Domain Adaptation of Translation Models for Multilingual Applications

Monica Rogati

CMU-CS-09-116

April 2009

Computer Science Department
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Yiming Yang, Chair
Jaime Carbonell, Chair
Jamie Callan
Salim Roukos, IBM TJ Watson

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Copyright © 2009 Monica Rogati

This research was sponsored by the National Science Foundation under grant numbers IIS-9618941, IIS-9873009, IIS-9982226, and EIA-9873009, and the International Business Machines (DARPA prime) grant number W0550432.

The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of any sponsoring institution, the U.S. government or any other entity.

Keywords: domain adaptation, machine translation, statistical machine translation, parallel corpora, domain specific, cross-language information retrieval, CLIR, criteria optimization, SMT, resource selection, training resource adaptation

Abstract

The performance of a statistical translation algorithm in the context of multilingual applications such as cross-lingual information retrieval (CLIR) and machine translation (MT) depends on the quality, quantity and proper domain matching of the training data. Traditionally, manual selection and customization of training resources has been the prevailing approach. In addition to being labor-intensive, this approach does not scale to the large quantity of heterogeneous resources that have recently become available, such as parallel text and bilingual thesauri in various domains. More importantly, manual customization does not offer a solution to efficiently and effectively producing tailored translation models for a mixture of heterogeneous target documents in various domains, topics, languages and genres. Translation models trained on a general domain do not work well in technical domains; models trained on written documents are not appropriate for spoken dialogue; models trained on manual transcripts can be sub-optimal for translating noisy transcripts produced by a speech recognizer; finally, models trained on a mixture of topics are not optimal for any of the topic-specific documents.

We seek to address this challenge by automatically adapting translation models (and implicitly parallel training resources) to specific target domains or sub-domains.

The high-level adaptation process involves automatically weighting and combining multiple translation resources, according to several criteria, in order to better match a target corpus or a specific domain sample. The criteria we examine include lexical-level domain match, translation quality estimates, size, and taxonomy representation. An orthogonal dimension in the adaptation process is the granularity level at which these criteria are measured and applied: from the collection level - under the assumption of homogeneous within-collection data - to the document level. The relative contribution of each criterion is subsequently determined by a model that can range from uniform weighting to a global non-linear optimization model trained on application specific evaluation data.

In this thesis, we examine how such adaptation applies to two important multilingual

applications: cross-lingual information retrieval and machine translation. In CLIR, we adapt translation models for domain-specific query translation; in MT, we adapt translation models to heterogeneous target corpora and compare them with previously studied target language model adaptation. We use our adaptation algorithms to enhance state-of-the-art systems, seeking to improve performance under different testing conditions and to reduce the demand for large amounts of domain specific parallel data. We also address the challenge of combining multiple criteria to rank parallel sentence candidates. We investigate Continuous Reactive Tabu Search (CRTS) [2], a global optimization method, as well as Reactive Affine Shaker (RASH) [6], an efficient algorithm which continuously adjusts its search area in order to identify a local minimum.

Our experiments in CLIR and statistical MT indicate that selecting training data based on the above-mentioned approaches allows a significant reduction in training data while preserving about 90% of the performance. This result significantly surpasses the random selection approach, and it holds for both CLIR and MT. As expected, the difference increases as the subdomain becomes more specific. Our optimized criteria weights considerably outperform the uniform distribution baseline, as well as lexical similarity adaptation.

Acknowledgements

The number of pages of this dissertation could easily double if I enumerated all people who have contributed to my computer science education. I can start, as cliché as it sounds, with my mother, who suppressed her curiosity for the next chapters of my sci-fi stories and encouraged me to pursue math and computer science as soon as she noticed my passion for it. Or with my father, who ignited that passion and fueled it with puzzles, guidance and late night BASIC programming. I'm grateful to both my parents - for being incessant trailblazers who had the foresight to spend $>$ 6 months' salary on a computer at a time when few people in my Romanian small town have even heard of one, and the strength to encourage me to apply for a scholarship thousands of (physical and cultural) miles away at the age of 17.

I'm grateful to my teachers, mentors and professors - to Mrs. Rădulescu and Vişinescu for math Olympiad preparations masterfully disguised as fun puzzles; to Debbie and Jody Gambles for making it possible to attend college in the US and to study computer science; to prof. Shapiro for suggesting graduate studies (although there were times in the past few years when my opinion on the subject had taken a dive); to Barak Pearlmutter for challenging countless paradigms in how I thought about computer science in particular and learning in general; to Marilyn Walker for showing me how exciting research can be; to my committee members, Salim Roukos and Jamie Callan for their valuable input and suggestions during the preparation of this dissertation; and finally, to Yiming Yang and Jaime Carbonell, my advisors, for their insights and support throughout my graduate studies.

I'm grateful to the skeptics - to the big city teacher who told me my small town education is not enough to pass the entrance exam to the computer science high school in Bucharest; and to my first US host family who told me I'd never graduate from a US high school. They (and K.L., their polar opposite) have taught me a valuable lesson about persistence, hope and resilience.

It has been a long journey - and throughout it all, I'm most grateful for the presence and constant support of the most extraordinary person I've met - Lucian Vlad Lita. I was incredibly lucky to have Lucian and his many hats, from his remarkable insights into research

problems during numerous professional discussions, to his moral support in tough times, to his advice and contributions during strategic and tactical planning, to his light hearted puns that made challenging days bearable.

I'm grateful to everybody who showed me that the journey doesn't end here - that it is a starting point of a fascinating time in a world-changing field.

Table of Contents

Chapter 1

Introduction

In our increasingly connected world, where information has become a de facto currency, communication across language barriers is a crucial and frequently encountered challenge. Facilitating information access on a global scale is one of the goals of modern technology, ranging from providing the communication medium, to its instant propagation across the globe and beyond, to transforming the information into a data stream that can be digested by information consumers - be it persons or computers. This last step often involves translation from one or multiple languages into a language that the information consumer is familiar with.

As the amount of data requiring processing and translation has become insurmountable, human translators are no longer a realistic solution from a cost and speed perspective. This problem is exacerbated particularly when the data is rooted in a technical domain with highly specific vocabulary - in which case both bilingual capabilities and domain expertise are required in a translator.

Multilingual applications and algorithms such as machine translation (MT) and cross-language information retrieval (CLIR) are part of a suite of automated tools facilitating cross-language communication. *Machine Translation* refers to the automated translation of (usually text) data from one natural language into another. The *Cross-Lingual Information Retrieval* (CLIR) problem consists of finding documents in a *target language* that are relevant to queries expressed in a (different) *source language*. The most popular (and, currently, the most successful) subset of these algorithms rely on (and take advantage of) massive quantities of previously translated examples, referred to as *parallel corpora*.

Parallel corpora are produced at an increasing rate - from commercial web sites to news sources to bilingual books to European Union or Canadian legislation and parliament proceedings, sources publish their information in several languages simultaneously. Leveraging

these resources (and their increasing rate of growth) is the advantage of corpus-based methods in the broader fields of CLIR and MT. However, the growth and scale of these resources, as well as the question of which resource (or subset thereof) to use for training these algorithm poses its own challenges. In the remainder of this chapter, we will discuss the importance of these challenges, as well as give a brief overview of the multilingual tasks we are addressing in this thesis.

1.1 Adapting Training Resources for Multilingual Applications

The performance of a corpus-based translation algorithm depends on the quality, quantity and proper domain matching of the training data. Translation models trained on a general domain would not work well in a technical domain; those trained on written-style documents would not work well on spoken dialogs; models trained on manual transcripts can be sub-optimal for translating noisy transcripts produced by a speech recognizer; a single model trained on the entire mixture of topics would not be optimal for any of the topic-specific documents.

Until recently, manual selection and customization of cross-lingual training resources has been the prevailing approach. In addition to being labor-intensive, this approach does not scale to the widely available heterogeneous resources on the Internet (parallel text and bilingual thesauri in various domains), and more importantly, does not offer a solution to automatically producing tailored translation models for a mixture of heterogeneous test documents differing by domain, topic, language and genre. The amount of electronically available data is increasing every day; bilingual web pages are harvested as parallel corpora as the quantity of non-English data on the web increases; online dictionaries of various qualities and in various domains become available; previously translated documents are automatically aligned, and time-aligned comparable news are published every day. These resources are different in size, quality, vocabulary, genre, other domain characteristics, and purchasing cost. They provide the potential of significantly enhancing the performance of corpus-based statistical methods for CLIR or MT.

More recently, domain-specific multilingual tasks such as patent cross-language retrieval [20] and domain-specific machine translation tasks have begun to gain attention in evaluation forums. The parallel corpus adaptation is, in these cases, largely manual, and restricted to domain specific data provided. We provide an overview of related research within this area,

and of domain adaptation literature in general in Section 1.2.

In general, manual selection and customization of training resources cannot scale to large amounts of training corpora, and cannot produce consistently good customized translation models when the test data vary constantly in domain, topics, and genres. Usually, documents come from a mixture of heterogeneous resources (newswires, TV broadcasts and radio programs) and differ by language, genre, topic coverage and signal-noise ratio (manual transcripts vs. automated transcripts by speech recognition). Statistical translation models trained on conversational dialogs (e.g., in radio broadcasts) would most likely be sub-optimal for translating written-style documents (newswire stories), and vice-versa. Similarly, models trained on clean text (written stories or manual transcripts) are not suitable for automated transcripts since systematic speech-recognition errors would not be reflected properly.

Once we move away from the news domain and into more technical domains, translation-specific disambiguation becomes more problematic, in that the most common translation is no longer the most desirable one. For example, the word *agent* should have different translation probabilities when the target corpus consists of newspaper stories (where the *agent* is more likely to be a person or entity) vs. medical literature or biological weapon articles (where the *agent* is more likely to be chemical or biological).

We argue that domain-dependent term variations can be captured by successfully matching training resources to target corpora. Our approach is to customize translation models to a domain, by automatically selecting the resources (dictionaries, parallel corpora) that are best for training for this particular topic. Such customized translation models have the potential of successfully capturing the specific, specialized vocabularies present in each domain or sub-domain, in addition to providing a tradeoff curve between parallel corpus quality and its domain-specific match.

In this thesis, we seek to develop an automated solution for domain adaptation of training resources. The adaptation process produces a weighted combination of several translation resources, according to multiple criteria, in order to better match a target corpus or a specific *domain sample*. A *domain sample* is defined as a set of unlabeled data points in the target domain. In both the cross-lingual information retrieval and the machine translation scenario, a domain sample is a collection of monolingual sentences that represent the target domain.

In our medical literature dataset, an article titled *Remodelage bronchique* (airway remodeling) was translated as *Re-drawing bronchique* when an otherwise popular parallel corpus (Europarl, [36]) was used as the training set, and *Bronchial Remodeling* when an adapted, medical domain corpus was used. In this thesis, we will show that the performance penalty

goes well beyond such anecdotal evidence, and we propose a multi-faceted adaptation solution.

The adaptation criteria we examine include lexical domain match, translation quality estimate, instance size and taxonomy representation. An orthogonal dimension in the adaptation process is the granularity level at which these criteria are measured and applied: from the collection level (under the assumption of homogeneous within-collection data) to the document level. The relative contribution of each criterion is subsequently determined by a model that can range from uniform weighting to global non-linear optimization algorithms trained on application-specific evaluation data.

We examine how such adaptation applies to two multilingual applications: cross-lingual information retrieval and machine translation. In CLIR, we adapt translation models for domain-specific query translation and in MT we adapt translation models to heterogeneous target corpora in the medical domain, and, where applicable, we also adapt translation models to a more specific subdomain.

Our goal is to use adaptation algorithms to enhance state-of-the-art systems in these application areas, seeking to improve performance under different testing conditions and to significantly reduce the training data necessary to obtain equivalent performance.

1.2 Motivation and Related Work

1.2.1 Corpus-Based Machine Translation and Domain Adaptation

Over the past few decades, corpus-based machine translation has been the preferred paradigm in machine translation. Its superior performance when compared to earlier, rule- and dictionary based approaches can also be attributed to the (gradual) availability of large quantities of parallel text. Example-based translation (EBMT) [5] uses the parallel text to identify previously translated phrases; and it incorporates existing linguistic resources and knowledge [5, 55], which is especially important in resource-poor languages. Statistical Machine Translation (SMT) [4, 29] uses a probabilistic approach to find the most likely translation of a sentence, given the sentence in the original language. Statistical machine translation as described in [4] has become the foundation for an entire research area, with constantly improving effectiveness and enhancements such as phrase-based decoding [53] and parsing-aware approaches [44].

Until recently, research in corpus-based machine translation has focused on general domains, with the corresponding evaluations being mostly news. However, domain-specific

translation is crucial for tasks such as translation of (and cross-lingual retrieval in) instruction manuals, medical articles or other technical literature. As we have shown in [46], the performance penalty in these multilingual tasks is staggering unless translation model adaptation is employed.

In the past two years, domain adaptation for NLP tasks has become an active research area [3, 38, 25, 23]. New domain adaptation tasks have surfaced: a shared CoNLL task on domain adaptation for parsing [24], a statistical MT workshop evaluation, as well as CLIR evaluations in NTCIR [19]. As an emerging research area explored after the proposal of this thesis, "domain adaptation" is now understood as the task of adapting a previously constructed model to a new domain, using only unlabeled in domain data. Blitzer examines various aspects of "domain adaptation" as defined above - its application to sentiment classification [27] and parsing [41], learning bounds [26] and representations [62, 3]. The results are mixed, ranging from "frustratingly easy" [23] to "frustratingly hard" [41] - mainly due to the differences in annotation orthogonal to the domain-specific characteristics of the data. Jiang and Zhai [25] focus on adapting pre-trained classifiers to a new domain using unlabeled data; Daumé III and Marcu [22] use both labeled and unlabeled data in the target domain and learn a mixture model to adapt from the source domain. Other NLP tasks where domain adaptation has been studied include capitalization restoration using an enhanced maximum entropy approach [9], and word-sense disambiguation [8].

While this active research area is directly relevant to this thesis, in that an existing model is adapted to a new domain, there are several fundamental differences in both the problem definition and the proposed approaches. The problem setting in the above work assumes the presence of *identified* in-domain data, labeled or unlabeled. The adaptation we perform in this thesis is at the level of the training data itself, while the model is subsequently re-trained as opposed to adapted post-training. The fundamental difference here is the assumption we make that in-domain (or quasi in-domain) training instances are available within the larger training data, and, although *they are not specifically labeled as such*, it is possible to identify these instances as in-domain by automated means. Once identified, the in-domain instances can be used to re-train the model, while eliminating the noise contributed by i.e. out-of-domain data with conflicting labels.

Another important difference refers to the multi-faceted characteristics of the training data that we can leverage. Some facets (or criteria) are specific to multilingual applications (for example, translation quality estimates - this can be generalized to label quality estimate in an expanded application space). As far as we know, none of the work referred to above uses additional criteria when adapting the pre-trained learner, as we do in the work presented

here.

In the specific space of multilingual applications (MT and CLIR), domain-specific corpora have started to surface in response to a growing need for domain-specific evaluation. For example, the NTCIR evaluation forum includes a patent-retrieval task, for which a bilingual patent corpus is provided. Our purpose, however, is to construct a highly adapted, domain-specific corpus out of existing corpora - parts of which pertain to a given domain.

Within the MT context, initial adaptation efforts focused on adapting the *target language model* (vs. the translation model) to the specific domain [3, 32], building on similar work for speech recognition [60]. More specifically, in [34], after the translation model has been trained, a language model is constructed for each of the test sentences by retrieving documents in the target language using CLIR. A back-off general English model is used to cover the non-domain specific terminology. The problem we examine in this thesis is the adaptation of the translation model itself (via training data adaptation), which is orthogonal and complementary to language model adaptation. In Chapter 5 of [31] Kauchak describes a method for learning machine translation example usefulness - a purpose similar to ours, although different in scope. In [31], Kauchak does not target the domain adaptation problem, but instead it focuses on ranking parallel sentences based on their performance as part of a training subset. Each example is randomly assigned to a subset, and its contribution to the subset performance is assumed to be linear. The example/parallel sentence ranking is derived by averaging the performance (i.e. BLEU scores) of all the subsets an example appears in. The differences observed are very small when compared to the differences observed when domain match is taken into account (i.e. 0.34 difference in BLEU score on a 0-100 scale), but they allow the author to identify features that characterize promising parallel sentences. This approach is not feasible for a larger dataset to select from, due to the continuous re-training and testing on random samples necessary in order to estimate the tion of each example.

In MT, several groups have informally given priority to parallel corpora or sentences that ensured vocabulary coverage at testing time, to the extent this information was available. [14] uses a method similar to [46] in order to filter parallel sentences that are closest to test sentences. This on-line, test-time adaptation technique does improve results if on-line response time permits its utilization. The above-mentioned MT/CLIR domain adaptation work only took into account the domain match as indicated by vocabulary and word distribution [14, 46]. However, especially in the case of today's heterogeneous parallel corpora, it is especially important to consider issues such as corpus and translation quality, noise, size distribution, redundancy, genre or other available metadata in additional to lexical similarity.

In this thesis, we explore the impact of these criteria on both online and offline adaptation. One of the main advantages of domain adaptation is the significant reduction in the amount of training data necessary. When the adaptation is performed offline, adaptation can be seen as an active-learning scenario, in that the selected data can be translated at a much smaller cost that an entire corpus.

In this thesis, we present an adaptation framework that incorporates several adaptation criteria and we show the individual (and combined) criteria effects on MT and CLIR performance.

In research areas directly related to multilingual applications, domain-specific language modeling has been used in speech recognition [33], with encouraging results. [32] used CLIR followed by MT to find domain-specific articles in a resource-rich language, in order to use them for language modeling in a resource-poor language.

Resource analysis and modeling has been previously studied in the context of federated search / distributed IR [7]. However, in distributed IR, the "resources" are (usually limited access) databases to be queried; here, our target collection is available and resources are defined as any aids in crossing the language barrier, including parallel corpora, dictionaries or MT systems. An overview of distributed IR can be found in [1] and [7].

1.2.2 Additional Relevant Research in CLIR

In cross-lingual information retrieval (CLIR), benchmark evaluations have shown that the performance of some systems has reached that of monolingual retrieval, as seen in TREC, NTCIR and CLEF [30, 11, 17]. The most successful corpus-based approaches combine the translation of queries and documents, or integrate translation in the retrieval models [11, 17, 18]. We use a similar approach, described in Chapter 5, which also performed well in multilingual retrieval evaluation forums [45].

In the past few years, CLIR has evolved from being an active academic research area in the first half of the decade, into large-scale implementation in industry. Currently, several online news sources and blogs identify CLIR as "a new Google algorithm" (i.e. in [15]), due to the launch of CLIR capabilities in Google in the summer of 2007.

Resource combination in CLIR

Previous work in CLIR addressed a problem related to its domain adaptation: choosing or weighting different translations, when more than one is available. A popular approach [12] is to concatenate translations obtained from different sources. This does not take into

account the target corpus and its domain, and it does not attempt to disambiguate query term translations. [13] combines the evidence for alternate translations by modifying a structured query method to use translation probabilities. This approach does take into account target corpus characteristics, but it uses already unified translation probabilities that match the target corpus usage. In this context, favoring a translation that is common in the target corpus, but improbable given the training corpus, is deemed undesirable. We argue that technical, domain-specific terms exhibit exactly this behavior when a general-purpose training corpus is used. Consequently, correctly disambiguating these terms by choosing the translation that is more common in the target corpus is not undesirable. [10] addresses the issue by retaining the top two translations that occur most frequently in the target collection. We accomplish this goal by incorporating domain-specific information into the translation probability.

Domain Specific Research in CLIR

Domain specific research in CLIR focused on the social sciences, with CLEF tasks using corpora such as GIRT (German) and RSS (Russian). Challenges for domain-specific CLIR, in particular the problem of distinguishing domain-specific meanings, have been noted in [35]. Newer tasks in NTCIR (an evaluation forum focused on Asian languages) focused on patent retrieval. All of these task identify in-domain data a priori, and in many cases the use of additional training data is prohibited.

1.3 Contributions

This thesis provides a solution to automatically adapting translation models to specific target corpora and domains using multiple criteria.

Specific contributions include:

- Existence proof by demonstration that automated means for domain specific training corpus selection for training translation mode ls to improve cross-language information retrieval and machine translation
- Developing, examining and evaluating **multiple criteria** for selecting and weighting translation training resources for multilingual applications
- A **flexible framework** for incorporating individual criteria (i.e. translation quality, instance size, lexical similarity, taxonomy information, other metadata) into a resource

adaptation model

- Integrating a high-performing statistical **machine translation** (MT) system into the domain adaptation framework by translation model adaptation. Exploring the benefits of adaptation by examining a) its impact on state-of-the-art MT systems and b) the effect of individual criteria on MT-specific issues.

- Examining the performance tradeoffs between traditional online training set adaptation and its more realistic **offline adaptation** counterpart in the context of MT.

- Developing methods for translation model adaptation, as specifically pertaining to the problem of **cross-language information retrieval** (CLIR). Demonstrating the advantages of these methods by examining their impact on state-of-the-art CLIR systems and the effect of individual adaptation criteria on CLIR-specific performance.

- Delivering an instrument for **cost reduction in an active learning context**. The labeling/translation cost is drastically reduced by identifying and focusing on the unlabeled (i.e. monolingual) training instances with the highest estimated utility given a domain sample.

- Providing a high performing statistical **CLIR system** and using it as an application within the domain adaptation framework.

- Compensating for inconsistent or particularly **poor parallel corpora** by integrating translation quality estimates into the selection model and showing the significant performance improvements.

- Presenting a flexible technique for leveraging **existing domain-specific resources** (dictionaries, taxonomies, parallel corpora) by simply including them in the pool of resources to be selected and weighted.

- Using global and local non-linear optimization methods such as Reactive Affine Shaker and Continuous Reactive Tabu Search in order to **find optimal weights dictating relative importance of above-mentioned criteria** incorporated in a domain adaptation method.

- Significantly alleviating the considerable performance penalty incurred when using a proven high-performance (but general domain) MT system on domain specific test data.

- Prioritizing training instances such that a **reduction in training resources** by more than 90% results in 10% performance difference, even when compared to using all of the available domain-specific data. More than one order of magnitude reduction in training data can be accomplished with little performance loss, or, in the case of online adaptation, a performance gain.
- A tool that, given a target corpus (or domain sample) and a pool of translation resources (parallel corpora, dictionaries etc.) will present a **customized parallel corpus** (or pre-trained translation model) tailored to the given domain sample.

The remainder of the thesis is organized as follows:

- Chapter 2 presents the parallel resource domain adaptation (PARDA) framework. It discusses adaptation criteria and their combination into a model.
- Chapter 3 includes the first of the multilingual applications - statistical machine translation system and data used.
- Chapter 4 continues by showing adaptation results on statistical machine translation.
- Chapter 5 describes the cross-language information retrieval system we have developed, including its performance and the data used.
- Chapter 6 shows the effects of corpus domain adaptation on domain specific cross-language information retrieval.
- Chapter 7 concludes the thesis by emphasizing its contributions and impact.

Chapter 2

Parallel Resource Domain Adaptation (PARDA) Framework and Adaptation Criteria

2.1 Overview

In the first chapter, we have motivated the need for a flexible framework that, given several parallel resources and a domain sample, produces a customized, domain adapted parallel resource.

Lexical similarity is the most direct adaptation criterion to ensure the domain match of the newly created parallel resource. However, especially in the case of today's heterogeneous parallel corpora, it is especially important to consider issues such as corpus and translation quality, noise, size distribution, redundancy, genre or other available metadata in additional to lexical similarity. In this chapter we present an adaptation framework that incorporates several adaptation criteria. In later chapters, we show the individual (and combined) criteria effects on MT and CLIR performance.

2.2 The Parallel Resource Domain Adaptation Process

The high-level adaptation process involves automatically weighting several training resources such as parallel and comparable corpora and dictionaries, according to several criteria, in order to better match a provided domain sample.

Figure 2.1 illustrates the high-level concept of adapting parallel resources to a given domain sample.

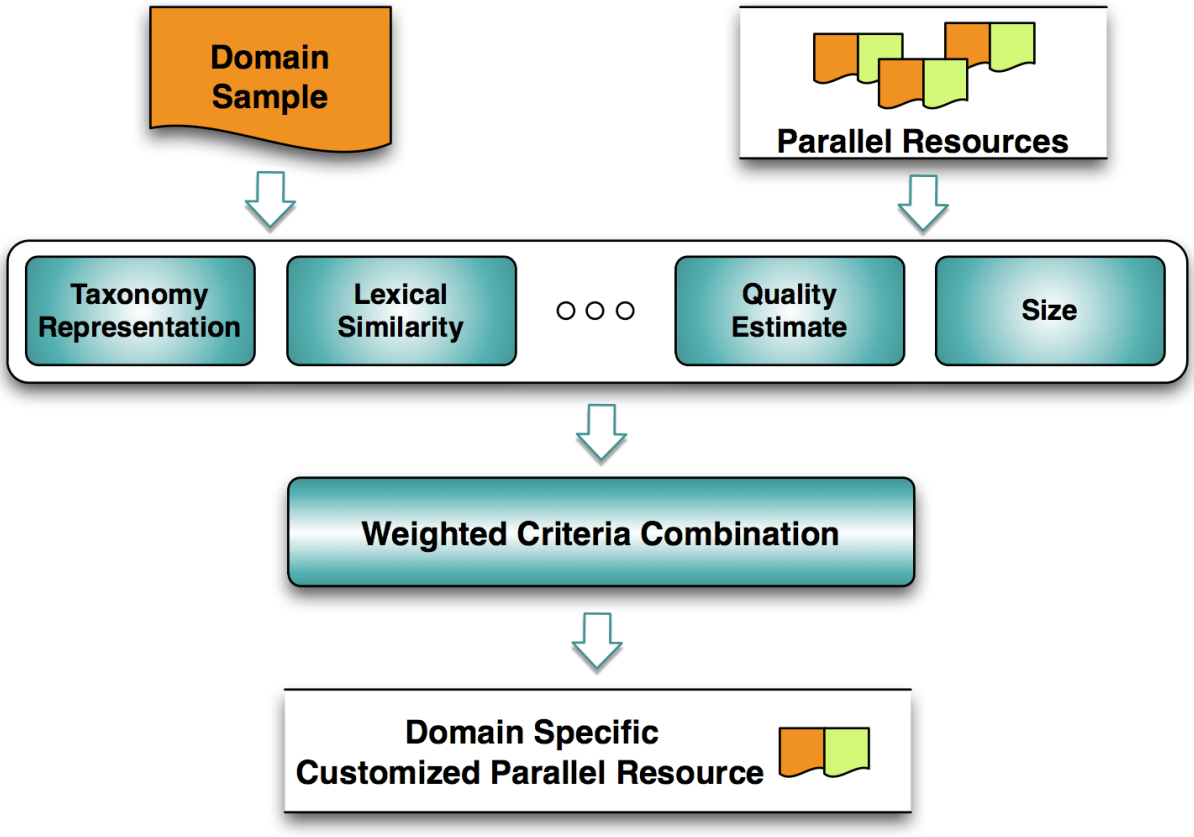

Figure 2.1: Adaptation Framework Overview

Given the domain sample and a parallel resource candidate (i.e. an entire corpus, or a particular parallel sentence, or a dictionary entry), the first component of the system produces a profile for each of the given resource. The profile includes features such as the lexical domain similarity to the target corpus, the language, the within-corpus topic

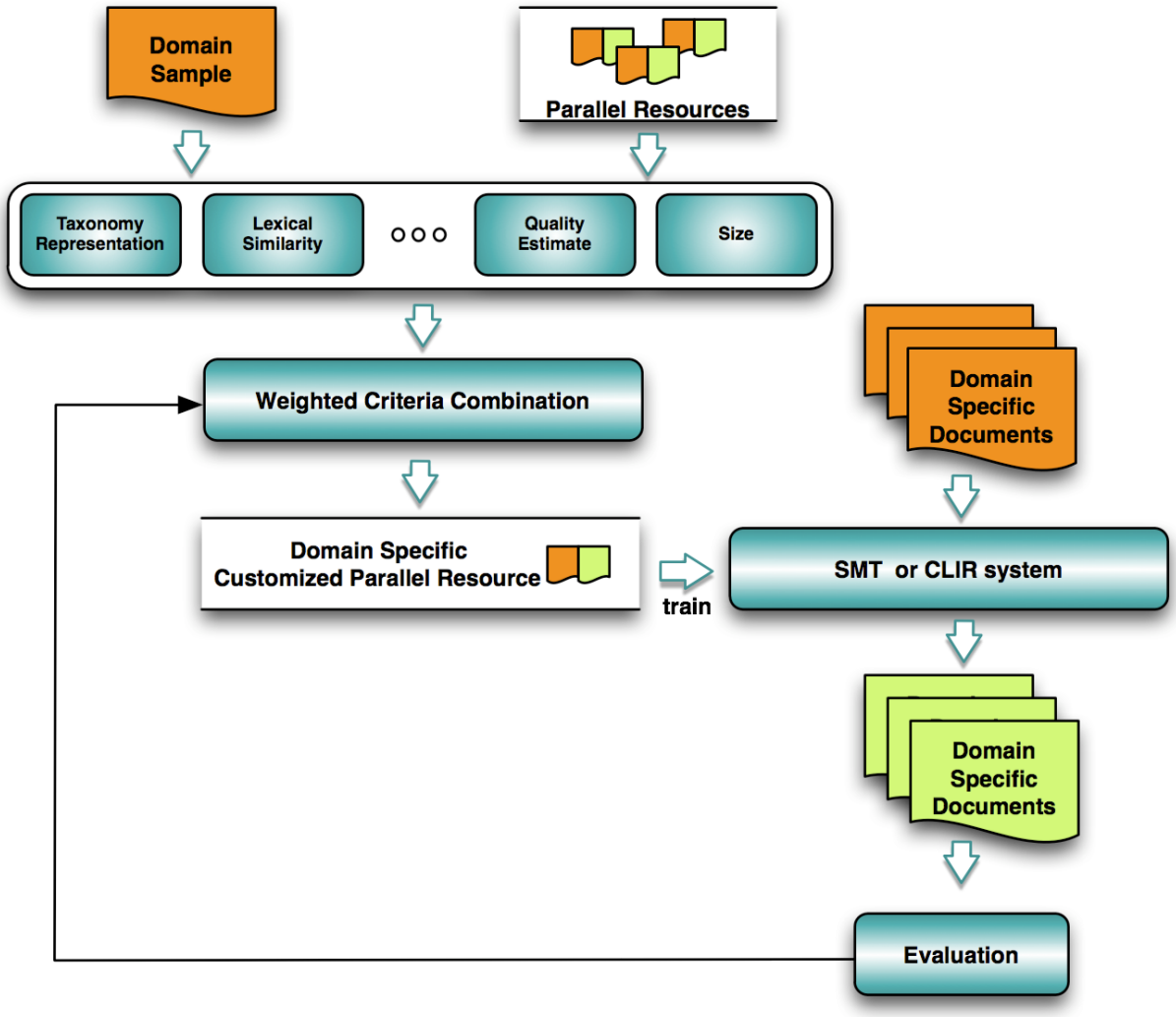

Figure 2.2: Adaptation Framework Overview - Integration and Evaluation

distribution, the corpus size and a translation quality estimate. More specifically, each of the above-mentioned criteria calculates a fitness score for the candidate. The criteria can be included or excluded, depending on their suitability to a particular task.

