Tuomas Talvensaari

Comparable Corpora in Cross-Language Information Retrieval

ACADEMIC DISSERTATION

To be presented, with the permission of the Faculty of Information Sciences of the University of Tampere, for public discussion in the B1097 Auditorium of the University on September 26th, 2008, at 12 noon.

DEPARTMENT OF COMPUTER SCIENCES
UNIVERSITY OF TAMPERE

A-2008-7
TAMPERE 2008

Tuomas Talvensaari

Comparable Corpora in Cross-Language Information Retrieval

DEPARTMENT OF COMPUTER SCIENCES
UNIVERSITY OF TAMPERE

A-2008-7

TAMPERE 2008

Acta Electronica Universitatis Tamperensis 779
ISBN 978-951-44-7490-3 (pdf)
ISSN 1456-954X
http://acta.uta.fi

ISBN 978-951-44-7459-0
ISSN 1459-6903

Supervisors: Professor Martti Juhola, Ph.D.
Department of Computer Sciences
University of Tampere
Finland

Academy Professor Kalervo Järvelin, Ph.D.
Department of Information Studies
University of Tampere
Finland

Docent Jorma Laurikkala, Ph.D.
Department of Computer Sciences
University of Tampere
Finland

Opponent: Docent Helena Ahonen-Myka, Ph.D.
Department of Computer Science
University of Helsinki
Finland

Reviewers: Professor Olli Nevalainen, Ph.D.
Department of Information Technology
University of Turku
Finland

Docent Jussi Karlgren, Ph.D.
Swedish Institute of Computer Science
Kista, Sweden

Department of Computer Sciences
FIN-33014 UNIVERSITY OF TAMPERE
Finland

ISBN 978-951-44-7459-0
ISSN 1459-6903

Tampereen yliopistopaino Oy
Tampere 2008

Abstract

Cross-language information retrieval (CLIR) enables users to express queries in a language different from the language of the documents to be retrieved. For example, a Finnish-speaking person could pose a query to a CLIR system in Finnish (the source language) to retrieve documents written in English (the target language). The language barrier is usually crossed by translating the query into the target language, after which the documents can be retrieved with the methods of monolingual information retrieval (IR).

Aligned text collections (corpora) are common query translation resources in CLIR. A *parallel corpus* is a collection where texts in one language are aligned with their translations in another language. The aligned texts of a *comparable corpus* are more loosely related. They are not translations, but share topics and include common vocabulary in the two languages. Both kinds of corpora can be used to train statistical translation models, but parallel corpora are preferred because more dependable translation knowledge can be derived from them. However, parallel corpora do not exist for all language pairs and domains. Hence, it is sometimes necessary to resort to noisier comparable corpora.

This thesis proposes new methods for the acquisition, alignment, and employment of comparable corpora. The acquisition method is based on language-aware focused web crawling, where web content written in specific languages and discussing specific topics of interest is obtained by employing the hyper-link structure of the web. In the alignment phase, the source language documents are used as CLIR queries to retrieve target language documents. The similarity of the query to the documents, and various other factors, are used as evidence to form alignments between the source and target language documents.

The constructed corpora were employed in query translation as a cross-language *similarity thesaurus*, a structure where target language words are ranked based on their similarity with a source language word that is given as input. The highest ranking words are assumed to be either translations of the input word or related to it in some other manner.

The methods were evaluated with extensive IR experiments that covered different language pairs, domains, and test data. The proposed CLIR approach was combined with approaches based on bilingual dictionaries. The combined approaches outperformed pure dictionary-based translation. In addition, the comparable corpus translation performed better in domain-specific CLIR than translation utilizing high-quality parallel corpora. This suggests that the proposed methods are particularly useful in domains where CLIR resources are scarce.

Acknowledgments

Foremost, I wish to thank my supervisors, Professor Martti Juhola, Ph.D., Academy Professor Kalervo Järvelin, Ph.D., and Docent Jorma Laurikkala, Ph.D., for their invaluable expertise, encouragement and support during the project.

The Department of Computer Sciences, headed by Professor Jyrki Nummenmaa, Ph.D., has been an enjoyable working environment. The department administration and senior collegues have made it possible to concentrate on the actual research work instead of bureaucracy. Of my younger collegues, I would like to thank the other members of the Data Analysis and Research Group (DARG), especially Jyrki Rasku, M.Sc., who helped in the evaluation process of the first publication of this thesis. The department's floorball team has provided much-needed recreation and companionship.

I also wish to thank the staff at the Department of Information Studies, especially Eija Airio, Heikki Keskustalo, and Ari Pirkola, who generously have provided me with their knowledge and technical resources.

This thesis was funded by the Tampere Graduate School in Information Science and Engineering (TISE), of which I am very grateful. Thanks also go to the Oskar Öflund Foundation and Emil Aaltonen Foundation for financial support.

Finally, I would like to thank my wife Katariina, and our daughters Maija and Inari for their love and support during these busy years.

Tampere, September 2008
Tuomas Talvensaari

Contents

Glossary

CLEF	Cross-Language Evaluation Forum.
CLIR	Cross-Language Information Retrieval.
Cocot	Comparable Corpus Translation program. Cocot uses an aligned corpus as a cross-language similarity thesaurus.
FITE-TRT	Frequency-based Identification of Translation Equivalents received from Transformation Rule based Translation.
GenWeb	The genomics WWW collection, built for Publication IV. The collection consist of English, German, and Spanish documents.
InQuery	IR system based on the inference network model of IR. *InQuery language* refers to the query language of the system.
IR	Information Retrieval.
JRC-Acquis	A parallel corpus consisting of legislative documents of the EU (Steinberger et al., 2006). Often referred to in the text as "JRC".
MAP	Mean Average Precision.
MT	Machine Translation
OOV	Out Of Vocabulary
QE	Query Expansion.
RATF	Relative Average Term Frequency, a measure for the discrimination power of a word.
TREC	Text REtrieval Conference.

TWOL	A word form normalization program based on the two-level morphological model by Koskenniemi (1983).
Utaclir	A dictionary-based query translation program developed in University of Tampere (Keskustalo et al., 2002).

Publications

I. Talvensaari, T., Laurikkala, J., Järvelin, K., and Juhola, M. (2006). A study on automatic creation of a comparable document collection in cross-language information retrieval. *Journal of Documentation* 62(3), 372–387.

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

III. Talvensaari, T., Juhola, M., Laurikkala, J., and Järvelin, K. (2007). Corpus-based CLIR in retrieval of highly relevant documents. *Journal of the American Society of Information Science and Technology* (JASIST) 58(3), 322–334.

IV. Talvensaari, T., Pirkola, A., Järvelin, K., Juhola, M., and Laurikkala, J. (2008). Focused web crawling in the acquisition of comparable corpora. *Information Retrieval* 11(5), 427–445.

V. Talvensaari, T. (2008). Effects of aligned corpus quality and size in corpus-based CLIR. In Ruthven, I. et al. (Eds.) *Advances in Information Retrieval: Proceedings of the 30th European Conference on IR Research, ECIR 2008*. Lecture Notes in Computer Science, vol. 4956, pp. 114–125. Springer-Verlag.

Chapter 1

Introduction

Information retrieval (IR) aims to provide means to find relevant documents to users' information needs. In a typical IR system, a user expresses his information need as a query, and the system searches a database for documents that are relevant to the query.

After the advent of WWW, IR systems have become crucial for practically every walk of life – business, education, entertainment, etc. – and the currently available WWW search engines answer users' needs more or less effectively. The amount of information available, through web or, e.g., corporate intranets, has exploded, and information is provided in an increasing variety of languages. Thus, there is an increasing demand for IR systems that can somehow cross language boundaries. With such systems, one could retrieve documents of various languages with a query expressed in only one language.

Cross-language information retrieval (CLIR) aims to achieve this, i.e., it aims to find relevant documents written in a language different from the query. The query language is referred to as the *source language*, and the language of the documents as the *target language*. The usual CLIR approach is to translate the query into the target language, and use the translated query to retrieve target language documents. In a typical CLIR usage scenario, the source language is the native language of the user, while the target language can be a language in which the user has only moderate skills. The user may be uncomfortable producing text in the target language – even typing a short query may be burdensome – while he may be able to read the retrieved documents. An IR system capable of cross-langauge retrieval would clearly be helpful for such a user.

The two most commonly used sources of query translation knowledge in CLIR are machine-readable dictionaries and multilingual corpora (Oard and Diekema, 1998; Kishida, 2005). In dictionary-based translation, source

language query words are replaced by their translation equivalents in a bilingual dictionary. Although straightforward, this approach has its problems, mainly untranslatable query keys (i.e., words missing from the dictionary) and translation ambiguity, meaning difficulty of choosing among translation alternatives (Pirkola et al., 2001a). For example, the Finnish word *kuusi* has two possible translations in English, *spruce* and *six*. Without context information, it is impossible to say which one is the correct translation.

In corpus-based methods, the translation knowledge is obtained statistically from the applied multilingual *corpus*, i.e. a collection of text. The corpora can be *aligned* or *unaligned*, depending on whether documents of the source language have been mapped to similar counterparts in the target language collection. Further, the corpora can be categorized based on the relatedness of the texts: a *parallel corpus* is a collection where pieces of source language text are mapped to their exact translations in the target language. For instance, the body of EU legislation is a parallel corpus, because the same laws are written in every official EU language. In a *comparable corpus*, on the other hand, the texts are not translations of each other, but related topically (Sheridan and Ballerini, 1996). The aligned documents can be, e.g., news articles about the same events, written in different countries.

Parallel or comparable corpora can be used to obtain translation knowledge because related cross-lingual word pairs appear in similar contexts in such collections. For example, words like *vaali* (*election*) and *äänestää* (*to vote*) probably appear in Finnish articles about the US presidential election. Similarly, corresponding words will probably appear in Swedish articles about the same events. Naturally, the problem of missing vocabulary also affects corpus-based translation – one cannot reliably translate football vocabulary with a parallel corpus consisting of EU's agricultural legislation. That is, the domain of the applied corpus has to match that of the translated queries.

The most reliable translation knowledge can be obtained from large parallel corpora. However, although numerous parallel corpora exist (see, e.g., Steinberger et al., 2006), they usually cover some rather general domain, e.g., legislation or the news domain. In more special domains (e.g., agriculture or genomics) one may have to resort to noisier comparable corpora. Moreover, these special domains have shortage of other CLIR resources as well: general-purpose dictionaries do not cover most of the technical vocabulary of such domains. For these reasons, the acquisition and use of comparable corpora remains a valid field in CLIR.

The study at hand concentrates on corpus-based methods in CLIR. It aims to answer the following research questions:

1. Is it possible to build an effective aligned comparable corpus from two

collections that are connected only by the domain they represent? The collections could be, for example, newspaper reports written in Finland and Germany in the same time period, or, Swedish and English Wikipedia pages about similar topics.

2. How should this alignment be done? What different indicators of similarity between a source language text and a text in the target language collection should be used?

3. How should the comparable corpus be applied in query translation?

4. How well does a CLIR system based on aligned comparable corpora perform compared to other translation approaches? How does such system perform with different languages that vary in, e.g., inflectional complexity? Why does such system perform as it does?

5. How should comparable corpus query translation be combined with other query translation methods?

6. How does such system manage in retrieving highly relevant documents, compared to other systems?

7. Could an aligned comparable corpus be mined from the web?

8. How does the domain of an aligned comparable corpus affect the performance of a system that uses the corpus as a translation resource?

9. How does the size, on one hand, and the quality of the alignments (that is, similarity of the aligned documents), on the other, affect translation quality?

This study consists of five publications, which are now briefly introduced (for a more in-depth summary of the publications, see Chapter 4).

Publication I The first publication introduces a novel way to create an aligned comparable corpus from two independent text collections in different languages. The publication addresses the first two research questions in the preceding list.

Publication II In this publication, the method for creating comparable corpora is further developed, and the created corpus is applied in query translation. The Comparable Corpus Translation program (Cocot) is introduced. The publication considers the questions 1–5.

Publication III In the third publication, the system proposed in the earlier studies is used in Finnish-Swedish CLIR. Further, non-binary relevance assessments (see Section 2.3.3) are used in the experiments to find how the system manages in retrieving highly relevant documents. The publication considers the questions 3–6.

Publication IV In the fourth publication, a method for obtaining domain-specific comparable corpora from the web is proposed. The method is used to acquire such corpora in the genomics domain, and the acquired corpora are used in query translation. The publication considers the questions 4, 5, 7, and 8.

Publication V In the last publication, the corpora acquired in the fourth publication are used to study the effects of alignment quality, size, and the domain of the translation corpus to CLIR performance. The publication considers the questions 4, 8, and 9.

In short, in this study I propose a new set of methods for acquiring and aligning comparable corpora that can be applied to any domain or language pair. The acquisition phase (Publication IV) involves *focused web crawling* (Chakrabarti et al., 1999), meaning acquiring web content specific to some topic by following web's hyperlink structure. In the alignment phase (Publications I and II), queries are formed of the acquired source language documents, which are then translated using some translation resource(s) available. The translated queries are then run against the acquired target language documents, and alignments are made based on the similarity (or probability) ranking of an IR ranking algorithm.

I also aim to show that employing comparable corpora is profitable from the point of view of CLIR performance. This is done by employing standard methods of the *laboratory model* of IR (see Section 2.3), which implies experimenting with a set of pre-defined search topics, a test document collection, and a set of relevance assessments, i.e., a list of relevant documents in the collection for each test topic. The model does not involve user interaction – the major argument against it – but it does allow for comparing the performance of different retrieval algorithms in a controlled setting. In the experiments, test topics are translated with the Cocot system (Publication II) and with various other translation approaches. The experiments are extensive, covering various language pairs and domains. I also apply non-binary relevance assessments (Publication III) to show that the proposed system manages very well in retrieving highly relevant documents. This is an important result from a user's perspective, since real users rarely have the patience to

go through documents that are only marginally related to the topic of the search. Overall, the results of the experiments are rather promising.

The study at hand suggests that using comparable corpora as source of translation knowledge is profitable in CLIR. This is especially true for domains that lack parallel corpora and other CLIR resources. The corpora can be acquired from the web with relatively few resources.

The introductory part of this thesis is organized as follows: Chapter 2 briefly introduces the field of information retrieval (IR), while Chapter 3 introduces cross-language information retrieval (CLIR). Chapter 4 summarizes the results of the individual articles, and Chapter 5 discusses them in depth and proposes points for future research.

Chapter 2

Information retrieval

According to Baeza-Yates and Ribeiro-Neto (1999), information retrieval (IR)

> deals with the representation, storage, organization of, and access to information items.

IR systems aim to give users access to items that provide information that is *relevant* to an information need which users express as a *query* to the system. The use of an IR system takes place in the context of a *user task* that can be, e.g., exploring previous scientific work for a research project, or trying to come up with a recipe for a dinner.

A more or less clear-cut division can be seen in IR research between the *system-oriented* approach, on one hand, and the *cognitive*, or *user-oriented* approach, on the other (Ingwersen and Järvelin, 2005). This study represents the former, which is centered around developing and evaluating *retrieval models* or algorithms. The problem of system-oriented IR is, in a nutshell, "how to find relevant documents to a query from a database of documents?" The latter approach studies the broader question "how to find information that helps people completing different tasks". The notion of relevance also differs in the two approaches – in the former, relevance is usually a relatively straightforward mapping between queries and documents, whereas in the latter, it is more subjective and situational, related to the user's cognitive state and the situation of the task at hand (Schamber et al., 1990). The two approaches are not contradictory, however. The problem of system-oriented IR can be seen as an important sub-problem of the user-oriented approach.

