INFIM
DEPARTMENT OF INFORMATION STUDIES AND INTERACTIVE MEDIA
UNIVERSITY OF TAMPERE

Data Driven Methods for Improving Mono- and Cross-lingual IR Performance in Noisy Environments

Antti Järvelin, Tuomas Talvensaari
& Anni Järvelin

2009
http://www.trim.fi
TRIM Research Notes 1
ISSN 1798-6001
ISBN 978-951-44-7929-8

Data Driven Methods for Improving Mono- and Cross-lingual IR Performance in Noisy Environments

Antti Järvelin[a,*], Tuomas Talvensaari[a], Anni Järvelin[a,1]

[a]*Department of Information Studies and Interactive Media, University of Tampere, FI-33014 University of Tampere, Finland*

Abstract

In cross-language information retrieval (CLIR), novel or non-standard expressions, technical terminology, or rare proper nouns can be seen as noise when they appear in queries or in the target collection, because they often are out-of-vocabulary (OOV) for dictionaries that are used to translate queries. In this paper, three data driven approaches to these problems are presented. The two first methods, the transformation rule based translation (TRT) method and the classified *s*-gram method, operate on string level. With them approximate matches of a query word can be recognized from the target document collection and included into the target query. In the third method, the corpus-based approach, comparable corpora are employed to derive translation knowledge that can be used to translate OOV words. Besides the overview of the methods, three case studies highlighting their practical applications in CLIR are presented. The methods are shown to be effective in OOV word translation (*s*-grams), in query translation without dictionaries between closely related languages (TRT and *s*-grams), and in query translation in a special domain (*s*-grams, TRT and corpus based methods).

Keywords: Cross-language information retrieval, noise, OOV words, TRT, *s*-grams, corpus based methods

1. Introduction

Noisy data introduces problems to many information retrieval applications. Out-of-vocabulary (OOV) words in cross-language information retrieval (CLIR) and OCR errors and historical spelling variants in historic document retrieval are examples of this. OOV words introduce noise in query translation as they cannot be translated with the common dictionary-based query translation tools. Many typical OOV words, such as proper names and technical terms, are usually important query keys (Pirkola, 1998), which makes their translation an essential question in CLIR. Such OOV words are often cross-lingual spelling variants in different languages, i.e., they share a common root but are rendered with different spelling of the underlying sounds. The cross-lingual spelling variants are similar (e.g. a Swedish word *heksaklorid* and its English variant *hexachloride*), which provides a good basis for the use of, e.g., approximate string matching to translate OOV words. Other techniques suggested for the OOV word translation include the TRT technique based on the use of statistically derived transformation rules, corpus-based translation of OOV words, where translation equivalents are automatically mined from the web (Cheng et al., 2004) and transliteration that has been used for phonetic translation of OOV words, e.g., between Arabic and English (AbdulJaleel and Larkey, 2003), Japanese and English (Fujii and Ishikawa, 2001), and English and Persian (Karimi et al., 2006).

CLIR between closely related languages can be seen as a special case of OOV word translation: When a large part of two languages vocabularies are orthographically related words – cross-lingual spelling variants – all the query words can be handled as OOV words with relatively good query translation quality as a result (Järvelin et al., 2006). Therefore CLIR between closely related languages can be seen as querying from noisy texts. This is also the case with

*Corresponding author.

Email addresses: antti.jarvelin@cs.uta.fi (Antti Järvelin), tuomas.talvensaari@uta.fi (Tuomas Talvensaari), anni.jarvelin@uta.fi (Anni Järvelin)

[1]*Current address:* Swedish Institute of Computer Science, Box 1263, SE-164 29 Kista, Sweden.

historic document retrieval (HDR), where the spelling variation introduced by the language development and OCR errors in the scanned collections cause similar problems (Robertson and Willett, 1992; Koolen et al., 2006).

From the information retrieval system's point of view, both spelling variation (be it cross-lingual or historical spelling variation), and OCR errors in the digitalized collection, can be seen as problems of searching relevant documents from noisy target collection. The queries can also contain noise in the form of novel or non-standard expressions, typos, misspellings, etc. In this paper three data driven approaches for dealing with noisy target collections and queries are presented, and their performance in various CLIR-related tasks are evaluated.

The first approach, the *classified s-gram matching*, is an approximate string matching technique, which generalizes the well known *n*-gram matching technique. Pirkola et al. (2002) and Keskustalo et al. (2003) showed that cross-lingual spelling variation can advantageously be modeled with the *s*-grams which thus can be effectively used for searching translation candidates for OOV words. Especially, they showed that in this task the classified *s*-gram matching outperformed other established approximate string matching techniques, such as edit distance, longest common subsequence, and *n*-grams. Section 3.1 presents extensive experiments where typical OOV words (proper nouns, technical terms) were translated with the classified *s*-gram method between 11 language pairs. The results indicate that *s*-grams outperform *n*-grams in OOV translation especially between remotely related languages. The performance of the *s*-gram technique as part of full CLIR system is also explored in Sections 3.2 and 3.3.

The second approach, called *Transformation Rule based Translation* (TRT), is a fuzzy translation technique based on statistical rules that model typical character changes between cross-lingual spelling variants. The technique is used in two steps: First the statistical rules are used to create intermediate forms of the source words. Then the intermediate forms are matched with their target language equivalents through approximate string matching. Toivonen et al. (2005) found that the two-step TRT technique performed clearly better than *n*-gram matching in translating technical terms and proper names. Pirkola et al. (2006) developed a word frequency-based technique for identifying the correct translations from the alternatives produced by TRT called *FITE* (frequency-based identification of translation equivalents). This technique performed well as a complement to dictionary-based CLIR in (Pirkola et al., 2006). In Section 3.2 experiments where TRT and classified *s*-grams were used in CLIR between closely related languages, are presented. Norwegian queries were translated into Swedish with the TRT method, and promising results were achieved. Section 3.3 presents experiments where the performance of the FITE-TRT technique is compared to other OOV-translation techniques investigated in this paper.

