Translation Techniques in Cross-Language Information Retrieval

DONG ZHOU
Trinity College Dublin and Hunan University of Science and Technology
MARK TRURAN
Teesside University
TIM BRAILSFORD
University of Nottingham
VINCENT WADE
Trinity College Dublin
and
HELEN ASHMAN
University of South Australia

Cross-language information retrieval (CLIR) is an active sub-domain of information retrieval (IR). Like IR, CLIR is centred upon the search for documents, and for information contained within those documents. Unlike IR, CLIR must reconcile queries and documents which are written in *different languages*. The usual solution to this mismatch involves *translating* the query and/or the documents before performing the search. Translation is therefore a pivotal activity for CLIR engines. Over the last 15 years, the CLIR community has developed a wide range of techniques and models supporting free text translation. This paper presents an overview of those techniques, with a special emphasis on recent developments.

Categories and Subject Descriptors: A.1 [**General Literature**]: Introduction and survey; H.3.1 [**Content Analysis and Indexing**]: Dictionaries; Thesauruses; Linguistic Processing; I.2.7 [**Natural Language Processing**]: Machine translation; Text Analysis; Language Parsing and Understanding

General Terms: Algorithms, Languages

Additional Key Words and Phrases: Cross-language information retrieval, query translation, document translation, bilingual dictionary, parallel corpora, machine translation, semantic model

1. INTRODUCTION

The World Wide Web is polyglot in nature. It expresses all human languages, it speaks all dialects. By comparison, the average Web user is often a monoglot, restricted to just one native language, or a handful at best. This mismatch places a

Author's address: Dong Zhou, School of Computer Science and Engineering, Hunan University of Science and Technology, Xiangtan, Hunan, China; email: dongzhou1979@hotmail.com.

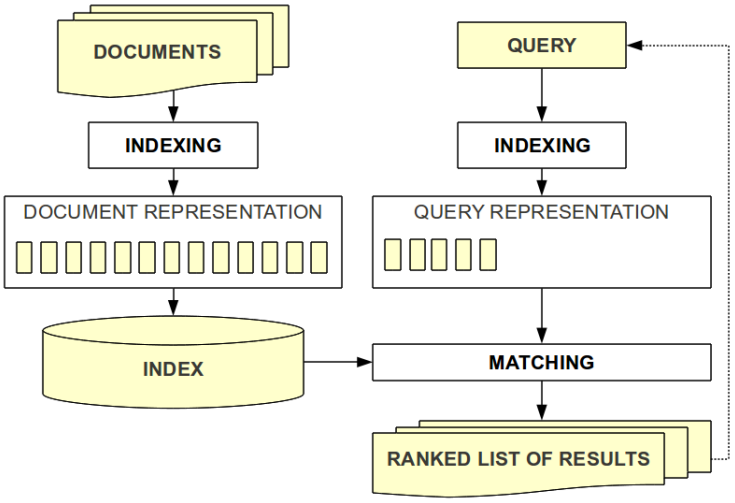

Fig. 1: Monolingual information retrieval

structural barrier between the user and swathes of globally available (yet linguistically impenetrable) information. One partial solution to this barrier is the development and progressive refinement of *cross-language information retrieval engines*. CLIR engines provide a mechanism through which information can be accessed[1] regardless of the language in which it is authored.

A cross-language information retrieval engine is a *specialisation* of a traditional information retrieval system. As illustrated in Figure 1,[2] traditional IR assumes the existence of a query (which expresses the user's information need) and a set of documents (known as a document collection or corpus). These components are processed into internal representations suitable for efficient comparison. The technique of deriving document representations from a corpus is known as *indexing*. Indexing involves extracting terms, phrases and concepts from the collection and recording this information in a format permitting rapid access. Query representations work in a very similar fashion, albeit on a much smaller scale. Using a variety of different approaches, these representations are subsequently compared to determine the 'best fit'. Documents that appear to match the query are then passed to the user, usually in the form of a ranked list. At this point, the user is often given an opportunity to respond to the subset of documents generated by his/her query, providing *feedback* that can iteratively improve the results of their retrieval operation.

[1]The CLIR community often replaces references to 'information retrieval' (i.e. the science of finding documents) with the more specific 'information access' (i.e., the science of finding documents and *rendering them usable*) [Peters and Sheridan 2001]. In this article, we treat these two terms as interchangeable.

[2]Figures 1-4 reproduced from (or inspired by) '*Best Practices in System-oriented and User-oriented Multilingual Information Access*', http://www.trebleclef.eu/getfile.php?id=254.

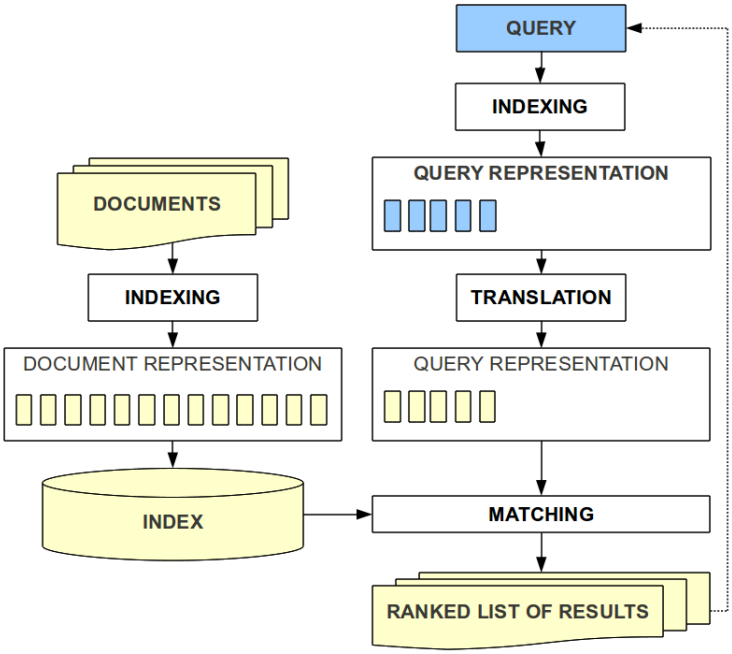

Fig. 2: Cross-language information retrieval utilising query translation

By comparison, in cross-language information retrieval, there is a *linguistic disparity* between the queries that are submitted and the documents that are retrieved. To resolve this disparity, CLIR engines are normally required to incorporate some facility for language translation, an obvious requirement if query representations and document representations are to be meaningfully compared. There are three general approaches to translation that can be employed at this point:

(1) Translate the query representation to match the document representations (as shown in Figure 2).

(2) Translate the document representations to match the query representation (as shown in Figure 3).

(3) Translate the document and the query representations into a *third* language or semantic space, which we will label *dual translation* (as shown in Figure 4).

Historically, the CLIR community has tended to favour *query translation*, probably because it offers a computationally economical solution to the mis-match problem. However, document translation and dual translation are still very much live research topics.

1.1 Scope of this paper

In this paper we survey the various translation techniques used in free-text cross-language information retrieval. The more limited field known as controlled-vocabulary

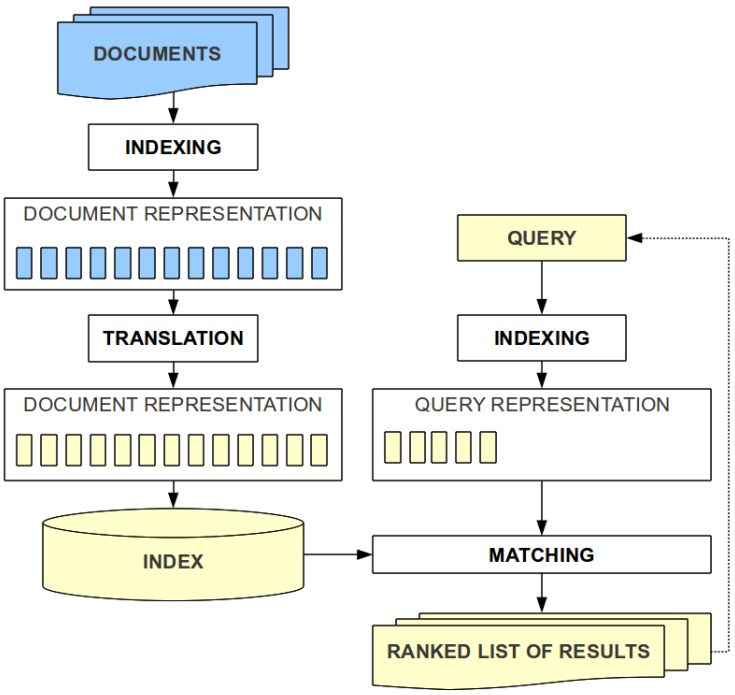

Fig. 3: Cross-language information retrieval utilising document translation

CLIR (in which a searcher is restricted to a fixed vocabulary relating to a constrained, multilingual domain) is not discussed, chiefly because it involves a set of indexing and translation procedures that are simply not applicable to the modern World Wide Web.[3] Furthermore, although the label of 'cross-language information retrieval' covers many diverse tasks and services, this survey concentrates on translation techniques as applied in *ad hoc* cross-language retrieval. This basic form of text retrieval remains the fundamental challenge for the field, providing the underlying foundation for all other CLIR applications [Peters and Sheridan 2001]. An exhaustive survey of cross-language information retrieval needs to cover a huge range of topics (e.g., indexing, query analysis, results filtering, feedback, language resources). This survey is *not* intended to be an exhaustive study.[4] Our goal is to provide a straightforward guide to the translation techniques and models currently used in CLIR.

[3][Oard and Dorr 1996] provide an informative discussion of early controlled vocabulary CLIR systems

[4]Readers interested in the broader picture are directed towards the various outcomes of the TrebleCLEF project (http://www.trebleclef.eu/). For example, '*Language Resources for Multilingual Information Access*' (http://www.trebleclef.eu/getfile.php?id=255), which surveys various language resources and natural language processing tools.

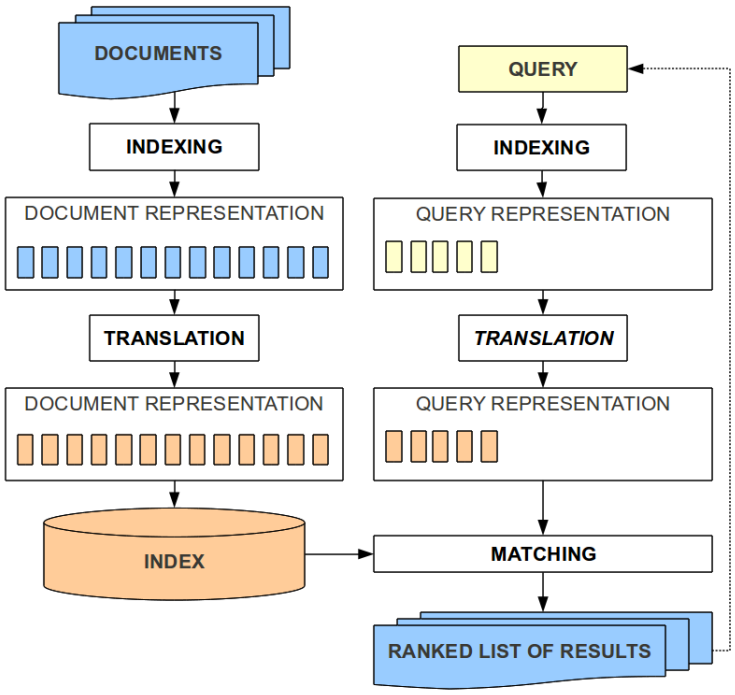

Fig. 4: Cross-language information retrieval utilising dual translation

1.2 Structure of the paper

The remainder of the paper is structured as follows - §2 provides a conceptual overview of the cross-language retrieval process, with an emphasis on issues relating to translation; §3 describes various methods which can be used to evaluate the output of a translation system in the context of CLIR; §4 outlines a taxonomy of translation models and techniques; §5-6 elaborate on this taxonomy (chiefly in relation to query translation), providing commentary and notation; §7 examines the subject of *document translation*; §8 concludes the survey with a look at the future of the field.

2. A CONCEPTUAL OVERVIEW OF THE CLIR PROCESS

A translation-oriented cross-language information retrieval engine can be conceptualized as a framework containing a number of discrete *modules*. These modules are serially connected, so that the output of one module constitutes the input to the next (see Figure 5). There are four major modules in most CLIR engines, as follows:

—A pre-translation module
—A translation module
—A post-translation module
—An information retrieval module

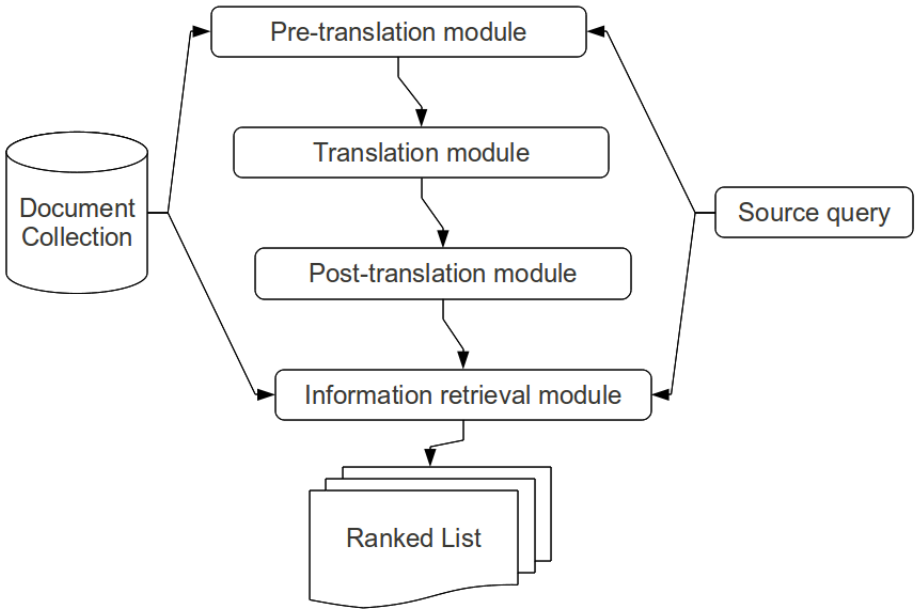

Fig. 5: CLIR - A conceptual framework

The following sections explain the purpose of each module in detail. Please note that all references to the '*source text*' in these sections encompass the query and/or the document collection.

