

APPROVAL SHEET

Title of Thesis: Textual Representations for Corpus-Based Bilingual Retrieval

Name of Candidate: Paul McNamee
Doctor of Philosophy, 2008

Thesis and Abstract Approved: _____
Charles K. Nicholas
Professor
Department of Computer Science and
Electrical Engineering

Date Approved: _____

ABSTRACT

Title of Thesis: Textual Representations for Corpus-Based Bilingual Retrieval

Paul McNamee, Doctor of Philosophy, 2008

Thesis directed by: Charles K. Nicholas, Professor
Department of Computer Science and
Electrical Engineering

The traditional approach to information retrieval is based on using words as the indexing and search terms for documents. However, word-based representations have difficulty addressing morphological processes that confound retrieval, such as inflection, derivation, and compounding. One part of this research investigates alternative methods for representing text, including a method based on overlapping sequences of characters called n-gram tokenization. N-grams are studied in depth and one notable finding is that they achieve a 20% improvement in retrieval effectiveness over words in certain situations.

The other focus of this research is improving retrieval performance when foreign language documents must be searched and translation is required. In this scenario bilingual dictionaries are often used to translate user queries; however even among the most commonly spoken languages, for which large bilingual lexicons exist, dictionary-based translation suffers from several significant problems. These include: difficulty handling proper names, which are often missing; issues related to morphological variation since entries, or query terms, may not be lemmatized; and, an inability to robustly handle multiword phrases, especially non-compositional expressions. These problems can be addressed when translation is accomplished using parallel collections, sets of documents available in more than one language. Using parallel texts enables statistical translation of character n-grams rather than words or stemmed words, and with this technique highly effective bilingual retrieval performance is obtained.

In this dissertation I present an overview of the field of cross-language information retrieval and then introduce the foundational concepts in n-gram tokenization and corpus-based translation. Then monolingual and bilingual experiments on test sets in 13 languages are described. Analysis of these experiments gives insight into: the relative efficacy of various tokenization methods; reasons for the effectiveness of n-grams; the utility of automated relevance feedback, in both monolingual and bilingual contexts; the interplay between tokenization and translation; and, how translation resource selection and size influence bilingual retrieval.

Textual Representations for Corpus-Based Bilingual Retrieval

by
Paul McNamee

Thesis submitted to the Faculty of the Graduate School
of the University of Maryland in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2008

To my parents, John P. and Judith R. McNamee. They taught me when I was young.

ACKNOWLEDGMENTS

This dissertation signals the end of my graduate education and I have many to thank. First I must express my gratitude to the members of my examination committee: Tim Finin, James Mayfield (reader), Charles Nicholas (chair), Sergei Nirenburg, and Douglas Oard (reader). They were generous with encouragement and constructive criticism that improved this work.

For over ten years James Mayfield has been my supervisor and colleague and I continue to work with him at the JHU Human Language Technology Center of Excellence in Baltimore. Not only did he serve on my committee, but his inspiration led to my developing an interest in information retrieval, and an enthusiasm for research that led to my pursuing a Ph.D. I have benefitted in many ways from our long research collaboration.

Part-time graduate students can be a challenge to supervise, and I am indebted to my research advisor, Charles Nicholas, for shepherding me through the long process of writing a dissertation.

Bruce Croft (University of Massachusetts) and Liz Liddy (Syracuse University) reviewed my proposal and offered advice at the 2nd SIGIR Doctoral Consortium.

I am particularly grateful to Carol Peters (Italian National Research Council), who has directed the Cross-Language Evaluation Forum workshops over the past nine years, and to the rest of the CLEF community. Without the unique multilingual datasets that were developed at CLEF, this research would have been unfathomably more difficult to conduct, if not impossible, and it would not be nearly as multilingual in scope.

The JHU Applied Physics Laboratory has been supportive of my graduate education, providing tuition support, a stimulating professional environment, and computing resources.

I also owe a great deal to the family and friends who have been supportive during the times that I have been preoccupied with this research. Thank you all.

TABLE OF CONTENTS

DEDICATION	ii
ACKNOWLEDGMENTS	iii
LIST OF FIGURES	ix
LIST OF TABLES	xi
Chapter 1 INTRODUCTION	1
1.1 Motivating CLIR	2
1.2 Mechanisms for Translation	4
1.3 Contributions	6
1.4 Information Retrieval Evaluation	8
1.5 History of CLIR	12
1.6 Foundational Research in CLIR	15
1.7 Statistical Translation Using Parallel Corpora	17
1.8 N-gram Tokenization	19
1.9 Research Overview	21
1.9.1 Alternative Tokenization Methods	23
1.9.2 Subword Translation	23

1.9.3	Summary	25
Chapter 2	RELATED WORK	26
2.1	Character N-grams	26
2.2	Parallel Texts	26
2.3	Phrase-Enhanced Retrieval	28
2.4	Bilingual Retrieval and Phrases	29
2.5	Alternatives to Translation	31
Chapter 3	METHODOLOGY	35
3.1	IR Test Sets	35
3.1.1	Cross Language Evaluation Forum	36
3.1.2	Experiment Design	37
3.2	Parallel Corpora	39
3.3	Retrieval Engine	40
3.3.1	HAIRCUT	40
3.3.2	Language Model Framework	40
3.4	Metrics and Statistical Testing	42
3.5	Other Software	43
Chapter 4	MONOLINGUAL EFFECTIVENESS OF N-GRAMS	44
4.1	Overview	44
4.2	Comparison with Words	46
4.2.1	Observations of N-gram Length	47
4.2.2	Observations of Language Variability	50
4.2.3	Morphological Processes	52

4.3	Comparison with Stemming	53
4.4	Comparison with Unsupervised Morphological Segmentation	55
4.4.1	Morphology Challenge	56
4.4.2	Morfessor	57
4.4.3	Experiments	57
4.5	N-gram Stemming	59
4.6	Conclusions	63
Chapter 5	ADDITIONAL N-GRAM EXPERIMENTS	64
5.1	Relevance Feedback	64
5.2	Skip N-grams	69
5.2.1	Previous Work	70
5.2.2	Examples and Performance Analysis	71
5.2.3	Experiments	73
5.3	Reasons for N-gram Effectiveness	76
5.3.1	Misspellings	76
5.3.2	Word-Spanning N-grams	77
5.3.3	Removing Morphology	77
5.4	Conclusions	83
Chapter 6	BILINGUAL EXPERIMENTS	85
6.1	Bilingual Methodology	85
6.2	Translation Corpora	86
6.2.1	Bible Corpus	87
6.2.2	JRC-Acquis Corpus	89
6.2.3	Europarl Corpus	90

6.2.4	Official Journal of the EU Corpus	90
6.3	Translation	91
6.4	Experimental Results	93
6.4.1	Translation Resource Effect on Retrieval Effectiveness	94
6.4.2	Comparing Tokenization Methods in CLIR	95
6.4.3	Size of Parallel Text	100
6.5	Pre-Translation Query Expansion	104
6.6	Synopsis	108
Chapter 7	CONCLUSION AND FUTURE WORK	110
7.1	Review	110
7.2	Impact	113
7.3	Limitations	114
7.4	Major Findings	115
7.5	Future Directions	115
Appendix A	CLEF BENCHMARKS	117
A.1	CLEF 2002	117
A.2	CLEF 2005	118
REFERENCES	120

LIST OF FIGURES

1.1	Google’s CLIR interface.	3
1.2	A sample precision-recall graph.	11
1.3	Monolingual IR performance as a function of n-gram length in eight languages (CLEF 2003 data). Note that languages here, and throughout the paper are abbreviated using ISO 639 codes (see Table 1.2).	21
2.1	Comparative efficacy of words and n-grams for bilingual retrieval making no attempt at translation.	33
3.1	English Query 62 from CLEF 2001.	38
5.1	Effect of the number of expansion terms on retrieval performance.	67
5.2	MAP variation based on relevance feedback parameter settings in English.	68
5.3	Comparative efficacy of 5-grams against words, when the order of letters in words has, and has not, been scrambled throughout the corpus.	82
6.1	Aligned passages from parallel sources in English, Spanish, and French.	88
6.2	Relative translation corpus effectiveness. English topics and Spanish documents.	94
6.3	Tokenization experiments using two query languages in three target languages.	101
6.4	Performance improvement with corpus growth.	102

6.5	Example of pre-translation query expansion in corpus-based bilingual retrieval.	105
6.6	Improvement in bilingual retrieval using pre-translation expansion and <i>acquis</i>	106
6.7	Improvement in bilingual retrieval using pre-translation expansion and <i>europarl</i>	107

LIST OF TABLES

1.1	IR-related tasks studied at various competition-style workshops. Two stars indicate a cross-language component.	14
1.2	Codes for language names from ISO 639-1.	22
1.3	4-gram translations for <i>milk</i> and <i>olympic</i>	24
3.1	Cross-language test collections.	36
3.2	Number of judged queries in CLEF test sets (2000-2007).	37
3.3	Number of topics with few relevant documents.	38
3.4	Large parallel collections.	39
4.1	Document frequencies for 5- and 6-grams in <i>real estate</i>	45
4.2	Comparing words and n-grams of various lengths.	48
4.3	5-gram effectiveness and linguistic complexity.	51
4.4	Morphological properties of CLEF languages.	54
4.5	Number of rewrite rules for Snowball implementations.	55
4.6	Comparing words, 4-grams, and 5-grams to Snowball stems.	55
4.7	Selected results (MAP) from Morpho Challenge 2007.	56
4.8	Data for unsupervised morphology experiments.	58
4.9	Comparing words, 4-grams, and 5-grams to Morfessor segments.	59

4.10	Average posting list length for words and n-grams (CLEF 2002 data).	60
4.11	Constituent n-grams by document frequency for <i>precaution</i> .	61
4.12	Examples of n-gram stemming.	62
4.13	Effectiveness of 7 tokenization methods, including n-gram stemming.	63
5.1	Relative gain from automated relevance feedback by tokenization method.	69
5.2	The templates for producing skipgrams with four letters and two skips.	72
5.3	Examples of skipgrams generated from ‘_crust_’.	72
5.4	Number of postings generated from skipgrams for <i>sesquipedalian</i> .	73
5.5	Skipgram results with 3 preserved letters.	74
5.6	Skipgram results with 4 or 5 preserved letters.	74
5.7	Gain from word-spanning n-grams.	78
5.8	Sample word transformations (CLEF 2000 English corpus).	79
5.9	Distinct indexing terms (CLEF 2000 English corpus).	80
5.10	Change observed by scrambling the letters in words.	81
6.1	Parallel texts used in experiments.	87
6.2	Bible versions used.	89
6.3	Sample one-best translations for <i>nuclear energy</i> .	93
6.4	Word-based CLIR using English topics and various aligned corpora.	96

6.5	Tokenization effects with English topics and <i>acquis</i> corpus.	97
6.6	Tokenization effects with English topics and <i>europarl</i> corpus.	98
6.7	Suitable topics for bilingual experiments on English documents.	99
6.8	Tokenization using <i>acquis</i> corpus bilingual retrieval on English documents.	99
6.9	Tokenization using <i>europarl</i> corpus bilingual retrieval on English documents.	99
6.10	Size of English <i>europarl</i> subsets in words.	103
6.11	Efficacy of pre-translation query expansion.	108
A.1	Comparison with CLEF 2002 monolingual results.	118
A.2	Comparison with CLEF 2002 bilingual results.	118
A.3	Comparison with CLEF 2005 monolingual results.	119

Chapter 1

INTRODUCTION

Trends in the past decade have increased the importance of robust and effective multilingual processing. Since the advent of the World Wide Web, growth in non-English content and web sites has dramatically increased the linguistic diversity of the Internet. Standards such as Unicode (2003), a means of representing character encodings for most of the world's languages, have received wide acceptance. Economic changes such as globalization, the outsourcing of jobs, and reliance on immigrant workers have fueled demand for multilingual technology. Political changes such as EU enlargement have increased the need for translation and management of multilingual data. Finally, the unpredictable nature of military and humanitarian crises (*e.g.*, conflicts in Serbia and Croatia, war in Afghanistan, genocide in Sudan, and tsunami relief in southeast Asia) demonstrate a continuing need for tools to process less commonly taught languages.

Cross-language information retrieval (CLIR) is concerned with the organization and retrieval of unstructured text where user queries can be expressed in a language different than the one(s) in which documents are written. Many interesting research issues arise, primarily as to the role and mechanism of translation, but also including how to summarize and present foreign language content to a non-fluent user, how the unique linguistic properties of each language should be addressed, and how repositories in multiple languages can

be searched and have results presented to the user in a transparent fashion. In this dissertation I am interested in the translation problems that arise in bilingual retrieval and solutions that scale across many language pairs.

1.1 Motivating CLIR


The undeniable utility of high-speed, good-quality information retrieval (IR) systems is revealed by the hundreds of millions of user queries submitted daily to commercial web search engines. The major search engines have been slow to provide support for querying in one language to obtain documents from multiple languages; however, in May 2007 Google released a prototype multilingual web search tool¹ (see Figure 1.1).

Several important communities require CLIR capabilities. Governments and geopolitical entities with diverse, multilingual populations (*e.g.*, Canada, India, Switzerland, the United Nations, and the European Union) can benefit from an ability to provide a uniform means of access to publications and records regardless of a user's native language. In the United States, there are immigrants with poor English skills who can benefit from software that enables information retrieval in their native language for purposes such as access to government assistance or health care. Intelligence analysts, both military and business, need an ability to search foreign language data. Medical, scientific, and legal researchers are interested in foreign publications and patent attorneys need to be able to search foreign archives.

CLIR is still beneficial even if no capability to translate retrieved documents in the user's native language is available. One's ability to read a foreign language is often greater than one's ability to write (*i.e.*, compose grammatical and correctly spelled queries); thus CLIR is helpful to those with some limited foreign language skills. And if an end-user

¹http://translate.google.com/translate_s

[Help](#)


[Text and Web](#)
[Translated Search](#)
[Dictionary](#)
[Tools](#)

Translated Search

Search for:
 Translated to: **Japón terremoto** - [Not quite right? Edit](#)

My language:
 Search pages written in:

[Translate and Search](#)

Translated results from Spanish web pages
Results 1 - 10 of about 678,000 for **Japón terremoto**.

<p>English translation</p> <p>An earthquake in Japan caused a radioactive leak at a nuclear plant... An earthquake in Japan caused a radioactive leak at a Japanese nuclear power plant, located in one of the most active seismic zones of the world, ... www.elmundo.es/elmundo/2007/07/16/internacional/1184557283.html - 30k - Cached</p> <hr/> <p>elmundo.es - A strong quake shakes Japan A strong earthquake of magnitude 7 on the Richter scale has shaken for two minutes northeastern Japan, in the area of Iwate and Miyagi, ... www.elmundo.es/elmundo/2003/05/26/sociedad/1053941654.html - 45k - Cached</p> <hr/> <p>ENERGY-JAPAN: Earthquake shakes nuclear expansion plans The radioactive leak at a nuclear power plant after the earthquake that struck Niigata Prefecture in Japan, triggered public alarm and arrested for... ipsnoticias.net/nota.asp?idnews=41549 - 60k - Cached</p>	<p>Original Spanish - Hide Spanish results</p> <p>Un terremoto en Japón causa una fuga radiactiva en una central ... Un terremoto en Japón causa una fuga radiactiva en una central nuclear Japón, que se encuentra en una de las zonas sísmicas más activas del mundo, ... www.elmundo.es/elmundo/2007/07/16/internacional/1184557283.html - 30k - En caché</p> <hr/> <p>elmundo.es - Un fuerte seísmo sacude Japón Un fuerte terremoto, de magnitud 7 en la escala de Richter ha sacudido durante dos minutos el noreste de Japón, en la zona de Iwate y Miyagi, ... www.elmundo.es/elmundo/2003/05/26/sociedad/1053941654.html - 45k - En caché</p> <hr/> <p>ENERGÍA-JAPÓN: Terremoto sacude planes de expansión nuclear Las fugas radioactivas en una planta nuclear tras el terremoto que sacudió la prefectura de Niigata, en Japón, desataron alarma pública y detuvieron de ... ipsnoticias.net/nota.asp?idnews=41549 - 60k - En caché</p>
--	--

FIG. 1.1. Google's CLIR interface.

cannot read a foreign language document, a high quality CLIR system can still identify the documents that are worth translating manually.

I tender as axiomatic the contention that better IR performance is desirable as it will lead to streamlined information access and reduce effort by users to find the information they desire. The question then is how can, and how should better bilingual retrieval performance be obtained? The answer will depend on how the best monolingual retrieval can be obtained, on how translation of queries (or documents) can be performed, and on the interaction between these two components. It is not clear whether monolingual retrieval can best be accomplished through IR systems and techniques carefully engineered for individual languages, or through systems that are only mildly customized for new languages, or by relying on approaches that are maximally language-neutral. If language-neutrality can result in equally effective IR performance this would be quite desirable, as it would reduce the amount of software engineering required to facilitate each new language. Recent studies have shown that a critical factor in improving cross-language retrieval is the quality of available translation resources (McNamee & Mayfield 2002; Demner-Fushman & Oard 2003). Obtaining and integrating translation resources for multiple languages is a difficult task, one complicated by language diversity. For example, different bilingual translation dictionaries may lemmatize (*i.e.*, normalize) word forms according to different rules and compounding languages will have rules or devices for incorporating multiword expressions.

1.2 Mechanisms for Translation

Three principal resources are used to accomplish translation for cross-language information retrieval: bilingual wordlists, mappings induced from aligned parallel corpora, and machine translation systems. Each is briefly summarized below.

Bilingual dictionaries sometimes include for each lemmatized form, information such as pronunciation, part of speech, and a definition, in addition to translations. However, simpler wordlists that only provide a mapping from each surface form to one or more translation equivalents are more commonly available in electronic form. Use of general-purpose wordlists creates several problems, including translation ambiguity, where the correct translation is uncertain, and poor coverage of multiword phrases. As a result, *palm tree* could be literally mistranslated into French *paume* (part of a hand) and *arbre* (tree), rather than *palmier*, confounding an IR system. Finally, most wordlists are unlikely to contain translations of names of people and locations. Pirkola *et al.* (2001) analyze the issues endemic to dictionary-based retrieval and report:

The main problems associated with dictionary-based CLIR are (1) untranslatable search keys due to the limitations of general dictionaries, (2) the processing of inflected words, (3) phrase identification and translation, and (4) lexical ambiguity in source and target languages.

Despite these problems, the application of bilingual wordlists to CLIR is computationally efficient (*i.e.*, table lookup is fast) and does not require extensive preprocessing as the use of corpora does. Furthermore, it is easy to add new translations to wordlists as they become available.

Parallel corpora are sets of documents with translations in another language. Common sources include religious or famous literary texts and publications of multinational organizations or governmental bodies with multilingual populations (*e.g.*, the EU, Canada, Switzerland). If a large number of documents can be aligned, that is, identified with their corresponding foreign language documents, then statistical methods can be applied to identify word correspondences. Difficulties in using corpora for CLIR include finding sufficiently large collections, aligning documents, obtaining good lexical coverage for expected queries or documents, and selecting a method for translation of terms. Regarding the last,

it is also the case that the translations produced from corpora have a degree of lexical uncertainty about them that is not an issue with manually compiled dictionaries. Parallel corpora were once rare and difficult to obtain in quantity, therefore bilingual dictionaries were more commonly used in CLIR, however, with increased availability this will become less the case.

Machine translation (MT) systems can potentially provide a grammatical, correct translation; however, such high quality MT is not commonplace. They are perhaps better able to capture longer sentence structure than the alternatives above, a fact reflected in the word n-gram evaluation models used for MT systems such as BLEU (Papineni *et al.* 2002). From a system developer's point of view machine translation is simple to use, as it is only a preprocessing step before monolingual retrieval. Historically MT systems have not been able to produce alternative translations that could be used to improve translation for the purpose of IR, but modern statistical systems no longer have this limitation. In the future, if fully automated, high-quality machine translation becomes widely available, then it may become the dominant approach in CLIR.

Common to any automated method of translation is the fact that lexical coverage is limited, thus any CLIR system will be faced with some out-of-vocabulary (OOV) words for which a translation is unknown. Without a robust method for coping with untranslatable terms there will remain some queries to which a CLIR system cannot appropriately respond.

1.3 Contributions

This research investigates alternative representations for text that can lead to improvements in retrieval efficacy, monolingually and bilingually. Rule-based stemming, statistical stemming, and character n-gram tokenization are considered as methods for addressing

morphological variation in monolingual settings. Then translation of subword units (*i.e.*, character n -grams) is explored using techniques based on parallel corpora. This method speaks directly to problems identified in dictionary-based bilingual retrieval, namely, coping with morphological variation, out-of-vocabulary words, and multiword phrases. The following claims will be investigated:

1. Effective multilingual text retrieval can be achieved without the costs and complexities introduced by language-specific processing.
2. Indexing using character n -grams is effective because n -grams provide lexical normalization, and the benefit of n -gram indexing is greatest in languages with high morphological complexity.
3. In cross-language information retrieval, translation need not be performed at the word level.
4. In corpus-based bilingual retrieval the relative advantage from using character n -grams as both indexing terms and units of translation is inversely proportional to resource size and quality.

In the balance of this chapter I shall: (1) explain how information retrieval performance can be formally evaluated; (2) describe the cross-language information retrieval problem; (3) introduce the foundational concepts of statistical translation and n -gram tokenization; (4) and, outline solutions to the translation problems in bilingual retrieval, which are explored in this dissertation. In Chapter 2 previous research that bears on corpus-based translation is reviewed. Chapter 3 describes the data sets and research methodology that will be followed. Chapter 4 compares character n -gram indexing to alternative methods for tokenization in a monolingual retrieval setting. Chapter 5 reports on additional experiments

that further illuminate reasons why n-grams are an effective choice of indexing method. Experiments in improving bilingual retrieval through the use of n-grams and other techniques are presented in Chapter 6. In Chapter 7 the main contributions of this dissertation are reviewed and discussed.

1.4 Information Retrieval Evaluation

Empiricism is the byword of modern information retrieval research. While user-centric evaluations are important and continue to this day, they are difficult to reproduce and are expensive, both financially and in human labor, to conduct. Studies that undertake to show improvements in retrieval performance (*i.e.*, accuracy, not efficiency) are usually based on fixed sets of queries that are run against a static collection of documents. The idea of controlling IR test collections in this way came from protocols developed by Cleverdon for the purpose of comparative evaluation of algorithms and systems (1967). At the time Cleverdon and his peers worked with IR test collections containing only a few thousand documents, which could be completely judged against a small set of queries.

An IR test collection consists of a set of documents, a set of user needs, or topics, and a set of judgments establishing which documents are and are not relevant for a given topic. Because of the significant expense of constructing exhaustive relevance judgments, the concept of pooling ranked lists from multiple retrieval systems to create shorter lists of documents that can be examined for relevance was introduced by several British researchers (Gilbert & Spärck Jones 1979). Since only top-ranked documents are evaluated, a substantial savings occurs as most documents are never examined for relevance to any topic. These enriched pools of documents presumably identify a large number of relevant documents and so become useful for comparative system evaluation.

In fact, this type of semi-automated judging is the dominant technique for developing

IR test sets today and is the principal method undertaken at the Text REtrieval Conference (TREC) series of annual workshops. TREC is a U.S. government program that invites researchers from around the world to participate in evaluations of IR systems (Voorhees & Harman 2005). TREC is organized by the National Institute for Standards and Technology (NIST) – in 2008 it is in its seventeenth year. When the conference began in 1992 it had a charter to explore how IR techniques of the time would succeed on the large-scale collections that were then becoming possible. In comparison to early test collections of approximately 3 MB of text, TREC collections have grown from 2 GB (in 1992), to 100 GB, to half of a terabyte in 2004. Collection size is now limited mainly by acquisition issues involving intellectual property and privacy. Pooling has been adopted by other large-scale IR evaluation workshops that were patterned after the TREC model, including: the Cross-Language Evaluation Forum workshop (CLEF), which is organized in Europe; the NII-NACSIS Test Collection for Information Retrieval Systems (NTCIR), which is funded by the Japanese government; and, the Forum for Information Retrieval Evaluation (FIRE), an initiative to study retrieval in Indian subcontinent languages.

There has been extensive experimentation to determine whether the technique of pooling successfully creates a test collection in which new IR systems can evaluate their techniques. The primary concern is that a system which did not contribute documents to the original pools may be unfairly penalized. Zobel set out to determine whether this was in fact the case and he ended up concluding that the document pools in the TREC collections were suitable for unbiased post-hoc system evaluation (1998). Researchers at NIST have also justified the use of pooling by demonstrating stability in system rankings over multiple trials that discount the relevant documents contributed by a single group (Voorhees & Harman 1998). More recently Sanderson and Zobel (2005) have posed the question of whether shallow document pools created for a larger set of document queries can lead to a more cost-effective IR test collection with greater sensitivity.

The notion of document relevance has generally been considered a binary relationship (*i.e.*, a document is relevant or not); however, there have been some proposals to establish multigrade relevance judgments that differentiate between highly relevant and only marginally relevant documents. Voorhees, the TREC project manager at NIST, has maintained that binary judgments are sufficient for comparative system evaluation. Though some researchers are concerned about annotator errors and inconsistencies in judging documents – inter-assessor agreement is typically only about 70% – it has been shown in the TREC evaluations that system rankings are stable despite errors in the assessment process (Voorhees 2000; Buckley & Voorhees 2004).

Given a set of relevance judgments one must determine how best to rate system performance. Each IR system produces a ranked list of documents, which is scored against the known judgments. The two metrics of *precision* and *recall* measure different aspects of retrieval performance. Precision measures the percentage of returned documents that are deemed relevant; it can be measured at fixed document levels, for example, after 10 or 50 documents have been examined. Recall quantifies the percentage of the relevant documents that are discovered in the ranked lists of retrieved documents. The two measures are somewhat at odds with one another as one can achieve good recall by constructing an enormous list of documents; however such a list would have very low precision. Conversely, a short list of high-scoring documents may have high precision but fail to find many relevant documents. Which measure is more important depends on the end application. For example, medical researchers or trial attorneys may put a high premium on finding essentially all documents that bear on their information request. On the other hand, with the advent of commercial web search, there has been a recent tendency to focus on precision at the top ten or so documents.

Precision can be plotted against recall to produce a precision-recall graph (see Fig. 1.2). These graphs are ubiquitous in IR research and convey graphically how a system or

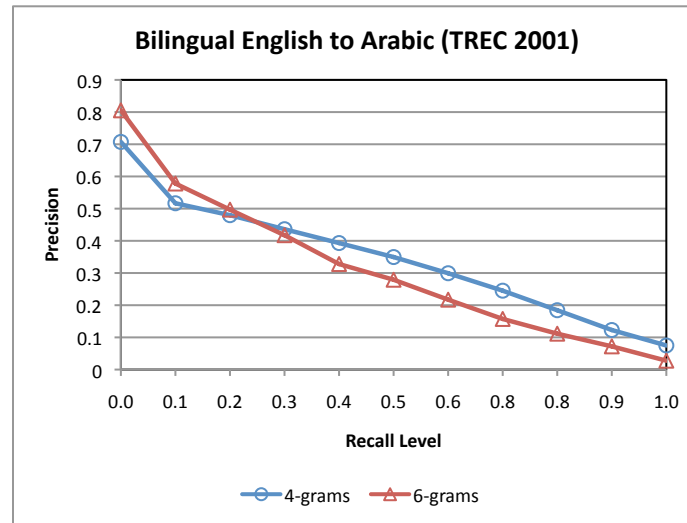


FIG. 1.2. A sample precision-recall graph.

algorithm performs. For example, in the figure the curve labeled 4-grams performs worse on the left side of the graph, indicating that fewer relevant documents are found in the high precision region of the graph. On the other hand, there is a crossover point beyond which 6-grams do comparatively worse at finding relevant documents. One thing that is not immediately obvious is how to best compare two curves and ascertain which one represents the more desirable performance. A widely adopted measure of retrieval performance was engineered to simplify comparative evaluation by combining the information from a precision-recall graph into a single numeric quantity. Average precision is the average of precision values determined after each relevant document is observed. It can be considered a rectangle rule approximation to determine the area under the precision-recall graph using numerical integration. Typically average precision is itself averaged over a set of different topics to produce mean average precision (MAP), the standard evaluation measure reported at TREC and by nearly all IR researchers.

1.5 History of CLIR

Like IR itself, CLIR research began decades ago. In a 1998 summary Oard and Diekema (1998) cite initiatives between 1969 and 1973 by Pigur, Pevzner, and Salton, that relied on multilingual thesauri to map query terms into another language. At that time the debate between controlled vocabulary search and free text search was ongoing and the use of controlled terminology was common. The field blossomed in the mid-1990s and in the next decade sessions on CLIR became commonplace at SIGIR conferences, a technical book on the subject was published (Grefenstette 1998), and several multilingual evaluations were held at TREC and elsewhere. Of course in this timeframe computer memory and hard disk space became dramatically less costly and larger collections of electronic texts became available. At the same time multinational corporations became increasingly interested in managing their vast repositories of text.