The third approach presented in this paper is to mine parallel or comparable texts from the web, and align them so that passages in the source language are mapped to similar passages (i.e., translations or, in the case of comparable texts, topically related texts) in the target language. The alignments in turn can be used to derive translation knowledge which can be used in CLIR query translation (Kraaij et al., 2003; Talvensaari et al., 2008). This kind of "real-life" training data can boost CLIR performance in situations when queries or the target collection are marred by noise such as novel or non-standard terminology or abbreviations. However, even in the vast WWW, it is hard to find clean parallel content for a given language pair and domain, and, consequently, it is often necessary to resort to noisier comparable texts. Thus, besides noisy queries or target collections, noise can also be induced into CLIR via noisy training data. In the experiments presented in Section 3.3, German queries belonging to a specific domain (genomics) were translated into English with various CLIR set-ups. The results indicate that noisier comparable corpora work well in OOV translation, especially in special domains.

The rest of this paper is organized as follows. Section 2 gives an introduction to the three methods presented in this paper. Section 3 discusses three case studies where the methods are applied to CLIR related problems. In Section 3.1 the performance of classified *s*-grams in OOV word matching is illustrated. Section 3.2 gives an example how TRT, and *s*-gram methods can be used in CLIR to avoid dictionary translation of queries between closely related languages. Corpus based methods for noisy translation are investigated in Section 3.3. Finally, Section 4 provides discussion and conclusions.

2. Data Driven Methods for Improving Query Performance

2.1. s-grams

The *classified s-gram matching* (Pirkola et al., 2002), is an approximate string matching technique generalizing the well know *n*-gram matching technique. *n*-grams have earlier been applied successfully e.g. in proper name matching (Pfeiffer et al., 1996) and historic word variant matching (O'Rourke et al., 1997).

Table 1: The gram classes for forming the s-grams of length two with different CCIs for the string "abracadabra".

CCI	Gram classes
$\{\{0\}\}$	$\{ab, br, ra, ac, ca, ad, da, ab, br, ra\}$
$\{\{1\}\}$	$\{ar, ba, rc, aa, cd, aa, db, ar, ba\}$
$\{\{2\}\}$	$\{aa, bc, ra, ad, ca, ab, dr, aa\}$
$\{\{1,2\}\}$	$\{ar, ba, rc, aa, cd, aa, db, ar, ba, aa, bc, ra, ad, ca, ab, dr, aa\}$
$\{\{0\},\{1,2\}\}$	$\{ab, br, ra, ac, ca, ad, da, ab, br, ra\}$, $\{ar, ba, rc, aa, cd, aa, db, ar, ba, aa, bc, ra, ad, ca, ab, dr, aa\}$

In the classified s-gram matching the strings to be compared are split into substrings of length n (n-grams) and then the proximity of the strings is defined as the overlap of the strings' n-grams using some proximity measure. The technique differs from the n-gram matching in two important aspects. Firstly, skipping characters is allowed when forming the substrings. Secondly, the substrings are produced using various skip lengths and are classified into sets called *gram classes* based on the number of characters skipped. 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. For example, CCI $\{\{0\},\{1,2\}\}$ denotes that two gram classes are to be formed from a string: one formed of adjacent characters ($\{0\}$) and one formed both by skipping one and two characters ($\{1,2\}$). Only the substrings belonging to the same gram class of the CCI are compared to each other, thus the name classified s-gram matching. Table 1 provides an example of forming s-grams of different CCIs for the string "abracadabra".

s-gram-based string proximity measures are based on strings' *s-gram profiles* which contain the information how many times each s-gram occurs in a given string (see (Järvelin et al., 2007) for details).

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

Definition 2. Let $w \in \Sigma^*$, $C \subset \mathbb{N}$ be 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. In fact, Järvelin and Järvelin (2008) show that in classified s-gram-based OOV word matching, proximity measures based on the binary gram class profile perform better or equally well as the proximity measures based on the general gram class profile. The binary gram class profile is defined by binarizing the gram class profile of Definition 2.

Definition 3. Let $w \in \Sigma^*$, and $C \subset \mathbb{N}$ 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 binary vector $B_{n,C}(w) = (B_C(w)[x]), x \in \Sigma^n$.

Various proximity measures can be used to calculate string proximities based on the general and binary gram class profiles. For example, the Dice's coefficient, which was the best performing proximity measure in the tests of Järvelin and Järvelin (2008), is defined for strings v and w by

$$D_{n,C}(v,w) = \frac{2B_C(v)^T B_C(w)}{\|B_C(v)\|^2 + B_C(v)^T B_C(w) + \|B_C(w)\|^2}, \quad (1)$$

where T denotes the transpose of a vector.

The classified s-gram matching is based on character combination indices, which contain at least one gram class. The CCI based string proximity measures are defined as the average proximity of strings' gram class proximities. That is, for a given CCI $\mathscr{C}$, gram length n, and a gram class proximity measure $\mathscr{P}$, the corresponding CCI based proximity measure $\mathscr{P}_{n,\mathscr{C}}(v,w)$ between the strings v and w is given by

$$\mathscr{P}_{n,\mathscr{C}}(v,w) = \frac{1}{|\mathscr{C}|} \sum_{C \in \mathscr{C}} \mathscr{P}_{n,C}(v,w). \quad (2)$$

Thus, for example, the CCI based Dice's coefficient is obtained by substituting $D_{n,C}(v,w)$ for $\mathscr{P}_{n,C}(v,w)$ in the Eq. (2).

It has been found beneficial to use padding spaces around the strings when forming s-grams (Keskustalo et al., 2003; Robertson and Willett, 1998). This helps to get the characters at the beginning and at the end of a string properly presented in string comparison.