The second component of the system decides the relative weights of each criterion (and criterion-specific score) in the profile and it can range from a simple model giving equal weights to all criteria, to directly using confidence scores generated using individual criterion, to a more sophisticated non-linear optimization module based on evaluation results when training data is available (the approach described in this thesis). In this setting, a key

observation is that the training of the relative weights depends on the available validation data. This module, described in Chapter 7, optimizes the corresponding performance of the resulting translation model for each multilingual application.

Once each parallel resource candidate has been assigned a weight, the domain-adapted resource is constructed by a) mixing the given resources proportionally with their weights, or b) ranking all the candidates and keeping the top $N\%$ as contributors in the new domain-adapted resource.

Figure 2.2 shows how the adaptation framework fits into training and evaluation. Once the domain adapted corpus is created, a multilingual (CLIR or SMT) system is trained and evaluated. In the learning phase, the results of the evaluation are fed back into the criteria combination module, optimizing their relative weights. In the evaluation phase, results are reported on CLIR/SMT performance given the pre-trained mixture model, and the resulting domain-adapted parallel training corpus.

The specifics of each criteria, as well as the combination mechanisms and weights are outlined below.

2.3 Adaptation Granularity

When constructing a new, customized translation training resource from existing resources, we have several choices for the granularity that is the most appropriate for the selection/weighting unit. In particular, when combining parallel corpora, we could make the following decisions concerning the basic unit of combination:

- Corpus level: An entire parallel corpus, or the translation model trained on such a corpus (macro adaptation)
- Document level: A document in any of the parallel corpora. Since many corpora are sentence-aligned, the document in this case is a sentence, or a dictionary entry in the case of dictionaries (micro adaptation)

- Cluster level: A cluster of documents originating in the same corpus. In this setting, the additional problem of cluster size, density and number introduces additional free parameters into the system but it has the potential of faster performance in the case of a pre-trained model.

2.3.1 Macro (Corpus Level) Adaptation

In macro adaptation, the basic unit used to construct a new, customized parallel resource is an entire parallel corpus (or a translation model trained on such corpus). The corpus is seen as a coherent unit, with consistent translation quality, within a single domain and/or topic, with documents belonging to the same genre. Any available meta-data varies from corpus to corpus, but is consistent within the corpus.

These assumptions are usually true for corpora obtained from the same source (for example, bilingual news in the same newspaper, a translated novel, or the Canadian Parliament proceedings). Moreover, this level of granularity has the advantage of allowing off-line training of the translation models followed by efficient on-line combination when required.

2.3.2 Micro (Document Level) Adaptation

In macro adaptation, we use entire weighted parallel corpora as resources to build a domain-specific translation model. This approach treats a parallel corpus as a homogeneous entity, an entity that is self-consistent in its domain and document quality. In this section, we propose that instead of weighting entire resources, we can select individual documents from multiple corpora in order to build a parallel corpus that is tailor-made to fit a specific target collection (e.g. a set of topically coherent documents to translate). In previous work [59] micro-adaptation lead to superior results when compared to macro-adaptation on a medical domain corpus.

We compute the domain match score between the target collection and each individual document in the parallel corpora for that respective language. Once this is computed for each document in the parallel corpora, only the top N most similar documents are kept for

training. Alternatively, the document score can also be incorporated into the translation model, eliminating the need for thresholding.

2.4 Adaptation Criteria

Several factors influence the extent an individual parallel corpus or translation resource contributes to a newly constructed domain-specific corpus. The vocabulary match (or the lexical similarity) is a prime indicator for a close domain match. Translation/corpus quality (i.e. the degree to which the corpus is actually parallel) is another feature we need to consider, especially when the quality and sources of the available corpora vary. The size of a corpus is important since it provides a measure of the quantity of the training data available. A resource can be projected into dimensions determined by taxonomies and genres; similarity along these dimensions can prove to be significant when adapting the translation model.

2.4.1 Lexical-Based Domain Similarity

Word-level (lexical) similarity is the low-hanging fruit of domain matching. A good overview of distributional similarity measures (as applied to word distributions) can be found in [39]. Usually used exclusively, word-level similarity is a crucial domain match criterion. For both online and off-line adaptation, we calculate the similarity of each candidate sentence in the parallel corpus to the domain sample. Similarity measures vary, starting with word overlap, cosine similarity and ending with the language-model information retrieval measure of the probability of a segment being generated by the domain sample. In this thesis, our results are shown using the latter, as implemented by INDRI [64], without pseudo-relevance feedback and using Dirichlet smoothing.

We examine a series of other similarity measures an outline their advantages and disadvantages in order to analyze their suitability to the task at hand (lexical similarity).

1. Cosine Similarity

$$cos(p, r) = \frac{\sum_w p(w)r(w)}{\sqrt{\sum_w p(w)^2}\sqrt{\sum_w r(w)^2}} \tag{2.4.1}$$

Cosine similarity does not depend on the length: This allows documents with the same content, but different length (i.e. a document duplicated over and over) to be treated identically which makes this the most popular measure for text documents. We have experimented with cosine similarity in the German language experiments presented in Chapter 5.

2. Jaccard's coefficient

   The binary version of Jaccard's coefficient measures the degree of overlap between two sets:

$$Jaac_b(p, r) = \frac{|pIr|}{|pYr|} \tag{2.4.2}$$

   Jaccard's coefficient binary definition can be extended to non-negative features:

$$Jac(p, r) = \frac{\sum_w p(w)r(w)}{\sum_w p(w)^2 + \sum_w r(w)^2 - \sum_w p(w)r(w)} \tag{2.4.3}$$

   which is equivalent to the binary version when the feature vector entries are binary.

   It retains the sparsity property of the cosine while allowing discrimination of collinear vectors.

3. Dice coefficient

   The Dice coefficient is monotonic in Jaccard's coefficient, so its inclusion would be redundant [39]

$$Dice(p, r) = \frac{2\sum_w p(w)r(w)}{\sum_w p(w)^2 + \sum_w r(w)^2} \tag{2.4.4}$$

4. Overlap (vocabulary coverage)

   Vocabulary coverage is particularly important in machine translation tasks, where the effect of an out-of-vocabulary word is more easily noticed. We have experimented with overlap in [46].

$$cov(p, r) = \frac{|pIr|}{|p|} \tag{2.4.5}$$

5. Euclidean distance

$$L_2(p, r) = \sqrt{\sum_w (p(w) - r(w))^2} \tag{2.4.6}$$

6. Hellinger distance

$$d_k(p, r) = \sum_w (\sqrt{p(w)} - \sqrt{r(w)})^2 \tag{2.4.7}$$

7. Kullback-Leibler divergence

   Also known as relative entropy, the KL Divergence can be seen as the distance between two distributions, although it is not symmetric.

$$KL(p, r) = \sum_w p(w) \log \frac{p(w)}{r(w)} \tag{2.4.8}$$

8. Jensen-Shannon divergence

   The Jensen-Shannon divergence is a symetrized and smoothed version of the KL divergence above.

$$JS(p, r) = \frac{1}{2} KL(p, \frac{1}{2}(p + r)) + \frac{1}{2} KL(r, \frac{1}{2}(p + r)) \tag{2.4.9}$$

9. Pearson Correlation

   In collaborative filtering, correlation is often used to predict a feature from a highly similar mentor group of objects whose features are known. The [0,1]-normalized Pearson correlation is defined as

$$s^{(P)}(x_a, x_b) = \frac{1}{2}\left(\frac{(x_a - \overline{x}_a)(x_b - \overline{x}_b)}{\parallel x_a - \overline{x}_a \parallel_2 \cdot \parallel x_b - \overline{x}_b \parallel_2} + 1\right) \tag{2.4.10}$$

   where $\overline{x}_a$ denotes the average feature value of over all dimensions.

10. Language Model Perplexity

    Given the n sentences in the domain sample $(S_1 \ldots S_n)$, we can calculate the perplexity of the language model trained on the parallel corpus p. The perplexity is defined as:

$$Perp = 2^{-\frac{1}{|p|}\sum_{i=1}^{n} \log P(S_i)} \tag{2.4.11}$$

    One important issue that we need to take into consideration is that the perplexity of a language model depends on its application domain. There is generally higher precision (and less ambiguity) in specialized fields [60].

Similarity Aggregation: Mean Reciprocal Rank

Once the similarity is computed, there are two possible views of the domain sample: either as being represented by the **centroid** of its constituent datapoints, or as being represented by the datapoints themselves. In the second approach, identified by **MRR** in the results chapter, a candidate training instance (i.e. the parallel corpus sentence being evaluated) is selected according to its nearest neighbors in the domain sample. We use the mean reciprocal rank over all domain sample datapoints in order to score a candidate (hence the MRR name). Each candidate data point c_j in the training data is scored, and the top candidates are subsequently selected. Intuitively, the score of a candidate encodes the aggregate usefulness/similarity to each of the domain sample data points (i.e. sentences) d_i.

Figure 2.3 provides the algorithm for aggregating the similarity scores between each candidate and the domain sample.

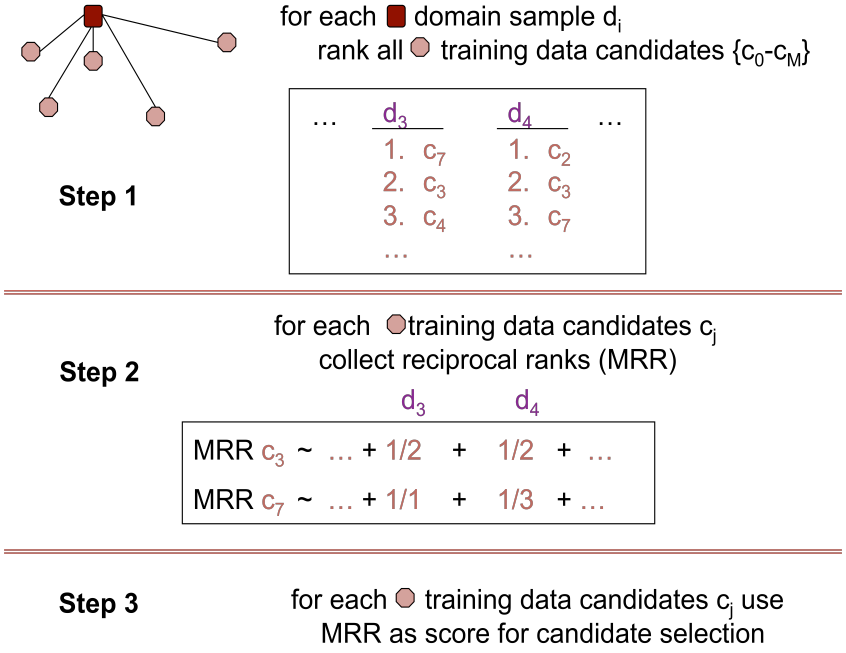

Figure 2.3: Mean Reciprocal Rank Calculation

We define the relevance (similarity) of training candidates C to a sentence in the domain sample d_i by ranking each candidate c_j according to its distance to d_i, $rank_{ji}$. The candidate score is then the reciprocal rank: $score_{ji} = 1/rank_{ji}$.

In the second step, we compute the overall score for each candidate given the entire domain sample D. The mean of all reciprocal rank scores is used as the aggregate score:

$$score_j = \frac{1}{|C|}\sum_{i=0}^{|D|} score_{ji} = \frac{1}{|C|}\sum_{i=0}^{|D|}\frac{1}{rank_{ji}} \tag{2.4.12}$$

, which is known as the mean reciprocal rank (MRR).

Finally, once an overall MRR score is computed for each candidate, the selection process focuses on the highest-scored top N candidates which are most relevant to the overall domain sample:

$$C_s = \{c_j = \arg\max_{c_j} score_j\}, where |C_s| = N \tag{2.4.13}$$

.

2.4.2 Translation Quality Estimate (TQE)

A monolingual corpus can vary in quality due to noise, speech recognition or OCR errors and encoding issues. A parallel corpus has the added possibility of poor translations, misaligned sentences, or completely wrong sentences undetected by, for example, a parallel corpus web mining software. In our case, parliament proceedings (carefully translated by humans) result in a high-quality parallel corpus, leading to more accurate translation models. In contrast, noisy corpora collected from the web, or comparable corpora known not to be exact translations of each other, lead to noisier translation models. This criterion is orthogonal to the domain match (or word-based domain similarity), and we need to find suitable tradeoffs when weighting the two.

There are several approaches to automated translation quality estimation, such as translation equivalence [63], length ratio variance, and bootstrapping+evaluation (described below). We have used length-ratio variance as a step function followed by bootstrapping and evaluation. In preliminary experiments, this method proved more stable than using translation probability stability (below) as a translation quality estimate. Manual approaches (such as corpus-level metadata based on the source of the translation, or using bilingual speakers to estimate a sample chosen for each corpus) can also be utilized but have not been the focus of this thesis.

Length/Ratio Variance as Translation Quality Estimate

When translating from a language to another, the ratio of the number of words in the original text vs. the translated text tends to remain constant for a certain language pair. We calculate the variance over the entire collection:

$$\frac{1}{q_c} = \sigma^2 = \frac{\sum_{i=1..|C|} (\lambda_i - \mu_\lambda)^2}{|C|} \tag{2.4.14}$$

where:

$|C|$ = the size of the collection in sentences or documents

λ_i = the ratio of the lengths (in words) of the i-th sentences in each half of C

μ_λ = the mean of λ_i in C

A large variance might indicate problems such as missing parts on either side of the translation, parsing or processing mistakes that occurred at the time of the corpus collection etc.

In the past, using this criterion as a filter did not lead to a significant improvement [57], but has been used in practice [47, 56].

Translation probabilities stability

We can examine how the term-to-term translation probabilities change when a random selection of documents is eliminated from the training corpus.

$$\frac{1}{q_c} = \frac{\sum_{i=1..K} (\partial_i - \mu_\partial)^2}{K} \tag{2.4.15}$$

where K is the number of folds/turns in eliminating documents, and

$$\partial_k = \frac{\sum_{i=1..|V_e|, j=1..|V_f|} (p(e_i|f_j) - p_k(e_i|f_j))}{|V_e| \times |V_f|} \tag{2.4.16}$$

A large change can indicate inconsistent translations and a lower parallel corpus quality.

Bootstrap and Evaluation

This is the method later used in Chapters 4 and 6, partly due to its ability to leverage existing general domain trained models. It consists of two steps: first, half of the parallel corpus is translated using another parallel corpus - if available, even if it is domain-mismatched - or itself. Then, an automatic MT evaluation measure such as BLEU (including modified BLEU) or NIST is applied to each sentence. This method has the disadvantage of using measures that have been shown to be unreliable at the sentence/segment level. However, according to the JHU MT workshop report [50] examining solutions for evaluation at the

sentence level, no alternative features are fully reliable at this granularity level. This method has the advantage of catching misaligned sentences that a length-based estimate might miss, as well as those the estimate does identify.

2.4.3 Size or Sentence Length

The size of the parallel corpus used for training has been shown repeatedly to affect performance in applications such as MT and CLIR [46]. It follows that size is a natural choice for the selection of a parallel corpus as a training resource. Anecdotaly, size has been used as the only criterion for the selection of training resources, since it implies a bigger known vocabulary size and more training examples, which affect the robustness of the trained model. However, size alone is not a good predictor of whether a particular parallel corpus would lead to the best performing translation model, even if the two corpora are in the same general domain [46].

The size of the selected sentence is important in that smaller sentences add little additional information, but overly-long sentences lead to less sharp co-occurrence probabilities used for the translation model. The size criterion can be used either through its value, or as a thresholding measure. In the experiments presented in this thesis, the size criterion has been tempered by the log function before combined with the other criteria.

2.4.4 Genre

Genre information can be extracted from a) metadata for the parallel corpora, and b) automatic classification by constructing a vector representation of a corpus projection into genre space. Genre classification has been the object of several studies [21]. However, it is not clear to what extent other features such as vocabulary match are already capturing the genre specific characteristics.

Examination of a corpus produced during our previous work showed that such features are being indeed captured (i.e. documents selected from the TED corpus in order to be used to translate spoken presentations showed spoken language characteristics). The results

presented later in the thesis do not identify genre as a separate dimension, since the genre of the data being used is homogeneous. Heterogeneous genre corpus exploration is left to future work.

2.4.5 Taxonomy Representation and other Meta-Data

Other criteria that have the potential to be used for domain adaptation depend on the specific metadata (if any) available. For example, PubMed makes MeSH categories available for its articles and abstract. When taxonomy information is missing for some corpora, classifiers trained on the corresponding taxonomy can be used to project these corpora into the multi-dimensional space defined by using the taxonomy as dimensions. This representation has the advantage of incorporating several sub-domains of the target corpus, along dimensions that humans deemed important.

Our previous work in this area involves using MeSH categories to represent documents in the medical domain, effectively using the taxonomy as an interlingua for cross-lingual information retrieval.

2.4.6 Redundancy

The quality of a parallel corpus and its associated segments also depends on its degree of redundancy. If a large quantity of data selected from a redundant corpus is itself redundant, its utility is diminished regardless of how closely it matches the domain.

Redundancy is traditionally measured by examining the lexical level similarity between the previously selected (or higher ranked) parallel sentences and a new candidate sentence. Using this criterion has the practical disadvantage of imposing an order on the selection process (i.e. two-staged re-ranking). A certain degree of redundancy is, however, needed by the statistical translation modules. We leave the task of fully exploring the effects of this criterion to future work.

2.5 Adaptation Criteria Combination

The criteria values are not probabilities and they are not independent. However, they can be normalized to [0, 1] in order to allow their combination. (Weighted) arithmetic average acts as a soft OR and allows i.e. misaligned sentences (poor TQE) to have a high score if i.e. the domain match is high, which is not a desirable effect. However, weighted geometric and harmonic mean give us the desired soft-AND effect; they have been used in the experiments presented in this paper.

While the simplest approach is to use equal weights for all criteria, the problem of assigning the relative importance to each criteria when ranking parallel sentence candidates requires optimizing a highly non-linear, non-convex multi-dimensional function. Since multiple local optima may exist, we require a global constrained optimization method. Global optimization strategies include branch-and-bound methods (however, most have the disadvantage of relying on information about the problem structure or on the availability of an analytic formulation), bayesian partition algorithms (where a prior on the problem dimensions is needed), genetic algorithms, adaptive stochastic search (e.g simulated annealing, which places a non-zero probability on moving away from the optimum) etc. Many of the above methods suffer from two main drawbacks: requiring a (fast) calculation of the objective function gradient, requiring smooth continuous functions or the availability of a formula, and/or requiring too many objective function calls. Since our particular function (corpus domain adaptation, then CLIR) is a fairly time-consuming black box, minimizing the function evaluation calls and not having to provide a gradient are important considerations.

Since it provides a satisfactory answer to the two considerations mentioned above, our method of choice is continuous reactive tabu search (CRTS) [2], as described below and in Chapter 7 .

2.6 Continuous Reactive Tabu Search for Criteria Optimization

We are optimizing CLIR average precision $f : \omega \rightarrow R$, where ω is the set of feasible points and a subset of R^n, defined by bounds on the n weights $w_i : 0 \leq w_i \leq 1$. Our function f's convexity and differentiability cannot be relied upon, therefore algorithms such as simple hill climbing are not recommended. We use CRTS [2], a global, deterministic, tabu-search based optimization method that uses the reactive affine shaker (RASH) [6] algorithm as its local optimization routine. CRTS's combinatorial optimization algorithm focuses on locating the set of promising boxes in the search space, and it initializes RASH while adapting box size and other search parameters.

More details on RASH and CRTS, as well as experimental results can be found in Chapter 7.

2.7 Online versus Offline Adaptation

The most common approach, in both language model and translation model domain adaptation, is to use information retrieval methods to find the nearest neighbors in a particular corpus for each of the test sentences. Then, a mixture model is built from a) the general language model or translation model and b) the adapted, domain-specific model (i.e. [46, 14]). We refer to this scenario as *online adaptation*; such adaptation permits very specific and accurate adaptation tailored to a specific test sentence or query. However, both the language model and translation model adaptation are time-consuming, as they involve searching and model retraining. One can argue that retraining the models when presented with a test sentence is not a realistic scenario. We therefore compare it with and explore the impact of what we refer to as *offline adaptation*: adaptation when the domain is known and a domain sample is available. We remind the reader that a ***domain sample*** is defined as a set of unlabeled datapoints in the target domain. In both the cross-lingual information retrieval

and the machine translation scenario, a domain sample is a collection of monolingual sentences that represent the target domain. The quantity and quality of the domain sample does affect the adaptation quality - but obtaining such data is not difficult, given that it need not be parallel. For example, monolingual technical manuals or monolingual medical articles are abundant.

One of the main advantages of domain adaptation is the significant reduction in the amount of training data necessary. When the adaptation is performed offline, adaptation can be seen as an active learning scenario, in that the selected data can be translated at a much smaller cost that an entire corpus. The sentence selection, in this case, can be performed out of available monolingual corpora (with criteria such as translation quality taken out of the equation) The traditional active learning labeling of examples is, in this case, the translation process itself. Similarly, re-training can be performed less often and requires fewer resources. Online adaptation, on the other hand, does not allow for the manual translation of monolingual sentences and is therefore less suited to an active learning scenario.

Chapter 3

PARDA for Statistical Machine Translation: System and Data

3.1 Overview

After introducing the PARDA framework in Chapter 2, we now focus on the two multilingual, corpus-based tasks we chose as its applications. The first task is statistical machine translation (MT), more specifically the corpus-based methods that rely on parallel corpora as training data in order to tackle the task of automatically translating written text from one language to another. We start by describing the datasets used for this task, as well as the MT system details. In the following chapter, we analyze the effects of automatically selecting the best training data subset, which is subsequently used by our corpus-based MT system.

In addition to showing the significant effect that selection has on the amount of parallel/training data necessary to produce satisfactory MT results, we address the experimental differences between a given domain (in our case, we use data in the medical field), and a sub-domain (in our case, the heart-related medical domain subset). We also contrast a) medical-domain adaptation results with results obtained using a general-domain corpus, and b) on-line and off-line adaptation.

We explore the effect of several individual RDA criteria: lexical-based mean reciprocal

rank (MRR), translation quality estimate (TQE), sentence size, as well as the taxonomy projection described in 2. We show that, while the MRR criterion is crucial, the addition of quality estimates and categories improves the MT results.

3.2 Phrase-Based Statistical Machine Translation: System Details

Although for CLIR we are building our own high performing system (discussed in Chapter 5), we have taken a different approach when a Statistical Machine Translation system was needed. We have used freely available, off the shelf, components in order to assemble our phrase-based Statistical MT system. In order to obtain state-of-the-art performance we have experimented with several baseline MT choices, using components such as GIZA++ [51], the CMU-Cambridge Statistical Language Modeling kit, the SRI Statistical language Modeling Toolkit, the ISI decoder, Pharaoh Phrase-based decoder, and the Carmel finite-state transducer kit.

The combination used in the experiments presented here is GIZA++ [51] and Pharaoh [53], without the costly minimum error rate training for parameter estimation [49]. The language model we use is a general English language model; we do not perform language model adaptation. This general purpose system has been trained on the European Parliament English-French corpus, tuned on a development set from the same corpus, and tested on a 2,000 sentences test set provided by the 2006 Statistical Machine Translation Workshop organizers [37]. This allows us to establish a system baseline by directly comparing the results with 2005 state-of-the art systems.

Figure 3.1 shows the performances of top 5 MT systems; the black bar is the baseline for the system we used. We present this high baseline in order to establish the system as a competitive one, and attribute poor performance in subsequent experiments to the test & training data (more specifically, to its domain mismatch).

The most popular evaluation metrics (and the ones we use in this thesis) are BLEU[52]

and modified BLEU (MBLEU) [66]. These evaluation metrics are fundamentally different from the CLIR metric in that they take into account the translation accuracy and the fluency of the resulting text when evaluating the quality of system-produced output. MBLEU is different from BLEU in that the n-gram level scores are combined using arithmetic mean instead of geometric, allowing more stable scores in short MT samples (e.g. sentences and paragraphs) when long n-grams do not match . Due to the n-gram influence in evaluation metrics, one possible similarity criteria to incorporate for MT is n-gram-level similarity (as opposed to unigram only). In this respect, MT differs from the other two application areas, in that the final fluency and grammatical correctness of the output is a factor in the evaluation. The evaluation metric used for the rest of the chapter is the BLEU score [52]; similar trends and effects are obtained when using modified BLEU.

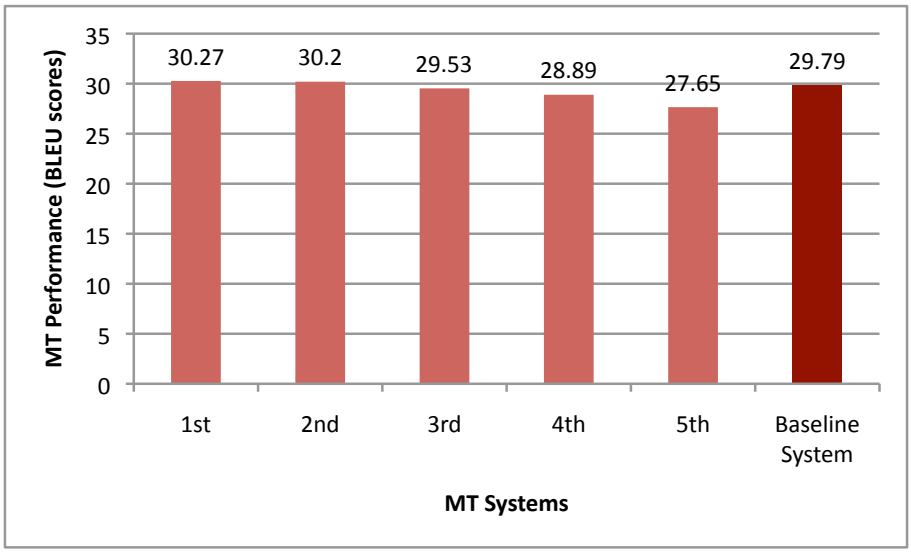

Figure 3.1: Top 5 MT workshop systems results (BLEU) on Europarl data (left) vs. our GIZA++ with Pharaoh baseline system (right, in black), demonstrating its competitive performance.