2.1 Pre-Translation Module

The pre-translation module is responsible for identifying, extracting and processing suitable linguistic units present in the source text. This activity can be broken down into four separate activities - tokenisation, stop word removal, stemming and term expansion. We discuss each activity in turn.

2.1.1 *Tokenisation.* Tokenisation is an attempt to recognise and isolate the various linguistic units present in the source text. The two main tokenisation procedures, which are very similar in computational terms, are *word segmentation* and *decompounding*. Segmentation is typically applied to Eastern Asian languages, while decompounding is usually reserved for agglutinative languages common to Europe.

Word segmentation is simply the process of separating the constituent terms in the text. This is relatively simple for languages that indicate explicit word boundaries using white space (e.g., English, French), but very difficult for languages in which terms are concatenated with adjacent terms (e.g., Chinese, Japanese). Queries and documents that belong to the latter group of languages usually require sophisticated segmentation tools. One approach to segmentation uses a maximal matching algorithm to identify potential segmented terms using a list of known

words.[5] Obviously this approach will fail if the potential segmented terms are not contained in the lexicon. Alternative approaches that match single characters (unigrams), pairs of characters (bigrams) or combinations of the two have been documented [Nie and Ren 1999; Shi et al. 2007; Kwok and Grunfeld 1996; McNamee and Mayfield 2004a].

Certain languages (e.g., German, Dutch, Russian) are rich in *compound words* (e.g., the German for *substitute coffee*, 'der Kaffee-Ersatz', is compounded to 'der Kaffeeersatz'). Unless a CLIR engine decompounds (separates) these words before translating the source text, retrieval effectiveness is likely to suffer. A representative decompounding algorithm can be found in [Chen 2002]. In that paper, the authors used a monolingual dictionary containing uncompounded German words in various forms. Text was segmented into the minimal number of words present in this base dictionary. If the algorithm found two (or more) possible decompositions of the same chunk of text, it would select the alternative with the highest probability, with probability values generated using corpus frequency analysis. The authors found that decompounding led to significantly better results when using German and Dutch document collections [Chen 2002]. Positive results were also demonstrated in [Hollink et al. 2004; Braschler and Ripplinger 2004] and [Alfonseca et al. 2008], where the authors used text extracted from Web anchors to improve the decompounding of various European languages. The combinatorial behaviour of compounds in the setting of cross-lingual retrieval was explored by [Hedlund 2002].

2.1.2 *Stop Word Removal.* Stop word removal is possibly the simplest natural language processing (NLP) technique used in cross-language information retrieval. Prepositions, articles, pronouns, conjunctions, common verbs and non-significant words are usually removed from the source text before it is translated. Removal of these terms is typically implemented using a *stop word list*. General guidelines for the production of these lists can be found in [Fox 1989]. Stop word lists for a number of major world languages have been developed and are freely available to CLIR researchers. Most approaches to translation in the context of CLIR, with the notable exception of machine translation (MT), incorporate stop word filtering.

2.1.3 *Stemming.* Stemming is the process of removing inflectional and derivational affixes. The output of this process is a *word stem*, which may or may not be a meaningful word. The technique of stemming was originally developed for monolingual IR. In that context, stemming is generally acknowledged as a tool for improving the effectiveness of a search engine because it tends to produce more potentially relevant documents [Hull 1996; Melucci and Orio 2003; Savoy 2007; Fautsch and Savoy 2009]. Early stemming algorithms were designed for the English language. Following their success, stemmers were developed for a number of other major world languages.

Stemmers can be divided into two distinct types. The first type is the *rule-based stemmer*. Rule-based stemmers capture language specific word-formation rules [Porter 1980]. Their development tends to be expensive, usually requiring an expert in linguistics. Various attempts have been made to automate this formalisa-

[5]A example of this type of segmenting tool can be found at http://www.mandarintools.com/

tion process by designing systems capable of automatically learning morphological transformation rules. For example, Snajder et al. [2008] developed a method to automatically acquire inflection rules for the Croatian language, and Moreau et al. [2007] described a similar system which used an analogy-based learning method to automatically detect morphological variants within documents. Majumder et al. [2007] proposed a corpus-based stemming algorithm based on string distance measurements and lexicon clustering.

The second type of stemmer is the *statistical stemmer*. This type of stemmer uses a variety of statistical methods to infer the word formation rules of a particular language [Oard et al. 2000; Melucci and Orio 2003; Bacchin et al. 2005]. Statistical stemmers have been shown to perform well in certain languages, but usually struggle with compound-rich lexicons (e.g., German, Dutch).

There is an alternative to stemming which is known as *lemmatisation* (i.e., the algorithmic process of determining the *lemma*, or canonical form, for a given term). This process involves applying a lexical-based stemmer, or *lemmatiser*, to each term in the source text. The main difference between a lemmatiser and a traditional stemming algorithm is that the former reduces a term to its base lexical form. Although this sounds quite promising in principle, the actual application of lemmatisers has produced mediocre results at best [Hollink et al. 2004]. These results are undoubtedly related to the complexity of the lemmatisation task. Reducing a term to its lemma can be a complicated operation requiring contextual knowledge, part-of-speech (POS) information and knowledge of the target grammar. For this reason, lemmatisers for languages poor in linguistic resources have been hard to acquire (although see [Loponen and Järvelin 2010]).

2.1.4 *Term Expansion.* Expansion occurs when additional terms are added to the source text to improve its quality, expressiveness or discriminatory power. A number of different expansion techniques have been described in the IR & CLIR literature. One common technique employs a machine readable thesaurus to locate expansion terms in lists of synonyms. Other approaches extract expansion terms from large collections of documents. A classic technique in this category is known as *pseudo-relevance feedback* (PRF) [Baeza-Yates and Ribeiro-Neto 2008]. In the (pre-translation) CLIR variant of PRF, the source text is used to search a document collection written in the *same language*. High weighted terms then extracted from the top n documents returned by this search (which are assumed to be relevant) and added to the source text, which is then optionally re-weighted [Rocchio 1971]. Subsequently, this expanded source text can be translated using any of the techniques discussed in §5-6.

2.2 Translation Module

The translation module is the core of the CLIR process chain. As a rule, this module will employ one of two general approaches to translation, namely:

(1) Direct translation

(2) Indirect translation

Furthermore, as mentioned previously, the translation module can operate in one of three modes:

query translation. The query is translated into the language in which the document collection is written.

document translation. At indexing time, the documents are translated into the same language as the query.

dual translation. Both the query and the document collection are translated into a third language (or semantic space) to enable comparison.

The relationship between these two general approaches to translation, and the three modes in which translation can operate, are explored at length in §5-6.

2.3 Post-Translation Module

The purpose of the post-translation module is to shape the output of the translation module into the final product. This process may involve *expanding* the translated text or *re-weighting* individual terms. Post-translation term expansion mirrors the procedure used for pre-translation term expansion, but expands the translated text using terms extracted from top ranked documents retrieved from the target or reference corpus [Singhal and Pereira 1999]. Post-translation query expansion is a very popular technique. Post-translation document expansion is still somewhat of a theoretical backwater, with little to demonstrate in the way of positive results [Levow and Oard 2000].

2.3.1 *Evaluation of Pre- and Post-Translation Expansion.* The relative merits of query expansion are a matter of some dispute. In [Ballesteros and Croft 1997] the authors championed the use of pre- and post-translation query expansion via PRF, reporting a significant increase in CLIR engine effectiveness over unexpanded queries, and better results for dual- versus single expansion. However, later work by Gey and Chen, which summarised experiments reported in TREC-9, revealed a disconcerting lack of consistency in the results obtained in PRF experiments. One paper reported a 42% improvement in effectiveness when using PRF, while a second described a result actually inferior to unexpanded queries [Gey and Chen 1998]. McNamee and Mayfield [2002] supported the finding in [Ballesteros and Croft 1997], although they did question the relative contributions of each expansion stage. In their study, pre-translation expansion led to the largest increases in retrieval effectiveness, although post-translation expansion was still useful because it detected poor translations (see also [Levow et al. 2005]). The confounding factor here seems to be the *size* of the machine readable dictionary (MRD) - the smaller the dictionary, the more pronounced the effect when pre-translation expansion is applied [McNamee and Mayfield 2002; Demner-Fushman and Oard 2003; Xu et al. 2001].

2.4 Information Retrieval Module

The information retrieval module performs the actual search, matching query representations against document representations and ranking the results. For the sake of completeness, the remainder of this section describes some of the dominant IR models in use today. This section is merely intended as a summary - readers requiring greater depth are referred to [Baeza-Yates and Ribeiro-Neto 2008; Manning et al. 2008] for further information.

2.4.1 *Boolean Model.* The oldest, and simplest, retrieval model is known as the *Boolean model* [Lancaster and Fayen 1973; Salton et al. 1983]. This model is based on set theory and Boolean algebra, and represents queries as Boolean expressions with precise semantics. Suppose a document has three terms, such that $d = \{s_1, s_2, s_3\}$. In the Boolean model this document could be represented as a conjunction of terms $d = s_1 \wedge s_2 \wedge s_3$. If a query contained two of these terms so that $q = (s_1 \wedge s_3) \vee s_2$, the similarity between d and q can be calculated using the logical implication $sim(d, q) = d \rightarrow q$.

This approach has two main drawbacks. Firstly, unlike the models described below, it is based on a *binary decision criterion* (i.e., a document is predicted to be either relevant or non-relevant). Secondly, it is not always easy to translate a specific information need into a Boolean expression.

2.4.2 *Vector Space Model.* The next major retrieval model is known as the *vector space model* [Salton 1971; Salton et al. 1975]. Unlike the binary output of a Boolean model, the vector space model ranks documents in decreasing order of a measure that corresponds to the *relevance* of each document to the query.[6] This is accomplished by assigning non-binary weights to the index terms found in the query and the documents. These weights are usually calculated using the popular *tf-idf weighting scheme* [Sparck Jones 1988]. Thereafter, the similarity of all documents in the collection to the query is computed using a distance measure. More formally, a document d and a query q are represented as vectors within a *high dimension information space*. The similarity of the document to the query $sim(q, d)$ is measured using the cosine of the angle between these vectors, as follows:

$$sim(q, d) = cos(\overrightarrow{q}, \overrightarrow{d}) = \frac{\overrightarrow{q} \bullet \overrightarrow{d}}{|\overrightarrow{q}| \times |\overrightarrow{d}|}$$

where $|\overrightarrow{d}|$ is the length of the vector representing d and $\overrightarrow{q} \bullet \overrightarrow{d}$ is the dot product. The $tf - idf$ weight of a term t in a document d is calculated as follows:

$$tf - idf_{s,d} = tf_{s,d} \times \log \frac{N}{df_s}$$

where $tf_{s,d}$ is the total number of occurrences of s in d (*term frequency*), df_s is the number of documents in the collection which contain s (*document frequency*) and N is the number of documents in the collection. Advantages of the vector space model tend to focus on *usability*. It is easier to represent naturalistic queries using a vector space model, many users prefer ranked retrieval over Boolean retrieval and those users tend to employ ranked retrieval more effectively than Boolean retrieval.

2.4.3 *Probabilistic Model.* Next we have the *probabilistic model*, which attempts to solve the problem of selective retrieval within a probabilistic framework. This means that the similarity between documents and queries is computed through a

[6] A document is considered to be relevant if it is a document that the user perceives as containing information of value with respect to their information need. The process of determining relevance for evaluative purposes is described in greater detail in §3.

probabilistic description of the ideal answer set. If we take the simplest binary independence retrieval model, the similarity between a document d and a query q is computed as:

$$sim(q,d) = \log \frac{P(rel|d,q)}{P(irrel|d,q)}$$

where *rel* and *irrel* denote relevance and irrelevance respectively. Where no relevant or irrelevant documents are available in advance, several methods can be used to estimate these probabilities [Baeza-Yates and Ribeiro-Neto 2008]. Non-parametric models are also studied by researchers. For example, [Amati and Van Rijsbergen 2002] introduced a probabilistic model which was based on the *observed divergence from randomness*. This model was derived in a purely theoretical way, by combining several different probability distributions.