2.2. TRT

The second approach, Transformation Rule based Translation (TRT), is a fuzzy translation technique developed for translation of OOV words in CLIR (Toivonen et al., 2005). It is based on the use of statistically generated rules that model typical character changes and correspondences between cross-lingual spelling variants within a language pair. The rules are created from a large set of cross-lingual spelling variants, that are typically extracted from dictionaries. The TRT technique can be used both as a complement to dictionary-based query translation or as the sole translation technique for translation between closely related languages. Even if the technique is based on the use of bilingual word lists, such as dictionaries, it can overcome many of the dictionary-based translation's problems as the transformation rules can be used for translating words that are not present in the word list used for their generation. The word lists used for the generation of TRT rules could probably also be mined from the Web.

The TRT technique has two steps: the transformation rules are combined to n-gram matching so that first all possible rules are used to create a set of intermediate translation candidates. Then n-grams are used to match these against the target document collection's index to discard bad intermediate translations that are not real words. The idea of the TRT and the generation of the transformation rules are described in more detail in (Toivonen et al., 2005).

A *transformation rule* contains source and target language characters and their context characters. 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* shows how reliable a rule is, in practice it is the frequency of a rule divided by the number of source words where the source substring of the rule occurs. Frequency and confidence factor are threshold factors that can be used for selecting the most reliable rules for the translation. An example of a Norwegian to Swedish rule is:

$$\text{for} \Rightarrow \text{för}\ [\text{beginning}, 132, 147, 89.80]$$

The rule means that the letter o, between r and f, is transformed into the letter ö in the beginning of words, with the confidence factor being 89.80. The confidence factor is the frequency of the rule (132) divided by the number of source words where the string occurs (147).

The performance of the TRT technique can be enhanced by combining it to the FITE (frequency-based identification of translation equivalents) technique, instead of n-grams. The FITE-TRT (Pirkola et al., 2006) is a statistical technique for the identification of correct transformation equivalents of source words obtained by TRT. The core idea of FITE is that except for the correct translation equivalents, the word forms yielded by TRT are malformed rather than real words, or they are rare words, e.g., foreign language words in the target language text. The equivalents belong to a language's basic lexicon and are much more common in the language that the other word forms. Therefore document frequencies of words can be used to find the best translation alternatives given by TRT.

2.3. Corpus Based Methods

Parallel or *comparable corpora* are often-used resources in CLIR query translation. A parallel corpus is a collection where texts in one language are aligned with their translations in another language. Comparable corpora, on the other hand, consist of texts that are not translations, but share similar topics. They can be, e.g., newspaper collections written in the same time period in different countries.

Parallel corpora are preferred to comparable corpora as translation resources because more dependable knowledge can be derived from them. However, a translation corpus has to fulfill two conditions before it can be of use: firstly, naturally, the source and target languages of the aligned collection have to match the source and target languages of the CLIR queries at hand. Secondly, the domain, or the topic of speech, of the translation corpus has to match the domain of the queries. For example, translating sports-related queries with a parallel corpus consisting of agricultural legislation would probably produce bad results.

The problem is that high quality parallel corpora do not exist for all language pairs and domains. Hence, it is sometimes necessary to resort to noisier comparable corpora. In CLIR query translation, comparable corpora can be utilized as a complementary resource to provide translations to words that are OOV for the more general resources. The use of noisier resources is acceptable in CLIR, because query translation is easier than machine translation (MT). Queries can be translated word-for-word, whereas in MT, syntactical knowledge is required in addition to lexical coverage.

There are many ways to utilize an aligned (i.e., parallel or comparable) corpus. For example, the alignments can be used in *cross-language query expansion* (Ballesteros and Croft, 1998), or they can be used as training data for statistical translation models, which in turn can be employed in translating queries (Kraaij et al., 2003).

Another approach is to use the corpus as a *cross-language similarity thesaurus*, meaning a structure in which target language words are ranked based on their calculated similarity with a source language query word that is given as input. The more often two words co-occur in the aligned documents, the more similar they are. The highest ranking words are assumed to be either translations of the input word or related to it in some other manner. This approach was pioneered by Sheridan and Ballerini (1996).

In similarity thesaurus calculation, the vector space model of information retrieval (Baeza-Yates and Ribeiro-Neto, 1999) can be used in an inverted way. In the classic vector model, documents are modeled as vectors whose features correspond to the words in the documents. The similarity of two documents (or the similarity of a query to a document, since queries can be seen as short documents) can then be calculated, e.g., with the cosine of the angle between the document vectors. In similarity thesaurus calculation, the source language word to be translated is thought of as the query, and target language words are retrieved as the answer. The aligned documents are thought of as the defining features of words, rather than the other way around, as in document retrieval. The distribution of a word across the documents determines its location in the semantic space defined by the documents. As in document retrieval, measures such as the cosine similarity can be used in defining the similarity between two words.

To use a comparable corpus as a similarity thesaurus, the documents of the two languages have to be aligned so that documents in the source language are mapped to documents in the target language that cover similar topics. Talvensaari et al. (2008) proposed a method where source language documents are used as queries that are translated into the target language with an initial dictionary. The translated queries are then run against the target documents with a document ranking algorithm, and a few highest ranking documents are aligned with the source document. Similarity score thresholds are used to filter out bad alignments. Consequently, only a subset of the original corpus is part of the aligned translation corpus.

Talvensaari et al. (2008) created a German-English comparable corpus in the genomics domain. The translation corpus was created by first retrieving genomics-related text in German and English from the web. This was done by means of language-aware focused web crawling. The texts were then aligned with the procedure described above. A total of 39,143 German paragraphs extracted from the web pages were aligned with 39,190 unique English paragraphs (some target documents were part of more than one alignments). The alignments were made in 1-to-*n* manner, meaning that one source document was aligned with one or more target documents. The comparable corpus created in this way was used as a similarity thesaurus in the experiments presented in Section 3.3.

Table 2: The six tested CCIs. CCI_0 corresponds to the *n*-grams.