3.3 Resource Domain Adaptation for Machine Translation

3.3.1 Datasets

MedTitle and MedTitleTest

One of the domain-specific corpus we are using is **MedTitle**, a collection of 505,000 titles of French medical articles, together with their English counterparts. The collection has been mined by the author from the PubMed online database, by querying for French articles for which the English translation of the title is available. **MedTitleTest** is the test set used in conjunction with MedTitle, and it contains 1,000 titles in each language. **MedTitleTest** is also parallel, and it is a randomly selected (and excluded) subset of **MedTitle**. The French half of **MedTitle** is used as input to the MT system, and the English half is used as the evaluation gold standard. Both **MedTitleTest** and **MedTitle** are parallel.

For these and the following datasets, we refer to each title as a *sentence* for consistency and comparison with other parallel datasets, although each title is also its own *document*, complete with its own metadata.

MedCat

A similar and highly overlapping parallel collection we have created is MedCat. Using the same methodology as in the MedTitle case, we have also included the MeSH [48] categories where available, and excluded articles for which taxonomy information was not available. This sentence-level metadata allows us to experiment with a "perfect" (human-quality) classifier using an extensive and established taxonomy. This method produced 423,702 parallel sentences and their corresponding categories.

MedHeart: A Subdomain Parallel Dataset

While MT adaptation to the medical domain in general is a crucial goal of this chapter, we also plan to examine the effect adaptation has when the domain in question is more focused.

We choose the ***Heart*** top-level subdomain, as identified by the MeSH categories. Out of the 403K sentences in the MedCat collection, approximately 30K were identified by their MeSH category as belonging to the ***Heart*** subdomain. After randomly selecting 1,000 of these as the test sentences, and subsequently deleting them from the MedCat collection, we use the remaining sentences in the MedCat collection as the pool of sentences to select from. A few of the 1,000 test sentences are shown in Table 3.1.

MedTitle-IT

We have also downloaded 177,000 Italian/English titles for which the French counterpart was not available (MedTitle-IT). We are using this corpus as a domain sample when the untranslated test set is not available, as is the case in offline adaptation.

Med* Validation and Testing

Our domain-specific test and validation set consists of 1,000 sentence pairs each, randomly selected and removed from the MedTitle (or MedCat, where applicable) collection. The reference translation were the parallel counterparts. Although we are aware not using multiple reference translations can be problematic, securing multiple translations for a technical domain is costly and is reserved for research focused on this effect.

Europarl

Another parallel corpus that has extensively been used for statistical machine translation training is Europarl [36], a collection of English/French sentence-aligned proceedings of the European Parliament. The domain mismatch between Europarl and MedTitle is significant - they differ in genre, vocabulary and even sentence length distribution. As we have seen in [46], this domain mismatch yields to an extremely poor performance when an Europarl-trained MT system is tested on MedTitle.

English (reference) translation of sampled sentences in the MedHeart test set
angioedema associated with the use of dihydropyridines
hormone replacement therapy and cardiology : don ' t dream !
syncope
diagnosis of coronary insufficiency in diabetics . when , how , why ?
calcium antagonists and arterial hypertension
anatomo-clinical conference . inflammatory syndrome , glomerulopathy and
abnormal pulmonary images in a patient with a starr ' s valve
myocardial infarction without q wave . clinical course characteristics and therapy
anterior interventricular revascularization using the internal mammary artery.
short and medium-term follow-up of 140 patients
prognostic value of the normalization of the exercise test under medical
treatment in coronary insufficiency
prognostic factors of hypereosinophilic syndrome . study of 40 cases
late effect of intracoronary urokinase . apropos of a case of recurrent
coronary thrombosis after angioplasty
vascular risk factors
quantitation of left ventricular function after an inferior or superior
myocardial infarct . comparative value of resting ejection fractions or after
effort attenuation of cardiocirculatory reactions induced by the ablation of
acoustic neurinomas (trans-labyrinthine approach)
left ventricle in noonan 's syndrome . electro-vecto-echo and angiocardiographic aspects
continuous emission doppler study with spectrum analysis in the evaluation
of aortic stenosis in adults . apropos of 30 cases
percutaneous aortic valvuloplasty by trans-septal approach
fetal cardiology : a new perspective in pediatric cardiology
significance of minor vectorcardiographic changes (the qrs complex)
bidimensional echocardiography in the search for the latent origin of cerebral
embolism

Table 3.1: MedHeart test set sample (20 out of $1,000$ sentences shown).

Dataset Summary

The pre-processing was identical for all data sets, after first eliminating PubMed-specific noise such as HTML tags etc. The character encoding was consistent across datasets. Table

3.2 summarizes the dataset properties[1]:

Dataset	**Size**[1]	**Genre**	**Source**
MedTitle	505K	Medical Article Titles	PubMed – authors' translations
MedTitle-IT	177K	Medical Article Titles	PubMed
Europarl	648K	(Spoken) Parliament Proceedings	Transcripts, professional translations
MedCat	424K	Medical Article Titles	PubMed – authors' translations
MedHeart	29K	Medical Article Titles	PubMed; test data is selected from the "heart" category

Table 3.2: Datasets: domains and characteristics.

[1] *size* is given in sentence pairs

Chapter 4

PARDA for Statistical Machine Translation : Experiments, Results and Discussion

4.1 Overview

After describing the statistical machine translation system used to evaluate the effects of parallel corpus adaptation, we now present the effects of automatically selecting its training data subset.

In addition to showing the significant effect that selection has on the amount of parallel/training data necessary to produce satisfactory MT results, we address the experimental differences between a given domain (in our case, we use data in the medical field), and a sub-domain (in our case, the heart-related medical domain subset). We also contrast a) medical-domain adaptation results with results obtained using a general-domain corpus, and b) on-line and off-line adaptation.

We explore the effect of several individual RDA criteria: lexical-based mean reciprocal rank (MRR), translation quality estimate (TQE), sentence size, as well as the taxonomy projection described in 2. We show that, while the MRR criterion is crucial, the addition of quality estimates and categories improves the MT results.

4.2 System Performance on Domain-Mismatched Data

We revisit results shown in [46], using the dataset described in Section 3.3.1. Figure 4.1 illustrates how the well-performing Europarl-trained system fails dramatically when out-of-domain.

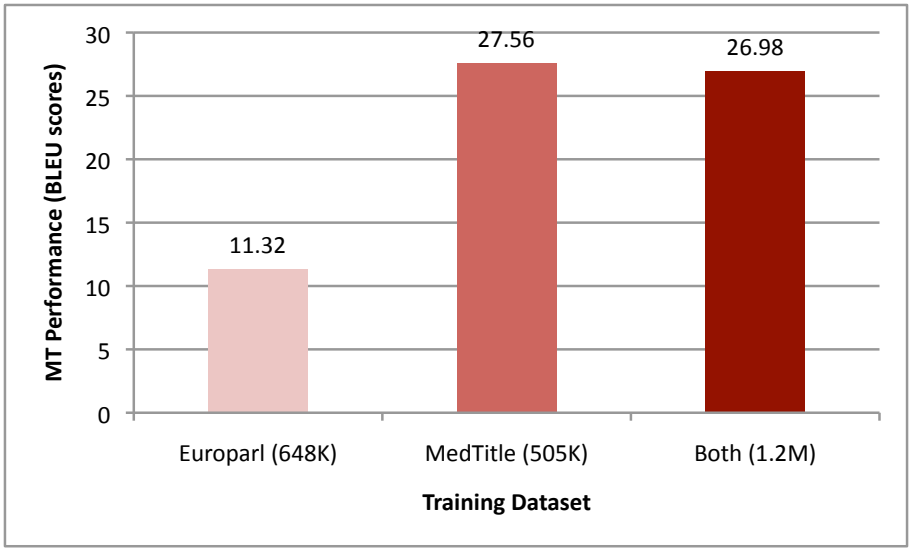

Figure 4.1: Translation performance (BLEU) after training on standalone corpora and testing on MedTitle-Test. Note that in-domain data performance is slightly better than combining it with out-of-domain data and more than doubling its size.

This result is significant in the motivation for domain adaptation not only as an optional enhancement leading to slight improvements, but crucial to effective performance. In order to allow the reader to get a better feel for the translation quality difference, we include a few translation examples in Table 4.1.

The italicized portions indicate where an additional translation reference would have been helpful to distinguish correct (but not exact) translations from actual errors.

4.2.1 MRR vs. Centroid Adaptation: Implementation Details

For both online and off-line adaptation, we calculate the similarity of each candidate sentence in the parallel corpus to the domain sample. As discussed in Chapter 2, similarity measures

Source	Text
Reference	sequencing of adjuvant treatment after surgery for invasive breast cancer : recognize the fragility of the patient
Trained on MedTitle	sequence of adjuvant treatments after surgery breast cancer : recognize the fragility of the guards
Trained on Europarl	sequence of treatments willing after surgery breast cancer : recognize the fragility of provides treatment
Reference	6 neurosyphilis cases : *value of cerebrospinal fluid analysis*
Trained on MedTitle	6 neurosyphilis : the contribution of the study of the cerebrospinal fluid
Trained on Europarl	six neurosyphilis : contribution to the study of liquid cephalorachidien
Reference	*airway remodeling*
Trained on MedTitle	bronchial remodeling
Trained on Europarl	the re-drawing bronchique
Reference	nystagmus and vibration test research of mechanisms , theoretical methods
Trained on MedTitle	nystagmus and test vibratory research on mechanisms , theoretical approach
Trained on Europarl	nystagmus and test vibratoire research on the mecanismes , theoretical approach

Table 4.1: The effects of training on a mismatched domain datasets: Examples.

vary, starting with word overlap, cosine similarity and ending with the language-model information retrieval measure of the probability of a segment being generated by the domain sample. Our results are shown using the latter, as implemented by INDRI [64], without pseudo-relevance feedback and using Dirichlet smoothing.

Once the similarity is computed, there are two possible views of the domain sample: either as being represented by the **centroid** of its constituent datapoints, or as being represented by the datapoints themselves. The second approach is described in detail in Section 2.4.1.

4.2.2 MRR vs. Centroid Adaptation: Results

The two approaches described in Section 2.4.1 (distance to one centroid vs. the mean reciprocal rank of the nearest neighbors of each domain sample instance) are compared in Figure 4.2. This figure covers the MedTitle dataset. Note that using the entire dataset yields the best results using MRR, but here we are interested in comparing the centroid and MRR conditions when the quantity of available data is smaller.

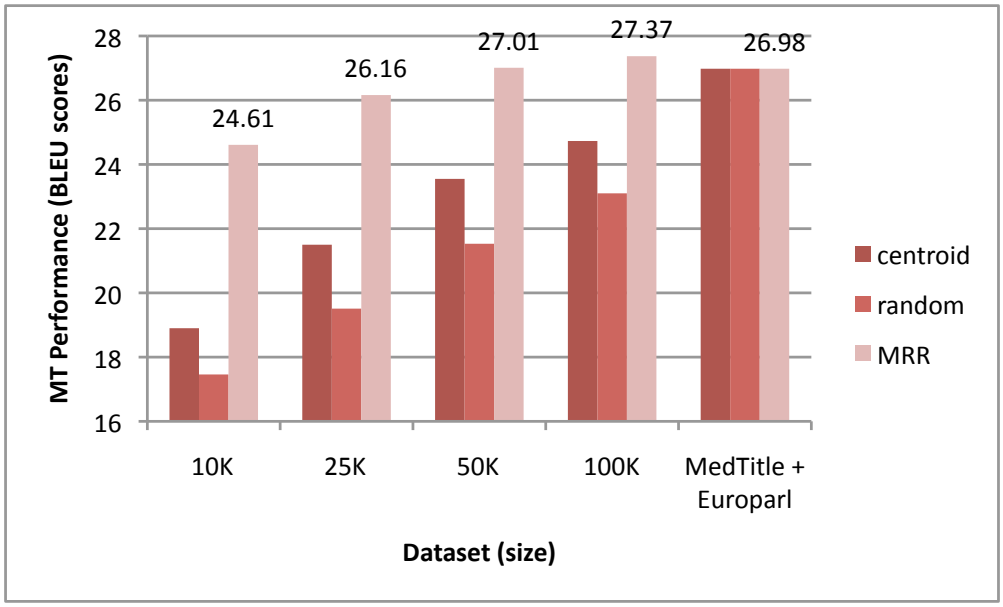

Figure 4.2: MT adaptation results on MedTitle (BLEU Score). We compare the one centroid and MRR approach at various sentence selection levels. A 1.4 difference in BLEU score is statistically significant.

Figure 4.2 shows how a 24-fold reduction in training examples has identical results to using all available data, and a 12-fold reduction yields to a performance improvement. To establish statistical significance, we have used the bootstrap resampling method described in [54]. With a 1,000 sentences test set, a difference of 1.4 in BLEU score was statistically significant (p value ≤ 0.05)

In Figure 4.3 we examine the tradeoff between the more realistic offline adaptation and online adaptation discussed in Section 2.7. We note that, in the case of offline adaptation, there is a performance cost to be paid in exchange for the reduction in training data. This

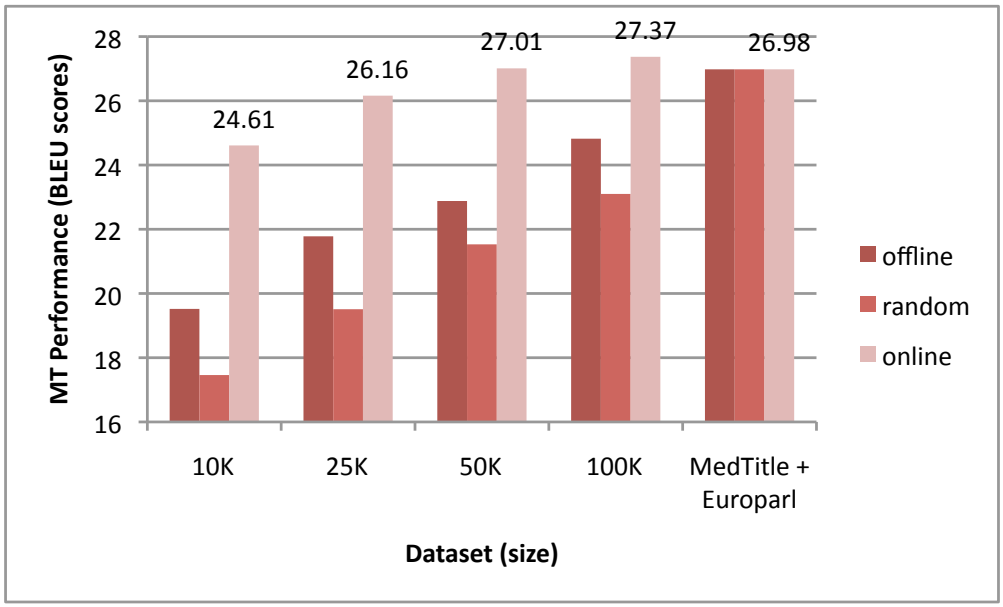

Figure 4.3: Online vs. offline MT adaptation results (BLEU scores) on MedTitle. We are using MedTitle-IT as the offline domain sample. Note that offline adaptation is still improving at 100K parallel sentences, but online adaptation appears to be saturating.

cost is alleviated by the use of additional criteria. We find the offline adaptation scenario more realistic, even more so when our experiments indicated that a large monolingual domain sample (177K sentences) had virtually identical performance with a 1,000-sentence domain sample randomly chosen from the 177K sentences.

4.2.3 The Effect of Individual Criteria on MT Adaptation

One-criterion Adaptation

To examine the effects of adaptation itself, we started by testing the RDA-MRR criterion (the criterion that measures lexical simlarity as described in 2.4.1). We also included an experimental setting that only used the taxonomy information (i.e. the MeSH categories), in order to properly separate each criterion effect. Both are shown in Figure 4.4, together with a baseline of sentences randomly selected from the MedCat training subset. As expected, the RDA-MRR criterion performed well, significantly reducing the number of sentences needed to reach close to peak performance.

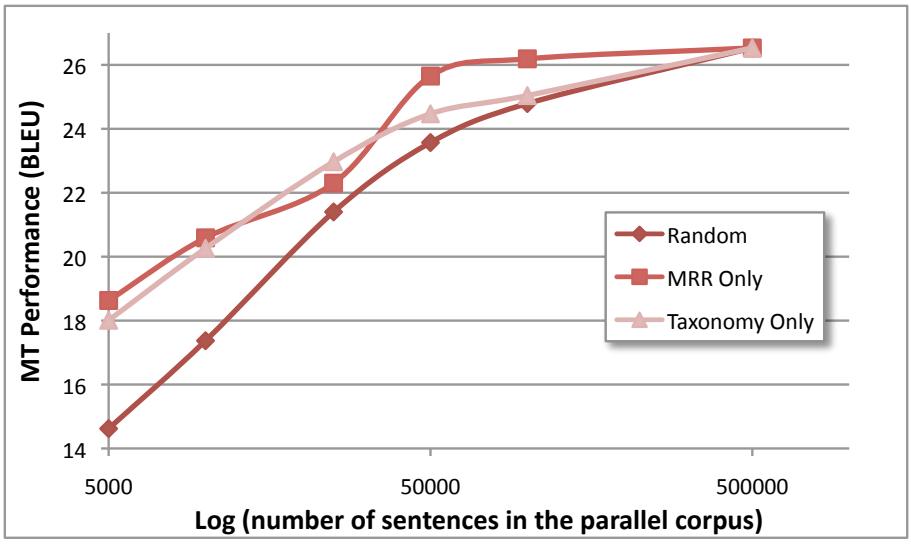

Figure 4.4: One criterion MT adaptation effect on the MedCat test dataset (BLEU score, semilog plot). Sentences are selected from the MedCat training set.

We notice significant improvement when RDA-MRR adaptation is used versus random selection of the training data. This effect is more pronounced at the smaller dataset levels, where it can be argued that adaptation performance has the most impact.

Similarly, we examine the case where selection is performed using the MeSH categories. When this taxonomy information is used alone to represent both the domain sample and the sentences used for selection, the selection has a similar effect to the RDA-MRR selection when the number of selected sentences is low. However, when the number of sentences is higher, the MRR criterion outperforms the taxonomy criterion, due to its richer representation power (i.e. more numerous and more diverse features).

Figure 4.5 examines the same one-criterion conditions for the heart subdomain (Med-Heart, described in Section 3.3.1.) Here, the adaptation effect is even more dramatic: the difference between randomly selected sentences and RDA-MRR selected sentences is astounding (7 BLEU points at the 5,000 sentence level - MRR increases the BLEU score from 17 to 24). Moreover, the MT performance when the training data is reduced by two orders of magnitude is reduced by less than 10%, even when the selection here is done in the medical corpus as it is in Figure 4.5. The gap between the MRR-only method and the taxonomy-only

method is also further increased.

Since the taxonomy and vocabulary-based representations of the same datapoints both lead to good adaptation performance, the next subsection examines the effects of their combination.

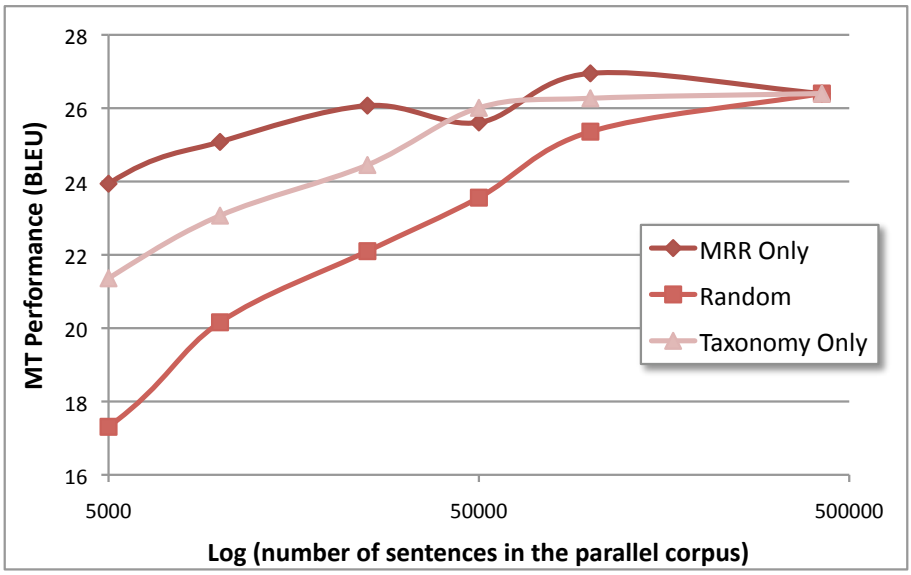

Figure 4.5: One criterion MT adaptation effect on the MedHeart dataset (BLEU score, semilog plot). On this more specific dataset, even 1K properly selected sentences lead to performance close to that obtained when training on 500K sentences. Sentences are selected from the MedCat training set.

Taxonomy/Categories as an Adaptation Criterion

The MedCat dataset is designed by taking advantage of the availability of human-assigned MeSH categories. Since Figure 4.4 shows that adaptation based on both the taxonomy representation and vocabulary representation improve MT performance, we combine the two criteria, as well as add the TQE criterion where applicable. The TQE criterion is added by including it in the pool of criteria to be combined as described in Section 2.5. Not that the TQE criterion cannot be used alone - well translated but irrelevant sentences are not helpful in a domain adaptation task. Figure 4.6 implies that adding the taxonomy information does improve the performance; however, the best performance was obtained when all the above

three criteria were used. To establish statistical significance, we have used the bootstrap resampling method described in [54]. With a 1,000 sentences test set, a difference of 1.4 in BLEU score was statistically significant (p value ≤ 0.05).

A similar result was obtained on the MedHeart dataset, in Figure 4.7. Table 4.2 shows the top selected sentences when all criteria are combined. One interesting observation is that, while the performance is similar at the 1K sentence level, only 33% of the sentences are common between the RDA-MRR condition and using all 3 criteria. This number is even lower for the top 100 sentences (19%); however, using 100 sentences to train an MT system is not realistic.

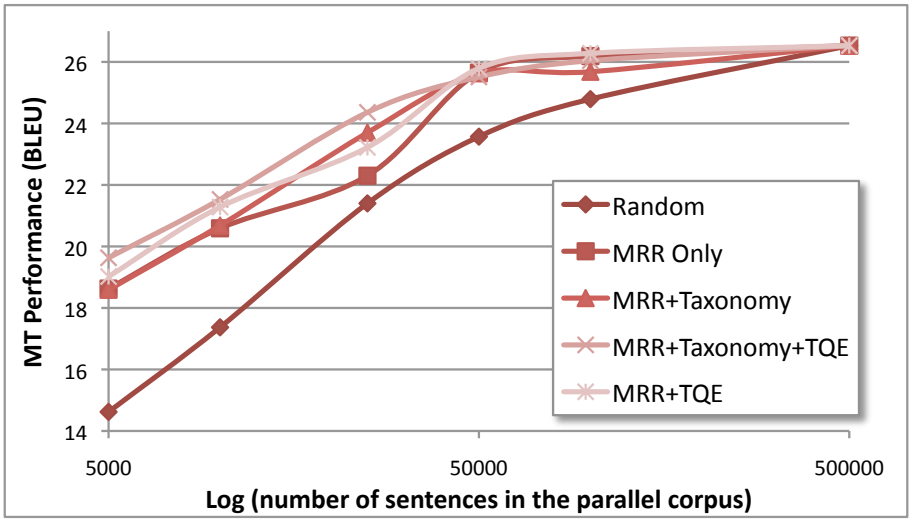

Figure 4.6: The effect of the taxonomy criterion on the MedCat dataset. The translation quality estimate criterion is included for completeness. Sentences are selected from the MedCat training set.

Translation Quality Estimate (TQE) as an Adaptation Criterion

The effect of adding the translation quality to the MRR and Taxonomy-based criterion is shown in Figure 4.8 and Figure 4.9 for the MedCat and MedHeart datasets, respectively. The improvement over the taxonomy-only case is larger than over the MRR-only case; both improvements are small. As discussed in Chapter 6, this small improvement is expected in

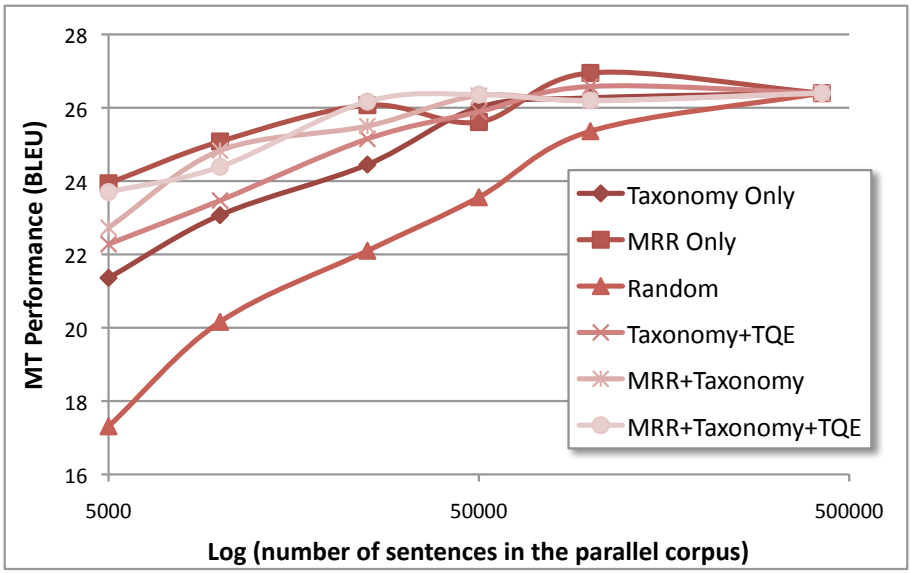

Figure 4.7: The effect of the taxonomy criterion on the MedHeart dataset. Sentences are selected from the MedCat training set.

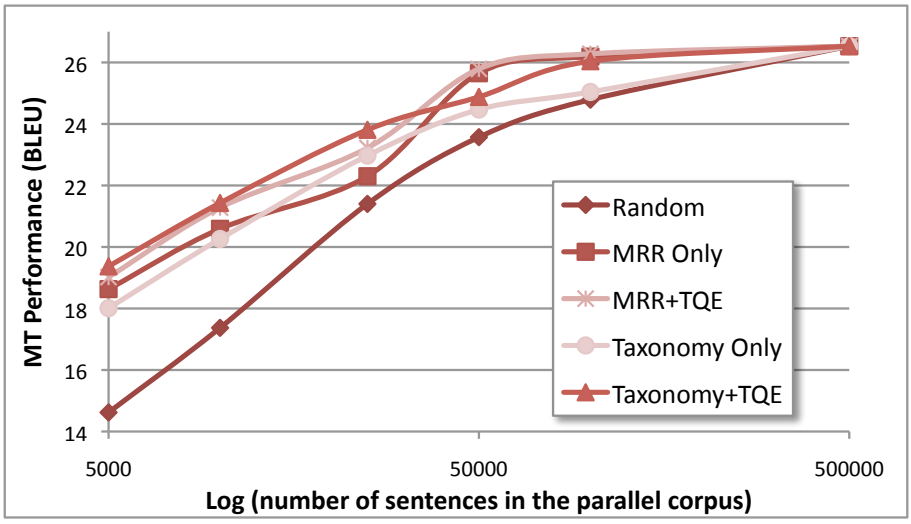

Figure 4.8: The effect of the TQE criterion on the MedCat dataset. Sentences are selected from the MedCat training set.

a corpus where translation quality is consistent from sentence to sentence.