Multilingual information access includes several related subfields. One is how to best perform monolingual retrieval in multiple or diverse languages. Much early work was done in this area. Another area is bilingual retrieval where one is concerned with IR between a source language and a target language. Finally, there is the issue of how to search a heterogeneous collection containing documents in many languages. For example, an English-speaking patent attorney may wish to find all international patents bearing on a particular invention, regardless of the language of the original patent. This last task differs from simpler bilingual retrieval in desiring to impart a single ranking on documents from multiple languages. The distributed IR flavor of this task creates problems because many traditional systems produce scores that are dependant on the collection used; when multiple collections are searched relevance scores may not be commensurable.

At TREC investigations into monolingual search in Spanish and Chinese were undertaken between 1994 and 1997. From 1997 to 1999 a track that explored cross-language

retrieval in four European languages, English, French, German, and Italian, was run. At that time there were concerns in the IR community that multilingual research at TREC was not attracting a sufficiently large number of international participants; most CLIR track participants were based in North America. Two international workshops emerged in 1999 and 2000 to encourage a greater diversity of research groups and to focus on regionally significant languages. The first, NTCIR (NII-NACSIS Test Collection for Information Retrieval Systems) is organized by the National Institute for Informatics (NII) in Tokyo. It focuses on Asian languages including Chinese, Japanese, Korean, as well as English. Since 1999, NTCIR has been run along the same lines as TREC and has held workshops approximately every 18 months. The other major evaluation workshop is CLEF (Cross Language Evaluation Forum), which is based in Europe. Based on the initial foray into European languages at TREC, over the past nine years CLEF has developed document collections of journalistic text in 13 European languages: Bulgarian, Czech, Dutch, English, Finnish, French, German, Hungarian, Italian, Portuguese, Russian, Spanish, and Swedish. Queries are available in the languages of the CLEF documents and occasionally in other languages, including: Amharic (one of the dominant languages in Ethiopia), Chinese, Greek, Indonesian, Korean, Japanese, and Thai.

Smaller initiatives have focused on particular languages, though not generally on a recurring annual basis. AMARYLLIS was an IR evaluation workshop focused on the French language (Landi *et al.* 1998). Retrieval in Russian was been studied at the ROMIP/RIRES evaluation (Dobrov *et al.* 2004). After the formation of NTCIR and CLEF, CLIR research at TREC continued for several years and studied both Chinese and Arabic, before ending in 2002. An evaluation for Indian languages called FIRE is scheduled for late 2008 and it will include test sets in Hindi, Bengali, and Marathi.

The NTCIR and CLEF evaluations are organized in a distributed fashion. Topics are created and then translated by a team of translators who generally are not participants in

	TREC	NTCIR	CLEF	TDT	DUC	MUC	TIDES	ACE
Ad Hoc IR	**	**	**				**	
Filtering	*		**	**			**	
Question Answering	*	**	**					
Speech Retrieval	*		**	**				
Text Summarization		*			*		**	
Video / Image	*		**					
Web Search	*	*	**					
Information Extraction						**	**	**

Table 1.1. IR-related tasks studied at various competition-style workshops. Two stars indicate a cross-language component.

an ongoing evaluation. After system responses are submitted relevance assessments are conducted by native speakers. This is in contrast to assessment at TREC where assessors often work in a uniform environment at a single location.

In addition to the ad hoc IR tasks conducted at TREC, NTCIR, and CLEF, each of these workshops has undertaken tracks that focus on other human language technologies (HLT) or other variants of IR. Table 1.1 lists for a variety of human language processing tasks the workshops that have held formal, open evaluations. The table is not complete; it leaves out several conferences (*e.g.*, CoNLL and SEMEVAL) and many areas in HLT such as automated speech recognition (ASR), machine translation, parsing, and word sense disambiguation. One interesting thing to note about the DARPA TIDES evaluation (Oard 2003) is that a surprise language evaluation was run in the languages of Cebuano and Hindi. In this evaluation participants were given only several weeks to amass linguistic resources and build functioning systems, including IR systems. This task illustrated the effort required to develop capabilities in less studied languages. More recently the Linguistic Data Consortium (LDC) developed ‘language packs’ for about a dozen minority languages (Simpson *et al.* 2008).

1.6 Foundational Research in CLIR

With the increasing availability of large, non-English test collections and the emergence of well-organized international evaluation workshops modeled after TREC (*e.g.*, CLEF and NTCIR), a significant body of research has been undertaken on the foundational issues for CLIR.

The issue of whether queries or documents should be translated was examined by McCarley (1999). Though he concluded optimal results would come from translating both, query translation is the dominant approach today. Document translation requires selecting the query language of interest in advance, which may be difficult in some applications. In addition to running MT software on an entire document collection, it is possible to translate documents word-by-word in the same manner as query translation is performed. McNamee and Mayfield examined the latter, which can be done in time linear in the collection size, and they obtained similar performance to query translation (McNamee & Mayfield 2003).

Due to the losses incurred by translation ambiguity and out-of-vocabulary (OOV) words, bilingual retrieval performance is expected to be worse than monolingual performance. In the late 1990s relative bilingual performance of 60-80% was considered good, and more recently results around 90% have become expected. To reduce the adverse effects encountered during translation, Ballesteros and Croft introduced pre-translation query expansion (1997), a technique where an initial query is expanded into a larger set of terms in the source language, all of which are then translated.

One problem encountered during translation is due to the fact that some words may have many different translations, while other words have only one or two possibilities. If all potential translations are used, the resulting query in the target language can become unbalanced and reflect greater weight on concepts with many acceptable translations. Pirkola *et al.* promote translation of structured queries to control this effect (2003). Others attempt

to select the single best translation alternative or to probabilistically weight translations (Monz & Dorr 2005).

The fact that translation resources are generally difficult to obtain has motivated investigation into how effective retrieval can be accomplished with limited resources. Ballesteros and Croft described experiments in transitive translation, whereby an intermediate language was used in the translation process (1998). This supports translation when resources for a particular language pair are not obtainable but both the source and target language can be mapped to a common language. Gollins and Sanderson introduced a refinement to this approach using multiple intermediate languages to reinforce the appropriate translations (2001).

Following work by Buckley *et al.* (2000), some have investigated using orthographic rules for pseudo translation (Toivonen *et al.* 2005) or attempting retrieval absent any language resources (McNamee & Mayfield 2004). The latter method appears to provide serviceable retrieval performance (*e.g.*, 50% of a good bilingual run), but only between related languages. More practically, investigations of how performance varies as a function of lexical coverage have been undertaken by several groups (Xu & Weischedel 2000; Franz *et al.* 2001; McNamee & Mayfield 2002; Demner-Fushman & Oard 2003). In their paper, Xu and Weischedel suggest that translations beyond the most common 30,000 words add little value, on average.

Synthesis of multiple translation sources was examined by Kraaij during the second CLEF evaluation (2001). He found that a combination of MT, bilingual wordlists, and corpora could achieve bilingual performance 97% of a monolingual baseline, significantly better than results obtained using any single resource.

Kishida has written a detailed summary of recent research in CLIR (Kishida 2005).

1.7 Statistical Translation Using Parallel Corpora

Today, in the early years of the information age, there are a seemingly endless number of electronic texts present on the Internet, some of which occur with translations. The field of machine translation underwent an upheaval in the mid-1990s as large parallel texts became available. Prior to that time the dominant approach was to use bilingual dictionaries and lexical ontologies to produce translations. In contrast to this knowledge-intensive paradigm, statistical machine translation (SMT) systems require as input parallel texts and very little additional information (Brown *et al.* 1993). Because the goal of a machine translation system is to produce a syntactically correct (and semantically appropriate) rendering of the input into the desired target language, it is important for an SMT system to translate phrasal structure correctly. In contrast, for the purpose of finding relevant documents, accurately translating concepts (*i.e.*, words or phrases) is sufficient. In other words “green ideas sleep furiously” and “sleep ideas furiously green” are equivalent in the bag-of-words model, which pervades IR systems.

SMT systems first preprocess parallel texts. A critical step is to split the input into sentences and to then align sentences with their translations. Sentences cannot always be put in a one-to-one correspondence with translations. After blocks of one, two, or three sentences are aligned with their mates in the other language a SMT system will try to learn the concepts of word order and fertility. Within a sentence words can appear in different order than their translated equivalents; for example, in Spanish adjectives follow a noun while in English adjectives usually precede the noun that they modify. Systems learn to transfer structure from the source to the target. Fertility describes how one word can be replaced with zero, one, or several words in a foreign language. Finally, the translation systems must decide how individual words are translated. As we mentioned above, translation for IR is simpler than translation for MT. With IR, we are principally concerned with

computing a list of the most likely candidate translations for a word.

Potential translations are proposed for individual words based on counting the frequencies of occurrence of the source language word and the target language words that appear in the translations of documents containing the source word. Likelihood scores, or probabilities can be determined using several information theoretic measures. Suppose in a collection of 10,000 aligned English/French sentence pairs the word *dog* appears 200 times (*i.e.*, in 2% of documents). In the 200 or so French documents aligned with the English documents containing *dog* we can count the frequencies of occurrence of each word that occurs. Suppose that *chien* appears in 225 documents in the entire French collection and 175 times in the 200 documents of interest. Pointwise Mutual Information (PMI) is one measure that can be used to score *chien* as a possible translation for *dog*. PMI is the log of the ratio between the joint probability and the product of prior probabilities. Besides PMI, one might use Cosine, Dice scores, the Chi-squared statistic, symmetric conditional probabilities, or other related measures (Och & Ney 2000). An example scoring the word *chien* is given below:

$$\begin{aligned}
 PMI(dog, chien) &= \log \left(\frac{P(dog, chien)}{P(dog) \times P(chien)} \right) \\
 &= \log \left(\frac{\frac{175}{200}}{\frac{200}{10000} \times \frac{225}{10000}} \right) \\
 &= \log \left(\frac{0.875}{0.0200 \times 0.0225} \right) \\
 &= 10.925
 \end{aligned}$$

Thus, *chien* receives a score of 10.925. Pointwise mutual information is an unbounded metric where larger values indicate greater degrees of association.

1.8 N-gram Tokenization

Character n-grams, sequences of n consecutive characters², have been used for a multiplicity of tasks in human language technology, including: spelling correction (Zamora, Pollock, & Zamora 1981); personal name matching (Zobel & Dart 1995), diacritics restoration (Mihalcea & Nastase 2002); and language identification (Cavnar & Trenkle 1994). Their use for IR dates to the mid-1970s when they were used primarily as a technique to decrease dictionary size. At that time $n = 2$ or $n = 3$ were typical lengths, and for a fixed alphabet size α , α^3 is almost certainly smaller than the size of a lexicon, which IR systems store in memory. Over time as memory constraints became less severe, research in the mid-1990s led to n-grams being considered as an alternative indexing representation to words or stemmed words (see (Damashek 1995)). There are many variations on n-gram indexing; in this dissertation I concentrate on overlapping character n-grams of a fixed length (typically $n = 4$ or $n = 5$). For the text `prime_minister` and $n = 7$ the resulting n-grams are: `_prime_`, `prime_m`, `rime_mi`, `ime_min`, `me_mini`, `e_minis`, `minist`, `ministe`, `inister`, and `nister_`. The single n-gram `ime_min` that occurs at the word boundary is a fairly distinct indicator of the query phrase ‘prime minister’ and it would not be generated from a sentence like ‘the finance minister ordered prime rib for lunch’ which might generate a false match using words alone as indexing terms.

A significant amount of work has been conducted comparing language-specific methods for tasks such as segmentation, stemming, and compounding. In earlier work with Mayfield (2004), I have promoted the use of character n-gram tokenization to accomplish these tasks in a language-neutral and easy to implement fashion. Our consistently high results in the CLEF evaluations suggest the method has promise. The use of character n-

²The term n-gram has two different meanings in human language technology. Unless otherwise indicated the term is used in this dissertation to denote sequences of characters, not sequences of words.

grams has not been considered a mainstream technique, except in some Asian languages where large ideographic character sets and unsegmented words create special problems. N-grams are commonly used in Chinese (Chen *et al.* 1997), Japanese (Ogawa & Matsuda 1997), and Korean (Lee & Ahn 1996). Following our earlier work with n-grams and our successes using them on European languages at CLEF, several European research groups have begun more seriously considering them, including the University of Neuchâtel (Savoy 2003), the University of Amsterdam (Hollink *et al.* 2004), and the University of A Coruña (Vilares, Oakes, & Ferro 2007).

While proponents cite their language independence and surrogate morphological normalization, n-gram tokenization can create problems in terms of accuracy and computational expense. The use of n-grams can cause unintentional conflation (primarily with shorter length n-grams). For example, the 4-gram *mini* could be generated from words like *dominion*, *feminist*, *miniature*, or *ministry*, among others. Still, while individual n-grams may be rather ambiguous, a set of n-grams is far less so. Moderate length n-grams, the ones that attain the highest IR performance, require more than five times the disk space and processing time of words or stemmed words. This performance issue is deserving of additional study (see (Mayfield & McNamee 2003; Miller *et al.* 2000)).

The most common practice is to select a single length of n and to produce substrings that overlap by a single character. Some practitioners have tried non-overlapping n-grams or select only those n-grams that do not span a word boundary. In my experience fixed length n-grams of length 4 or 5 that span word boundaries and overlap by a single character work well across a variety of alphabetic languages. In Figure 1.3, performance as a function of n-gram size is shown in several European languages. Relevance feedback was not employed in these runs; however it can indeed be used with effect.

Looking at the data in Figure 1.3, it is apparent that 5-grams outperform simple words,

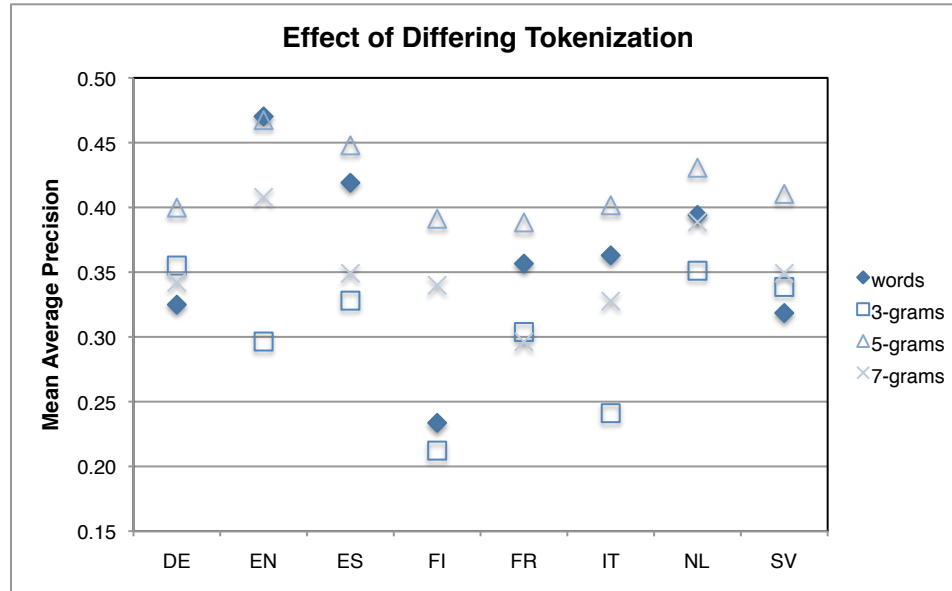


FIG. 1.3. Monolingual IR performance as a function of n-gram length in eight languages (CLEF 2003 data). Note that languages here, and throughout the paper are abbreviated using ISO 639 codes (see Table 1.2).

except for English (EN). In the more morphologically complicated languages (*e.g.*, German and Finnish) n-grams have a decided advantage.

The convention used throughout this manuscript is to use abbreviations for language names using the two-letter codes described in the international standard ISO 639-1:2002³. In Table 1.2, codes are listed for languages that have been used in CLEF, NTCIR, or TREC evaluations, along with information about the number of native speakers (Katzner 1999).

1.9 Research Overview

The research in this dissertation aims to improve monolingual and bilingual retrieval performance through improvements in tokenization and novel exploitation of parallel cor-

³<http://www.iso.org/>

Digraph	Language	Speakers (millions)	Major Countries Where Spoken
AM	Amharic	20	Ethiopia
AR	Arabic	215	Algeria, Bahrain, Egypt, Iraq, Jordan, Kuwait, Lebanon, Libya, Morocco, Oman, Qatar, Saudi Arabia, Sudan, Syria, Tunisia, United Arab Emirates, Yemen
BG	Bulgarian	8	Bulgaria
CS	Czech	10	Czechoslovakia
DE	German	100	Austria, Germany, Switzerland
EL	Greek	10	Cyprus, Greece
EN	English	350	Australia, Bahamas, Canada, Great Britain, Ireland, Jamaica, New Zealand, United States
ES	Spanish	325	Argentina, Belize, Bolivia, Chile, Costa Rica, Cuba, Dominican Republic, Ecuador, El Salvador, Guatemala, Honduras, Mexico, Nicaragua, Panama, Paraguay, Peru, Spain, Uruguay, Venezuela
FI	Finnish	5	Finland
FR	French	75	Belgium, Canada, France, Switzerland
HU	Hungarian	10	Hungary, Rumania
ID	Indonesian	150	Indonesia
IT	Italian	60	Italy, Switzerland
JP	Japan	125	Japan
KR	Korean	65	Korea
NL	Dutch	15	The Netherlands
PT	Portuguese	170	Brazil, Portugal
RU	Russian	155	Armenia, Azerbaijan, Belarus, Estonia, Georgia, Kazakhstan, Kyrgyzstan, Latvia, Lithuania, Moldova, Russia, Tajikistan, Turkmenistan, Ukraine, Uzbekistan
SV	Swedish	8	Sweden
ZH	Chinese	1100	China, Malaysia, Taiwan, Thailand, Singapore, Vietnam

Table 1.2. Codes for language names from ISO 639-1.

pora. Two techniques in particular will be explored: (1) alternative methods for monolingual tokenization including the use of character n-grams, and (2) corpus-based translation of character n-grams.

1.9.1 Alternative Tokenization Methods

Alternatives to exact matching of words have long been considered in IR. As 75% of the world's languages have concatenative morphology, the most commonly implemented approach to word normalization for retrieval is stemming. Stemming algorithms attempt to remove affixes and leave root forms as the indexing terms. In this research test sets for a large number of languages are used to compare different approaches to tokenization, including plain words, a popular stemming algorithm, and character n-grams.

1.9.2 Subword Translation

This research extends the idea of n-gram indexing for the purpose of bilingual retrieval by attempting to translate n-grams themselves, rather than words or stemmed words, using aligned corpora. Table 1.3 contains several examples showing the promise of this idea. Matches on substrings that are smaller than words can provide partial matches for multiword phrases or named entities for which no exact correspondence would be found in either a translation dictionary or even the given parallel text.

Subword translation has the potential to solve the problems identified by Pirkola *et al.* (2001), which were discussed earlier. First, because we propose to use n-grams that are typically smaller than individual words, some resilience to variations in morphology is built in. For example, whether a query is 'juggling', 'juggler', or 'juggled', the non-suffix n-grams should translate well. Second, even if a word is unknown (*i.e.*, is an out-of-vocabulary term), we can likely translate it piecemeal through n-grams that touch the morpheme and affixes. Shorter n-grams will enable partial translation of surnames and technical terminol-

Pair	Source	Target	Words
French to Dutch	lait	melk	lait melk
Italian to Spanish	latt	lech	latte leche
Italian to Spanish	atte	acte	
French to Dutch	olym	olym	olympique olympisch
French to Dutch	ique	isch	
Italian to Spanish	olim	olim	olimpico olimpico
Italian to Spanish	pico	pico	

Table 1.3. 4-gram translations for *milk* and *olympic*.

ogy due to matches of word fragments. Third, multiword phrases might be approximated with translation of word-spanning n-grams. For example, consider the central 5-grams of the English phrase prime minister (*i.e.*, *ime_m*, *me_mi*, and *e_min*). The derived translations of these English 5-grams into French are *er_mi*, *_mini*, and *er_mi*, respectively. This seems to work as expected for the French phrase “premier ministre”, although the method is not foolproof. Consider n-gram translations from the phrase “communist party” (*parti communiste*): ‘_commu’ (mmuna), ‘commu’ (munau), ‘ommun’ (munau), ‘mmuni’ (munau), ‘munis’ (munis), ‘unist’ (unist), ‘nist_’ (unist), ‘ist_p’ (ist.p), ‘st_pa’ (l_re_), ‘t_par’ (rtie_), ‘_part’ (_part), ‘party’ (rtie_), and ‘arty_’ (rtie_). The word-spanning n-grams in this particular multiword phrase do not seem to appropriately handle the French inversion of English adjective/noun ordering.

The proper translation of an n-gram is an elusive concept: there is typically no single, correct answer. To quantify the quality of an n-gram translation we adopt a functional view of meaning. We define the meaning of an indexing term broadly as the range of documents the term allows us to access. Given this definition, a good translation of an indexing term is a term in the target language that means the same thing, that is, one that provides access to target language documents that are similar to those accessible through the source language

term. Similarity can be defined in a variety of ways, such as describing the same concepts, equally relevant, or even direct translations if a parallel corpus is in use for evaluation. This view of meaning does not commit to words as indexing terms; it admits the use of stems, phrases, or any other type of indexing term. In this research the translation of an n-gram is a target language term that provides access to target language documents that are similar to the source language documents accessible through the source n-gram.

1.9.3 Summary

In this chapter I laid out the problem being investigated in this dissertation: how to improve monolingual and cross-language information retrieval performance using techniques including non-word tokenization and translation using aligned parallel corpora. Techniques based on character n-gram indexing are studied in Chapters 4 and 5. Bilingual retrieval experiments are presented in Chapter 6. In Chapter 2 key relevant literature is first reviewed.

Chapter 2

RELATED WORK

2.1 Character N-grams

In the previous chapter background information about character n-gram indexing was provided and several points about their use were made based on my long collaboration in this area with Dr. James Mayfield. The notion of translating n-grams using parallel corpora is as far as I know, completely original. The closest work of which I am aware is that of Toivonen *et al.* (2005) which rank candidate translations of words or phrases using ‘fuzzy matching’; matches between words are computed in part using transformations from source language character n-grams of length two or three to the target language.

2.2 Parallel Texts

Immediately upon the discovery of the Rosetta Stone in 1799 by Napoleonic forces it was realized that the parallel scripts could be used to decipher Egyptian hieroglyphics (Andrews 1981). Within 25 years of its discovery, the work of Sir Thomas Young and Jean-François Champollion accomplished just that. Two decades later Sir Henry Rawlinson unraveled the mystery of cuneiform writing based on his study of the Behistun Inscription (Adkins 2004). In modern times the growing availability of parallel electronic texts has fueled the development of statistical machine-translation systems. Obtaining, preprocessing,

aligning, and exploiting parallel texts has become an important task in human language technology. Since the mid-1990s two books devoted to parallel texts have been written. The manuscript by Melamed (2001) largely focuses on bitext alignment and its evaluation; however it also discusses applications such as identification of non-compositional compounds and attempts to improve translation through word sense identification. The book by Veronis (2000) is an excellent collection of papers describing the breadth of research at that time.

The application of parallel corpora to bilingual retrieval was initiated with work by Landauer and Littman (1990) using Canadian parliamentary proceedings, which are produced in English and French. They demonstrated their techniques, based on latent semantic indexing, through mate finding and by qualitatively examining candidate word translations. Numerous studies in cross-language retrieval have since been conducted using corpora as the principal method for translation (Nie, Simard, & Foster 2000; Franz *et al.* 2001; Xu, Fraser, & Weischedel 2001)¹.

Resnik *et al.* examined combination of translation lexicons and presented a backoff method that uses stemmed forms when surfaces forms do not match (Resnik, Oard, & Levow 2001).

Because of the difficulties in obtaining and aligning parallel corpora, there has been some investigation into just how much translation can be accomplished using documents that are comparable but not true translations (Franz, McCarley, & Roukos 1998). Such a corpus can be created by obtaining contemporaneous journalistic text, since many events will be reported in the news in independent newspapers in different languages. Our research departs from prior work in its novel translation of overlapping character n-grams, rather than words or stemmed words.

¹For other examples, consult the proceedings of the CLEF, NTCIR, and TREC workshops

2.3 Phrase-Enhanced Retrieval

Identification of good phrases has seemed important for information retrieval systems because intuitively, multiword units appear to clarify a concept more distinctly than the same phrase represented only as individual words. For example, *united states* and *machine gun* both mean something rather more specific than their individual words would indicate. Whether phrases can in fact improve IR performance has been a long debated issue; one that I believe has yet to be decided. Over fifteen years ago Salton and Buckley argued against the use of phrases, citing problems in finding a sufficiently large number of phrases and finding good phrases without introducing poor ones (1988). They stated:

In reviewing the extensive literature over the past 25 years in the area of retrieval system evaluation, the overwhelming evidence is that the judicious use of single-term identifiers is preferable to the incorporation of more complex entities extracted from the texts themselves or obtained from available vocabulary schedules.

It should be remembered that their opposition at that time was based on studies conducted on the only available IR collections of the time, which were uniformly small by present standards. Fagan investigated syntactic and statistical methods of augmenting indexes with phrases but saw only small improvements (1987). On the larger TREC collection Voorhees and Harman survey the effect of phrases among the best performing automatic ad hoc runs in their synopsis of the TREC-7 ad hoc track (1998). The relative improvement due to phrases was measured in four systems and reported to be “minimal, 3.6%, 2%, and 2%” on average query performance, but about 30% of queries did benefit from the technique.

There has been some recent work in monolingual retrieval that is promising. Metzler and Croft presented a model based on Markov Random Fields that allows for combining evidence from multiple features. They examined both ordered and unordered² word

²They define unordered bigrams as word pairs that occur within a window of k words.

bigrams, finding that “the combination of both ordered and unordered features led to noticeable improvements in mean average precision” (Metzler & Croft 2005). Relative gains on the order of 10% were observed for large web collections.

The studies mentioned above examined retrieval performance in an automated IR system. It could be the case that with a user explicitly indicating important phrases, such as by the web search engine convention of enclosing phrases in quotation marks, IR systems would find more substantial improvement due to phrasal processing. Vechtomova took up the study of nominal phrases for interactive query expansion using data from the TREC 2004 HARD task (2006). She saw small improvements at high precision levels (*e.g.*, precision at 5 documents), but the major contribution of the work is the analysis of types of phrases based on *stability*, and a model for term weighting motivated by these observations. Three classes were identified:

- phrases where the constituent terms only occur with each other (*e.g.*, *Burkina Faso*);
- phrases composed of words that frequently occur together, and where the phrasal structure is rigid (*e.g.*, *Mad Cow Disease*); and,
- combinations of terms where flexibility is permitted, such as substitution of words or reorderings (*e.g.*, *animal cruelty* or *cruelty to animals*).

2.4 Bilingual Retrieval and Phrases

Even though phrasal processing does not seem to improve monolingual retrieval, some have argued that it still may be important for cross-language information retrieval because of the problem of translation ambiguity, where query terms may have more than one meaning and more than one translation. Ballesteros and Croft investigated both phrasal translations and various query expansion techniques to mitigate losses incurred due to ambigu-

ity (1997). They found that phrasal translation did not improve performance over word-by-word translation; however they did find that query expansion techniques, notably pre-translation query expansion, could substantially reduce errors in dictionary-based translation. Pre-translation expansion involves performing an initial source language retrieval to identify documents pertinent to the query; from these documents additional terms are extracted, as in pseudo relevance feedback (Harman 1992), and used to augment the original query terms. A larger set of source language query terms is better able to cope with losses due to unknown or poorly chosen translations. In earlier work I was able to demonstrate that the relative efficacy of pre-translation query expansion depends on the quality (*i.e.*, the coverage) of a translation resource and that inferior resources benefit proportionally more than high quality translation resources (McNamee & Mayfield 2002).

Hull and Grefenstette (1996) reported on experiments in French to English retrieval where they compared word-for-word translation (both automated and manual) with translation that utilized manually produced phrase translations for the query set. Use of the human phrasal translations resulted in a roughly 40% relative improvement. While this dramatic result cannot be easily obtained using a fully automated system, it does motivate effective translation of multiword expressions. In their analysis the authors write:

Our experimental results demonstrate that recognizing and translating multiword expressions is crucial to success in MLIR. This is in distinct contrast to monolingual IR, where identifying noun phrases or word pairs generally helps but does not produce dramatic gains in average performance. The key difference is that the individual components of phrases often have very different meanings in translation, so the entire sense of the phrase is often lost. This is not always the case, but it happens often enough to make correct phrase translation the single most important factor in our multilingual experiments.

This is compelling support for an improved method for phrase translation. But to date little success has been reported using phrases for bilingual retrieval. I believe this is because most CLIR systems rely on general-purpose wordlists that have spotty coverage of

multiword expressions and because wordlists seldom have proper names and technical terminology, which are important in a large class of queries. A recent study that did show promise in translating OOV words and phrases used the Web as a huge corpus and looked for patterns particular to Chinese writing where an unusual name is often followed by a parenthetical translation in English (Zhang & Vines 2004); unfortunately, this phenomenon is not common with other language pairs. Other proposed methods for translating multiword expressions are based on part-of-speech (POS) tagging or parsing, which require tools to be developed in each language of interest.