CCI_0	$\{\{0\}\}$	CCI_3	$\{\{0\},\{0,1\}\}$
CCI_1	$\{\{0,1\}\}$	CCI_4	$\{\{0\},\{1\},\{1,2\}\}$
CCI_2	$\{\{0,1,2\}\}$	CCI_5	$\{\{0\},\{1,2\}\}$

3. Case Studies

3.1. Out-of-Vocabulary Word Translation with s-grams

To demonstrate the value of the classified *s*-gram matching technique in the OOV word translation, a set of 271 typical OOV words was collected and translated between 11 language pairs using classified *s*-gram matching. These OOV words were mostly technical terms from the domains of biology, medicine, economics and technology, but also a list of geographical names obtained from (Keskustalo et al., 2003) was included. The search keys were expressed in seven languages (English, Finnish, French, German, Italian, Spanish, and Swedish) and were translated into four target languages (English, German, Finnish, and Swedish). English was combined to 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. Examples of English OOV words in the collection include *adrenalin*, *Pyongyang*, and *zygote*.

Target word lists (TWLs), from where the translations for the OOV words were searched from, consisted of CLEF 2003 (Peters, 2003) document collections' indices for the target languages. 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 CLEF'03 collection was considerably larger 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 with the TWOL morphological analyzer by Lingsoft Ltd. The words not recognized by the morphological analyzer were indexed as 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 exactly one correct translation for each search key in the TWLs.

The gram length was set to two in this experiment, as it has been found suitable with the classified *s*-grams (Pirkola et al., 2002; Keskustalo et al., 2003). The Dice's coefficient was used as the proximity measure between the strings, because it seems to be the best performing proximity measure with the classified *s*-grams (Järvelin and Järvelin, 2008). Totally six CCIs were tested (see Table 2). The tested CCIs contain also *n*-grams which correspond to CCI_0.

For each search key 100 best translations were produced, with exception of ties at the last place when all translations within the cohort of equal proximity values were included into the result set. Translations ranked lower were not taken into consideration. This is well motivated as taking more than 2-4 translation candidates into a query tends to deteriorate the query performance (Hedlund et al., 2004). To compare the CCIs, the average precision (AP) was calculated for each language pair and CCI at three different levels: among top 2, top 5 and top 100 highest ranked translation candidates. The top 2 and top 5 levels were the most interesting ones, as more translation candidates would deteriorate the query performance. 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 statistical significance of the differences between the *n*-grams and different CCIs inside each language pair was tested with Friedman test.

The results are presented in Table 3 when the top 5 translation candidates are considered. The results for top 2 and top 100 translation candidates gave the same order for the CCIs the overall APs being slightly lower in top 2, and slightly higher in top 100. The differences between the CCIs were largest with language pairs that were linguistically remotely related. Thus the classified *s*-gram matching technique performed better than the *n*-grams in noisy environments, i.e., when the languages were not closely related. When the environment was less noisy (the languages are more closely related), the differences between the classified *s*-gram matching and *n*-gram matching diminished. CCIs 4 and 5 are examples of how the classification of skip indices into suitable gram classes benefits the technique as the differences between these CCIs and *n*-grams were almost always statistically highly significant.

Table 3: The APs among top 5 translation candidates for all CCIs and language pairs. The best performing CCI for each language pair is on bold. Note that CCI_0 corresponds to *n*-grams. The performance gain achieved by using of CCI_1-CCI_5 compared to *n*-grams of CCI_0 is marked into parenthesis. Statistically highly significant differences ($p < 0.001$) between CCI_0 (*n*-grams) and the other CCIs are marked with ** and statistically significant differences ($p < 0.01$) with *.

Language	CCI_0	CCI_1	CCI_2	CCI_3	CCI_4	CCI_5
EN-FI	0.37	0.43 (+0.07)**	0.42 (+0.05)**	0.43 (+0.06)	**0.45 (+0.08)****	**0.45 (+0.08)****
FI-EN	0.40	0.44 (+0.04)	0.42 (+0.02)	0.45 (+0.05)**	0.45 (+0.05)**	**0.46 (+0.06)****
IT-EN	0.45	0.50 (+0.05)**	0.48 (+0.03)**	0.50 (+0.04)**	**0.53 (+0.07)****	0.52 (+0.06)**
SP-EN	0.53	0.54 (+0.01)	0.53 (+0.00)	0.55 (+0.02)	0.56 (+0.03)**	**0.57 (+0.04)****
EN-SW	0.55	0.56 (+0.01)	0.55 (+0.00)	0.56 (+0.01)	**0.57 (+0.02)****	0.56 (+0.02)*
SW-EN	0.54	0.56 (+0.02)	0.55 (+0.01)	0.57 (+0.03)**	**0.59 (+0.05)****	0.58 (+0.04)**
GE-EN	0.57	0.60 (+0.03)*	0.60 (+0.03)*	0.61 (+0.04)**	**0.63 (+0.06)****	**0.63 (+0.06)****
EN-GE	0.59	0.61 (+0.03)	0.59 (+0.00)	0.60 (+0.01)	0.61 (+0.03)**	**0.63 (+0.05)****
FR-EN	0.68	0.71 (+0.03)**	0.69 (+0.01)	0.71 (+0.03)**	**0.72 (+0.04)****	0.71 (+0.03)**
GE-SW	0.74	0.75 (+0.00)	0.73 (−0.01)	0.75 (+0.01)	**0.77 (+0.02)****	0.76 (+0.02)*
SW-GE	0.75	0.75 (+0.00)	0.73 (−0.02)*	0.76 (+0.01)	0.76 (+0.02)	**0.77 (+0.02)***
MEDIAN	0.55	0.56 (+0.01)	0.55 (+0.00)	0.57 (+0.02)	**0.59 (+0.04)**	0.58 (+0.04)

3.2. CLIR Between Closely Related Languages without Dictionary Translation

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 method. Closely related languages typically share a high number of spelling variants. If the number of the shared cross-lingual variants is high enough, query translation can be handled by cheaper and simpler fuzzy translation techniques.

Norwegian and Swedish are closely related Scandinavian languages: Around 90 % of the vocabularies of the languages are similar having only some orthographical and inflectional differences (Barˋodal et al., 1997). The TRT technique and the *s*-gram matching technique were tested in query translation from Norwegian to Swedish. The goal was to reach translation quality that would enable CLIR effectiveness comparable to dictionary-based translation.