Figure 2.1 presents the *laboratory model of IR*, the theoretical framework for system-oriented IR research (Ingwersen and Järvelin, 2005). The components of a working IR system are in the center of the picture, and the

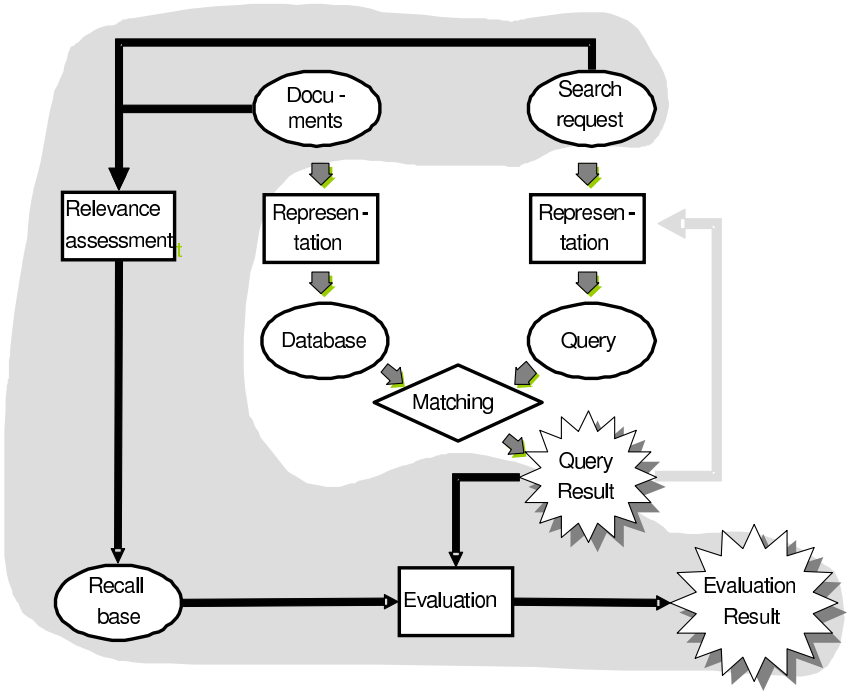

Figure 2.1: The laboratory model of IR according to Ingwersen and Järvelin (2005).

evaluation components are in the shaded area. An IR system applies a *retrieval model* that comprises of the internal representations of queries and documents, and the specification of a matching algorithm. The matching specification defines the way in which the document and query representations are compared to measure the relevance of the documents to the queries.

In this chapter, the *vector space model of IR* is examined next. It is perhaps the best-known IR model, and it is applied in the present study in the Cocot query translation program (see Section 4.2). In Section 2.2, the InQuery query language is presented briefly. It is an example of an advanced query language that can incorporate various different IR models. The language is also applied in the experiments of this study. In Section 2.3, the evaluation methods of the laboratory model are discussed. In the last section of the chapter, IR issues that arise from natural language, are discussed.

2.1 Vector space model of IR

IR systems aim to "predict" whether information items are relevant to a user query. To be able to make this prediction, an IR system has to have some premises about how different characteristics of the items correspond to relevance. For example, IR systems usually assume that the more frequently a word appears in a document, the better that word describes what the document is about. Consequently, when a word appears in a query, the documents where the word appears frequently are considered more relevant than other documents. As a further example, a web search engine might make assumptions about the importance of web pages based on the number of incoming hyperlinks to that page. These premises, whether explicated or not, constitute the retrieval model that the system applies.

Retrieval models can roughly be divided into exact match models and best match models (Belkin and Croft, 1987). Exact match models, such as models based on Boolean logic, return only documents that exactly match some well-defined query. Best match models, such as the vector space model and the *probabilistic model of IR*, on the other hand, can return documents that only partly correspond to the query.

As noted earlier, a retrieval model consists of three factors: document representation, query representation, and a matching algorithm (see Figure 2.1). In the vector space model, documents and queries are represented by vectors whose elements represent document features, that is, words, phrases etc. The relevance of a document to a query is measured by the *cosine similarity* of the document and query vectors. This definition of relevance constitutes the matching specification of the model. The following presentation of the vector space model is based on a text by Salton (1988), who invented the model.

2.1.1 Document-word matrix

In the vector space model, the documents of a collection form a matrix, where rows represent documents, and columns represent words in the documents:

$$\mathbf{A} = \begin{array}{c} \\ \mathbf{D_1} \\ \mathbf{D_2} \\ \vdots \\ \mathbf{D_m} \end{array} \begin{array}{c} \begin{array}{cccc} \mathbf{T_1} & \mathbf{T_2} & \dots & \mathbf{T_n} \end{array} \\ \begin{bmatrix} w_{11} & w_{12} & \dots & w_{1n} \\ w_{21} & w_{22} & \dots & w_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{m1} & w_{m2} & \dots & w_{mn} \end{bmatrix} \end{array} \tag{2.1}$$

The element w_{ij} of this *document-word matrix* (or *document-term matrix*) represents the *weight* of the jth word of the collection in the ith document

of the collection. As can be seen, there are n distinct words ($1 \leq j \leq n$) and m documents ($1 \leq i \leq m$) in this collection. The ith document of the collection is represented by the vector

$$\mathbf{D_i} = (w_{i1}, w_{i2}, \ldots, w_{in}).$$

Similarly, a query formulated by the user can be expressed as the vector

$$\mathbf{Q} = (q_1, q_2, \ldots, q_n),$$

where q_i is the weight of the ith word of the collection in the query.

In the vector model, the relevance of the document D_i to the query Q can be defined as the similarity of the vectors representing them, which can be calculated, for example, with the cosine of the angle between the vectors, that is,

$$\mathrm{sim}(\mathbf{Q}, \mathbf{D_i}) = \frac{\sum_{j=1}^{n} q_j \cdot w_{ij}}{\sqrt{\sum_{j=1}^{n} q_j^2 \cdot \sum_{j=1}^{n} w_{ij}^2}}. \tag{2.2}$$

For any query and document vector, it holds that $0 \leq \mathrm{sim}(\mathbf{Q}, \mathbf{D_i}) \leq 1$. In a document ranking system, this similarity would have to be calculated between the query and all of the documents in the collection. After the calculation, the documents would be sorted according to the similarity score. The resulting rank of documents would then be presented to the user.

2.1.2 The tf.idf weight

What, exactly, are the weights in the document-word matrix? In a straightforward solution, they could be the frequencies of the words in the documents, that is, the jth word appears w_{ij} times in the ith document. In an even more straightforward case, the weights would be binary – 1 when a word appears in a document, and 0 in the other case. In both cases, for any reasonably-sized collection, most of the elements of the matrix will be zero, since most of the words in a collection appear in a relatively few documents, and conversely, a document usually contains only a small portion of the words of the collection.

Usually, though, the weights are more elaborate, and they are based on notions of the correlation of word frequency and its "importance" to the document's topic. Salton and Buckley (1988) define three factors relating to word frequency that should be taken into account in a word weighting scheme:

1. Words that frequently appear in a document most likely describe the topic of the document.

2. Words that have a good discrimination value (see Section 2.4.2) should be rewarded. Conversely, words that are bad discriminators, that is, appear in a lot of documents, should be penalized.

3. Longer documents have more distinct words, and words appear more frequently in long documents than in shorter documents. This does not, however, mean that long documents are *a priori* more relevant than shorter documents.

Correspondingly, Salton and Buckley proposed a weighting scheme that comprised of three components:

1. *Term frequency*, tf_{ij}, is the number of times the jth word appears in the ith document.

2. *Inverse document frequency* (idf) is inversely proportional to a word's *document frequency*, df_j, that is, the number of documents the jth word appears in.

3. The *normalization factor* normalizes the weight in proportion to the length of the document. In the classic vector space model, the length of the document vectors serves as the normalization factor. The Equation 2.2 can be written as

$$\text{sim}(\mathbf{Q}, \mathbf{D_i}) = \frac{\sum_{j=1}^{n} q_j \cdot w_{ij}}{\|\mathbf{Q}\| \cdot \|\mathbf{D_i}\|}, \tag{2.3}$$

   where $\|\mathbf{Q}\|$ and $\|\mathbf{D_i}\|$ are the norms – or the Euclidean lengths – of the query and document vectors.

A typical *tf.idf* weighting scheme (without the normalization component) would be, for example,

$$w_{ij} = tf_{ij} \cdot \log \frac{m}{df_j}, \tag{2.4}$$

where m is the number of documents in the collection.

2.1.3 Pivoted document length normalization

As noted earlier, the cosine normalization (see Equation 2.3) is a straightforward way to account for the varying document lengths in word weighting. However, in practice, cosine normalization is proven to be "too harsh" on long documents. In other words, it makes the retrieval of longer documents less probable than the probability of their relevance (Singhal et al., 1996).

Consequently, Singhal et al. (1996) proposed a *pivoted document length normalization* scheme. The scheme assumes a *pivot* point in document length; documents longer than *pivot* are retrieved less probably than their probable relevance. Conversely, documents shorter than *pivot* have a higher probability to be retrieved than their probability of relevance. The pivoted scheme "tilts" the normalization function so that documents shorter than *pivot* are normalized more harshly than before, while documents longer than *pivot* are normalized more leniently than in cosine normalization. The pivoted normalization factor for the ith document would be

$$(1.0 - \alpha) + \alpha \cdot \frac{\|\mathbf{D_i}\|}{\overline{\|\mathbf{D}\|}}, \tag{2.5}$$

where α, or *slope* is a parameter of the scheme ($0 < \alpha < 1$), and $\overline{\|\mathbf{D}\|}$ is the average length of the document vectors.

The pivoted scheme can be incorporated into the similarity function of Equation 2.2 in the following way:

$$\mathrm{sim}(\mathbf{Q}, \mathbf{D_i}) = \frac{\sum_{j=1}^{n} q_j \cdot w_{ij}}{\|\mathbf{Q}\| \cdot \left((1 - \alpha) + \alpha \cdot \frac{\|\mathbf{D_i}\|}{\overline{\|\mathbf{D}\|}}\right)} \ . \tag{2.6}$$

The pivoted normalization is applied only to the document vector, while cosine normalization is applied to the query vector. This is done because the length of queries varies considerably less than the length of documents. It should also be noted that applying the pivoted scheme causes the weights to no longer lie between 0 and 1.

2.2 The InQuery query language

The InQuery IR system is based on the inference network model of IR (Turtle and Croft, 1991). In the model, relevance is seen as belief in, or probability, that a document satisfies an information need. Various different document and query representations can be used as "evidence" to infer whether this belief holds. Consequently, the model can incorporate various IR models, e.g., the vector space model or Boolean querying. InQuery's query language reflects this flexibility – it can be used for free-text querying, for strictly structured queries with Boolean and word proximity operators, and anything in between. In this study, only two InQuery operators were used predominantly, namely the *#sum* and *#syn* operators.

InQuery attaches words with a belief value which is approximated by the following modification of the *tf.idf* (see Equation 2.4) weight (Kekäläinen

and Järvelin, 1998):

$$0.4 + 0.6 \cdot \frac{tf_{ij}}{tf_{ij} + 0.5 + 1.5 \cdot \frac{dl_j}{adl}} \cdot \frac{\log \frac{m+0.5}{df_i}}{\log (m + 1.0)} \quad , \tag{2.7}$$

where dl_j is the length of document j, measured in number of unique words, and adl the average document length in the collection. The *#sum* operator is the default operator of free text queries, and it evaluates the belief value of the query as the average belief in the query words. The *#syn* operator causes InQuery to treat the enclosed expressions as synonymous. The belief value of a word enclosed within a *#syn* operator is calculated as

$$0.4 + 0.6 \cdot \frac{\sum_{i \in S} tf_{ij}}{\sum_{i \in S} tf_{ij} + 0.5 + 1.5 \cdot \frac{dl_j}{adl}} \cdot \frac{\log \frac{m+0.5}{df_S}}{\log (m + 1.0)} \quad , \tag{2.8}$$

where S is the set of search keys enclosed within the *#syn* operator, and df_S the number of documents containing at least one key of the set S.

The *#syn* operator facilitates *concept-based querying* in a best-match context. In concept-based querying, the information need is analyzed to recognize the central concepts, or aspects, of the request. The central aspects will be represented by separate *facets* in the query. The recognized aspects are further analyzed to find the linguistic expressions (that is, words or phrases) that define the concepts. In the third step, the expressions and the conceptual relations are expressed as a query, using the syntax of the query language at hand. These three steps correspond to the *conceptual*, the *linguistic*, and the *occurrence* level of conceptual querying (Järvelin et al., 1996). Originally, concept-based queries were predominant in Boolean IR systems. Later, stronger query structuring was introduced also to best-match models, InQuery being an example.

For a simplified example, let us assume a user who is willing to find documents about nuclear accidents. At the conceptual level, two concepts may be recognized: NUCLEAR and ACCIDENT (capitalized to separate the concepts from their natural language representations). The concept NUCLEAR may be expressed in the linguistic level by the expressions *nuclear*, *atomic energy* and *fission power*; whereas ACCIDENT could be expressed by *accident* or *disaster*. In a Boolean system, the query – or the occurrence level expression – could be formulated as

> (nuclear OR prox(atomic energy) OR prox(fission power)) AND
> (accident OR disaster),

where *prox* is the proximity operator, that is, it matches when the words enclosed in it are found close to each other in a document. The query comprises of two facets which represent the two aspects of the information need. Now, with the InQuery language, the occurrence level expression could be

> #sum(#syn(nuclear #3(atomic energy) #3(fission power))
> #syn(accident disaster)),

where *#n* is the proximity operator of the InQuery language, which allows the words within it to be at most n words apart from each other. In this InQuery query, the expressions representing a concept are marked as synonymous. For comparison, a weakly structured query in the InQuery language would be

> #sum(nuclear atomic energy fission power accident disaster) .

Strong query structuring has been shown to be beneficial in both monolingual IR (Kekäläinen, 1999) and CLIR (Pirkola, 1998).

2.3 IR evaluation

Figure 2.1 presented the laboratory model of IR (Ingwersen and Järvelin, 2005). The evaluation part of the model involves a set of documents – the *test collection* – on one hand; and a set of search requests – *topics* – on the other. The *relevance assessments* link these two sets; for each topic, they consist of a set of pointers to documents of the test collection that are relevant to the topic. The *recall base* is a set of pairs $\langle i, R_i \rangle$, where R_i is the set of relevant documents (or rather, pointers to such documents) for topic number i.

As an example of a test collection, the CLEF (Cross-Language Evaluation Forum) consortium offers a multilingual news article collection for CLIR research. The collection consists of 3 million news documents in 13 languages (Peters, 2006). For example, the English sub-collection consists of news documents by Los Angles Times and Glasgow Herald from the years 1994–1995. In each new annual "campaign", a few dozen new test topics are introduced. An example of a test topic of the CLEF collection is presented in Figure 2.2. Usually, when queries are constructed from the topics, the "narration" part is omitted. Sometimes only the "title" field is used. Further, redundant phrases such as "find documents on" in the example topic are usually removed. The sample topic has 51 relevant documents in the Los Angeles Times collection.

```
<top>
<num> C042 </num>
<EN-title> U.N./US Invasion of Haiti </EN-title>
<EN-desc> Find documents on the invasion of Haiti by U.N./US
soldiers. </EN-desc>
<EN-narr> Documents report both on the discussion about the decision
of the U.N. to send US troops into Haiti and on the invasion itself.
They also discuss the direct consequences. </EN-narr>
</top>
```

Figure 2.2: Example topic from the CLEF collection.