2.4.4 *Language Model.* Next, we have the *language model*. The language model was introduced by Ponte and Croft in 1998 [Ponte and Croft 1998] and has been extremely popular with IR researchers ever since [Miller et al. 1999; Song and Croft 1999; Berger and Lafferty 1999; Jin et al. 2002; Gao et al. 2004; Lavrenko and Croft 2001]. The basic idea here is to estimate the probability of a query given a document language model (see [Liu and Croft 2005; Zhai 2009] for surveys of language models and their uses). Formally, a basic statistical language model scores the relevance of a document d to a query q as

$$sim(q,d) = P(s_1, ..., s_n \in q|d) \approx \prod_{i=1}^{n} P(s_i|M_d)$$

with an approximation step that assumes term independence and use of a unigram language model for documents (M_d). To reduce the products in this equation to summations, cross-entropy estimations are calculated as follows:

$$P(q|d) = \sum_{s_i \in V} P(s_i|M_q) \log P(s_i|M_d)$$

where M_q is the unigram language model for queries and, since d and q are in the same language, s_i occurs in the shared vocabulary V. To solve the unseen terms problem, various smoothing techniques have been proposed [Zhai and Lafferty 2001]. Studied in the monolingual retrieval environment, the language model has demonstrated comparable effectiveness to the traditional vector space and probabilistic models.

2.4.5 *Other Models.* In this section we briefly describe a number of alternative retrieval models that have been adopted by CLIR researchers. We begin with an extension of the conventional vector space model. One of the weaknesses of this model lies in its use of terms as the orthogonal basis of the vector space. This assumption is problematic because terms are rarely semantically independent. In [Wong et al. 1985] the authors proposed a 'Generalized Vector Space Model' (also known as 'the dual space' model [Sheridan and Ballerini 1996]) which addressed this weakness. In the GVSM the index term vectors are assumed to be linearly

independent but not pairwise orthogonal. Given a term-document matrix A (with rows representing terms and columns representing documents) a query could be transformed to $A^{'}\overrightarrow{q}$ and a document could be transformed to $A^{'}\overrightarrow{d}$, where $A^{'}$ is transpose of the matrix A. The retrieval criterion is then defined to be:

$$sim(q,d) = cos(A^{'}\overrightarrow{q}, A^{'}\overrightarrow{d})$$

It is fairly easy to extend this model to encompass cross-language information retrieval. Assuming we have a bilingual parallel corpus, we form two matrices A and B so that A is a term-document matrix in the query language, B is a matrix in the document language, and the columns in A and B contain matching pairs of documents found in the training corpus [Sheridan and Ballerini 1996; Carbonell et al. 1997].

Another interesting approach is essentially a fusion model that combines sets of results generated using different indexing and retrieval models [Savoy 2004; 2005]. Since each model tends to identify dissimilar groups of pertinent and non-pertinent items, merging the results generally leads to an increase in retrieval effectiveness when compared with a single model used in isolation. Savoy et al. have experimented with various fusion operators, including simple summation, the round-robin approach, logistic regression and the so-called 'Z score' [Savoy 2004; 2005](see also [Shaw and Fox 1994]). The overall effectiveness of this model has made it very popular in both monolingual and cross-language information retrieval.

3. EVALUATION

In this section we describe some common approaches which can be used when evaluating the quality of a translation system in the context of cross-language information. We discuss evaluation at this specific point in the paper to provide a common vocabulary for the comparison of various techniques and translation models in the following sections.

Evaluating the effectiveness of a query and/or document translation system usually involves assessing the *retrieval effectiveness* of the CLIR engine associated with it. The standard mechanisms for this sort of assessment all rely upon the availability of large Cranfield-style test collections [Cleverdon 1991]. A Cranfield-style test collection normally consists of a document corpus, a set of search topics, and a matched set of assessments. The document corpus is often provided as a set of semi-structured XML documents. Each document in this set will consist of several text fields (e.g., title, abstract, keywords), and a unique document identification number. The search topics will describe a number of search tasks. These tasks are often categorised as 'short' query tasks and 'long' query tasks. The long query tasks are more detailed, but short query tasks tend to replicate realistic Web queries.

Relevance assessments are manually derived assessments representing the relevance of each document in the corpus to each topic. Producing these assessments is a time-consuming and expensive process. Because test collections are generally large, only a fraction of the documents relevant to each query are scored. The standard approach to selecting this subset is known as as 'pooling' [Kuriyama et al. 2002]. A pooling operation selects only the top k documents returned by a number of different CLIR systems for manual assessment. The CLIR engines used during a

pooling run are usually the same CLIR engines that need to be evaluated. Monolingual runs and interactive runs (either cross-language or monolingual) can also be used to enrich the pools.

The production of test collections has always been one the main responsibilities of CLIR conferences and workshops. The Text Retrieval Conference (TREC),[7] sponsored by the National Institute of Standards and Technology (NIST), was started in 1992 as part of the TIPSTER Text program. Its purpose was to support research within the IR community by providing the infrastructure necessary for large-scale evaluation of text retrieval methodologies. TREC had a very important influence on CLIR, especially in its formative years, and hosted one of the earliest competitive CLIR tracks in 1997 (TREC-6). DARPA was another early proponent of CLIR, launching the TIDES (Translingual Information Detection, Extraction, and Summarisation) program.

Other CLIR conferences of note include the Cross-Language Evaluation Forum (CLEF[8]), which concentrates on European languages, and the Forum for Information Retrieval Evaluation (FIRE[9]), a CLEF-2007 spin-off chiefly concerned with Indian languages (Hindi, Telugu, and Malayalam). Along a similar line, NTCIR[10] was founded in 1999. It has been responsible for a series of evaluative workshops designed to enhance CLIR research in Pacific Rim languages (e.g., Chinese, Japanese, Korean).

Researchers with access to a test collection can retrieve documents using the topics provided and then measure the retrieval effectiveness of their system using the relevance judgements. The three basic measures that are applied at this stage are precision, recall, and the F-measure (also known as the F1 score). Precision is the fraction of retrieved documents that are relevant. Recall is the fraction of relevant documents that are retrieved. The F-Measure is the harmonic mean of precision and recall. These are *set-based measures* which are computed using unordered sets of documents. When working with *ranked retrieval results* a number of other measurements may come into play [Manning et al. 2008]. In these circumstances, precision can be measured using a relatively low number of retrieved results (e.g., precision 10 results). Mean average precision (MAP) can provide a single figure measure of quality across recall levels. Normalized discounted cumulative gain (NDCG) can be deployed in situations involving non-binary notions of relevance. There are many other measurements which can be applied [Baeza-Yates and Ribeiro-Neto 2008]. Once measured, the retrieval effectiveness of a CLIR system is often compared with a *monolingual baseline*. Tests of statistical significance, such as the widely used Wilcoxon signed-rank test, are commonly used when interpreting results [Hull 1993].

4. TRANSLATION TECHNIQUES IN CLIR

The following sections (§5-6) outline the state of the art in free text translation in the context of CLIR. They also enumerate a taxonomy that codifies the work done

[7] http://trec.nist.gov/
[8] http://www.clef-campaign.org/
[9] http://www.isical.ac.in/∼clia/
[10] http://research.nii.ac.jp/ntcir/

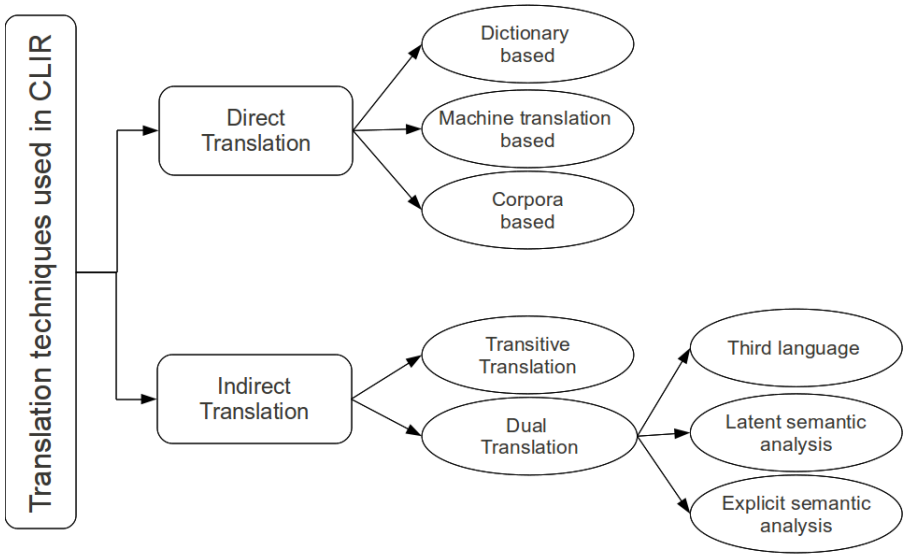

Fig. 6: Translation techniques in CLIR - A Taxonomy

so far. A graphical overview of this taxonomy is shown in Figure 6. As illustrated, translation in CLIR is divided into two basic categories. There are techniques that effect *direct* translation. These will use translation resources like bilingual dictionaries, machine translation systems and parallel corpora. Then there are techniques that employ *indirect* translation. These approaches exploit an intermediate language to translate the source text, or match query and document representations via dual translation.

Although the differences between query translation and document translation in terms of retrieval effectiveness are arguable [Oard and Hackett 1997; Oard 1998; McCarley 1999] there is an obvious gulf between the two techniques in terms of computational requirements. For this reason, the majority of CLIR systems in existence today practice query translation. Therefore, in §5-6, our examination of translation techniques in CLIR will tend to focus on query translation systems (with the obvious exception of §6.2, which addresses dual translation systems). We reserve our discussion of document translation proper until §7.

5. DIRECT TRANSLATION

Direct translation systems exploit bilingual dictionaries, parallel corpora and machine translation algorithms to translate the source text. We discuss each of these translation resources below.

5.1 Dictionary-Based Translation

Machine-readable bilingual dictionaries have become increasingly available and are often used in the translation modules of CLIR engines. A dictionary-based approach to translation is relatively simple (when compared to the alternatives), but suffers

from two major weaknesses:

—Ambiguity
—Lack of coverage

These weaknesses are described in detail below.

5.1.1 *Ambiguity.* The first major problem to affect systems employing dictionary-based translation is *ambiguity*. Bilingual dictionaries will usually contain multiple translations for any given query term. Selecting the 'correct' translation from a list of competing candidate terms is a crucial but decidedly non-trivial task [Ballesteros and Croft 1998; Aljlayl and Frieder 2001]. Solutions to this selection task have specialised towards two mutually exclusive techniques. Some techniques try to find the single best translation for each term in the query (*single selection translators*). Other techniques are open to the possibility of more than one translation of each query term, addressing the issue of ambiguity via the re-weighting of terms (*multiple selection translators*).

—*Single Selection Translators*

Early single selection systems addressed the problem of ambiguity in a primitive way by simply selecting the first translation offered by the dictionary. This strategy exploits the fact that in some bilingual dictionaries the most commonly used translation is listed first [Ballesteros and Croft 1998; Aljlayl and Frieder 2001]. This basic disambiguation strategy has obvious shortcomings and was soon replaced by more sophisticated techniques exploiting *term co-occurrence* statistics [Adriani 2000; Ballesteros and Croft 1998; Gao et al. 2002; Jang et al. 1999; Maeda et al. 2000; Liu et al. 2005; Gao and Nie 2006]. The hypothesis grounding the use of term co-occurrence data in this context states that the correct translations of individual query terms will tend to co-occur as part of a sub-language while incorrect translations will not. In other words, this approach should be able to determine the most likely translation for a given query by examining the pattern of term co-occurrence within some representative text collection (e.g., the World Wide Web [Maeda et al. 2000] or a monolingual corpora [Ballesteros and Croft 1998; Gao and Nie 2006]).

However, there is a problem with this general approach. The terms in the query are *mutually dependent*. Therefore, for each of these terms, we need to select the translation most consistent with the translations of the remaining terms. This sort of recursive, global optimisation is computationally prohibitive for even shortest of queries [Gao and Nie 2006]. A common workaround, used by several researchers working on this particular problem [Adriani 2000; Ballesteros and Croft 1998; Liu et al. 2005; Gao and Nie 2006], substitutes the following greedy algorithm:

(1) Given a source query q, for each word in q, acquire the set of all translation alternatives $T_i = \{t_{i,1}, t_{i,2}, ..., t_{i,m}\}$ from the translation resources
(2) For each set T_i do
    (a) For each translation $t_{i,m} \in T_i$, define the similarity measurement between the translation word $t_{i,m}$ and the other set $T_j (T_j \neq T_i)$ as the sum of the similarities between $t_{i,m}$ and each word in the set T_j as

$$sim(t_{i,m}, T_i^{'}) = \sum_{\forall t_{j,n} \in T_j} sim(t_{i,m}, t_{j,n})$$

(b) Compute the cohesion score for $t_{i,m}$ as

$$co(t_{i,m}) = \sum_{\forall i \neq j} sim(t_{i,m}, T_j)$$

(c) Select the term t_i in T_i with the highest cohesion score

$$t_i = argmax_{t_{i,m}} co(t_{i,m})$$

In the equation shown above, cohesion is calculated using an undefined similarity measurement. This measurement can be implemented using a wide variety of interchangeable algorithms. Commonly used algorithms exploit mutual information (MI), Dice's coefficient, the log likelihood ratio and the Chi-square test [Maeda et al. 2000; Gao et al. 2001; Gao et al. 2002; Liu et al. 2005; Adriani and Wahyu 2005]. Less common algorithms have examined the effectiveness of the EMIM (Expected Mutual Information Measure) weighting measure [Adriani 2000; Ballesteros and Croft 1998; Gao et al. 2001], Hidden Markov Models [Federico and Bertoldi 2002] and expectation-maximisation (EM) algorithms [Monz and Dorr 2005]. Some researchers have attempted to exploit multilingual relations to improve the quality of query translation. Zhou et al. [2008] modelled translation candidates and their relationships as a directed graph in which the similarity between terms acts collectively to 'elect' the best translation candidates (see also [Zhou et al. 2008]). This type of graph-based disambiguation can also exploit monolingual relations [Cao et al. 2007]. Finally, there have been various attempts to capture term dependencies through consideration of word order in the combinations of translated query terms (e.g., [Jang et al. 1999]).