A typical CLIR test setting with a subset of 50 search topics and the test collection from Cross-Language Evaluation Forum (CLEF) 2003 (Cross-Language and More -track) was used. Norwegian search topics were not included in the CLEF test environment. Therefore English test topics were translated into Norwegian by a native speaker. The document collection and the search topics were lemmatized prior to the dictionary translation and the monolingual Swedish query. No morphological preprocessing was done prior to the fuzzy translation. The stop words were removed.

Test queries were formed from the title and description fields of the test topics. The *s*-gram translation was done by translating the Norwegian topics into Swedish by matching the *s*-grams of the topic words against the Swedish document collection's index words' *s*-grams. CCI $\{\{0\},\{1,2\}\}$ was used and the gram length was set to two (digrams). The four best matches were selected as translations for each word.

The transformation rules were created by translating a part of the Swedish document collection's index to Norwegian with the GlobalDix dictionary by Kielikone Ltd. The final Norwegian to Swedish word-pair list had 3,058 unique word-pairs of non-identical words, with a maximum edit distance value of half of the length of the longer word and with minimum length of four characters. A confidence factor of 50 % and a low frequency threshold of 2 were used. The TRT was done as follows: First all the possible (fitting) TRT rules were applied to each source word to create intermediate translations. The intermediate translation with highest frequency and confidence factor was then selected and matched against the target document collection with *n*-grams. The four highest ranked keys from the result list of *n*-gram matching were selected for the final queries.

The GlobalDix dictionary by Kielikone Ltd. was used for the dictionary translation. All the translations for each topic word were selected to the query. To tackle the ambiguity problem common in dictionary-based CLIR the queries were structured according to the Pirkola method (Pirkola, 1998) that is known to be an effective way to disambiguate CLIR queries (Pirkola, 1998; Sperer and Oard, 2000). Also a "minimal effort" monolingual Swedish

Table 4: The MAPs and the differences between the query translation techniques for the Norwegian to Swedish CLIR tests.

Technique	Dicbase	*s*-grams	TRT
MAP	0.36	0.30	0.29
Difference to Dicbase	-	−16.7 %	−18.1 %

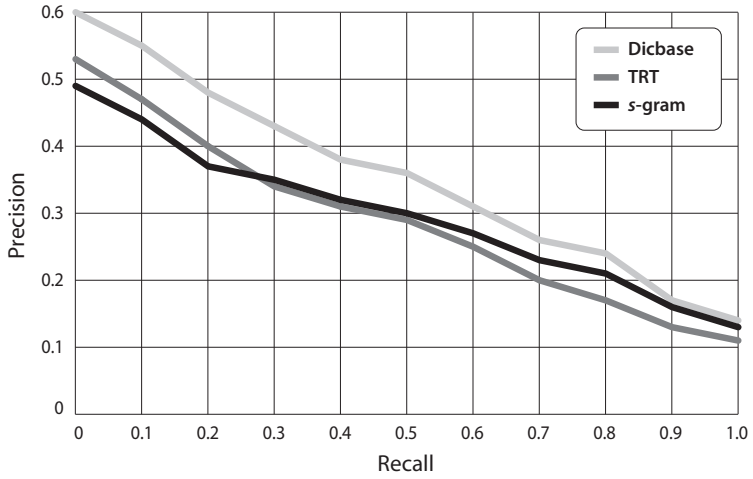


Figure 1: Precision-recall graph over the standard recall levels for the Norwegian to Swedish CLIR tests.

baseline query was run. The topic words were lemmatized but no compound splitting was done. This monolingual baseline performed slightly worse than the dictionary baseline and thus only the results for the dictionary baseline are reported here. The performance of the translation techniques was measured using mean average precision (MAP). The statistical significance of the results was tested using Friedman test (Conover, 1999) on inversed ranks.

The results are presented in Table 4 and in Fig. 1. Both the TRT and the *s*-grams achieved on average over 80 % of the dictionary baseline. The differences in terms of MAP were not statistically significant between the tested query translation approaches. This is a result that shows that fuzzy translation is a promising and interesting approach to query translation between closely related languages.

The results are the same when comparing the precision at the higher ranks, i.e., at the 0-0.1 recall level (Fig. 1). Again, the TRT and the *s*-grams achieved over 80 % of the dictionary baseline's performance. The differences between the translation techniques were not statistically significant here either. The documents placed at the top of the result list are the most important ones from the practical user perspective.

The *s*-grams and the TRT performed quite equally. TRT performs better at low recall levels (0-0.2) but has slightly lower MAP than the *s*-grams. However, the transformation rules were generated from a relatively small word-pair list, with general vocabulary from newspaper texts. This list seemed to be insufficient for generating enough high frequency transformation rules. The lack of high quality rules probably affected negatively the TRT technique's translation results.

3.3. CLIR Based on Noisy Comparable Corpora

To test the performance of the CLIR methods presented in this article, a series of German-English CLIR experiments was performed using the TREC 2004 Genomics track (Hersh, 2005) topics. The 50 topics were first translated into German by a native German speaker, who also has a background in molecular biology. Of the 50 topics, 20 were used as training topics for the *s*-gram and corpus-based experiments, and the remaining 30 topics were used in the actual experiments. The MEDLINE collection of 4.6M English medical documents was used as the test collection.

Prior to applying the different query translation approaches, the topic words were lemmatized with the TWOL lemmatizer by Lingsoft Ltd., and stop words were removed. Also, compound words – very common in German – were split with TWOL.

Some of the TREC topics contain key vocabulary that cannot, or should not, be translated (e.g., "Find articles about function of FancD2."), which, it could be argued, make the topics too "easy" for CLIR experiments. However, most topics also contain lots of important query words that should be translated to achieve acceptable CLIR performance (e.g., "Find protocols for generating transgenic mice."). It should be noted that such variation in the difficulty of CLIR queries also occurs in more general domains. In the news domain, for example, queries often have proper names that need not be translated.