Adriani and van Rijsbergen investigated the use of phrase identification to improve bilingual retrieval performance (2000). One nice thing about their study is that they explicitly compared an automated disambiguation technique to the use of phrasal translation and found a 10-12% relative advantage using phrasal translation. Against this improvement it should be noted that only 24 queries were studied on only the Associated Press (English) subset of the TREC dataset. It is possible that different results would be observed in a larger test set.

Unlike the study by Ballesteros and Croft, Adriani and van Rijsbergen did not attempt to translate phrases from user queries, but rather attempted to form likely target language phrases by examining the multiple translations of individual query words. Using transitivity, phrases longer than just two words could be constructed.

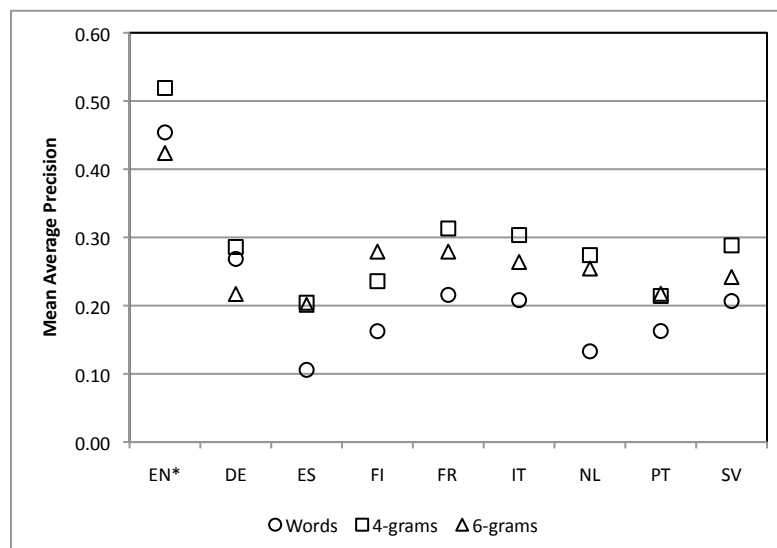
2.5 Alternatives to Translation

Translation resource availability varies widely across language pairs and for some pairs no resources may be available. An alternative to query translation in such cases is to leave the query untranslated. This approach seems peculiar at the outset, but reasonably good results can be obtained without translation when the source and target languages

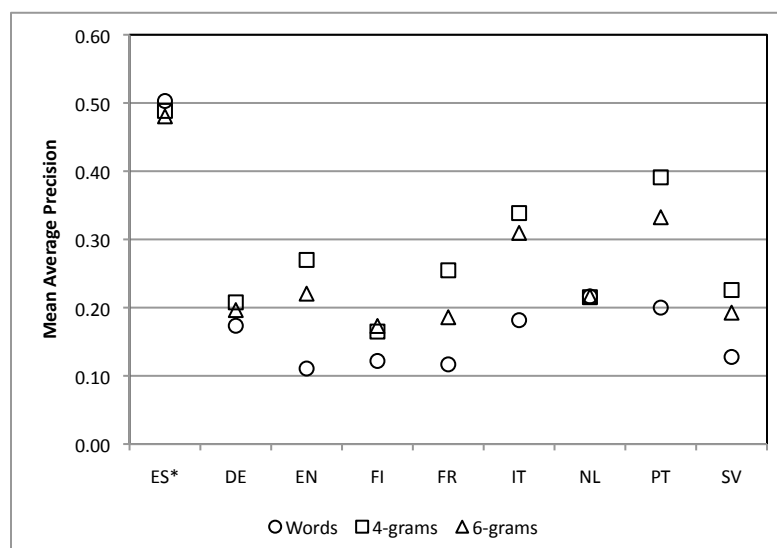
are similar. This is because of morphological cognates, words in two languages with the same meaning and orthographic representation (*e.g.*, *automobile* in English and French), or because of loan words. Scenarios exist in which the non-translation approach might be useful. For example, suppose we want to search a collection of Galician documents using a query expressed in English. We might be able to obtain reasonable accuracy by translating the English query into Portuguese, for which adequate resources are available, and relying only on cognate matches to transfer from Portuguese to Galician.

Buckley *et al.* (2000) examined retrieval of French from English queries by treating English as misspelled French. Natural English-French cognates were augmented by manually devised rules to spell-correct English words to French words of similar spelling. They estimated that 30% of non-stopwords could be transformed automatically in this fashion. Unpredictably, this approach was enormously successful relative to other cross-language approaches of the time, achieving the highest scoring automatic cross-language run at the TREC-6 evaluation (mean average precision of 24%, which was about 60% of monolingual). The authors pointed out that this technique is only useful for related languages. Buckley *et al.* made use of a translation lexicon, spelling correction, and cognates.

In earlier work with Mayfield (2004) I was able to show that n-gram tokenization greatly improves the accuracy of the no-translation approach, because the number of matching n-grams across related languages is significantly greater than the number of words that match exactly. Simply using n-gram matches instead of word cognates can double the efficacy of this approach. The use of pre-translation query expansion may result in even better performance but we have not explored this. Figure 2.1 shows the accuracy of the no-translation approach for a variety of language pairs. English (upper) and Spanish (lower) document collections were searched using queries prepared in other languages by human translators. For comparison monolingual conditions are shown at the left edge of the charts. With character 4-grams untranslated Portuguese queries achieve performance 70%



(a) English documents



(b) Spanish documents

FIG. 2.1. Comparative efficacy of words and n-grams for bilingual retrieval making no attempt at translation.

as good as monolingual Spanish queries.

A hybrid approach that leverages both spelling correction and n-gram matching was taken by Toivonen *et al.* who used small n-grams ($n = 2$ or $n = 3$) in concert with statistically derived rules for mapping orthography to achieve translation of out-of-vocabulary words between closely related languages (2005). This work can be distinguished from that study in several ways. First, our method of subword translation should be effective even in languages without a common alphabet. Second, Toivonen *et al.* used a bilingual dictionary, while our approach requires a parallel corpus. Third, we have already reported results validating the efficacy of our translations on bilingual test sets (McNamee & Mayfield 2005) while they examined translation accuracy alone.

Another approach for tackling out-of-vocabulary words was investigated by Knight and Graehl (1998). They converted English words into Japanese phonetic equivalents (katakana) using a generative model. The technique worked well for technical word and borrowed words such as geographic names.

Having reviewed the relevant literature on the translation in CLIR, Chapter 3 will describe the experimental methods and test sets that will be used in the experimental work that follows.

Chapter 3

METHODOLOGY

The goal of this research is to investigate new techniques using non-word representations to improve multilingual retrieval, and in concert with parallel corpora, to address the major problems in dictionary-based bilingual retrieval, namely coping with morphological variation, out-of-vocabulary words, and multiword phrases. To test various hypotheses it will be necessary to identify and develop certain resources such as IR test sets, parallel texts suitable for inducing translations in the language pairs of interest, and a retrieval engine with support for translation and n-gram tokenization. Specific algorithms will have to be tested for statistical significance and appropriate protocols will have to be developed so that success, or failure, will be measurable.

3.1 IR Test Sets

To verify and quantify the efficacy of the proposed techniques it is necessary to leverage existing cross-language test collections where possible. Creating new test collections is challenging and labor-intensive because of the need for human judgments. Several large-scale cross-language evaluations that are potentially appropriate for evaluating this research are summarized in Table 3.1. These test collections are mainly newswire. Though there are many potentially useful collections in Table 3.1, I plan to concentrate on the CLEF

Name	Queries	Query Languages	Target Languages
TREC 9	50	EN ZH	ZH
TREC-10,11	75	AR EN FR	AR
CLEF 2000-2007	400	AM BG CS DE EL EN ES FI FR ID IT JP NL PT RU SV TH ZH	BG CS DE EN ES FI FR HU IT NL PT RU SV
NTCIR-2,3	100	EN JP KR ZH	EN JP KR ZH

Table 3.1. Cross-language test collections.

collections.

3.1.1 Cross Language Evaluation Forum

Starting in 2000 the Cross-Language Evaluation Forum (CLEF)¹ has conducted annual evaluations for multilingual retrieval in European languages. Each year ad hoc test collections were developed and typically about fifty topics were created per year. The queries were chosen to balance regional, national, and international topics.

The document collections are somewhat contemporaneous, coming from the years 1994, 1995, or 2002. Over time new sources were occasionally added to the collections in several of the languages, thus the relevance judgments for a given language and year might not include documents that were added in subsequent years. The test data covered only four languages in 2000 but as of 2007 test sets existed for thirteen languages. The present experiments are notable for using such a large number of queries in a panoply of languages.

Not every language was studied during each year. In English and French, which received the most study, there are over 300 available queries, but Czech and Russian only have around 50. The remaining languages have around 150 queries. Information about the number of available queries in the test collections is given in Table 3.2. Queries for which

¹<http://www.clef-campaign.org/>

		2000	2001	2002	2003	2004	2005	2006	2007	Total
BG	Bulgarian						49	50	50	149
CS	Czech								50	50
DE	German	37	49	50	56					192
EN	English	33	47	42	54	42	50	49	50	367
ES	Spanish		49	50	57					156
FI	Finnish			30	45	45				120
FR	French	34	49	50	52	49	50	49		333
HU	Hungarian						50	48	50	148
IT	Italian	34	47	49	51					181
NL	Dutch		50	50	56					156
PT	Portuguese					46	50	50		146
RU	Russian				28	34				62
SV	Swedish			49	53					102
# Languages		4	6	8	9	5	5	5	4	
Maximum		40	50	50	60	50	50	50	50	400

Table 3.2. Number of judged queries in CLEF test sets (2000-2007).

no relevant documents were identified in the pools did not contribute to these counts.

The topic sets for two languages contain a large number of topics that have a small number of known relevant documents. In Finnish about 40% of the topics had fewer than 5 relevant documents. For the Russian collection, which has the smallest document collection, the percentage is 73%. This might suggest that comparisons between experimental conditions on the Russian test set might require special evaluation measures or testing for significance. The number of topics with few relevant documents is given in Table 3.3.

3.1.2 Experiment Design

Queries consist of three fields named *title*, *description*, and *narrative*. The title field is a set of two or three keywords that succinctly express the query. The description is a grammatical sentence that often repeats the keywords in the title. The narrative usually consists of a few sentences that amplify or clarify the information that is being sought.

	#Docs	All topics	≤ 4	≤ 3	1 or 2	1
BG	69k	149	23	16	10	3
CS	82k	50	5	3	2	0
DE	295k	192	23	17	13	8
EN	170k	367	86	63	50	34
ES	453k	56	11	9	6	3
FI	55k	120	47	42	30	16
FR	178k	333	64	56	34	18
HU	50k	148	17	7	5	2
IT	157k	181	41	29	15	6
NL	190k	156	22	14	11	1
PT	107k	146	25	21	17	8
RU	17k	62	45	39	28	17
SV	143k	102	18	13	12	17

Table 3.3. Number of topics with few relevant documents.

Title: Northern Japan Earthquake
Description: Find documents that report on an earthquake on the east coast of Hokkaido, northern Japan, in 1994.
Narrative: Documents describing an earthquake with a magnitude of 7.9 that shook Hokkaido and other northern Japanese regions in October 1994 are relevant. Also of interest are tidal wave warnings issued for Pacific coastal areas of Hokkaido at the time of the earthquake. Documents reporting any other earthquakes in Japan are not relevant.

FIG. 3.1. English Query 62 from CLEF 2001.

A sample query is shown in Figure 3.1. The normal mode at CLEF has been to require participating systems to submit a run based on the title and description fields alone. Such runs are often described with the abbreviation TD. The experiments in Chapters 4-6 are based on TD runs.

In the experiments which follow performance is measured using the number of queries appropriate to each language given in Table 3.2. While this decision will generally prevent direct comparison with previously published experiments from a single year, this choice has the advantage of providing the largest number of samples and therefore give the greatest

Name	Size (MB)	Languages	Genre	Source
Hong Kong news/laws	50	EN ZH	Legal, News	LDC 2000T46 and LDC 2000T47
Europarl v3	197	DA DE EL EN ES FI FR IT NL PT SV	Parliamentary oration	Philip Koehn
JRCAcquis v3	202	BG CS DA DE EL ES ET FI FR HU IT LT LV MT NL PL PT RO RU SK SL SV	EU Laws	Joint Research Centre
UN	250	EN ES FR	UN conferences	LDC 94T4A
Canadian Hansard	400	EN FR	Legislative discourse	LDC 95T20
OJ EU	590	DA DE EL EN ES FI FR IT NL PT SV	Governmental affairs (written)	Developed during this research from texts at http://europea.eu.int/

Table 3.4. Large parallel collections.

statistical power when comparing different methods of tokenization. In Appendix A results using the methods in this research are compared with results from the CLEF 2002 and 2005 workshops.

3.2 Parallel Corpora

An equally important consideration in guiding this research is the availability of parallel texts. I am aware of the resources listed in Table 3.4, which can be used. Consistent with the intention to focus on the CLEF test sets and their coverage of many languages I plan to focus on the Europarl (Koehn 2003), JRC-Acquis, and Official Journal of the European Union corpora. Additional detail about the parallel texts used is given in Chapter 6 “Bilingual Experiments.”

3.3 Retrieval Engine

To conduct experiments a search engine that is capable of working with large multilingual test sets is needed. Working with a state of the art system is preferable to give greater credence to the empirical results. The HAIRCUT system (McNamee & Mayfield 2003) has been used in multiple international evaluations and achieved consistently high results.

3.3.1 HAIRCUT

The Hopkins Automated Information Retriever for Combing Unstructured Text (HAIRCUT) information retrieval system was developed at the Johns Hopkins University Applied Physics Laboratory. The software is written in Java and it supports modern IR techniques, including the language modeling retrieval framework (Hiemstra 2001; Ponte & Croft 1998), which has become increasingly popular in recent years. HAIRCUT enables n-gram tokenization, automated relevance feedback, and both dictionary and corpus-based translation, which are essential techniques for the experiments in this dissertation. The system has only limited support for phrasal processing as it does not store within-document positional information for terms, nor does it implement nextword indexing as proposed by Bahle et al. (2002).

3.3.2 Language Model Framework

In the language model approach to retrieval documents are ranked for their relevance to queries based on a generative model. Specifically the probability that is being estimated is the maximum likelihood estimate that a relevant document, D , could be generated from a unigram language model based on the query, Q , is $P(D|Q)$. Because queries tend to be much shorter than documents it is very difficult to estimate this probability directly,

therefore Bayesian inversion is applied:

$$(3.1) \quad P(D|Q) = \frac{P(Q|D)P(D)}{P(Q)}$$

If we make the assumption that *a priori* all documents are equally likely to be relevant regardless of Q we obtain:

$$(3.2) \quad P(D|Q) = \frac{P(Q|D)}{P(Q)}$$

Now for the purpose of ranking of documents in decreasing likelihood of relevance we can omit the prior probability of the query, Q , leaving:

$$(3.3) \quad P(D|Q) \propto P(Q|D)$$

Making the naïve Bayes assumption (*i.e.*, that terms are independent):

$$(3.4) \quad P(D|Q) \propto \prod_{t \in Q} P(t|D)$$

Relative document term frequency is a reasonable estimate for $P(t|D)$. However, in this form the model only gives non-zero scores to documents that contain all of the query terms t_1, \dots, t_n . This corresponds to a strict Boolean model with AND semantics. If a document contains a synonym of a query word instead of one of the exact words then $P(D|Q)$ would be zero because one of the terms is missing and a $P(t|D)$ term is zero. To enable more permissive matching smoothing can be applied where document term frequencies are mediated by a generic model of language from a corpus, C . When all terms are present then the highest scores will result; this is roughly analogous to extended Boolean or coordinate-level ranking. Using Jelinek-Mercer smoothing (1980), or linear interpolation, a parameter

λ can be introduced to reflect the importance of the term's prescence in the document, Equation 3.4 can be written:

$$(3.5) \quad P(D|Q) \propto \prod_{t \in Q} \lambda P(t|D) + (1 - \lambda)P(t|C)$$

It remains to estimate $p(t|C)$. In HAIRCUT the mean relative document term frequency is used for each term t .

Several implementation details remain. One issue is dealing with words that are duplicated in a query. My approach is to treat each instance independently and simply multiply the probability twice. Another issue is how to select a value for λ . A single value could be used for all terms – this is what is done in HAIRCUT. It turns out that performance is fairly insensitive to the value selected as long as values near 0 or 1 are avoided (Zhai & Lafferty 2004). Though performance could be improved slightly by optimizing choice of λ as a function of tokenization, a smoothing constant of 0.5 was used in these experiments. Finally, because the multiplied probabilities become very small, the calculation is performed in log-space to avoid floating-point underflow.

3.4 Metrics and Statistical Testing

Quantitative evaluation using mean average precision (MAP) is used to compare different retrieval algorithms. MAP is the standard evaluation metric in information retrieval research; it combines precision and recall into a single value and it is stable and sensitive to small changes in ranking relevant documents.

Uninterpolated average precision can be computed by summing precision values after each relevant document is found² and dividing by the total number of relevant documents.

²Relevant documents that are not retrieved are considered to have a precision of zero

Mean average precision is simply the arithmetic mean of average precision over a set of queries. Queries with no known relevant documents do not affect the calculations.

For the CLEF test sets, the pooling process is believed to have created judgments capable of supporting post-hoc retrieval experiments (Hiemstra & van Leeuwen 2002). Given the relatively complete judgments MAP is a suitable measure, however alternative metrics have been developed when this assumption is less valid (Yilmaz & Aslam 2006). To simply analysis across multiple languages, the arithmetic mean of MAP from several languages can be calculated; I occasionally refer to this quantity as panlingual mean average precision, or PMAP.

Software to evaluate performance is needed. I use the `trec_eval` package³ which computes mean average precision and other IR measures of performance. Additionally, software developed at the University of Massachusetts to perform statistical analyses on IR experiments was made available by Prof. James Allan. This software computes p-values using the sign, Student's *t*-, and Wilcoxon rank tests. Cormack and Lynam (2007) compare the use of the Student's *t*-test and the Wilcoxon test for information retrieval experiments using data from the TREC 2004 Robust track. Wilcoxon had more discriminating power, but larger error in its significance values. They recommend the paired *t*-test.

In the experiments presented in the following chapters testing for statistical significance was performed using the paired *t*-test.

3.5 Other Software

To preprocess parallel texts a number of Unix and Perl scripts were developed and alignment software developed by Church (1993) was used.

³Available at http://trec.nist.gov/trec_eval/

Chapter 4

MONOLINGUAL EFFECTIVENESS OF N-GRAMS

If character n-grams are not effective for monolingual retrieval, then there is much less reason to think they will be a good choice for improving bilingual processing, which is the principal goal of this research. Therefore in this chapter the use of character n-gram tokenization in thirteen European languages is examined. N-grams are compared to traditional words, rule-based stemming, and several additional methods of tokenization. The translation issues that arise in bilingual retrieval are deferred until Chapter 6.

4.1 Overview

As described in Chapter 1 the phrase “n-gram tokenization” refers to overlapping sequences of n characters. For example, the phrase ‘real estate’ can be represented using 5-grams by the set: *_real*, *real_*, *eal_e*, *al_es*, *l_est*, *_esta*, *estat*, *state*, and *tate_*. This example illustrates several features, namely, the leading and trailing spaces that generate the n-grams *_real* and *tate_*, the n-grams that span word boundaries such as *al_es*, and the conflation inherent in this form of tokenization since the n-gram *state* could have been generated from words like apostate, interstate, prostatectomy, state, stateswoman, unstatesmanlike, and hundreds of other terms. Although such conflation is rampant with n-grams (the n-gram *real_* could have come from the word surreal) I suggest that the redundancy provided

5-gram	DF	6-gram	DF
_real	45166	_real_	19928
real_	21440	real_e	4064
eal_e	4522	eal_es	3333
al_es	4389	al_est	4064
l_est	4770	l_esta	4322
_esta	17009	_estat	5953
estat	7034	estate	6103
state	56533	state_	38685
tate_	40277		

Table 4.1. Document frequencies for 5- and 6-grams in *real estate*.

by the technique counteracts the negative effect of less discriminating terms. Thus while many documents may have the 5-gram *state*, few will have *_real*, *al_es*, and *state* unless the phrase *real estate* is present.

To illustrate this point compare the document frequencies of the component 5-grams for *real estate* in Table 4.1. Statistics were obtained from the CLEF 2004 English document collection. The words *real* and *estate* occurred together in 3562 documents; however the number of documents with the phrase is less, approximately 3300. This is indicated by the fact that the 6-gram *eal_es* was present in only 3333 documents. The 6-grams generated 30 or so false matches for phrases like: *Dr. Neal estimates*, *ideal Estee Lauder*, *your ideal escape*, and *a deal estimated at*. The word-spanning 5-grams are a little less predictive of the phrase compared to their corresponding 6-grams.

Alternatives for generating n-grams other than with overlapping, fixed size windows that span word boundaries include using only word-internal n-grams, using multiple, or variable length character spans, and removing the restriction that characters must be adjacent (Järvelin, Järvelin, & Järvelin 2007). Preliminary work has indicated that there is not a large difference in the effectiveness of word-spanning versus word-internal n-grams (McNamee & Mayfield 2004); however by spanning adjacent words there is potential to

capture phrase-like information and to decrease conflation, thus there is some justification for retaining them. In the extant literature fixed length n-grams appear to be the dominant approach.

In the remainder of this chapter I discuss experiments that directly compare the use of n-gram tokenization to other approaches. In Section 4.2 traditional words are compared to various lengths of n-grams and in Section 4.3 stemmed word forms are examined. N-grams are a language-neutral technique and in Section 4.4 I compare them to another language-independent method of normalization. I also describe a synthetic replacement for stemming that was inspired by n-gram tokenization (Section 4.5). Finally, this chapter's findings are summarized in Section 4.6.

The data sets used in these experiments were described in Section 3.1 in the previous chapter.

4.2 Comparison with Words

The use of ordinary words (*i.e.*, space delimited tokens) as indexing terms has several advantages: in alphabetic languages they are easy to identify, they do not present computational difficulties (*i.e.*, dictionaries of 1 million terms are trivially represented on modern hardware), and most significantly, they are transparent to the end user (*i.e.*, with Boolean semantics only documents containing the exact query terms are returned). In contrast, with n-gram indexing it is possible, albeit unlikely, that documents are returned which contain none of the input words. Words are not a perfect solution. Indexing based on unnormalized word forms does not enable matching morphologically related words, and this is not the only problem.

In some languages lengthy compound words can be formed which inhibit exact matching. Decompounding algorithms can be utilized, but these must balance desirable splitting

with over splitting. For example, splitting the German *kindergarten* into *kinder* and *garten* is probably too far; once separated there is a loss of meaning. In languages such as Chinese and Japanese where spaces are not used to indicate word boundaries an errorful segmentation process is often performed to identify words. N-grams are a conceptually simpler approach.

Data comparing words and n-grams is presented in Table 4.2. Statistical testing with the paired t -test is performed by examining the average precision scores for individual queries. The number of trials varies from 50 to 333 depending on the number of available topics in the specific language being examined. Significant improvements or degradations with $p < 0.05$ are indicated with Δ or ∇ , respectively. Likewise when $p < 0.01$ significant gains are indicated with \blacktriangle and losses are indicated with \blacktriangledown .

When interpreting these results comparisons using mean average precision should not be made between two different languages. While the topics are drawn from a common set of 400 topics, each language uses a subset pertaining to those years for which relevance judgments were created. Even for a single topic scores are not commensurable across different languages because the topics are translations (*i.e.*, they are not identical) and of even more significance, the document collections are different and thus for any given topic it may be easier or harder to find a relevant document in one language compared to another due to the different sizes of the collections or their differing coverage of a subject.

4.2.1 Observations of N-gram Length

The data in Table 4.2 show that 4-grams and 5-grams both significantly outperform plain words as indexing terms. 5-grams have higher mean average precision in all 13 languages and statistically significant improvements over words were observed in 12 of the 13 cases ($p < 0.05$). Averaged across all languages 5-grams obtain a substantial 21%

Lang	N	Words	3-grams	4-grams	5-grams	6-grams	7-grams
BG	149	0.2164	0.2271 (+4.9%)	0.3105 [▲] (+43.5%)	0.2820 [▲] (+30.3%)	0.2528 [▲] (+16.8%)	0.2161 (-0.1%)
CS	50	0.2270	0.2792 [△] (+23.0%)	0.3294 [▲] (+45.1%)	0.3223 [▲] (+42.0%)	0.2918 [▲] (+28.6%)	0.2536 (+11.7%)
DE	192	0.3303	0.3188 (-3.5%)	0.4098 [▲] (+24.1%)	0.4201 [▲] (+27.2%)	0.3961 [▲] (+19.9%)	0.3632 [△] (+10.0%)
EN	367	0.4060	0.2588 [▼] (-36.3%)	0.3990 (-1.7%)	0.4152 (+2.3%)	0.3903 (-3.9%)	0.3556 [▼] (-12.4%)
ES	156	0.4396	0.3010 [▼] (-31.5%)	0.4597 (+4.6%)	0.4609 [△] (+4.9%)	0.4252 (-3.3%)	0.3621 [▼] (-17.6%)
FI	120	0.3406	0.3591 (+5.4%)	0.4989 [▲] (+46.5%)	0.5078 [▲] (+49.1%)	0.4692 [▲] (+37.8%)	0.4323 [▲] (+26.9%)
FR	333	0.3638	0.2544 [▼] (-30.1%)	0.3844 [△] (+5.7%)	0.3930 [▲] (+8.0%)	0.3660 (+0.6%)	0.3201 [▼] (-12.0%)
HU	148	0.1976	0.2778 [▲] (+40.6%)	0.3746 [▲] (+89.6%)	0.3624 [▲] (+83.4%)	0.3335 [▲] (+68.8%)	0.3030 [▲] (+53.3%)
IT	181	0.3749	0.2177 [▼] (-41.9%)	0.3738 (-0.3%)	0.3997 [△] (+6.6%)	0.3669 (-2.1%)	0.3217 [▼] (-14.2%)
NL	156	0.3813	0.3326 [▼] (-12.8%)	0.4219 [▲] (+10.7%)	0.4243 [▲] (+11.3%)	0.3960 (+3.9%)	0.3663 (-3.9%)
PT	146	0.3162	0.2213 [▼] (-30.0%)	0.3358 (+6.2%)	0.3524 [▲] (+11.5%)	0.3223 (+1.9%)	0.2834 [▼] (-10.4%)
RU	62	0.2671	0.3252 (+21.8%)	0.3406 [▲] (+27.5%)	0.3330 [△] (+24.7%)	0.3181 (+19.1%)	0.3028 (+13.4%)
SV	102	0.3387	0.3244 (-4.2%)	0.4236 [▲] (+25.1%)	0.4271 [▲] (+26.1%)	0.4004 [▲] (+18.2%)	0.3713 (+9.6%)
PMAP		0.3230	0.2844 (-12.0%)	0.3894 (+20.5%)	0.3924 (+21.5%)	0.3637 (+12.6%)	0.3270 (+1.2%)

Table 4.2. Comparing words and n-grams of various lengths.

relative improvement in retrieval effectiveness¹.

4-grams and 5-grams were statistically indistinguishable in eight languages. 4-grams attain a higher score than 5-grams in Bulgarian[▲], Czech, Hungarian^Δ, and Russian, but are statistically worse in English[▼], Italian[▼], and Portuguese[▼]. Tokenization with 4-grams instead of words led to significant improvements in 9 of the 13 cases. On average the 4-grams perform just marginally below 5-grams, experiencing a 20% relative gain over words for the 13 languages compared to 21% for 5-grams.

If it were possible, setting n to some intermediate value between four and five (*e.g.*, 4.75) might prove to be the most effective option. This leads one to consider the question of whether some orthographic manipulation could lead to better performance than the use of a single n -gram length on ordinary text. For example, substitutions like $ck \rightarrow k$, $ss \rightarrow s$, or $ee \rightarrow e$ might be interesting to explore as a means of altering morpheme length in the hopes of a favorable impact on retrieval accuracy. Indexing with multiple n -gram lengths has been attempted (McNamee & Mayfield 2004), but it did not lead to gains.

As n -gram length decreases from $n = 4$ to $n = 3$ performance dramatically worsens. On average 3-grams exhibit a 12% decrease in mean average precision relative to words. And compared to words performance was statistically worse in seven languages, although statistically significant gains over words were seen in two languages, Czech and Hungarian.

The longest n -grams considered, 6-grams and 7-grams, are more effective than words; however, the relative improvements are noticeably less than those observed with 4-grams and 5-grams.