English half of the selected parallel pair
valvular prosthesis in children . indications and results
hemodynamic effects of milrinone in the treatment of cardiac insufficiency after heart surgery with extracorporeal circulation
paroxysmal supraventricular tachycardia with complete atrio-ventricular block or dissociation
circulatory assistance with double-purpose tube of aspiration in the left ventricle and reinjection in the aorta introduced in the apex of the heart by sub-xyphoid abdominal approach
transluminal coronary ngioplasty : immediate and short-term results . apropos of 302 dilated vessels
endocavitary fulguration in the treatment of ventricular tachycardia complicating myocardial infarction
surgical plasty of the coronary trunks : an alternative to bypass techniques
effects of molsidomine during the cold test in stable coronary insufficiency under beta-blocker treatment
automatic analysis of polygraph recordings of sleep
antihypertensive action and inhibition of tissue conversion enzyme by ramipril , perindopril and enalapril in the spontaneously hypertensive rat (shrsp)
post-infarction anterior left ventricular aneurysms . echocardiographic and hemodynamic study of the nonaneurysmal contractile zone
effects of epinephrine upon the circulatory system of the 17-day-old rat fetus .
emergency coronary surgery after transluminal angioplasty . immediate results and long-term outcome of 100 operations
rhythm disorders in the acute phase of myocardial infarct and their treatment
results of radiofrequency ablation of the atrioventricular junction in patients with refractory atrial arrhythmia and severe impairment of the left ventricular systolic function

Table 4.2: Top selected MedHeart sentences using the Springer queries as the domain sample. The criteria used here are RDA-MRR+TQE+Taxonomy.

The Use of Sentence Length as an Adaptation Criterion

To examine the effect that selected sentence length has on machine translation quality, we used the top 100K domain-specific sentences, selected according to the RDA-centroid criterion. We have divided them into 4 approximately equal bins based on sentence length,

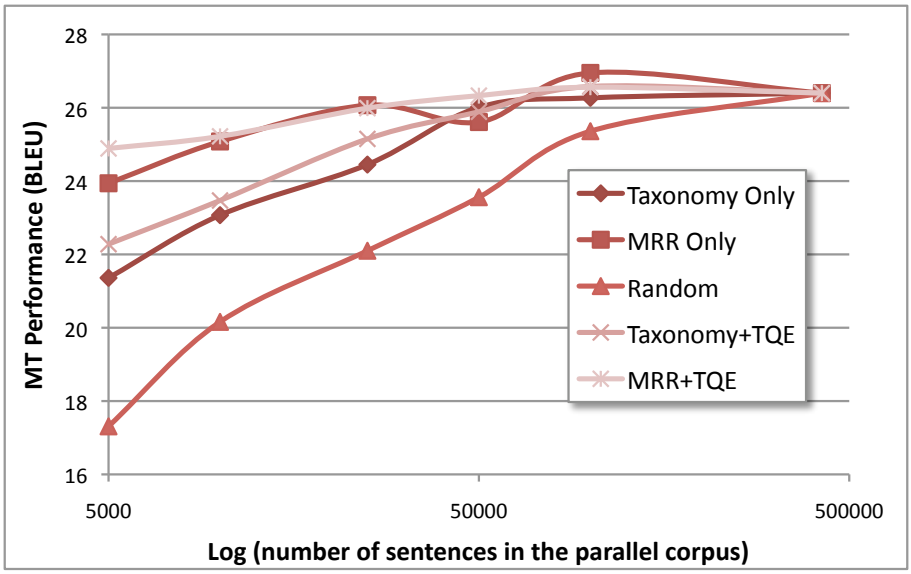

Figure 4.9: The effect of the TQE criterion on the MedHeart dataset. Sentences are selected from the MedCat training set.

then truncated the bins to bring them to the same size. This has resulted in 20K sentences in each size bin.

As in the case of sentence length, we noticed that adding the TQE criterion performs better than RDA-Centroid alone. These experiments were performed using the centroid approach; the results change when using the MRR approach, on the test set, at the 100K level (shown in Table 4.3)

Criterion	**Online-BLEU**	**Online-MBLEU**
MRR+TQE(BLEU)	26.47	32.03
MRR+TQE(NIST)	26.88	32.47
MRR Only	27.37	32.69

Table 4.3: Using NIST vs. BLEU as translation quality estimates. (MedTitle, 100K selected sentences).

The results in Table 4.3 were obtained selecting a 100K sentence pair parallel corpus. Modified BLEU scores are also presented the effects are similar.

Online vs. Offline Adaptation and Criteria Combination

We combined normalized length, TQE and RDA-MRR using harmonic and geometric means. Weighted geometric and harmonic mean give us the desired soft-AND effect. The criteria were equally weighted for simplicity/generality. Table 4.4 shows that, while the averaging method does not have a significant effect, offline adaptation is improved by adding additional criteria to the RDA-MRR domain match. The already high-performing online adaptation is not improved in the case of the MedTitle dataset and at 100K selection level; however, we have shown improvements at lower dataset size in Figure 4.8

Criterion	**Online-BLEU**	**Offline-BLEU**
Harmonic	26.50	25.46
Geometric	26.64	25.33
Random	23.1	23.1
MRR Only	27.37	24.82

Table 4.4: Combining size, quality and similarity using various combination methods. These are results obtained on MedTitle, at the 100K selection level.

4.2.4 Statistical Machine Translation Adaptation: Conclusions

In this chapter, we explore the effect of several individual RDA criteria on statistical MT performance. We examine mean reciprocal rank (RDA-MRR) and RDA-Centroid, translation quality estimate (TQE), sentence size, as well as the taxonomy projection criterion described in Chapter 2. We show that, while the RDA-MRR criterion is crucial, the addition of quality estimates and categories improves the MT results, especially when using a small to moderate amount of data. Another tradeoff we explore is that between on-line (done at testing time) and off-line (done before testing, assuming the availability of a domain sample) adaptation. We also address the experimental differences between a given domain (the medical field), and a sub-domain (in our case, the heart-related medical domain subset). Here, our selection method allows a two-order of magnitude reduction in training data with only a 10% BLEU-score decrease, vs. a 35% decrease for randomly-selected but in-domain data.

Chapter 5

PARDA for Cross-Language Information Retrieval: System and Data

5.1 Overview

This chapter describes our corpus-based approach to CLIR, presenting system details as well as results in international cross-lingual evaluation forum tasks. By establishing it as a competitive system in this chapter, we proceed to examine it in a multilingual task in the domain-adaptation evaluation shown in Chapter 6. In this chapter, we also specify the two CLIR systems used to examine domain adaptation evaluation tasks, and we introduce the domain specific data used in the adaptation experiments shown in Chapter 6.

5.2 The CLIR task

The Cross-Lingual Information Retrieval (CLIR) problem consists of finding documents in a *target language* that are relevant to queries expressed in a (different) *source language*. In CLIR, system performance is measured by comparing a ranked list of the documents identified as relevant by the system with a pre-determined set of human-labeled relevant

documents. The metric used is *Mean Average Precision* (MAP), which averages the performance over all queries. MAP is also the metric used in monolingual information retrieval. This allows a direct comparison between the cross-lingual and monolingual scenarios, and in many cases the cross-language performance is close to that of the monolingual case. Initially, this has been attributed to the query expansion effect that corpus (or dictionary) based translation introduces - however, this effect is maintained even with monolingual query expansion [12].

As discussed in previous chapters, the CLIR methods we use are statistical methods based on *parallel corpora* - text data presented in two different languages. In the following section we present our implementation of a high-performing CLIR system and its results.

5.3 CLIR Systems

We have built a corpus-based CLIR system, using query expansion both before and after translation. A significant difference from machine translation (MT) or online dictionary based approaches is that instead of using a rank-based cutoff (i.e. the first or first two variants for each word) we are using *all* translations weighted by their translation probability. The most successful CLIR corpus-based approaches combine the translation of queries and documents, or integrate translation in the retrieval models [11][17][18]. Our approach is similar, and it has the welcome side effect of providing a very focused query expansion.

Our approach consists of 4 general steps. Given source (query) language L_1 and the target (document) language L_2,

1. Expand the query in L_1 using pseudo-relevance feedback
2. Translate the query, while preserving the relative weights from 1.
3. Expand the query in L_2 using pseudo-relevance feedback
4. Retrieve documents in L_2

Here, pseudo-relevance feedback is the process of retrieving documents and adding the terms of the top-ranking documents to the query for expansion. We used simplified Rocchio positive feedback as implemented by Lemur [64] - the number of documents used for feedback (20) is based on CLEF multi-language training data. Our corpus-based methods differ only in the translation step, as described below.

5.3.1 Weighted Model 1 (WM1)

IBM's statistical machine translation Model-1 (or simply "Model 1") [4] uses a sentence-aligned training corpus to compute the term-term translation probabilities across two languages. The translation probability from term s (in L_1) to term t (in L_2) is defined as:

$$p(t|s) = \lambda_t^{-1} \sum_a P(S_s, a|S_t) \sum_{j=1}^{m} \delta(s, s_j)\delta(t, t_{a_j}) \tag{5.3.1}$$

where λ_t is a normalization factor, a is an alignment of cross-lingual term-term translation, S is a sentence in the L_1 or L_2 half of the parallel corpus, m is the number of tokens in the sentence in L_1, and the second summation is the number of times s aligns with t in the corresponding alignment.

A matrix of translation probabilities is initialized and updated iteratively (we set the iteration number to 10 in our experiments). We use the resulting matrix to translate each query from L_1 to L_2: for each query word in the source language (L_1), the entire vector of the corresponding target terms (in L_2) is used in the translation, with the normalized probability as the weight of each target term. We named this method "Weighted Model 1" to distinguish it from using only the top target word in the translation of each source word.

Our approach is similar to IBM's and BBN's approaches to CLIR [16] except that the translation is not integrated in the retrieval model; only the query is translated. We found that this method performed well in CLIR benchmark evaluations [45].

5.3.2 Chi-square Statistic (CHI)

Chi-squared statistics are commonly used to measure the dependence between terms and categories in text classification; we are including it as a measure for term-term similarity between a source language term (s) and a target language term (t). CHI measures the dependence between s and t using four counts: A, B, C and D, where A is the number of passages (sentences or documents, depending on how the parallel training corpus is aligned) in which s and t co-occur, B is the number of passages s occurs without t, C is the number of passages t occurs without s, and D is the number of passages where none of them occur:

$$\chi^2(s,t) = \frac{(A+B+C+D) \times (AD-CB)^2}{(A+C) \times (B+D) \times (A+B) \times (C+D)} \tag{5.3.2}$$

This calculation results in a matrix of term-term associations, which we use for query translation in the same manner as the matrix of translation probabilities in WM1. The advantage of this calculation is its efficiency, compared to that of WM1. The effectiveness of CHI vs. WM1 in CLIR is a question worth examining; in [46] we show that the effectiveness is comparable. We also show that point-wise mutual information (PMI, below) is both efficient and effective in our experiments.

5.3.3 Point-wise Mutual Information (PMI)

Point-wise mutual information is another common choice for measuring the empirical association between two variables (in our case, two terms across languages). The metric is defined as:

$$PMI(s,t) = P(s,t) \log \frac{P(s,t)}{P(s)P(t)} \tag{5.3.3}$$

The main difference between CHI and PMI is that PMI measures the positively correlated dependence while CHI counts both the positively and negatively correlated dependencies. With respect to our task, translating a term from one language to another, PMI appears to be a more appropriate measure since we do not want to consider t as a translation of s if

the joint probability of the two terms in human translations is too low. The same argument applies to Information Gain (IG). In terms of computation, the two methods are equally efficient since the joint and marginal probabilities used in computing PMI can be easily derived from the counts of A, B, C and D defined in Section 5.3.2

5.3.4 Weighted SYSTRAN (WSYS)

Although not a corpus-based method, we are including this approach in order to provide a comparison with a general-purpose machine translation system that is used as a strong baseline in standard evaluation benchmarks such as CLEF [12]. We use SYSTRAN online to translate each query after the expansion using local feedback. In order to have a fair comparison, and not put SYSTRAN at a disadvantage, we preserve the term weights before the translation, and propagate the weight of each word to its translations. Post-translation query expansion is also included in the process and is identical to that of our corpus-based methods. Note that, unlike in the case of our corpus-based methods, morphological processing of a query has to be postponed until the query is translated.

5.4 CLIR Systems: Performance and Uses

In order to test the performance of our CLIR system and compare it with other state-of-the-art CLIR sysems all over the world, we have participated in CLEF 2003, a cross-lingual evaluation forum similar to TREC, specialized in multilingual tasks.

In addition to participating in the (non-domain specific) multilingual retrieval tasks, we have included our CLIR system as a module in a cross-lingual question answering task.

5.4.1 Cross-Lingual Question Answering

We have participated in the cross-lingual question answering task at CLEF 2003 [12] by combining our cross lingual IR system with a monolingual QA system [40]. After tuning the combined systems on available question/answer datasets, we focused on participating

in the cross-lingual French-to-English QA CLEF task. The CLIR system produced both a list of relevant documents as well as a translated expanded query with corresponding weights for each word. These were subsequently used by the monolingual QA system by applying a term weighted proximity measure to candidate answers of a type determined by a question classification module. Evidence for a particular answer is then combined with that of identical or similar answers in order to compute its final score.

Our hybrid CLQA system was not a high performing one - however, when compared to other participating systems (at the time, similarly performing), it showed the advantages of a modular approach that allows weighted queries and weighted query translation.

5.4.2 CLEF 2003: Bilingual and Multilingual Results

The retrieval tasks we participated in were two bilingual tracks ($DE \rightarrow IT$, $IT \rightarrow ES$) and the four-language multilingual task. The retrieval task here was not domain specific, which conferred an advantage to commercial general-purpose translation systems. However, our CLIR system's performance in CLEF 2003 was consistently in the top 5.

According to [12], the main CLEF multilingual corpus consists of sets of documents in different European languages but with common features (e.g., same genre and time period, comparable content). The CLEF corpus includes both newswires and national newspapers and most collections cover the period 1994-1995. There were 60 queries, and the relevance judgments were provided by CLEF.

As training data, we have used the European Parliament proceedings 1996-2001 [36]. It includes versions in 11 European languages: Romance (French, Italian, Spanish, Portuguese), Germanic (English, Dutch, German, Danish, Swedish), Greek and Finnish. We have also prepared German-Italian and Italian-Spanish versions for the two bilingual CLEF tasks we participated in. We preprocessed the parallel corpora and CLEF documents by eliminating punctuation, stopwords, and document sections disallowed in the task description. We have used the Porter stemmer for English and the rule-based stemmers and stopword lists provided by J. Savoy [61]. After stemming, we have used 5-grams as a substitute for German word

decompounding. Detailed parameter settings and further analysis can be found in [45].

Figures 5.1 and 5.2 show the system performance compared to other participants in the multilingual and bilingual task, respectively. among the systems in this evaluation, using IBM Model 1 to build the translation matrix used in the query vector multiplication is specific to our system. This performance establishes our CLIR approach as a competitive system - a necessary condition in order to allow its use as an RDA evaluation task.

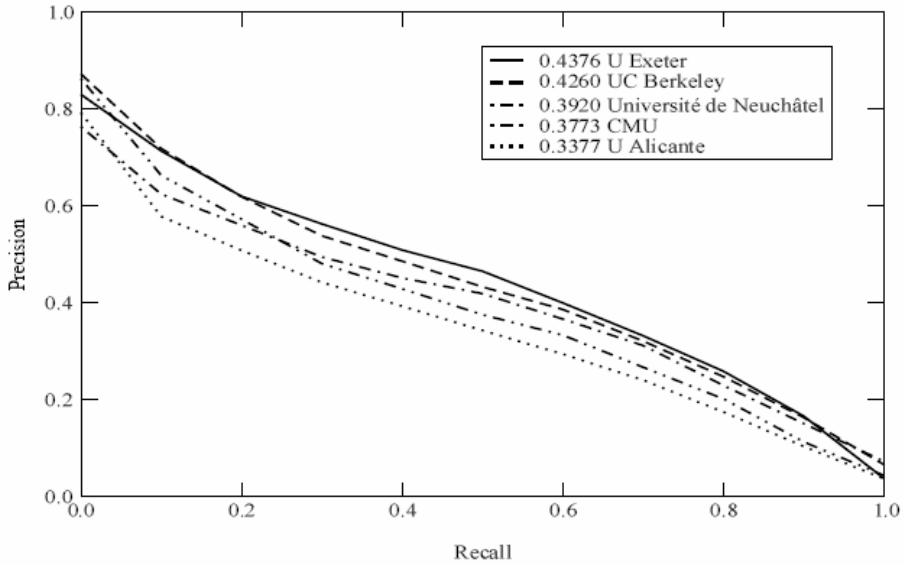

Figure 5.1: Top systems in CLEF 2003 – multilingual task.

5.5 CLIR Adaptation Data: Domain Specific Datasets and Parallel Corpora

Since the CLIR methods we use are statistical methods based on *parallel corpora*, the *domain adaptation* problem refers to adapting the parallel corpus (i.e. the system training resource) to the domain of the queries. As we will see in Chapter 6, good CLIR systems suffer a significant performance degradation when used in a new domain.

In order to evaluate the impact of domain adaptation on domain-specific CLIR, the evaluation dataset would need to contain domain-specific documents in the target language,

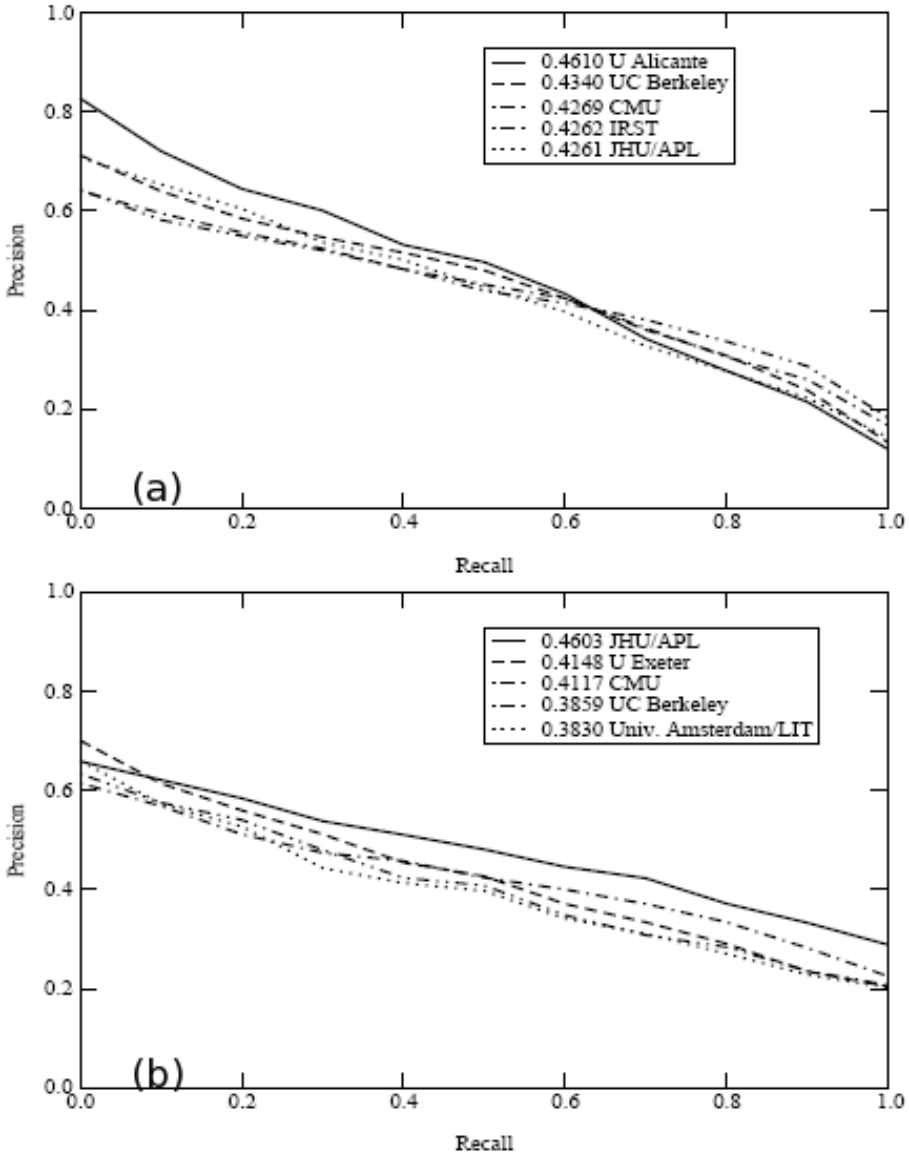

Figure 5.2: Our CLIR system (denoted by CMU) was one of the top performers in CLEF.

as well as domain-specific queries and their relevance judgments. These evaluation corpora are extremely difficult and costly to construct, due to the specialized skills and effort involved in designing the queries and making the relevance judgments. We have leveraged one such corpus in the medical domain by translating the queries in multiple languages. We have augmented the Springer corpus [46], which is a product of the MUCHMORE project, an international effort concerned with cross-lingual retrieval in the medical domain. It consists of

9,640 documents (titles plus abstracts of medical journal articles) in English and in German, with 25 queries in both languages, and relevance judgments made by native German speakers who are medical experts and are fluent in English. In Table 5.2 we show the English version of these queries. The Springer corpus is unique in its combination of domain-specificity and availability of multilingual queries with relevance judgments made by highly qualified medical professionals. Other parallel corpora such as Europarl are neither domain specific nor do they have multilingual queries with relevance judgments.

In its original incarnation, the parallel corpus was split into two subsets - one (4,688 documents) for training, and the remaining subset (4,952 documents) as the test set. In addition to the German queries, the dataset has been augmented with French, Italian and Spanish versions of the queries, translated by human domain experts. In [46], the corpus has been split into training and testing, so that the German half can be used for training. In this thesis, we used the French version of the queries (seen in Table 5.1) and the entire English corpus as the target corpus, since no German language training was needed.

In addition to Springer, we are also using the domain specific parallel corpora described in Chapter 3 (MedTitle, MedCat, MedTitle-IT). We remind the reader that these corpora consist of 500K+ titles/sentences of medical articles in French and English. Note that the training corpus contains titles only, whereas the target corpus consists of articles. In the case of CLIR, it can be argued that titles are closer to queries, therefore MEDTITLE as training data implies a fortunate genre match.

5.5.1 CLIR Adaptation Data: Corpus Degradation for Translation Quality Criterion Evaluation

In order to systematically explore the translation quality estimate criterion presented in Chapter 2, we set out to simulate parallel corpus translation mistakes. Since the parallel corpus is assumed to be translated by humans, the mistakes are inherently different from typical machine translation errors. While detection of machine translation errors is more

Query #	French Query
1	Traitement arthroscopique des lésions des ligaments croisés
2	Complications de l'intervention arthroscopique
3	Pathophysiologie et prophylaxie de l'arthrofibrose
6	Épidémiologie du VIH, évaluation du risque
9	Indications et limites d'analgésie à la demande
10	Amorçage avec des myorelaxants non dépolarisants
19	Complications aprs la cholécystectomie par laparoscopie
29	Trombopénie provoquée par l'héparine, diagnostic et gestion
66	Diagnostic de la maladie de Lyme
69	Traitement de l'infarctus aigu du myocarde
71	Ablation du cathéter et cartographie cardiaque
78	Approche de diagnostique pour les blessures de l'épaule
81	Diagnostic différentiel de fertilité
88	Approche de la correction des malformations orthopédiques
91	Traitement du carcinome spino-cellulaire
95	Maladies associées avec les diabtes insulino-dépendants
99	Thérapie de la douleur chronique du dos
103	Traitement de la tachycardie ventriculaire
104	Indication pour un défibrillateur interne à synchronisation automatique
108	Cause de la dysphagie
109	Traitement de la surdité de perception
112	Réparation chirurgicale de l'anévrysme aortique
115	Complications chez les patients souffrant de troubles psychosomatiques
124	Nouvelle approche de la chirurgie des ligaments croisés

Table 5.1: Translated French queries.

broadly studied (starting with i.e. [43]), human mistakes are not. One exception is TransCheck [28], a prototype system developed at RALI. It detects several classes of mistakes humans make and it involves extensive human translator research.

Human translation mistakes include:

- Mistranslations
- Sentence alignment errors
- Omissions and/or insertions

Query #	Original Query
1	Arthroscopic treatment of cruciate ligament injuries
2	Complications of arthroscopic interventions
3	Pathophysiology and prophylaxis of arthrofibrosis
6	HIV epidemiology , risk assessment
9	Patient-controlled analgesia indications and limits
10	Priming with non-depolarizing muscle relaxants
19	Complications after laparoscopic cholecystectomy
29	Heparin induced thrombocytopenia , diagnosis and management
66	Diagnostic in lyme disease
69	Treatment of acute myocardial infarction
71	Catheter ablation and cardiac mapping
78	Diagnostic approach in injuries of the shoulder
81	Differential diagnosis in infertility
88	Approach of the correction of deformities in orthopedics
91	Treatment of squamous cell carcinoma
95	Associated diseases with insulin dependent diabetes mellitus
99	Therapy in chronic low back pain
103	Treatment of ventricular tachycardia
104	Indication for implantable cardioverter defibrillator (icd)
108	Cause of dysphagia
109	Treatment of sensorineural hearing loss (snhl)
112	Complications of surgical repair of aortic aneurysm
115	Treatment of psychosomatical patients
124	New approach in cruciate ligament surgery

Table 5.2: Original Springer queries.

- Deceptive cognates ("false friends")
- Grammatical errors (not addressed in the current model)
- Geo-cultural errors (not addressed in the current model)

We are examining and modeling the first three types of errors; the rest are left to future work.

Sentence alignment errors are specific to parallel corpora: a text segment in language A is designated as being the translation of another text segment in language B that is nearby

the true translation.

Misalignments are easily detected by the quality estimate criterion; in fact, most misalignments are eliminated at the parallel corpus construction stage ([47])

Error models simulating human *mistranslation and omission* could span a wide range of approaches. Simple models include randomly omitting or replacing a word with another vocabulary word. More complex models could take into account language model probabilities at the unigram and n-gram level, replace words with synonyms instead of random vocabulary words, and consider domain-specific priors. Learning-based models are trained on datasets of tagged human translations and learn the type of errors made, their frequency and occurrence context. TransCheck [28] concentrates on detecting such errors - to our knowledge, it does not attempt to generate them.

In our experiments, we are using a simple error model (i.e. randomly replacing a word with another vocabulary word). However, more sophisticated error generation models as the ones mentioned above can easily be incorporated into the RDA framework.

In order to systematically explore the effect of the translation quality estimate (TQE) criterion, we degraded the English (target) half of the parallel corpus using the error model described above. The experiments in this chapter use F-1 [65] as the quality criterion estimate. The relationship between the quality estimate criterion and the degradation value is explored in Figure 5.3

Here, we have ranked the $100,000$ sentences that the lexical-level similarity criterion selected by their Translation Quality Estimate (in this case, the F-1 score) at various degradation levels. As expected, we observe a very strong correlation between F-1 score and the degradation level, due to the word substitution error model discussed above.

We can use this correlation to our advantage. If the F-1 measure is an accurate estimate of translation quality, it can be used as part of the input features to the learning algorithm that decides how to weigh the different criteria. For example, the average F-1 score can determine to what extent the quality estimate criterion should be used: if the translation is very good, the criterion is not needed and can have a lower weight; if, on the other hand,

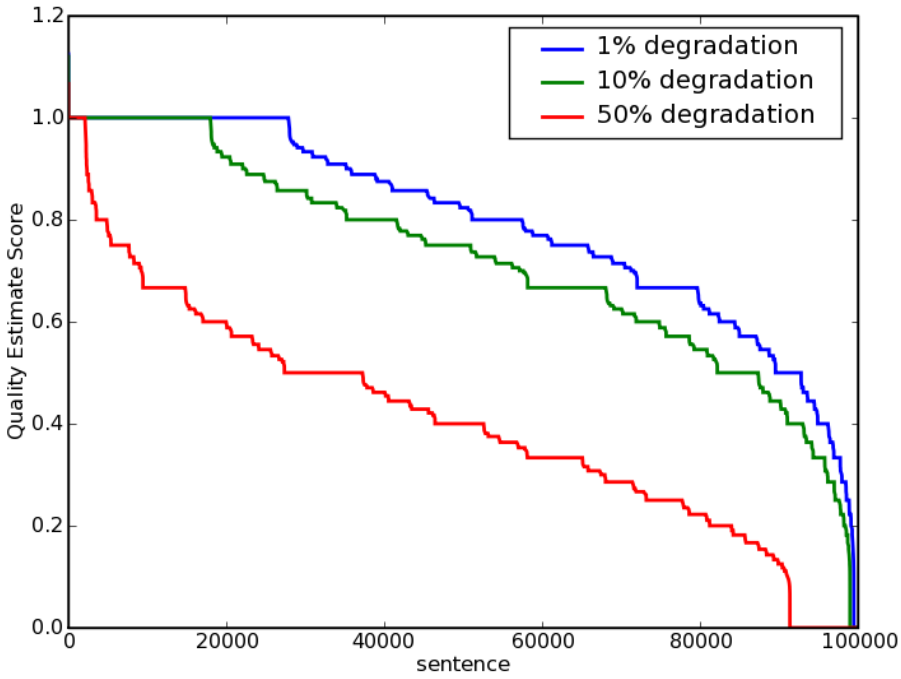

Figure 5.3: Translation Quality Estimate vs. Degree of Degradation. The X axes represents the number of degraded sentences in the parallel corpus. The sentences are ranked by quality estimate, as calculated by their F-1 measure relative to the non-degraded value.

the translation quality is estimated to be poor, the quality criterion is crucial.