As a baseline CLIR approach we used the Utaclir dictionary-based query translation program (Keskustalo et al., 2002). In the experiments, Utaclir employed the GlobalDix German-English dictionary of about 29,000 entries. The dictionary is quite small, but provides a suitable baseline for other methods to improve upon. Utaclir applies the Pirkola query structuring method (Pirkola, 1998) to neutralize the effect of ambiguous translations.

Five translation resources were compared to Utaclir: *s*-grams, FITE-TRT, similarity thesaurus based on a domain specific web corpus (GenWeb), and Google translate and BabelFish machine translation systems. The output of the Utaclir program was also combined with the similarity thesaurus system, *s*-grams and FITE-TRT in order to translate the words which were OOV for Utaclir's dictionary translation in hope of increasing the effectiveness of the combined system. Also the performance of the untranslated source language queries was tested as a baseline for the other techniques. A total of eleven different CLIR approaches were applied:

Ger Untranslated German queries used as a baseline to see whether translation of the queries was needed in the first place.

UC Queries were translated using Utaclir.

GenWeb The German-English genomics web corpus (see Sec. 2.3) used as a similarity thesaurus to translate the queries.

FITE The FITE-TRT method alone.

Google The Google Translate machine translation system.

BabelFish The BabelFish machine translation system.

***s*-grams** 4-grams with CCI $\{\{0\},\{1,2\}\}$. The parameters (including the gram length and the CCI) were selected using the training set. Tanimoto coefficient was used as the proximity measure, and top three three translations for each word were included into the queries. Padding was used at the both ends of the strings, and the queries were structured using Pirkola method.

UC+GenWeb The similarity thesaurus is applied to compensate for the limited dictionary of Utaclir. The queries are first translated with Utaclir, after which words that are OOV for Utaclir are translated with the genomics comparable corpus.

UC+*s*-grams The combined output of UC and *s*-grams (similarly as UC+GenWeb) used as the target query.

UC+FITE The combined output of UC and FITE-TRT (similarly as UC+GenWeb) used as the target query.

The results are presented in Table 5 and in Figs. 2 and 3. Table 5 depicts the performance of the mentioned CLIR approaches in mean average precision (MAP) and precision after 10 retrieved documents (P@10). Figure 2 shows the PR-curves for six approaches. The *s*-gram, UC+*s*-gram, FITE-TRT, and the GenWeb approaches were left out to clarify the figure. However, *s*-gram and UC+*s*-gram followed closely the UC curve, FITE-TRT followed closely the UC+FITE-TRT curve and GenWeb the UC+GenWeb curve. Figure 3 presents the performance of the translation methods in more detail by showing, for each query, the difference in average precision to the untranslated German queries.

Based on the MAPs, the best approach was the combination of Utaclir and FITE-TRT, where the OOV words of the dictionary-based translation were translated using the FITE-TRT technique. Only slightly worse was the combination

Table 5: Results for the German-English genomics CLIR experiments.

Approach	MAP	P@10 docs	Resource[1]
UC+FITE-TRT	0.228[3]	0.393	D+TRT
UC+GenWeb	0.221[2]	0.343	D+CC
GenWeb	0.219[3]	0.361	CC
FITE-TRT	0.208[2]	0.323	TRT
Google	0.207[4]	0.380	MT
BabelFish	0.171	0.277	MT
UC	0.163	0.240	D
UC+*s*-grams	0.162	0.213	*s*-grams+D
s-grams	0.159	0.217	*s*-grams
Ger	0.157	0.197	none

[1] CC = comparable corpus, D = dictionary, TRT = transformation rule, MT = machine translation

[2] Significantly better than Ger, and *s*-grams (Friedman test, $p < 0.01$)

[3] Significantly better than Ger, *s*-grams, and UC+*s*-grams (Friedman test, $p < 0.01$)

[4] Significantly better than Ger, *s*-grams, UC+*s*-grams, and UC (Friedman test, $p < 0.01$)