In all languages except Russian, the following trends were observed to be strongly significant ($p < 0.01$): (1) 3-grams performed worse than 4-grams; (2) 5-grams were better

¹ Average Precision (AP) measures performance for a single topic. Mean Average Precision is computed by micro-averaging AP across a set of topics. At the bottom of Table 4.2 the averages labelled PMAP are macro-averages of MAP across different conditions, in this case multiple document languages.

than 6-grams; and, (3) 6-grams were better than 7-grams.

4.2.2 Observations of Language Variability

The results on the Hungarian collection are of particular note: n-grams are a better choice than words for $3 \leq n \leq 7$. Amazingly 4- and 5-grams were 80% more effective than words.

N-grams were able to provide at least a 25% relative improvement in Bugarian, Czech, German, Finnish, Hungarian, Russian, and Swedish. Notably absent from this list are any of the Romance languages. In fact, n-grams (*e.g.*, 5-grams) have the least advantage in English and in French, Italian, and Spanish.

Linguistic typology appears to affect the success of n-gram tokenization. One hypothesis that would account for this is that n-gram effectiveness is tied to morphological complexity. Though such methods are not without controversy among linguists, there have been studies that attempted to quantify morphological complexity using principles from information theory (Juola 1998; Kettunen *et al.* 2006).

Juola examined translations of Biblical texts and erased morphology from each in the following way. Each word (or type) in a text is replaced with a unique symbol, a randomly selected integer. After this has been done to the entire text the words that normally exhibit morphological regularity, such as *jump*, *jumped*, *jumping*, no longer bear an obvious relationship to one another any more than do the numbers 18, 5429, and 1641. Juola then compared languages based on the ratio of the compressibility of the original text to the compressibility of the morphologically degraded text; the program *gzip* was used as a way of approximating the Kolmogorov complexity of the texts. Kettunen *et al.* followed the approach described by Juola and performed a similar analysis using translations of the European Union Constitution in 21 languages and the program *bzip2*.

In Table 4.3 data is presented that show for each language: (1) the mean word length,

	Word Length	Juola Ratio	Kettunen Ratio	5-gram Gain
BG	5.02	0.9717		30.31%
CS	5.38		1.0867	41.98%
DE	5.98		1.1660	27.19%
EN	4.68		1.0529	2.27%
ES	4.89		1.0624	4.85%
FI	7.23	1.1253	1.1637	49.09%
FR	4.79	1.0117	1.0622	8.03%
HU	5.99	0.9949	1.1421	83.40%
IT	5.08		1.0518	6.62%
NL	5.17		1.1189	11.28%
PT	4.89		1.0676	11.45%
RU	5.93			24.67%
SV	5.26	1.0456	1.1252	26.10%
ρ	0.7771	0.9054	0.6761	

Table 4.3. 5-gram effectiveness and linguistic complexity.

by token, for the CLEF corpora; (2) the ratio that Juola computed to indicate morphological complexity (larger indicates greater complexity), if available; (3) Kettunen *et al.*'s corresponding ratio for the language, if available²; and (4) the relative improvement observed with 5-gram tokenization. The three estimates of morphological complexity can be used to rank languages by inferred complexity. Similarly the relative gains attained using 5-grams instead of words can also be used to order languages from those that gain much (*e.g.*, Hungarian and Finnish) down to those that gain little (*e.g.*, English and Spanish). The table also gives Spearman rank correlation coefficients, which show moderate to large correlations between each of the three estimates of morphological complexity and the gains attainable with 5-grams.

²Kettunen *et al.*'s ratios tend to be slightly higher, but for languages in common the agreement in rankings is good.

4.2.3 Morphological Processes

Linguistic phenomena such as polysemy, where a word can have multiple meanings, and synonymy, where the same concept can be expressed with different word choices, complicate information retrieval. Additionally, failure to normalize morphologically related words (*e.g.*, *swimmer*, *swam*, *swimming*), can prevent matches in full-text retrieval. Though they are sometimes difficult to separate from one another, three broad classes of morphological processes result in surface forms that impair effective retrieval: *inflection*, *derivation*, and *word formation*.

Inflectional morphemes add information to root morphemes such as number (*e.g.*, *dog/dog+s*; *fox/fox+es*) and gender (*e.g.*, *act+or/act+ress*, though English does not often inflect for gender). Other functions such as negation (*e.g.*, *un+happy*) and comparison (*e.g.*, *fast/fast+er/fast+est*) can be indicated with inflectional (or grammatical) morphemes, though sometimes these are expressed through function words (*e.g.*, *not happy*). The process of adding inflectional morphemes by attaching them to root morphemes is called *agglutination*. Some languages separate each morpheme into separate words (*e.g.*, Chinese and Vietnamese), and these languages are termed *isolating*. However, affixation, the use of prefixes and suffixes to attach morphemes is extremely common. Languages that do this extensively are termed *agglutinative*. Languages vary in the degree of inflection and lie somewhere on the spectrum from isolating to strongly agglutinative. English nouns only have two cases (singular and plural), but in Finnish, a highly agglutinative language, nouns can have fifteen different cases.

Derivational morphology transforms words from one syntactic class into another. For example *compute* (verb) can produce *computer* (noun), or *boy* (noun) can become an adjective through addition of the suffix *-ish*.

There are variety of other methods for producing new words in a language, including:

- *foreign borrowing*: *ombrelli* (Italian) becomes *umbrella* (English); *quiche* and *trompe l'oeil* are borrowed from French.
- *acronyms*: USA, NASA, IRS, and IBM are all derived from the initial letters in their corresponding phrase.
- *clipping*: compression of *professor* to *prof*, or *gymnasium* to *gym*.
- *blending*: fusion of component words into a shortened single form, such as *brunch* from *breakfast* + *lunch*.
- *compounding*: concatenation of two or more words to form a new word (e.g., *pick-pocket*, *airport*, *airplane*, *girlfriend*, *mother-in-law*, *red-hot*, *underachieve*). Like agglutination, compounding is more productive in some languages than others, and noun-noun compounding is a feature of Germanic languages.

Table 4.4 characterizes the CLEF languages based on significant morphological processes that can affect retrieval.

4.3 Comparison with Stemming

Stemming is an intentional conflationary technique designed to group together morphological variants of a word. This is frequently desirable because if a query contains a gerund such as *kayaking*, it is likely that documents that have the words *kayak* or *kayaker* are also of interest. Even a relatively benign tokenization procedure like case-folding can create errors during retrieval, and unfortunately all stemming algorithms will produce errors. For example, the popular Porter stemmer (1980) conflates *generic*, *generous*, and *generation*, words that do not belong in the same equivalence class.

On early English collections (e.g., Cranfield, Medlars, and CACM) Harman found little advantage in stemming (Harman 1991); however Krovetz found measurable differences

	Language family	Highly Inflective	Compounding
Bulgarian	Slavic	Yes	
Czech	Slavic	Yes	
Dutch	Germanic		Yes
English	Germanic		
Finnish	Finno-Urgic	Yes	Yes
French	Romance		
German	Germanic		Yes
Hungarian	Finno-Urgic	Yes	
Italian	Romance		
Portuguese	Romance		
Russian	Slavic	Yes	Yes
Spanish	Romance		
Swedish	Germanic		Yes

Table 4.4. Morphological properties of CLEF languages.

using CACM, NPL, TIME, and WEST (Krovetz 1993) which he ascribed to addressing derivational morphology. In English, where the number of inflectional forms is low, the observed differences are not particularly large. Hull reported average absolute improvements of 1% to 3% (Hull 1996).

To assemble stemmers for each of the CLEF languages would require a lot of effort. Fortunately Porter developed a compiler named *Snowball*³ which can take a specification for a stemming algorithm and render a C or Java implementation of the stemmer. Snowball stemmers have been developed for a number of the CLEF languages, though there is no support for Bulgarian, and Czech at present. Due to trouble running Snowball in Hungarian, Portuguese and Russian, attention was focused on the remaining eight CLEF languages. Table 4.5 lists the number of transformation rules for the implementation in each language.

The data in Table 4.6 compares the relative effectiveness of words, stemmed words, and n-grams of lengths four and five. Stemming is clearly advantageous and outperforms

³Available from <http://snowball.tartarus.org/>

	DE	EN	ES	FI	FR	IT	NL	SV
# rules	27	105	197	88	141	195	32	49

Table 4.5. Number of rewrite rules for Snowball implementations.

Lang	Snowball	Words	4-grams	5-grams
DE	0.3695	0.3303▼(-10.6%)	0.4098▲(+10.9%)	0.4201▲(+13.7%)
EN	0.4373	0.4060▼(-7.2%)	0.3990▼(-8.8%)	0.4152▼(-5.0%)
ES	0.4846	0.4396▼(-9.3%)	0.4597▼(-5.1%)	0.4609▼(-4.9%)
FI	0.4296	0.3406▼(-20.7%)	0.4989▲(+16.3%)	0.5078▲(+18.2%)
FR	0.4019	0.3638▼(-9.5%)	0.3844▼(-4.4%)	0.3930 (-2.2%)
IT	0.4178	0.3749▼(-10.3%)	0.3738▼(-10.5%)	0.3997 (-4.3%)
NL	0.4003	0.3813▼(-4.8%)	0.4219 (+5.4%)	0.4243▲(+6.0%)
SV	0.3756	0.3387▼(-9.8%)	0.4236▲(+12.8%)	0.4271▲(+13.7%)
Average	0.4146	0.3719 (-10.3%)	0.4214 (+1.6%)	0.4310 (+4.0%)

Table 4.6. Comparing words, 4-grams, and 5-grams to Snowball stems.

plain words in all eight languages. On average words are worse by -10%, although the loss ranges in each language from -5% (Dutch) to -21% (Finnish). On average n-grams achieve higher mean average precision than stems; 4-grams gain 1% and 5-grams gain 4%. However, the effectiveness varies appreciably by language in a similar way to what was previously observed with raw words. Both lengths of n-grams had lower scores than stemmed words in English, French, Italian, and Spanish, but higher scores in Dutch, German, Finnish, and Swedish, which are compounding languages. Unfortunately, comparison between techniques on the higher complexity languages (*e.g.*, Czech and Hungarian), where n-grams yielded the largest gains against words, was not possible.

4.4 Comparison with Unsupervised Morphological Segmentation

There has been substantial recent work in unsupervised morphological analysis. Since rule-based stemmers are not available for all languages, unsupervised segmentation may be valuable. In fact, it is possible that unsupervised segmentation might outperform widely

	English 2005	Finnish 2004	German 2003
Top system	0.3943	0.4915	0.4729
Morfessor	0.3882	0.4412	0.4571
No analysis	0.3123	0.3274	0.3228

Table 4.7. Selected results (MAP) from Morpho Challenge 2007.

used stemmers such as Porter’s Snowball algorithm.

4.4.1 Morphology Challenge

In 2005 an evaluation was held for unsupervised segmentation of words into morphemes, for example, splitting a word like *seabirds* into *sea+bird+s*. Only segmentations of the original word were allowed and letters could not be substituted or dropped, even when it would make sense to identify a morphological root. Thus *flies* would not be segmented into *fly+es*.

A second evaluation was held in 2007 with two aims. The first was to produce not merely a segmentation, but an analysis of words, for example, determining that *cats* has the root *cat* and the *+plural* attribute. The second goal was to extrinsically evaluate the use of unsupervised morphological analysis for information retrieval. The information retrieval task used one year of CLEF data in English, Finnish, and German. Using the Lemur toolkit with Okapi term weighting nearly all of the analyses produced by competing systems, which could contain both segmentations and attributes, significantly outperformed the baseline condition where words were left unaltered (Kurimo, Creutz, & Turunen 2007). (See Table 4.7.) The fact that unsupervised language-neutral methods lead to large improvements in a few languages motivates further study using a much larger set of CLEF benchmarks.

One of the leading approaches was the Morfessor algorithm which is described next.

4.4.2 Morfessor

Morfessor⁴ is designed to accomodate languages with concatenative morphology and it does not restrict the number of morphemes that can be present in a single word. The input to the algorithm is a list of words in the language, possibly with frequencies of occurrence. The output is a segmentation for each vocabulary word (e.g., *affectionate* represented as *affect+ion+ate*). The algorithm minimizes a cost function composed of two parts; one piece is based on how well the model represents the observed data, and the other part measures the length of the segments (or codewords) that made up the model's vocabulary (Creutz & Lagus 2002). The procedure is entirely language independent.

4.4.3 Experiments

To test whether Morfessor segments might be a more effective form of tokenization I conducted experiments on all of the CLEF datasets described in Section 3.1. During indexing a word was split into segments (e.g., *affect+ion+ate*) and a posting entry was added for each piece (e.g., *affect*, *ion*, and *ate*). Undesirable confluations from segments from unrelated words like *ion* in the word *ionized* (*ion+ized*) do occur; however when this happens with affixes this cannot cause great harm because such terms are so common (i.e., they are stopwords, or have low IDF) and therefore are downweighted by the retrieval engine.

The default parameters for Morfessor were used as described in a technical report (Creutz & Lagus 2005). Case folding was performed and all 'words' containing digits (i.e., mainly numbers) were not segmented. A trivial one-line modification to the source code was made to enable use on non Latin-1 encoded text.

Table 4.8 gives the size of the input word lists for each language, along with the per-

⁴A Perl implementation is available from <http://www.cis.hut.fi/projects/morpho/>

Lang	Year	Surface Forms	Segmented	Segments / Word	Unique Segments
BG	2007	318757	90.2%	1.37	36263
CS	2007	456132	90.8%	1.31	54580
DE	2003	1180517	90.4%	1.34	77790
EN	2003	280853	93.4%	1.57	21839
ES	2003	545343	94.6%	1.59	37616
FI	2003	973231	93.3%	1.29	81818
FR	2003	264439	92.0%	1.55	26421
HU	2007	535263	93.0%	1.29	46897
IT	2003	370199	94.0%	1.51	26353
NL	2003	678033	92.6%	1.33	57760
PT	2006	420152	93.2%	1.51	35473
RU	2004	248493	92.1%	1.25	29584
SV	2003	494110	92.6%	1.39	42771

Table 4.8. Data for unsupervised morphology experiments.

centage of words that are segmented, the mean number of segments per segmented word, and the number of unique segments. The unique segments form the lexicon for the inverted index. Most words, (*i.e.*, between 90 and 95%) are split into multiple segments. The average number of segments per word is highest for the Romance family and English (above 1.50), and in the range of 1.29 to 1.39 for the other languages. As a first approximation, the number of unique segments produced is about one-tenth the size of the input word list.

Words, 4-grams, and 5-grams are compared to the segments produced by Morfessor in Table 4.9. Compared to words, the segments led to gains in 11 of 13 languages; English and Italian were the languages where word indexing received a higher score. On average, a 10% improvement in mean average precision was observed when segments were used for indexing instead of words. Segments achieved more than a 20% relative improvement in Bulgarian, German, and Russian, and over 40% in Czech and Hungarian.

In Table 4.2 it was shown that 4-grams and 5-grams have roughly a 20% advantage

Lang	Morfessor	Words	4-grams	5-grams
BG	0.2703	0.2164 [▼] (-19.9%)	0.3105 [▲] (+14.9%)	0.2820 (+4.3%)
CS	0.3215	0.2270 [▼] (-29.4%)	0.3294 (+2.5%)	0.3223 (+0.2%)
DE	0.3994	0.3303 [▼] (-17.3%)	0.4098 (+2.6%)	0.4201 (+5.2%)
EN	0.4018	0.4060 (+1.1%)	0.3990 (-0.7%)	0.4152 (+3.3%)
ES	0.4451	0.4396 (-1.2%)	0.4597 (+3.3%)	0.4609 (+3.6%)
FI	0.4018	0.3406 [▼] (-15.2%)	0.4989 [▲] (+24.2%)	0.5078 [▲] (+26.4%)
FR	0.3680	0.3638 (-1.1%)	0.3844 ^Δ (4.5%)	0.3930 [▲] (+6.8%)
HU	0.02921	0.1976 [▼] (-32.3%)	0.3746 [▲] (+28.2%)	0.3624 [▲] (+24.1%)
IT	0.3474	0.3749 ^Δ (+7.9%)	0.3738 ^Δ (+7.6%)	0.3997 [▲] (+15.1%)
NL	0.4053	0.3813 [▽] (-5.9%)	0.4219 (+4.1%)	0.4243 (+4.7%)
PT	0.3287	0.3162 (-3.8%)	0.3358 (+2.2%)	0.3524 ^Δ (+7.2%)
RU	0.3307	0.2671 [▽] (-19.2%)	0.3406 (+3.0%)	0.3330 (+0.7%)
SV	0.3738	0.3387 (-9.4%)	0.4236 [▲] (+13.3%)	0.4271 [▲] (+14.3%)
PMAP	0.3605	0.3230 (-10.4%)	0.3894 (+8.0%)	0.3923 (+8.8%)

Table 4.9. Comparing words, 4-grams, and 5-grams to Morfessor segments.

over words, therefore the segment-based indexing achieves approximately half of the gain observed with n-grams. However, these averages are skewed by a few languages where very large improvements occur when n-grams are used (*e.g.*, Finnish and Hungarian).

By examining rows in Table 4.6 and comparing with Table 4.9 it can be seen that rule-based stemming, in the lower-complexity languages for which it is available, does better than unsupervised segmentation. Parameter tweaking might improve the unsupervised segmentation, but this would conflict with the goal of language neutrality.

4.5 N-gram Stemming

The drawback of n-grams is not in retrieval accuracy, but rather in higher query execution time and storage requirements. Each character of a text begins a new n-gram, so an n-gram representation of a text contains many more indexing terms than does a word or stem representation. The number of posting entries for a document (or the number of terms in a query) become a function of the number of characters, not the number of words. Not

	Words	3-grams	4-grams	5-grams	6-grams	7-grams
Dutch	53	4678	1002	250	86	38
English	128	6803	1251	293	90	36
Finnish	10	2041	396	97	34	16
French	69	3038	642	170	63	30
German	40	5979	1089	259	89	40
Italian	65	4189	913	240	84	36
Spanish	98	6198	1112	286	102	46
Swedish	35	3763	572	131	44	20

Table 4.10. Average posting list length for words and n-grams (CLEF 2002 data).

only does this produce larger indexes, it also increases the number of disk seeks required to locate all of the postings for a query. As if this were not bad enough, for a typical n-gram the length of the postings list is longer than that of a typical word or stem. Therefore the data transfer time to read an entire term's posting list is greater. Average values of posting list length for several languages and n-gram sizes are shown in Table 4.10. For the most effective lengths of n-grams (*i.e.*, $n = 4$ and $n = 5$) the number of posting entries ranges from between 2 and 40 times longer than for words.

It would be desirable if one could take advantage of the simple approach to tokenization found with n-grams to simulate stemming in a language-neutral way without paying a concomitant performance penalty. A single carefully chosen n-gram from each word might serve as an adequate stem substitute. The most discriminating and semantically valuable n-gram probably comes from the morphologically invariant portion of the word. Since affixes that indicate a particular morphological variation (*e.g.*, -ation, -ing) will be repeated across many different words, they will be commonplace and will exhibit low IDF. Thus, a reasonable method of selection would be to choose a word-internal n-gram with the highest inverse document frequency (IDF) as a word's surrogate stem.

In Table 4.11 document frequencies are given for the n-grams that make up the word *precaution*. Frequently occurring n-grams such as *_pre* and *tion* occur in a majority of

4-grams	DF	5-grams	DF
<i>-pre</i>	97185	<i>-prec</i>	8556
<i>prec</i>	13633	<i>preca</i>	1452
<i>reca</i>	12047	<i>recau</i>	846
<i>ecau</i>	56736	<i>ecaut</i>	844
<i>caut</i>	4227	<i>cauti</i>	4200
<i>auti</i>	9307	<i>autio</i>	4201
<i>utio</i>	28857	<i>ution</i>	27891
<i>tion</i>	146622	<i>tion_</i>	136345
<i>ion_</i>	147873		

Table 4.11. Constituent n-grams by document frequency for *precaution*.

documents; here the collection contained 166754 documents. Often the n-grams around the morphological root are the least common of the n-grams that span the word as is the case with *caut*.

Like any technique for stemming, errors of over-conflation and under-conflation will occur. In Table 4.12 examples are given of several English words, their Snowball stems, and their least common n-gram stems.

In preliminary work on the CLEF 2002 test set this method of n-gram selection showed statistically significant improvements in English and Finnish (Mayfield & McNamee 2003). Here those results are extended by covering additional CLEF languages and a greater number of test topics. Table 4.13 shows how stems identified from the least common n-grams compare to the tokenization alternatives previously investigated using mean average precision. It appears that with $n = 4$ or $n = 5$, n-gram stemming is comparable to Morfessor segmentation. Like Morfessor segments, performance of the least frequent n-gram stems is better than words, worse than full n-gram indexing, and a bit worse than Snowball stemming (in those languages that are supported by the stemmer). The row marked ‘8 Langs’ gives an average for the 7 methods across the 8 languages for which Snowball stems could

Word	Snowball stem	4-gram stem	5-gram stem
author	author	hor_	thor_
authored	author	hore	thore
authorized	author	oriz	orize
authorship	authorship	orsh	horsh
reauthorization	reauthor	oriz	eauth
eat	eat	_eat	_eat_
eating	eat	_eat	_eati
eater	eater	_eat	_eate
eaten	eaten	_eat	_eate
juggle	juggl	jugg	juggl
juggled	juggl	jugg	juggl
jugglers	juggler	jugg	ggler

Table 4.12. Examples of n-gram stemming.

be compared.

Selecting a single n-gram per word results in an inverted file with the same number of postings as a word or stem index, and in queries that have the same number of terms as for words or stems; thus, the performance penalty paid by full n-gram indexing is ameliorated. The technique requires a priori knowledge of n-gram frequencies, but calculating such frequencies is straightforward given a monolingual collection in the target language. Table lookup can be used so that only a small constant factor is added to index creation and query processing times.

This method of indexing can also be viewed as a form of pruning an n-gram index since one n-gram per word is chosen and others, including word-spanning n-grams, are essentially discarded. Other methods of pruning a full n-gram index might also be effective, such as selecting two n-grams for each word, selecting the least common n-gram and its left and right neighbor n-gram, or pruning terms at the document-level rather than per word as in Carmel *et al.* (2001).

Lang	words	stems	morfessor	4-stem	5-stem	4-grams	5-grams
BG	0.2164		0.2703	0.2822	0.2442	0.3105	0.2820
CS	0.2270		0.3215	0.2567	0.2477	0.3294	0.3223
DE	0.3303	0.3695	0.3994	0.3464	0.3522	0.4098	0.4201
EN	0.4060	0.4373	0.4018	0.4176	0.4175	0.3990	0.4152
ES	0.4396	0.4846	0.4451	0.4485	0.4517	0.4597	0.4609
FI	0.3406	0.4296	0.4018	0.3995	0.4033	0.4989	0.5078
FR	0.3638	0.4019	0.3680	0.3882	0.3834	0.3844	0.3930
HU	0.1976		0.2921	0.2836	0.2668	0.3746	0.3624
IT	0.3749	0.4178	0.3474	0.3741	0.3673	0.3738	0.3997
NL	0.3813	0.4003	0.4053	0.3836	0.3846	0.4219	0.4243
PT	0.3162		0.3287	0.3418	0.3347	0.3358	0.3524
RU	0.2671		0.3307	0.2875	0.3053	0.3406	0.3330
SV	0.3387	0.3756	0.3738	0.3638	0.3467	0.4236	0.4271
All	0.3230		0.3605	0.3518	0.3466	0.3894	0.3923
8 Langs	0.3719	0.4146	0.3928	0.3902	0.3883	0.4214	0.4310

Table 4.13. Effectiveness of 7 tokenization methods, including n-gram stemming.

4.6 Conclusions

N-gram tokenization is a language-neutral technique that is very effective in a mono-lingual setting. In 13 European languages n-grams of various lengths were compared and $n = 4$ and $n = 5$ were found to be optimal, and nearly equally effective. Compared to the use of unnormalized words as indexing terms a 21% relative improvement was observed. N-grams also yielded better performance than several approaches to stemming. The advantage with n-gram indexing appears greatest in morphologically complex languages. Using n-gram tokenization incurs a performance penalty, but a method for trading off some of the gains with n-grams by deleting the least valuable n-grams was introduced; alternatively this can be viewed as a form of stemming which totally removes the additional disk and run-time penalty associated with n-grams.

Next we explore several additional facets of n-gram tokenization in Chapter 5 before taking up the study of tokenization and translation issues in bilingual retrieval in Chapter 6.

Chapter 5

ADDITIONAL N-GRAM EXPERIMENTS

In Chapter 4 a variety of alternatives to word-based indexing were studied. Of all the tokenization alternatives examined, character n-grams resulted in the highest overall accuracy across thirteen languages. More than a 20% improvement over words was observed.

In this chapter three additional monolingual investigations using n-gram tokenization are conducted: (1) the use of automated relevance feedback with n-grams; (2) templatic skip n-grams; and (3) reasons to explain the utility of n-grams. Experiments are run using the same tests sets used in Chapter 4, and n-grams of lengths $n = 4$ and $n = 5$ are emphasized as they proved most effective in the previous studies.

5.1 Relevance Feedback

Automated relevance feedback is a technique that attempts to obtain improved retrieval results by modifying an initial query based on the highly ranked documents initially retrieved (Harman 1992). The method is motivated by observing users modify initial queries upon reading retrieved passages. For example upon scanning documents retrieved for a query “*doping scandals*” a user might add keywords like *steroids* and *testosterone*. Typically, frequently observed terms from highly ranked documents are added to the initial query vector and a subsequent document ranking is calculated. Term weights can also be

adjusted to modify the relative contribution of each term in the computation. The two-step procedure can be performed in a transparent fashion so the user perceives the system as performing a single function. While significant gains in performance can sometimes be obtained, these improvements must be measured against the increased run-time costs incurred because of multiple retrieval passes.

Because n-grams are a very different representation of a document and single n-grams often have less information content than traditional words, it is not clear that the methods for assigning term relevance values or performing query expansion should be the same. Here we look at the efficacy of relevance feedback as a function of the number of top-ranked documents used and the number of expansion terms.

To create a modified query vector for a second retrieval pass the following procedure is performed. The initial query is used to search for top-ranked documents. Then candidate terms are identified as being statistically important in top-ranked documents compared to lowly ranked documents (*i.e.*, terms that are topical, but not likely to be highly relevant). A metric derived on mutual information is used and a set number of terms is selected and then weighted in a modified query vector. The term selection metric is:

$$(5.1) \quad (P_{local}(t) - P_{global}(t)) \times idf(t)^{1.25}$$

where for term t , P_{local} and P_{global} are based on relative document frequencies and $idf(t)$ is inverse document frequency¹.

Figure 5.1 contains plots that illustrate the effect of the number of query terms in the expanded query in six of the CLEF languages. Mean average precision is charted on the vertical axis of each plot and the number of query terms (25 to 800) is on the horizontal. The vertical axes have different scales in Figure 5.1 (a-f), but the scores in MAP are not

¹ $IDF(t) = \log_2(\frac{N}{df(t)})$

commensurable across languages. Here the 20 top-ranked documents were used to select good expansion terms. Each plot displays curves for words, 4-grams, and 5-grams.

Across the six charts several trends are discernible:

- N-grams achieve maximal performance after a moderate number of query terms are used.
- When word-based tokenization is used performance declines as the number of expansion terms increases.
- The drop off in performance with additional expansion terms is greater for words than with n-grams.

These observations are also apparent in the other CLEF languages.

We explore the combination of how many top documents should be used as well as the number of query terms. Looking at just the English data set we examined a range of top documents (10, 15, 20, 25, and 30) and a range of query terms (again 25 to 800). In Figure 5.2 three graphs reveal the interrelationship between these two parameters for (a) words, (b) 4-grams, and (c) 5-grams.

For all three token types a smaller number of top-ranked documents should be used for query expansion. But Figure 5.2 (a) is visibly different from Figure 5.2 (b) and (c) and it is clear that the optimal number of query terms is different. With words 25 terms is optimal, and with n-grams 200 terms is a good choice, in English and in other languages.

It would make sense that n-grams, being more conflationary, would require a greater number of terms for topical cohesion.