5.5.2 Corpus and Query Pre-Processing

In our experiments, we have used several stages of pre-processing, applied to the target corpus, the parallel corpus, as well as the query. These pre-processing stages consisted of:

- case normalization
- selective punctuation removal
- rule-based (Porter) stemming for both English and French. For the German experiments, we have used overlapping character 6-grams.
- target corpus indexing using the INDRI search engine [64].

To explore the effect of the pre-processing order in our selection-CLIR pipeline, we examine two different pre-processing conditions. In the first experimental setting (S1), the selection of the parallel sentences to be used for training was done before stopword removal and stemming, in order to mimic the machine translation selection process. In the second experimental setting (S2), the queries and parallel corpora were stemmed and stopped before the selection process began. This setting is closer to that of corpus-based CLIR.

5.6 CLIR Adaptation: Systems Used

5.6.1 Baseline (MT-based) System (BMT)

We used our existing statistical machine translation system (discussed in Chapter 3) as the basis for constructing a simple CLIR engine . The system uses GIZA++ and the Pharaoh decoder [53], and a general English language model. This baseline, machine-translation based system (BMT) first trains its MT component using an RDA-selected subset of the MEDTITLE corpus. The parameter setting remains unchanged from the MT task. Then, the trained system is used to translate the pre-processed queries from the source language to the target language.

Once translated, the queries are used to retrieve documents in the INDRI-indexed target corpus. Pseudo-relevance feedback is used at this stage in order to facilitate a more direct comparison with our high-performing CLIR system below. The BMT system only serves as a baseline in order to compare its performance (and adaptation sensitivity) to that of a true CLIR system (PMI).

5.6.2 PMI-based CLIR system (PMI)

Our second system is the corpus-based CLIR system described earlier in this chapter. PMI is a true CLIR system in the sense that intermediary results are intended for computer, as opposed to human consumption: instead of choosing the best word or phrase, all weighted alternatives are included. More specifically, the function used to calculate word-to-word

similarity is PMI.

There are several fundamental differences between the experimental settings described in [46] and those described in this chapter. Although the same system is used, the results obtained reflect the following changes:

a. Different source language (German vs. French). This implies different morphological processing and degree of appropriateness of the query terms

b. Different target corpora (half of the available documents in [46], vs. all of them here). This implies a different set of relevant documents

c. A completely different training set (MEDTITLE vs. half of the Springer corpus)

5.6.3 Parameter Values and Design Decisions

Where possible, we have kept the parameter values consistent when using the same CLIR system. Due to data scarcity and the numerous degrees of freedom such a system entails, we used conservative, generally accepted parameter values that have worked well across tasks and data collections. Since this thesis' main goal is not CLIR parameter optimization, we preferred to avoid the risk of overfitting. In BMT, the parameters were the ones used for the machine translation experiments in Chapter 3. When blind feedback was used, the parameters were identical to those of the PMI system. In PMI, the parameters used were tuned on the CLEF 2002 data collection, and values with solid performance across languages were selected. These parameters include: the number of documents and words used for blind feedback (both on the source and target side), and the minimum number of parallel instances in which two words co-occur in order to be included in the translation candidate list. The choices were 5, 20 and 3 respectively.

5.7 Chapter Summary

In this chapter, we have described various aspects of our high-performing CLIR systems, as well as their performance in international evaluation forums. We have introduced CLIR-specific evaluation datasets and parallel corpora, obtained by augmenting existing medical CLIR evaluation datasets like Springer.

Chapter 6

PARDA for Cross-Lingual Information Retrieval: Experiments and Results

6.1 Overview

In Chapter 1 we presented a succinct overview of the cross-lingual information retrieval (CLIR) problem, evaluation metrics and domain-specific challenges. In Chapter 5, we described in detail our corpus based CLIR systems and presented results on general-domain (i.e. CLEF evaluation) datasets, in order to establish our CLIR system as a high-performing system in the general domain. As in the case of statistical machine translation, using our general-domain CLIR system on a medical domain evaluation dataset results in significantly degraded results (explored in Section 6.9.1).

In this chapter, we focus on the specific application of our RDA framework to the CLIR problem, taking advantage of the PMI CLIR system described in Chapter 5 and the domain-specific Springer CLIR dataset discussed in [46] and in Section 5.5. More specifically, our framework automatically selects the best training data subset, which is subsequently used in our corpus-based CLIR methods.

In addition to showing the significant effect that selection has on the amount of parallel/training data necessary to produce satisfactory CLIR results, we compare two different

corpus-based CLIR systems and two different preprocessing settings.

We explore the effect of several individual RDA criteria: mean reciprocal rank (MRR), translation quality estimate (TQE) and sentence size. We show that, while the MRR criterion is crucial, the TQE criterion fulfills a very important role when the parallel corpus is noisy.

6.2 Lexical Similarity (RDA-MRR) Effect on CLIR Performance

We remind the reader that the MRR criterion (see Section 2.4.1) is a sentence selection criterion based on its mean reciprocal rank when ranked by lexical similarity to the domain sample.

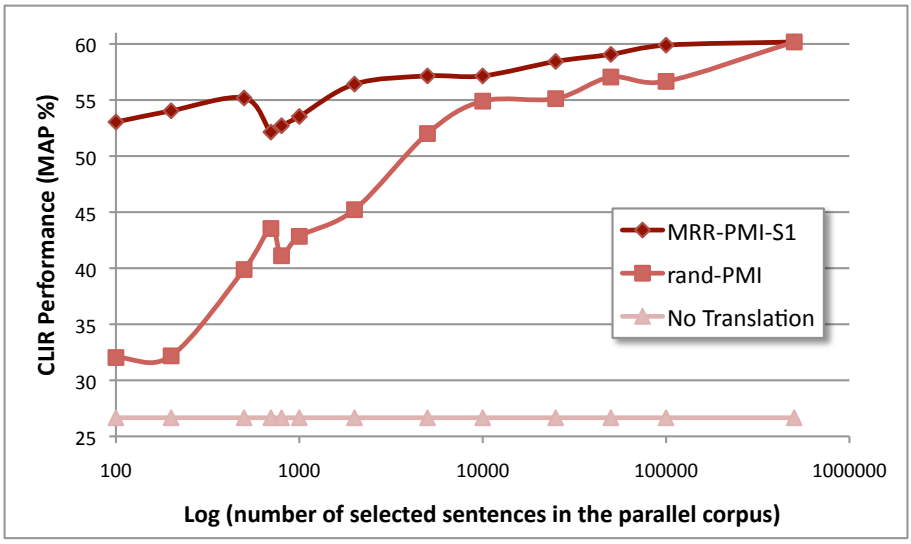

Figure 6.1: Springer FR-EN CLIR Results: Mean Average Precision when the parallel corpus is a) randomly selecting sentences out of MEDTITLE, b) selected from MEDTITLE using Mean Reciprocal Rank, and c) absent. Medical terminology similarity in the two languages yields a high no-translation baseline. Note that the full corpus performance in this case is the rightmost point of the PMI lines. In this particular case, the performance for 100 selected sentences is equivalent to 10,000 randomly selected sentences.

Figure 6.1 shows CLIR results on the Springer dataset [46], with French as the source language and English as the target language. The X-axis represents (on a log scale) the

number of parallel sentences selected by each method to obtain the respective CLIR result.

Due to the medical terminology similarity in the two languages, morphological processing alone with no translation leads to results that are higher than normally observed for this setting. The random setting is a purposely high baseline, due to the random selection being performed in a corpus that is already domain-specific (MEDTITLE).

The highest performance is obtained when the entire MEDTITLE corpus is utilized. We remind the reader that the corpus contains approximately half a million article titles in French, with their English counterparts. The important observation emerging out of this log-scale plot is that 90% of the performance is obtained at the **100 sentences** level, when they are selected using MRR. In other words, when RDA-MRR is used, the parallel corpus translation cost is reduced by more than 90%, while the CLIR performance level is reduced by only 10%. The same result is shown in Figure 6.2, which answers the following question: Given that the CLIR performance using this dataset has an upper bound in the performance obtained when the training corpus consists of all MEDTITLE documents, what percentage of this performance can be obtained with significantly fewer documents?

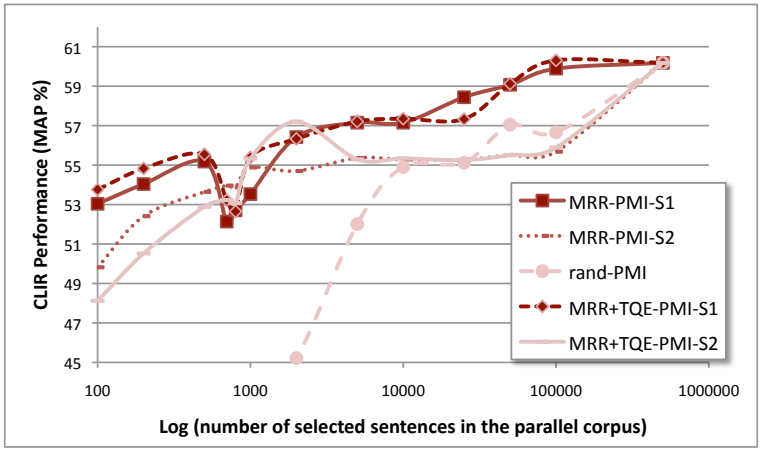

Figure 6.2: How many parallel sentences are needed to reach a fraction of the best performance? (semilog plot; this plot quantifies the relative (percent of peak) performance vs. absolute performance in Figure 6.1.)

6.3 Example Queries: Expansion and Translation

Given the queries in Table 5.1, the top selected documents in the S1 setting (Section 5.5.2) are shown in Table 6.1. Some documents can be traced directly to the query they support; others are selected due to high mean reciprocal rank across several queries.

Rank	English half of the selected parallel pair
1	surgic repair chronic instabl cruciat ligament knee cyst
2	cruciat ligament knee treatment simpl radio guid punctur
3	arthroscop treatment chronic anterior shoulder instabl
4	treatment plant injuri caus marin anim caledonia approach
5	indigen treatment diagnosi treatment lyme diseas children
6	letter canadian societi pediatr binder syndrom maxillo nasal
7	dysostosi orthoped malform complic cholecystectomi laparoscopi
8	avoid low back pain women diagnosi treatment role intern
9	committe dermatolog patient control analgesia studi fentanyl
10	requir burnt patient acut phase unusu dysphagia esophag
11	tuberculosi epidemiolog atherosclerot cardiovascular risk hiv
12	infect patient map radiofrequ ablat form peri atriotomi flutter
13	complex tachycardia myocardi infarct criteria ventricular
14	tachycardia diagnosi develop heparin induc thrombocytopenia
15	biolog clinic aspect result associ depolar depolar muscl relax
16	analysi case myocardi ruptur acut myocardi infarct treatment
17	renal osteodystrophi physiopatholog secondari effect aneurysm
18	transvers arch aorta ruptur left lung surgic cure case report
19	author transl occup rehabilit patient mental psychosomat disord

Table 6.1: Top selected MEDTITLE documents using the Springer queries as the domain sample. The criterion used here is MRR, and the selection/pre-processing order setting is S1 (see Section 5.5.2). The documents are processed and tokenized.

However, the documents are severely altered at the 50% degradation level. Table 6.2 shows the selected documents after the degradation process. Note that the quality criterion has not been applied. The low document quality is reflected in low CLIR scores. This effect is mitigated by the addition of the quality criterion - the results of this selection are shown in Table 6.3.

The composition of the documents changes significantly when the quality criterion is

Rank	English half of the selected parallel pair – with 50% degradation
1	surgic infants resettlement 6.30 epidemiology discrimination
2	knee legitimises inauguration wake knee treatment simpl boot
3	guid punctur arthroscop treatment opportunist countries
4	shoulder instabl treatment tamils injuri caus marin air surface
5	1977 mobiles treatment beef thereby vitro explained utri diseas
6	children letter attribution societi pediatr binder template
7	them ibrahim dysostosi formulations peres obligation
8	cholecystectomi impartially avoid low back prepares vain
9	diagnosi treatment unworkable recognising committe eliminates
10	patient control analgesia stay endeavouring communications
11	honestly advised acut swear
12	unusu catalyst definitive tuberculosi photos atherosclerot
13	biotechnological smell isis rent patient map executed ablat
14	form peri atriotomi complexion penalise undcp myocardi infarct
15	criteria ventricular tachycardia diagnosi develop lawlessness
16	induc rearing biolog clinic aspect result formerly sub-regional
17	depolar egunkaria deterioration analysi concentration affairs
18	ruptur fixed myocardi delivering ries redressing osteodystrophi
19	physiopatholog secondari effect twenty-one transvers disbanded
20	question klamt repudiate lung surgic conservatism
21	case earliest author transl
22	occup autonomous exaggerated helped psychosomat o flourish

Table 6.2: Top selected MEDTITLE documents using the Springer queries as the domain sample. The English half shown here have been degraded using the simple error model described above, at the 50% level. The criterion used here is MRR, at the selection/pre-processing order setting is S1. The documents are processed and tokenized.

added to the 50% degraded documents (Table 6.3). More specifically, the selected documents are either ones where the degradation was kept to a minimum, or where the less significant (less domain-specific) words were the ones replaced by the degradation process, leaving the content words intact.

The parallel corpus selection affects query expansion as well as selection. In Figure 6.3, and 6.4 we show the two example queries (number 1 and number 71 in Table 5.2 and 5.1) after their expansion. The Y- axis shows the proportional (normalized) weight (in percent) that

Rank	English half of the selected parallel pair – with 50% degradation and with quality criterion
1	complic cholecystectomi laparoscopi avoid
2	role consulates committe dermatolog arthroscop treatment
3	chronic wasted bases instabl diagnosi effect subclin burst
4	spraying case analysi societal eventualities pop acut myocardi
5	infarct binder syndrom 12.00 nasal dysostosi enter
6	old-fashioned treatment plant optimising caus marin hormones
7	fancy approach indigen treatment undernourished transl unusu
8	dysphagia providers bic immunotherapi optimist reintroduce
9	squamou staggering carcinoma ey auditori local speech australia
10	bicycles infringement studi case perceptu deaf map
11	inundated gendarmerie form today atriotomi flutter complex week
12	myocardi infarct impairment scandal heralded
13	cyst cruciat macrofinancial knee scenes bodes eliminating guid
14	punctur occup candidate patient complies psychosomat crashed
15	innovate 300,000 laparoscop cholecystectomi vascular biliari
16	complic cages control gujarat studi keen requir a5-0162 patient
17	solely phase submit disrupted anterior instabl
18	shoulder opportunity treatment lyme gbp optimising yearly
19	anadian attwooll exposes latent hiring inherent demand
20	converging psychiatrist art displac cruciat manufacturer
21	creeping flexion normal ecologists hoping apolipoprotein level
22	insulin depend nostalgic patient 800 glycosyl
23	hemoglobin fructosamin
24	epicardi undermine case miracles tachycardia caus sordid blood
25	myocardi infarct origin smallest surgic approach lyme options
26	biolog diagnosi treatment

Table 6.3: Top selected MEDTITLE documents using the Springer queries as the domain sample. The English half shown here have been degraded using the simple error model described above, at the 50% level. The criterion used here is MRR and TQE (translation quality estimate), and the selection/pre-processing order setting is S1. The documents are processed and tokenized.

the X-axis word has in the expanded query. The two settings shown are when the expansion has been done in 1) the entire MEDTITLE corpus and 2) the top 100 selected sentences (shown in Table 6.1). For both settings, the number of documents used for expansion was 5,

and the number of words, 20. The words are sorted according by the average weight between the two settings, and only the top 14 are shown.

The 100 selected sentences allow the top terms to behave similarly to when the entire corpus is used. Moreover, the top terms added to the original query are very relevant. However, this is expected since finding 5 relevant sentences in 100 sentences selected for query set relevance is not a difficult task. For one particular query (Example Query 1 in figure 6.3), the task is even easier since there are similar queries in the Springe dataset, which over-weights arthroscopic-related documents when MRR is used.

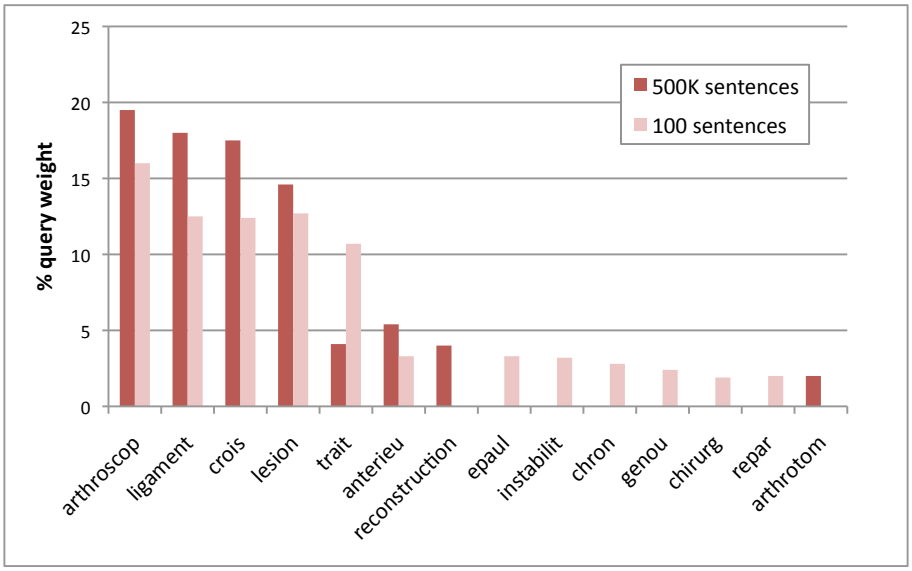

Figure 6.3: Example Query 1: Expanded Query using the entire MEDTITLE parallel corpus vs. a 100-sentences RDA-MRR selected corpus. The Y-axis shows the proportional (normalized) weight (in percent) that the X-axis word has in the expanded query. Note that the top weighted words are the same for using the entire corpus or 100 sentences, and the words introduced by 100-sentence expansion include some query-relevant words such as "genou", "chirurg", "repar".

The difference is more pronounced after the expanded queries have been translated (in figure 6.5 and 6.6) . In these queries, the 100-(selected) sentences and 500K-sentences corpora have been used for both the expansion phase as well as the PMI-based translation phase. Note that although the small corpus introduces noise ("author"), and leaves some terms

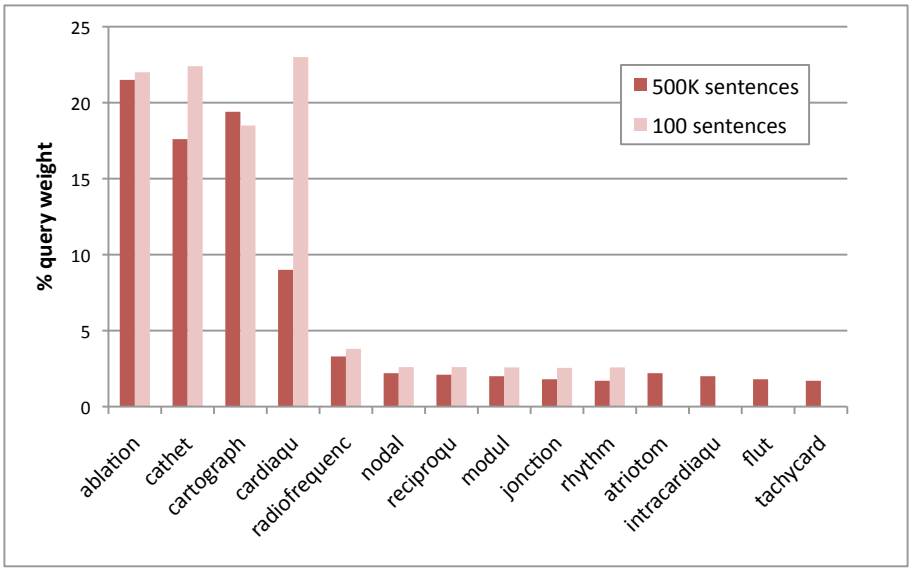

Figure 6.4: Example Query 72: Expanded Query using the entire MEDTITLE parallel corpus vs. a 100-sentences RDA-MRR selected corpus. The Y-axis shows the proportional (normalized) weight (in percent) that the X-axis word has in the expanded query.

untranslated (anterieur, cardiaq), the top terms are the same as with the 5,000 times larger corpus, and are translated correctly. This greatly influences the CLIR performance, which is very high given the 100-sentences parallel corpus.

6.4 BMT vs. PMI: Comparing the Two CLIR Systems

We compared the two CLIR methods discussed in Chapter 5. BMT, based on machine translation, translates the query after training a statistical machine translation system. The second system (PMI) projects the weighted query in the other language using corpus-based similarity statistics.

The parallel corpora used to train the two alternatives are the same, when the same experimental conditions are shown in the graphs. Figure 6.7 shows the results when both system are trained with parallel corpora which have been stemmed after selection. The selection here has been done using MRR. The PMI method yields significantly better results.

Since the performance difference can be attributed to several factors, we decided to

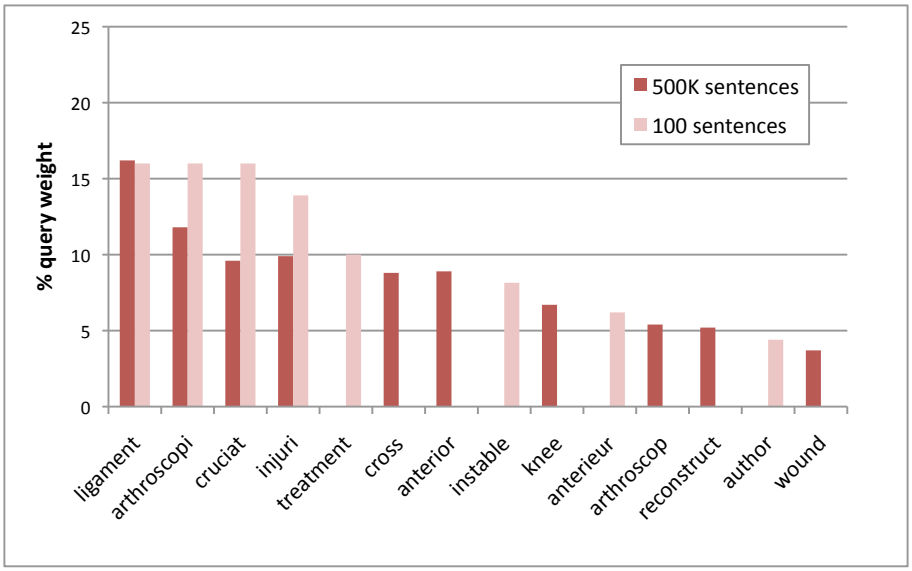

Figure 6.5: Example Query 1: Translated Query using PMI and the entire MEDTITLE parallel corpus vs. a 100-sentences RDA-MRR selected corpus. The Y-axis shows the proportional (normalized) weight (in percent) that the X-axis word has in the query.

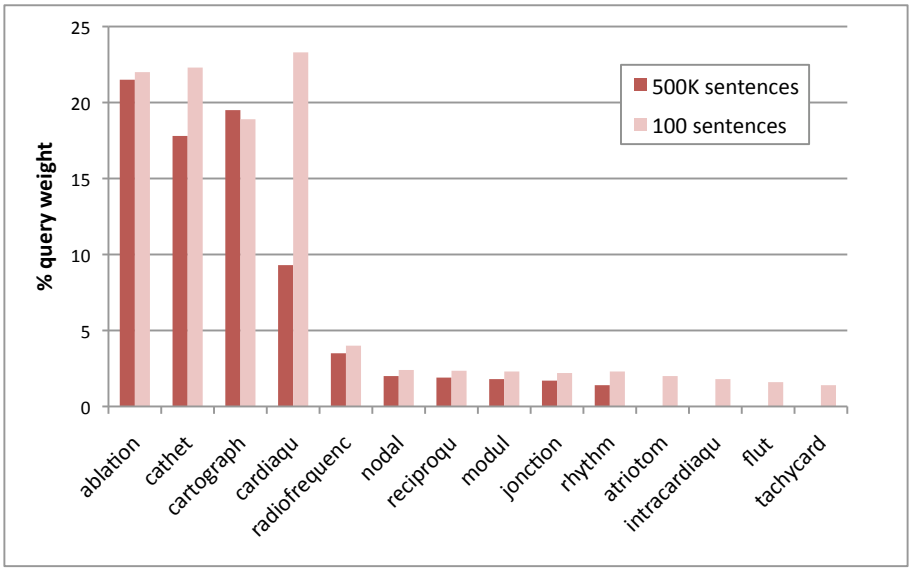

Figure 6.6: Example Query 72: Translated Query using PMI and the entire MEDTITLE parallel corpus vs. a 100-sentences RDA-MRR selected corpus. The Y-axis shows the proportional (normalized) weight (in percent) that the X-axis word has in the query.

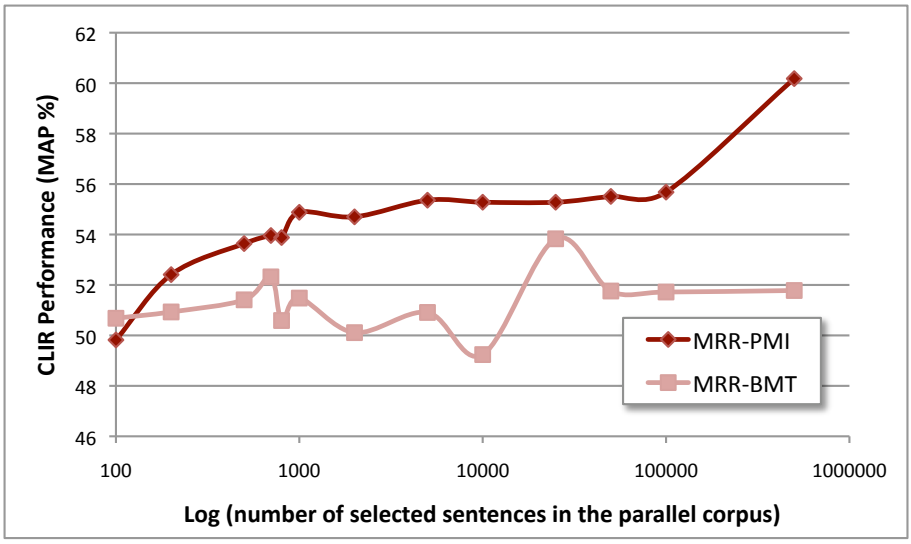

Figure 6.7: Springer FR-EN CLIR Results: Comparing PMI and BMT when the parallel corpus is selected from MEDTITLE using MRR.

examine the most likely contributors other than the system itself: a) the query expansion (and weighting) effect and b) the general language model being present or absent in the final step of the translation.

We modified the BMT system to accept the same expanded queries (together with word-level weights) that the PMI system used. The weights are preserved through the final retrieval step and are being attributed to a phrase wherever the translation is a phrase.

The results are presented in Figure 6.8. Figure 6.8 shows that introducing query expansion before the translation step results in lower CLIR performance. This effect (which is contrary to the effect observed in the PMI system) can be attributed to a) the expanded query being translated by a phrase-based decoder like Pharaoh, and b)introducing noisy terms that cannot be successfully under-weighted by the (post-translation) retrieval engine.