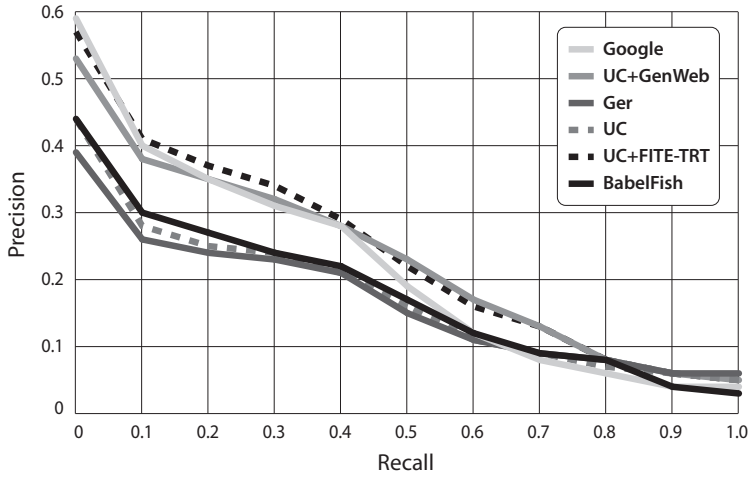
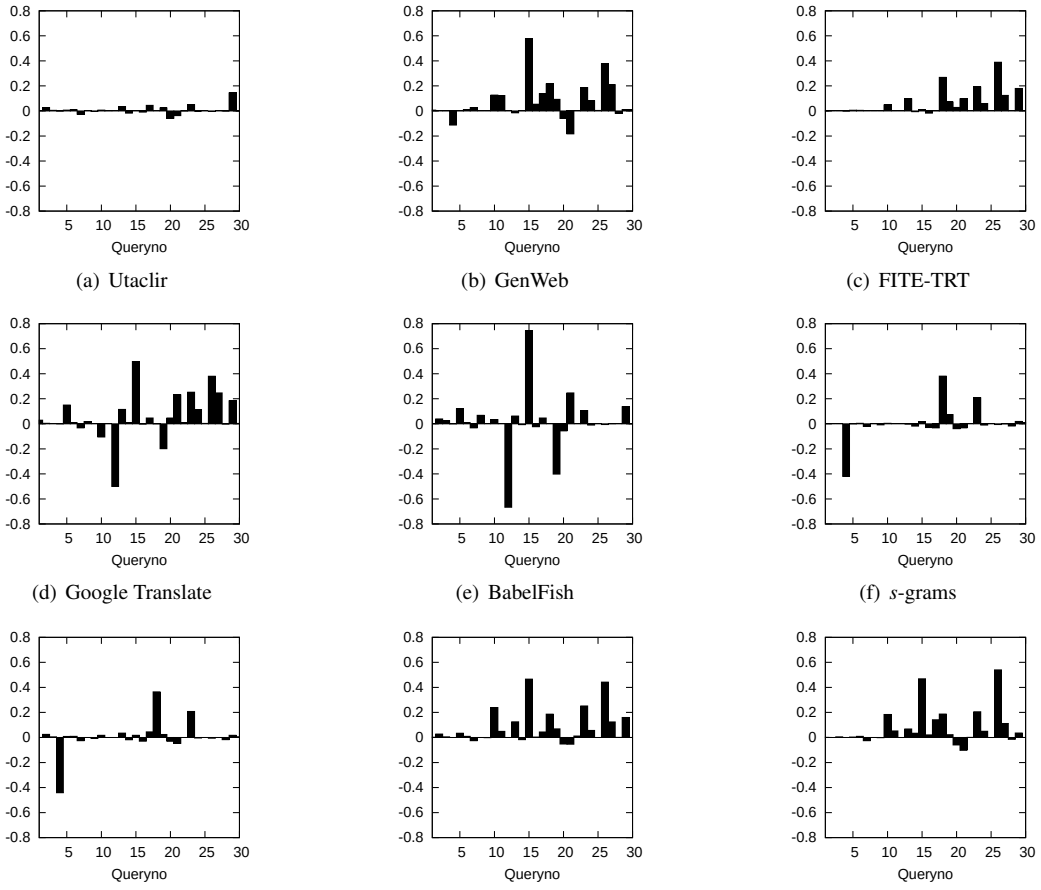


Figure 2: Precision-recall graph over the standard recall levels for the German-English genomics CLIR experiments.



(a) Utaclir (b) GenWeb (c) FITE-TRT

(d) Google Translate (e) BabelFish (f) *s*-grams

(g) UC+*s*-grams (h) UC+FITE-TRT (i) UC+GenWeb

Figure 3: Query-by-query difference in average precision to untranslated German queries.

of Utaclir and GenWeb. These two were closely followed by the GenWeb and FITE-TRT translations alone and then the Google Translate MT system. These techniques seemed to be useful translation resources – the difference between them and the non-translated baseline query was statistically significant at level $\alpha = 0.01$. Based on the PR-graph of Fig. 2, these approaches performed quite equally on the lower recall levels. On higher recall levels Google Translate drops below the UC+FITE-TRT and UC+GenWeb approaches.

Utaclir dictionary translation alone performed only slightly better than the untranslated German baseline. Actually, in 2/3 of the queries, Utaclir simply performed badly – in 8 of the 30 queries it performed worse than the non-translated baseline, and in 12 more queries the average precision was below 0.1.

Utaclir only added little to the performance of FITE-TRT or GenWeb. The combination was more beneficial for FITE-TRT (+0.02 in MAP) than for GenWeb, where the increase in MAP was only 0.002. For about half of the queries, the combinations of UC+FITE-TRT and UC+GenWeb performed similarly to the better of the single techniques involved. UC+FITE-TRT performed better than any of the techniques alone for 9 queries and UC-GenWeb only for 5 queries.

The *s*-gram translation techniques seemed useless if not harmful. Combinations of *s*-grams and Utaclir performed worse than Utaclir alone and the 4-grams performed only slightly better than the non-translated German baseline. The difference was not statistically significant. Rest of the *s*-grams performed worse than the non-translated German baseline.

4. Discussion

Many "real-life" document collections contain noise which is problematic for the traditional IR systems. The noise might be introduced to the target document collection due the spelling errors in the original collection or from other sources such as OCR errors in scanned collections. In CLIR, besides these "normal" sources of noise, also OOV words can be seen as a source of noise, because they cannot be translated. Typical OOV words such as technical terminology and proper names are often cross-lingual spelling variants between languages. Likewise, novel or non-standard expressions, typos, misspellings etc introduce noise in CLIR.

The studies presented in this paper show that data-driven translation techniques can perform well in noisy CLIR tasks. In situations where general translation resources such as machine readable dictionaries or MT systems are unavailable or insufficient, these techniques can be used in query translation instead of the standard resources or in combination with them. Three data-driven translation techniques were presented and evaluated in the article using three case studies.

In the first study, *s*-gram matching was used to translate typical OOV words between 11 language pairs. The *s*-grams were compared to the *n*-grams and it was found that *s*-grams performed better than the *n*-grams in noisy environments, i.e. when the languages were not closely related. The overall performance of the *s*-grams was good suggesting that translating OOV words with *s*-gram matching is feasible.

The second study then compared *s*-gram matching and the TRT technique to dictionary-based translation in query translation between closely related languages. *s*-grams and the TRT technique performed equally well, reaching about 80 % of the dictionary-based translation's MAP. This suggests that data-driven translation techniques are promising for query translation between closely related languages.

Finally, in the third study German queries from TREC 2004 genomics track were translated into English using several data-driven and general translation resources. The data-driven translation resources performed well and the results suggested that general domain translation resources are not very useful for the translation of domain specific queries. Instead, rule-based techniques that rely on word frequencies (FITE-TRT) and techniques based on the use of domain specific web corpora (GenWeb) can be useful resources. Google Translate performed better than the dictionary based translation, probably because the translation is based on large corpora which produces a larger vocabulary than that available for Utaclir (even if still general in domain). BabelFish instead is based on the rule-based Systran MT system, which relies more on "dictionaries" (simple or multiword lexical entries + disambiguation rules) than on corpora, which probably explains why its performance is closer to that of Utaclir than to that of Google Translate.