Relevance feedback does not benefit words and n-grams equally. In Table 5.1 baseline runs (*i.e.*, without RF) and those using feedback are presented. With words, 10 documents and 25 terms were used and with 4-grams and 5-grams, 10 documents and 200 terms were

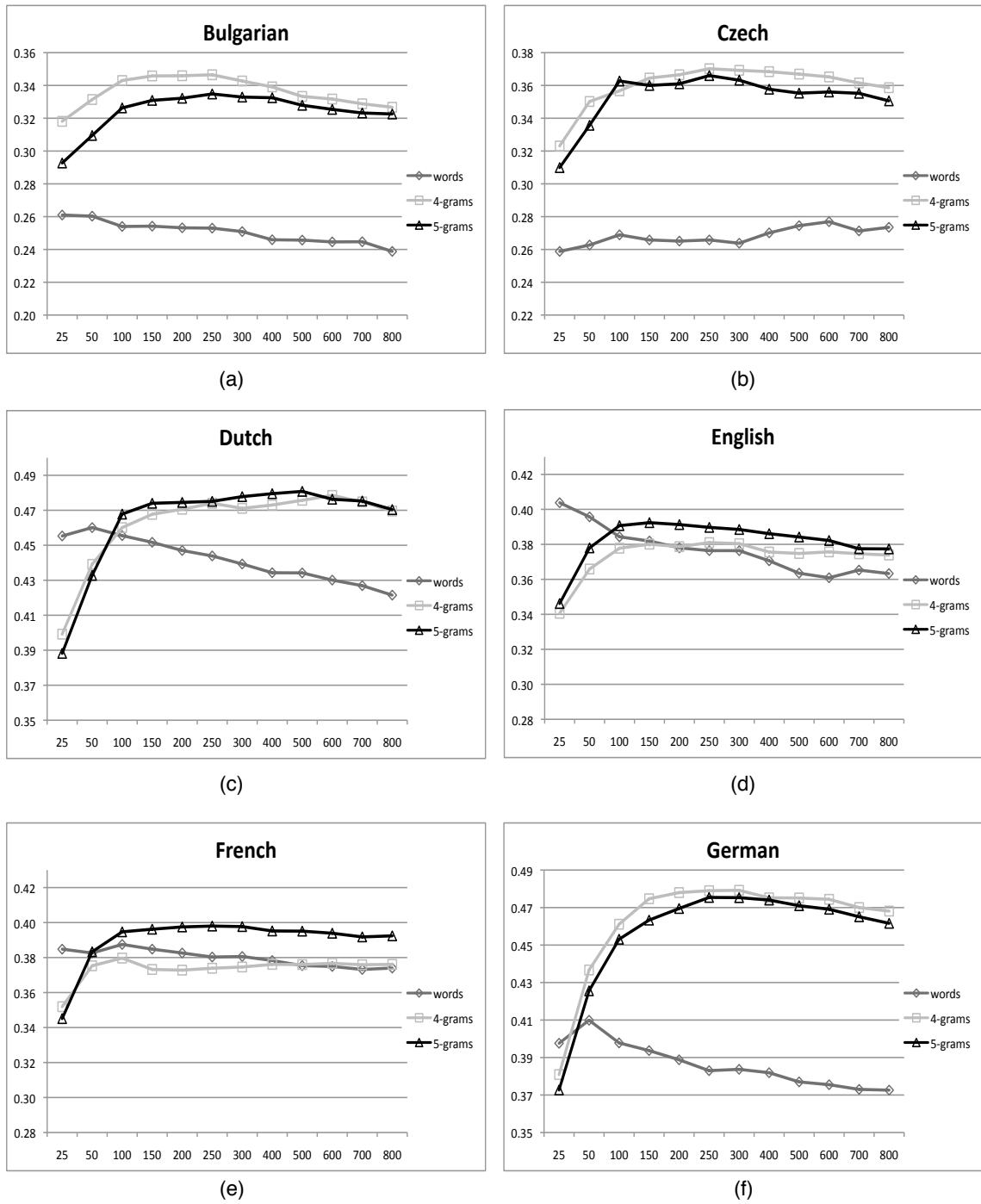
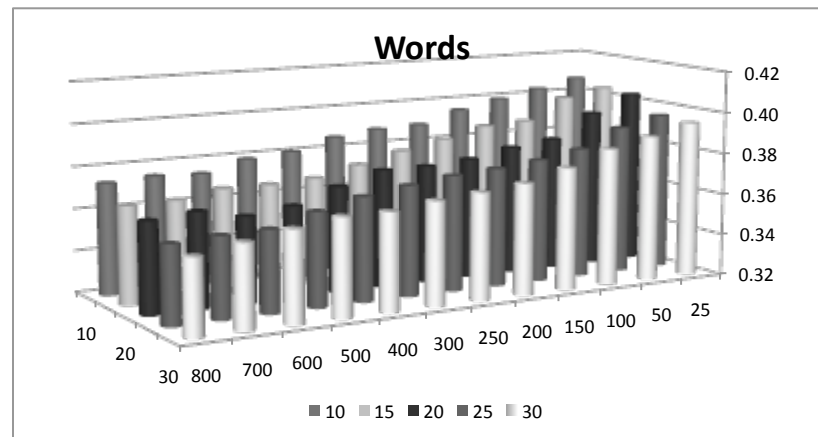
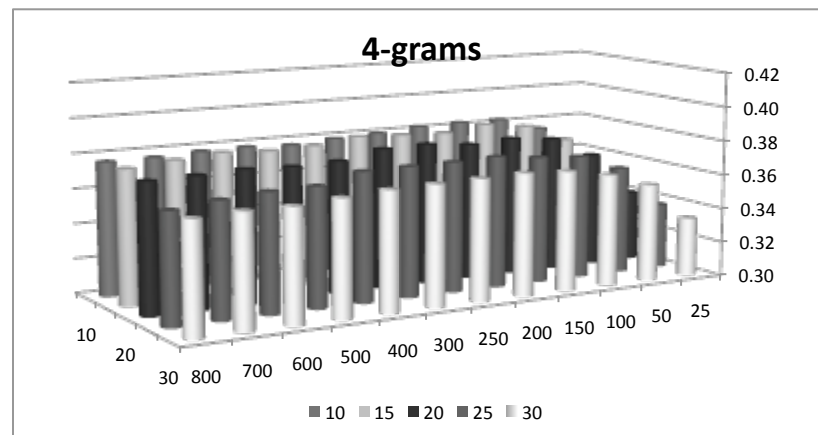


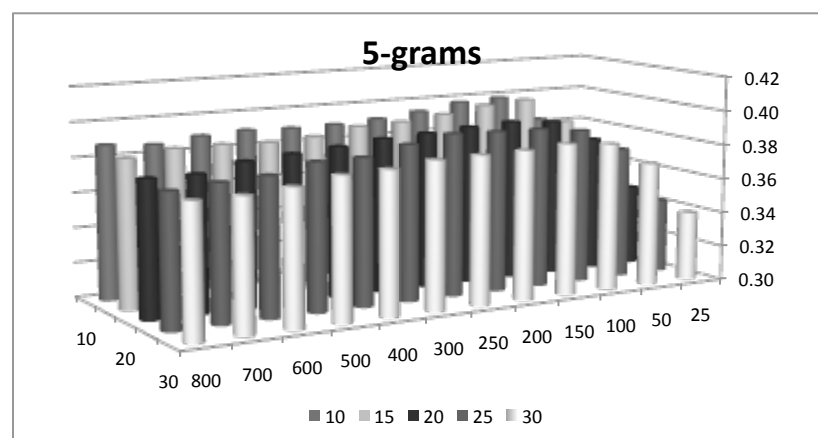
FIG. 5.1. Effect of the number of expansion terms on retrieval performance.



(a)



(b)



(c)

FIG. 5.2. MAP variation based on relevance feedback parameter settings in English.

	Baseline			With Relevance Feedback		
	words	4-grams	5-grams	words	4-grams	5-grams
BG	0.2164	0.3105	0.2820	0.2622 (+21.2%)	0.3658 (+17.8%)	0.3473 (+23.2%)
CS	0.2270	0.3294	0.3223	0.2786 (+22.7%)	0.3601 (+9.3%)	0.3664 (+13.7%)
DE	0.3303	0.4098	0.4201	0.4101 (+24.2%)	0.4709 (+14.9%)	0.4738 (+12.8%)
EN	0.4060	0.3990	0.4152	0.4073 (+0.3%)	0.3823 (-4.2%)	0.3939 (-5.1%)
ES	0.4396	0.4597	0.4609	0.4894 (+11.3%)	0.4910 (+6.8%)	0.5040 (+9.4%)
FI	0.3406	0.4989	0.5078	0.3154 (-7.4%)	0.4206 (-15.7%)	0.4424 (-12.8%)
FR	0.3638	0.3844	0.3930	0.3902 (+7.3%)	0.3810 (-0.8%)	0.4047 (+3.0%)
HU	0.1976	0.3746	0.3624	0.2605 (+31.8%)	0.4112 (+9.8%)	0.4156 (+14.7%)
IT	0.3749	0.3738	0.3997	0.4245 (+13.2%)	0.3889 (+4.0%)	0.4099 (+2.5%)
NL	0.3813	0.4219	0.4243	0.4575 (+20.0%)	0.4751 (+12.6)	0.4902 (+15.5%)
PT	0.3162	0.3358	0.3524	0.3633 (+14.9%)	0.3575 (+6.5%)	0.3912 (+11.0%)
RU	0.2671	0.3406	0.3330	0.1905 (-28.7%)	0.2319 (-31.9%)	0.2308 (-30.7%)
SV	0.3387	0.4236	0.4271	0.3365 (-0.7%)	0.4035 (-4.7%)	0.4136 (-3.2%)
PMAP	0.3230	0.3894	0.3923	0.3528 (+9.2%)	0.3954 (+1.5%)	0.4064 (+3.6%)

Table 5.1. Relative gain from automated relevance feedback by tokenization method.

used. The gains observed with relevance feedback are more substantial for words than they are for n-grams. Averaged across the 13 languages, a nearly 9% improvement in mean average precision is observed with words; however gains for n-grams are much less, only 4% for 5-grams, and 2% for 4-grams. While n-grams still possess an advantage in accuracy, when relevance feedback is employed, the performance gap narrows.

5.2 Skip N-grams

Consider the present tense conjugation of the Spanish verb *contar* (*to count*): *cuento*, *cuentas*, *cuenta*, *contamos*, *contáis*, and *cuentan*. Such inflectional variation can cause lexical mismatches that would impair retrieval, and character n-grams are unlikely to be a complete solution to this problem since the 1st and 2nd person plural forms do not share longer n-grams with the other forms². Similarly the English verb “to swim” has past tense

²Table 4.2 showed that short n-grams are not effective for retrieval. Here *nta* is the only matching 3-gram.

swam and perfect tense swum; none of the forms have n-grams in common³. Similar problems also happen with nouns, for example, in Welsh *plentyn* (*child*) and its plural, *plant* (*children*). Yet each of these examples contain patterns that could enable matching. Using a dot symbol to indicate one or more skipped characters, expressions *c•nt*, *sw•m*, and *pl•nt* would match all these forms.

Other approaches to canonicalize related morphological forms for retrieval are possible, but not investigated here. For example, if all vowels were mapped to ‘a’ and adjacent vowels truncated to a single one, n-gram matches would be available in each of the instances mentioned above.

5.2.1 Previous Work

At the word level skip n-grams have been proposed in the speech recognition community as a means of coping with data sparseness in language modelling (Guthrie *et al.* 2006; Siu & Ostendorf 2000) and for use in n-gram based MT evaluation (Lita, Rogati, & Lavie 2005). In information retrieval the focus has been on letter-based skipgrams. Pirkola *et al.* (2002) have proposed n-grams with skips⁴ to match terminology for cross-language information retrieval in languages sharing a common alphabet. For example, the English word *calcitonin* can be matched to its Finnish translation *kalsitoniini*, supported in part by matches like *l•t* and *n•n*. Mustafa (2004) proposed a similar method for monolingual Arabic language processing, where infix morphological changes are common. He identified relevant dictionary terms using bigrams with and without a single skip character and a Dice coefficient to compare sets of bigrams. Järvelin *et al.* (2007) formalized the notion of skipgrams and investigated methods of comparing lexical terms; however, they focused on the case where a single skip is formed by deleting contiguous letters. This makes sense when

³Not counting bigram *sw*

⁴They use the term s-grams. I refer to these as skip n-grams or skipgrams.

only bigrams are considered – in this case the only place to skip characters is between the first and last letters of the (skip) bigram.

But character skipgram methods can be generalized even further by including the possibility of multiple non-adjacent skips within a single word. In this research skipgrams are considered as an alternative method for tokenization that might support matches across morphologically related words, with application for both monolingual and bilingual retrieval. Two variations of skipgrams are considered: where a sequence of deleted characters is indicated with a special character in the resulting n-grams (*e.g.*, ●) or where no explicit indication that letters were removed is placed in the resulting string.

5.2.2 Examples and Performance Analysis

The notation $s_{n:i-j}$ will be used to indicate skip n-gram processing where n characters are preserved and $k \in i, i + 1, \dots, j$ characters can be deleted; the deleted characters need not be contiguous. Skipgrams can be generated by examining a window of $n + j$ characters and removing all combinations of $i, i + 1, \dots, j$ characters. Without loss of generality we stipulate that $n \geq 2$ and prohibit removal of the initial and final letters⁵. Placing a dot over the s , we can use $\dot{s}_{n:i-j}$ to denote skipgrams with deleted letters replaced with a special symbol. Unless stated otherwise words will be padded with underscores to indicate word boundaries. Consistent with the work in Chapter 4 word-spanning skipgrams will be generated.

Letting X represent a preserved letter and ● a deleted one, the templates for producing $s_{4:2}$ skipgrams for a six-letter word are given in Table 5.2. Examples of skipgrams generated for the isolated word *crust* are given in Table 5.3. In the table the twelve entries for the $s_{4:2}$ and $\dot{s}_{4:2}$ skipgrams consist of six entries starting with the first pad character ‘_’ and

⁵Omitting the initial or final letters would be equivalent to working with a length of $n - 1$. Or $n - 2$ if both were left out.

X●●XXX
X●X●XX
X●XX●X
XX●●XX
XX●X●X
XXX●●X

Table 5.2. The templates for producing skipgrams with four letters and two skips.

$s_{3:0-2}$	0: _cr, cru, rus, ust, st_ 1: _ru, _cu, cus, crs, rst, rut, ut_, us_ 2: _us, _rs, _cs, cst, cut, crt, rt_, rs_, ru_
$\dot{s}_{3:0-2}$	0: _cr, cru, rus, ust, st_ 1: _●ru, _c●u, c●us, cr●s, r●st, ru●t, u●t_, us●_ 2: _●●us, _●r●s, _c●●s, c●●st, c●u●t, cr●●t, r●●t_, r●s●_, ru●●_
$s_{4:0-2}$	0: _cru, crus, rust, ust_ 1: _rus, _cus, _crs, cust, crst, crut, rst_, rut_, rus_ 2: _ust, _rst, _rut, _cst, _cut, _crt, cst_, cut_, cus_, crt_, crs_, cru_
$\dot{s}_{4:0-2}$	0: _cru, crus, rust, ust_ 1: _●rus, _c●us, _cr●s, c●ust, cr●st, cru●t, r●st_, ru●t_, rus●_ 2: _●●ust, _●r●st, _●ru●t, _c●●st, _c●u●t, _cr●●t, c●●st_, c●u●t_, c●us●_, cr●●t_, cr●s●_, cru●●_

Table 5.3. Examples of skipgrams generated from ‘_crust_’.

six more starting at the ‘c’; each set of six matches the templates from Table 5.2.

An individual word in a document results in just one inverted file posting. N-grams are a more redundant representation than words, and for a word of length l (in isolation), generate $l - n + 1$ postings list entries (or $l - n + 3$ with padding). But skipgrams are an even more productive representation. A skipgram with n preserved letters and k skipped letters generates $\binom{n+k-2}{k}$ strings per starting position, and thus for a word of length l , there are:

$$(5.2) \quad (l - n + k - 1) \times \binom{n + k - 2}{k}$$

Indexing term	Postings	Indexing term	Postings
2-grams	15	4-grams	13
$s_{2:0-1}$	28	$s_{4:0-1}$	48
$s_{2:0-2}$	39	$s_{4:0-2}$	109
$s_{2:0-3}$	45	$s_{4:0-3}$	196
3-grams	14	5-grams	12
$s_{3:0-1}$	39	$s_{5:0-1}$	55
$s_{3:0-2}$	71	$s_{5:0-2}$	149
$s_{3:0-3}$	107	$s_{5:0-3}$	310

Table 5.4. Number of postings generated from skipgrams for *sesquipedalian*.

postings list entries per word. With multiple skips this becomes:

$$(5.3) \quad (l - n + k - 1) \times \sum_{k=i}^j \binom{n + k - 2}{k}$$

This will certainly become a problem for large k .

To illustrate the extraordinary redundancy that skipgram indexing affords, Table 5.4 lists the number of postings added to an inverted file as a results of seeing the isolated word *sesquipedalian*. Note that when word or stemmed words were used as indexing terms only a single posting is generated, but when sequences of 5 letters with 0 to 3 skips are used (*i.e.*, $s_{5:0-3}$) then over 300 indexing terms are generated from this single word!

5.2.3 Experiments

Despite possible efficiency concerns, it would be nice to know how the skipgram technique compares to traditional n-gram indexing, which is examined next. Because traditional, skipless n-grams (*i.e.*, $s_{4:0}$ and $s_{5:0}$) have been found effective, it will make sense to give strong consideration to setting $i = 0$. For pragmatic (*i.e.*, efficiency) reasons we will want to keep $j \leq 3$.

The seven CLEF languages with greater morphological complexity were used to com-

		0 or 1 skip		0 – 2 skips	
Lang	5-grams	$s_{3:0-1}$	$\dot{s}_{3:0-1}$	$s_{3:0-2}$	$\dot{s}_{3:0-2}$
BG	0.2820	0.2101	0.2633	0.1595	0.2606
CS	0.3223	0.2451	0.3161	0.2011	0.3106
DE	0.4201	0.2843	0.3499	0.2259	0.3452
FI	0.5078	0.3267	0.4027	0.2594	0.3977
HU	0.3624	0.3304	0.3613	0.2910	0.3547
RU	0.3330	0.2639	0.3066	0.2289	0.3213
SV	0.4271	0.3173	0.3664	0.2580	0.3606
Average	0.3792	0.2825	0.3380	0.2320	0.3358

Table 5.5. Skipgram results with 3 preserved letters.

		0 or 1 skip		0 – 2 skips		0 or 1 skip	
Lang	5-grams	$s_{4:0-1}$	$\dot{s}_{4:0-1}$	$s_{4:0-2}$	$\dot{s}_{4:0-2}$	$s_{5:0-1}$	$\dot{s}_{5:0-1}$
BG	0.2820	0.2828	0.2919	0.2510	0.2695	0.2656	0.2629
CS	0.3223	0.3202	0.3267	0.2990	0.3051	0.3086	0.3027
DE	0.4201	0.3837	0.4094	0.3437	0.3902	0.3927	0.3939
FI	0.5078	0.4744	0.4974	0.4176	0.4743	0.4717	0.4700
HU	0.3624	0.3685	0.3693	0.3472	0.3524	0.3438	0.3397
RU	0.3330	0.3216	0.3346	0.2918	0.3265	0.3165	0.3173
SV	0.4271	0.3958	0.4142	0.3697	0.3976	0.3983	0.4025
Average	0.3792	0.3639	0.3776	0.3314	0.3594	0.3567	0.3556

Table 5.6. Skipgram results with 4 or 5 preserved letters.

pare the use of skipgram indexing to traditional overlapping n-grams. No relevance feedback was applied, and only *title+description* runs were performed as is the case throughout this dissertation. Tables 5.5 and 5.6 present mean average precision for skipgrams with between 3 and 5 preserved letters and up to 2 skip positions. The indexing methods that outperform character 5-grams are emboldened.

The skipgram variants that retain 3 letters are not generally competitive with the traditional 5-gram baseline. In Czech and Hungarian, skipgrams with 0 and 1, or 0, 1, and 2 letters that are replaced with a wildcard symbol, experience only a slight loss. Character 3-grams were studied in Chapter 4 and they did not perform on par with 4- and 5-grams,

therefore it is not too surprising that skipgrams with 3 retained letters do not outperform 5-grams. Of these variants $\dot{s}_{3:0-1}$ is the highest performing; averaged across these seven languages only about a 10% decline is observed.

With longer classes of skipgrams results rivaling basic 5-grams are observed. On average $\dot{s}_{4:0-1}$ is equivalent to 5-grams (0.3792 vs. 0.3776) and yields higher performance in Bulgarian, Czech, Hungarian, and Russian. Of the runs where skipgrams showed an improvement, only the skip 4-grams with at most one skip yielded significant improvements ($p < 0.01$). 5-grams with skips were not as competitive as the skip 4-grams.

Two caveats should be kept in mind in understanding these results. First, in these experiments the skipgrams were allowed to span word boundaries. It is possible that word-spanning skipgrams introduce harmful pollution in the indexing representation and therefore the word-internal variant should also be considered. Second, the skipgram classes investigated always included traditional n-grams (*i.e.*, with zero skips) in addition to the strings created from skips⁶.

Summing up these results, we can say that skipgrams perform reasonably well, but they are not demonstrably more effective than plain character n-grams. With their significantly higher disk space and query time costs, it would be hard to advocate their use based on these findings. Still, skipgrams are a highly redundant representation of text and may still have application to retrieval in Semitic languages, due to infix morphology, or in OCR'd documents where the letter error rate may be high. They may also be useful for specialized applications such as spam detection in email where spelling errors are deliberately introduced in an attempt to obscure the fact that marketing is being attempted (*e.g.*, matching *di\$count* and *discount*). Investigating these conjectures is beyond this scope of this dissertation.

⁶It is possible, of course, to perform skipgram indexing where basic n-grams are not included in the representation. No such experiments were conducted in this research

5.3 Reasons for N-gram Effectiveness

There are a number of factors that could be the underlying cause of the 20% improvement that was reported in Chapter 4. The gains observed with n-grams could be due to:

- robustly coping with spelling variations (*e.g.*, *Jacobsson/Jacobssen* or *color/colour*) or misspellings because of the redundancy that comes from having multiple indexing terms that occurs from different sections of a word;
- the word-spanning n-grams that provide evidence about word adjacency;
- handling morphological variation, including inflectional changes and compounding;
- combination of the above, or an unidentified factor.

While spelling normalization and word-spanning n-grams likely have a beneficial effect, morphological variation would seem like a major cause since it was observed in Chapter 4 that n-grams were most advantageous in languages with greater estimated morphological complexity.

5.3.1 Misspellings

It is a challenging exercise to design an experiment to explore the effect of spelling errors and variants on retrieval effectiveness. While an error model could be applied to a document collection to introduce degradations, such models introduce artificialities that would obscure the results.

However even without conducting such an experiment we can reasonably conjecture that spelling mistakes alone cannot account for a large difference in effectiveness on the order of 20%. This is because the CLEF test collections are news corpora and spelling error rates in journalistic text are on the order of 1 word in 2000 (Church & Gale 1991).

Even at 25 words per thousand, a rate measured in secondary school writing (Mitton 1987), spelling errors would not account for such a large change in retrieval effectiveness.

5.3.2 Word-Spanning N-grams

Examining the contribution of word-spanning n-grams is straightforward; by generating only word-internal n-grams and comparing the retrieval efficacy between the two conditions, the contribution of the word-crossing n-grams can be measured. Leading and trailing word boundaries are still identified with an underscore character (*e.g.*, `_four` and `four_`), but n-grams like `our_s` (from `four score`) that contain letters from adjacent words are not produced.

Table 5.7 compares the two types of n-gramming. Averaged across the languages the 4-grams improve from 0.3830 to 0.3851 (+0.5%) when word-spanning n-grams are included, as was the normal mode in Chapter 4. Performance for 5-grams goes from 0.3832 to 0.3880, a +1.3% relative improvement. Both gains are very slight, and it is clear that providing surrogate phrasal information cannot be a principal reason for the superior performance with n-gram indexing. I conjecture that 4-grams are simply too short to convey strong cues of a multiword phrase and that is why word-crossing 5-grams see a slightly larger percent improvement.

5.3.3 Removing Morphology

In an effort to establish whether or not coping with morphological processes such as inflection, derivation, and compounding is the prime reason behind n-gram's monolingual effectiveness, we can attempt to remove morphology from language and see what changes occur. Inspired by Juola's work in degrading morphology (1998) a method of altering every word in the lexicon will be performed and retrieval experiments can be run against indexes created using word-based or n-gram-based tokenization on the transformed words. If the

	Internal		Spanning	
Lang	4-grams	5-grams	4-grams	5-grams
BG	0.3016	0.2866	0.3105	0.2820
CS	0.3329	0.3245	0.3294	0.3223
DE	0.4045	0.4129	0.4098	0.4201
EN	0.3948	0.4037	0.3990	0.4152
ES	0.4578	0.4662	0.4597	0.4609
FI	0.5006	0.4882	0.4989	0.5078
FR	0.3796	0.3886	0.3844	0.3930
HU	0.3714	0.3490	0.3746	0.3624
IT	0.3672	0.4053	0.3738	0.3997
NL	0.4141	0.4050	0.4219	0.4243
PT	0.3367	0.3475	0.3358	0.3524
RU	0.3610	0.3393	0.3406	0.3330
SV	0.4126	0.4234	0.4236	0.4271
Average	0.3894	0.3923	0.3873 (+0.5%)	0.3877 (+1.2%)

Table 5.7. Gain from word-spanning n-grams.

relative advantage of character n-grams disappears this will be strong evidence that it is by addressing morphology that n-grams improve on word-based indexing.

It remains to decide how to manipulate each word in the lexicon to subtract morphological regularity. Each surface form must be modified in a consistent fashion throughout the corpus so that an input query word still matches all occurrences in the document collection. Simply sorting the letters in alphabetical order would be consistent, but not effectual in removing morphology; a word like *stroke* and its past tense would be represented as *ekorst* and *dekorst*, which share too much in common. However, by randomly shuffling the order of the characters in each word, each word can be transmuted in a way that preserves its length and removes orthographic regularity. Affixes like *pre-* or *-ing* will become much less apparent and the morphemes in related words (*e.g.*, *golfed* and *golfing*) will become difficult, if not impossible, to detect.

This method of lexical transformation should be adequate, if not quite perfect for our

Original Word	DF	Shuffled Form	DF
ate	613	aet	1316
eat	2459	tae	2459
tea	741	aet	1316
team	16605	tema	16605
meat	1217	maet	1217
luau	20	luua	20
lull	119	lull	119
golfer	258	legfro	258
golfed	5	dofegl	5
golfing	97	ligfgon	97
golfball	2	gaboflll	2

Table 5.8. Sample word transformations (CLEF 2000 English corpus).

purposes; small words (*e.g.*, *I*, *be*) and those with mostly repeated characters (*e.g.*, *oology* or *lull*) will bear a strong resemblance to their original forms after scrambling the letters. The effect on word-based indexing should be minimal, although some increase in polysemy is possible due to manufactured confluations in the transformed representations. This might happen because anagrams, such as *team* and *meat*, could become cognates through shuffling if each was converted to *eamt*. The probability of this occurring depends on the number of anagrams, their length in characters, and the number of duplicate letters. This method of removing morphology will not distinguish between morphological processes such as inflection and compounding; some types of morphology may have a more significant impact on retrieval than others, but this experiment will not explain the relative contribution of different morphological processes in a language.

Table 5.8 lists several examples of words and their permuted forms, along with each's respective document frequency. The first group illustrates that additional confluations will occur as both *ate* and *tea* are transformed to *aet*, which has a document frequency near the sum of the number of documents that the original terms appeared in. Anagrams *team* and *meat* remain separate in the transformed space. No constraint was imposed to ensure

	Regular	Shuffled
Words	235713	230662
4-grams	146223	284381
5-grams	788187	1866806

Table 5.9. Distinct indexing terms (CLEF 2000 English corpus).

that shuffled forms differed from their original strings; while this is unlikely with longer terms, the word *lull*, which only has four possible forms depending on where the letter ‘u’ is positioned, is an example of a word left unaltered. Finally, the last grouping in Table 5.8 demonstrates how related forms of the lexeme *golf* lack any resemblance in their scrambled representations.

In Table 5.9 the number of dictionary terms is given for both the original and shuffled indexes for the CLEF 2000 English corpus. While the number of word forms actually decreases slightly for word-based indexing, because of the anagram conflation mentioned above, the number of distinct n-grams about doubles. This is because with morphology effectively removed from the language, orthographic sequences are much less regular.

Now we examine whether the relative effectiveness of n-grams changes when the letters of words are randomly scrambled. Table 5.10 shows how performance varies when word-based indexing, and character n-grams of lengths 4 and 5 are used on both unaltered words and when the letters in each words are randomly shuffled around. Several trends are identical in each language, and are reflected in the average as well:

- No change occurs when using space-separated words as indexing terms.
- The n-grams of both lengths perform markedly worse, suffering a 27% decline in mean average precision, averaged over all languages. Performance falls to the level of words, or for the morphologically simpler languages, below that of word-based indexing.

Lang	Regular			Shuffled		
	Words	4-grams	5-grams	Words	4-grams	5-grams
BG	0.2164	0.3105	0.2820	0.2164	0.1709	0.1697
CS	0.2270	0.3294	0.3223	0.2270	0.2111	0.2283
DE	0.3303	0.4098	0.4201	0.3303	0.2832	0.2770
EN	0.4060	0.3990	0.4152	0.4063	0.3647	0.3592
ES	0.4396	0.4597	0.4609	0.4375	0.3797	0.3709
FI	0.3406	0.4989	0.5078	0.3406	0.3127	0.3107
FR	0.3638	0.3844	0.3930	0.3635	0.3230	0.3226
HU	0.1976	0.3746	0.3624	0.1978	0.1861	0.1858
IT	0.3749	0.3738	0.3997	0.3747	0.3363	0.3310
NL	0.3813	0.4219	0.4243	0.3813	0.3298	0.3231
PT	0.3162	0.3358	0.3524	0.3165	0.2700	0.2715
RU	0.2671	0.3406	0.3330	0.2671	0.2250	0.2441
SV	0.3387	0.4236	0.4271	0.3387	0.2893	0.2860
Average	0.3230	0.3894	0.3923	0.3229 (-0.0%)	0.2832 (-27.3%)	0.2831 (-27.9%)

Table 5.10. Change observed by scrambling the letters in words.