The second factor that could contribute to the better performance of the PMI system is the presence of the language model. Since CLIR queries are not destined for human consumption, their post-translation fluency is not relevant. Moreover, it could hurt by favoring less relevant terms that contribute to fluency. In order to test this hypothesis, we set the weight of the language model to zero at decoding time. The results are presented in

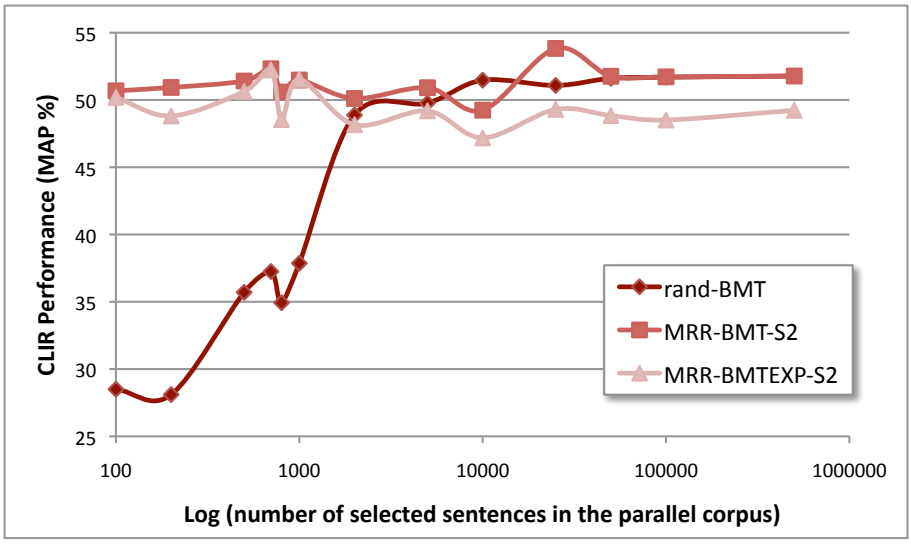

Figure 6.8: Springer FR-EN CLIR Results: Comparing query expansion effects when using the BMT system.

Figure 6.9.

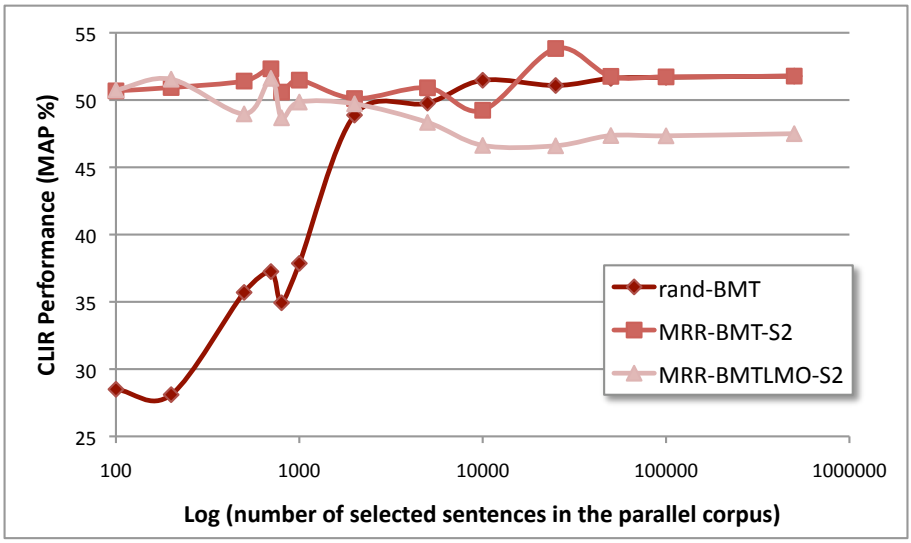

Figure 6.9: Springer FR-EN CLIR Results: the impact of the language model when using BMT.

We notice that the lack of a language model hurt, instead of improving the CLIR results. By examining the resulting queries and comparing them with the ones produced with the language model, we notice many queries show several spurious terms introduced by the

translation model, but filtered out by the language model.

Since neither the language model nor the query expansion is responsible for the performance improvement, we continue using the dedicated, true CLIR system (PMI-based) for future experiments in this chapter, instead of the baseline, machine-translated queries system (BMT) that did not match its performance.

6.5 Comparing the Two Pre-Processing Settings

In this section, we examine the effect of the S1 and S2 preprocessing settings described in Section 5.5.2.

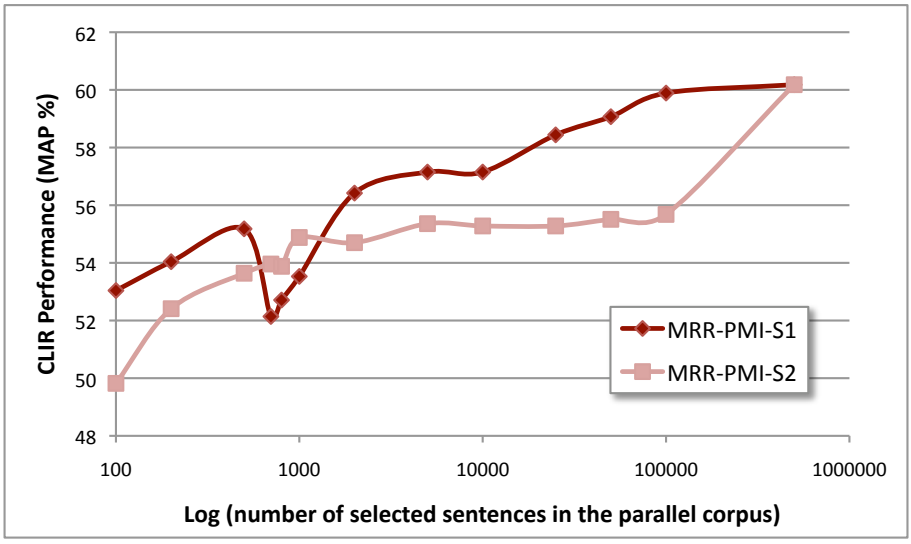

Figure 6.10: Springer FR-EN CLIR Results: Mean Average Precision when the parallel corpus is stopped and stemmed before selection (S2) or after (S1).

We used two pre-processing settings, identified by S1 and S2. One (S1) performs the stemming and stopping after the selection process, closely resembling the machine translation setting. The second (S2) performs the selection after stopping and stemming, similar to traditional corps-based CLIR.

This allows us to explore the effect of the pre-processing order. The results are somewhat

surprising, given this is a CLIR task: S1 had better performance. We offer as an interpretation the possibility that the more specific terminology present before the stemming-induced conflation increase the precision of the selection process.

6.6 Using Sentence Size as a Selection Criterion

In this section we consider the sentence (and corpus) size as a selection criterion. Longer sentences have more information, and the length criterion may also be needed to counterbalance the perfect quality scores short sentences might have, akin to a brevity penalty. However, longer sentences are harder to align word-to-word, and the length needs to be balanced with the other criteria. Moreover, it can be argued that, in the cases where RDA is used as an active learning device, longer sentences carry a higher translation cost which should be factored into the overall sentence score. Figure 6.11 shows CLIR performance when sentence size is used as a threshold after the initial MRR-based selection. More specifically, sentences shorter than a specific threshold (on the X axis) are eliminated from the training data.

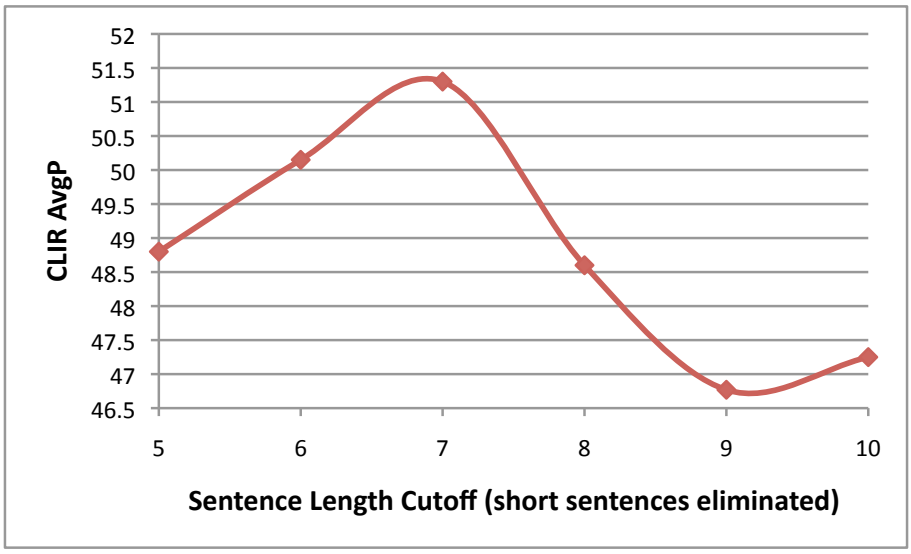

Figure 6.11: CLIR performance when using sentence size as a threshold. The number of selected sentences is always 100. Sentences shorter than a specific threshold (on the X axis) are eliminated from the training data, which results in lower quality sentences for high cutoff values.

We notice that the elimination of short sentences does improve performance - however, once we eliminate medium-sized sentences the drop in quality is noticeable. This is a predictable effect, since when too many top scoring (MRR-wise) sentences are eliminated, the performance of the CLIR system suffers. The exact break-even point for this criterion depends on several factors, one of the most important being the cost function associated with data translation and processing speed.

6.7 The Effect of the Translation Quality Estimate (TQE) Criterion on CLIR Results

The translation quality estimate criterion is added to the selection process using the geometric mean combination function (see Section 2.5 for more details). In particular, the specific measure used here is sentence averaged word-level F-1, calculated between the English half of the parallel corpus and the translated French half of the parallel corpus (bootstrap and evaluation in Section 2.4.2). The MT system used for this purpose is one trained on the entire MEDTITLE dataset.

Adding the quality estimate criterion does not affect the CLIR results significantly. This is the expected result when the quality of the translation and/or alignment is consistent throughout the corpus. In the case of MEDTITLE, the translation quality is high throughout the corpus, and since the titles were not further broken down into sentences, the alignment problems were minimal.

However, the potential of the quality criterion is better reflected in results when the quality varies throughout the corpus. As discussed in Section 5.5.1, we simulate this setting with various degrees of parallel corpus degradation. The results for this experimental setting are described in Section 6.7.1.

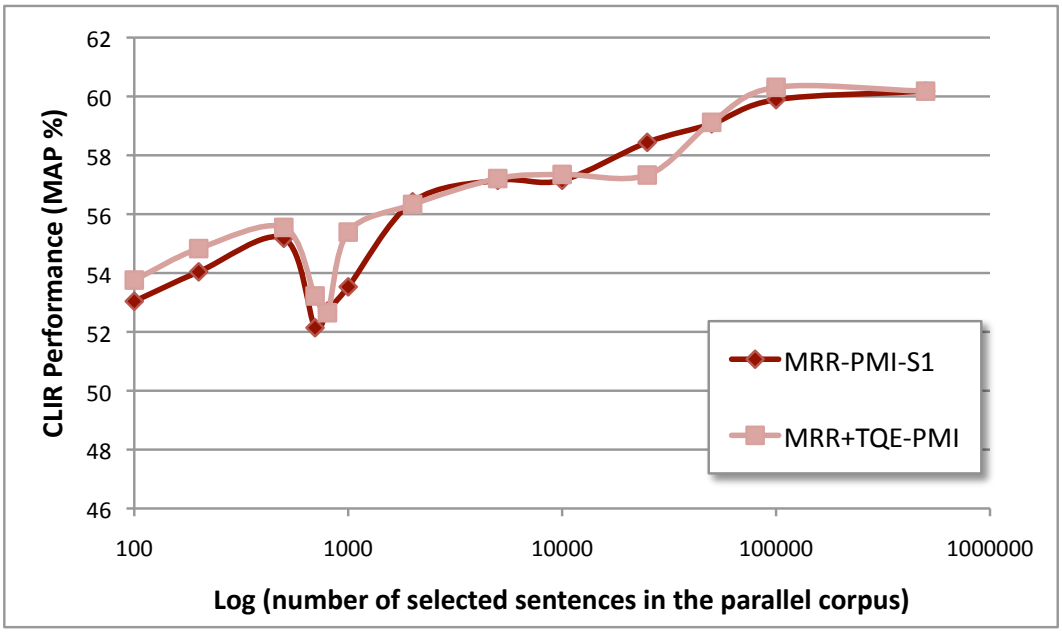

Figure 6.12: Springer FR-EN CLIR Results: Mean Average Precision when the parallel corpus is selected from MEDTITLE using Mean Reciprocal Rank vs. using the Mean Reciprocal Rank and the Translation Quality Estimate criterion. The dip is due to a noisy selected sentence that lead to a misleading translation. The difference between the two test conditions is not statistically significant: The TQE criterion does not improve CLIR performance when the corpus translation quality is high and homogenous; we explore a different setting in the next section.

6.7.1 Effects of the Translation Quality Estimate (TQE) Criterion Given a Low Quality Parallel Corpus

We remind the reader that the translation quality estimate criterion refers to the quality of the parallel corpus (i.e. human) translations. As described in section 5.5.1, in order to systematically explore the effect of the TQE criterion we have degraded the quality of the parallel corpus by replacing a certain percentage of the English words with random vocabulary words.

Figure 6.13 examines the effect of said degradation on the CLIR results. It is an interpolated heatmap, with datapoints collected at the 1%, 5%, 10%, 30% and 50% degradation level (on the x axis), and at 100, 200, 300, 500, 700, 800 and 1000 selected sentences (on

the y axes). The temperature represents mean average precision (MAP) - the red regions represent better performance than the blue regions. The experimental setting here is using the PMI system, with the S1 preprocessing scheme.

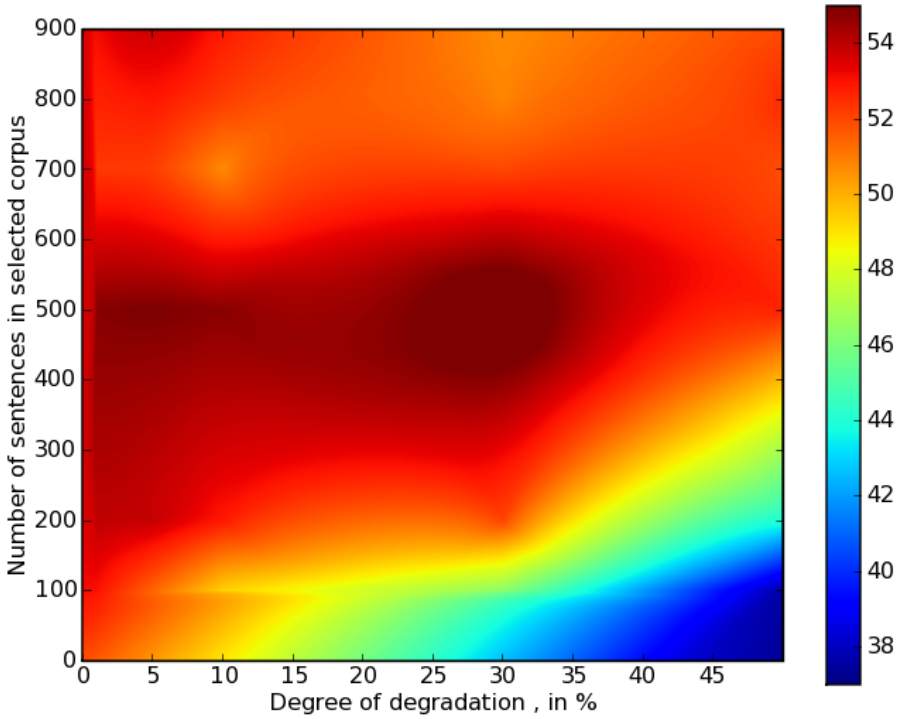

Figure 6.13: CLIR adaptation results for various degrees of corpus degradation (best viewed in color). The thermometer legend shows the corresponding MAP value for each color - red indicates high performance, blue low. The CLIR performance suffers at high levels of corpus degradation, but only when few sentences are selected.

As expected, when the parallel corpus has a higher degree of degradation and is small, the CLIR performance decreases dramatically (see Figure 6.13).

Maintaining the same experimental settings as before, we add the quality estimate criterion, using weighted geometrical mean as the combination method between TQE and MRR (see Section 2.5 for more details). For simplicity, these experiments use equal weights for the two criterion.

Figure 6.14 examines the effect of adding the quality criterion to the selection process, for various degrees of degradation. As before, it is an interpolated heatmap, with datapoints

collected at the 1%, 5%, 10%, 30% and 50% degradation level (on the x axis), and at 100, 200, 300, 500, 700, 800 and 1000 selected sentences (on the y axes). The temperature represents mean average precision (MAP). The temperature scale is the same as Figure 6.13 in order to make the figures directly comparable.

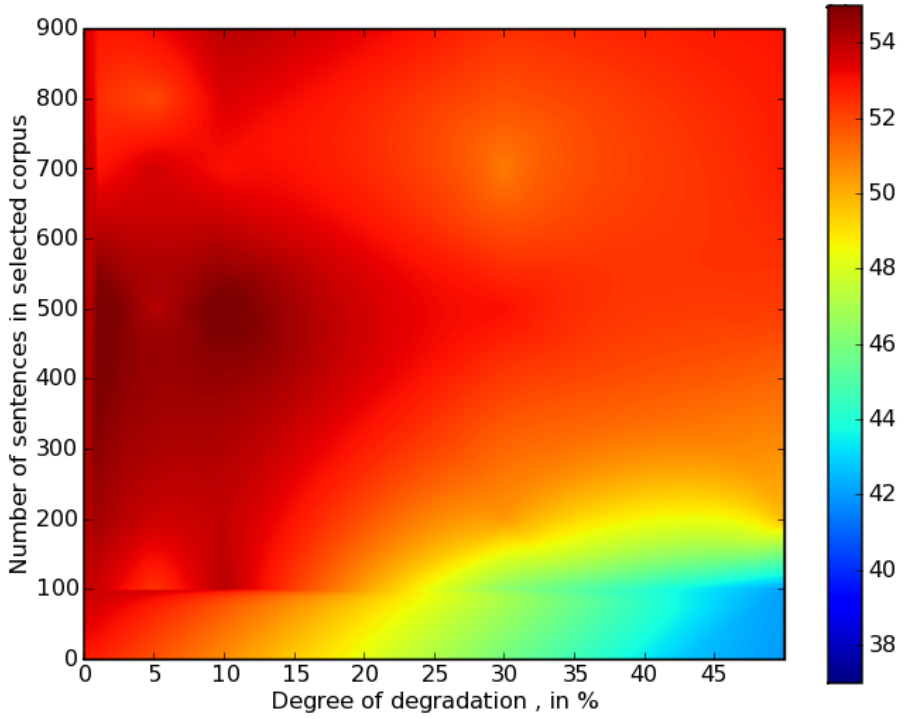

Figure 6.14: The effect of adding the quality criterion to the selection process, for various degrees of degradation (S1) (best viewed in color). The thermometer legend shows the corresponding MAP value for each color - red indicates high performance, blue low. The CLIR performance is significantly improved for high degradation values.

We notice that the previous region of concern (high degradation, small parallel corpus) has greatly improved. This implies that, even though the TQE criterion does not significantly improve (or hurt) the results when the quality is high (or consistent) throughout, it offers superior protection against moderately or severely mistranslated texts.

Figure 6.15 highlights the difference between Figure 6.14 and Figure 6.13 .

Here, warm (red and yellow) values are indicative of an improvement when using the quality criterion vs. not using it, and cold (blue) values show a decrease in performance.

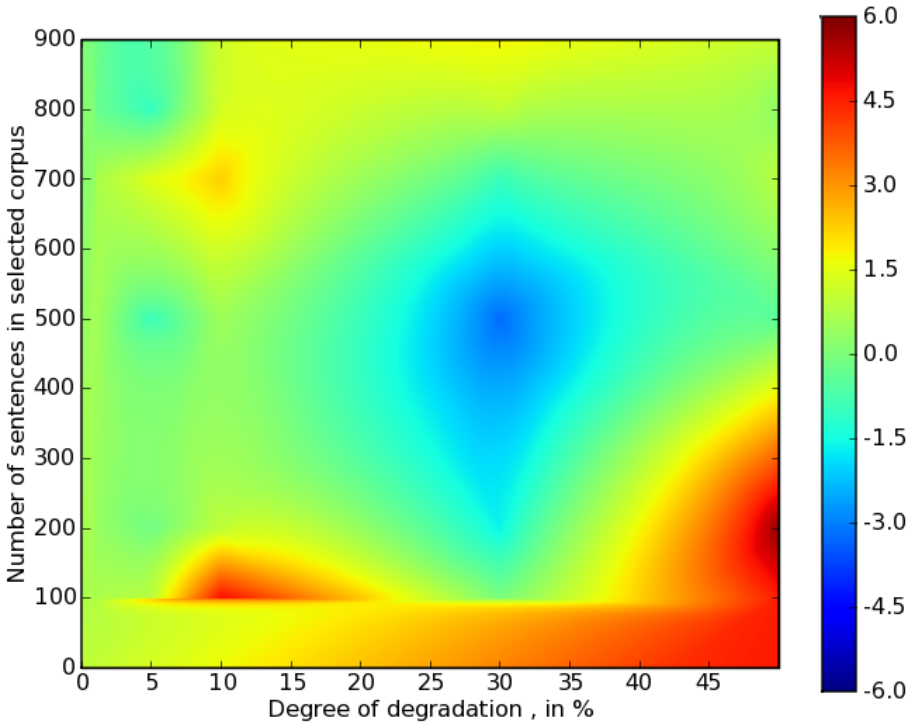

Figure 6.15: The improvement obtained by adding the TQE criterion (best viewed in color).

Green values indicate no change in either direction.

6.7.2 TQE: Related Observations and Results

Figure 6.16 shows the correlation between translation quality measures (such as F1 and modified BLEU) used to evaluate the Springer queries, and the subsequent CLIR performance (MAP). The correlation is weaker than expected, with F1 having better correlation than MBLEU. This suggests that we cannot use MBLEU of F1 as a strong predictor for CLIR MAP - other factors, such as expansion effects and relevant document set size affect CLIR performance.

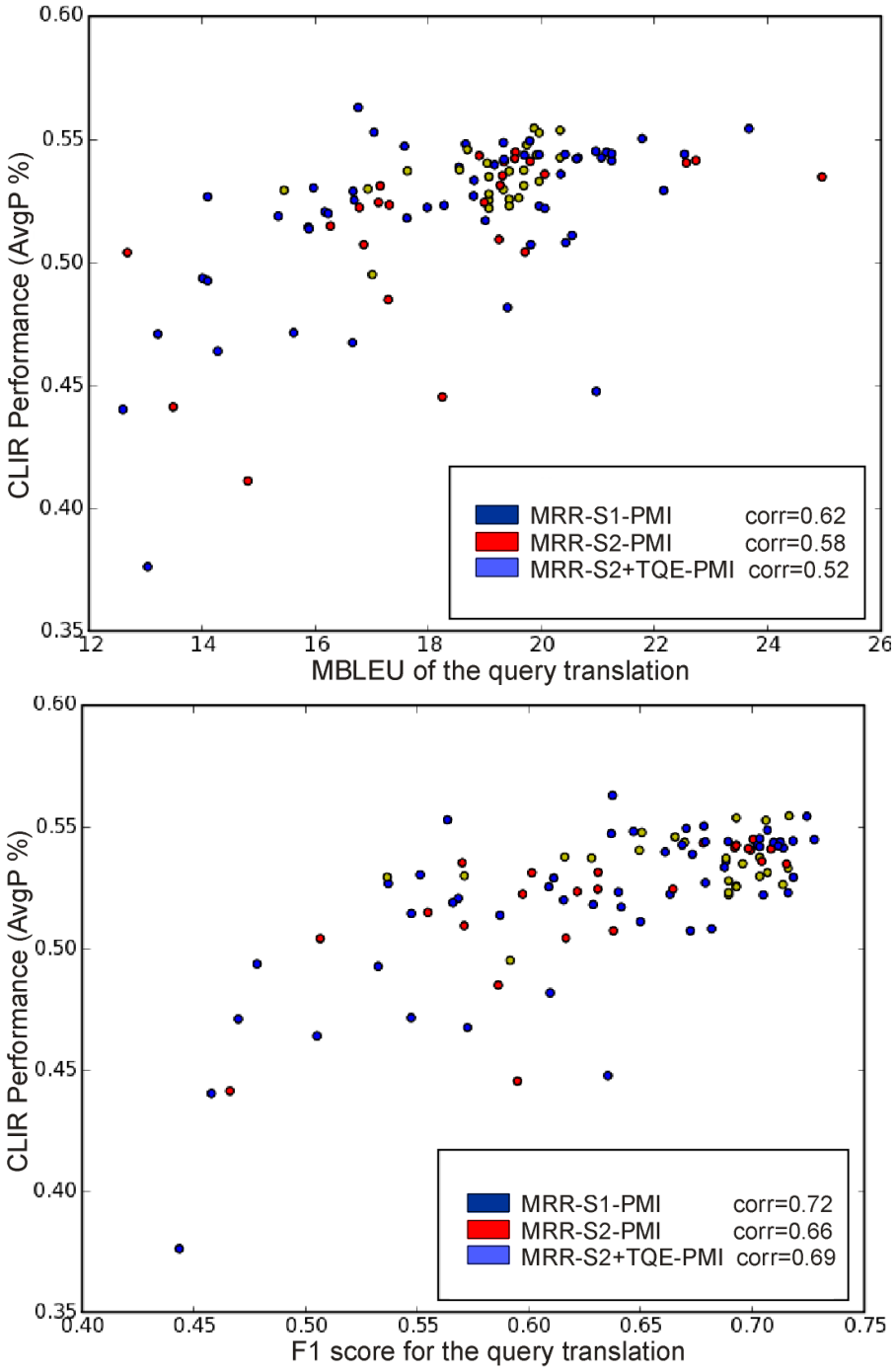

Figure 6.16: Correlation coefficient between CLIR results and query translation Modified BLEU value, and average F1 value.

6.8 CLIR Results: Comparing Previous Experimental Settings

In order to provide an overall picture of the various factors and experimental conditions that affect CLIR performance, Figure 6.17 shows the results for all the condition and criteria discussed above. Figure 6.18 zooms into the zone where selection has the most impact - when the number of sentences is limited to 100-1,000.

The most surprising result is the impact RDA-MRR has on CLIR results when little data is used, especially when compared to random selection (be it in the medical domain, but not in the specific sub-domain).

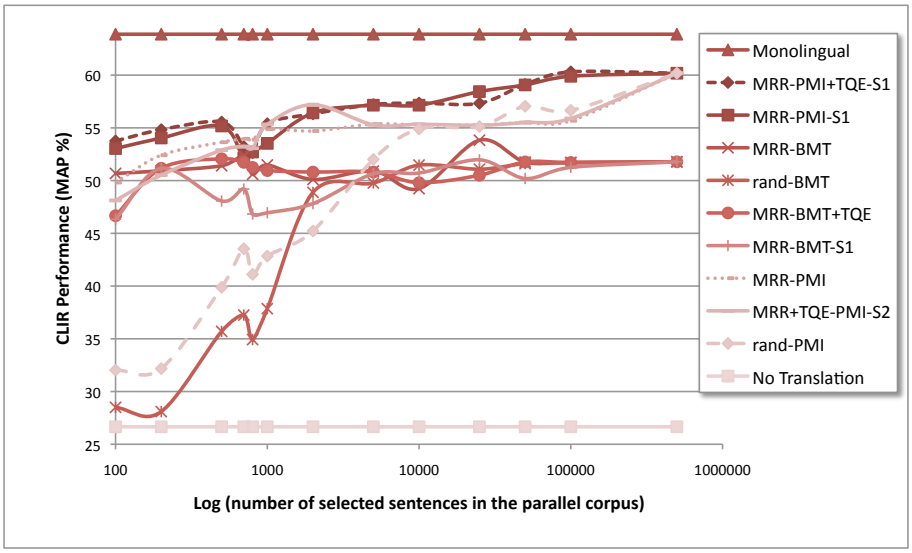

Figure 6.17: Springer FR-EN CLIR Results: Comparing all experimental settings.