It is critically important that single translator systems select the *correct translation unit*. The commonest units of translation are (from smallest to largest) - the n-gram,[11] the word stem, the word, the phrase and the sentence. Generally speaking, the accuracy of a translation will improve as the size of the translation unit increases, but the coverage of a typical bilingual dictionary will drop (see §5.1.2). The use of sentence length translation units is extremely rare, due to the average size of queries, but the phrasal unit has often proved critical [Hull and Grefenstette 1996; Ballesteros and Croft 1997; 1998; Meng et al. 2000; Meng et al. 2001]. The procedure for translating phrases typically uses a dual-pass procedure. In the first pass, phrases in the query are identified and translated if their translations are found in the bilingual dictionary. In the second pass, the remaining terms in the query are translated word by word. The efficacy of the first pass is obviously dependent on the phrasal content of the dictionary.

There have been a number of attempts to identify the most accurate unit of translation given the limited size of query strings [Gao et al. 2001; Gao et al. 2002]. In one experiment, the authors used a noun phrase (*np*) translation model based on template patterns alongside a probabilistic framework to capture the dependency of adjacent/non-adjacent query terms [Gao and Nie 2006]. In this translation model noun phrases were identified in the source queries using statistical methods, then

[11] An n-gram is a sub-string of n characters. A skip-gram is a digram (or longer sequence of letters) formed from non-adjacent characters. This concept was developed to address the problem of cross-lingual spelling variation [Pirkola et al. 2002] (see also [Pirkola et al. 2003])

translated using probabilities obtained from the target language model. Given a np in source language e and a set of candidates of np in target language c, the best translation in target language np_c^* can be obtained by:[12]

$$np_c^* = argmax_{np_c} P(np_c|np_e) = argmax_{np_c} P(np_c)P(np_e|np_c)$$

where $P(np_c)$ is the prior probability of the target language (estimated using a trigram language model) and $P(np_e|np_c)$ is the translation probability. Estimation required both monolingual and bilingual corpora, as did calculation of the dependency model. The np translation template z was introduced as a hidden variable so that

$$np_c^* = argmax_{np_c} P(np_c) \sum_z P(z|np_c)P(np_e|z, np_c)$$

This enabled the automatic extraction of translation templates. The authors concluded that larger and more specific units of translations will always be superior provided the underlying translation models are well trained.

—*Multiple Selection Translators*

Multiple selection translators do not try to find the single best translation for each of the various query terms. Instead, they are open to the possibility of more than one translation of each query term. *Structured query translation* is perhaps the commonest application of this general approach. The concept of structured query translation was introduced by Hull and developed to fruition by Pirkola [Pirkola 1998; Hull 1997]. Pirkola implemented a system that used the INQUERY [Broglio et al. 1993] synonym operator to select multiple translation candidates for a query term. This operator was originally designed to support monolingual thesaurus expansion. Pirkola used it in the following way:

$$TF(q, d) = \sum_{i=1}^{n} TF(t_i, d)$$

$$DF(q) = |\bigcup_{i=1}^{n} \{d|t_i \in d\}|$$

where q is a query term, t_i is a set of translation alternatives, d is a document, TF is the sum of the translation synonyms of the source term in that document, and DF is the total number of documents that contain at least one term from t_i. This calculation proved to be computationally expensive (see also [Pirkola et al. 2003]). Later optimisations [Kwok 2000; Darwish and Oard 2003] include replacing the union operator with a sum

[12]This formula is based on Shannon's noisy-channel coding theorem [Shannon and Weaver 1963]. Please refer to Appendix A for a detailed description of this theorem as applied to statistical machine translation.

$$DF(q) = \sum_{i=1}^{n} DF(t_i)$$

and defining the maximum document frequency of any replacement term as

$$DF(q) = MAX_{i=1}^{n}[DF(t_i)]$$

An extension to the Pirkola method can be found in [Darwish and Oard 2003]. In this paper the authors describe a probabilistic approach that integrates the translation likelihood of computing the TF and DF of a source query term. Using techniques drawn from statistical machine translation,[13] translation probabilities were used in the following way:

$$TF(q, d) = \sum_{i=1}^{n} TF(t_i, d) \times P(t_i|q)$$

$$DF(q) = \sum_{i=1}^{n} DF(t_i) \times P(t_i|q)$$

where $P(t_i|q)$ refers to the probability of a query term q translating into document term t_i. In this equation, the values derived for TF or DF can be weighted using the best available replacement probability (through all translation alternatives). The authors experimented by weighting TF in isolation, DF in isolation, and a combination of the two. A combined weighting scheme led to the best retrieval effectiveness. Overall the results suggested that a probabilistic approach to structured query translation model will significantly outperform the basic structured method for query terms, assuming a large number of translation alternatives.

An alternative to the Pirkola method is known as the *balanced* approach [Levow and Oard 2000; Leek et al. 2000]. This technique involves balancing the weight of multiple translation alternatives to avoid placing high weight on terms with many translations (as these tend to be common terms). There are two basic implementations of the balanced approach, one based on replication and the other based on re-weighting. Balanced and structured query translation were compared in [Oard and Wang 2001; Meng et al. 2000]. The authors concluded that although the balanced approach was more sensitive to rare translations, structured query translation was superior provided the researchers had operational control of the retrieval module. This was necessary because Pirkola's method required that aggregation be applied individually to the TF and DF components of the calculation, not to the term weights. The balanced translation approach should only be selected when the IR engine is used essentially as a 'black box'.

Multiple selector translators can use *bidirectional translation* to improve the quality of their output. Bidirectional translation is a technique for re-weighting translation probabilities that merges a ranked list of documents retrieved using a translated query with a ranked list of translated documents retrieved using the original

[13] See Appendix A

source query. McCarley [1999] was the first to pioneer this technique. He reported a statistically significant improvement in mean average precision with bidirectional translation when compared with unidirectional translation (in either direction). This finding was subsequently confirmed by a number of other researchers who extended his work [Kang et al. 2004; Boughanem et al. 2002; Wang and Oard 2006].

Wang and Oard [2006] proposed an approach that extended bidirectional translation using synonymy relations. The analytical framework they employed was extremely similar to the probabilistic structured translation method described above. For each word q in query language S, assume that a set of terms t_i in document language T is known and shares the searchers intended meaning for term s with some probability $P(q \longleftrightarrow t_i)$. Now calculate the TF and DF as follows:

$$TF(q,d) = \sum_{t_i} P(q \longleftrightarrow t_i) \times TF(t_i,d)$$

$$DF(q) = \sum_{t_i} P(q \longleftrightarrow t_i) \times DF(t_i)$$

where $P(q \longleftrightarrow t_i)$ is calculated using:

$$P(q \longleftrightarrow t_i) = \sum_{mean_j} P(mean_j|q) \times P(mean_j|t_i)$$

and where $P(mean_j|q)$ denotes the probability that term q has meaning $mean_j$ and $P(mean_j|t_i)$ denotes the probability that term t_i has meaning $mean_j$. Hoping to find a computational model in which meaning representations were aligned across languages, the authors selected their synonymous terms from various automatically generated resources. Their experiment produced a statistically significant improvement in mean average precision over the state of the art.

For single selection translators, cross-language information retrieval is essentially a two-step process - external translation, followed by monolingual retrieval. In the current section we see some examples of a slightly different process which involves *embedding* the translation module within the information retrieval system. This merger exposes the internal state of the IR engine to the translation module and moves us closer towards a more unified CLIR framework. To understand this 'embedding' process we need to revisit to the statistical language model. As discussed in §2.4.4, cross-entropy estimation in the statistical language model is calculated as follows:

$$P(q|d) = \sum_{s_i \in V} P(s_i|M_q) \log P(s_i|M_d)$$

In the next step, as described by [Berger and Lafferty 1999], the translation probabilities for terms in the same language are modelled as relationships:

$$P(s_i|M_d) = \sum_{s_i \in V_e} P(s_i|s_j) P_{ML}(s_j|M_d)$$

Obviously, this process of inter-term estimation can be extended to CLIR by training a translation model $P(t_i|s_j)$ that captures the relationships between terms in two different languages [Kraaij et al. 2003; Nie 2010] so that:

$$P(t_i|M_{q_e}) = \sum_{s_j \in V_e} P(t_i|s_j) P_{ML}(s_j|M_{q_e})$$

where M_{q_e} is the language model for the source query and t_i is a term in the target language. The ranking of documents is then computed using the following formula:

$$P(q_e|d_c) = \sum_{t_i \in V_c} \sum_{s_j \in V_e} P(t_i|s_j) P_{ML}(s_j|M_{q_e}) \log P(t_i|M_{d_c})$$

This model was trained on parallel documents automatically mined from the web. It out-performed a MT system trained on similar resources and achieved around 90% of monolingual IR effectiveness (w.r.t English and French) and 80% of monolingual effectiveness (w.r.t. English and Italian).[Xu et al. 2001; Xu and Weischedel 2005] achieved similar results in related experiments.

Other researchers have concentrated on the development of new *estimation methods*. In [Lavrenko et al. 2002] the authors described a technique for estimating an accurate topic model in the target language, starting with a query in the source language. Given a query $q = s_1...s_k$, $R(q)$ is the set of target documents that are relevant to that query. Effective ranking of these documents could be achieved if there was a way of estimating the relevance model of q, in another words, the set of probabilities $P(t_j|R_q)$ to the word occurrence of every word t_j in the target vocabulary. $P(t_j|R_q)$ denotes the probability that a word sampled at random from a relevant document would be the word t_j. If we knew beforehand which documents were relevant, estimation of these probabilities would be straightforward. However, we do not normally know the membership of this set. A reasonable way to approximate this probability is to use the joint probability of observing the word t_j together with query words $s_1...s_k$:

$$P(t_j|R_q) \approx P(t_j|q) = \frac{P(t_j, s_1...s_k)}{P(s_1...s_k)}$$

Now we need to adapt the original relevance model, which was designed for a monolingual setting. *Cross-language estimation* works as follows: suppose t_j is a word in source language, and $s_1...s_k$ are words in another language. There are two possible strategies at this point. The first strategy uses a parallel corpus (i.e., a set of document pairs d_e, d_c where d_e is a document in source language and d_c is a document in the target language). Assume Ψ is the set of corresponding distribution pairs Ψ_{d_e}, Ψ_{d_c}, we can estimate the joint probability with:

$$P(t_j, s_1...s_k) = \sum_{\{\Psi_{d_e}, \Psi_{d_c}\} \in \Psi} P(\{\Psi_{d_e}, \Psi_{d_c}\}) P(t_j|\Psi_{d_c} \prod_{i=1}^{k} P(s_i|\Psi_{d_e}))$$

using the parameters described above. The second strategy uses a *statistical lexicon* to estimate the translation probability $P(s_i|t_j)$ for every word in the source and

target language with $P(s_i|\Psi_{d_c})$ computed through:

$$P(s_i|\Psi_{d_c}) = (1-\lambda)P(s_i) + \lambda \sum_v P(s_i|t_j)P_{ML}(t_j|\Psi_{d_c})$$

where the summation iterates through all the words t_j in the vocabulary of the target language, $P(s_i|t_j)$ is the translation probability derived from the statistical lexicon, $P_{ML}(t_j|\Psi_{d_c})$ is the number of times t_j occurs in d_c (divided by the length of d_c) and $P(s_i)$ is the background probability of s_i computed over a large corpus. Lavrenko et al. [2002] reported an experiment using the TREC-9 cross-language collection [Voorhees and Harman 2000] in which lexicon-derived translation probabilities produced an impressive 93%-98% of the monolingual baseline.

5.1.2 *Coverage.* The second major problem to affect systems employing dictionary-based translation is related to vocabulary coverage. Certain types of words (e.g., newly coined terms, technical terms, compound words, proper names, acronyms, abbreviations) are under-represented in machine readable bilingual dictionaries. These *out-of-vocabulary* (OOV) terms can severely degrade the retrieval effectiveness of a CLIR engine, especially when the source queries being translated are very short. Early solutions to the coverage problem advocated the use of domain-specific bilingual dictionaries. These dictionaries delivered access to uncommon vocabularies and technical terms [Pirkola 1998], but they were extremely costly to produce (although see [Korn et al. 2005]). Stemming the source language component of a bilingual dictionary partially addressed this issue, but the problem of incomplete coverage persisted [Resnik et al. 2001]. For this reason, research quickly moved away from domain-specific resources and towards *transliteration*.