Figure 5.3 plots the percent change in performance for each language when character 5-grams are used instead of ordinary words. The triangles indicate regular 5-grams and circles are used for 5-grams that are generated from the documents with permuted words. Languages are ordered left-to-right by the magnitude of the decline in performance.

These results give strong evidence that is the ability of overlapping character n -grams to capture regularity across morphologically related words forms that gives them their primary advantage. This is consistent with the observation reported in Chapter 4 that n -grams are more powerful in morphologically richer languages.

If it is the isolation of the root morpheme (or in compounds, roots) that is key, then these findings also suggest why longer length n -grams such as $n = 6$ and $n = 7$ are less effective than $n = 4$ and $n = 5$: longer sequences of characters are not focused on morphemes and fail to match some inflected allomorphs.

This also gives hope that the computational expense incurred with n -gram indexing

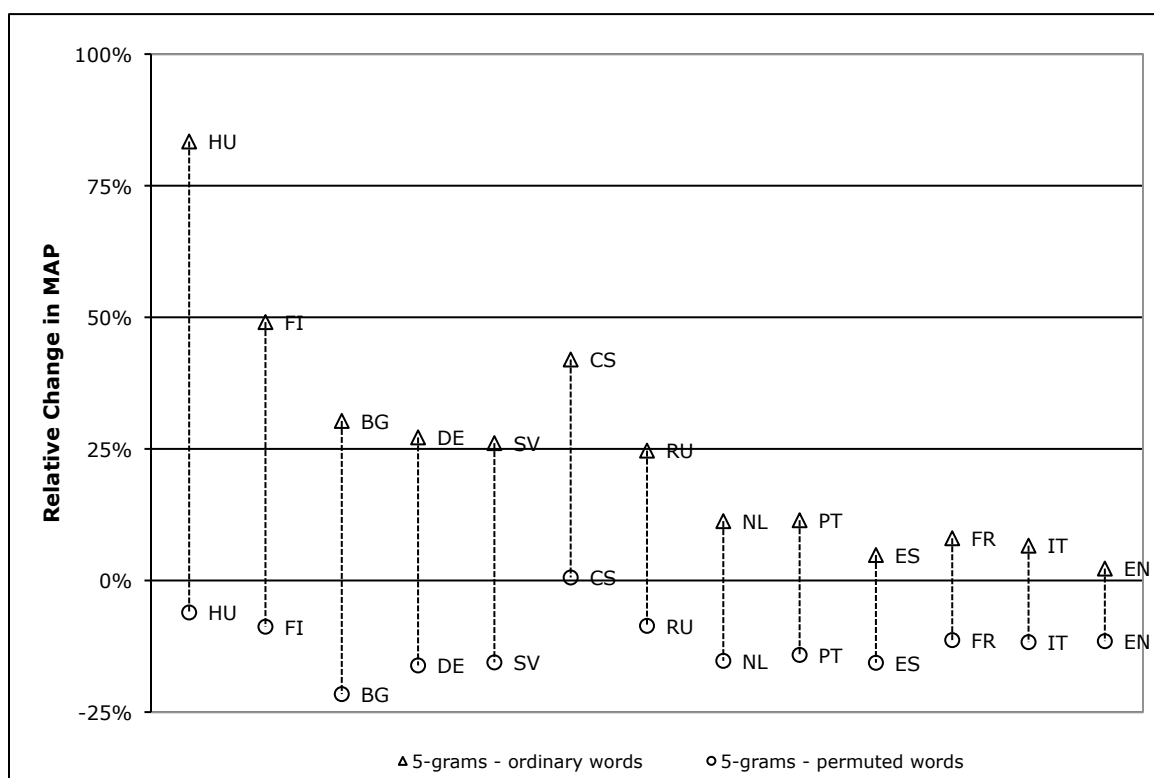


FIG. 5.3. Comparative efficacy of 5-grams against words, when the order of letters in words has, and has not, been scrambled throughout the corpus.

can be substantially reduced through aggressive pruning based on detecting morphological roots.

If morphological variability is the key factor, than matching of proper names should not be greatly effected by the use of n-grams, but many names are not consistently spelled, and it may be that n-grams have some advantage with queries involving proper nouns.

Further investigation of these conjectures is beyond the scope of this dissertation.

5.4 Conclusions

Three additional areas in monolingual n-gram retrieval were explored in this chapter. Relevance feedback using n-grams was studied and it was found that n-grams require a greater number of expansion terms than words for optimal results. Also, the performance gap between words and n-grams narrows when automated feedback is performed, though n-grams still maintain an distinct advantage.

A relatively new, little explored variant of n-gram processing was investigated as a full indexing option. Skipgrams proved effective, but did not result in demonstrable improvements, except in a few cases. It may be that skipgrams would provide gains in Semitic languages, which have root and template morphology, or with documents containing many errors, such as electronic documents obtained through optical character recognition.

Section 5.3 studied the question of why n-grams have a performance advantage over plain words. An experiment to remove morphological regularity from a language was designed, and the results were strongly suggestive that the fundamental reason n-grams are more effective is because they control for morphological variation. This explanation also explains a variety of previously observed phenomena, namely:

- that n-grams yield greater improvements in more morphologically complex languages;

- n-grams of lengths 4 and 5 (about the size of root morphemes) are most effective;
- and, the fact that relevance feedback helps words more than n-grams could be explained by the fact that feedback with words can bring in related word forms (*e.g.*, if the word *golf* appears in a top-ranked documents, then words like *golfer* or *golfing* are reasonable expansion terms).

These results confirm that n-grams are effective indexing terms in monolingual applications and they give insight into the causal factors behind their success. In Chapter 6 the use of n-gram tokenization in bilingual retrieval is explored.

Chapter 6

BILINGUAL EXPERIMENTS

Using the same test collections used in Chapters 4 and 5, this chapter explores the use of alternative methods for tokenization where the aim is to improve retrieval performance when queries and documents are in different languages. Corpus-based translation of n-grams instead of words or stemmed words is shown to produce significant gains in retrieval effectiveness.

6.1 Bilingual Methodology

In this chapter the general approach taken for CLIR is to map source language queries into a form that can be matched against a target language index. The target language index could use any of the tokenization alternatives presented in Chapters 4 and 5; however emphasis will be given to word indexing and the use of character 5-grams. The same methods presented in Chapter 3 for matching queries and documents monolingually, namely the use of a statistical language model similarity metric as implemented in the HAIRCUT retrieval engine, are applied here.

Except in Section 6.5 where the issue is addressed, no use is made of pre-translation query expansion or post-translation automated relevance feedback.

The mapping, or translation, of queries can be performed through several means. For

example, machine translation software, if available, could be used to transform the original query into a more or less grammatical rendering in the language of the document collection. Or word-for-word translation using a bilingual wordlist could be performed. However, this dissertation is concerned with translation methods based on parallel texts.

Parallel collections are texts where documents have translations in another language. Often a translation in only one other language is available; however, multiply-aligned collections exist, especially for governmental documents. For the languages of the CLEF document collections several large multilingual corpora are available. Large collections are valued because with a greater amount of translated text the problems of data sparseness and untranslatable terms are lessened.

To support translation, it is desirable to align portions from the source language half of the parallel data with matching passages from the target side. For statistical machine translation word-level alignments are sought. However aligning larger units such as sentences or paragraphs requires less sophisticated techniques. In some cases coarse alignments can be provided from external information, as is normally the case with religious texts due to verse markings, but in general, alignments are computed automatically using dynamic programming methods that seek an optimal alignment.

6.2 Translation Corpora

Four distinct sources of parallel text are used in this chapter: *bible*, *acquis*, *europarl*, and *ojeu*. In some experiments we also consider a combined corpus, *all*, that combines data from each of the four sources. The sizes and sources of the parallel data are given in Table 6.1. These sources include translations in a number of languages. Here consideration is restricted to only those languages for which a CLEF document collection with relevance judgments exists, namely, Bulgarian (BG), Czech (CS), Dutch (NL), English (EN), Finnish

Name	Size	Wrds/doc	Genre	Source
<i>bible</i>	785k words, 4 MB	25.3	Religious	http://unbound.biola.edu/
<i>acquis</i>	32M words, 202 MB	26.3	EU law (1958 to 2006)	http://wt.jrc.it/lt/acquis/
<i>europarl</i>	33M words, 197 MB	25.5	Parliamentary oration (1996 to 2006)	http://www.statmt.org/europarl/
<i>ojeu</i>	84M words, 593 MB	34.5	Governmental affairs (written)	Developed during this re- search from texts available at http://europea.eu.int/
<i>all</i>	150M words, 995 MB	30.1	Mixed	Composite of bible, acquis, europarl and ojeu

Table 6.1. Parallel texts used in experiments.

(FI), French (FR), German (DE), Hungarian (HU), Italian (IT), Portuguese (PT), Russian (RU), Spanish (ES), and Swedish (SV). A sample passage from each of the four corpora is presented in English, French, and Spanish in Figure 6.1. Each of the corpora is described in greater detail below.

6.2.1 Bible Corpus

The *bible* corpus is based on the 66 books in the Old and New Testaments; it does not include the deuterocanonical books. Multiple biblical translations are available in some of the languages; however only a single translation was selected for these experiments. The selected translations are presented in Table 6.2. As the CLEF documents use contemporary language, preference was given to modern translations, for example, the American Standard Version (1901) was chosen for English instead of the King James Version (1611). Alignments at the verse level are used; there are 31103 verses in the English text.

EN: Therefore was the name of it called Babel; because Jehovah did there confound the language of all the earth: and from thence did Jehovah scatter them abroad upon the face of all the earth.
ES: Por esto fué llamado el nombre de ella Babel, porque allí confundió Jehová el lenguaje de toda la tierra, y desde allí los esparció sobre la faz de toda la tierra.
FR: C'est pourquoi son nom fut appelé Babel (confusion); car l'Éternel y confondit le langage de toute la terre, et de là l'Éternel les dispersa sur toute la face de la terre.

(a) Bible – Genesis 11:9

EN: (24) In order to contribute to the conservation of octopus and in particular to protect the juveniles, it is necessary to establish, in 2006, a minimum size of octopus from the maritime waters under the sovereignty or jurisdiction of third countries and situated in the CEEAF region pending the adoption of a regulation amending Regulation (EC) No 850/98.
ES: (24) A fin de contribuir a la conservación del pulpo, y en particular para proteger a los juveniles, es necesario establecer, en 2006, una talla mínima para el pulpo procedente de las aguas marítimas bajo la soberanía o jurisdicción de terceros países y situadas en la región CPACO a la espera de la adopción de un reglamento que modifique el Reglamento (CE) no 850/98.
FR: (24) Afin de contribuer à la conservation du poulpe et en particulier de protéger les juvéniles, il est nécessaire d'établir, pour 2006, une taille minimale du poulpe des eaux maritimes relevant de la souveraineté ou de la juridiction de pays tiers et situées dans la région de la COPACE jusqu'à l'adoption d'un règlement modifiant le règlement no 850/98.

(b) Acquis – from Council Regulation (EC) No 51/2006, 22 December 2005

EN: Mr President, the tsunami tragedy should be no less significant to the world's leaders and to Europe than 11 September.
ES: Señor Presidente, la tragedia del maremoto no debe ser menos importante para los dirigentes mundiales y para Europa que el 11 de septiembre.
FR: Monsieur le Président, la tragédie du tsunami ne doit pas avoir moins d'importance aux yeux des dirigeants du monde et de l'Europe que celle du 11 septembre.

(c) Europarl – Parliamentary proceedings of 12 January 2005.

EN: 11. Trafficking in women for sexual exploitation. A4-0372/97. Resolution on the Communication from the Commission to the Council and the European Parliament on trafficking in women for the purpose of sexual exploitation (COM(96)0567 - C4-0638/96). The European Parliament,
ES: 11. Trata de mujeres con fines de explotación sexual. A4-0372/97. Resolución sobre la Comunicación de la Comisión al Consejo y al Parlamento Europeo sobre la trata de mujeres con fines de explotación sexual (COM(96)0567 - C4-0638/96). El Parlamento Europeo,
FR: 11. Traite des femmes à des fins d'exploitation sexuelle. A4-0372/97. Résolution sur la communication de la Commission au Conseil et au Parlement européen sur la traite des femmes à des fins d'exploitation sexuelle (COM(96)0567 - C4-0638/96). Le Parlement européen,

(d) Official Journal of the European Communities – C 14/39 16 December 1997

FIG. 6.1. Aligned passages from parallel sources in English, Spanish, and French.

Language	Version	Year of Publication
Czech	Czech Ecumenical Translation	1978
Dutch	Dutch Staten Vertaling	1750
English	American Standard Version	1901
Finnish	Pyhä Raamattu	1992
French	Ostervald (revised)	1996
Italian	Riveduta	1927
German	Schlachter-Bibel (revised)	1951
Portuguese	Almeida Atualizada	unclear
Russian	Synodal Translation	1827
Spanish	Reina Valera (revised)	1909
Swedish	Swedish Church Bible	1917

Table 6.2. Bible versions used.

6.2.2 JRC-Acquis Corpus

The European Commission's Joint Research Centre (JRC) developed the JRC-Acquis (version 3) corpus for use in computational linguistics research (Steinberger *et al.* 2006). This parallel text is based on EU laws comprising the *Acquis Communautaire*. Translations are available in 22 languages, making this one of the larger parallel corpora in existence given both its size and number of supported languages. The English portion of the *acquis* data includes 1.2 million aligned passages containing over 32 million words. This is approximately 40 times larger than the Biblical text.

Alignments are at roughly the sentence level – 85% of the alignments correspond to a single sentence in both source and target language. Alignments are possible between any two language pairs. Two different sets of passage alignments are provided by the JRC. In this research the alignments produced by the *Vanilla* algorithm¹ were used.

¹Vanilla is based on an implementation of Gale and Church's algorithm (1991) and the software is available from <http://nl.ijs.si/telri/vanilla/>.

6.2.3 Europarl Corpus

The Europarl corpus (version 3) was assembled by Philipp Koehn to support experiments in statistical machine translation (Koehn 2005). The documents consist of verbal dialog from the official proceedings of the European Parliament. The data are available in 11 languages: Danish, Dutch, English, Finnish, French, German, Greek, Italian, Portuguese, Spanish, and Swedish. While alignments are possible between any language pair, Koehn has made precomputed alignments between English and the other 10 languages available, and these alignments were used for the experiments reported in this chapter. The alignments are also based on the algorithm by Church and Gale (1991).

The *europarl* corpus is comparable in size to the *acquis* corpus, and contains about 33 million words.

6.2.4 Official Journal of the EU Corpus

The Official Journal of the European Union (OJEU) is published in all of the official languages. The Journal publishes a variety of material, including legislation, directives, informative reports, and judgments from the Court of Justice. The Journal covers a broad range of topics including typical governmental issues such as agriculture, trade, and foreign relations. The translated documents are published online in Adobe's Portable Document Format (PDF).

To support this research the parallel corpus was created by downloading documents dating from January 1998 through April 2004 (just prior to the Enlargement). It was necessary to convert PDF documents to plain text and this was accomplished using the *pdftotext* tool, which can output text in the ISO-8859-1 (Latin-1) encoding. In this manner documents in Dutch, English, Finnish, French, German, Italian, Portuguese, Spanish, and Swedish were obtained. The documents were segmented into pages and into paragraphs

consisting of a small number of sentences (typically 1 to 3); however this process was complicated by the fact that many documents have outline or tabular formatting. The mean number of words per “document” is 34.5, which is notably longer than the other sources. Alignments were produced using Church’s *char_align* software (Church 1993).

Due to complexities of decoding the PDF, some of the accented characters were not extracted properly as can be observed in Figure 6.1 (d). This is mainly a problem with the earlier material in the collection.

In total about 85 million words (in excess of 500 MB) of text per language was obtained. This is over twice the size of either the *acquis* or *europarl* collections.

6.3 Translation

After creating the pairwise-aligned corpora described above, it is necessary to create parallel indexes for each corpus using the tokenization styles of interest. As an example, for the aligned English-Spanish documents in the Europarl data, separate indexes are created using words, stems, and n-grams, in both English and Spanish. With these indexes it is possible to translate query terms between source and target languages. It is also possible to translate across tokenization types to determine, for example, the target language character 5-gram that best corresponds to a source language stemmed word. For the experiments in this dissertation, only translation to terms that were generated using the same tokenization method is considered.

Candidate translations were extracted as follows. First, taking a query term as input, where a term can be a word, n-gram, etc., documents containing the term in the source language subset of the aligned collection are identified. A limit of 10000 documents is used for reasons of efficiency and because in earlier experiments it did not appear that performance was enhanced appreciably when a greater number of documents was considered.

If no document contains this term, then it is left untranslated. Second, the corresponding documents in the target language subset of the aligned corpus are identified. Third, using a statistic such as pointwise mutual information, the candidate translation with the highest score replaces the original source language query term. For these experiments the same term selection metric from Section 5.1 was used. When all query terms have been processed in this fashion the transformed query vector is used for retrieval in the target language collection of interest.

Before discussing experiments using aligned corpora for bilingual retrieval, examples of translation using these corpora are given. Because n-grams are a conflationary technique, on average n-grams have greater ambiguity than words. As a result, often no single correct answer exists. For example, should a translation for an n-gram like *minis* be based on the corresponding n-grams from administration or feminist?

Sample translations using words and n-grams are given in Table 6.3. Diacritical marks were removed during indexing, which is why they do not appear in the table. Several observations can be made. First, a word is sometimes translated as a larger expression in a compounding language. For example, nuclear becomes *ydinvoiman* in Finnish (nuclear power) and *karnvapen* in Swedish (nuclear weapon). Second, the n-grams that correspond to a whole English word (*e.g.*, *clear*) due to the choice of *n*, tend to be translated as the equivalent word in the target language (*e.g.*, *_clar* in Spanish or *_clai* in French). Third, around the boundary between the two source words, target language n-grams are produced that also span words (*e.g.*, in French *ie_nu* from *énergie nucléaire*, or *rnene* from *kernenergie* in German). And the spanning n-grams appropriately model word order changes, such as adjective-noun reversal (*e.g.*, in French and Spanish adjectives follow the nouns they modify).

Subword translation, the direct translation of n-grams, may offer a solution to the key obstacles in dictionary-based translation that were mentioned in Chapter 1. Word nor-

English	DE	ES	FI	FR	IT	NL	SV
nuclear energy	kernenergie energie	nuclear energia	ydinvoiman energia	nucleaire energie	nucleare energia	nucleaire energie	karnvapen energi
_nucl	_kern	uclea	_ydin	nucle	uclea	_kern	_karn
nucle	_kern	uclea	_ydin	nucle	uclea	_kern	_karn
uclea	_kern	uclea	_ydin	uclea	uclea	_kern	_karn
clear	_klar	_clar	selva	_clai	_chia	duide	_tydl
lear_	_klar	_clar	selva	_clai	_chia	duide	_tydl
ear_e	rnene	a_nuc	ydine	ie_nu	leare	kerne	_karn
ar_en	nener	a_nuc	ydine	ie_nu	leare	nener	_karn
r_ene	energ	energ	energ	energ	_ener	energ	energ
_ener	energ	energ	energ	energ	_ener	energ	energ
energ	energ	energ	energ	energ	energ	energ	energ
nergy	ergie	energ	nergi	energ	energ	nergi	nergi
ergy_	ergie	energ	nergi	energ	energ	nergi	nergi

Table 6.3. Sample one-best translations for *nuclear energy*.

malization is not essential since sub-word strings will be compared. Translation of multiword expressions can be approximated by translation of word-spanning n-grams. Out-of-vocabulary words, particularly proper nouns, can be partially translated by common n-gram fragments or left untranslated in close languages. Additionally, since the lexical coverage of translation resources is a critical factor for good CLIR performance, the fact that almost every n-gram has a translation should improve performance. This last point can be put another way: there are few out-of-vocabulary n-grams, at least for $n = 4$ and $n = 5$.

6.4 Experimental Results

In this section we report on CLIR experiments using the aligned parallel corpora described in Section 6.2. Specifically examined are the relative efficacy of different translation resources, the effect of alternative tokenization methods on cross-lingual retrieval effectiveness, and the relationship between corpus-size and CLIR performance.

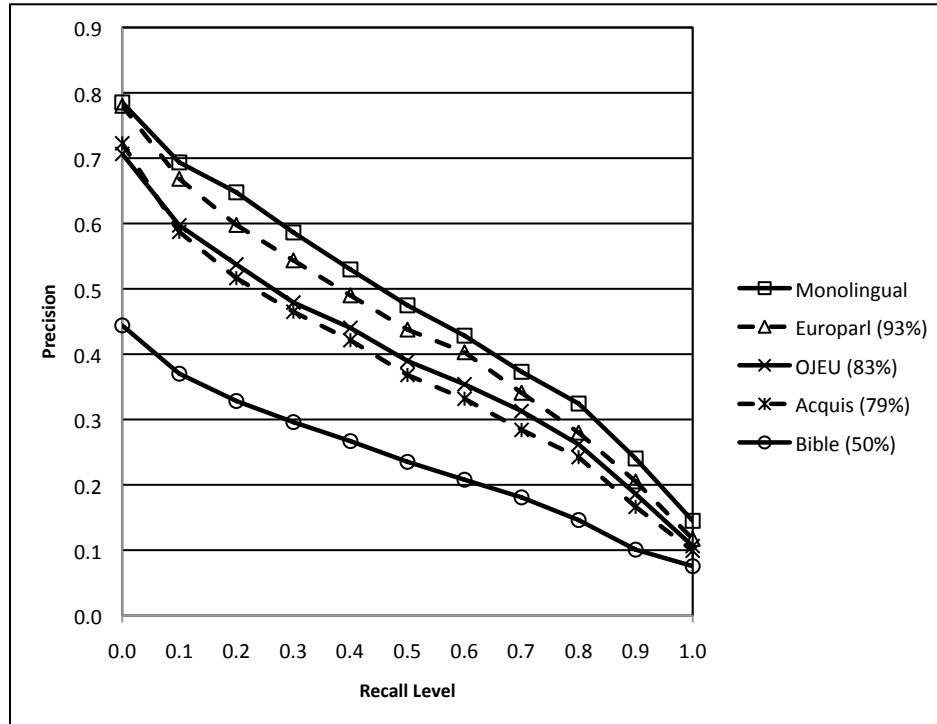


FIG. 6.2. Relative translation corpus effectiveness. English topics and Spanish documents.

6.4.1 Translation Resource Effect on Retrieval Effectiveness

First the relationship between translation source and bilingual retrieval effectiveness is studied. For these retrieval experiments English was used as the sole source language and attention was focused on the eight languages supported by each of the four translation corpora.

In Figure 6.2 we see that the choice of translation corpus can have a substantial effect on bilingual retrieval. The figure shows a precision-recall graph using English topics to search Spanish documents with 5-gram tokenization. Each curve corresponds to use of a different parallel corpus for translation. With the *europarl* corpus performance is obtained that is 93% as good as when human translated queries are used (*i.e.*, monolingual retrieval).

Table 6.4 reports mean average precision when word-based tokenization and translation was performed. For comparison the corresponding performance using topics in the target language (*i.e.*, the monolingual condition) is also given.

There is a conspicuous difference in effectiveness depending on which source of parallel data is used. As expected the smallest bitext, *bible*, performs the worst. Averaged across the eight languages only 39% relative effectiveness is seen compared to monolingual performance. Both *acquis* and *europarl* are roughly 40 time larger in size than *bible* and both do significantly better; however *europarl* is clearly superior and achieves 75% of monolingual effectiveness. Though nearly twice the size, *ojeu* fails to outperform *europarl* and just barely beats *acquis*. There could be many reasons for this. Difficulty in converting the OJEU data from PDF to text, problems making good alignments, and the substantially greater length of aligned passages (cf. Table 6.1), all could have adverse effect.

When combining all sources (*all*), performance is lower than when *europarl* alone is used. And *europarl* outperforms each other source for each for each of the 8 languages, as well as in aggregate. The same relative ordering of {*europarl*, *all*, *ojeu*, *acquis*, *bible*} was also observed when stemmed words and character 5-grams were the methods of tokenization and translation, and this comparison is the subject of the next section.

6.4.2 Comparing Tokenization Methods in CLIR

In this section bilingual retrieval is compared when both indexing and translation is performed using words, stems, or character n-grams. As 5-grams were the most effective method identified in Chapter 4 they, not 4-grams, are used in these experiments.

Before describing these experiments it is worth pointing out that the functions of tokenization and translation can be separated. For example, first query translation could be conducted using words as translation units, and then (word-internal) character n-grams could be generated from the translated words for use in subsequent target-language re-

Target	Mono	<i>bible</i>	<i>acquis</i>	<i>europarl</i>	<i>ojeu</i>	<i>all</i>
DE	0.3303	0.1338	0.1802	0.2427	0.1937	0.2075
ES	0.4396	0.1454	0.2583	0.3509	0.2786	0.2847
FI	0.3406	0.1288	0.1286	0.2135	0.1636	0.1583
FR	0.3638	0.1651	0.2508	0.2942	0.2600	0.2633
IT	0.3749	0.1080	0.2365	0.2913	0.2405	0.2567
NL	0.3813	0.1502	0.2474	0.2974	0.2484	0.2505
PT	0.3162	0.1432	0.2009	0.2365	0.2157	0.2216
SV	0.3387	0.1509	0.2111	0.2447	0.1861	0.2000
Average	0.3607	0.1407	0.2142	0.2714	0.2233	0.2303
		39.0%	59.4%	75.3%	61.9%	63.9%

Table 6.4. Word-based CLIR using English topics and various aligned corpora.

trieval. Preliminary experiments suggested that actually using n-grams for translation was the more effective approach and that is the focus of this section.

Experiments were conducted for three scenarios, for language pairs where: (1) English was the source language; (2) where English was the target language; and, (3) for a few pairs where English was neither the source or target language. Additional language pairs should be studied, however this is left for future work.

Experiments with English as the Source Language

Of the available corpora, *acquis* has the broadest coverage for the CLEF target collections, supporting 11 of the 12 bilingual pairs possible with English as the source language. Only Russian, a non-EU language, is missing. Table 6.5 reports mean average precision using English topics that were translated with the *acquis* corpus. Integrated tokenization and translation using words, Snowball stems, and character 5-grams are compared using the same methods in Chapters 4 and 5. Statistical testing was performed using the paired *t*-test, and triangles indicate statistically significant improvement over words with $p < 0.01^{\Delta}$ (or with $p < 0.05^{\Delta}$).

Sizeable improvements are seen when stems and n-grams are used not only for tok-

Target	Words	Snowball	5-grams
BG	0.0591		0.0898 [▲] (+51.9%)
CS	0.1107		0.2479 [▲] (+123.9%)
DE	0.1802	0.2097 [△] (+16.4%)	0.2952 [▲] (+63.8%)
ES	0.2583	0.3072 [▲] (+18.9%)	0.3661 [▲] (+41.7%)
FI	0.1286	0.1755 [△] (+36.5%)	0.3552 [▲] (+176.2%)
FR	0.2508	0.2733 [△] (+9.0%)	0.3013 [▲] (+20.1%)
HU	0.1087		0.2224 [▲] (+104.6%)
IT	0.2365	0.2656 [△] (+12.3%)	0.2920 [▲] (+23.5%)
NL	0.2474	0.2249 (-9.1%)	0.3060 [▲] (+23.7%)
PT	0.2009		0.2544 [▲] (+26.6%)
SV	0.2111	0.2270 (+7.5%)	0.3016 [▲] (+42.9%)
Average	0.1811		0.2756 (+63.5%)
Average-7	0.2161	0.2405 (+13.1%)	0.3168 (+56.0%)

Table 6.5. Tokenization effects with English topics and *acquis* corpus.

enization of queries and documents, but also as the unit of translation. In Sections 4.2 and 4.3 it was shown that stemming and n-gram tokenization resulted in 11% and 21% relative improvements versus words, respectively, in monolingual retrieval. Here stems obtain a comparable 13% improvement over words, but with n-grams, much larger changes are observed: on average 5-grams are 63% better. Both methods of controlling for morphology experience larger improvements in the more morphologically complex languages (*e.g.*, Finnish, German, and Hungarian). Not shown in Table 6.5 is whether or not the greater performance of 5-grams over stems is significant or not. The gains in Dutch, German, Finnish, Spanish, and Swedish were significant with $p < 0.01$.