2.3.1 Recall and precision

When a proposed IR algorithm is evaluated, it is applied to either document or query preprocessing, document-query matching, or all of these, depending on the algorithm. A baseline algorithm is also applied to the same part of the system. The query performance of each of the tested methods and the baseline is evaluated by matching the query results to the recall base. Various performance metrics, that are usually based on *recall* and *precision* are used in the evaluation.

Let R be the set of relevant documents for a test topic, and A the set of documents retrieved for the topic by some proposed algorithm. *Recall* is the fraction of the relevant documents that have been retrieved, i.e.

$$Recall = \frac{|R \cap A|}{|R|}.$$

Precision, on the other hand, is the fraction of the documents retrieved that are relevant, that is,

$$Precision = \frac{|R \cap A|}{|A|}.$$

2.3.2 Derived measures

Table 2.1 presents a ranking of retrieved documents after a test query has been executed. The query was formed by applying some IR method to a test topic that has 12 relevant documents in the test collection. For each of the 20 top ranking documents, its relevance to the topic is depicted in the table. Let us examine the result set cumulatively, document-by-document, starting from the top. The highest ranking document is not relevant to the topic, but the document ranked second is. This document corresponds to

Table 2.1: Cumulative recall and precision values for a test retrieval run, $|R| = 12$

Rank	Relevant	Recall	Precision
1	no		
2	yes	0.08	0.5
3	yes	0.17	0.67
4	no		
5	yes	0.25	0.6
6	no		
7	yes	0.33	0.57
8	no		
9	yes	0.42	0.56
10	no		
11	yes	0.5	0.55
12	no		
13	yes	0.58	0.54
14	yes	0.67	0.57
15	yes	0.75	0.6
16	yes	0.83	0.63
17	no		
18	no		
19	yes	0.92	0.58
20	yes	1	0.6

8.33% of all the relevant documents, and 50% of the documents encountered so far have been relevant. That is, at *recall level* 0.08, precision is 0.5. The *average precision* of the query is the mean of precisions at each recall level, i.e., at each relevant document. For each relevant document not retrieved, zero precision is added to the calculation. For this query, $avgp = (0.5+0.67+0.6+0.57+0.56+0.55+0.54+0.57+0.6+0.63+0.58+0.6)/12 \approx 0.58$. The *mean average precision* (MAP) of a test run is average precision averaged over all queries.

Precision at different *document cut-off values* is also a useful metric. For the example query, *precision at 10 documents* (P@10) would be $5/10 = 0.5$, because 5 of the 10 highest ranking documents are relevant. The *R-precision* of a query means precision after $|R|$ documents, which for this query would be $6/12 = 0.5$. Usually, only the average values over all of the queries are presented for these metrics. Precision among the highest ranking documents is important because real users rarely have the patience to go through dozens

Table 2.2: Interpolated precision values for the retrieval run of Table 2.1

Recall	Precision
0.0	0.67
0.1	0.67
0.2	0.63
0.3	0.63
0.4	0.63
0.5	0.63
0.6	0.63
0.7	0.63
0.8	0.63
0.9	0.6
1.0	0.6

of documents to find a relevant one. From this point of view, precision at high recall levels is really not that important, especially if there are a lot of relevant documents. However, there are situations where high recall is important, and accordingly, precision at high recall levels should also be high.

In addition to these single value summaries, precision is often presented at 11 *standard recall levels* $0.0, 0.1, 0.2, \ldots 1.0$. The "real" recall levels (as shown in the third column of Table 2.1) will have to be interpolated to the standard levels. The precision at a standard recall level i is the maximum precision at any real recall level greater than or equal to i. Precision at the interpolated recall levels for the example query are shown in Table 2.2.

When precision at each standard level is averaged over all queries, we can depict the performance of the queries with a graph where precision is plotted against the recall levels. Figure 2.3 presents an example, where the performance of five IR methods is depicted.

2.3.3 Generalized recall and precision

In the previous section, relevance is assumed to be a binary relation between a search request and a document. That is, a document is either relevant or not relevant to a given request. This assumption has been criticized for being unrealistic (Sormunen, 2002; Voorhees, 2001) – in a real search task, a user assesses the documents in a more multi-leveled manner. Moreover, Sormunen (2002) argues that the relevance assessments in traditional IR tests collections – such as CLEF (Peters, 2006) or TREC (Voorhees, 2006) –

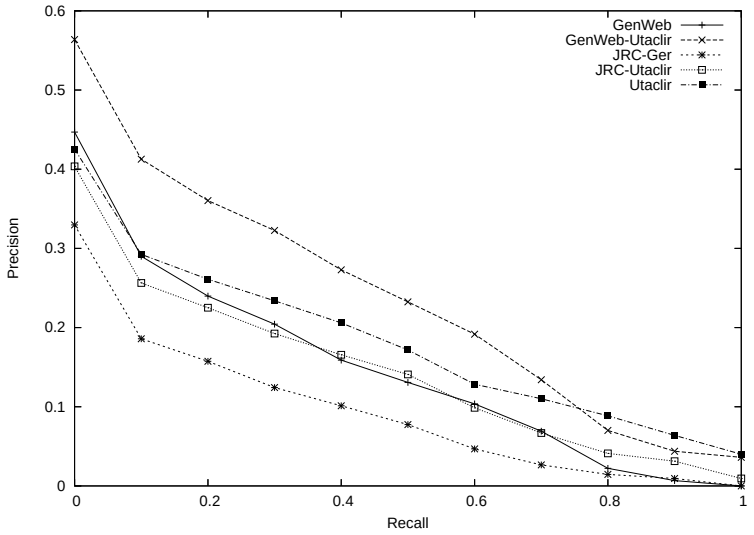

Figure 2.3: Interpolated precision at 11 recall points for five IR approaches.

have been too liberal. In other words, some of the documents judged relevant in such collections are only marginally related to the topic in question. Using such liberal relevance assessments in IR evaluation might skew the results in favor of systems that actually do not perform particularly well.

Consider, for example, two IR systems, A and B. The systems are evaluated in a laboratory experiment. On one of the test topics, the systems perform equally well according to average precision. When the result sets are examined, however, it is discovered that system A has mostly retrieved documents that are only marginally relevant to the topic, whereas system B has managed to retrieve highly relevant documents. For a real user, system B would have been more valuable, but the experiment results do not indicate this.

Graded relevance assessments are thus employed to gain more reliable results in IR evaluation. In the third publication of this study, a recall base where relevance assessments were based on a four-point scale, was used. The scale was introduced by Sormunen (1994), and it is portrayed in Table 2.3.

Using graded relevance assessments calls for performance metrics that can take into account the different relevance levels. One way is to use standard recall and precision, but have separate recall bases for the different relevance levels. In this way, the performance can be examined separately for each level. Alternatively, different relevance threshold levels can be applied. For example, using the above relevance scale, three relevance threshold levels can be defined:

Table 2.3: Four-point relevance scale by Sormunen (1994).

0	irrelevant	The document does not contain any information about the topic
1	marginally relevant	The document only points to the topic. It does not contain any other information, with respect to the topic, than the description of the topic.
2	fairly relevant	The document contains more information than the description of the topic but the presentation is not exhaustive. In the case of a topic with several aspects, only some of the aspects are covered by the document.
3	highly relevant	The document discusses all of the themes of the topic. In the case of a topic with several aspects, all or most of the aspects are covered by the document.

1. *Liberal level*, where documents of relevance levels 1–3 are considered relevant.

2. *Regular level*, where documents of levels 2 and 3 are relevant.

3. *Stringent level*, where only documents of relevance level 3 are considered relevant.

To gain a general picture over all the relevance levels, *generalized recall and precision* (Kekäläinen and Järvelin, 2002) can be used. Let A be the set of documents retrieved from a database D in response to some query, $A \subseteq D$. Further, let $r(d)$ be the relevance score of the document d in relation to some test topic, $0 \leq r(d) \leq 1$. In the four-point scale of Table 2.3, it would be natural to set the scores $0, 0.33, 0.66$ and 1 for the relevance levels 0, 1, 2 and 3, respectively. The generalized recall gR may now be computed as

$$gR = \sum_{d \in A} r(d) / \sum_{d \in D} r(d),$$

and generalized precision gP as

$$gP = \sum_{d \in A} r(d) / |A|.$$

With the generalized recall and precision, it is possible to use the same kinds of derived metrics as with regular recall and precision. Further, the

relevance scores can be adjusted, for example, to give more weight to highly relevant documents.

2.4 Natural language and IR

Various characteristics of natural language cause problems in IR. Imagine a "basic" IR system that users can query with queries consisting of a list of words. For each query, the system would scan the documents in its database and look for words appearing in the query, and then return a list of those documents where such words were found. There are many reasons related to natural language that would cause problems for such a system (Ingwersen and Järvelin, 2005; Pirkola, 1999).

The following list presents an array of such problems, and also some of the proposed solutions. In an IR system, these problems are mostly addressed in the *indexing* stage, that is, when the documents are transformed into the internal representation format of the IR model in question (see Figure 2.1). To successfully match queries against documents, the same operations should also be performed on queries.

Ambiguity Words can be ambiguous, i.e., they may have various different meanings. *Homonyms* are words that are spelled similarly, but have different meanings (e.g., *bear*). *Polysems*, on the other hand, are words that have multiple, but related, senses. In the basic system, an ambiguous query word can match to irrelevant documents that include the word in a sense different from that intended by the user. *Query expansion* can resolve ambiguity by bringing additional search keys that in turn bring more context to the query. Different *word sense disambiguation* methods have also been proposed.

Inflection Inflected forms of the query words are not found by the basic system. *Word form normalization* can be used in IR to resolve problems brought by word inflection.

Compounds If a query word appears as a part of a compound word, it is not found by the system. Different *decompounding* techniques can be used.

Phrases Imagine a query that includes the phrase *computer monitor*. The basic system evaluates the expressions "Many of us sit in front of a computer monitor every day" and "A computer can be programmed to monitor the voltage signals" equally relevant, although the phrase does

not appear in the latter sentence. *Phrase recognition* is widely studied and it can be used to index whole phrases, in addition to single words.

Synonymy Different expressions may refer to the same concept, and documents that use synonyms of the query words are not necessarily retrieved by the example system. Query expansion can be used to include synonymous expressions in queries.

Anaphors Anaphors (such as the word *he* in the expression "This is Bill. He is a plumber.") do not match a query that include the antecedent. Different techniques for *anaphor resolution* have been proposed.

Affixes Prefixes and postfixes hide the root word from the system. Word form normalization can work to strip suffixes.

Varying semantic significance Some words have more descriptive power than others. In English text, words such as *the* and *a* have no semantic significance whatsoever, and documents should usually not be retrieved based solely on such words. Such words can be omitted altogether from indices by using *stoplists*. Also, word weighting techniques, such as the *tf.idf* weight, aim to reward words with high semantic significance.

In the following sections some of the above mentioned techniques – those that are relevant for this study – will be examined more closely.

2.4.1 Word form normalization

In the basic example system, word inflection causes recall to drop, because inflected forms of the query words are not retrieved by the system. However, if the inflected forms in documents and queries would be normalized to a common "root form", the drop in recall could be avoided. There are two main approaches to word form normalization, namely *stemming* and *lemmatization*.

Stemming refers to a language-specific algorithmic process, in which words are stripped of their inflectional suffixes. The Porter (1980) stemmer is probably the most common English stemmer. For example, the words *retrieved* and *retrieves* are stemmed by the Porter stemmer into the root form *retriev*. As can be seen, the root forms are not necessarily "real words". This fact is usually hidden from the user, because stemming only affects the internal representations of documents and queries. Stemmers are also widely available for languages other than English (Porter, 2001).

Lemmatization, on the other hand, transforms inflected words to *lemmas*, i.e., the base form (or the "dictionary form") of a word. The obvious difference to stemming is that lemmatization produces real words. The more fundamental difference is that lemmatizers need a dictionary of the language in question, whereas stemmers are usually based on transformation rules. The performance of a lemmatizer is limited by the size of the dictionary – for example, new technical terms or proper nouns are often missing from dictionaries.

Both approaches can hurt query precision: Greedy stemming can produce common root forms for words that are related only morphologically. For example, *generate* and *generation* are normalized to *generat* by the Porter stemmer. On the other hand, ambiguous word forms can be lemmatized to many base forms, as is exemplified by the Finnish inflected word *hauista*. The word may be an inflection of any of the words *haku* (*retrieval*), *hauki* (*pike*) or *hauis* (*biceps*).

Hull (1996) showed that, in general, stemming is beneficial in English IR. Further, Airio (2006) found that, perhaps surprisingly, stemming achieves comparable performance with lemmatization in monolingual IR, even with morphologically complex languages such as Finnish. However, in CLIR, lemmatization performed better.

Lemmatizers can also be used in *decompounding*, that is, splitting compound words to their constituents. Some languages, such as German and Finnish, are highly compounding, whereas English is an example of a phrase-oriented language. Decompounding can increase recall, because query words may be constituents of related compounds. For example, take a Finnish query that includes the word *karhu* (*bear*), and a document that contains the word *harmaakarhu* (*grizzly bear*). On the other hand, including the unrelated compound constituent *harmaa* (*grey*) in the query may hurt precision.

2.4.2 Frequency-based word selection

As noted earlier, words differ in their descriptive power. For example, in the sentence "Obama surges past Clinton in Democratic race", *Obama* and *Clinton* describe what the sentence is "about", whereas *past* and *in* do not. As early as Luhn (1958) discussed the *discrimination value* of words, and noted that the words that carry the most information, that is, discriminate between documents, are in the middle of the frequency spectrum. That is to say, words that appear very frequently on one hand, and very rarely, on the other, have small discrimination power. Going back to the example sentence, *in* and *past* are clearly more frequent in English text in general than *Obama* or *Clinton*.

Very frequent words are often omitted all together from an index by applying *stoplists*, meaning lists of words to be "stopped", that is, excluded (Fox, 1990). As for the rare words, it has been noted that in large document collections, most of the unique words in the collection appear in but a few documents. For instance, in one of the collections used in this study, the CLEF (Peters, 2006) L.A. Times collection, 36% of words appear only once in the collection. Such anomalies are called *hapax legomena* (Greek for "read only once"), which may be rare proper nouns, misspellings or errors brought by optical character recognition (OCR), etc. Salton and McGill (1983) suggest excluding such rare words from the index.

Removing common and rare words are mainly done to save computational resources. However, in document retrieval, stopword removal can hurt query performance. Consider a database of the work of Shakespeare for which stopword removal has been applied. The famous phrase "to be or not to be" consists entirely of stopwords, and hence a query consisting of the phrase would return an empty result!

Word weighting is a more elegant way to account for varying descriptive power. For example, the *tf.idf* weight (see Section 2.1.2) penalizes frequently appearing words. The *tf.idf* weight is a document-specific measure. The RATF value (*relative average term frequency*), proposed by Pirkola et al. (2001b), is an example of a collection-wide measure for discrimination value. The RATF value of the jth word of the collection is calculated as

$$\mathrm{RATF}_j = (cf_j/df_j) \cdot \mathrm{C}/\ln(df_j + \mathrm{SP})^p,$$

where df_j is the document frequency of the word; cf_j its *collection frequency*, that is, the number of times the word appears in the collection; SP and p are collection-specific parameters. The constant C scales the product to a more convenient value, C = 1000 was used in this thesis.

2.4.3 Query expansion

A document may discuss a topic with various alternative concepts and synonymic expressions. Therefore, it is hard for users to formulate queries that would cover all of the possible vantage points to the topic (Kekäläinen, 1999). Furthermore, user queries usually consist of but a few words. Such short queries can result in low recall, because documents using alternative vocabulary are not retrieved. *Query expansion* (QE) is a technique where additional search terms are added to queries in order to enhance recall (Efthimiadis, 1996).