Transliteration involves identifying similarities in the orthographic structures of two languages. These similarities are subsequently used to generate rules specifying how sub-strings written in one language are spelled in another [Buckley et al. 2000]. Transliteration is a *directional* process. Forward transliteration involves converting a word in the source language into an approximate equivalent in the target language [Kang and Choi 2000]. Backward transliteration is the process that converts the transliterated word back to its original form [Goto et al. 2004; Jeong et al. 1999]. Both techniques have been successfully adopted by CLIR researchers, with forward transliteration generally preferred as it is easier to apply.

In this particular context, the process of transliteration has two stages. The first stage is essentially a matching exercise during which strings from the source and target languages are paired together using a variety of methods [Keskustalo et al. 2003; Pirkola et al. 2002; Pirkola et al. 2003]. These methods range in complexity from the relatively simple (e.g., treating the source terms as potentially mis-spelled terms belonging to the target language [Buckley et al. 2000]) to complex techniques utilising n-gram analysis,subsequence matching, string-string distance measurements and statistical models [Robertson and Willett 1998; Zobel and Dart 1995; Melamed 2000; AbdulJaleel and Larkey 2003; McNamee and Mayfield 2004b]). The second stage of transliteration involves exploiting these orthographic mappings to generate transliterations for all OOV terms.

The type of transliteration described above only really works when the languages share a similar character set (e.g., English and French). Transliteration between

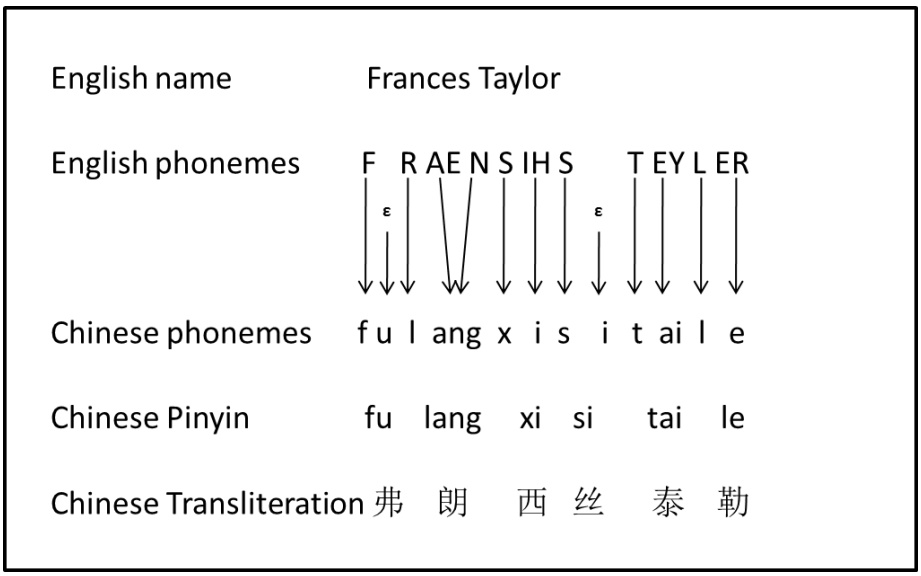

Fig. 7: English to Chinese transliteration

languages with dissimilar character sets (e.g., Chinese and Russian) requires a process known as *phonetic mapping*. Phonetic mappers generate rules representing the phonetic presentation of a language. During the mapping process, all proper nouns (usually names) are transformed into a corresponding *phonetic sequence*. This phonetic sequence is matched with a phonetic sequence in the target language, then transformed into a final translation [Fujii and Ishikawa 2001; Gao et al. 2005; Kang and Kim 2000; Knight and Graehl 1998; Qu et al. 2003; Virga and Khudanpur 2003]. Figure 7 provides an example of this type of transliteration, extracted from [Virga and Khudanpur 2003]. Qu et al. [2003] described a transliteration method which used a phonetic dictionary and a set of probabilistic rules to create English-Japanese mappings. In the same year [Virga and Khudanpur 2003] reported a generative statistical model that uses statistical machine translation techniques to "translate" the phonetic representation of an English name to Chinese.

Recent work on the problem of OOV terms has exploited the World Wide Web (WWW). Lu et al. [2002] were early proponents of this approach (see also [Lu et al. 2004]). They described a technique that mined multilingual HTML anchor text to translate newly coined terms and proper names. Their approach used a probabilistic model to determine the best translation for an OOV term using co-occurrence data. Given a set of URLs $U = u_1, u_2, \ldots, u_n$ which link to the same Web page, the degree of similarity between source term s and target translation t is estimated using:

$$P(s \longleftrightarrow t) = \frac{P(s \bigcap t)}{P(s \bigcup t)} = \frac{\sum_{i=1}^{n} P(s \bigcap t|u_i)P(u_i)}{\sum_{i=1}^{n} P(s \bigcup t|u_i)P(u_i)}$$
$$\approx \frac{\sum_{i=1}^{n} P(s|u_i)P(t|u_i)P(u_i)}{\sum_{i=1}^{n}[P(s|u_i) + P(t|u_i) - P(s|u_i)P(t|u_i)]P(u_i)}$$

The approximation step described above assumes that s and t are independent with respect to u_i.

Recent extensions of this Web-based approach to query translation have tended to focus on the *gratuitous translations* provided by Web page authors for the convenience of their readers (e.g., proper names or technical terms followed by a translation in parenthesis). These incidental translations can be identified and extracted using fairly simple pattern-matching techniques [Cheng et al. 2004; Zhang and Vines 2004; Zhang et al. 2005; Huang et al. 2005; Cao et al. 2007]. In most cases these patterns are manually constructed, although automatic pattern generation is feasible [Shi 2010; Zhou et al. 2007; 2008]

Wikipedia, the online encyclopaedia, can also be used for OOV translation. Each Wikipedia article has hyper-links joining it to versions of the same article written in other languages. A number of researchers have exploited these cross-lingual associations to translate unknown terms, assuming that article names linked in this fashion are mutual translations [Jones et al. 2008].[14] Results, thus far, are limited, but the potential for future development seems clear [Su et al. 2007; Lin et al. 2010; Schönhofen et al. 2008].

Finally, it should be noted that untranslated OOV terms are not always deleterious. Several studies have shown that the presence of an untranslated OOV term can sometimes improve retrieval effectiveness rather than impair it [Kishida 2008]. Lee et al. [2010] studied this phenomenon using a machine learning methodology. The authors discovered that it was better to leave an OOV term untranslated (and hope that pre- or post-translation query expansion 'fixed' the problem) than mine an inadequate translation.

5.2 Machine Translation

Machine translation (MT) is the automatic translation of free-text from one natural language to another. MT systems have become extremely popular in CLIR over the last 5-10 years. In part, this popularity is due to the wide availability of MT systems and the linguistic resources required to train them. However, it can also be attributed to the excellent results obtained in experiments. For example, in CLEF-2009, CLIR systems using machine translation achieved the equivalent of 90%-99% of the monolingual baseline on English, French and German collections [Ferro and Peters 2009].

The recent dominance of off-the-shelf MT systems, especially in time-constrained CLIR campaigns (in which a series of tracks designed to test different aspects of mono- and cross-language information retrieval engines are provided), cannot be overstated [Agirre et al. 2009; Ferro and Peters 2009; Sakai et al. 2008; Sakai et al. 2010]. In CLEF 2009, 7 out of the 10 participants used some sort of statistical MT

[14]This paper forms part of the *MultiMatch* project, an attempt to develop a prototype multilingual/multimedia search engine which enables 'users to explore and interact with online accessible cultural heritage content, across media types and languages boundaries' (http://www.multimatch.eu/). Readers interested in the broader context of CLIR are encouraged to review the outcomes of this project. The CACAO initiative, which aims to deliver cross-language access to catalogues and on-line libraries, is another useful resource for CLIR researchers (http://www.cacaoproject.eu/)

system,[15] with the *Google Translate API* heavily represented.[16] This API outperformed the rest of the field by a very clear margin. The best performing non-Google MT system managed just 70% of the MAP achieved by Google-based systems [Leveling et al. 2009], and results for the Google participants rivalled those recorded by markedly more sophisticated systems [Anderka et al. 2009]. This prompted the organisers of CLEF 2009 to ask the (rather mischievous) question - '*Can we take this as meaning that Google is going to solve the cross-language translation resource quandary?*'.

Despite these recent successes, most researchers would agree that off-the-shelf MT systems are still some distance from solving the 'CLIR problem'. This is due to a number of specific problems that impair their suitability for query translation, as follows:

(1) The effectiveness of off-the-shelf statistical MT systems is heavily dependent on the languages involved. For resource-poor languages (e.g., Thai) or language pairs with little in common (e.g., English and Chinese), CLIR effectiveness can be as low as 50% of the monolingual baseline [Adriani and Wahyu 2005; Kwok 1999; Zhou et al. 2008]. This unfortunate fact was confirmed in a recent study using two separate machine translation systems [Dolamic and Savoy 2010]. Furthermore, certain pairings of closely related, resource-rich languages may also produce poor retrieval effectiveness when MT is applied. This was revealed in a series of CLIR experiments that compared the results for an MT system working with various European language pairs [Savoy and Dolamic 2009]. The authors found that searching a collection of French documents using a Google translation of a German query produced significantly worse results than a translated English query on the same collection.

(2) The output of an off-the-shelf MT system is usually one word per query term. This sort of literal mapping ignores the availability of multiple expressions in the target language, leaving some translations incomplete. Crucial factors here could be the quality of the statistical translation lexicon used by the MT system, or the quality of the MT system itself ([Zhu and Wang 2006; Parton et al. 2008; Xu and Weischedel 2000].

(3) MT systems generally pay too much attention to syntactic structure, which is largely unimportant when translating queries. Conversely, they usually ignore OOV terms, which often have a significant impact on retrieval effectiveness.

5.3 Corpus-based Translation

A *parallel text* is a document written in one language together with its translation in another. Large collections of parallel texts are referred to as *parallel corpora*. Parallel corpora can be acquired from a variety of sources. International organisations such as the United Nations and the European parliament[17] publish a huge volume of parallel documentation every year in a wide variety of languages. At the national level, the same approach has been adopted by the Canadian Parliamen-

[15] Please see Appendix A.
[16] http://translate.google.com/
[17] http://www.europarl.europa.eu/

t[18] (French and English) and the Hong Kong Legislative Council[19] (Chinese and English). The Bible also makes a significant contribution, providing an important resource for low-density languages [Chew et al. 2006]. The World Wide Web is another obvious source for parallel corpora [Chen et al. 2004; Chen and Nie 2000; McEwan et al. 2002; Yang and Li 2002]. Experiments conducted by [Resnik 1998; 1999; Resnik and Smith 2003; Nie et al. 1999] used simple but effective heuristics to mine parallel Web pages from a Web crawl (e.g., structural clues, anchor text)

Parallel corpora are commonly used in cross-language information retrieval to translate queries. The basic technique involves a side-by-side analysis of the corpus producing a set of translation probabilities for each term in a given query. These translation probabilities are usually generated using the model described in Appendix A. One of the important characteristics of this model is that it makes no assumptions whatsoever about word order during the training process - a source sentence could be translated into a sentence of any length, and one position in the target sentence can be aligned to any position in the source sentence. The probability of aligning a word at a specific position in the source sentence is dependent on presence of the corresponding word in the target sentence. This simplified model has proved to be very effective in the CLIR environment, in which the word order of queries is relatively unimportant, and the accidental selection of loose translation relations (i.e., high related terms in vicinity of the actual translation) is sometimes desirable.

The corpus-based approach to query translation can be supplemented by other resources. For example, [Nie 1998; 2010] tried to select the top n translation words t_i with the highest translation probabilities conditioned by the query q, so that:

$$P(t_i|q) = \sum_{s_i \in q} P(t_i|s_i)P(s_i|q) \propto \sum_{s_i \in q} P(t_i|s_i)$$

having assumed that the probabilities q are the same for every word in s_i. Their translation model was trained on materials released by the Canadian Parliament, supplemented by a small bilingual dictionary and some simple corpus statistics. This combination proved to be quite effective, and, after reasonable parameter tuning, outperformed the results obtained using corpora alone. Similar positive results combining parallel corpora with bilingual dictionaries (and other factors) have been observed in [He and Wu 2008; Federico and Bertoldi 2002; Darwish and Oard 2003].

One of the key disadvantages of the corpus-based approach to query translation is the difficulty inherent in obtaining suitable document collections. Parallel corpora can be extremely time consuming to produce, even when restricted to specific information domains (e.g., legislative, medical) [Kashioka et al. 2003; Sun et al. 2002; Zhang et al. 2005]. This disadvantage has been partially addressed by new strategies leveraging the World Wide Web, but these techniques are only successful in relation a subset of known languages - certain low frequency languages have yet to produce enough material to train a translation model. One possible solution in-

[18] http://www.parl.gc.ca/
[19] http://www.legco.gov.hk/

volves the use of *comparable corpora*. A comparable corpus is a combination of texts which are composed independently but share the same communicative function and theme [Sheridan and Ballerini 1996]. Given a set of terms in one language, comparable corpora can be used to identify contexts which contain equivalent or related expressions in another language [Sheridan and Ballerini 1996; Peters and Picchi 1996; Braschler and Ripplinger 2004; Franz et al. 1999]. Equivalence is usually established using co-occurrence statistics, with paired terms stored in a similarity thesaurus. Experiments with the granularity of the alignment (e.g., document-level, page-level, passage-level, sentence-level) suggests there is an inverse relationship between granularity and overall retrieval effectiveness [Franz et al. 1999].