Earlier *europarl* was shown to be the most effective translation resource for bilingual retrieval, so we next compare words, stems, and n-grams when *europarl* is the parallel corpus used. *Europarl* supports a smaller set of languages, only the EU-languages from before the 2004 Enlargement. Table 6.6 reports mean average precision when English topics are translated using the *europarl* corpus. As with *acquis* large changes are observed due to the tokenization method selected, though the relative gains are smaller with the better transla-

Language	Words	Snowball	5-grams
DE	0.2427	0.2646 (+9.0%)	0.3519 [▲] (+45.0%)
ES	0.3509	0.3721 (+6.0%)	0.4294 [▲] (+22.4%)
FI	0.2135	0.2488 (+16.5%)	0.3744 [▲] (+75.4%)
FR	0.2942	0.3233 [▲] (+9.9%)	0.3523 [▲] (+19.7%)
IT	0.2913	0.3132 (+7.5%)	0.3395 [▲] (+16.5%)
NL	0.2974	0.2897 (-2.6%)	0.3603 [▲] (+21.1%)
PT	0.2365		0.2931 [▲] (+23.9%)
SV	0.2447	0.2534 (+3.6%)	0.3203 [▲] (+30.9%)
Average	0.2714		0.3527 (+31.9%)
Average-7	0.2764	0.2950 (+7.1%)	0.3612 (+33.0%)

Table 6.6. Tokenization effects with English topics and *europarl* corpus.

tion resource. Stems gain 7% over words; 5-grams gain 32%. It is difficult to make comparisons between the two corpora using aggregate performance across languages because the set of languages varies between the two translation sources. On the seven common languages where stems were used, 5-grams averaged 0.3168 (+56% relative improvement over words) with *acquis* and 0.3612 (+33% vs. words) with *europarl*.

When using *europarl* for translation n-grams achieve higher performance but see a decrease in their relative advantage vis-à-vis words. Since the gains with n-grams are magnified when lower quality translation resources are used, this suggests that when only poor resources are available they should be given strong consideration.

Experiments with English as the Target Language

When English is the language of the document collection and queries are expressed in other languages, the number of usable topics depends on the number of judged English topics that have been manually translated and which contain at least one relevant document. The number of available topics by source language is given in Table 6.7.

Mean average precision is reported in Tables 6.8 and 6.9 for experiments using multilingual topics to search English documents; translation was accomplished using the *acquis*,

	BG	CS	DE	ES	FI	FR	HU	IT	NL	PT	RU	SV
# Topics	124	50	268	268	218	317	124	268	218	183	77	218

Table 6.7. Suitable topics for bilingual experiments on English documents.

Source	Words	Snowball	5-grams
BG	0.0747		0.1258 [▲] (+68.4%)
CS	0.1568		0.2007 [△] (+28.0%)
DE	0.2339	0.2457 (+5.0%)	0.2800 [△] (+19.7%)
ES	0.2779	0.2852 (+2.6%)	0.2959 (+6.5%)
FI	0.2277	0.2428 (+6.6%)	0.2752 (+20.9%)
FR	0.2937	0.2970 (+1.1%)	0.2957 (+0.7%)
HU	0.1479		0.2148 [▲] (+45.2%)
IT	0.2742	0.2897 (+5.7%)	0.3057 [△] (+11.5%)
NL	0.2541	0.2521 (-0.8%)	0.3226 [▲] (+27.0%)
PT	0.2323		0.2644 [△] (+13.8%)
SV	0.2701	0.2611 (-3.3%)	0.3067 (+13.6%)
Average	0.2221		0.2625 (+23.2%)
Average-7	0.2617	0.2677 (+2.4%)	0.2974 (+14.2%)

Table 6.8. Tokenization using *acquis* corpus bilingual retrieval on English documents.

Source	Words	Snowball	5-grams
DE	0.2984	0.3288 [▲] (+10.2%)	0.3318 (+11.2%)
ES	0.3367	0.3474 (+3.2%)	0.3435 (+2.0%)
FI	0.3084	0.3124 (+1.3%)	0.3414 (+10.7%)
FR	0.3181	0.3303 (+3.8%)	0.3318 (+4.3%)
IT	0.3214	0.3258 (+1.4%)	0.3650 [▲] (+13.6%)
NL	0.2855	0.2734 (-4.2%)	0.3634 [▲] (+27.3%)
PT	0.2670		0.2926 [△] (+9.6%)
SV	0.3119	0.3189 (+2.2%)	0.3666 [△] (+17.5%)
Average	0.3059		0.3420 (+12.0%)
Average-7	0.3115	0.3196 (+2.6%)	0.3491 (+12.4%)

Table 6.9. Tokenization using *europarl* corpus bilingual retrieval on English documents.

and *europarl* data, respectively. As before, words, stems, and 5-grams are compared. In comparison to the results in Table 6.5, changes due to choice of tokenization are smaller. For example, when English queries were used to search Finnish documents 5-grams resulted in a 176% relative improvement (see Table 6.5), but in Table 6.8 we see only a 21% gain when Finnish queries were used to search English documents. These data may signal that the function of morphological normalization is more critical for the document language in bilingual retrieval than it is for the language of the query.

Experiments with Other Language Pairs

The CLEF 2004 evaluation encouraged study of bilingual retrieval where English was not used as either the source or target language. For that evaluation I conducted experiments using the same essential approach used in this research. In Figure 6.3 the relative effectiveness of corpus-based, one-best translation of words, stemmed words, and character 5-grams is compared. Three target languages are used, Finnish, French, and Portuguese with two different source languages each. 5-grams outperform words in each case and they outperform stems in 5 of 6 pairs. As before, the effects are most striking when a morphologically complex target language is involved; with Finnish documents 50% gains are attained.

6.4.3 Size of Parallel Text

Intuitively CLIR performance should improve with the quality of bilingual translations. And presumably translation quality improves with larger amounts of parallel text. However, it may be that n-grams are still effective when only limited amounts of parallel text are available. One reason why this might be so is because for a fixed size of parallel text, more instances of a morphological root, like *walk*, will be observed than will instances of words such as *walked* or *walking*. Additional observations should provide more evidence

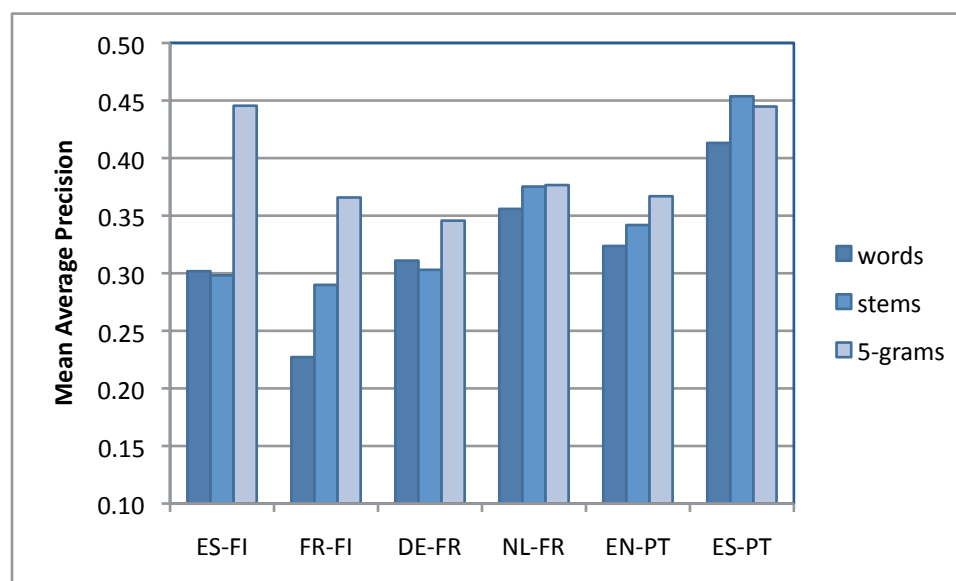


FIG. 6.3. Tokenization experiments using two query languages in three target languages.

and therefore make translation easier given limited training data.

To investigate how translation corpus size effects bilingual retrieval I subsampled the *europarl* corpus and used these smaller subcorpora for translation. The entire corpus is 33 million words in size, and samples of 1%, 2%, 5%, 10%, 20%, 40%, 60%, and 80% were made based on counting documents, which for *europarl* is equivalent to counting sentences. Samples were taken by processing the data in chronological order. A random sample of sentences taken without respect to time might result in greater lexical diversity because data in temporal order will contain repeated references to topics and entities; however, such a body of text would not be obtained through a natural collection effort, and the goal here is to model resource size.

In Figure 6.4 (a-h) the effect of using larger parallel corpora is plotted. English is the source language and each of the eight *europarl* languages is used as the target language. Mean average precision is on the vertical axes and for visual effect, the chart for each

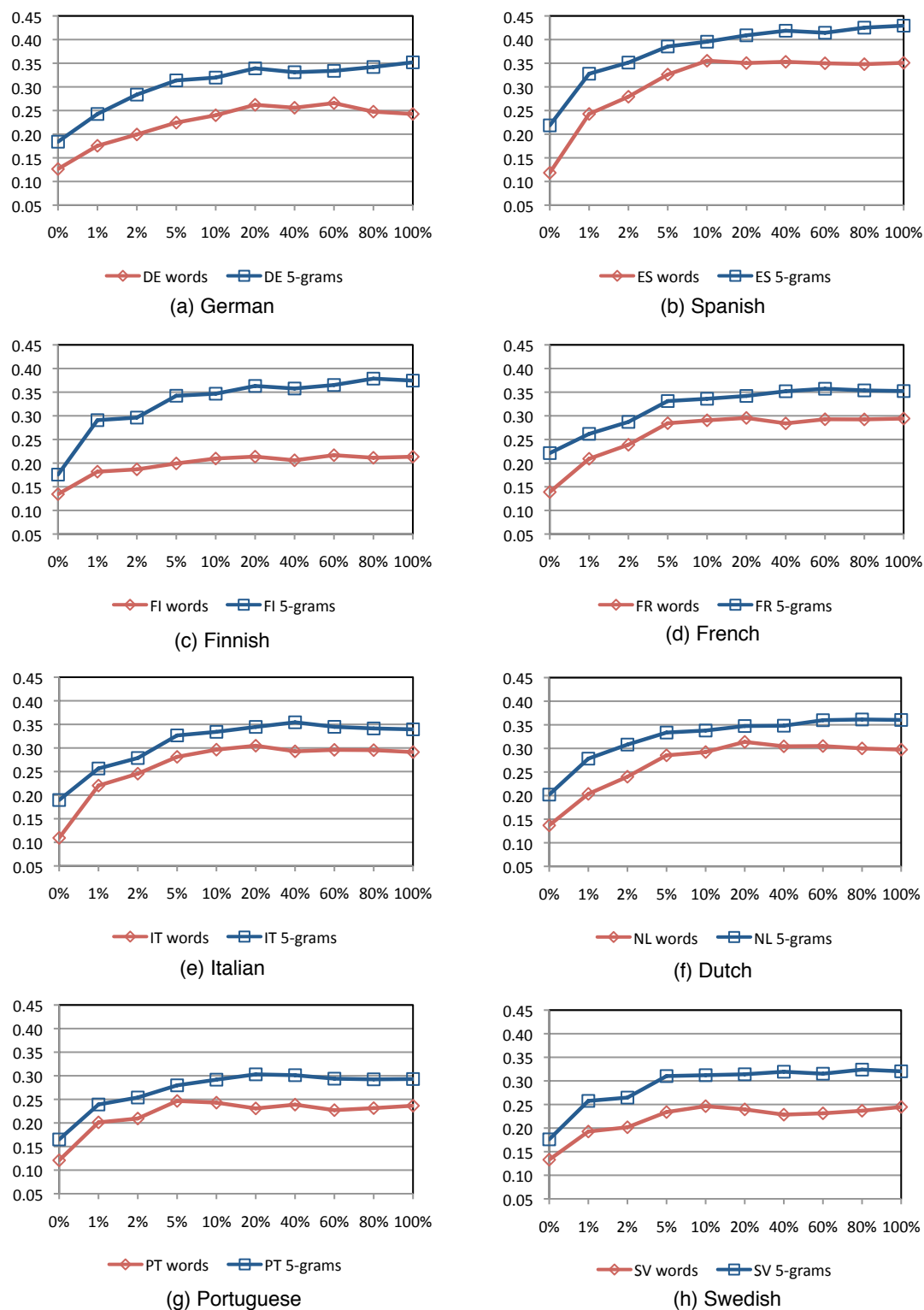


FIG. 6.4. Performance improvement with corpus growth.

	1%	2%	5%	10%	20%	40%	60%	80%
Unique	21k	31k	59k	70k	101k	145k	188k	223k
Total	317k	637k	1.58M	3.19M	6.33M	12.7M	19.0M	25.4M

Table 6.10. Size of English *europarl* subsets in words.

language pair uses the same scale. The horizontal axes are not to scale. The general shape of the curves is to rise quickly as increasing subsets from 1% to 10% are used and to flatten as size increases further. The deceleration of growth with increasing corpus size can be explained by Heap's Law², which characterizes the decrease in incremental vocabulary size as additional text is observed. As the number of topics in these tests sets³ is small, once the topical vocabulary is covered well enough to produce accurate translations, further gains would not be expected. Similar results have been obtained in the few studies that have sought to quantify bilingual retrieval performance as a function of translation resource size (Xu & Weischedel 2000). In the higher complexity languages such as Dutch, German, and Finnish, n-grams appear to be gaining a slight improvement even when the entire corpus is used; this makes sense as the vocabulary size is greater in those languages.

The data for the 0% condition were based on cognate matches for words and 'cognate n-grams' that require no translation. The figure reveals that even a small amount of parallel text quickly improves performance. Sizes for the subsets are given in Table 6.10. The 2% condition is roughly the size of *bible*, which performs poorly, though much of this may be due to genre and topic. For example, the Biblical text does not contain the words *nuclear* or *energy* and thus is greatly disadvantaged for a topic about nuclear power. In each of the language pairs 5-grams dominate words. In fact, in each language words using any of the *europarl* subsets never outperform 5-grams using only 5% of the available data.

² $V = kN^\beta$ where V is vocabulary size and N is a corpus size. Typically, k lies between 10 and 100 and β is between 0.4 and 0.6 (Baeza-Yates & Ribeiro-Neto 1999).

³Variable by language, but less than 400.

6.5 Pre-Translation Query Expansion

Translation of query terms is a noisy and imperfect process. When queries are short or translation resources are poor, then failure to properly translate a query term can lead to abysmal results on a particular topic. To increase the robustness of bilingual retrieval performance, expansion of query vectors prior to translation has been proposed and found to be effective (Ballesteros & Croft 1997). A typical method for expanding query vectors is to first perform a retrieval against a source language document collection, and then extract terms from top-ranked documents. In this section an experiment is conducted to ascertain whether pre-translation query expansion is effective in the scenarios examined in the previous section, namely corpus-based translation of n-grams.

The *acquis* and *europarl* corpora are used for translation from English queries to eight target languages: Dutch, German, Finnish, French, Italian, Portuguese, Spanish, and Swedish. The English document collection was used for an initial retrieval pass and either 50 words or stems, or 200 5-grams were extracted from top-ranked documents. The number of terms extracted was chosen based on the results from Section 5.1. These terms were then translated using the parallel corpus, and finally retrieval was performed against the target language document collection. There are several places where the choice of tokenization can be made: initial search, pre-translation expansion, term translation, and in target language retrieval. Thus it is possible to use a single approach to tokenization throughout, or to mix-and-match multiple representations. The process is illustrated in Figure 6.5

The figure shows how an English query could be used to search Spanish documents. On the left side of the figure English documents are searched for pre-translation query term expansion; this phase is absent in normal bilingual retrieval. On the top side of the figure word-based processing is exemplified and below, the corresponding n-gram method is depicted. There are several points where it is possible to switch tokenization. For example,

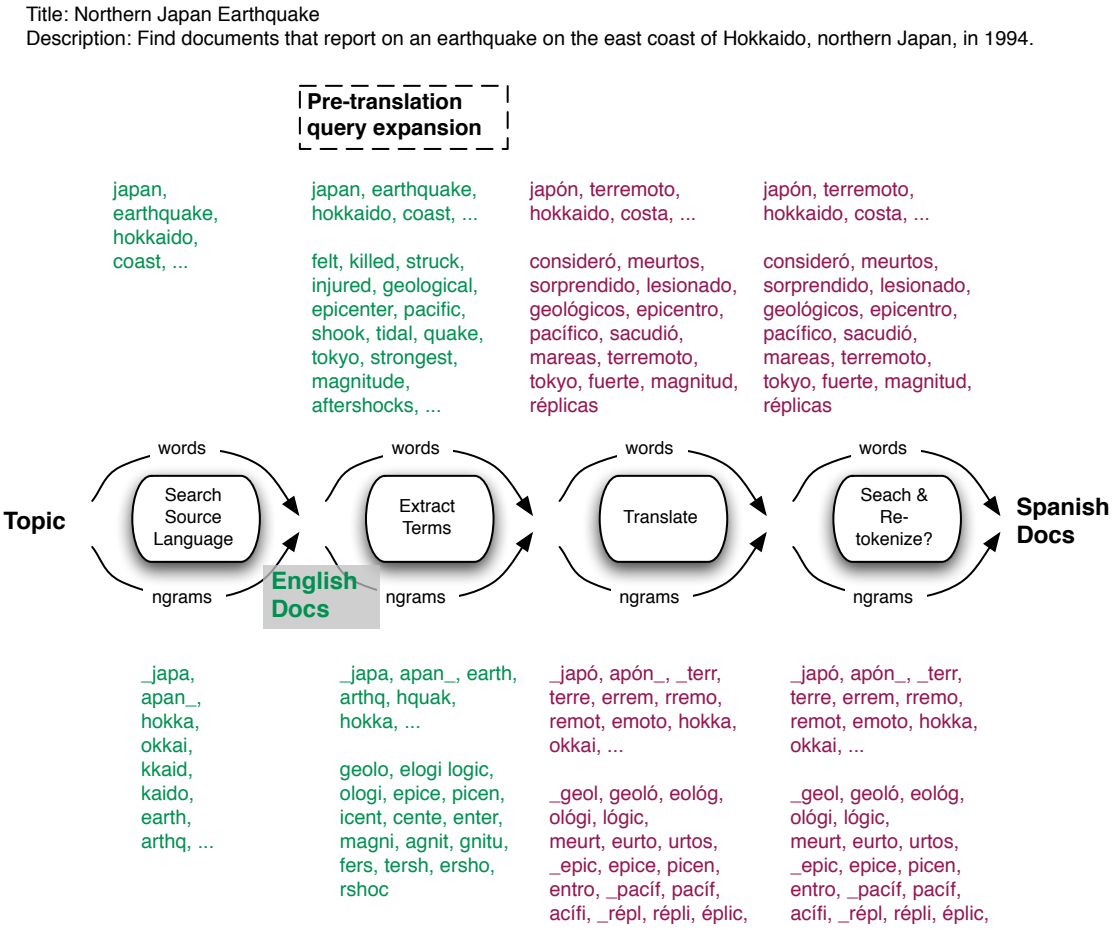
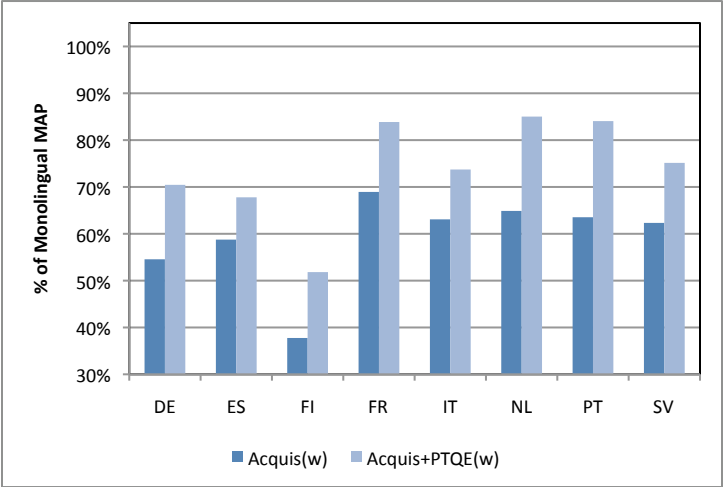


FIG. 6.5. Example of pre-translation query expansion in corpus-based bilingual retrieval.

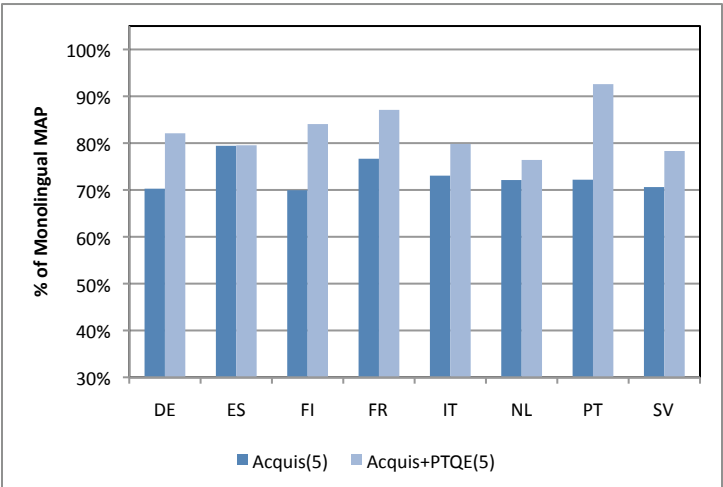
words could be extracted in the query expansion phase and those could be used to produce n-grams, which could then be translated.

In our experiments pre-translation query expansion is found to be quite useful. Effectiveness increases both for words or 5-grams, using either *acquis* or *europarl*. Figures 6.6 (using *acquis*) and 6.7 (*europarl*) chart performance relative to a monolingual baseline for (a) words and (b) 5-grams. Table 6.11 reports MAP averaged across seven languages⁴ for

⁴Stemming wasn't available in Portuguese

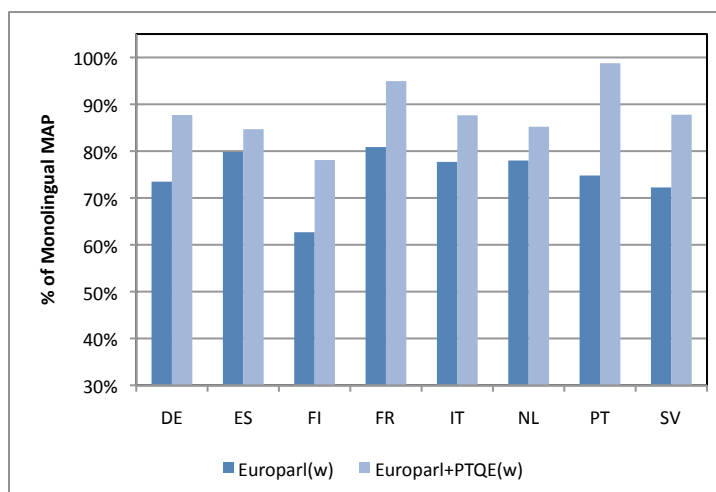


(a) English topics, Acquis, Words

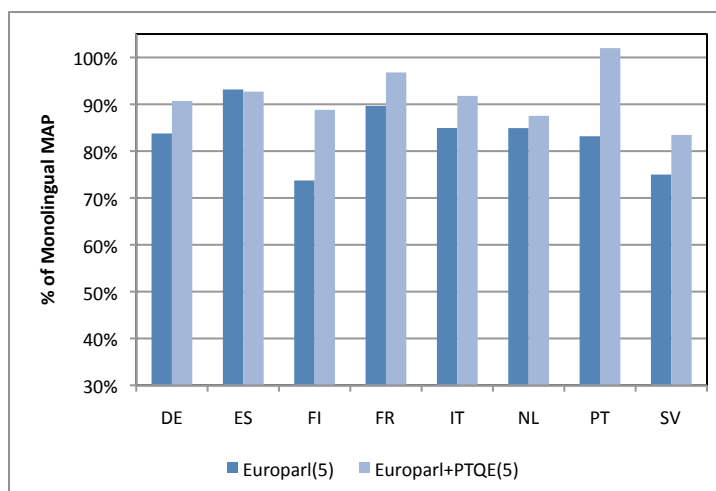


(b) English topics, Acquis, 5-grams

FIG. 6.6. Improvement in bilingual retrieval using pre-translation expansion and *acquis*.



(a) English topics, Europarl, Words



(b) English topics, Europarl, 5-grams

FIG. 6.7. Improvement in bilingual retrieval using pre-translation expansion and *europarl*.

		Mono 5-grams	Standard Bilingual			Pre-Translation Expansion		
			words	stems	5-grams	words	stems	5-grams
acquis	PMAP	0.4333	0.2161	0.2405	0.3168	0.2668	0.2889	0.3512
	%mono		49.9%	55.5%	73.1%	61.6%	66.7%	81.1%
	%change					+23.4%	+20.1%	+10.9%
europarl	PMAP	0.4333	0.2764	0.2950	0.3612	0.3177	0.3397	0.3906
	%mono		63.8%	68.1%	83.4%	73.3%	78.4%	90.2%
	%change					+15.0%	+15.2%	+8.2%

Table 6.11. Efficacy of pre-translation query expansion.

words, stems, and 5-grams.

Application of the technique resulted in an 8% to 23% improvement over a comparable bilingual run based on the methods of Section 6.4. The relative improvement was largest when the weaker translation resource (*i.e.*, *acquis*) was used, or when an inferior tokenization method was used (*e.g.*, words). The fact that gains are diminished when higher quality translation corpora or n-grams are used suggests that the benefit of the technique is chiefly due to reducing the negative effects of out-of-vocabulary terms.

6.6 Synopsis

This chapter explored the subject of corpus-based bilingual retrieval and it examined a number of issues, including: the utility of subword translation; the relative bilingual effectiveness of alternative tokenization methods; the importance of parallel corpus selection and size; and, the use of pre- and post-translation query expansion. Key results include:

- Size is not the most important factor in corpus-based bilingual retrieval; the quality of alignments and genre are crucial.
- 5-grams outperform stems and words in bilingual retrieval when these methods of tokenization are used both for translation and document indexing. Large relative

improvements were observed with 5-grams compared to words. Gains over 50% bilingually were obtained, which is much larger than the 21% gain observed monolingually in Chapter 4. The relative advantage of n-grams depends significantly on the morphological complexity of the target language and on the caliber of the translation corpus.

- When only limited parallel data is available for translation, n-grams are markedly more effective than words. Using a subsample of only 5% of available data from the highest performing translation resource, *europarl*, 5-grams outperformed plain words using any amount of parallel data.
- Mitigating translation losses through query expansion is effective when sub-word translation is used; moderate gains of 8% to 11% were observed when pre-translation expansion was undertaken with n-gram tokenization.

Chapter 7

CONCLUSION AND FUTURE WORK

This chapter summarizes the key contributions of this dissertation, mentions limitations of the experiments conducted, and outlines directions for future work.

7.1 Review

This dissertation set out to explore language-independent methods for effective monolingual and bilingual text retrieval. Four claims were put forth:

1. Effective multilingual text retrieval can be achieved without the costs and complexities introduced by language-specific processing.
2. Indexing using character n-grams is effective because n-grams provide lexical normalization, and the benefit of n-gram indexing is greatest in languages with high morphological complexity.
3. In cross-language information retrieval, translation need not be performed at the word level.
4. In corpus-based bilingual retrieval the relative advantage from using character n-grams as both indexing terms and units of translation is inversely proportional to resource size and quality.

To establish these whether or not these claims are valid experiments were designed and conducted to evaluate information retrieval performance using test sets in 13 European languages.

In Chapter 4 the focus was on approaches to monolingual retrieval and the central method examined was tokenization using overlapping character n -grams. Comparisons were made to plain words, rule-based stemming, statistically derived morphemes, and a select n -gram for each word (*i.e.*, n -gram stems). Experimental results showed that n -grams were the most effective monolingual technique and over the suite of test sets an average 21% improvement in mean average precision was obtained compared to plain words. A relative gain of 11% was achieved against a popular rule-based stemming algorithm. N -grams lengths of $n = 4$ and $n = 5$ were found to be optimal and equivalent in performance. N -gram tokenization is a language-independent method that does not require language-specific customization or parameter training to be effective. Collectively these findings provide ample evidence for Claim 1.

In Chapter 4 it was also observed that morphological complexity of the language is a key factor to consider when choosing a method of tokenization, and a strong correlation was shown between several metrics of morphological complexity and the improvement possible with n -gram indexing. In Romance languages n -grams revealed little or no monolingual advantage compared to stemmed words, but in languages such as Finnish and Hungarian, dramatic improvements were achieved. In Section 5.3 an experiment was conducted that established a causal link between the morphological complexity of a language and n -gram effectiveness. The experiment degraded the morphological variation in languages and found that word-based methods remained unaffected while n -gram techniques lost their advantage entirely. These results confirm Claim 2.

Chapter 6 addressed the major goal of this dissertation, namely determining whether alternative tokenization methods and corpus-based translation could improve bilingual re-

trieval performance. Empirical results support this and validate the third hypothesis. The gains observed with n-gram tokenization were amplified when n-grams were used both as indexing terms and the unit of translation (Section 6.4). Average gains over 50% were obtained when subword translation was compared to words. These experimental results, and the work using pre-translation query expansion in Section 6.5, show the advantages of integrating translation and tokenization.