QE keys can be added by using a *thesaurus*, i.e. a structure where relations between words and concepts are presented. Traditionally, thesauri are

manually created, and manually employed by users or information service experts. *Similarity thesauri* can be part of IR systems to facilitate automatic QE. Similarity thesauri are created automatically by learning co-occurrence data from a large collection of text (Qiu and Frei, 1993; Jing and Croft, 1994).

Relevance feedback is a QE technique where new query keys are automatically extracted from relevant documents. The relevant documents can be picked after an initial search either manually by the user, or automatically by the IR system. In the latter case, in which the system assumes the highest ranking documents to be relevant, the process is called *local feedback* or *pseudo relevance feedback*. Pseudo relevance feedback can be elegantly incorporated into the vector model: for example, a centroid vector of the relevant documents can be calculated and added to the original query vector (Buckley et al., 1994). This causes the query vector to "move" towards the relevant documents in the document space.

Keskustalo et al. (2006) propose a technique where top ranking documents "vote" for expansion keys. The "candidates" are the semantically most significant words from each relevant document. The significance is calculated with the RATF value (see Section 2.4.2). This QE technique, the "RATF-based pseudo relevance feedback" is used in this study (see Publication II).

Views differ on whether QE actually improves IR performance significantly. Kekäläinen (1999) found that strong query structuring is vital for QE success, especially when a large number of new keys are added. She also found that using the synonym structure of the InQuery language (see Section 2.2) to represent facets is advantageous in QE.

Chapter 3

Cross-language information retrieval

Cross-language information retrieval (CLIR) aims to find relevant documents to a query that is expressed in a language different from the documents. The language of the query is referred to as the *source language*, and the language of the documents as the *target language*. Historically, CLIR research started with the pioneer work of Salton (1969), but it took until the late 1990's for CLIR to really establish itself (see Grefenstette (1998a) for early work).

The CLIR process differs from the IR process only in the respect that the language boundary must somehow be crossed. This can be done by translating either the queries to the target language, or the documents to the source language. In CLIR, the former approach is more common. Query translation is simpler than document translation, because queries are usually much shorter than documents. Also, syntactic knowledge need not be considered in query translation, which makes it possible to use rather simple algorithms and resources. Furthermore, the translated documents would have to be indexed before retrieval. However, the brevity of the queries also may cause problems, because the lack of context in typical queries increases *translation ambiguity*. Another argument in favor of document translation is that the translation of documents can be made off-line, unlike query translation. (Grefenstette, 1998b; Kishida, 2005). This study, however, concentrates on CLIR based on query translation.

Basically, therefore, CLIR can be viewed as "normal" IR which involves additional steps in the query processing phase, and it can be set within the laboratory IR framework (see Figure 2.1). This also implies that CLIR shares the same natural language-related problems with IR, as well as problems stemming from query translation.

CLIR can be useful in various different usage scenarios and for users with

varying language skills. Users with moderate or non-active skills in the target language may be able to understand text in the target language, but are often unable to produce it, i.e., express queries with it. Such users could benefit from a CLIR system based on query translation. Further, a fluently multilingual person could benefit from a CLIR system where the query would be translated into more than one language. The system could save him the time and labor of having to produce queries for each desired language. Such a system could produce separate result sets for each language, or it could merge the results into one list of documents. Merging is a non-trivial problem of multilingual IR, and it is not addressed in this thesis.

A CLIR system could be helpful also for a user with little or no skills in the target language. Imagine, e.g., an inventor who would like to know if inventions similar to his exist at all in the world. He could make a cross-lingual web search, and get documents written, perhaps, in a totally unfamiliar language. He could examine the documents, e.g., by looking at pictures and other language-independent clues. Alternatively, the retrieved documents could be translated with a MT system.

Airio (2008) experimented with users of varying target language skills and found out that cross-lingual web search based on query translation is helpful, particularly for users with non-active or moderate skills in the target language. This group of users got better retrieval results with translated queries than with queries that they produced directly in the target language. However, the quality of the translation resource (i.e., the dictionary in this case) also played an important part in the results.

In the following sections, different CLIR query translation approaches will be reviewed. The main approaches are dictionary-based translation, machine translation, corpus-based methods, and cognate matching (Oard and Diekema, 1998; Kishida, 2005).

3.1 Dictionary-based CLIR

In dictionary-based translation, a machine-readable bilingual dictionary is used to replace the source language query words with their target language counterparts. Pirkola et al. (2001a) listed problems of dictionary-based translation. The problems are mostly common to all CLIR approaches:

1. Out-of-vocabulary (OOV) words. No dictionary is complete – especially technical terms, proper nouns, and novel expressions are often missing. The same goes for compound words in compound-prone languages, such as German and Finnish. In some cases, the target language entirely lacks an adequate translation for a source language word.

For example, Finnish has a wide array of terms describing snow of different variety. Translating them with the English word *snow* loses some of the meaning in them.

2. Lexical ambiguity in source and target languages. A source language word can have many senses, which translate to different words in the target languages. For example, the Finnish word *maali* can be translated as *goal* or *paint* in English. This phenomenon is called *translation ambiguity*. Further, the translated words may be ambiguous in the target languages. *Goal* can be a sports-related term (as *maali* usually is in Finnish), as well as having the sense *purpose* or *aim*. Thus, ambiguity can increase many-fold in the translation process.

3. Inflected source language words. Dictionary entries are usually in their base forms, whereas queries may have inflected words.

4. Phrase identification and translation of phrases. Usually, phrases do not occur as head entries in dictionaries, and consequently are OOV. Even if phrases are found in a dictionary, they must first be recognized and extracted from the query.

The last two problems are common to monolingual IR, and can be resolved with similar approaches. For example, query words will have to be stemmed or lemmatized before matching them with the dictionary. Similarly, phrase recognition and decompounding can be performed prior to translation. The first two problems, OOV words and translation ambiguity, are inherent to CLIR, and are the main reasons for CLIR performance to be on average significantly worse than monolingual IR. Dictionary-based translation can be viewed as the baseline method of CLIR. The following sections review approaches to solve the above problems characteristic of dictionary-based translation and CLIR in general.

3.2 Cognate matching

The OOV problem can be eased by employing *cognate matching*. Many proper nouns and technical terms – i.e., words most often missing from dictionaries – are very similar across languages. Often they are *cognates*, meaning words with similar etymological background. This is true, for example, for the Finnish-English word pair *informaatio–information*, or for the German-English pair *konstruktion–construction*.

When cognate matching is applied to query translation, OOV query words are matched against target language words which have been extracted from

a target language corpus. The most similar target language word, or a few of the most similar, can then be chosen as "translations". The similarity can be calculated, e.g., with edit distance or s-gram matching. s-grams (Järvelin et al., 2007) are a generalization of n-grams. Unlike n-grams, s-grams allow skipping over characters when character strings are decomposed into gram sets. For example, the string *informaatio* decomposes into digrams $\{if, no, fr, om, ra, ma, at, ai, to\}$ when one character is skipped. The distance between the gram sets of two strings can be measured, e.g., by the Jaccard distance. s-grams have outperformed regular n-grams in CLIR experiments (Pirkola et al., 2002).

A more advanced way to translate cognates is to apply transformation rules that capture stereotypical variation between languages. For example, the letter k at the beginning of a German word often changes to c in English (e.g., *konstruktion* $\rightarrow$ *construction*). Pirkola et al. (2003) mined such rules from bilingual dictionaries. However, translation based solely on transformation rules can produce a lot of nonsense words and other bad translations. Accordingly, Pirkola et al. (2006) used frequency data of the target language to choose the most probable translation candidates. Their technique, FITE-TRT (Frequency-based Identification of Translation Equivalents received from Transformation Rule based Translation), is also applied in this study. The technique requires much more resources than simple cognate matching, and is computationally more intense, but produces more accurate results.

3.3 Machine translation

Machine translation (MT) aims to provide human-readable translations of natural language texts. This makes it arguably a much harder task than query translation, since queries can be translated word-by-word, and the translations need not be seen by the user. However, since MT has to choose "the correct" translation alternative for each word, it involves word sense disambiguation, which is lacking in simple dictionary translation. For short queries, though, there is often too little context for MT systems to infer the correct translation alternative. Early results (Ballesteros and Croft, 1998) indicated that MT performs worse than dictionary-based methods in CLIR.

3.4 Corpus-based CLIR

In corpus-based CLIR approaches, the translation knowledge is derived from parallel or comparable corpora. As noted in Chapter 1, parallel corpora are preferred because they provide more accurate translation knowledge. However, because of the scarcity of parallel corpora, comparable corpora are often used in CLIR.

As noted in Chapter 1, the aligned texts of a comparable corpus are not translations of each other, but related topically (Sheridan and Ballerini, 1996). For example, consider collections of articles from a Finnish and a Swedish newspaper from the same time period. A lot of the topics and events covered in the Finnish newspaper would also be covered in the Swedish newspaper. A comparable corpus could be created by finding, for each article in the Finnish collection, a document in the Swedish collection that discussed the same event or topic. In this case, the Finnish collection is the *source collection* of the comparable corpus, and the Finnish articles form the set of *source documents*. By contrast, the Swedish collection is the *target collection* that consists of *target documents*. (Of course, the roles of the collections could be reversed.)

It is not realistic to expect to find a pair for every source document, because not all of the events and topics covered in the Finnish newspaper would be covered in the Swedish one. Hence, the number of document pairs in the comparable corpus would be smaller than the number of source documents. Note also that the alignments need not be pairs of documents: a source document could also be aligned with a set of similar target documents.

In CLIR literature, comparable corpora sometimes refer to unaligned collections (see, e.g., Rapp (1999)). To continue with the above example, the Finnish and Swedish collections would form a comparable corpus as such, without the alignments, by this definition. In this thesis, though, comparable corpora are aligned collections.

There are various ways to utilize parallel or comparable corpora. In *cross-language pseudo relevance feedback*, the source language query first retrieves documents monolingually from the source language documents of the aligned corpus. Then, the top n documents assumed to be relevant are exchanged with their alignment pairs, and QE keys are extracted from them. The obtained keys are then used as the target language query. If the aligned texts are parallel, the target language query most likely contains translations of source language query keys, and, importantly, also good target language expansion keys. Davis and Dunning (1995) pioneered this approach.

In a much similar vain, Ballesteros and Croft (1998) employed a parallel corpus to disambiguate dictionary translation. If a word with multiple

dictionary translations appeared in a query, they retrieved source language documents with the original query, and replaced a few dozen highest ranking documents with their alignment pairs. Then they ranked the target document words according to their discriminative power. The translation alternatives that had the highest rank were selected to appear in the target queries.

Another approach in corpus-based CLIR is to learn cross-lingual word associations from parallel or comparable corpora, in effect creating a *cross-language similarity thesaurus* (Sheridan and Ballerini, 1996; Braschler and Schäuble, 1998). This is much like a regular automatic association thesaurus (see Section 2.4.3), except that an aligned cross-lingual corpus is used as training data.

The idea is to calculate similarity scores between a source language query word and words in the target language documents of the aligned corpus – the most similar target language words are assumed to be translations or related words. This approach was first proposed by Sheridan and Ballerini (1996), and later employed, e.g., by Braschler and Schäuble (1998) and Yang et al. (1998). Molina-Salgado et al. (2002) applied a similarity thesaurus in bigram translation, i.e., they translated pairs of consecutive words in addition to single words. This was done in order to emulate phrase translation, but the results were poor. The similarity thesaurus technique is applied in this study and it is further discussed in Section 4.2.

Several studies have seen the query translation problem as the problem of maximizing the *translation probability* $p(t|s)$, where s is a source language word, and t a target language word. In other words, the target language word that has the highest probability of being the translation of s is sought for. The translation probabilities can be estimated, e.g., by using a dictionary, or by learning from aligned corpora using the machine translation models proposed by Brown et al. (1993), i.e., the IBM models 1–5. The simplest one of these models, Model 1, is the most often used in CLIR. Model 1 does not concern with the order of words in sentence translation, which is usually adequate in CLIR, because query translations need not be syntactically correct sentences (Nie et al., 1999). The translation probabilities can further be incorporated into probabilistic IR models (Xu et al., 2001). In such models, the entire CLIR process (not only query translation) is elegantly portrayed by a probabilistic model.

An unaligned corpus can also be utilized as a translation resource. Rapp (1999) and Fung and Yee (1998) proposed a method where context vectors were calculated for each word in source and target corpora. A context vector of a word includes other words that appear frequently near the word. A subset of the words in the source language vectors are translated with a "seed

dictionary". Then, to translate a source language word that is not found in this seed dictionary, its context vector is compared to the target language context vectors with some distance metric. The target word with the most similar vector is assumed to be the translation. The main difference of this approach to methods that employ aligned corpora is in the context that is used to define the "semantic space" of the words. In aligned corpora, the aligned entities (i.e., documents, paragraphs or sentences) provide the context, whereas in Rapp's approach, the context is provided by the neighboring words.

3.5 Obtaining aligned corpora

Early corpus-based CLIR studies applied parallel corpora such as United Nations documents (Davis and Ogden, 1997; Ballesteros and Croft, 1998), or the Canadian Hansard corpus of Canadian parliament proceedings in English and French (Nie, 1998). The parallel texts were aligned with algorithms such as the one proposed by Gale and Church (1991).

From quite early on it was evident that the freely available parallel corpora were not sufficient: their lexical coverage was limited, and they were not available for all language pairs. Some researchers turned their attention to the fast-burgeoning WWW (Kilgarriff and Grefenstette, 2003). Resnik (1999) and Nie et al. (1999) proposed techniques for finding parallel content on the web. They used language-independent clues, such as similar structure and length, to detect whether two web pages were parallel. Yang and Li (2004) and Shi et al. (2006) further advanced these techniques. Another approach to use the web as a translation corpus is to obtain translation knowledge from *mixed-language* web pages, i.e., pages where lots of words, especially proper nouns and technical terms, are expressed in two languages. This is very common in Japanese, Chinese, and Korean, where English translations are often provided for "foreign" words. This approach has not been applied outside these languages, however (see, e.g. Zhang et al. (2005); Cheng et al. (2004)).

The results by Resnik and Nie et al. broadened the horizon for employing parallel corpora in CLIR, but despite the enormity of the WWW, parallel content is limited mainly to official text or, e.g., web sites of multinational companies. For special domains and language pairs, resources were still scarce.

Comparable corpora were seen as at least a partial solution to this problem. After all, it would seem much easier to find collections that only share similar topics than to find parallel texts. Furthermore, the requirement for

translation quality is more relaxed in CLIR than in MT, which allows the use of noisier learning data. Sheridan and Ballerini (1996) aligned Italian and German news reports by the Swiss news agency SDA by combining reports with similar dates and meta-descriptors. Although produced by the same agency, the documents were not parallel. Braschler and Schäuble (1998) aligned news stories by SDA and the American news agency AP. They used source documents as queries to retrieve target documents. Alignments were then made by using matching dates and query-document similarity values. Utsuro et al. (2002) used date-based alignment in creating comparable corpora from news reports in the web.

However, all three of the above studies on comparable corpora assumed that matching could be made depending on some meta-data, i.e., publication dates or content descriptors. Also, even though not parallel, the used collections were news collections from the same time period. News text does not, however, sufficiently cover the vocabulary of more technical domains, and text of most other domains cannot be aligned based on matching dates. Thus, even after the pioneer work on comparable corpora, the questions of sufficient lexical coverage and availability for different language pairs remained.