6.9 CLIR Adaptation: Language Variation Results

In this thesis, we have applied the same CLIR method for both German and French, with minor differences in data pre-processing. In this section, we present results obtained with a preliminary version of our system on German/English data. While the main body of the thesis uses French as the query language, the adaptation results are similar across the two

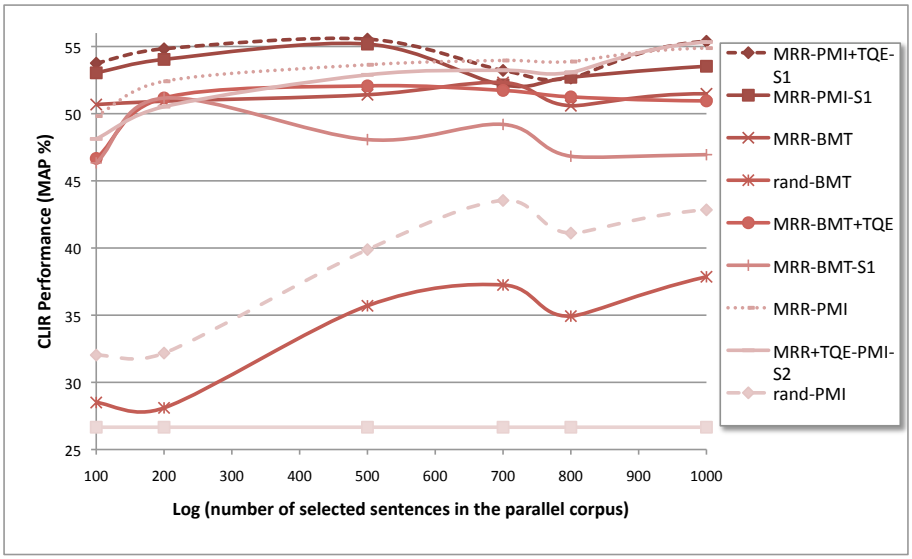

Figure 6.18: Springer FR-EN CLIR Results: Comparing all experimental settings: Selecting up to 1,000 sentences.

languages (French/German).

For CLIR training and testing we have used the Springer corpus (introduced in Section 5.5). We split this parallel corpus into two subsets, and used the first subset (4,688 documents) for training, and the remaining subset (4,952 documents) as the test set in all our experiments. We applied an alignment algorithm to the training documents, and obtained a sentence-aligned parallel corpus with about 30K sentences in each language. The sentence-aligned version of the Springer training set was used in the experiments presented here.

In addition to Springer, we have used four other English-German parallel corpora for training:

- NEWS is a collection of 59Kx2 parallel sentences extracted from news stories, downloaded from the web and covering the 1996-2000 period. It is available for download at
  http://www.isi.edu/ koehn/publications/de-news/.

- WAC is a small (60Kx2 sentences) parallel corpus obtained by mining the web. It is

more general and it does not focus on a particular domain.

- EUROPARL ([36]) is a parallel corpus introduced in Section 3.3.1.
- MEDTITLE-DE, another product of the MUCHMORE project, is an English-German parallel corpus consisting of 549K paired titles of medical journal articles. These titles were gathered from the PubMed online database (*http://www.ncbi.nlm.nih.gov/PubMed/*)

Table 6.4 presents a summary of the five training corpora characteristics.

Name	**Approximate Size** (sentences, words)	**Domain**
NEWS	59Kx2, 2M	news
WAC	60Kx2, 1.1M	mixed
EUROPARL	665Kx2, 35M	politics
SPRINGER	30Kx2, 0.9M	medical
MEDTITLE-DE	550Kx2, 21M	medical

Table 6.4: Characteristics of English-German parallel training corpora.

6.9.1 Applying General Domain Models to the Medical Domain

Our preliminary translation model adaptation results are obtained by using a very simple method for automatically choosing and weighting the training resources to adapt them to the target collection in the medical domain. We were able to show $5-10\%$ improvements in average precision over corpus-based approaches that used all available resources, and very significant improvements over general purpose MT systems such as SYSTRAN. Additionally, we have shown the need for proper selection of training resources, as seen in Figure 6.19.

CLIR results in the medical domain with different training corpora show that choosing the largest collection (EUROPARL), using all resources available without weights (ALL), and even choosing a large collection in the medical domain (MEDTITLE) are all sub-optimal strategies. This motivates our search for a more sophisticated adaptation model, as presented in the rest of this thesis.

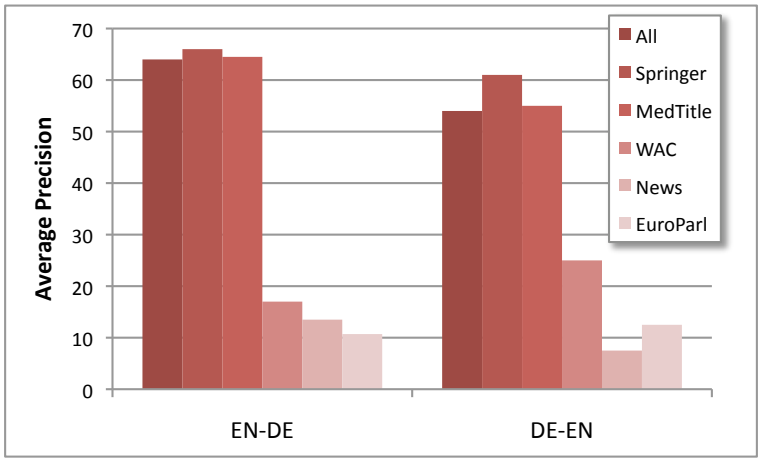

Figure 6.19: CLIR results in the medical domain with different training corpora. Notice the large differences obtained with domain-mismatch training data. Choosing the largest collection (EUROPARL), using all resources available without weights (ALL), and even choosing a large collection in the medical domain (MEDTITLE) are all sub-optimal strategies.

Data Preprocessing

We have eliminated all punctuation, stopwords and numbers for both German and English, and we have used the English Porter stemmer. To simulate German decompounding, we first stemmed the German portion of the corpus and we split the words into 5-grams. The German stemmer and stopword list was provided by [61].

Selection Criterion: Vocabulary Coverage

A naive, but quick method to measure domain similarity in these experiments is by using vocabulary overlap between the training corpus and the domain sample as a domain match approximation. Table 6.5 shows the vocabulary coverage with respect to the training collection. The training vocabulary coverage is calculated as follows:

$$Cov(train, sample) = \frac{|V_{train} \cap V_{sample}|}{|V_{train}|} \tag{6.9.1}$$

Name	**DE Coverage (%)**	**EN Coverage (%)**
NEWS	27.9	14.5
WAC	28.8	12.5
EUROPARL	6.6	1.8
SPRINGER	57.7	35.4
MEDTITLE	10.8	3.4

Table 6.5: Vocabulary coverage of training corpora with respect to the Springer test set

Selection Criterion: Cosine Similarity

The second simplest idea to approximate domain matching is to use the cosine similarity between the testing and training corpus (TFIDF term weights). Note that these documents are very short (sentences), which means the document length is fairly constant and TF and DF tend to be close. Table 6.6 shows the cosine similarity between the five training corpora and the test set.

Name	**DE COS (%)**	**EN COS (%)**
NEWS	44.08	33.90
WAC	54.67	33.84
EUROPARL	49.82	36.22
SPRINGER	99.29	90.89
MEDTITLE	55.20	72.94

Table 6.6: Cosine similarity of training corpora with respect to the Springer test set

Combining Translation Resources

Our previous approach has been to use vocabulary coverage or cosine similarity as the weight for each translating resource when combining them, instead of using it as a criterion for the selection threshold. Each parallel corpus produces a similarity matrix, using one of the methods outlined in Chapter 5. A new similarity matrix is produced from their linear combination, using the vocabulary coverage or cosine similarity as the corresponding weights. In practice, it is only necessary to calculate this linear combination for dictionary entries

present in the queries.

This approach has the advantage that it does not require relevance judgments and existing queries to learn the weights. We examine the robustness of this approach in Table 6.7

Using Vocabulary Coverage or Cosine Similarity as Weights

Figures 6.21 and 6.20 show that both vocabulary coverage and cosine similarity between the training and target corpus are positively correlated with retrieval performance, as measured by the average precision over all queries. However, MEDTITLE-DE is a significant outlier in this respect (when considering vocabulary coverage in Figure 6.20). We believe this to be because of the different spelling normalization applied to MEDTITLE-DE when it was downloaded from the web; document length did not have a significant effect. However, this disparity is alleviated in Figure 6.21, because the high TF-IDF terms are medical terms where fewer spelling variations occurred.

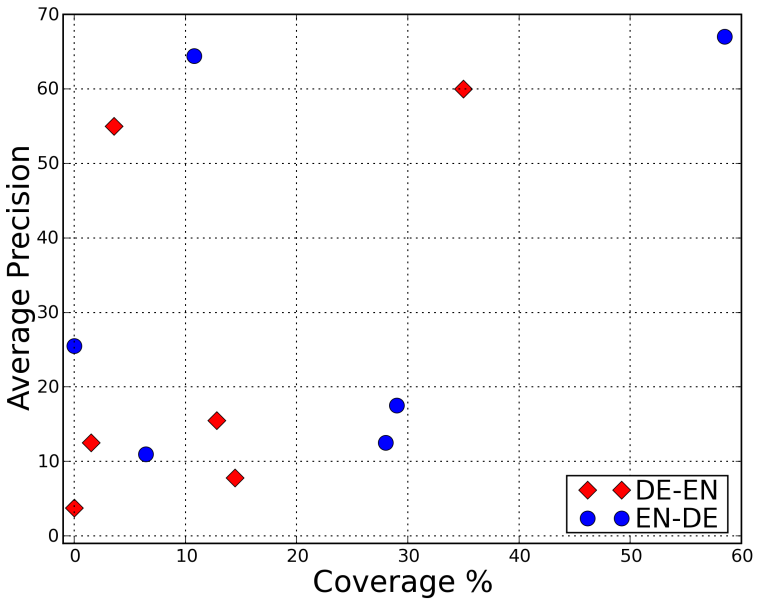

Figure 6.20: CLIR performance (of PMI) vs. the training-set vocabulary coverage

We wish to examine the robustness of using vocabulary coverage or cosine similarity between the target collection and the parallel corpus as linear combination weights. To that end, it is insufficient to experiment with weighting the five corpora mentioned above. For this

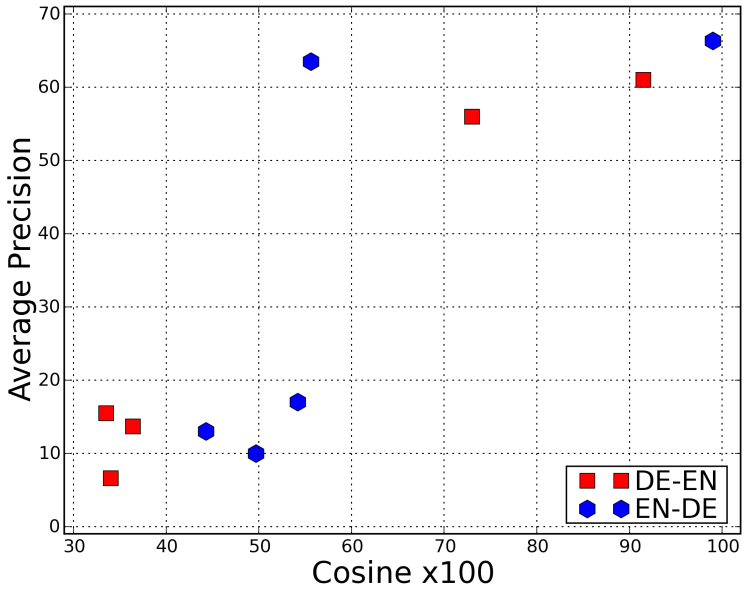

Figure 6.21: CLIR performance (of PMI) vs. the cosine similarity

approach to be robust, the combination should be consistent in not performing significantly worse than the best collection available – as given by an oracle – and in not performing worse than using all collections with equal weight. Naturally, the combination should also perform better than the expected performance of a randomly selected collection, but in our case this straw man baseline is not needed.

In Table 6.7 and its corresponding experiments, we simulate the availability of 5, 4 and 3 training collections from the five we have described above. We did not examine resource pairs, since the resource selection problem becomes trivial in this case. There are a total of 16 testing conditions: 1 way to choose all 5, 5 ways to choose 4, and 10 ways to choose 3 (order does not matter).

The first column enumerates which training collections we are allowed to select from and/or to weight. The respective collections (WAC, MEDTITLE, SPRINGER, EUROPARL, NEWS) are represented by their initials. All 16 combinations are shown for both DE-EN and EN-DE.

In this table, COV represents performance when training set coverage was used as the weight; COS represents performance when the cosine similarity was used as the weight; EQ

Available Resources	DE-EN				DE-EN			
	COV	**COS** (%impr over EQ)	**EQ**	**Best Single Collection (Oracle)**	**COV**	**COS** (% impr over EQ)	**EQ**	**Best Single Collection (Oracle)**
WESNM	59.61	57.34(5.52%)	54.34	60.56(S)	66.36	66.52 (5.23%)	63.21	66.47(S)
ESNM	60.33	59.37(4.23%)	56.96	60.56(S)	66.51	67.5 (2.61%)	65.78	66.47(S)
WENM	40.24	55.26(5.94%)	52.16	55.1(M)	48.46	61.62 (2%)	60.41	63.5(M)
WESM	59.07	57.7(5.61%)*	54.63	60.56(S)	66.92	67.13 (4.75%)	64.08	66.47(S)
WESN	60.15	58.79(9.74%)*	53.57	60.56(S)	65.35	59.47(9.64%)	54.24	66.47(S)
WSNM	59.89	57.88(3.67%)	55.83	60.56(S)	66.71	67.44(5.49%)	63.93	66.47(S)
WNM	40.71	54.83(2.29%)	53.60	55.1(M)	52.49	62.26(-0.24%)	62.41	63.5(M)
WES	59.77	58.82(7.49%)	54.72	60.56(S)	65.74	64.91(19.76)*	54.20	66.47(S)
WEN	18.79	19.01(0.21%)	18.97	15.39(W)	22.22	22.83(15.88)	19.70	17.26(W)
WSM	59.27	57.25(3.11%)	55.52	60.56(S)	66.33	66.86(2.76)	65.06	66.47(S)
ESM	60.32	58.87(0.56%)	58.54	60.56(S)	66.98	66.44 (-1.24%)	67.28	66.47(S)
WEM	44.01	55.42(2.74%)	53.94	55.1(M)	56.07	61.78(0.29%)	61.60	63.5(M)
ESN	60.88	60.08(4.63%)	57.42	60.56(S)	66.72	66.56(11.02%)*	59.95	66.47(S)
SNM	60.31	59.57(1.72%)	58.56	60.56(S)	67.10	67.19(0.97%)	66.54	66.47(S)
ENM	48.39	55.55(0.34%)	55.36	55.1(M)	51.25	63.15(0.66%)	62.73	63.5(M)
WSN	59.86	59.68(7.33%)	55.60	60.56(S)	65.08	65.02(10.74%)*	58.71	66.47(S)

Table 6.7: Collection Availability Simulation for Macro Adaptation

represents using all available resources with equal weights; and Best Single Collection shows the performance of the single best one training corpus, if it were possible to know which one is the best in advance. Numbers in bold highlight the best performance of the four conditions. The star indicates statistical significance within the 95% confidence interval, using the paired t-test.

From this table, we see that cosine similarity is a better weighting measure than vocabulary coverage, which is to be expected. This simple method is robust when we simulate the availability of different collections; however, we believe more research is needed to explore different weighting criteria.

Summarizing across all 16 testing conditions, we observe that our strategy accounts for a $4-5\%$ improvement over using all resources with no weights, for both retrieval directions. It is also very close to the "oracle" condition, which chooses the best collection in advance.

Document Level Selection

In the previous section, we used cosine similarity between training and target corpora as respective weights when building a translation model. This approach treats a parallel corpus as a homogeneous entity, an entity that is self-consistent in its domain and document quality. In this section, we propose that instead of weighting entire resources, we can select individual sentences from these corpora in order to build a parallel corpus that is tailor-made to fit a

specific target collection.

In addition to proposing individual documents as the unit for building custom-made parallel corpora, in this section we start exploring the criteria used for individual document selection by examining the effect of ranking documents using the length-normalized Okapi-based similarity score between them and the target corpus.

In this section, as well as in the rest of the thesis, we focus on a lower granularity level: individual documents in the parallel corpora. In the case of sentence-aligned corpora, a "document" is a sentence. Possible intermediary granularity levels include document(or sentence) clusters and paragraphs; higher granularity levels include collection clusters. We seek to construct a custom parallel corpus, by choosing individual documents which best match the domain sample. We compute the similarity between the test collection (in German or English) and each individual document in the parallel corpora for that respective language. We have a choice of similarity metrics, but since this computation is simply retrieval with a long query, we start with the Okapi model [58], as implemented by the Indri search engine [64]. Although the Okapi model takes into account average document length, we compare it with its length-normalized version, measuring per-word similarity. The two measures are identified in the results shown below by "Okapi" and "Normalized". Once the similarity is computed for each document in the parallel corpora, only the top N most similar documents are kept for training. They are an approximation of the domain(s) of the test collection. Selecting N has not been an issue for this corpus (values between $10 - 75\%$ were safe). However, more generally, this parameter can be tuned to a different test corpus as any other parameter. Alternatively, the document score can also be incorporated into the translation model, eliminating the need for thresholding.

We start by selecting individual documents that match the domain of the test collection. We examine the effect this choice has on domain-specific CLIR.

Using Okapi Weights to Build a Custom Parallel Corpus

Figures 6.22 and 6.23 compare the two document selection strategies to using all available documents, and to the ideal (but not truly optimal) situation where there exists a "best" resource to choose and this collection is known. By "best", we mean one that can produce optimal results on the test corpus, with respect to the given metric. In reality, the true "best" resource is unknown: as seen above, many intuitive choices for the best collection are not optimal.

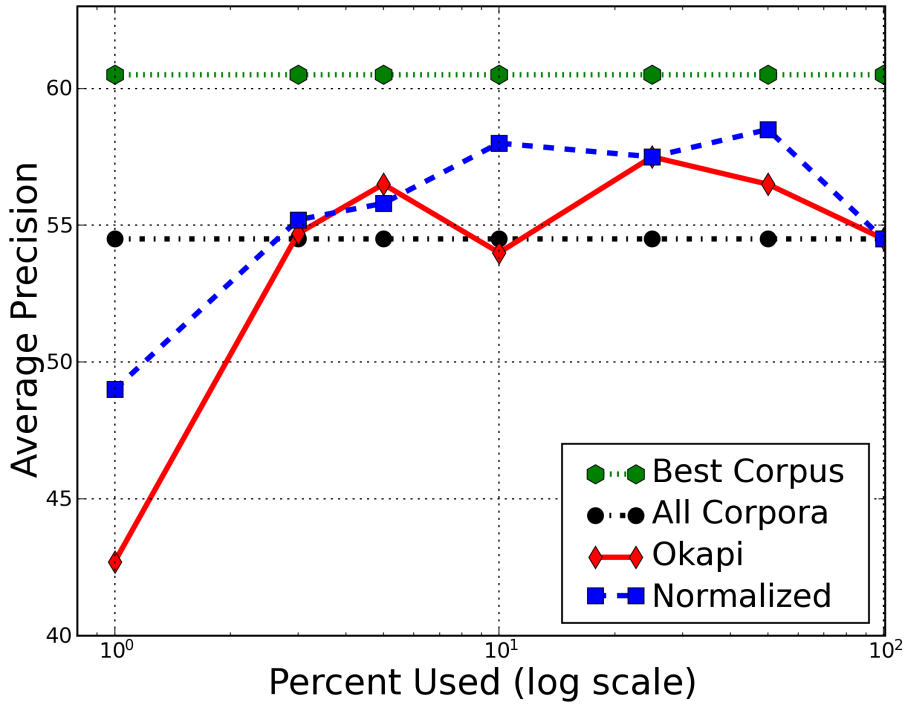

Figure 6.22: CLIR DE-EN performance vs. Percent of Parallel Documents Used. Best Corpus is given by an oracle and is usually unknown.

Notice that the normalized version performs better and is more stable. Per-word similarity is, in this case, important when the documents are used to train translation scores: shorter parallel documents are better when building the translation matrix. Our strategy accounts for a $4-7\%$ improvement over using all resources with no weights, for both retrieval directions. It is also very close to the "oracle" condition, which chooses the best collection

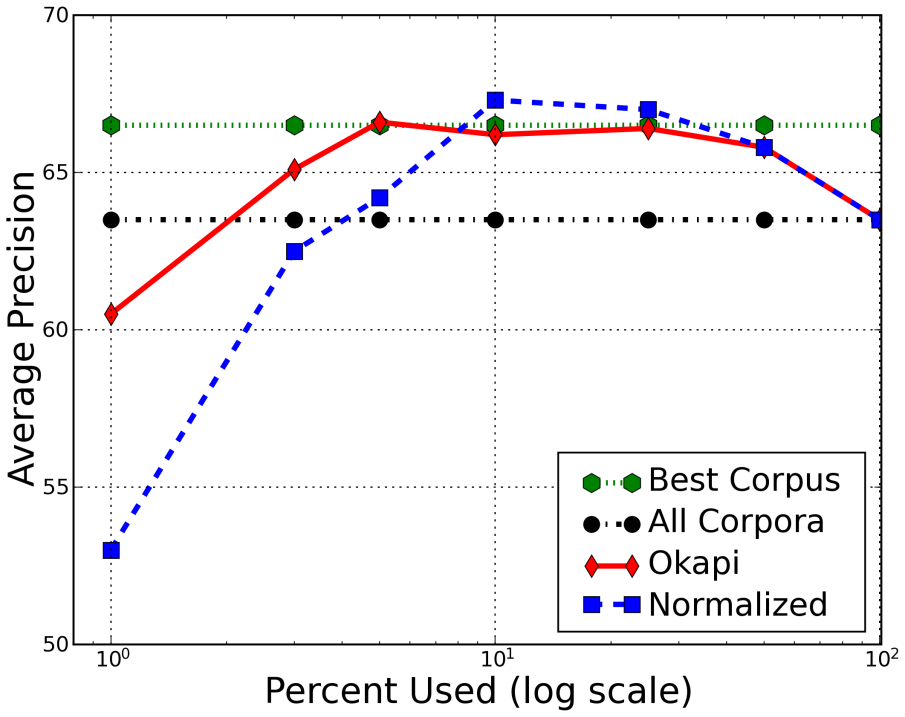

Figure 6.23: CLIR EN-DE performance vs. Percent of Parallel Documents Used. "Best Corpus" is given by an oracle and is usually unknown.

in advance. More importantly, by using this strategy we are avoiding the sharp performance drop when using a mismatched, although very good, resource (such as EUROPARL).

Offline adaptation consists of building profiles for all translation resources and the target corpus, then combining the resources into a customized translation model. The (online) query is then translated with the resulting translation model.

To avoid confusion, it is helpful to remember that in IR settings the true test data are the queries, not the target documents. The documents are available off-line and can be (and usually are) used for training and system development. In other words, by matching the training corpora and the target documents we are not using test data for training.

The results presented in this section show that choosing the largest collection, using all resources available without weights, and even choosing a relatively large parallel collection in the medical domain are all sub-optimal strategies. This results is maintained across the two languages, with different corpora and different adaptation methods.

6.10 Chapter Summary and Conclusion

In this chapter we have examined the effects of domain adaptation of the parallel corpus, using cross-lingual information retrieval as the evaluation task.

The particular characteristic differentiating CLIR from other multilingual tasks is its advantage of not having humans as the post-translation end-users. This property allows greater flexibility when generating the translation: i.e., instead of premature optimization and lexical selection, the post-translation result (the translated query) can remain in its weighted-token form. When this flexibility is taken away (i.e. when using the BMT system), the results suffer accordingly.

We have shown that, when data is costly to translate or when domain specific data is available in limited quantities, appropriate selection is crucial. The results obtained under these circumstances show that 100 RDA selected sentences can lead to 90% of the performance obtained with 5,000 times more data in the same general domain.

We have also examined the impact of two other criteria in the RDA framework. The length of a sentence does impact the CLIR results; however, whether this criterion can be successfully exploited depends on the training data generation cost function The other criterion we examined was TQE - translation quality estimate. While ineffective when the parallel corpus quality is high, the criterion significantly improves CLIR performance when data is scarce and of low quality.

Selection using the various RDA criteria is crucial when CLIR training data generation is expensive - as it usually is the case with domain-specific data. When using the RDA framework to perform document selection, the CLIR performance is similar to when using three orders of magnitude more data. This result surpasses the random selection baseline by a significant margin.

Chapter 7

Criteria Optimization for Cross-Lingual IR Domain Adaptation

7.1 Overview

Automatic subset selection from a parallel corpus significantly improves cross-lingual information retrieval (CLIR) performance, in addition to increasing its efficiency. Our selection method extracts relevant training data by incorporating additional criteria (i.e. estimated corpus quality, taxonomy projection and size) in addition to lexical-based criteria. The challenge lies in combining these criteria using a meaningful scoring function that can be used for ranking parallel sentence candidates. We choose weighted geometric mean for its soft-AND properties, and we optimize criteria weights by wrapping the CLIR task in an optimization shell. We start by exploring local optimization - in particular, the reactive affine shaker (RASH) method, an efficient algorithm which continuously adjusts its search area in order to identify a local minimum. However, due to the indeterminate nature of the search space convexity properties, we also investigate continuous reactive tabu search (CRTS), a global optimization method. We use a large parallel corpus in the medical domain to examine the effect of adaptation criteria and their combination on CLIR performance. In our experiments, 100 selected sentences yield 90% of the performance obtained with 5,000 times more

in-domain parallel sentences. Our optimized criteria weights considerably outperform the uniform distribution baseline, as well as lexical similarity adaptation, the preferred approach in statistical machine translation

7.2 Problem Description and Motivation

In previous chapters we have shown the impact each adaptation criterion has on CLIR performance. We have shown that, especially in the case of today's heterogeneous parallel corpora, it is especially important to consider issues such as corpus and translation quality, noise, size distribution, redundancy, genre or other available metadata in additional to lexical similarity. In this context, the results presented in Chapter 6 and 3 were obtained using equal criterion weights whenever criteria were combined.

In this chapter, we enhance the adaptation framework by focusing on the challenge of combining these criteria: in particular, on the problem of assigning the relative importance to each criteria when ranking parallel sentence candidates.

The problem of assigning criteria weights requires optimizing a highly non-linear, non-convex multi-dimensional function. Since multiple local optima may exist, we require a global constrained optimization method. Global optimization strategies include branch-and-bound methods (however, most have the disadvantage of relying on information about the problem structure or on the availability of an analytic formulation), bayesian partition algorithms (where a prior on the problem dimensions is needed), genetic algorithms, adaptive stochastic search (e.g simulated annealing, which places a non-zero probability on moving away from the optimum) etc. A more detailed description of global optimization methods and heuristics, as well as pointers to related work and comprehensive surveys can be found at [42]. Many of the above methods suffer from two main drawbacks: requiring a (fast) calculation of the objective function gradient, requiring smooth continuous functions or the availability of a formula, and/or requiring too many objective function calls. Since our particular function (corpus domain adaptation, then CLIR) is a fairly time-consuming black box, minimizing the function evaluation calls as well as avoiding to provide a gradient are

important considerations.

Our method of choice is continuous reactive tabu search (CRTS), as described in [2]. This is a global, deterministic, tabu-search based optimization method that uses the reactive affine shaker (RASH) [2] method as its local optimization routine. CRTS's combinatorial optimization algorithm focuses on locating the set of promising areas in the search space, and it initializes RASH while adapting search parameters and area size.

In the results presented in this chapter, we combine criteria using weighted geometric mean, chosen for its soft-AND properties, and we optimize criteria weights by wrapping the CLIR task in a CRTS optimization shell.

The questions we aim to explore in future sections are:

- Does information other than lexical domain match help domain adaptation for CLIR? If so, to what extent?
- Can we learn the relative importance of such adaptation criteria to produce optimal performance in domain specific CLIR tasks?
- What are the trade-offs and limitations of using a local vs. global optimization method?

7.3 Continuous Reactive Tabu Search for Criteria Optimization

We are optimizing CLIR mean average precision $f : \omega \rightarrow R$, where ω is the set of feasible points and a subset of R^n, defined by bounds on the n weights w_i: $0 \leq w_i \leq 1$. Our function f's convexity and differentiability cannot be relied upon, therefore algorithms such as simple hill climbing are not recommended. We use CRTS [2], a global, deterministic, tabu-search based optimization method that uses the reactive affine shaker (RASH) [6] algorithm as its local optimization routine. CRTS focuses on locating the set of promising boxes in the search space, and it initializes RASH while adapting box size and other search parameters.

7.3.1 RASH

RASH [6] is an adaptive random optimization algorithm for functions of continuous variables. For our purposes, the crucial advantages of RASH are a) its support for functions that are discontinuous or non-differentiable, requiring only the availability of the optimized function value at a given point and b) the assumption that the main computational cost lies in the function evaluation. In our case, the function evaluation is the entire adaptation, training and testing process - clearly a very expensive function call. Other than being a local optimization algorithm, RASH is therefore perfectly suited for black-box, computationally expensive CLIR adaptation and evaluation.

The RASH original pseudocode is included in Figure 7.1. The main idea behind RASH is to adapt the step size and direction to maintain the largest possible movement per function evaluation, given uniformly distributed "trial points". After an initial function evaluation at a randomly chosen point (x), a new point is generated within the search region by sampling it with a uniform probability distribution, and testing function improvement at both $x+\delta$ and, if not successful, at $x-\delta$. This choice drastically reduces the probability of generating two consecutive unsuccessful samples ([6]). Then, the search region is adapted to the outcome of the tentative point. If the sampling is successful (i.e. the new function value is better), the region is expanded along the promising direction; if not, the region shrinks. The search area undergoes an affine (i.e. linear followed by translation) transformation so that the successful point becomes the center of the region.

7.3.2 CRTS

CRTS[2] is a hybrid between a combinatorial optimization component (reactive tabu search (RTS)) and RASH, the local optimization algorithm (RASH).

RTS activates RASH within a sub-region of the search space when it estimates that the sub-region contains a good local optimum. RTS uses an iterative modified greedy search to bias the search towards good f values; the tabu component refers to the prohibition of visiting previously seen points. The CRTS hybrid creates a tree of search boxes (where the

procedure affine_shaker

comment: *The constant adjustment factors for the search region* Γ *are* $\rho_{expand} = 2$, $\rho_{compress} = 1/2$.

Initialize :

$t \leftarrow 0$ *(iteration counter)*;

$X \leftarrow$ *(initial random point)*;

$\forall j \quad \vec{b}_j \leftarrow \vec{e}_j \times (1/4) \times$ (range of j-th variable) where $\vec{e}_j$ is the canonical basis of R^N

repeat

("double shot" strategy:)

$\vec{\delta} = \sum_j rand_j \times \vec{b}_j$ *($rand_j$ = random numbers in (-1,1))*

if $f(X + \vec{\delta}) < f(X)$ **then**

$X \leftarrow X + \vec{\delta}$ *(first "shot" successful)*

$P \leftarrow I + (\rho_{expand} - 1)\frac{\vec{\delta}\vec{\delta}^T}{\|\vec{\delta}\|^2}$ *(expand Γ)*

$\forall i \ \vec{b}_i \leftarrow P \vec{b}_i$

else if $f(X - \vec{\delta}) < f(X)$ **then**

$X \leftarrow X - \vec{\delta}$ *(second "shot" successful)*

$P \leftarrow I + (\rho_{expand} - 1)\frac{\vec{\delta}\vec{\delta}^T}{\|\vec{\delta}\|^2}$ *(expand Γ)*

$\forall i \ \vec{b}_i \leftarrow P \vec{b}_i$

else *(no success)*

$P \leftarrow I + (\rho_{compress} - 1)\frac{\vec{\delta}\vec{\delta}^T}{\|\vec{\delta}\|^2}$ *(compress Γ)*

$\forall i \ \vec{b}_i \leftarrow P \vec{b}_i$

$t \leftarrow (t + 1)$ *(increment iteration counter)*

until convergence criterion is satisfied

Figure 7.1: Original RASH pseudocode. See explanation in Section 7.3.1.

children are subdividing a box), and generates a path to reach a leaf. RTS identifies the promising boxes after sampling them, and launches RASH. CRTS is fully described in [2]; the RASH original pseudocode is included in Figure 7.2.

7.3.3 CLIR System Adaptation Specifics

We optimize criteria weights by wrapping the CLIR task in a CRTS optimization shell. Due to the repeated evaluation of the objective function in the optimization process, a fast-training CLIR system is desirable. The PMI system, described in Chapter 5 is extremely fast and since encouraging results are observed when the training parallel corpus is small,

```
procedure reactive_tabu_search
(Initialize the data structures for tabu:)
t ← 0 (iteration counter); T_F^(0) ← 1/N (fractional prohibition period); t_T ← 0 (last time T_F was changed);
C ← ∅ (set of often-repeated configurations); R_ave ← 1 (moving average of repetition interval);
X^(0) ← random X ∈ X (initial configuration); X_b ← X^(0) (best so far X); f_b ← f(X^(0)) (best so far f);
repeat
    (See whether the current configuration is a repetition:)
    escape ← memory_based_reaction(X^(t))     (see Fig. 9)
    if escape = DO_NOT_ESCAPE then
        μ ← arg min_{ν∈A^(t)} f(ν(X^(t)))
        X^(t+1) = μ(X^(t))
        Λ(μ) ← t        (A^(t) and T^(t) are therefore implicitly changed)
        (Update time, and best_so_far:)
        t ← (t + 1)
        if f(X^(t)) < f_b then
            f_b ← f(X^(t))
            X_b ← X^(t)
    else
        diversify_search
until f_b is acceptable or maximum no. of iterations reached


function diversify_search
comment: A sequence of random steps, that become tabu as soon as they are applied.
    repeat for i = 1 to Max(2, n_max N/4)     (at least two moves)
        σ_i ← move corresponding to a random step        (see text for details)
        X^(t+1) ← σ_i(X^(t))
        Λ(σ_i) ← t    ( A^(t) and T^(t) are therefore changed)
        (Update time, and best_so_far:)
        t ← (t + 1)
        if f(X^(t)) < f_b then
            f_b ← f(X^(t))
            X_b ← X^(t)
```

Figure 7.2: Original CRTS pseudocode. See explanation in Section 7.3.2.

but well chosen, we choose 100 sentences as the size of the training corpus. Adaptation effects are most important at this size, and are crucial when adaptation is used as an active learning scenario (e.g. when the system chooses monolingual sentences to be translated by a very costly domain expert). The dimensions used in this experiment are MRR, translation quality estimate (bootstrap and evaluation) (TQE) and size (log (sentence length)).

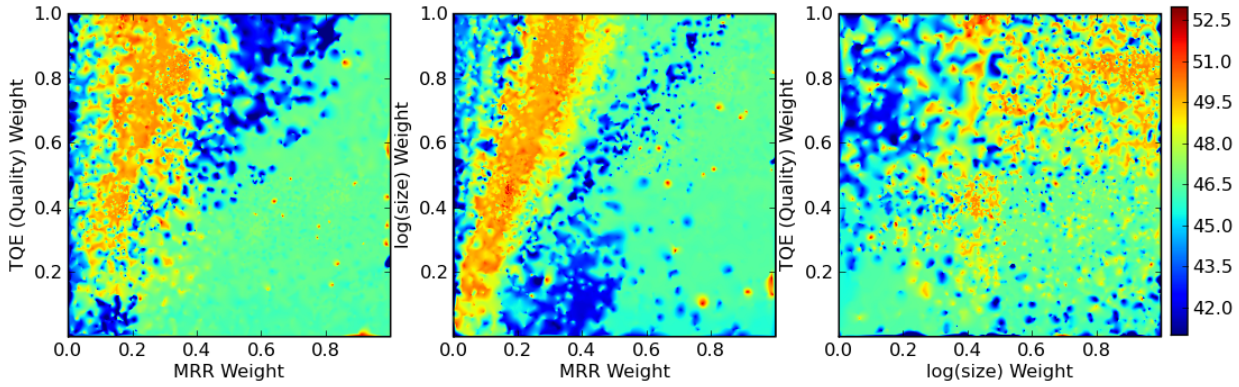

Figure 7.3: 2-D projections of the 3-D search parameter space and the corresponding CLIR performance. The color represents average precision of the CLIR system when trained with 100 sentences selected using the criteria weights shown on the X and Y axes. MRR represents lexical domain match; TQE is the estimate translation quality for the candidate sentence; size is the (log of) the length of the sentence. (best viewed in color).

7.3.4 Using RASH and CRTS to Optimize Criteria Weights: Experiments and Results

In this section, we compare CRTS with its local optimization component, the Reactive Affine Shaker algorithm. Since the objective function evaluation (CLIR average precision, see beginning of this section) is costly, we are trying to minimize the number of iterations - therefore, quick convergence to a good local optimum is a desirable feature. Figure 7.4 shows how fast each of the two optimization methods find a better objective function value.

CRTS samples the search space faster and quickly (in less than 10 iterations) finds a good, although not optimal, value. RASH, on the other hand, finds the global optimum faster than CRTS. However, in most cases RASH cannot be assumed to find the global optimum, since it is a local optimization method.

We explore the (training) parameter space sampled by CRTS and RASH in Figure 7.3. The three figures are interpolated heatmaps, with warmer colors representing better average precision on the training set. In each of the three figures, the data is projected from the three dimensions (MRR, TQE and size) into two. Values projected into the same point are

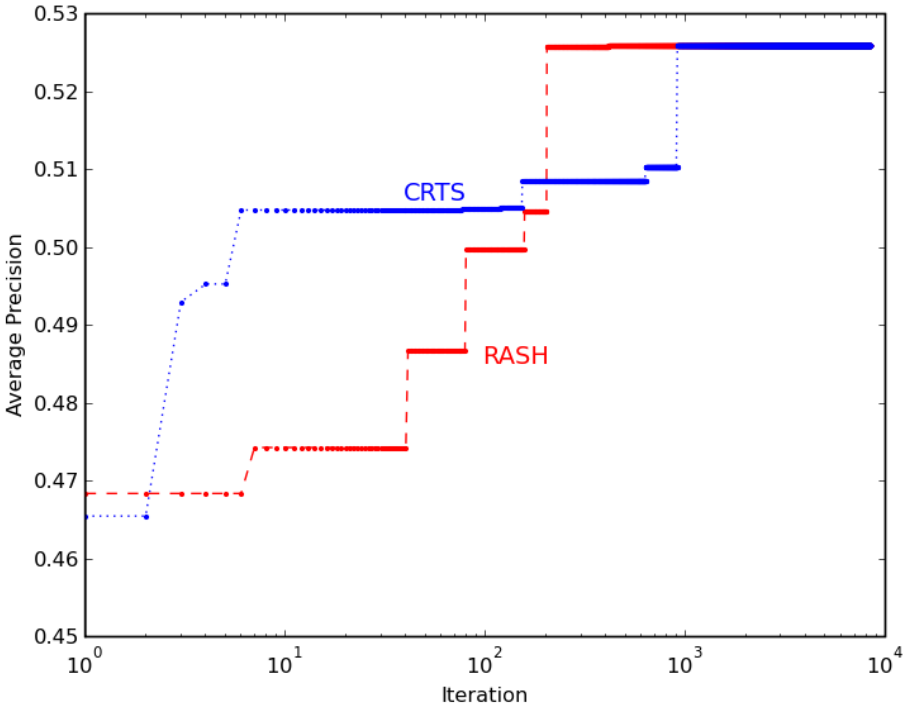

Figure 7.4: RASH vs. CRTS: Best function values found at each iteration. The X axis shows on a log scale the number of iterations (objective function evaluations), and the Y axis shows mean average precision on the training queries. The step function behavior corresponds to changes in ranking of the relevant documents.

averaged.

MRR	TQE	Size
0.17	1.0	0.46

Table 7.1: Final criteria weights optimized by CRTS & RASH on the training data.

	MRR Only	Equal	Optimized
training	0.4953	0.4538	0.5204
testing	0.4642	0.4667	0.5259

Table 7.2: CLIR Average precision for criteria combination experiments for both training and testing scenarios. Optimized weights yield better results than giving equal importance to all criteria, or than using lexical similarity alone. The optimized criteria weights found in the training process are shown in Table 7.1.

When interpreting the heatmaps in Figure 7.3, it is important to keep in mind the underlying 3-D surface. For example, the third plot (TQE weight vs. log(size) weight) helps explain the red and blue band in the first plot (TQE weight vs. MRR weight): the red band is the projection of CLIR performance when the weight of the size criterion is large, while the blue band is the projection of CLIR performance when the weight of the size criterion is small. It is interesting to observe that, if the MRR weight is non-zero (in other words, the domain match is not ignored), better results are obtained when its corresponding weight is kept small relative to quality estimate and size. However, the question is whether these results optimized on training data allow us to take advantage of non-MRR criteria when adapting and evaluating the test queries.

In this case, both RASH and CRTS reached the same optimum. The optimized criteria weights found in the training process are shown in Table 7.1.

We use them to adapt a 100-sentences parallel corpus to the test queries, then we evaluate the CLIR performance. This performance, along with two high baselines, is shown in Table 7.2.

It is important to note that the "equal weights" baseline, which we have used earlier in this section to mix criteria, significantly underperforms the MRR-only condition. This condition is, in turn, outperformed by our optimized criteria weights.

Chapter 8

Conclusions

In this dissertation we have focused on domain specific, corpus-based multilingual applications (more specifically, cross-language information retrieval and statistical machine translation). We have shown that high-performing, but general-domain systems have disappointing results when confronted with this class of problems. This poor performance pinpoints *domain adaptation* as a necessary component in statistical translation-based systems. Although manual domain adaptation is possible and has been the prevailing approach, our goal is to tackle this issue in an automated fashion.

To that end, we have introduced a multi-faceted adaptation framework that addresses the problem of *translation model adaptation* in the context of multilingual applications based on parallel corpora. The Parallel Resource Domain Adaptation (PARDA) framework allows the integration of several characteristics of parallel corpora (or parallel aligned units) by using them as criteria in a flexible scoring model. Its goal is that of automated selection of the best building blocks for assembling a high quality, domain specific, customized parallel resource. The evaluation characteristics or criteria include: lexical similarity between the selection candidate and the domain sample, candidate translation quality estimate, candidate size, and similarity between the taxonomy projections of the candidate and the domain sample.

We have introduced a new, million-sentence medical domain parallel corpus, annotated with MeSH taxonomy information, and we have used it to examined the effect that each criterion (and their combination) has on both CLIR and MT performance. By adapting the

translation model, we significantly alleviate the considerable performance penalty incurred when using a proven high-performance (but general domain) CLIR or MT systems on domain specific test data.

When data is costly to translate - as it usually is the case with domain-specific data - or when only limited quantities are available, appropriate selection is crucial. For CLIR, the results obtained under these circumstances show that relatively few selected sentences can lead to 90% of the performance obtained with significantly more *in-domain* data. We have explored the effect of several adaptation criteria on CLIR performance before criteria optimization. In particular, experiments show that the translation quality criterion can compensate for inconsistent or poor quality parallel corpora when integrated in the selection model. In our particular domain specific machine translation task, our selection method allows a two-order of magnitude reduction in training data with only a 10% BLEU-score decrease, vs. a 35% decrease for an equivalent quantity of *in-domain* data.

We have also addressed the experimental differences between a given domain (the medical field), and a sub-domain (in our case, the heart-related medical domain subset). We have also explored the use and effect of a domain-specific taxonomy or ontology in the training set selection.

A challenging problem in the PARDA framework is learning the relative importance of miscellaneous criteria incorporated in a domain adaptation method for cross-language IR. We have used global and local non-linear optimization methods in order to find optimal weights assigned to criteria such as lexical similarity, corpus quality and instance size. We have compared two learning methods, one local (Reactive Affine Shaker) and one global (continuous reactive tabu search) and their effect on optimizing the criteria weights used in CLIR medical domain adaptation. Our optimization methods (RASH & CRTS) allow us to find better mixing criteria weights. On the test set, optimized weights outperform the two high baselines represented by a) giving criteria equal importance and b) using lexical similarity alone.

Through the PARDA framework, we have provided a tool that, given a domain sample

and a pool of parallel resources, produces a customized parallel corpus tailored to the given domain sample. Through a flexible, trainable set of evaluation criteria, we accomplish the two-fold goal of 1) significantly alleviating the considerable performance penalty incurred by general-domain systems, and 2) prioritizing training instances such that a significant reduction in training resources results in 10% performance difference, even when compared to using all of the available domain-specific data.

8.1 Future Directions

There are many possible future directions that can build upon our domain adaptation work. They include:

- Exploring various other adaptation criteria in order to establish their importance and impact within the PARDA framework. Such criteria include redundancy, genre, data sources etc.

- Cross-domain experiments with additional exploration of sub-domains, similar to the work presented here for a medical sub-domain.

- Exploring the stopping criterion problem: i.e. *how much* data should one select? What are the tradeoffs in speed and accuracy relevant to this issue?

- One interesting application of our adaptation framework is as an active learning tool - i.e. to what extent using PARDA on monolingual collections allows the rapid, short cycle training of domain-specific translation models? In the case of domain-specific translation, it is important to minimize the quantity of data to be translated, as the cost per unit is likely higher than general domain translation.

- Generalizing the selection problem to an instance weighting problem taking all criteria into account. This is particularly useful when documents or entire collections (as opposed to sentences) are the granularity at which the adaptation process is performed.

8.2 Impact and Significance to the Broader Research Community

Domain specific translation and cross-language retrieval are crucial for a variety of applications: translation of technical manuals, patents, and medical records and literature. The impact of a properly domain-adapted translation model is significant in these areas, and this thesis provides a flexible and adaptable mechanism for automated constructions of such models.

Translating domain-specific data using human translators is expensive, since the translators need to be trained in a specific domain's vocabulary. Generally, the more technical the domain, the more expensive the translation. The impact of domain-adapted training resources therefore *increases* as the domain becomes more technical, and the availability of a collection of algorithms to select training data tailored to a given domain sample leads to significant cost savings in both time and resources.

A significant contribution of this work to the broader research community is challenging the conventional wisdom that more training data is always better. We have shown that having the *right* training data is crucial, and that it is possible for multilingual applications to yield superior results with less data. We have built an argument for domain adaptation as a standard pre-processing step that should be performed and experimented with as part of the data preparation phase of any application relying on parallel corpora as training data.

Bibliography

[1] Ricardo Baeza-Yates and Berthier Ribeiro-Neto, *Modern information retrieval*, Addison Wesley / ACM Press, 1999.

[2] R. Battiti and G. Tecchiolli, *The continuous reactive tabu search: Blending combinatorial optimization and stochastic search for global optimization*, 1995.

[3] J. Blitzer, R. McDonald, and F. Pereira, *Domain adaptation with structural correspondence learning*, EMNLP, 2006.

[4] P.F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer, *The mathematics of statistical machine translation: Parameter estimation*, Computational Linguistics **19** (1993), no. 2, 263–311.

[5] Ralph D. Brown, *EBMT Tutorial*, The Association for Machine Translation in the Americas, 2002.

[6] Mauro Brunato and R. Battiti, *The reactive affine shaker: a building block for minimizing functions of continuous variables*, 2006.

[7] J. Callan, F. Crestani, and M. Sanderson, *SIGIR 2003 workshop on distributed information retrieval*, SIGIR Forum, vol. 37, 1, no. 2, 2003.

[8] Yee Seng Chan and Hwee Tou Ng, *Estimating class priors in domain adaptation for word sense disambiguation*, The Association for Computational Linguistics Conference (ACL), 2006.

[9] Ciprian Chelba and Alex Acero, *Adaptation of maximum entropy capitalizer: Little data can help a lot*, Conference on Empirical Methods in Natural Language Processing (EMNLP), 2004.

[10] A. Chen, H. Jiang, and F. Gey, *Combining multiple sources for short query translation in chinese-english cross-language information retrieval*, Fifth International Workshop on Information Retrieval with Asian Languages, 2000.

[11] Aitao Chen and Fredric C. Gey, *Combining query translation and document translation in cross-language retrieval*, Cross Language Evaluation Forum (CLEF), 2003, pp. 108–121.

[12] CLEF, *Workshop of the cross-lingual evaluation forum*, http://www.clef-campaign.org.

[13] K. Darwish and D. Oard, *CLIR experiments at maryland for TREC-2002: Evidence combination for arabic-english retrieval*, Text REtrieval Conference (TREC), 2002.

[14] M. Eck, S. Vogel, and A. Waibel, *Language model adaptation for statistical machine translation based on information retrieval*, International Conference On Language Resources And Evaluation (LREC), 2004.

[15] Search Engine Land Blog Entry, *"http://searchengineland.com/070516-180352.php"*, Search Engine Land Blog Entry, 2007.

[16] M. Franz, J. S. McCarley, and S. Roukos, *Ad hoc and multilingual information retrieval at IBM*, Text REtrieval Conference (TREC), vol. NIST Special Publication 500-242, 1, no. November, 1998, pp. 157–168.

[17] M. Franz and J.S. McCarley, *Arabic information retrieval at IBM*, Text REtrieval Conference (TREC), 2002.

[18] A. Fraser, J. Xu, and R. Weischedel, *TREC 2002 cross-lingual retrieval at BBN*, Text REtrieval Conference (TREC), 2002.

[19] A. Fujii, N. Kando, and M. Iwayama, *Building a test collection for associative patent retrieval in NTCIR-4*, NTCIR-4, 2005.

[20] Atsushi Fujii, Makoto Iwayama, and Noriko Kando, *The patent retrieval task in the fourth NTCIR workshop*, SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval (New York, NY, USA), ACM, 2004, pp. 560–561.

[21] Jade Goldstein, Gary M. Ciany, and Jaime G. Carbonell, *Genre identification and goal-focused summarization*, CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management (New York, NY, USA), ACM, 2007, pp. 889–892.

[22] Hal Daume III and Daniel Marcu, *Domain adaptation for statistical classifiers*, Journal of Artificial Intelligence Research, 26, 2006, pp. 101–126.

[23] Hal Daum III, *Frustratingly easy domain adaptation*, The Association for Computational Linguistics Conference (ACL), 2007.

[24] S. Kubler R. McDonald J. Nilsson S. Riedel J. Nivre, J. Hall and D. Yuret, *The CoNLL 2007 shared task on dependency parsing*, Shared Task - Conference on Natural Language Learning - CoNLL, 2007.

[25] Jing Jiang and ChengXiang Zhai, *Instance weighting for domain adaptation in NLP*, The Association for Computational Linguistics Conference (ACL), 2007.

[26] Alex Kulesza Fernando Pereira John Blitzer, Koby Crammer and Jennifer Wortman, *Learning bounds for domain adaptation*, Advances in Neural Information Processing Systems(NIPS), 2008.

[27] Mark Dredze John Blitzer and Fernando Pereira, *Biographies, bollywood, boom-boxes, and blenders: Domain adaptation for sentiment classification*, The Association for Computational Linguistics Conference (ACL), 2007.

[28] J. Jutras, *An automatic reviser: The transcheck system*, 2000.

[29] D. Marcu K. Knight, *Machine translation in the year 2004*, International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2005.

[30] Noriko Kando, *Overview of the third NTCIR workshop*, Working notes of the Third NTCIR Workshop Meeting. Part I: Overview, 2002.

[31] David Kauchak, *Contributions to research on machine translation*, Ph.D. thesis, University of California, San Diego, 2006.

[32] S. Khudanpur and W. Kim, *Using cross-language cues for story-specific language modeling*, International Conference on Spoken Language Processing, 2002, pp. 513–516.

[33] Khudanpur, S. and Kim, W., *A maximum entropy language model to integrate n-grams and topic dependencies for conversational speech recognition*, the IEEE International Conference on Acoustics, Speech and Signal Processing, 1999, pp. 553–556.

[34] W Kim and S. Khudanpur, *Language modeling adaptation using cross-lingual information retrieval*, International Conference On Language Resources And Evaluation (LREC), 2004.

[35] M. Kluck and F. Gey, *The domain-specific task of clef - specific evaluation strategies in cross-language information retrieval*, Cross Language Evaluation Forum (CLEF), 2000.

[36] Philipp Koehn, *Europarl: A multilingual corpus for evaluation of machine translation*, Draft, Unpublished.

[37] P. Kohen and C. Monz, *NAACL 2006 workshop on statistical machine translation*, NAACL 2006 Workshop on Statistical Machine Translation, 2006.

[38] M. Lease and E. Charniak, *Parsing biomedical literature*, International Joint Conference on Natural Language Processing (IJCNLP), 2005.

[39] Lillian Lee, *Measures of distributional similarity*, 37th Annual Meeting of the Association for Computational Linguistics, 1999, pp. 25–32.

[40] Lucian Vlad Lita, Monica Rogati, and Jaime G. Carbonell, *Cross lingual qa: A modular baseline in CLEF 2003*, Cross Language Evaluation Forum (CLEF), 2003.

[41] Partha Pratim Talukdar Kuzman Ganchev Joao Graca Mark Dredze, John Blitzer and Fernando Pereira, *Frustratingly hard domain adaptation for parsing*, Shared Task - Conference on Natural Language Learning - CoNLL, 2007.

[42] Wolfram MathWorld, *Global optimization*, http://mathworld.wolfram.com/ GlobalOptimization.html.

[43] I. Melamed, *Automatic detection of omissions in translations*, 1996.

[44] I. Dan Melamed, *Statistical machine translation by parsing*, The Association for Computational Linguistics Conference (ACL), 2004.

[45] Monica Rogati and Yiming Yang, *Multilingual information retrieval using open, transparent resources in CLEF 2003*, Cross Language Evaluation Forum (CLEF), 2003.

[46] Monica Rogati and Yiming Yang, *Resource selection for domain-specific cross-lingual IR*, ACM SIGIR Special Interest Group on Information Retrieval Conference (SIGIR 2004), 2004.

[47] Jian-Yun Nie, Michel Simard, Pierre Isabelle, and Richard Durand, *Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the Web*, SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval (New York, NY, USA), ACM, 1999, pp. 74–81.

[48] NIH, *Medical subject headings (MESH)*, http://www.nlm.nih.gov/mesh, 2007.

[49] Franz Josef Och, *Minimum error rate training in statistical machine translation*, ACL '03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics (Morristown, NJ, USA), Association for Computational Linguistics, 2003, pp. 160–167.

[50] Franz Josef Och, Daniel Gildea, Sanjeev Khudanpur, Kenji Yamada, Alex Fraser, Shankar Kumar, David Smith, Katherine Eng, Viren Jain, Zhen Jin, and Dragomir Radev, *Final report of Johns Hopkins 2003 summer workshop on syntax for statistical machine translation*, Summer Workshop on Syntax for Statistical Machine Translation, 2004.

[51] Franz Josef Och and Hermann Ney, *Improved statistical alignment models.*, The Association for Computational Linguistics Conference (ACL), 2000.

[52] K. Papineni, S. Roukos, T. Ward, and W. Zhu, *BLEU: a method for automatic evaluation of machine translation*, 2001.

[53] Philipp Koehn, *Pharaoh: A beam search decoder for phrase-based statistical machine translation models.*, AMTA, 2004, pp. 115–124.

[54] Philipp Koehn, *Statistical significance tests for machine translation evaluation*, Conference on Empirical Methods in Natural Language Processing (EMNLP) (Barcelona, Spain), 2004.

[55] Ralph D. Brown, *Transfer-rule induction for example-based translation*, MT Summit VIII Workshop on Example-Based Machine Translation, 2001, pp. 1–11.

[56] Philip Resnik, *Parallel strands: A preliminary investigation into mining the web for bilingual text*, The Association for Machine Translation in the Americas, 1998, pp. 72–82.

[57] Philip Resnik and Noah A. Smith, *The Web as a parallel corpus*, Computational Linguistics, vol. 29, 1, no. 3, 2003, pp. 349–380.

[58] Stephen E. Robertson, Steve Walker, Micheline Hancock-Beaulieu, Aarron Gull, and Marianna Lau, *Okapi at TREC*, Text REtrieval Conference, 1992, pp. 21–30.

[59] Monica Rogati and Yiming Yang, *Customizing parallel corpora at the document level*, Annual Meeting of the Association for Computational Linguistics (ACL 2004), 2004.

[60] Ronald Rosenfeld, *Adaptive statistical language modeling: A maximum entropy approach*, Ph.D. thesis, Carnegie Mellon University, 1994.

[61] Jacques Savoy, *A stemming procedure and stopword list for general french corpora*, J. Am. Soc. Inf. Sci. **50** (1999), no. 10, 944–952.

[62] Koby Crammer Shai Ben-David, John Blitzer and Fernando Pereira, *Analysis of representations for domain adaptation*, Advances in Neural Information Processing Systems(NIPS), 2007, pp. 137–144.

[63] Noah Smith, *Detection of translational equivalence*, Technical report 4253, University of Maryland College Park Computer Science Department, College Park, MD, 2001.

[64] H. Turtle T. Strohman, D. Metzler and W. B. Croft, *Indri: A language model-based search engine for complex queries*, International Conference on Intelligence Analysis, 2005.

[65] Yiming Yang and Jan O. Pedersen, *A comparative study on feature selection in text categorization*, 1997.

[66] Ying Zhang and Stephan Vogel, *Measuring confidence intervals for the machine translation evaluation metrics*, International Conference on Theoretical and Methodological Issues in Machine Translation (TMI 2004), Baltimore, MD USA, October 4-6, 2004, 2004.