Other experiments were conducted that revealed the effects of parallel corpus selection and size on efficacy. When translation resource quality is low or corpus-size is small, the advantage of n-gram processing is magnified (as Claim 4 asserted). It was also observed that the morphological complexity of the target language had a more pronounced effect on retrieval effectiveness than did the complexity of the topic language in bilingual retrieval (Section 6.4).

Several areas were also explored that did not directly speak to these specific research questions.

In Chapter 5 several additional facets of n-gram tokenization were explored. The use of automated relevance feedback was studied and it was shown that a greater number of expansion n-grams were required to achieve maximal performance compared to when words are used as indexing terms. Also, the relative benefit of n-gram indexing diminishes somewhat when relevance feedback is was applied.

A novel form of n-gram indexing based on n-grams with skipped letters was introduced in Section 5.2, but substantial gains were not obtained. Skipgrams achieved 3% to 4% improvements in a few languages such as Bulgarian and Hungarian, but their potential might be greatest for languages with root and template morphology or for collections with high letter error rate, such as scanned image documents.

7.2 Impact

Stemming algorithms have been studied for over four decades. It is nearly an article of faith that they are important for retrieval; however the gains are relatively small in morphologically simpler languages, including English. This research examined several methods of word normalization and will contribute to a better understanding of the importance of transforming surface forms into more effective indexing units, and of the relationship between linguistic typology and the benefits of lexical regularization.

The use of n-grams has not been widely adopted in alphabetic languages, partly because of the negligible benefit in English, partly because of reaction to earlier reports (Damashek 1995), and also because of concerns about efficiency. This research extends the dialog about n-gram effectiveness through extensive experimentation. For languages with fewer available linguistic resources or with complex morphology the advantages of character n-grams are compelling. The experiments on n-gram stemming demonstrate that aggressive pruning can resolve efficiency concerns and still provide beneficial lemmatization; however a spectrum of possibilities remain to be explored.

Kraaij (2004) and Talvensaari (2007) have pointed out the advantages of parallel corpora over wordlists in CLIR and this work has gone a bit further in showing how much utility there is even in small corpora of only a few hundred thousand words.

Businesses that serve multilingual clientele and organizations concerned with organizing the world's information, irrespective of document language, will benefit from these insights.

In terms of the number of languages and language pairs studied, this dissertation may be the most extensive investigation of the impact of tokenization on retrieval effectiveness. Over 11,000 base runs (*i.e.*, experimental conditions) were produced. While the focus here was on ad hoc retrieval and experiments were based on newswire, these techniques should

be suitable for other applications including text classification, targeted advertising, web search, and scenarios where document text is generated from other media (*e.g.*, automatically recognized speech or scanned documents).

7.3 Limitations

Pragmatic considerations constrained the experimental work that could be undertaken. Existing IR test sets were used and these experiments were based exclusively on European languages using CLEF tests sets. In other work I have undertaken experiments in Arabic, Bengali, Chinese, Farsi, Hindi, Korean, Marathi, and Japanese, and consistently found n-gram tokenization to be a sensible choice. As has often been done at CLEF, only title and description topics were analyzed, but there is no reason to suspect that results would be significantly different with shorter, or longer topic statements.

Analysis was focused on mean average precision as the measure of effectiveness. This choice is not controversial, but it is worth pointing out that metrics such as precision at 10 documents or geometric mean average precision also could be examined. The large number of languages and experimental conditions made concentrating on quantitative measures of performance a logical choice; however, if time had permitted, analysis of qualitative factors that influence retrieval effectiveness would have been desirable.

Algorithms for bilingual retrieval are more complex than their monolingual counterparts. The experiments in Chapter 6 were based on 1-best term translation, translation of queries, not documents, and corpus-based translation. While this certainly restricted the landscape we believe that the conclusions reached will remain applicable as other avenues are pursued.

7.4 Major Findings

The principal results of this research include:

- Character n-gram tokenization is a language-neutral technique that is effective in both monolingual and bilingual settings. Good choices of n-gram length in European languages are $n = 4$ and $n = 5$.
- Proper tokenization is important for languages with greater morphological complexity. The principal reason why n-grams are effective is because they more robustly index root morphemes.
- The quality of alignments and textual genre are important factors in corpus-based CLIR, which are at least as significant as corpus size.
- Conventional techniques such as automated relevance feedback and pre-translation query expansion are effective with n-gram tokenization.
- When good textual representations are used for both document tokenization and term translation, highly accurate bilingual retrieval performance can be attained. 5-grams are remarkably effective, and under certain conditions can achieve relative improvements of over 50% compared to the use of word-based processing.

7.5 Future Directions

One minor direction to pursue would be to extend these results to a greater number of languages, particularly languages written in different scripts, or with rich morphological constructs (*e.g.*, vowel harmony, or reduplication), or languages with few linguistic or translation resources.

A more significant line of inquiry would be to examine the use of skip n-grams for applications such as highly complex languages or on scanned image documents. However, this textual representation showed only limited promise on low-error text such as newswire, and it is not clear that the additional robustness of the technique will warrant the additional computational expense that is required.

Recently techniques have been developed that exploit bidirectional evidence in parallel corpora to improve lexical translation probabilities in statistical machine translation and CLIR (Wang & Oard 2006). Though designed for words, this method combined with the use of weighted k-best translation may result in improvements in n-gram-based bilingual retrieval.

The most challenging area to explore, but one with the greatest potential impact would be phrase-enhanced bilingual retrieval. To date only limited gains have been shown with automated methods, and it is possible that as with monolingual retrieval, no significant improvements can be realized. However, a clear understanding of why automated phrase-based methods are unsuccessful in improving on single-term representations has not yet been put forward.

Appendix A

CLEF BENCHMARKS

As part of my experimental design I made several choices that made it more difficult to compare the results reported in Chapters 4-6 with published results from past CLEF evaluations. I combined the topic sets from multiple years to increase the sensitivity of the experiments. I also refrained from engineering practices such as combining evidence from multiple retrieval runs that typically lead to boosts in performance results (McCabe *et al.* 2001). Here I report several results for the CLEF 2002 and 2005 test sets using retrieval runs from this dissertation alongside with results from the official workshop proceedings.

A.1 CLEF 2002

IR Test sets in eight languages were developed during the CLEF 2002 campaign: Dutch, English, German, Finnish, French, Italian, Spanish, and Swedish; however, English was not an official monolingual language. Braschler (2002) describes the evaluation and discusses its results.

Table A.1 lists mean average precision for a 5-gram run using relevance feedback, as described in Chapter 5, and for the top-ranked run reported in the workshop proceedings. Similarly, Table A.2 provides a comparison for bilingual runs. The 5-gram bilingual runs used the *europarl* corpus with pre-translation query expansion as described in Chapter 6.

	5-grams	Top@CLEF	Team
Dutch	0.4979 (-1.0%)	0.5028	JHU/APL
Finnish	0.4581 (+12.0%)	0.4090	U Neuchâtel
French	0.4603 (-11.3%)	0.5191	Berkeley
German	0.4713 (-10.0%)	0.5234	Berkeley
Italian	0.4482 (-11.9%)	0.5088	F. U. Bordoni
Spanish	0.5441 (0.0%)	0.5441	U Neuchâtel
Swedish	0.4244 (-1.7%)	0.4317	JHU/APL

Table A.1. Comparison with CLEF 2002 monolingual results.

	5-grams	Top@CLEF	Team
Dutch	0.4033 (+14.7%)	0.3516	JHU/APL
Finnish	0.3907 (+93.8%)	0.2016	U Tampere
French	0.4039 (-18.2%)	0.4935	U Neuchâtel
German	0.4034 (-15.2%)	0.4759	Berkeley
Italian	0.3455 (-15.5%)	0.4090	Berkeley
Spanish	0.4538 (-5.2%)	0.4786	U Neuchâtel
Swedish	0.3318 (+10.5%)	0.3003	JHU/APL

Table A.2. Comparison with CLEF 2002 bilingual results.

The techniques advocated in this dissertation have been competitive and resulted in consistently high rankings of my JHU/APL submissions at previous CLEF evaluations. Those submissions were based on the HAIRCUT retrieval engine described in Chapter 3 using character n-gram tokenization.

From the tables it can be seen that in several cases my dissertation runs or earlier JHU/APL runs either were or improve upon the official top-ranked submissions. When 5-grams lag behind the top result the decrease is generally in the range of 10% to 15%.

A.2 CLEF 2005

The CLEF 2005 workshop tested ad hoc retrieval for a smaller set of languages: Bulgarian, English, French, Hungarian, and Portuguese. Di Nunzio (2005) *et al.* analyze

	5-grams	Top@CLEF	Team
Bulgarian	0.3130 (-2.3%)	0.3203	JHU/APL
French	0.4105 (-2.6%)	0.4214	JHU/APL
Hungarian	0.4130 (+0.4%)	0.4112	JHU/APL
Portuguese	0.3898 (+0.6%)	0.3875	U Neuchâtel

Table A.3. Comparison with CLEF 2005 monolingual results.

the results of the ad hoc track in detail. Table A.3 gives monolingual performance using mean average precision. The JHU/APL runs had the highest official score in three of four cases. Two of the four cases (Hungarian and Portuguese) were improved upon using runs produced for the experiments in Chapter 4.

These remarks are only intended to offer an informal, qualitative comparison.

REFERENCES

- [1] Adkins, L. 2004. *Empires of the Plain: Henry Rawlinson and the Lost Languages of Babylon*. Thomas Dunne.
- [2] Adriani, M., and van Rijsbergen, C. J. 2000. Phrase identification in cross-language information retrieval. In *RIAO*, 520–528.
- [3] Andrews, C. 1981. *The Rosetta Stone*. British Museum Press.
- [4] Baeza-Yates, R. A., and Ribeiro-Neto, B. 1999. *Modern Information Retrieval*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.
- [5] Bahle, D.; Williams, H. E.; and Zobel, J. 2002. Efficient phrase querying with an auxiliary index. In *SIGIR*, 215–221. ACM.
- [6] Ballesteros, L., and Croft, W. B. 1997. Phrasal translation and query expansion techniques for cross-language information retrieval. In *SIGIR*, 84–91. ACM.
- [7] Ballesteros, L., and Croft, W. B. 1998. Resolving ambiguity for cross-language retrieval. In *SIGIR*, 64–71. ACM.
- [8] Braschler, M. 2002. CLEF 2002 - Overview of Results. In Peters, C.; Braschler, M.; Gonzalo, J.; and Kluck, M., eds., *CLEF*, volume 2785 of *Lecture Notes in Computer Science*, 9–27. Springer.
- [9] Brown, P. F.; Pietra, S. A. D.; Pietra, V. J. D.; and Mercer, R. L. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics* 19(2):263–311.

- [10] Buckley, C., and Voorhees, E. M. 2004. Retrieval evaluation with incomplete information. In *SIGIR '04: Proceedings of the 27th annual international conference on Research and development in information retrieval*, 25–32.
- [11] Buckley, C.; Mitra, M.; Walz, J. A.; and Cardie, C. 2000. Using clustering and superconcepts within SMART: TREC 6. *Information Processing and Management* 36(1):109–131.
- [12] Carmel, D.; Cohen, D.; Fagin, R.; Farchi, E.; Herscovici, M.; Maarek, Y. S.; and Soffer, A. 2001. Static index pruning for information retrieval systems. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, 43–50. New York, NY, USA: ACM.
- [13] Cavnar, W., and Trenkle, J. 1994. N-gram-based text categorization. In *SDAIR94*, 161–169.
- [14] Chen, A.; He, J.; Xu, L.; Gey, C., F.; and Meggs, J. 1997. Chinese text retrieval without using a dictionary. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Chinese Language Retrieval*, 42–49.
- [15] Church, K. W., and Gale, W. A. 1991. Probability scoring for spelling correction. *Statistics and Computing* 1(2):93–103.
- [16] Church, K. W. 1993. Char_align: A program for aligning parallel texts at the character level. In *ACL*, 1–8.
- [17] Cleverdon, C. W. 1967. The cranfield tests on index language devices. In *Aslib proceedings*, volume 19, 173–192.

- [18] Cormack, G. V., and Lynam, T. R. 2007. Validity and power of t-test for comparing map and gmap. In *Proceedings of ACM SIGIR*, 753–754.
- [19] Creutz, M., and Lagus, K. 2002. Unsupervised discovery of morphemes. In *Proceedings of the ACL-02 workshop on Morphological and phonological learning*, 21–30. Morristown, NJ, USA: Association for Computational Linguistics.
- [20] Creutz, M., and Lagus, K. 2005. Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0. Technical report, Helsinki University of Technology.
- [21] Damashek, M. 1995. Gauging similarity with n-grams: Language-independent categorization of text. *Science* 267:843–848.
- [22] Demner-Fushman, D., and Oard, D. W. 2003. The effect of bilingual term list size on dictionary-based cross-language information retrieval. In *HICSS*, 108.
- [23] Di Nunzio, G. M.; Ferro, N.; Jones, G. J. F.; and Peters, C. 2005. CLEF 2005: Ad Hoc Track Overview. In Peters, C.; Gey, F. C.; Gonzalo, J.; Müller, H.; Jones, G. J. F.; Kluck, M.; Magnini, B.; and de Rijke, M., eds., *CLEF*, volume 4022 of *Lecture Notes in Computer Science*, 11–36. Springer.
- [24] Dobrov, B.; Kuralenok, I.; Loukachevitch, N.; Nekestyanov, I.; and Segalovich, I. 2004. Russian information retrieval evaluation seminar. In *LREC 2004 Proceedings*.
- [25] Fagan, J. L. 1987. *Experiments in automatic phrase indexing for document retrieval: a comparison of syntactic and non-syntactic methods*. Ph.D. Dissertation, Department of Computer Science, Cornell University, Ithaca, US.
- [26] Franz, M.; McCarley, J.; Ward, T.; and Zhu, W. 2001. Quantifying the utility of parallel corpora. In Croft, W. B.; Harper, D. J.; Kraft, D. H.; and Zobel, J., eds., *Proceedings*

- of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-01)*, 398–399. New York: ACM Press.
- [27] Franz, M.; McCarley, J. S.; and Roukos, S. 1998. Ad hoc and multilingual information retrieval at IBM. In *TREC*, 104–115.
- [28] Gale, W. A., and Church, K. W. 1991. A program for aligning sentences in bilingual corpora. In *Proceedings of the 29th annual meeting on Association for Computational Linguistics*, 177–184. Morristown, NJ, USA: Association for Computational Linguistics.
- [29] Gilbert, H., and Spärck Jones, K. 1979. Statistical bases of relevance assessment for the 'ideal' information retrieval test collection. Technical report, Computer Laboratory, University of Cambridge.
- [30] Gollins, T., and Sanderson, M. 2001. Improving cross language retrieval with triangulated translation. In Croft, W. B.; Harper, D. J.; Kraft, D. H.; and Zobel, J., eds., *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-01)*, 90–95. New York: ACM Press.
- [31] Grefenstette, G., ed. 1998. *Cross-Language Information Retrieval*. Kluwer International Series on Information Retrieval. Kluwer Academic Publishers.
- [32] Guthrie, D.; Allison, B.; Liu, W.; Guthrie, L.; and Wilks, Y. 2006. A closer look at skip-gram modelling. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*, 1222–1225.
- [33] Harman, D. 1991. How effective is suffixing? *JASIS* 42(1):7–15.
- [34] Harman, D. 1992. Relevance feedback revisited. In Belkin, N.; Ingwersen, P.; and Pejtersen, A. M., eds., *Proceedings of the 15th Annual International Conference on*

Research and Development in Information Retrieval, SIGIR Forum, 1–10. New York, NY, USA: ACM Press.

- [35] Hiemstra, D., and van Leeuwen, D. A. 2002. Creating an information retrieval test corpus for dutch. In Theune, M.; Nijholt, A.; and Hondorp, G. H. W., eds., *Computational Linguistics in the Netherlands 2001. Selected Papers of the 12th meeting of Computational Linguistics in the Netherlands (CLIN 2001), Enschede, The Netherlands*, volume 45 of *Language and Computers - Studies in Practical Linguistics*, 133–147. Amsterdam, The Netherlands: Rodopi.
- [36] Hiemstra, D. 2001. *Using Language Models for Information Retrieval*. Ph.D. Dissertation, University of Twente.
- [37] Hollink, V.; Kamps, J.; Monz, C.; and de Rijke, M. 2004. Monolingual document retrieval for european languages. *Inf. Retr* 7(1-2):33–52.
- [38] Hull, D. A., and Grefenstette, G. 1996. Querying across languages: A dictionary-based approach to multilingual information retrieval. In Frei, H.-P.; Harman, D.; Schäuble, P.; and Wilkinson, R., eds., *Proceedings of the Nineteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 49–57. New York: ACM Press.
- [39] Hull, D. A. 1996. Stemming algorithms: A case study for detailed evaluation. *Journal of the American Society of Information Science* 47(1):70–84.
- [40] Järvelin, A.; Järvelin, A.; and Järvelin, K. 2007. S-grams: Defining generalized n-grams for information retrieval. *Information Processing and Management* 43(4):1005–1019.

- [41] Jelinek, F., and Mercer, R. 1980. *Pattern Recognition in Practice*. North Holland. chapter Interpolated Estimation of Markov Source Parameters from Sparse Data, 381–402.
- [42] Juola, P. 1998. Measuring linguistic complexity: the morphological tier. *Journal of Quantitative Linguistics* 5(3):206–213.
- [43] Katzner, K. 1999. *The Languages of the World*. London: Routledge.
- [44] Kettunen, K.; Sadeniemi, M.; Lindh-Knuutila, T.; and Honkela, T. 2006. Analysis of EU languages through text compression. In *FinTAL*, 99–109.
- [45] Kishida, K. 2005. Technical issues of cross-language information retrieval: A review. *Information Processing and Management* 41:433–455.
- [46] Knight, K., and Graehl, J. 1998. Machine transliteration. *Computational Linguistics* 24(4):599–612.
- [47] Koehn, P. 2003. *Noun Phrase Translation*. Ph.D. Dissertation, University of Southern California.
- [48] Koehn, P. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT Summit*.
- [49] Kraaij, W. 2001. TNO at CLEF-2001: Comparing translation resources. In Peters, C.; Braschler, M.; Gonzalo, J.; and Kluck, M., eds., *CLEF*, volume 2406 of *Lecture Notes in Computer Science*, 78–93. Springer.
- [50] Kraaij, W. 2004. *Variations on language modeling for information retrieval*. Ph.D. Dissertation, Centre for Telematics and Information Technology, PO Box 217, 5700 AE Enschede, The Netherlands.

- [51] Krovetz, R. 1993. Viewing morphology as an inference process. In *Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 191–202. ACM Press.
- [52] Kurimo, M.; Creutz, M.; and Turunen, V. 2007. Overview of Morpho Challenge in CLEF 2007. In Nardi, A., and Peters, C., eds., *Working Notes of the CLEF 2007 Workshop*.
- [53] Landauer, T., and Littmann, M. 1990. Fully automatic cross-language document retrieval using latent semantic indexing. In *Proceedings of the 6th Conference of the UW Centre for the New Oxford English Dictionary and Text Research*, 31–38.
- [54] Landi, B.; Kremer, P.; Schibler, D.; and Schmitt, L. 1998. Amaryllis: an evaluation experiment on search engines in a french-speaking context. In *Proceedings of the First International Conference on Language Resources (LREC)*, 1211–1214.
- [55] Lee, J. H., and Ahn, J. S. 1996. Using n -grams for korean text retrieval. In Frei, H.-P.; Harman, D.; Schäuble, P.; and Wilkinson, R., eds., *Proceedings of the Nineteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 216–224. New York: ACM Press.
- [56] Lita, L. V.; Rogati, M.; and Lavie, A. 2005. Blanc: learning evaluation metrics for mt. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 740–747. Morristown, NJ, USA: Association for Computational Linguistics.
- [57] Mayfield, J., and McNamee, P. 2003. Single n -gram stemming. In *SIGIR*, 415–416. ACM.

- [58] McCabe, M.; Chowdhury, A.; Grossman, D.; and Frieder, O. 2001. System fusion for improving performance in information retrieval systems. *Information Technology: Coding and Computing, 2001. Proceedings. International Conference on* 639–643.
- [59] McCarley, J. S. 1999. Should we translate the documents or the queries in cross-language information retrieval? In *ACL*.
- [60] McNamee, P., and Mayfield, J. 2002. Comparing cross-language query expansion techniques by degrading translation resources. In *SIGIR*, 159–166. ACM.
- [61] McNamee, P., and Mayfield, J. 2003. Scalable multilingual information access. In Peters, C.; Braschler, M.; Gonzalo, J.; and Kluck, M., eds., *CLEF*, volume 2785 of *Lecture Notes in Computer Science*, 207–218. Springer.
- [62] McNamee, P., and Mayfield, J. 2004. Character N-gram tokenization for european language text retrieval. *Information Retrieval* 7(1-2):73–97.
- [63] McNamee, P., and Mayfield, J. 2005. Translating pieces of words. In Baeza-Yates, R. A.; Ziviani, N.; Marchionini, G.; Moffat, A.; and Tait, J., eds., *SIGIR 2005: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Salvador, Brazil, August 15-19, 2005*, 643–644. ACM.
- [64] Melamed, I. D. 2001. *Empirical Methods for Exploiting Parallel Texts*. Cambridge, MA: MIT Press.
- [65] Metzler, D., and Croft, W. B. 2005. A markov random field model for term dependencies. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and Development in Information Retrieval (SIGIR 2005)*, 472–479.

- [66] Mihalcea, R., and Nastase, V. 2002. Letter level learning for language independent diacritics restoration. In *Proceedings of the Sixth Conference on Natural Language Learning (CoNLL)*, 105–111.
- [67] Miller, E.; Shen, E.; Liu, J.; and Nicholas, C. 2000. Performance and scalability of a large-scale N -gram based information retrieval system. *Journal of Digital Information* 1(5):1–25.
- [68] Mitton, R. 1987. Spelling checkers, spelling correctors and the misspellings of poor spellers. *Information Processing and Management* 23(5):495–505.
- [69] Monz, C., and Dorr, B. J. 2005. Iterative translation disambiguation for cross-language information retrieval. In Baeza-Yates, R. A.; Ziviani, N.; Marchionini, G.; Moffat, A.; and Tait, J., eds., *SIGIR 2005: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Salvador, Brazil, August 15-19, 2005*, 520–527. ACM.
- [70] Mustafa, S. H. 2004. Character contiguity in n -gram based word matching: the case for arabic text searching. *Information Processing and Management* 41:819–827.
- [71] Nie, J.-Y.; Simard, M.; and Foster, G. F. 2000. Multilingual information retrieval based on parallel texts from the web. In Peters, C., ed., *CLEF*, volume 2069 of *Lecture Notes in Computer Science*, 188–201. Springer.
- [72] Oard, D. W., and Diekema, A. R. 1998. Cross-language information retrieval. *Annual Review of Information Science and Technology* 33:223–256.
- [73] Oard, D. W. 2003. The surprise language exercises. *ACM Transactions on Asian Language Information Processing (TALIP)* 2(2):79–84.
- [74] Och, F. J., and Ney, H. 2000. Improved statistical alignment models. In *ACL*.

- [75] Ogawa, Y., and Matsuda, T. 1997. Overlapping statistical word indexing: A new indexing method for japanese text. In *SIGIR*, 226–234. ACM.
- [76] Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 311–318.
- [77] Pirkola, A.; Hedlund, T.; Keskustalo, H.; and Järvelin, K. 2001. Dictionary-based cross-language information retrieval: Problems, methods, and research findings. *Inf. Retr* 4(3-4):209–230.
- [78] Pirkola, A.; Keskustalo, H.; Leppänen, E.; Käsälä, A.-P.; and Järvelin, K. 2002. Targeted s-gram matching: a novel n-gram matching technique for cross- and monolingual word form variants. *Inf. Res* 7(2).
- [79] Pirkola, A.; Puolamäki, D.; and Järvelin, K. 2003. Applying query structuring in cross-language retrieval. *Inf. Process. Manage* 39(3):391–402.
- [80] Ponte, J. M., and Croft, W. B. 1998. A language modeling approach to information retrieval. In *SIGIR*, 275–281. ACM.
- [81] Porter, M. F. 1980. An algorithm for suffix stripping. *Program* 14:130–137.
- [82] Resnik, P.; Oard, D.; and Levow, G. 2001. Improved cross-language retrieval using backoff translation. In *HLT '01: Proceedings of the first international conference on Human language technology research*, 1–3. Morristown, NJ, USA: Association for Computational Linguistics.
- [83] Salton, G., and Buckley, C. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing and Management* 24(5):513–523.
- [84] Sanderson, M., and Zobel, J. 2005. Information retrieval system evaluation: Effort, sensitivity, and reliability. In *ACM SIGIR'05*.

- [85] Savoy, J. 2003. Cross-language information retrieval: experiments based on CLEF 2000 corpora. *Inf. Process. Manage* 39(1):75–115.
- [86] Simpson, H.; Cieri, C.; Maeda, K.; Baker, K.; and Onyshkevych, B. 2008. Human language technology resources for less commonly taught languages: Lessons learned toward creation of basic language resources. In *Proceedings of the LREC 2008 SALT MIL Workshop*.
- [87] Siu, M., and Ostendorf, M. 2000. Variable n-grams and extensions for conversational speech language modelling. *IEEE Transactions on Speech and Audio Processing* 8(1):63–75.
- [88] 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 *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC-2006)*, 2142–2147.
- [89] Talvensaari, T.; Juhola, M.; Laurikkala, J.; and Jarvelin, K. 2007. Corpus-based cross-language information retrieval in retrieval of highly relevant documents. *Journal of the American Society for Information Science and Technology* 58(3):322–334.
- [90] Toivonen, J.; Pirkola, A.; Keskustalo, H.; Visala, K.; and Järvelin, K. 2005. Fuzzy translation of cross-lingual spelling variants using transformation rules. *Information Processing and Management* 41:859–872.
- [91] Unicode Consortium. 2003. *The Unicode Standard Version 4.0.0*. Addison–Wesley.
- [92] Vechtomova, O. 2006. Noun phrases in interactive query expansion and document ranking. *Information Retrieval* 9(4):399–420.

- [93] Véronis, J., ed. 2000. *Parallel Text Processing: Alignment and Use of Translation Corpora*. Dordrecht: Kluwer Academic Publishers.
- [94] Vilares, J.; Oakes, M. P.; and Ferro, M. V. 2007. *Character N-Grams Translation in Cross-Language Information Retrieval*, volume 4592. Berlin / Heidelberg / New York: Springer-Verlag. 217–228.
- [95] Voorhees, E. M., and Harman, D. K. 1998. Overview of the seventh text REtrieval conference (TREC-7). In *Text REtrieval Conference (TREC) TREC-7 Proceedings*. Department of Commerce, National Institute of Standards and Technology. NIST Special Publication 500-242: The Seventh Text REtrieval Conference (TREC 7).
- [96] Voorhees, E. M., and Harman, D. K. 2005. *TREC: Experiment and Evaluation in Information Retrieval*. Digital Libraries and Electronic Publishing. MIT Press.
- [97] Voorhees, E. M. 2000. Variations in relevance judgments and the measurement of retrieval effectiveness. *Inf. Process. Manage* 36(5):697–716.
- [98] Wang, J., and Oard, D. W. 2006. Combining bidirectional translation and synonymy for cross-language information retrieval. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, 202–209. New York, NY, USA: ACM.
- [99] Xu, J., and Weischedel, R. 2000. Cross-lingual information retrieval using hidden Markov models. In Schütze, H., and Su, K.-Y., eds., *Proceedings of the 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing*. Somerset, New Jersey: Association for Computational Linguistics. 85–103.
- [100] Xu, J.; Fraser, A.; and Weischedel, R. M. 2001. TREC 2001 cross-lingual retrieval at BBN. In *Text REtrieval Conference (TREC) TREC-2001 Proceedings*.

- [101] Yilmaz, E., and Aslam, J. A. 2006. Estimating average precision with incomplete and imperfect judgments. In *CIKM '06: Proceedings of the 15th ACM international conference on Information and knowledge management*, 102–111. New York, NY, USA: ACM.
- [102] Zamora, E. M.; Pollock, J. J.; and Zamora, A. 1981. The use of trigram analysis for spelling error detection. *Inf. Process. Manage* 17(6):305–316.
- [103] Zhai, C., and Lafferty, J. 2004. A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.* 22(2):179–214.
- [104] Zhang, Y., and Vines, P. 2004. Using the web for automated translation extraction in cross-language information retrieval. In Sanderson, M.; Järvelin, K.; Allan, J.; and Bruza, P., eds., *SIGIR 2004: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Sheffield, UK, July 25-29, 2004*, 162–169. ACM.
- [105] Zobel, J., and Dart, P. 1995. Finding approximate matches in large lexicons. *Software - Practice and Experience* 25(3):331–345.
- [106] Zobel, J. 1998. How reliable are the results of large-scale information retrieval experiments? In *SIGIR'98*, 307–314.