The *s*-gram translations' performance in the third study was a disappointment. The *s*-grams introduced a lot of noise into the queries, which caused the *s*-gram queries to perform worse than the non-translated baseline query. The noise was due to the fact that based on the training results relatively many (3-5) *s*-gram translations were used in the queries for each source word. It seems that the translation was simply too difficult a task for the *s*-grams – if the correct translations were found at all, they were found at low ranks in the result list (e.g. rank 5) and therefore good performance was not possible.

The case studies presented in this paper concentrated on the CLIR applications of the presented techniques. However, the techniques are data independent, and can thus be applied also to other domains. Especially, the TRT technique and the classified *s*-gram technique could be utilized in the mono-lingual IR to overcome the noise introduced by spelling or OCR errors. The methods could also be applied to historic document retrieval where the OCR errors and spelling variation caused by the language evolution introduce similar problems.

Acknowledgements

The authors wish to thank Professor Kalervo Järvelin from the University of Tampere for his constructive comments on the manuscript. The first author was funded by the Finnish Cultural Foundation, the second author was funded by the Academy of Finland (project name "Focused Web Crawling"), and the third author by the Tampere Graduate School of Information Science and Engineering (TISE). The TWOL was provided by the Lingsoft Ltd. The GlobalDix was provided by Kielikone Ltd.

References

AbdulJaleel, N., Larkey, L. S., 2003. Statistical transliteration for English-Arabic cross language information retrieval. In: Proceedings of the 12th International Conference on Information and Knowledge Management. pp. 139–146.

Baeza-Yates, R. A., Ribeiro-Neto, B., 1999. Modern Information Retrieval. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

Ballesteros, L., Croft, W. B., 1998. Resolving ambiguity for cross-language retrieval. In: Proceedings of the 21st ACM SIGIR Conference. pp. 64–71.

Barˋodal, J., Jörgensen, N., Larsen, G., Martinussen, B., 1997. Nordiska: Våra Språk Förr och Nu. Studentlitteratur, Lund, Sweden.

Cheng, P.-J., Teng, J.-W., Chen, R.-C., Wang, J.-H., Lu, W.-H., Chien, L.-F., 2004. Translating unknown queries with web corpora for cross-language information retrieval. In: Proceedings of the 27th ACM SIGIR Conference. pp. 146–153.

Conover, W. J., 1999. Practical Nonparametric Statistics, 3rd Edition. Wiley, New York, NY, USA.

Fujii, A., Ishikawa, T., 2001. Japanese/English cross-language information retrieval: Exploration of query translation and transliteration. Computers and the Humanities 35 (4), 389–420.

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 (1–2), 99–119.

Hersh, W. R., 2005. Report on the TREC 2004 genomics track. SIGIR Forum 39 (1), 21–24.

Järvelin, A., Järvelin, A., 2008. Comparison of *s*-gram proximity measures in out-of-vocabulary word translation. In: Proceedings of the 15th International Symposium on String Processing and Information Retrieval (SPIRE 2008). Vol. 5280 of LNCS. Springer, pp. 75–86.

Järvelin, A., Järvelin, A., Järvelin, K., 2007. s-grams: defining generalized n-grams for information retrieval. Inf. Process. Manage. 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 Infromation Retrieval Workshop (DIR 2006). pp. 25–32.

Karimi, S., Turpin, A., Scholer, F., 2006. English to persian transliteration. In: Crestani, F., Ferragina, P., Sanderson, M. (Eds.), Proceedings of the 13th International Symposium on String Processing and Information Retrieval (SPIRE 2006). Vol. 4209 of LNCS. Springer, Berlin, Germany, pp. 255–266.

Keskustalo, H., Hedlund, T., Airio, E., 2002. Utaclir : general query translation framework for several language pairs. In: Proceedings of the 25th ACM SIGIR Conference. pp. 448–448.

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 2003). Vol. 2857 of LNCS. Springer, Berlin, Germany, pp. 252–265.

Koolen, M., Adriaans, F., Kamps, J., de Rijke, M., 2006. A cross-language approach to historic document retrieval. In: Proceedings of 28th European Conference on Information Retrieval ECIR'06. Vol. 3936 of LNCS. Springer, pp. 407–419.

Kraaij, W., Nie, J.-Y., Simard, M., 2003. Embedding web-based statistical translation models in cross-language information retrieval. Comput. Linguist. 29 (3), 381–419.

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

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

Pfeiffer, U., Poersch, T., Fuhr, N., 1996. Retrieval effectiveness of proper name search methods. Inf. Process. Manage. 32 (6), 667–679.

Pirkola, A., 1998. The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. In: Proceedings of the 21st ACM SIGIR Conference. pp. 55–63.

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 at: `http://InformationR.net/ir/7-2/paper126.html`.

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. pp. 1043–1049.

Robertson, A. M., Willett, P., 1992. Searching for historical word-forms in a database of 17th-century english text using spelling-correction methods. In: Proceedings of the 15th ACM SIGIR Conference. pp. 256–265.

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

Sheridan, P., Ballerini, J. P., 1996. Experiments in multilingual information retrieval using the SPIDER system. In: Proceedings of the 19th ACM SIGIR Conference. pp. 58–65.

Sperer, R., Oard, D. W., 2000. Structured translation for cross-language information retrieval. In: Proceedings of the 23rd ACM SIGIR Conference. pp. 120–127.

Talvensaari, T., Pirkola, A., Järvelin, K., Juhola, M., Laurikkala, J., 2008. Focused web crawling in acquisition of comparable corpora. Information Retrieval 11.

Toivonen, J., Pirkola, A., Keskustalo, H., Visala, K., Järvelin, K., 2005. Translating cross-lingual spelling variants using transformation rules. Inf. Process. Manage. 41, 859–872.