More recently, there have been efforts to mine parallel sentences from unaligned comparable corpora (Munteanu and Marcu, 2005; Fung and Cheung, 2004) in order to broaden lexical coverage of MT systems. However, as strictly parallel corpora are not mandatory in CLIR (at least when other, complementary resources are used), it remains an open question whether these methods are necessary in CLIR. It could be argued that employing only parallel subsets of comparable corpora squanders the evidence brought by other, less-than-perfect alignments.

3.6 Combined approaches

Combining different CLIR resources, i.e., dictionaries, corpora, cognate matching etc. provides more evidence to query translation, and hence should result in better CLIR performance. "More evidence" in CLIR means both broader lexical coverage and reduced ambiguity because good translation alternatives are likely to outnumber bad ones when multiple sources of knowledge are employed (Gey et al., 2001; Braschler, 2004; Savoy, 2004).

The combination can be made either prior to document retrieval or after it. In the former, more popular, approach, outputs of the different resources are combined to form a single target language query. The different resources can be weighted by weighting the subsets of query keys produced by the dif-

ferent resources. In the latter approach, each translation approach produces a query, and the results of the queries, i.e., document ranks, are merged.

The Utaclir system (Keskustalo et al., 2002), widely used in this study, combines dictionary-based translation, s-gram matching of OOV words and strong structuring of the target language queries with the Pirkola method (Pirkola, 1998). The Pirkola method implies combining translation alternatives with InQuery's *#syn* operator (see Section 2.2), i.e., treating the translation alternatives as synonyms. This neutralizes the effect of translation ambiguity.

Query expansion is found to be useful also in CLIR. McNamee and Mayfield (2002) found that pre-translation QE is advantageous especially when translation resources are scarce. Expansion keys may "save" queries from OOV words, because expansion keys related to OOV words may be in-vocabulary, i.e., translatable. Also, expansion keys provide more context which reduces translation ambiguity.

Chapter 4

Results

The research problems stated in Chapter 1 can be divided into three subsets: problems concerning the acquisition and alignment of comparable corpora (problems 1, 2, and 7); problems related to obtaining translation knowledge from comparable corpora (problem 3); and problems related to comparable corpora as a part of CLIR systems (problems 4-9). Accordingly, in the next sections, experiments related to these subsets of problems, and their results, are summarized.

Various document collections were used in the experiments, either as a test collection for IR laboratory experiments (see Figure 2.1), or as part of a comparable corpus. Table 4.1 shows the collections used, along with their size and purpose of use. The news collections are all part of the collection of the Cross-Language Evaluation Forum (CLEF, see Peters (2006)). The genomics collections were created for Publication IV (see Section 4.1.1), and the MEDLINE corpus is the test collection of the genomics track of the TREC conference (Hersh, 2005). Apart from these collections, the JRC-Acquis parallel corpus (Steinberger et al., 2006) was used in publications IV and V. The corpus consists of legislative documents of the European Union.

4.1 Creation of comparable corpora

Figure 4.1 depicts the process of creating comparable corpora. The process consists of the acquisition phase, where texts in specific languages are obtained, and the alignment phase, where the obtained texts are aligned. The acquisition phase applies focused web crawling to gather the texts. In the alignment phase, the source language documents are processed into queries, which then are translated into the target language with some available translation resource(s). The acquired target language documents are queried with

Table 4.1: Document collections used in the study

Source	Domain	Language	Documents	Purpose[1]	Publication
Aamulehti	News	Finnish	54,851	CC	I,III
L.A. Times	News	English	113,005	TC,CC	I, II, V
TT[2]	News	Swedish	142,819	CC	II,III,V
Göteborgs-Posten	News	Swedish	72,858	TC	III
Helsingborgs Dagblad	News	Swedish	88,478	TC	III
GenWeb-Eng	Genomics	English	149,500	CC	IV, V
GenWeb-Spa	Genomics	Spanish	30,800	CC	IV
GenWeb-Ger	Genomics	German	84,200	CC	IV, V
MEDLINE	Medical	English	4,591,008	TC	IV,V

[1] TC = test collection, CC = comparable corpus
[2] Tidningarnas Telegrambyrå, Swedish news agency

the resulting queries, using, e.g., a probabilistic ranking algorithm. Alignments are then made between the source document and the top ranking target document(s). Score thresholding is used to filter out low-quality alignments, i.e., pairs whose topics do not match adequately. Further, date-based matching can be used if the publication date of the documents is a relevant matching criterion, as it is in, e.g., news documents.

In the following sections, the acquisition phase and the alignment phase are discussed in depth.

4.1.1 Acquiring comparable texts from the web

The acquisition phase was not addressed until the fourth publication of this study. In the preceding publications, the news document collections of the CLEF campaign (Peters, 2006) were used as translation corpora. The CLEF corpora are news collections that cover approximately the same time period in the mid 1990's, which makes them ideal for alignment. However, outside of the news domain it is unlikely that such ideal collections can be found.

In Publication IV, a method for acquiring comparable corpora from the web is proposed. It is based on focused web crawling, i.e., searching web content belonging to a specific topic by employing the hyper link structure of the web (Chakrabarti et al., 1999). The topical crawling approach was chosen because comparable corpora are needed to compensate for the limitations of general resources, such as general-purpose dictionaries, which do not cover vocabulary of special domains.

The method is outlined in Figure 4.2. Before the actual crawl, domain-

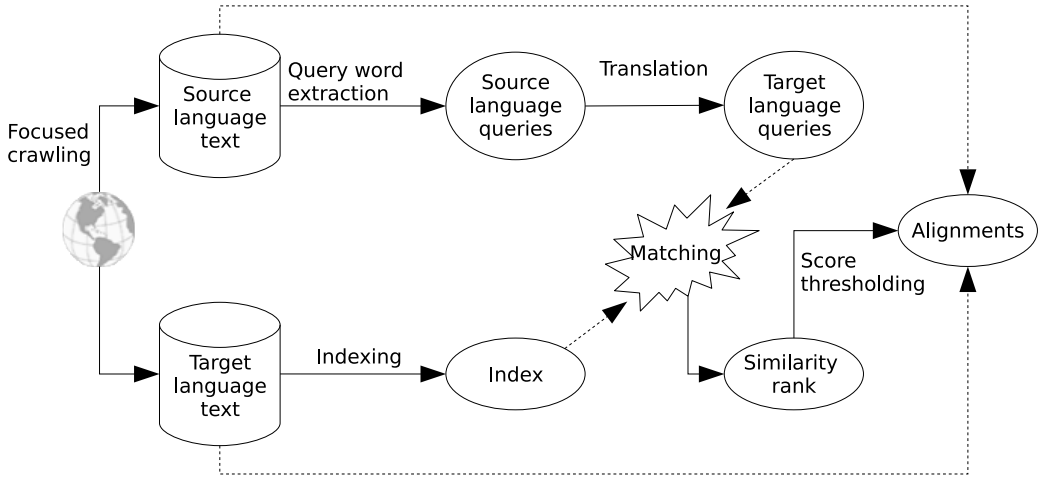

Figure 4.1: The process of acquiring and aligning comparable corpora

specific vocabularies are semi-automatically gathered from the web for all the wanted languages. The vocabularies play an important part in the process: they are used in finding the seed URLs of the crawl, and as "driver queries" to steer the crawling process to pages that contain text of the wanted topic. A set of seed URLs for each language is established by using the gathered vocabularies to query the web with, e.g., Google. A priority queue that holds the URLs of the to-be-visited pages is initialized with the seed URLs.

The actual crawl proceeds in the following way. One by one, the head URL of the URL queue is removed and the page pointed to by the URL is fetched. The text paragraphs of the page are extracted, and the language of the paragraphs is detected. If the language of a paragraph is one of the sought ones, the paragraph is matched against the driver query which consists of the domain vocabulary of that particular language. If the driver query similarity of the paragraph exceeds a threshold, the paragraph is saved to disk to wait for the alignment phase.

The out-links of each fetched page are extracted and scored. The score is based on matching the driver query against the link's anchor text (i.e., the text inside the HTML *a* tags). Also, the driver query similarity of the entire page, and average driver query similarity of pages belonging to the same host, are factored into the score. The URL queue is prioritized based on the scores.

In the proposed method, paragraphs, instead of whole pages, are used for the following reasons: Firstly, typical web pages have lots of content that does not belong to the topic of the page: navigation bars, contact information etc. In statistical translation, it is essential that words appear in their sentential

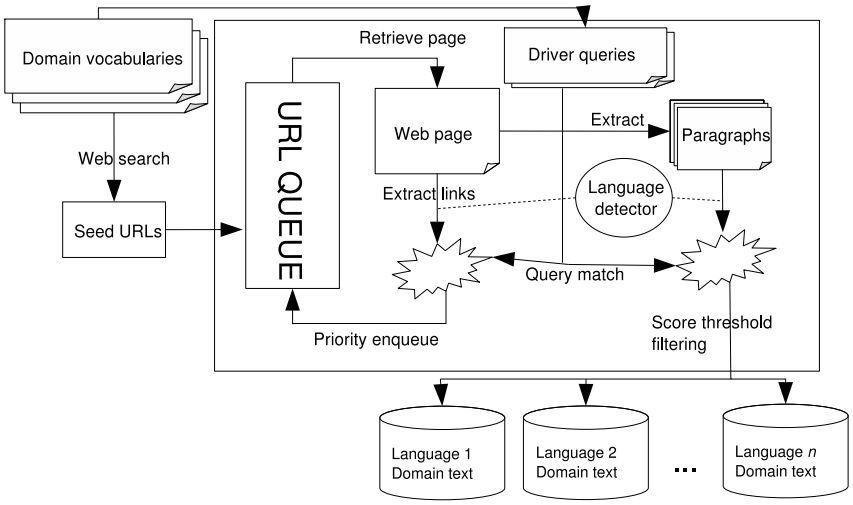

Figure 4.2: The crawling process

Table 4.2: Sizes of the acquired corpora

Language	Size (MB)	Words ($\cdot 10^6$)	Paragraphs
English	154	21.5	149,500
Spanish	25	3.5	30,800
German	73	8.8	84,200

context, which is often not the case with this kind of "functional content" of a web page. Secondly, some pages have text in multiple languages, while a single paragraph is usually written in only one language. Thirdly, the alignments were also made on paragraph, not document, level. One paragraph usually expresses a single concise idea, and thus suits better as a provider of context in statistical translation than a web page in its entirety.

The proposed method was used in Publication IV to gather English, German, and Spanish text in the genomics domain. Table 4.2 depicts the sizes of the acquired corpora.

The acquired corpus, the GenWeb corpus, was aligned to provide translation knowledge for Spanish-English and German-English query translation. The Cocot program (see Section 4.2), that uses aligned corpora as a cross-language similarity thesaurus, employed the alignments in experiments that consisted of two distinct set-ups. Firstly, standard laboratory CLIR experiments were performed with the topics of the genomics track of the 2004 TREC conference (Hersh, 2005). Secondly, word translation tests were performed, in which individual genomics-related words were extracted from the

topics and translated with the genomics web corpus on one hand, and with the JRC-Acquis parallel corpus, on the other hand.

In the IR experiments, Spanish and German translations of the TREC topics were transformed into queries, which were then translated into English with various CLIR systems. The target collection, the MEDLINE collection of medical abstracts and citations, was then queried with the translated queries. Standard performance measures, such as mean average precision (MAP), precision at a low recall level, and the 11-point interpolated precision, were reported. Cocot (CC) with the GenWeb corpus was combined with the Utaclir (UC) query translator (see Section 3.6). This was done because, realistically, GenWeb would be a complementary resource, its purpose being to cover technical vocabulary that are OOV for general resources. The combination performed better than UC alone, which suggests that acquiring web-based comparable corpora is indeed worthwhile. However, the difference was significant only in the German-English runs. The performance of the combination UC-CC was comparable to a machine translation system and to a combination where Utaclir was complemented by the FITE-TRT rule-based translation system (see Section 3.2). This latter combination did particularly well in the Spanish-English runs.

In the word translation tests, genomics-related vocabulary was translated with the Cocot-GenWeb system. The same words, extracted from the TREC genomics topics, were also translated with Cocot that employed the JRC-Acquis parallel corpus. Both Spanish and German words were translated. The tests aimed to prove that in special domains, it is not sufficient to use general-purpose resources, even of high quality, such as the JRC corpus. A measure for "translation goodness" was proposed: in short, a good translation appears relatively more frequently in the relevant documents of the topic it is extracted from, than in the other documents of the target collection. The tests clearly proved that the GenWeb could provide more good translations of the domain vocabulary than the JRC parallel corpus. In fact, more than half of the words were OOV for the JRC corpus, both in Spanish and German tests.

The experiments in Publication V are also relevant in evaluating the proposed method for acquiring translation corpora. The German genomics topics were translated, among others, with the combination Utaclir-Cocot. As translation corpus, Cocot utilized the GenWeb and the JRC corpora. Utaclir-Cocot with GenWeb performed significantly better than Utaclir-Cocot with JRC. This is consistent with the findings of the word translation tests.

4.1.2 Aligning comparable corpora

The following notational conventions are used in the rest of this chapter: *Tuples* are denoted between angle brackets. For example, let a and b be tuples: $a = \langle 1, 2, 3 \rangle$, $b = \langle 4, 5 \rangle$. The *length of tuple* a is $\text{len}(a) = 3$. *Tuple concatenation* is denoted by $\bowtie$. For example, $a \bowtie b = \langle 1, 2, 3, 4, 5 \rangle$. Tuple components are given by their indices using the notation $a[i]$ for the ith component of tuple a. For example, $b[2] = 5$.

Let $d^S \in C^S$ and $d^T \in C^T$ be documents in the source and target collections, respectively [1]. The *alignment candidate set* of source document d_i^S is defined as

$$AC_i = \{\langle d^T, \sigma \rangle | d^T \in C^T \wedge \sigma = sim(d_i^S, d^T) \wedge \sigma \geq \theta\}, \tag{4.1}$$

i.e., AC_i consists of tuples that contain the target documents whose similarity to d_i^S exceeds some threshold θ, paired with the similarity scores. Now, the set of alignments can be defined as the function

$$A(C^S, C^T) = \{\langle d_i^S, D_i \rangle | d_i^S \in C^S \wedge D_i = \tau\left(AC_i, R\right)\}, \tag{4.2}$$

That is, the alignments are a set of pairs consisting of a source document, and a *hyper document* D_i, which actually is a tuple τ of target documents defined by the set AC_i and the constant R. This constant is the maximum number of target documents aligned with a single source document.

Finally, we can define the tuple τ recursively as follows:

$$\tau\left(AC, R\right) = \tau\left(AC, 1, R\right), \tag{4.3}$$

where

$$\begin{aligned}
\tau\left(AC, i, R\right) &= \langle d^T \rangle \bowtie \tau\left(AC - \{\langle d^T, \sigma \rangle\}, i + 1, R\right) : \\
&\quad \langle d^T, \sigma \rangle \in AC \wedge \neg\exists \langle d^{T\prime}, \sigma' \rangle \in AC : \sigma' > \sigma \\
&\quad \text{if } AC \neq \emptyset \wedge i \leq R \\
\tau\left(AC, i, R\right) &= \langle\rangle \\
&\quad \text{otherwise.}
\end{aligned}$$

That is to say, the tuple $\tau\left(AC_i, R\right)$ contains at most R target documents whose similarity to d_i^S exceeds θ, arranged in descending order of similarity to d_i^S.

The alignment procedure proposed in this study comprises of the following steps:

[1] This presentation is somewhat more accurate than the one in Publication IV, on which it is based on.