5.4 Summary

To summarise, the crucial factor for all of the direct translation techniques appears to be the quality of the translation resources that are available. Language resources with poor coverage will generally lead to poor retrieval effectiveness. Resources can be enriched (using some of the techniques described above) but this will only provide a partial solution. Where high quality translation resources are available, the retrieval effectiveness of direct translation systems is now extremely high.

As described above, systems that utilise machine readable dictionaries will take one of two mutually exclusive approaches to the selection of translation terms. Single selection translation systems have recorded some highly impressive experimental results (e.g., [Gao et al. 2001; Gao and Nie 2006], where the authors described a single selection system using a MRD that exceeded the monolingual baseline). On the negative side, single selection translators risk a significant 'hit' in terms of retrieval effectiveness should their initial selection prove incorrect, plus they forego the opportunity to expand the query with translation alternatives.

Multiple selection translators can produce results that approach the monolingual baseline (e.g., [Wang and Oard 2006] described an experiment using English and French resources in which a multiple selection translator managed 97% of monolingual retrieval effectiveness). This type of system usually benefits from an 'expansion effect' when translation alternatives are added to the translated query. Furthermore, in the context of an enriched query, incorrect translations of single terms are generally less critical (which compares favourably with the single selection translation approach). The disadvantages of multiple selection translation tend to cluster around term weighting process. Systems based on Pirkola's structural model do not incorporate translation probabilities into their calculations, and this may lead to the inaccurate weighting of terms. Translators based on later extensions to this method (i.e., the probabilistic and bidirectional models) can incorporate these translation probabilities, but they are limited by the availability of the parallel texts and dictionaries required to generate them.

The embedded approach to translation is an interesting development. This type of system simplifies the CLIR framework presented in Figure 5 by merging the retrieval and translation modules. Results from embedded systems have been positive, approaching the monolingual baseline [Lavrenko et al. 2002]. However, this approach is not without its problems. Separate tools are required to train the translation model (which serves a very different purpose to the retrieval model) and this process may result in the loss of non-translation related terms. Furthermore, the

Table I: Summary of direct translation systems

Technique	Authors	Notes
Single Translation Selector		
Phrasal translation	Ballesteros & Croft 1997	Using larger translation units instead of word by word translation
Co-occurrence statistics from target documents	Ballesteros & Croft 1998	Disambiguation based on target document collection instead of separate training corpus
Co-occurrence statistics incorporating word order	Jang et al. 1999	An attempt to capture term dependencies
Co-occurrence statistics using a greedy algorithm	Adriani 2000	Describes a simplified way to select the best translation using co-occurrence statistics
Co-occurrence statistics from Web documents	Maeda et al. 2000	Describes how to use Web corpora during disambiguation of the translation
Statistical noun phrase translation	Gao et al. 2001, Gao et al. 2006	The integration of noun phrases into the translation model
Decaying co-occurrence model	Gao et al. 2002	Considers term-term distance to capture more accurate relations
Syntactic dependency	Gao et al. 2002, Gao et al. 2006	Captures syntactic dependencies to increase the accuracy of translations
HMM-based query translation model	Federico & Bertoldi 2002	Integrating the translation model with co-occurrence information
Maximum coherence model	Liu et al. 2005	An attempt to capture translation dependencies
Iterative expectation-maximisation	Monz and Dorr 2005	Describes use of a dynamic decision making process (as opposed to greedy algorithm)
Markov chain model for query translation	Cao et al. 2007	Extends query translation to query expansion using monolingual relations
Graph-based query disambiguation	Zhou et al. 2008	Captures translation and query term dependencies using graph-based analysis
Multiple Translation Selector		
Structured query translation	Pirkola 1998	Query structuring, plus separation of TF and DF
Bidirectional translation	McCarley 1999	Bidirectional translation combined with results merging
Sum-based structured query translation	Kwok 2000	Structured query translation with a modified DF component
Balanced translation	Levow & Oard 2000, Leek 2000	Balancing the relatively high weights of uncommon translations
Probabilistic structured query translation	Darwish and Oard 2003	A probabilistic model in which the translation probabilities can be included
Bidirectional translation by meaning-matching	Wang and Oard 2006	Meaning-match model that maps the translation to synonymy knowledge
The smoothed document model	Xu et al. 2001, Xu & al. 2005	Embedded translation using the smoothed document model
Relevance model	Lavrenko et al. 2002	Coarser-grained embedded translation with an expansion effect
Statistical translation model	Kraajj et al. 2003	Embedded translation using language models in the retrieval framework

Table II: Summary of direct translation systems (cont.)

Technique	Authors	Notes
Transliteration		
Treating source query as 'mis-spelled' query	Buckley et al. 2000	Substring spelling corrections for similar languages
Targeted s-gram matching	Pirkola et al. 2002	Categorical matching based on character contiguity
Fuzzy matching	Pirkola et al. 2003	N-gram fuzzy matching based on automatically generated rules
Matching based on non-adjacent digrams	Keskustalo et al. 2003	Matching based on bigrams composed from adjacent/non-adjacent characters
Statistical phonetic mapping	Qu et al. 2003	Dictionary based phonetic mapping using probabilistic mapping
Statistical phonetic translation	Virga & Khudanpur 2003	Statistical matching and translation of phonetic representations
Statistical learning model	Cao et al. 2007	Statistical learning model based on strictly monotonic alignment
Coverage - external resources		
Anchor text mining	Lu et al. 2002, Lu et al. 2004	Mining translations from hyperlinks
Web mining for OOV terms	Zhang et. al. 2004, Cheng et al. 2004	Mining translations from monolingual Web pages
Mining for OOV terms using patterns	Zhou et al. 2008	Mining the Web for translations using automatically generated patterns
Wikipedia-based query translation	Schonhofen et al. 2008	Translations mined from wikipedia hyperlinks, plus query disambiguation.
MRD query translation plus Wikipedia	Jones et al. 2008	Translation using domain-specific dictionaries and wikipedia-mined phrases
Statistical patterns for OOV terms	Shi 2010	Mining the Web for OOV terms using statistically generated patterns.
MT & Corpus-based Translation		
Probabilistic translation model	Nie et al. 1998	Translation probabilities calculated using a parallel corpus
MT based translation model	Franz et al. 1999	MT based translation model trained on a parallel corpus and a comparable corpus
Machine translation for queries	Kwok 1999	Using an MT system for 'quick' translations of queries
HMM-based query translation model	Federico & Bertoldi 2002	Integrating translation model with co-occurrence information in parallel corpus
Translation enhancement	He and Wu 2008	Revising translation probabilities using relationships in parallel corpus

quality of the training corpus operates as a constraint on the system as a whole, particularly noticeable when the corpus is not tightly bound to the test collection. Finally, since the relevance model relies on PRF rather than a translation model, the absence of finely grained translations at the point where relevance is evaluated makes it difficult to significantly improve overall retrieval effectiveness.

Transliteration is a supplemental technique that can be used when language resources are incomplete. As such, it can be used by single or multiple selection translation systems. In two separate studies, orthographic mapping of OOV terms have increased the retrieval effectiveness of a baseline system (i.e., MRD only) by over 60% [AbdulJaleel and Larkey 2003; Qu et al. 2003]. However, this technique

has an obvious weakness. It cannot be applied unless the languages involved share a similar character set. Phonetic mappers may offer a possible solution to this weakness, but different (and ambiguous) pronunciations of the same word in a given language can complicate their implementation. An alternative approach to OOV terms which mines the web for gratuitous translations shows great promise, but is currently only viable for certain language pairings (e.g., Asian languages and English).

The advantages of machine translation include the wide availability of high quality, inexpensive MT systems, and their demonstrated effectiveness in experimental trials - MT systems have been routinely achieving near monolingual effectiveness for several years [Ferro and Peters 2009]. They are also very popular in the CLIR research community, which simplifies the process of developing and troubleshooting a new MT-based system. The disadvantages of machine translation tend to focus on the issue of control - off-the-shelf MT systems typically deny researchers the opportunity to influence various operational factors significant during the experimental process (e.g., training resources, translation output). Furthermore, off-the-shelf systems usually produce one translation term per source term. This means that MT-based systems suffer all the negative aspects of single selection translation. Finally, even when quality language resources are available, the end product is by no means guaranteed [Savoy and Dolamic 2009].

Corpus-based translation has a number of notable advantages. Parallel corpora (authored in various major world languages) can be obtained from a number of different sources. Once obtained, they can be supplemented with other language resources (e.g., machine readable dictionaries). Furthermore, there is also a useful expansion effect inherent in sentence-sentence alignment that can enrich the translated text. Where parallel corpora for specific language pairs is not available, use of a comparable corpus may be indicated.

In terms of similarities between direct translation techniques, there is an obvious kinship between machine translation and the corpus-based techniques. Both these approaches require large document collections, and both approaches utilise statistical translation and statistical language models. However, they should not be confused. Statistical MT was developed to perform fluency-critical, term-to-term translations of entire text passages. This means that researchers using statistical MT systems to translate queries are effectively using a 'black box' that maps words. Researchers using corpus-based systems, on the other hand, have iteratively refined their translation models to address the problem of very short query strings (in which word order and syntax are largely irrelevant).

A summary of direct translation techniques can be found in Tables I and II. These tables contain a subset of the papers cited in the section above, ordered by general approach and the date of publication. Each entry in this table is, in our opinion, the *seminal* paper for a specific translation technique. Readers interested in learning more about these techniques should refer to Appendix B, which outlines the experimental settings used in each case (e.g., the translation resources, the languages involved, the test collections, the underlying IR system).

6. INDIRECT TRANSLATION

Indirect translation is a common solution when there is a shortage (or absence) of resources supporting direct translation. Indirect translation relies upon the use of an *intermediary* which is placed between the source query and the target document collection. In the case of *transitive translation*, the query will be translated into an intermediate language (or indeed languages) to enable comparison with the target document collection. In the case of *dual translation* systems, both the query and the document representations are translated into the intermediate language. The intermediate language itself can be a concrete human language or *abstract* in nature (e.g., a shared semantic space or conceptual interlingua). We will discuss each of these approaches in turn.

6.1 Transitive Translation

As mentioned above, transitive translation relies upon the use of a pivot language that acts as an intermediary between the source query and the target document collection [Ruiz et al. 1999; Ballesteros and Sanderson 2003; Lehtokangas et al. 2004; Kishida and Kando 2005; Gey 2007]. Sometimes more than one pivot language is employed. Gollins and Sanderson used a *triangulated* transitive approach employing two pivot languages to address the problems of translation ambiguity [Gollins and Sanderson 2001]. The following is a description of their transitive process, which assumes the use of German queries, English documents, Spanish and Dutch pivot languages:

> "If translating a German query word 'fisch', a Spanish translation dictionary suggests two terms 'pez, pescado' and the Dutch gives 'vis'. Taking each of these in turn, translating the Spanish terms to English gives 'pitch, fish, tar, food, fish', while Dutch to English gives 'pisces the fishes, pisces, fish'. Each of the transitive translations introduced much translation error largely due to word sense ambiguity. If we take the term that is in common from the two transitive translations, we have 'fish', a good unambiguous translation of the original German word [Gollins and Sanderson 2001]."

In later work, Mayfield and McNamee [2004] demonstrated that transitive retrieval without direct translation does not suffer the drop-off in retrieval quality sometimes reported for transitive retrieval with direct translation, and that triangulation combining multiple transitive runs without direct translation outperforms direct translation-free retrieval. In another interesting study, Kraaij [2003] examined the efficacy of pivot languages in the context of probabilistic retrieval. He concluded that the crucial factor for a probabilistic CLIR engine utilising transitive translation was the lexical coverage of the complete translation chain, a chain heavily dependent on the weakest translation resource. Given adequate translation resources, a CLIR engine exploiting a pivot language can achieve results approaching (but not exceeding) a direct translation system (see further [Lehtokangas et al. 2008], where a transitive system scored between 85%-93% of the monolingual baseline).

The application of triangulated translation to structured queries has been somewhat controversial. In [Lehtokangas et al. 2004] the authors compared unstruc-

tured/structured bilingual, transitive, and triangulated translations of various queries. They concluded that structured queries always outperformed unstructured queries, and that triangulation is actually harmful when query structuring is employed. In stark contrast to this result, an earlier paper by [Ballesteros and Sanderson 2003] reported that the combination of query structuring and an intersected triangulation of multiple pivot languages was the most effective combination in terms of retrieval.

6.2 Dual Translation

Dual translation systems attempt to solve the query-document mismatch problem by translating the query representation *and* the document representations into some 'third space' prior to comparison. This 'third space' can be another human language, an abstract language or a conceptual interlingua. This general category also includes translation techniques that induce a *semantic correspondence* between the query and the documents in a cross-language dual space defined by the documents.

One of the earliest published dual translation systems used a technique known as *latent semantic indexing* (LSI) [Blei et al. 2003; Landauer et al. 1998]. LSI is a well established technique for extracting concepts from a text corpus [Deerwester et al. 1990; Dumais 1993; 1995]. It is based on *singular value decomposition* (SVD), a technique developed for linear algebra. A full SVD is a loss-free decomposition of a term-document matrix A, which is decomposed into two orthogonal matrices X and V and a diagonal matrix Σ:

$$A = X\Sigma V^{'}$$

Estimating less singular values and their corresponding singular vectors leads to a low-rank (denoted as k) approximation to the original matrix A:

$$A \approx X_k \Sigma_k V_k^{'}$$

so that documents are no longer represented by terms but by latent concepts. New queries (denoted as q) are represented in terms of these latent concepts by folding them into the LSI model:

$$\overrightarrow{q_k} = \Sigma_k^{-1} U_k^{'} \overrightarrow{q}$$

Documents are folded into this semantic space in the same way as queries, enabling cosine correlation as follows:

$$Score(q, d) = cos(\overrightarrow{q_k}, \overrightarrow{d_k})$$

Cross-language LSI requires a text corpus consisting of comparable documents. These documents are merged and then added to the document-term matrix. This results in a multi-lingual feature space, to which the standard SVD algorithm can be applied. Dumais et al. [1997] described an experiment in which this approach led to monolingual-equivalent retrieval effectiveness. Unfortunately, this experiment used a non-standard test collection. Working with an orthodox corpus, Mori et al. [2001] reported that the effectiveness of LSI-based retrieval engines was always lower

than systems using direct translation. Furthermore, low effectiveness was usually accompanied by high computational cost. Latent Dirichlet allocation (LDA) [Blei et al. 2003], a probabilistic technique analogous to latent semantic analysis, suffers from the same problems as LSI [Cimiano et al. 2009].

As an alternative to latent semantics, a translation system can use *explicit semantic analysis* (ESA). ESA is simply another method for computing the semantic relatedness between words. ESA indexes documents in relation to some pre-existing external knowledge base (e.g., ODP[20]). This index indicates how strongly a given word in a document is associated with a specific article in the knowledge base. In this model, each article in the knowledge base is regarded as a concept. As with the latent semantic model, two words or two texts can be semantically related whether or not they share the same vocabulary.

The development of ESA has always been closely associated with the Wikipedia corpus [Anderka et al. 2009; Cimiano et al. 2009; Potthast et al. 2008; Sorg and Cimiano 2008]. As mentioned above, this corpus provides a rich set of bidirectional hyperlinks joining parallel articles [Sorg and Cimiano 2008]. This makes Wikipedia an ideal testing ground for ESA. Given a document d in language e, and assuming the availability of a linking function $lanlink(e \rightarrow c)$ mapping an article of Wikipedia W_e to its corresponding article in Wikipedia W_c, that document can be indexed with respect to another language c, by transforming the vector $\Phi_e(\overrightarrow{d})$ into a corresponding vector in the space occupied by articles of W_c:

$$\Psi_{e \rightarrow c} : \Re^{|W_e|} \rightarrow \Re^{|W_c|}$$

This linking function is calculated as follows:

$$\mathcal{F}_{e \rightarrow c} \langle s_1^e, ..., s_{|W_e|}^e \rangle = \langle t_1^c, ..., t_{|W_c|}^c \rangle$$

where

$$t_x^c = \sum_{y \in \{y* | lan-link_{e \rightarrow c}(w_{y*}) = w_x\}} s_y^e$$

with $1 \leq x \leq |W_e|, 1 \leq y \leq |W_c|$. Therefore, in order to get the representation of a document d in language e with respect to Wikipedia W_c, it is simply a case of computing the function:

$$\mathcal{F}_{e \rightarrow c}(\Phi_e(\overrightarrow{d}))$$

The score produced by this model can subsequently be calculated as the cosine similarity:

$$Score(q, d) = cos(\Phi_c(\overrightarrow{q_c}), \mathcal{F}_{e \rightarrow c}(\Phi_e(\overrightarrow{d_e})))$$

As described by [Anderka and Stein 2009], ESA was inspired by the application of the generalized vector space model (§2.4.5) to the specific problems of cross-language information retrieval [Sheridan and Ballerini 1996; Carbonell et al. 1997].

[20]http://www.dmoz.org/

Table III: Summary of indirect translation systems

Technique	Authors	Notes
Transitive Translation		
Triangulated translation	Collins & Sanderson 2001	Transitive translations using several pivot languages
Probabilistic, transitive translation	Kraajj 2003	Transitive translation used in a probabilistic model
Triangulation	Mayfield and McNamee 2004	Triangulation without direct translation combining multiple transitive runs
Transitive translation using dictionaries	Lehtokangas et al. 2004	Transitive translation with structured query translation
Dual Translation		
Dual-space based retrieval	Sheridan & Ballerini 1996	Dual-space based retrieval with a comparable corpus
Cross-language LSA	Dumais et al. 1997	Using an SVD-based latent semantic model in CLIR
Transitive translation through interlingua	Ruiz et al. 1999	Transitive translation through conceptual interlingua (WordNet)
Cross-language LSA with multiple word spaces	Mori et al. 2001	Dividing a large bilingual corpus into smaller sub-corpora
Cross-language ESA	Potthast et al. 2008, Sorg & Cimiano 2008	ESA using corresponding Wikipedia articles written in different languages
LSA and ESA	Cimiano et al. 2009	A comparison between these two techniques

Concrete implementations of the ESA model have recently produced results comparable with LSI-based systems [Cimiano et al. 2009; Anderka et al. 2009].

Experiments using an abstract pivot language to enable dual translation are extremely rare [Ruiz et al. 1999; Kishida and Kando 2005]. In [Ruiz et al. 1999] the authors constructed a conceptual interlingua using Wordnet, a lexical database for the English language.[21] This dual-translation methodology involved attaching synonymous terms from several different languages to sets of Wordnet cognitive synonyms (synsets), then mapping directly between the interlingual representations of documents and user queries authored in multiple languages. Unfortunately, initial results were disappointing.

6.3 Summary

To summarise, the transitive approach to translation is a viable option when the resources required for direct translation are unavailable. CLIR researchers have achieved near monolingual retrieval effectiveness using transitive methods, which is impressive [Gollins and Sanderson 2001]. Where high quality language resources are available, direct translation will almost always produce superior results.[22] It is interesting to note that most of the studies using transitive translation have thus far selected resource-rich language pairs, probably to facilitate comparison with direct translation systems. It is hoped that future work will explore the transitive

[21] http://wordnet.princeton.edu/

[22] A notable exception to this rule was reported in [Lehtokangas et al. 2004], where the authors used structured query translation to good effect when resolving translation ambiguity.

approach in relation to resource-poor languages.

When considering approaches to query translation that rely on inducing semantic correspondence, the key challenge seems to be cost. In the LSI model, complexity grows linearly with the number of dimensions and the number of documents. No successful experiment using more than one million documents has ever been reported [Manning et al. 2008]. An attempt to parcel a large document collection into sub-sets (which were subsequently indexed as *sub-spaces*) encountered difficulties during re-integration [Mori et al. 2001]. Due to these problems, future applications for LSI in cross-language retrieval may well be limited.

The ESA model suffers from the same drawbacks as its latent cousin. Computational cost is still very high, placing a practical cap on the number of dimensions that can be used (10,000 is the standard figure [Potthast et al. 2008]). Furthermore, overall effectiveness can degrade when articles are distributed across a high number of dimensions. Cimiano et al. [2009] described a trial in which the entire Wikipedia collection was used to build a semantic space. One document mapped to all ten thousand dimensions, very few of which were truly semantically related. The retrieval effectiveness of systems using LSA or ESA seems to be considerably lower than other indirect translation systems. Semantic approaches to query translation, though attractive in theory, are still some distance from practical use.

Table III provides a condensed view of the techniques discussed in §6. As in §5, we have selected a representative paper for each translation approach. Experimental settings (e.g., the translation resources, document collections) have been relegated to Appendix B.

7. DOCUMENT TRANSLATION

Throughout §5-6 we concentrated on CLIR engines that translate queries into the same language as the target document collection. In the following section we discuss an alternative approach, much less popular but equally valid, which translates the document collection into the same language as the query. To date, there has been very little work done in this particular area. However, this does not indicate that document translation is in some sense infeasible. In many way, it is easier to translate a full document than a query. Firstly, the translation process is simplified by the additional contextual information present in longer documents but absent from queries. Secondly, while it is critically important that each term in a very short query is correctly translated, the significance of a single term in longer portions of text is much lower. Unfortunately, these two advantages tend to be outweighed by the prohibitive effort required to translate an entire document collection.

According to the literature that is available, the most efficient approach to document translation usually involves processing the test collection with some sort of fully automatic machine translation system [Oard and Hackett 1997; Oard 1998; McCarley 1999](see also [Oard and Ertunc 2002]). Early experiments with MT-based document translation suggested it could (for certain fuzzily-defined applications) be just as effective as query translation [Oard and Hackett 1997; Oard 1998]. This result was subsequently confirmed by [Franz et al. 1999], who compared document and query translation using a commercial MT system and found no significant differences between the two. Interestingly, [McCarley 1999] downplayed the signifi-

cance of the discussion altogether, stating that the critical factor for CLIR engines was not the choice between query and document translation, but the relationship between the language pairs themselves.

It is important to realise that document translation can utilise all of the direct and indirect translation techniques discussed above, with little or no modification. To illustrate this statement, let us consider the application of just one of those techniques. In §5.1.1 we discussed the embedded query translation model reported in [Kraaij et al. 2003]. In the same paper, Kraaij et al. proposed an extension to this model designed to reduce translation noise which folded document translations into the calculations as follows:

$$P(s_i|M_{d_c}) = \sum_{t_j \in V_c} P(s_i|t_j) P_{ML}(t_j|M_{d_c})$$

where M_{d_c} is the language model for the document and s_i is a term in the target language. In this extension, the ranking of documents was calculated using the following formula:

$$P(q_e|d_c) = \sum_{s_i \in V_e} \sum_{t_j \in V_c} P(s_i|M_{q_e}) \log P(s_i|t_j) P_{ML}(t_j|M_{d_c})$$

Using this document translation model, Kraaij et al. achieved around 90% of monolingual IR effectiveness (English-French), out-performing a MT system. Hopefully, this quick example demonstrates the ease with which techniques developed for query translation can be re-purposed by document translation systems.

8. CONCLUSION

In this survey we have outlined the various types of techniques that can be used when translating queries and/or documents in the context of cross-language information retrieval. The systems we have described stand at the intersection of several complementary research fields including information retrieval, computational linguistics, information theory and natural language processing. The research field that they belong to is innovative, productive and, most of all, necessary. There is a global need for *information access* which transcends the dictates of language. Cross-language information retrieval systems offer a reasonable, technically feasible mechanism through which access can be provided.

Looking to the future, we anticipate a steady increase in the quality and quantity of translation resources available to researchers. Another broad theme may be the emergence of a more unified model that merges the translation and retrieval functions of a CLIR system.

APPENDIX

A. STATISTICAL MACHINE TRANSLATION

Statistical MT generates translations using statistical models whose parameters are derived from the analysis of bilingual text corpora. Inspired by Shannon's noisy channel coding theorem [Shannon and Weaver 1963], Brown et al. published the seminal paper on this topic in [Brown et al. 1993]. In that paper, the authors

calculate translation probabilities in the following way - given a sentence S_e in one language, the system's task is to find a translation S_c in a target language such that the probability $P(S_c|S_e)$ is maximised. This probability can be estimated by multiplying the *a priori* probability $P(S_c)$ and a conditional probability $P(S_e|S_c)$ using the Bayes rule, so that:

$$S_c = \arg\max_{S_c} P(S_c|S_e) = \arg\max_{S_c} \frac{P(S_e|S_c)p(S_e)}{P(S_c)} = \arg\max_{S_c} P(S_e|S_c)P(S_c)$$

Therefore, translation involves two discrete steps. The first step involves estimating the probability $P(S_c)$, which specifies the likelihood that S_c is generated. This is usually estimated using n-gram analysis [Brown et al. 1993]. The second step estimates $P(S_e|S_c)$, which encodes the probability that S_c is the the translation of S_e. CLIR researchers often select IBM's model 1 for this second step [Brown et al. 1993; Och and Ney 2003]. Model 1 is a fairly simple word-based translation model that ignores more sophisticated variables like word position, distortion and fertility. Readers interested in learning more about statistical MT are directed to [Brown et al. 1993; Lopez 2008; Goutte et al. 2009] for further details.

APPENDIX

B. EXPERIMENTAL DETAILS

In §5-6 we identified a number of landmark publications which could be considered representative of specific translation techniques. We presented these influential publications in Tables I-III. Tables IV-VI provide supplemental information which may be useful to readers interested in experimental design. When reading these tables, please note that we have used the appropriate ISO language codes throughout.[23] In the vast majority of cases, the test collections used were contributed by one of the following evaluation bodies:

[23]Recommendation ISO 639-1:2002 - Codes for the representation of names of languages – Part 1: Alpha-2 code, http://www.loc.gov/standards/iso639-2/iso639jac.html

Name	Description	Resource Link
TREC	Text REtrieval Conference	http://trec.nist.gov/
NTCIR	NII Test Collection for IR Systems	http://research.nii.ac.jp/ntcir/index-en.html
CLEF	cross-language Evaluation Forum	http://www.clef-campaign.org/
Kompas	kompas Indonesia corpus	http://www.kompas.com/
TDT	Topic Detection and Tracking	http://projects.ldc.upenn.edu/TDT/
SDA	Swiss News Agency	http://www.sda.ch/
Multext	Multilingual Text Tools and Corpora	http://aune.lpl.univ-aix.fr/projects/MULTEXT/
JRC-Acuis	EU Joint Research Center-Acquis Multilingual Parallel Corpus	http://langtech.jrc.it/JRC-Acquis.html

Most of the information retrieval systems used in these papers are open source or freeware, as follows:

Name	Resource Link
INQUERY	[Callan et al. 1992]
SMART	[Salton 1971]
Namazu	http://www.namazu.org/
Okapi	http://www.soi.city.ac.uk/ andym/OKAPI-PACK/
Lemur	http://www.lemurproject.org/
Indri	http://www.lemurproject.org/indri/
PIRCS	http://ir.cs.qc.edu/pircs.html
PRISE	[Oard 1999]
BNN	[Miller et al. 1999]
HAIRCUT	[McNamee et al. 2002]
Zettair	http://www.seg.rmit.edu.au/zettair/
Hungarian Academy of Science	[Benczur et al. 2003]
SPIDER	[Schäuble 1993]

Finally, here is an overview of the translation resources used in the papers we have selected. Most are still publicly available.

Name	Resource Link
Collins	http://www.collinslanguage.com/
An English-Indonesian Dictionary	http://www.amazon.com/Indonesian-English-Dictionary-John-M-Echols/dp/0801421276
EDR dictionary	http://www2.nict.go.jp/r/r312/EDR/index.html
EDICT dictionary	http://www.csse.monash.edu.au/ jw-b/edict.html
Linguistic Data Consortium (LDC)	http://www.ldc.upenn.edu/
DING dictionary	http://www-user.tu-chemnitz.de/ fri/ding/
Hansard	http://www.publications.parliament.uk/pa/cm/cmhansrd.htm
United Nations	http://www.un.org
EUROPARL	http://www.europarl.europa.eu/
CEDICT dictionary	http://www.mdbg.net/chindict/chindict.php?page=cedict
Transperfect MT system	http://www.otek.com.tw
Babelfish MT system	http://babelfish.yahoo.com/
EuroWordNet	http://www.illc.uva.nl/EuroWordNet/
CELEX database	http://www.kun.nl/celex/
Kielikone dictioanry	http://www.kielikone.fi/en/
An English-Spanish Lexicon	http://www.activa.arrakis.es
HKNews	http://library.ust.hk/res/beyond/News/Hong_Kong_SAR/
SYSTRAN MT system	http://www.systransoft.com
CETA dictionary	MRM Corporation, Kensingston,MD

When the resource is marked "N/A", the authors did not explicitly name the resources used.

ACKNOWLEDGMENTS

This research was partially supported by a PHD scholarship from the University of Nottingham and funding from the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (www.cngl.ie) at University of Dublin, Trinity College. The authors would also like to thank the anonymous reviewers who significantly improved the quality of this manuscript during preparation.

REFERENCES

AbdulJaleel, N. and Larkey, L. S. 2003. Statistical transliteration for English-Arabic cross language information retrieval. In *Proceedings of the Twelfth International Conference on Information and Knowledge Management*. ACM Press, New York, NY, USA, 139–146. 956890.

Adriani, M. 2000. Using statistical term similarity for sense disambiguationin cross-language information retrieval. *Inf. Retr. 2,* 1, 71–82.

Adriani, M. and Wahyu, I. 2005. The performance of a machine translation-based English-

Table IV: Direct translation systems - experimental settings

Authors	Lang.	Corpus	IR System	Translation Resources
Single Translation Selector				
Ballesteros & Croft 1997	es, en	TREC	INQUERY	Collins MRD
Ballesteros & Croft 1998	es, en	TREC AP	INQUERY	Collins MRD
Jang et al. 1999	ko, en	TREC-6	SMART	N/A
Adriani 2000	id, en	TREC AP, Kompas	INQUERY	English-Indonesian Dictionary
Maeda et al. 2000	ja, en	NTCIR 1	Namazu	EDR and EDICT MRD
Gao et al. 2001	zh, en	TREC 5, 6, & 9	SMART	In-house MRD
Gao et al. 2002	zh, en	TREC 9	Okapi	In-house MRD
Gao et al. 2006	zh, en	TREC 5, 6, & 9	Okapi	In-house MRD
Federico & Bertoldi 2002	it, en	CLEF 2000 & 2001	N/A	Collins MRD
Liu et al. 2005	zh, en	TREC AP, WS & DOE	SMART	Linguistic Data Consortium (LDC) MRD
Monz and Dorr 2005	de, en	CLEF 2003	N/A	DING MRD
Cao et al. 2007	zh, en	TREC 5, 6, 9, NTCIR 3	Lemur	LDC MRD
Zhou et al. 2008	zh, en	NTCIR 3, 4, 5 & 6	Lemur	LDC MRD
Multiple Translation Selector				
Pirkola 1998	fi, en	TREC	INQUERY	In-house MRD
McCarley 1999	fr, en	TREC 6, 7	N/A	Hansard and United Nations corpora
Kwok 2000	zh, en	TREC	PIRCS	LDC MRD
Levow & Oard 2000	zh, en	TDT 3	PRISE	LDC MRD & zh-en MRD
Leek 2000	zh, en	TDT 3	BNN IR system	LDC MRD
Darwish and Oard 2003	ar, en	TREC 2002	N/A	In-house training data
Wang and Oard 2006	zh, fr, en	TREC 5, 6 & CLEF 2001-3	Perl search engine	EUROPARL and various resources
Xu et al. 2001, Xu & al. 2005	es, zh, en	TREC 5, 9	N/A	LDC MRD, CETA MRD, HKNews, SYSTRAN MT
Kraajj et al. 2003	fr, it, en	CLEF 2000-02	N/A	Parallel corpus mined from the Web
Lavrenko et al. 2002	zh, en	TREC 9	N/A	LDC MRD, CETA MRD, HKNews corpus

Table V: Direct translation systems - experimental settings

Authors	Lang.	Corpus	IR System	Translation Resources
Transliteration				
Buckley et al. 2000	de, fr, zh, en	TREC 6	SMART	Various resources
Qu et al. 2003	ja, en	CLEF 2001, 2002	N/A	EDICT MRD
Virga & Khudanpur 2003	zh, en	TDT 2	HAIRCUT	LDC MRD
Cao et al. 2007 (b)	zh, en	TREC 5, 6 & 9	N/A	LDC MRD
Coverage - external resources				
Lu et al. 2004	zh, en	NTCIR 2	N/A	LDC MRD, Web
Cheng et al. 2004	zh, en	NTCIR 2	N/A	LDC MRD, Web
Zhang & Vine 2004	zh, en	NTCIR 4	Zettair	LDC & CEDICT MRD, Web
Zhou et al. 2008	zh, en	NTCIR 3, 4, 5 & 6	Lemur	LDC MRD, Web
Schonhofen et al. 2008	de, hu, en	CLEF 2007	In-house search engine	Wiktionary
Shi 2010	zh, en	TREC 5, 6	Lemur	LDC MRD, Web
MT & Corpus-based Translation				
Nie et al. 1998	fr, en	TREC 6	N/A	Hansard corpora
Franz et al. 1999	de, fr, it, en	TREC 7	N/A	SDA corpora, Babelfish MT system
Kwok 1999	zh, en	TREC	PIRCS	TREC, Transperfect MT system
Federico & Bertoldi 2002	it, en	CLEF 2000 & 2001	N/A	Collins MRD
He and Wu 2008	zh, en	TDT 4, 5	Indri	In-house MRD

Indonesian CLIR system. In *CLEF 2005: Workshop on Cross-Language Information Retrieval and Evaluation*. Vienna, Austria.

Agirre, E., Di Nunzio, G. M., Ferro, N., Mandl, T., and Peters, C. 2009. CLEF 2008: Ad hoc track overview. In *Proceedings of the 9th Cross-language Evaluation Forum Conference on Evaluating Systems for Multilingual and Multimodal Information Access*. CLEF'08. Springer-Verlag, Aarhus, Denmark, 15–37.

Alfonseca, E., Bilac, S., and Pharies, S. 2008. Decompounding query keywords from compounding languages. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*. Association for Computational Linguistics, Columbus, Ohio, 253–256.

Aljlayl, M. and Frieder, O. 2001. Effective Arabic-English cross-language information retrieval via machine-readable dictionaries and machine translation. In *CIKM '01: Proceedings of the tenth International Conference on Information and Knowledge Management*. ACM, New York, NY, USA, 295–302.

Amati, G. and Van Rijsbergen, C. J. 2002. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Trans. Inf. Syst. 20*, 357–389.

Anderka, M., Lipka, N., and Stein, B. 2009. Evaluating cross-language explicit semantic analysis and cross querying. In *Proceedings of the 10th Cross-language Evaluation Forum Conference on Multilingual Information Access Evaluation: Text Retrieval Experiments*. CLEF'09. Springer-Verlag, Corfu, Greece, 50–57.

Table VI: Indirect translation systems - experimental settings

Authors	Lang.	Corpus	IR System	Translation Resources
Transitive Translation				
Gollins & Sanderson 2001	de, es, nl, en	TREC 8	N/A	EuroWordNet, CELEX database
Kraajj 2003	de, fr, it, nl, en	CLEF 2001	N/A	Parallel corpus mined from the Web
Mayfield and McNamee 2004	de, fi, fr, it, nl	CLEF 2003	HAIRCUT	N/A
Lehtokangas et al. 2004	de, es, fi, en	CLEF 2000, 2001	INQUERY	Kielikone MRD
Ruiz et al. 1999	fr, en	TREC 8	N/A	WordNet
Dual Translation				
Sheridan & Ballerini 1996	fr, it, en	SDA	SPIDER	SDA Corpora
Dumais et al. 1997	fr, en	N/A	N/A	Hansard
Mori et al. 2001	ja, en	NTCIR 2	N/A	NTCIR 1 corpus
Potthast et al. 2008	de, en	JRC-Acuis	N/A	Wikipedia
Sorg & Cimiano 2008	de, fr, en	CLEF 2008	N/A	Wikipedia
Cimiano et al. 2009	de, fr, en	Multext, JRC-Acuis	N/A	Wikipedia

Anderka, M. and Stein, B. 2009. The ESA retrieval model revisited. In *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '09. ACM, New York, NY, USA, 670–671.

Bacchin, M., Ferro, N., and Melucci, M. 2005. A Probabilistic Model for Stemmer Generation. *Information Processing & Management 41,* 1 (January), 121–137.

Baeza-Yates, R. and Ribeiro-Neto, B. 2008. *Modern Information Retrieval*, 2nd ed. Addison-Wesley Publishing Company, USA.

Ballesteros, L. and Croft, W. B. 1997. Phrasal translation and query expansion techniques for cross-language information retrieval. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, New York, NY, USA, 84–91.

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

Ballesteros, L. and Sanderson, M. 2003. Addressing the lack of direct translation resources for cross-language retrieval. In *Proceedings of the Twelfth International Conference on Information and Knowledge Management*. ACM Press, New York, NY, USA, 147–152.

Benczur, A., Csalogany, K., Fogaras, D., Friedman, E., Sarlas, T., Uher, M., and Windhager, E. 2003. Searching a small national domain A preliminary report. In *In Proceedings of the 12th International World Wide Web Conference (WWW)*.

Berger, A. and Lafferty, J. 1999. Information retrieval as statistical translation. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, New York, NY, USA, 222–229.

Blei, D. M., Ng, A. Y., and Jordan, M. I. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res. 3*, 993–1022. 944937.

Boughanem, M., Chrisment, C., and Nassr, N. 2002. Investigation on disambiguation in CLIR: aligned corpus and bi-directional translation-based strategies. In *Revised Papers from the*

Second Workshop of the Cross-Language Evaluation Forum on Evaluation of Cross-Language Information Retrieval Systems. Springer-Verlag, 158–168.

Braschler, M. and Ripplinger, B. 2004. How effective is stemming and decompounding for German text retrieval? *Inf. Retr. 7,* 3-4, 291–316.

Broglio, J., Callan, J. P., and Croft, W. B. 1993. INQUERY system overview. In *Annual Meeting of the ACL-Proceedings of a Workshop on held at Fredericksburg, Virginia: September 19-23, 1993*. Fredericksburg, Virginia, 47–67.

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

Buckley, C., Mitra, M., Walz, J., and Cardie, C. 2000. Using clustering and superconcepts within SMART: TREC 6. *Inf. Process. Manage. 36,* 1, 109–131.

Callan, J. P., Croft, W. B., and Harding, S. M. 1992. The INQUERY retrieval system. In *Proceedings of the Third International Conference on Database and Expert Systems Applications*. Springer-Verlag, Valencia, Spain, 78–83.

Cao, G., Gao, J., and Nie, J.-y. 2007. A system to mine large-scale bilingual dictionaries from monolingual Web. In *11th Machine Translation Summit (MT Summit XI)*. Copenhagen, Denmark, 57–64.

Cao, G., Gao, J., Nie, J.-Y., and Bai, J. 2007. