Advances in Information Retrieval Theory*, 342-345. Springer Berlin Heidelberg.

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval*. Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA.

Beitzel, S., Jensen, E., & Grossman, D. (2003). A Survey of retrieval strategies for OCR text collections. In *Proceedings of Symposium on Document Image Understanding Technology*, 145-151.

Braun, L., Wiesman, F., & Sprinkhuizen-Kuyper, I.. (2002). Information retrieval from historical corpora. In *Proceedings of the 3rd Dutch-Belgian Information Retrieval Workshop*, 106–112.

Bremer-Laamanen, M. (2001). A Nordic digital newspaper library. *International preservation news* no. 26 (Dec, 2001), 18-20.

Buckley, C., Salton, G., Allan, J., & Singhal, A. (1995). Automatic query expansion using SMART: TREC 3. *NIST special publication sp*: 69-69.

Dahl, Ö. (2000). Språkets enhet och mångfald [The unity and diversity of language]. Studentlitteratur, Lund.

Ernst-Gerlach, A**.,** & Fuhr, N. (2007). Retrieval in text collections with historic spelling using linguistic and spelling variants. In *Proceedings of the 7th Joint Conference on Digital Libraries JCDL'07*, 333-341.

Gotscharek, A., Reffle, U., Ringsletter, C., Schulz, K., & Neumann, A.(2011). Towards information retrieval on historical document collections: the role of matching procedures and special lexica. *International Journal on Document Analysis and Recognition (IJDAR)*, 14(2), 159-171.

Harding, S. M., Croft, W. B., & Weir, C. (1997). Probabilistic retrieval of OCR degraded text using n-grams. In *Research and advanced technology for digital libraries*, 345-359. Springer Berlin Heidelberg.

Hauser, A., Heller, M., Leiss, E., Schulz, K. U., & Wanzeck, C. (2007). Information access to historical documents from the early new high German period. In *Proceedings of IJCAI-07 Workshop on Analytics for Noisy Unstructured Text Data*, 147-154.

Hedlund, T. (2002). Compounds in dictionary-based cross-language information retrieval. *Information Research*, 7(2).

Hedlund, T., Airio, E., Keskustalo, H., Lehtokangas, R., Pirkola, A., & Järvelin, K. (2004). Dictionary-based cross-language information retrieval: Learning experiences from CLEF 2000-2002. *Information Retrieval* 7(1-2), 99-119.

Holley, R. (2009). How good can it get? Analysing and improving OCR accuracy in large scale historic newspaper digitisation programs. *D-Lib Magazine*, 15(3-4), [http://www.dlib.org/dlib/march09/holley/03holley.html].

Häkkinen, K. (1994). Agricolasta nykykieleen. Suomen kirjakielen historia [From Agricola to modern language. The history of standard Finnish]. WSOY, Helsinki.

Järvelin, A., & Järvelin, A. (2008). Comparison of *s*-gram Proximity Measures in Out-of-Vocabulary Word Translation. In *Proceedings of 15th Symposium on String Processing and Information Retrieval (SPIRE'08)*: 75-86.

Järvelin, A., Järvelin, A., & Järvelin, K. (2007). *S*-grams: Defining Generalized *n*-grams for Information Retrieval. *Information Processing & Management*, 43(4): 1005-1019.

Järvelin, A., Kumpulainen, S., Pirkola, A., & Sormunen, E. (2006). Dictionary-independent translation in CLIR between closely related languages. In *Proceedings of the 6th Dutch-Belgian Information Retrieval workshop (DIR'06)*: 25-32.

Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4): 442-446.

Kantor, P., & Voorhees, E. (2000). The TREC-5 Confusion Track: Comparing Retrieval Methods for Scanned Text. *Information Retrieval, 2*, 165-176.

Karlsson, F. (1983). Suomen kielen äänne- ja muotorakenne [Phonology and morphology of Finnish]. WSOY, Helsinki, 1983.

Kekäläinen, J., & Järvelin, K. (2000). The co-effects of query structure and expansion on retrieval performance in probabilistic text retrieval. *Information Retrieval*, 1(4): 329 - 344.

Kempken, S., Luther, W., & Pilz, T. (2006). Comparison of distance measures for historical spelling variants. In *Artificial Intelligence in Theory and Practice*, IFIP Series 219: 295-304. Springer US.

Keskustalo, H., Pirkola, A., Visala, K., Leppänen, E., & Järvelin, K. (2003). Non-adjacent Digrams Improve Matching of Cross-Lingual Spelling Variants. In *Proceedings of the 10th International Symposium on String Processing and Information Retrieval (SPIRE'03)*: 252-265.

Kettunen, K. (2009). Reductive and generative approaches to management of morphological variation of keywords in monolingual information retrieval: An overview. *Journal of Documentation*, 65(2): 267-290.

Kettunen, K. (2013) Managing word form variation of text retrieval in practice - Why language technology is not the only cure for better IR performance? *Trends in information management* 9(1).

Kettunen, K., & Airio, E. (2006). Is a morphologically complex language really that complex in full-text retrieval? In *Advances in Natural Language Processing, LNAI* 4139: 411 - 422.

Kettunen, K., Airio, E., & Järvelin, K. (2007). Restricted inflectional form generation in management of morphological keyword variation. *Information Retrieval*, 10(4-5): 415-444.

Kettunen, K., & Baskaya, F. (2011). Stemming Finnish for Information Retrieval–Comparison of an Old and a New Rule-based Stemmer. In*: Proceedings of the 5th Language & Technology Conference (LTC 2011)*: 476–480.

Kettunen, K., McName, P., & Baskaya, F. (2010). Using syllables as indexing terms in full-text information retrieval. In Proceedings of Baltic HLT: 225-232.

Koolen, M., Adriaans, F., Kamps, J., & De Rijke, M. (2006). A cross-language approach to historic document retrieval. In *Advances in Information Retrieval: Proceedings of 28th European Conference on IR Research (ECIR'06)*: 407-419.

Koskenniemi, K. (1983). Two-level morphology: A general computational model for word-form recognition and production. University of Helsinki, Department of General Linguistics, Helsinki. Publications 11.

Liu, L-M., Babad, Y., Sun, W., & Chan, K-K. (1991). Adaptive post-processing of OCR text via knowledge acquisition. In *Proceedings of the 19th annual conference on computer science (CSC'91)*: 558-569.

Loponen, A., & Järvelin, K. (2010). A Dictionary- and Corpus-Independent Statistical Lemmatiser for Information Retrieval in Low-Resource Languages. In: Agosti, M. & al. (Eds.), Multilingual and Multimodal Information Access Evaluation, Proceedings of the International Conference of the Cross-Language Evaluation Forum, CLEF 2010: 3-14. Heidelberg: Springer.

McNamee, P., & Mayfield, J. (2004). Character n-gram tokenization for European language text retrieval. *Information Retrieval*, 7(1-2): 73–97.

McNamee, P., Nicholas, C., & Mayfield, J. (2009). Addressing morphological variation in alphabetic languages. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*: 75-82.

Mitra, M., & Chaudhuri, B. (2000). Information retrieval from documents: A survey. *Information retrieval*, 2(2-3): 141-163.

Mittendorf, E., & Schäuble, P. (1996). Measuring the effects of data corruption on information retrieval. In *Proceedings of the SDAIR'96 Conference*: 179 –189.

Navarro, G. (2001). A guided tour to approximate string matching. *ACM Computing Surveys*, 33(1): 31–88.

Niemikorpi, A. (1991.) Suomen kielen sanaston dynamiikkaa [Dynamics of the Finnish Vocabulary]. Vaasa: University of Vaasa. Acta Wasaensia 26.

O'Rourke, A., Robertson, A., & Willett, P. (1997). Word variant identification in old French. *Information research*, 2(4).

Paik, J., Kettunen K., Pal, D., & Järvelin, K. (2013). Frequent Case Generation in Ad hoc Retrieval of Three Indian Languages: Bengali, Gujarati and Marathi. In *Multilingual Information Access in South Asian Languages*: *Second International Workshop, FIRE 2010, Gandhinagar, India, February 19-21, 2010 and Third International Workshop, FIRE 2011, Bombay, India, December 2-4, 2011, Revised Selected Papers*: 38-50.

Pilz, T., Luther, W., Fuhr, N., & Ammon, U. (2006). Rule-based search in text databases with nonstandard orthography. *Literary and Linguistic Computing*, 21(2): 179-186.

Pilz, T., Ernst-Gerlach, A., Kempken, S., Rayson, P., & Archer, D. (2008). The identification of spelling variants in English and German historical texts: Manual or automatic? *Literary and Linguistic Computing*, 23(1): 65-72.

Pirkola, A. (1998). The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. In: *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*: 55-63.

Pirkola, A., Keskustalo, H., Leppänen, E., Känsälä, A-P., & Järvelin, K. (2002). Targeted s-gram matching: a novel n-gram matching technique for cross- and monolingual word form variants. *Information Research*, 7(2).

Pirkola, A. & Toivonen, J. & Keskustalo, H. & Järvelin, K. (2007). Frequency-based identification of correct translation equivalents (FITE) obtained through transformation rules. *ACM Transactions on Information Systems* (TOIS) 26(1).

Raitanen, I. (2012). "Etsikäät hywää ja älläät pahaa." Tiedonhakumenetelmien tuloksellisuuden vertailu merkkivirheitä sisältävässä historiallisessa sanomalehtikokoelmassa [Comparison of the effectiveness of information retrieval methods in historical newspaper collections containing character errors]. Pro Gradu (Master's thesis), University of Tampere, Finland.

Robertson, A. M. & Willett, P. (1993). A comparison of spelling correction methods for the identification of word forms in historical text databases. *Literary and linguistic computing*, 8(3): 143-152.

Robertson, A. M. & Willett, P. (1998). Applications of n-grams in textual information systems. *Journal of Documentation*, 54(1): 48–69.

Savoy, J. & Naji, N. (2011). Comparative information retrieval evaluation for scanned documents. In *Proceedings of the 15th WSEAS international conference on Computers*: 527-534.

Scholer, F. & Turpin, A. (2009). Metric and relevance mismatch in retrieval evaluation. In *Proceedings of the 5th Asia Information Retrieval Symposium (AIRS'09):* 50-62.

Sormunen, E. (2000). A Method for measuring Wide Range Performance of Boolean Queries in Full-Text Databases. Doctoral Thesis. Tampere: University of Tampere. Acta Electronica Universitatis Tamperensis.

Sormunen, E. (2002). Liberal Relevance Criteria of TREC – Counting on Negligible Documents? In *Proceedings of the 25th annual international ACM SIGIR conference on Research and Development in Information Retrieval*: 324-330.

Taghva, K., Borsack, J., & Condit, A. (1994). Results of applying probabilistic IR to OCR text. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*: 202-211.

Ullman, J. R. (1977). A binary n-gram technique for automatic correction of substitution, deletion, insertion and reversal errors in words. *The Computer Journal*, 20(2): 141–147.

White, R. & Marchionini, G. (2007) Examining the effectiveness of real-time query expansion. *Information Processing & Management*, 43(3): 685-704.