1. For each source document, extract the n words best describing the topic of the document. This can be done in the following way: first, order the words in the document by decreasing frequency in the document. Second, calculate the RATF value (see Section 2.4.2) for each word, and filter out words with RATF values that are below a threshold. The RATF value measures the discrimination value, i.e., semantic significance of a word. Select the top n remaining words to represent the document. In the experiments, n was usually between 20 and 30.

2. For each source document d_i^S, translate the selected words with some available translation resource, e.g., a dictionary.

3. Query the target documents with the translated queries, using, e.g., a probabilistic ranking algorithm.

4. For each query, examine the result set and choose at most R target documents whose similarity to the query exceeds θ into the hyper document D_i. Additionally, if publication date of the documents is relevant (as in news documents), dates can be used in filtering prospective target documents.

Prior to Publication IV, the alignments were always 1-to-1, i.e., the size of each hyper document was 1. In training tests not published in Publication IV, the 1-to-n alignments outperformed 1-to-1 alignments, supposedly because they provide more evidence for translation. Moreover, in publications II and III, where news collections were aligned, the fourth step involved iterations over different "date windows". A date window of n means that the aligned documents may be published at most n days apart from each other. Instead of one similarity threshold, three thresholds ($\theta_1 < \theta_2 < \theta_3$) were applied in the date-based alignment. In the first iteration round, only target documents published on the same day as the source document were searched for, and threshold θ_1 was applied. On the second and third rounds, the date window was incremented from zero to one and two, respectively, θ_2 being the threshold. On the fourth round, date-based filtering was abandoned, and only score thresholding with θ_3 was applied. In a way, the iteration algorithm trusts the date-based evidence more than score-based evidence: target documents published near the source document are preferred, even with lower similarity to the source query. This seems reasonable, because the similarity score reflects the similarity of the source and target documents in an imperfect way: especially the translation step can bring noise to the process by way of OOV words and translation ambiguity.

In Publication I different alignment schemes were compared. A small subset of the Aamulehti collection was aligned with the L.A. Times collection. The Utaclir query translation program was used in the translation step, as is done in the other publications of this study. The tested alignment schemes varied in the target document filtering stage, i.e., stage four of the above list. Three different schemes were tested:

1. Unrestricted alignment. The highest ranking document was chosen as the alignment pair.

2. Date-restricted search. Search was restricted to target documents inside a date window. Window sizes one and two were tested separately.

3. Combined approach. First, a date-based search was made and a score threshold was applied. If no alignment pair was found, an unrestricted search was performed.

The schemes were evaluated by randomly choosing 100 alignment pairs created by each scheme, and assessing the similarity of the document pairs with a 5-point similarity scale proposed by Braschler and Schäuble (1998). The scale ranged from "unrelated" to "same topic". The combined scheme minimized the number of unrelated alignments, and was deemed the best approach. It should be noted that the combined approach of Publication I was not exactly the one proposed in later publications and described earlier in this section. Yet, the results showed that combining evidence from various sources (i.e., ranking, dates, similarity score etc.) is advantageous in creating alignments.

Further, Publication I discussed reasons for bad alignments. In the first step, i.e., selecting query words from source documents, word lemmatization and decompounding (which was done because the source language was Finnish, a highly agglutinative language) often brought in extraneous keys and thus caused ambiguity. In the translation step, the ambiguity was further increased by translation ambiguity. Also, proper nouns and other important search keys were often OOV for the dictionary-based Utaclir. In short, bad alignments were mostly due to reasons that impair CLIR performance in general (see Section 3.1).

In Publication II, where the iterative date-based scheme was proposed, different levels for the similarity thresholds (θ_1, θ_2, and θ_3) were experimented with and evaluated with the 5-point scale. The source collection was the Swedish TT collection, and the target collection was the L.A.Times corpus. For each tested alignment approach, a sample of 500 alignments was evaluated. A restrictive threshold approach, where the threshold levels were relatively high, produced better alignments than a permissive approach, where

Table 4.3: The aligned corpora created in this study

Source collection	Target collection	Source documents	Alignments	Unique target documents	Align level[1]	Style	Publication
TT	L.A.Times	72,260	13,142	5,404	D	1-to-1	II
Aamulehti	TT	55,298	12,045	7,422	D	1-to-1	III
GWG[2]	GWE[3]	84,200	30,087	30,049	P	1-to-n	IV
GWS[4]	GWE	30,800	16,073	21,664	P	1-to-n	IV
TT	L.A.Times	72,260	12,579	7,732	D	1-to-n	V

[1] D = document, P = paragraph
[2] GenWeb-Ger
[3] GenWeb-Eng
[4] GenWeb-Spa

the thresholds were lower. Further, normalizing the similarity scores with a factor proportional to the length of the target language queries was found beneficial. This is because the similarity scores of ranking algorithms (InQuery in this case) depend on the query length. InQuery, for example, gives higher scores to short queries. In the alignment procedure, this caused some very short source documents to be aligned, although they were unrelated to the target documents.

Table 4.3 presents statistics about the aligned corpora created for this study. The number of source documents is the theoretical maximum for the number of alignments. The actual number of alignments is far below this maximum for all corpora, because for many source documents, a topically matching target document could not be found. The number of unique aligned target documents is far below the number of alignments in the corpora where 1-to-1 alignments were used. This reflects the fact that the alignments were not bijective, i.e., some target documents were part of more than one alignments. This is also true in the genomics corpora, but since they are aligned in 1-to-n manner, the number of unique target documents is much greater. This is beneficial for translation knowledge extraction, because more target documents means more evidence for translation.

4.2 Similarity thesaurus translation

As mentioned in Section 3.4, a cross-language similarity thesaurus involves calculating similarity scores between source language query words and words in the target language documents of an aligned corpus. "Similar" cross-lingual word pairs co-occur frequently in the aligned documents, because the documents are either translations of each other, or, in case of a comparable corpus, they share similar topics. The vector space model of IR (see Section 2.1) can be employed in calculating the scores, only this time documents are seen as features which "describe" or "define" the words, not the other way around. The documents define a semantic space, and a word's location in the space is defined by its occurrence pattern in the document set. Intuitively, this corresponds to the notion that the meaning of a word is defined by the contexts it is used in.

As in document retrieval, *tf.idf* style weights can be used to measure the importance of documents to words, i.e., the amount that a document contributes to the meaning of a word. The considerations concerning the different components of the weighting formula (see Section 2.1.2) should be taken into account also in weighting documents in a similarity thesaurus:

1. The importance of a feature is proportional to its frequency. In a similarity thesaurus, this means that the more times a document includes a word, the more it contributes to its meaning.

2. However, the importance of a feature (i.e., document) is inversely proportional to the number of other words it "describes". That is, a long document deals with many concepts and words, while its contribution to the meaning of individual words is smaller than that of a shorter document.

3. Words with lots of features, i.e., words that appear in a lot of documents are not necessarily better translations than rarer words.

The Cocot (Comparable Corpus Translation program) system, developed for this study, uses aligned cross-language corpora as cross-language similarity thesaurus for query translation. It calculates the feature weight w_{ij} for document d_j that describes the word s_i as follows:

$$w_{ij} = \begin{cases} 0 & \text{if } tf_{ij} = 0 \\ \left(0.5 + 0.5 \cdot \frac{tf_{ij}}{Maxtf_j}\right) \cdot \ln\left(\frac{NT}{dl_j}\right) & \text{otherwise} \end{cases}, \tag{4.4}$$

where tf_{ij} is the frequency of s_i in document d_j, $Maxtf_j$ the largest term frequency in d_j, dl_j the number of unique words in the document. NT can

be the number of unique words in the collection, or its approximation. This *tf.idf* modification is adapted from Sheridan and Ballerini (1996) who also used it in similarity thesaurus calculation.

For a hyper document D in which a word s_i appears, the weight is

$$W_i = \sum_{r=1}^{len(D)} \frac{w_{ir}}{\ln(r+1)} \ , \tag{4.5}$$

where w_{ir} is the weight of the word s_i in the document $D[r]$, i.e., the rth document in the hyper document D. The document is the rth most similar target document to the source document d^S. The less similar the documents d^S and $D[r]$ are (i.e., the higher the value r), the less $D[r]$ can be trusted as a source of translation knowledge. This is echoed in the equation above.

Finally, the similarity between a query word s_i and a word s_j appearing in the target hyper documents can be calculated as

$$sim(s_i, s_j) = \frac{\sum_{\langle d^S, D\rangle \in A} w_i \cdot W_j}{\|\mathbf{s_i}\| \cdot \left((1-\alpha) + \alpha \cdot \frac{\|\mathbf{s_j}\|}{\overline{\|\mathbf{T}\|}}\right)} \ , \tag{4.6}$$

where A is the set of alignments (see Equation 4.2), $\|\mathbf{s_i}\|$ the norm of the vector representing the word s_i, and $\overline{\|\mathbf{T}\|}$ is the mean of the target word vector lengths. The formula applies the pivoted vector length normalization scheme (Singhal et al. (1996), see Equation 2.5). As in document retrieval, the scheme is applied to compensate for the "over-normalization" of long feature vectors.

In Publication III, the performance of Cocot using the pivoted scheme was compared to Cocot that used "uncorrected" cosine normalization. A total of 52 Finnish CLEF topics were translated into Swedish using the Finnish-Swedish translation corpus of Aamulehti and TT documents (see Table 4.3). The recall base had graded relevance assessments. The relevance levels were the same as in Table 2.3. The graded levels were collapsed into three binary ones: liberal, regular and stringent. Generalized recall and precision was also calculated (see Section 2.3.3). Both normalization schemes performed evenly with liberal and regular levels, but in the stringent level the pivoted scheme performed better (24.6 vs. 18.7 in MAP), although the difference was not statistically significant.

When a word q is translated, the score in Equation 4.6 is calculated between q and every word appearing in the target documents. After this, the target language words can be ranked according to the scores. A word cut-off value (WCV) determines how many target words are chosen as translations. Further, score thresholding is used to filter out bad translations – after all, the

similarity score echoes the confidence in translation. Thus, the *translation set* $T(q)$ of a source word q consists of target language words s_i^T so that $T(q) = \{s_1^T, s_2^T, \ldots, s_n^T | sim(q, s_i^T) > \theta, n \leq \text{WCV}\}$, where θ is the applied score threshold. Cocot employs the Pirkola method (Pirkola, 1998) in query structuring, i.e., a translation set is enclosed within the *#syn* operator of the InQuery language (see Section 2.2). In CLIR training experiments this kind of query structuring outperformed structures where translation alternatives were weighted based on their similarity scores.

Next, the translation procedure of Cocot is explained in more detail. The most important file structures of Cocot are explained first, and then the word similarity calculation algorithm is presented.

Cocot uses the following index structures for fast data look-up:

- The source language *lexicon* includes, for every word s_i^S in the source documents, a pointer to the *inverted file*. For now, the pointer can be thought of as the index i for s_i^S.

- The words in the source language documents are indexed in the inverted file, which is a structure where words are mapped to the list of their *postings*, i.e., occurrences. In Cocot's inverted file, for each word s_i^S in the lexicon, there is a set $P_i = \{\langle d_j^S, tf_{ij}\rangle | tf_{ij} > 0\}$, where tf_{ij} is the frequency of s_i^S in the source document d_j^S.

- The target collection is indexed in a *document index*, which, for every target document d_m^T contains the set $Q_m = \{\langle s_l^T, tf_{lm}\rangle | tf_{lm} > 0\}$, where tf_{lm} is the frequency of word s_l^T in the target document d_m^T. That is, Q_m is the set of indexed words in the document, along with their frequencies.

- The *alignment file* contains, for each source document d_j^S, the document IDs of the documents belonging to the hyper document D_j, i.e., the ordered set of target documents aligned with d_j^S (see Equation 4.2). Note that for some source documents, it may be that D_j is empty, because not all source document are aligned.

Algorithm 1 depicts the way Cocot calculates the similarity scores between an input word q and the words in the target documents. After the described procedure, the target language words are sorted based on the scores, after which the set $T(q)$ can be constructed, based on the sorted words and parameters θ and WCV. It should be noted that the run-time calculation of word vector length and several other parameters of the weighting formula (Equation 4.4) would be costly. Therefore, these values are calculated in the indexing phase, and are stored on disk.

Algorithm 1: Calculating word similarities in a similarity thesaurus

Input: source language query word q
Output: hash S of similarity scores, $S_l = sim(q, s_l)$
$S \leftarrow$ empty hash
retrieve index i for word q from the source language lexicon
if *q not in lexicon* **then**
  return S
end
retrieve P_i from the inverted file
for $\langle d_j^S, tf_{ij} \rangle \in P_i$ **do**
  calculate weight w_{ij} (Eq. 4.4)
  retrieve hyper document D_j from the alignment file
  $W \leftarrow$ empty hash table `/* W is for calculating the 'hyper weights' according to Eq. 4.5 */`
  for $r \leftarrow 1$ **to** *len(D_j)* **do**
    $m \leftarrow$ document ID of target document $D_j[r]$
    retrieve set Q_m from document index
    for $\langle s_l^T, tf_{lm} \rangle \in Q_m$ **do**
      calculate weight w_{lm} (Eq. 4.4)
      if *W_l not in W* **then**
        $W_l \leftarrow 0$
      end
      $W_l \leftarrow W_l + w_{lm} / (\log (r + 1))$ (Eq. 4.5)
    end
  end
  foreach *W_l in W* **do**
    if *S_l not in S* **then**
      $S_l \leftarrow 0$
    end
    $S_l \leftarrow S_l + \frac{w_{ij} \cdot W_l}{\|\mathbf{s_i}\| \cdot \left((1-\alpha) + \alpha \cdot \frac{\|\mathbf{s_l}\|}{\|\mathbf{T}\|} \right)}$ (Eq. 4.6)
  end
end
return S

4.3 Comparable corpora in CLIR

How, then, should comparable translation corpora be applied in CLIR? The findings of this study suggest that combining comparable corpus translation with other resources is beneficial.

In Publication II, 91 Swedish CLEF topics were translated into English with various CLIR system set-ups. The test collection was the L.A. Times collection. Cocot used the Swedish-English TT-L.A. Times translation corpus (see Table 4.3). The target collection of the experiments and the target collection of the translation corpus were the same. The problem of "training with the test set" was avoided by removing the aligned target documents from the test collection and the recall base. Standard IR measures of performance (i.e., MAP, P@10, R-Precision, and 11-point interpolated precision) were reported. The following system set-ups were applied in the experiments:

Monolingual The original English CLEF topics were transformed to queries to provide the monolingual baseline.

CC Cocot, employing the TT-L.A. Times translation corpus, used to translate the Swedish queries into English.

UC The Utaclir query translation program employed to translate the Swedish queries.

PRF+UC Utaclir with pre-translation pseudo relevance feedback. The PRF technique applied is explained in depth in Publication II.

CC-UC Cocot was applied first to the Swedish queries. Words that were OOV for Cocot were then translated with Utaclir.

PRF+CC-UC The previous set-up with pre-translation pseudo relevance feedback.

UC-CC Queries were first translated with Utaclir, after which OOV words were translated with Cocot.

UC+CC All query words were translated with both Utaclir and Cocot.

By all measures, the combination approaches, save for PRF+CC-UC, performed quite evenly, while UC and CC were somewhat clearly behind them. However, significant differences were not found, although difference between UC and some of the combination approaches (namely PRF+UC, CC-UC, and UC-CC) was nearly significant ($0.06 \leq p \leq 0.10$). The poor performance of PRF+CC-UC was somewhat surprising, since pre-translation

expansion has generally proved to be a successful technique in CLIR (see, e.g., McNamee and Mayfield (2002)). The reason for the poor performance is perhaps noise. Since a noisy resource (i.e., a comparable corpus) is used, the noise in the target queries increases with the number of keys translated by Cocot.

In general, however, the experiments showed that combining comparable corpus translation with dictionary translation outperformed those approaches that used either of these resources alone. The reasons for this were analyzed by examining the "improvement sets" of the different approaches. The improvement set I_M was the set of queries in which the approach M performed significantly better (over 5 % absolute difference in average precision) than UC, which was considered a baseline CLIR approach. The analysis was based on the assumption that the overall improvement of UC-CC over UC was based on succeeding in translation of OOV words, because in UC-CC, Cocot was used to translate Utaclir's OOV words. The CC-UC approach, on the other hand, had a larger improvement set than UC-CC ($|I_{UC-CC}| = 19, |I_{CC-UC}| = 29$) which also included most of UC-CC's improvement set ($|I_{UC-CC} \cap I_{CC-UC}| = 14$).

Now, consider a query that belonged to the set $I_{CC-UC} - I_{UC-CC}$, i.e., a query that performed significantly better than UC in the CC-UC approach, but not in the UC-CC approach (this set had, of course, $29-14 = 15$ queries). This query consisted of three subsets of translated keys – K_U, K_C, K_O, i.e., keys translated by Utaclir, keys translated by Cocot, and keys that were OOV for both programs. Any of these sets (but not all of them) could have been an empty set. The difference of these sets between approaches UC-CC and CC-UC is crucial in the analysis. The set K_O was the same for both approaches, and keys that were in K_C for UC-CC were in K_C also for CC-UC, because same Cocot parameters were used for both approaches. The difference was that some of the keys that had been in K_U in UC-CC moved to K_C in the CC-UC approach, because Cocot was applied first in the latter approach. The translations of the set K_U were either correct translations or incorrect ones caused by translation ambiguity, because K_U consisted of dictionary translations. The query performed better in the CC-UC approach, because the ambiguity of the translations that moved from K_U to K_C was reduced by Cocot either by providing only the correct translations, or by bringing in related expansion keys. The latter option is highly probable, because Cocot not only produces translations, but related words frequently appearing in similar contexts. Hence, *the improvement of some of the queries can be attributed to the expansion keys brought in by Cocot.*

It should be noted that although this analysis revealed 15 queries (out of 91) that benefit from expansion keys, the figure is most probably higher

in reality. The aim of the improvement set analysis was not to measure the amount of improvement that expansion keys bring – this would require a deeper analysis of the actual queries – but to show that such an improvement actually exists.

4.3.1 Comparable corpora and highly relevant documents

In Publication III, experiments with graded relevance assessments were made (see Section 2.3.3). Part of these experiments, concerning the vector normalization factor of Cocot, were already discussed in Section 4.2. In the experiments, 52 Finnish CLEF topics were used as queries and translated into Swedish. The target test collection consisted of about 160,000 newspaper articles by the Swedish newspapers Göteborgs-Posten and Helsingborgs Dagblad. Again, Utaclir was used as a baseline CLIR approach, which was compared to Cocot and the combination of Cocot and Utaclir. Cocot employed the Finnish-Swedish comparable corpus Aamulehti-TT (see Table 4.3).

The performance of the systems was measured in MAP and the 11-point interpolated precision, using three relevance threshold levels, i.e., stringent, regular, and liberal. The generalized versions of the measures were also reported (see Section 2.3.3). The findings echo those of Publication II, in the respect that the combination of Utaclir and Cocot improves over dictionary-based translation. This was observed on all relevance threshold levels, as well as with the generalized measures. The interesting finding was that Cocot alone was significantly better than Utaclir on the stringent relevance level, which seems to indicate that the qualities of corpus-based translation (i.e., expansion keys along with translations) support the retrieval of highly relevant documents particularly well. On the stringent relevance level, the "Cocot alone" approach even matched the monolingual baseline in MAP.

A further analysis was made on the reasons for failing to retrieve highly relevant documents. For each translation approach (i.e., Cocot, Utaclir, and the combination), 50 highly relevant documents that were ranked lower than the rank 50 were tried to "rescue" back to the top 50 with various measures. For Utaclir, the rescue operation required mostly adding exact translations and related words, which suggests that Utaclir would benefit from a larger dictionary and pre-translation QE. Cocot, on the other hand, required more word removals, indicating that, besides bringing translations and good expansion keys, Cocot also produces extraneous keys which harm query performance. This seems to support the explanation given earlier to

poor performance of Cocot with pre-translation query expansion: the more keys Cocot is given to translate, the more harmful keys it also produces. After some "saturation point", the harmful keys perhaps outweigh the good ones.

4.3.2 Properties of aligned corpora and CLIR performance

Research problems 8 and 9 (see Chapter 1) dealt with characteristics of aligned corpora that affect CLIR performance; namely the domain of the corpus, the quality of the alignments, and the size of the corpus. The effect of these three qualities on CLIR performance were examined in Publication V.

Two topic sets and language pairs were experimented with. In the Swedish-English runs, 70 CLEF news topics were used as queries and translated from Swedish into English. Of these, 30 were used to train the parameters of Cocot. The target collection was the L.A. Times collection. In the German-English runs, 50 topics from the genomics track of the TREC conference (Hersh, 2005) were employed, of which 20 were used as training queries for Cocot. The test collection was the MEDLINE collection.

Two different translation corpora were applied to both of the topic sets. For the Swedish queries they were the Swedish-English TT-L.A. Times corpus, particularly the version with 1-to-n alignments (see Table 4.3), and the JRC-Acquis parallel corpus. The German queries were translated with the GenWeb corpus (Table 4.3) on one hand, and the JRC-Acquis parallel corpus, on the other. These corpora varied with respect to the "topical nearness" to the translated queries: for example, in the German runs, the GenWeb was topically near the genomics-related queries, whereas the JRC corpus was quite far from them. Further, the alignment quality of the corpora also varied: the JRC parallel corpus is, by definition, a collection of aligned translations, whereas GenWeb and the CLEF corpus are only comparable corpora. Figure 4.3 depicts the four translation corpora with respect to the alignment quality and topical nearness.

The following five CLIR approaches were tested for both language pairs:

- Utaclir alone. This was, again, the baseline CLIR approach.

- Cocot alone using the JRC corpus. For the Swedish queries, the Swedish-English alignments of the corpus were used, and, naturally the German-English alignments were used for the German queries.

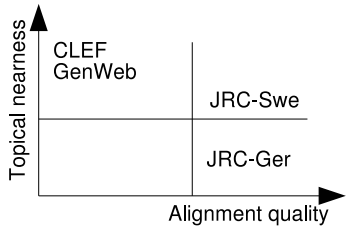

Figure 4.3: Alignment quality and topical nearness of the used translation corpora

- Cocot alone using a comparable corpus. For the Swedish runs, the corpus was the TT-L.A.Times corpus, and for the German runs, it was the German-English GenWeb corpus.
- Utaclir with Cocot that used the JRC parallel corpus. The target language query consisted of the concatenated output of the two programs.
- Utaclir with Cocot that used a comparable corpus.

Again, for both topic sets, the combined approaches performed better than the single resource approaches. In the German runs, Utaclir combined with the GenWeb Cocot was clearly better than the other approaches. Both of the approaches using the JRC parallel corpus (i.e., Cocot alone and Utaclir with Cocot using JRC) performed significantly worse than this combination, indicating that in special domains, general resources are not sufficient, even though they are of high quality as the JRC corpus is. In the Swedish runs, the Utaclir-Cocot combination approaches achieved comparable performance. This was expected, because the JRC corpus was this time topically closer to the queries.

The effect of corpus size on CLIR performance was examined by iteratively removing random alignments from the four translation corpora. On the smallest level, each corpus had 500,000 source language words. The performance of Cocot employing the cut-down corpora was measured on every 500,000 words onwards, until each corpus reached its full size. Figure 4.4 plots MAP against the size of the corpora, measured in the number of source language words. The curves indicate that the effect of the size is not linear; the performance fluctuates between size levels (though not significantly), and performance seems to reach its maximum level already after about a million source language words. However, size seems to matter, because there is a significant difference between the lowest and the highest level of size for all corpora, save for the German JRC, which performs badly on all size levels.

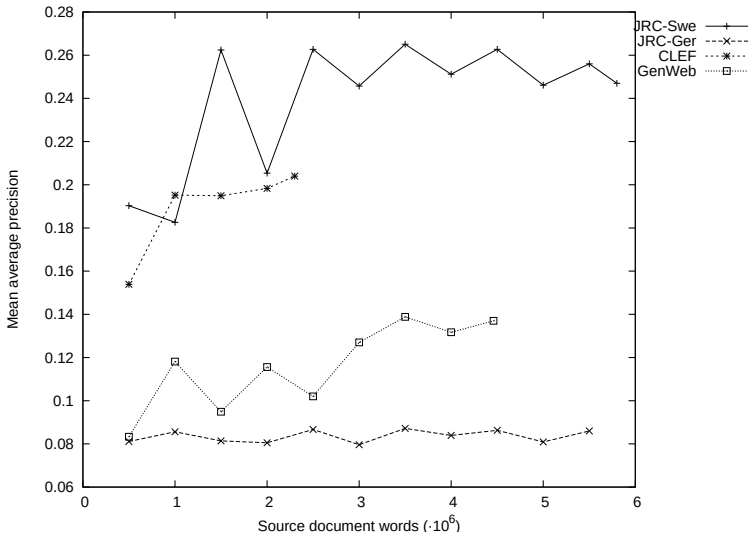

Figure 4.4: MAP plotted against corpus size for four translation corpora.

Chapter 5

Discussion

In this study, I proposed novel methods for the acquisition, alignment, and employment of comparable corpora in CLIR, and showed, with extensive experiments, that the proposed methods are beneficial for CLIR performance. The acquisition phase involved language-aware focused web crawling. The alignment method was based on using the source documents as CLIR queries, which were run against the target collection after they had been translated. Score thresholding and date windowing – when applicable – were used in deciding whether alignments were created. The constructed comparable corpora were used as cross-lingual similarity thesauri to provide translations for query words. This involved using the vector space model of IR in an inverted way, retrieving and ranking target language words as "answers" to source language query words. Cut-off values and score thresholds were used to delimit the size of the translation sets.

The experiments showed that the proposed approaches are best suited to special domains for which lexical resources are scarce. The domain in the experiments of this thesis (genomics) is only an example of a domain that can benefit from CLIR based on comparable corpora. Domains with less economical or political significance, that are not dealt with in "official" contexts, could be major beneficiaries. Also, CLIR between languages with few resources could be boosted.

Further, it was found that comparable corpora work best as a complementary resource for higher quality resources (e.g. dictionaries and parallel corpora) that provide translations for more general vocabulary. It was also shown that, compared to dictionary-based translation, the proposed approach reduces translation ambiguity by, e.g., producing not only translations of the query keys, but also related words that serve as query expansion keys.

The novel contributions of this thesis are as follows.

- *Applying focused crawling in the acquisition of comparable corpora.* Us-

ing focused crawling in this context is well-motivated, because domain-specific vocabulary is often missing from CLIR translation resources, and documents containing such vocabulary can be obtained with focused crawling.

- *Further development of the document alignment scheme of Braschler and Schäuble (1998).* Braschler and Schäuble mainly used date-based filtering, which is suitable for news documents. Date-based alignment is not applicable to all kinds of documents, however. Furthermore, date-based filtering aims to align two documents that report on the same event, although news reports can be related in other ways as well. For example, reports about earthquakes occurring independently (i.e., at different times and in different locations) contain common vocabulary, even though they are not directly related. These kind of relations are lost if one resorts only to date-based filtering. Further, Braschler and Schäuble only made 1-to-1 alignments.

- *Introducing new features to similarity thesaurus calculation.* Firstly, the pivoted vector normalization scheme was used to normalize the feature vectors. Previously, the scheme has been only applied to document retrieval. Secondly, alignment-phase evidence was used in calculating the feature weights for the hyper documents: the weights were inversely proportional to the alignment rank of the target documents (see Equation 4.4).

Next, problems related to the proposed methods, as well as possible future work, will be discussed. To begin with, the languages used in the experiments varied in many ways. For example, Finnish is a highly inflectional language, as opposed to, e.g., English and Spanish. Also, Finnish (as well as German) is a highly compounding language compared to English. The results of the experiments seem to hold regardless of the agglutinative nature of the languages used. However, this may well be largely due to the preprocessing techniques used. The TWOL lemmatizer program (Koskenniemi, 1983) was used in the experiments in normalizing and decompounding the words of the translation corpora. How different would the results be, if, e.g., stemming would be used instead?

The method for acquisition of comparable corpora is robust and it can be applied to different languages and domains. However, many improvements could be made to the crawling model, although it should be remembered that the contribution of this thesis is not in the crawler implementation, but in applying focused crawling to acquire comparable corpora. For example, the

acquisition of seed URLs could be made more automatic, using, e.g., freely available topical web directories, such as Yahoo.

The alignment method requires choosing at least one score threshold, which can be done via sampling the alignments on different threshold levels and evaluating the quality of the created alignments. This, of course, means manual work, but one working day may suffice in creating the alignments for one language pair (see Publication II). This is reasonable, considering that the aligned corpus created in this way could be a lasting CLIR resource.

The alignment method is dependent on two external resources: the initial dictionary that translates the queries and the retrieval engine used to index and search the target collection. In my opinion, this dependence is not harmful, because both of the resources can be quite easily acquired. General dictionaries are freely available in the web, and different open source search engines can be used in the latter task. At first, it might seem redundant to create translation resources by using another translation resource. However, the results clearly indicate that the role of the newly created translation corpus is to broaden the lexical coverage of a CLIR system, i.e., to provide translations for words that previously have been OOV.

A more significant question is, arguably, whether creating the alignments is necessary in the first place. As mentioned in Section 3.4, unaligned comparable corpora have been used as a translation resource by applying context vectors. The use of context vectors instead of alignments could be justified by looking at the alignment statistics of this study in Table 4.3. At most half of the original source documents are actually used in the aligned corpora. This happens because similar enough contexts are not found in the target collection, but it also means that most of the occurrences of a source language word to be translated are "thrown away". Hence, evidence for translation is wasted. In the context vector approach (Rapp, 1999), all occurrences of a source language word are used as evidence to calculate the context vectors.

However, the quality of this evidence is debatable, and it depends largely on the similarity of the original source and target collections. Both Rapp (1999) and Fung and Yee (1998) used newspaper collections that overlap in publication date at least partially. Would their approach work as well for corpora such as the genomics web corpora acquired for this study? To my knowledge, experiments where alignment-based and context vector-based approaches are compared have not been published, and this would be a an interesting topic for future studies. Also, a combination of the two approaches could be proposed. The alignments, that supposedly contain the higher quality evidence, could be used as primary evidence, while context vectors could provide supportive proof for translation. The details of such a model remain to be worked out.

In a working CLIR system, the employed translation corpora would have to be updated more or less frequently to keep up with new vocabulary. The current method for acquiring the corpora might prove to be slightly inefficient in this respect. Even though a new crawl would be initiated with fresh seed URLs, it is highly likely that the crawler would sooner or later end up on many pages that have already been visited in earlier crawls. To get more recent content more effectively, one could, e.g., "tap" into frequently updated web content, such as RSS feeds. The availability of such resources for different languages and domains is not self-evident, though.

The alignment method, as opposed to the acquisition phase, is better suited for dynamic corpora. When creating alignments for the updated corpus it would be possible to use the alignment parameters (i.e., score thresholds) from the previous alignments. This is because the new material would largely share the essential properties with the old corpus: the topic and the language of the texts would be the same, as well as perhaps the typical length of the documents. This would mean that the manual sampling work to choose the alignment parameters would have to be done only once for a given corpus.

As noted earlier, the similarity thesaurus uses the target document ranks from the alignment phase in its translation model: the lower the rank, the less the words of a document are weighted (see Equation 4.4). It would be quite easy to include the alignment scores in the model as well; after all they also provide a measure of the reliability of the translation evidence. This could also provide a convenient way to combine parallel and comparable corpora in a single model: parallel alignments (i.e., document pairs that are exact translations of each other) could be given some maximum similarity score, while the comparable alignments could be scored in proportion to the alignment scores. In this way, for example, the JRC-Acquis parallel corpus could be combined with the GenWeb corpus fruitfully to translate queries related to the genomics domain: the JRC corpus would provide translations for the more general vocabulary, while the GenWeb corpus could translate the genomics-related words.

In the IR experiments reported in the study, it was found that the proposed system works well, especially as a complementary CLIR resource. The results agree with the results of many previous studies (e.g. Savoy (2004); Braschler (2004)), which have shown that combining different translation resources is beneficial in CLIR. Hence, the relatively poor performance of the "Cocot alone" approaches in some of the experiments is not alarming. In the experiments, the combination of the different resources was done with rather simple methods, e.g., by concatenating outputs of two different query translation programs. The performance of the combination approaches could

perhaps be improved by, e.g., weighting the outputs of the different programs differently. For example, the translations produced by a dictionary-based program could be weighted more heavily than those produced by Cocot, because comparable corpora are noisier resources than dictionaries. Alternatively, different resources could be combined in a single translation model. For example, dictionary translation could be incorporated into the model proposed earlier, in which parallel and comparable corpora would be combined. In fact, a bilingual dictionary can be thought of as a parallel corpus where source language words are aligned with their translation alternatives.

In summary, this study proposed methods for creating and using comparable corpora in CLIR. The experiments of Publications I-V, and their results, suggest that the proposed methods can broaden the lexical coverage of a CLIR system, and, consequently, improve CLIR effectiveness. This seems to be particularly true for queries that contain domain-specific vocabulary that are OOV for general translation resources.

Personal contributions

In the collaborative publications (publications I-IV), Tuomas Talvensaari wrote the articles, except for Publication I, which was written by Martti Juhola and T.T.. The sections on statistical testing were written by Jorma Laurikkala, who also conducted the statistical tests on the experimental results. T.T. designed the experimental setups with help from other writers. In Publication IV, the word translation tests were designed by Kalervo Järvelin and Ari Pirkola. The measure for translation goodness in the same publication was proposed by K.J.. The experiments in all of the publications were conducted by T.T.. T.T. did the programming required for the publications, i.e, the retrieval engine for Publication I, the Cocot query translation program, and the focused crawler of Publication IV. T.T. also wrote various small-scale applications needed in data preprocessing.

Bibliography

Airio, E. (2006). Word normalization and decompounding in mono- and bilingual IR. *Inf. Retr.*, 9(3):249–271.

Airio, E. (2008). Who benefits from CLIR in Web retrieval? *Journal of Documentation*. Accepted for publication.

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

Ballesteros, L. and Croft, W. B. (1998). Resolving ambiguity for cross-language retrieval. In *SIGIR '98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 64–71, New York, NY, USA. ACM.

Belkin, N. J. and Croft, W. B. (1987). Retrieval techniques. *Annual Review of Information Science and Technology*, 22:109–145.

Braschler, M. (2004). Combination approaches for multilingual text retrieval. *Inf. Retr.*, 7(1-2):183–204.

Braschler, M. and Schäuble, P. (1998). Multilingual information retrieval based on document alignment techniques. In *ECDL '98: Proceedings of the Second European Conference on Research and Advanced Technology for Digital Libraries*, pages 183–197, London, UK. Springer-Verlag.

Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., and Mercer, R. L. (1993). The mathematics of statistical machine translation: parameter estimation. *Comput. Linguist.*, 19(2):263–311.

Buckley, C., Salton, G., Allan, J., and Singhal, A. (1994). Automatic query expansion using SMART: TREC 3. In Harman, D., editor, *The Third Text REtrieval Conference (TREC 3)*, pages 69–80. NIST and ARPA.

Chakrabarti, S., van den Berg, M., and Dom, B. (1999). Focused crawling: a new approach to topic-specific web resource discovery. In *WWW '99: Proceedings of the Eighth International Conference on World Wide Web*, pages 1623–1640, New York, NY, USA. Elsevier North-Holland, Inc.

Cheng, P.-J., Teng, J.-W., Chen, R.-C., Wang, J.-H., Lu, W.-H., and Chien, L.-F. (2004). Translating unknown queries with web corpora for cross-language information retrieval. In *SIGIR '04: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 146–153, New York, NY, USA. ACM Press.

Davis, M. and Dunning, T. (1995). A TREC evaluation of query translation methods for multi-lingual text retrieval. In Harman, D., editor, *The Fourth Text REtrieval Conference (TREC 4)*, pages 483–497. National Institute of Standards and Technology.

Davis, M. W. and Ogden, W. C. (1997). Quilt: implementing a large-scale cross-language text retrieval system. In *SIGIR '97: Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 92–98, New York, NY, USA. ACM.

Efthimiadis, E. N. (1996). Query expansion. *Annual Review of Information Science and Technology*, 31:121–187.

Fox, C. (1990). A stop list for general text. *SIGIR Forum*, 24(1-2):19–21.

Fung, P. and Cheung, P. (2004). Multi-level bootstrapping for extracting parallel sentences from a quasi-comparable corpus. In *COLING '04: Proceedings of the 20th International Conference on Computational Linguistics*, page 1051, Morristown, NJ, USA. Association for Computational Linguistics.

Fung, P. and Yee, L. Y. (1998). An IR approach for translating new words from nonparallel, comparable texts. In *Proceedings of the 17th International Conference on Computational Linguistics*, pages 414–420, Morristown, NJ, USA. Association for Computational Linguistics.

Gale, W. A. and Church, K. W. (1991). A program for aligning sentences in bilingual corpora. In *ACL '91: Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pages 177–184.

Gey, F. C., Jiang, H., Petras, V., and Chen, A. (2001). Cross-language retrieval for the CLEF collections - comparing multiple methods of retrieval. In *CLEF '00: Revised Papers from the Workshop of Cross-Language Evaluation Forum on Cross-Language Information Retrieval and Evaluation*, pages 116–128, London, UK. Springer-Verlag.

Grefenstette, G., editor (1998a). *Cross-Language Information Retrieval*. Kluwer Academic Publishers, Norwell, MA, USA.

Grefenstette, G. (1998b). The problem of cross-language information retrieval. In Grefenstette (1998a), pages 1–9.

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

Hull, D. A. (1996). Stemming algorithms: a case study for detailed evaluation. *J. Am. Soc. Inf. Sci.*, 47(1):70–84.

Ingwersen, P. and Järvelin, K. (2005). *The Turn: Integration of Information Seeking and Retrieval in Context (The Information Retrieval Series)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Järvelin, A., Järvelin, A., and Järvelin, K. (2007). *s*-grams: Defining generalized *n*-grams for information retrieval. *Inf. Process. Manage.*, 43(4):1005–1019.

Järvelin, K., Kristensen, J., Niemi, T., Sormunen, E., and Keskustalo, H. (1996). A deductive data model for query expansion. In *SIGIR '96: Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 235–243, New York, NY, USA. ACM.

Jing, Y. and Croft, W. B. (1994). An association thesaurus for information retrieval. Technical Report UM-CS-1994-017, Amherst, MA, USA.

Kekäläinen, J. (1999). *The effects of query complexity, expansion and structure on retrieval performance in probabilistic text retrieval*. PhD thesis, University of Tampere.

Kekäläinen, J. and Järvelin, K. (1998). The impact of query structure and query expansion on retrieval performance. In *SIGIR '98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 130–137, New York, NY, USA. ACM.

Kekäläinen, J. and Järvelin, K. (2002). Using graded relevance assessments in IR evaluation. *Journal of the American Society for Information Science and Technology*, pages 1120–1129.

Keskustalo, H., Hedlund, T., and Airio, E. (2002). Utaclir: general query translation framework for several language pairs. In *SIGIR '02: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 448–448, New York, NY, USA. ACM Press.

Keskustalo, H., Järvelin, K., and Pirkola, A. (2006). The effects of relevance feedback quality and quantity in interactive relevance feedback: A simulation based on user modeling. In *Advances in Information Retrieval*, volume 3936 of *Lecture Notes in Computer Science*, pages 191–204. Springer.

Kilgarriff, A. and Grefenstette, G. (2003). Introduction to the special issue on the web as corpus. *Comput. Linguist.*, 29(3):333–347.

Kishida, K. (2005). Technical issues of cross-language information retrieval: a review. *Inf. Process. Manage.*, 41(3):433–455.

Koskenniemi, K. (1983). *Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production*. Number 11 in Publications of the Department of General Linguistics. University of Helsinki.

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2):159–165.

McNamee, P. and Mayfield, J. (2002). Comparing cross-language query expansion techniques by degrading translation resources. In *SIGIR '02: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 159–166, New York, NY, USA. ACM Press.

Molina-Salgado, H., Moulinier, I., Knudson, M., Lund, E., and Sekhon, K. (2002). Thomson legal and regulatory at CLEF 2001: Monolingual and bilingual experiments. In *CLEF '01: Revised Papers from the Second Workshop of the Cross-Language Evaluation Forum on Evaluation of Cross-Language Information Retrieval Systems*, pages 226–234, London, UK. Springer-Verlag.

Munteanu, D. S. and Marcu, D. (2005). Improving machine translation performance by exploiting non-parallel corpora. *Comput. Linguist.*, 31(4):477–504.

Nie, J.-Y. (1998). TREC-7 CLIR using a probabilistic translation model. In *TREC 7*, pages 482–488.

Nie, J.-Y., Simard, M., Isabelle, P., and Durand, R. (1999). Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the web. In *SIGIR '99: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 74–81, New York, NY, USA. ACM Press.

Oard, D. W. and Diekema, A. R. (1998). Cross-language information retrieval. *Annual review of Information Science and Technology (ARIST)*, 33:223–256.

Peters, C. (2006). What happened in CLEF 2006? In Peters, C., Clough, P., Gey, F. C., Karlgren, J., Magnini, B., Oard, D. W., de Rijke, M., and Stempfhuber, M., editors, *CLEF 2006*, volume 4730 of *Lecture Notes in Computer Science*, pages 1–10. Springer.

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*, pages 55–63, New York, NY, USA. ACM Press.

Pirkola, A. (1999). *Studies on Linguistic Problems and Methods in Text Retrieval*. PhD thesis, University of Tampere.

Pirkola, A., Hedlund, T., Keskustalo, H., and Järvelin, K. (2001a). Dictionary-based cross-language information retrieval: Problems, methods, and research findings. *Inf. Retr.*, 4(3-4):209–230.

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

Pirkola, A., Leppänen, E., and Järvelin, K. (2001b). The RATF formula (Kwok's formula): exploiting average term frequency in cross-language retrieval. *Information Research*, 7(2). Available at http://InformationR.net/ir/7-2/paper127.

Pirkola, A., Toivonen, J., Keskustalo, H., and Järvelin, K. (2006). FITE-TRT: a high quality translation technique for OOV words. In *SAC '06:*

Proceedings of the 2006 ACM Symposium on Applied Computing, pages 1043–1049, New York, NY, USA. ACM Press.

Pirkola, A., Toivonen, J., Keskustalo, H., Visala, K., and Järvelin, K. (2003). Fuzzy translation of cross-lingual spelling variants. In *SIGIR '03: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 345–352, New York, NY, USA. ACM.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, 14(3):130–137.

Porter, M. F. (2001). Snowball: A language for stemming algorithms. Available at http://snowball.tartarus.org/texts/introduction.html, accessed December 4 2007.

Qiu, Y. and Frei, H.-P. (1993). Concept based query expansion. In *SIGIR '93: Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 160–169, New York, NY, USA. ACM.

Rapp, R. (1999). Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 519–526, Morristown, NJ, USA. Association for Computational Linguistics.

Resnik, P. (1999). Mining the web for bilingual text. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 527–534, Morristown, NJ, USA. Association for Computational Linguistics.

Salton, G. (1969). Automatic processing of foreign language documents. In *Proceedings of the 1969 Conference on Computational Linguistics*, pages 1–28, Morristown, NJ, USA. Association for Computational Linguistics.

Salton, G. (1988). *Automatic text processing.* Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage.*, 24(5):513–523.

Salton, G. and McGill, M. J. (1983). *Introduction to Modern Information Retrieval.* McGraw-Hill, Inc., New York, NY, USA.

Savoy, J. (2004). Combining multiple strategies for effective monolingual and cross-language retrieval. *Inf. Retr.*, 7(1-2):121–148.

Schamber, L., Eisenberg, M., and Nilan, M. S. (1990). A re-examination of relevance: toward a dynamic, situational definition. *Inf. Process. Manage.*, 26(6):755–776.

Sheridan, P. and Ballerini, J. P. (1996). Experiments in multilingual information retrieval using the SPIDER system. In *SIGIR '96: Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 58–65, New York, NY, USA. ACM Press.

Shi, L., Niu, C., Zhou, M., and Gao, J. (2006). A DOM tree alignment model for mining parallel data from the web. In *ACL '06: Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the ACL*, pages 489–496, Morristown, NJ, USA. Association for Computational Linguistics.

Singhal, A., Buckley, C., and Mitra, M. (1996). Pivoted document length normalization. In *SIGIR '96: Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 21–29, New York, NY, USA. ACM Press.

Sormunen, E. (1994). *Vapaatekstihaun tehokkuus ja siihen vaikuttavat tekijät sanoma-lehtiaineistoa sisältävässä tekstikannassa.[The effectiveness of free-text searching in full-text databases containing newspaper articles and abstracts].* Number 790 in Research Publications. Technical Research Centre of Finland, Espoo, Finland.

Sormunen, E. (2002). Liberal relevance criteria of TREC : counting on negligible documents? In *SIGIR '02: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 324–330, New York, NY, USA. ACM.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D., and Varga, D. (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *LREC'2006:Proceedings of the 5th International Conference on Language Resources and Evaluation.*

Turtle, H. and Croft, W. B. (1991). Evaluation of an inference network-based retrieval model. *ACM Trans. Inf. Syst.*, 9(3):187–222.

Utsuro, T., Horiuchi, T., Chiba, Y., and Hamamoto, T. (2002). Semi-automatic compilation of bilingual lexicon entries from cross-lingually relevant news articles on WWW news sites. In *AMTA '02: Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation: From Research to Real Users*, pages 165–176, London, UK. Springer-Verlag.

Voorhees, E. M. (2001). Evaluation by highly relevant documents. In *SIGIR '01: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 74–82, New York, NY, USA. ACM.

Voorhees, E. M. (2006). Overview of TREC 2006. In *Proceedings of the 15th Text REtrieval Conference (TREC 2006)*.

Xu, J., Weischedel, R., and Nguyen, C. (2001). Evaluating a probabilistic model for cross-lingual information retrieval. In *SIGIR '01: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 105–110, New York, NY, USA. ACM.

Yang, C. C. and Li, K. W. (2004). Building parallel corpora by automatic title alignment using length-based and text-based approaches. *Inf. Process. Manage.*, 40(6):939–955.

Yang, Y., Carbonell, J. G., Brown, R. D., and Frederking, R. E. (1998). Translingual information retrieval: learning from bilingual corpora. *Artif. Intell.*, 103(1-2):323–345.

Zhang, Y., Huang, F., and Vogel, S. (2005). Mining translations of OOV terms from the web through cross-lingual query expansion. In *SIGIR '05: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 669–670, New York, NY, USA. ACM.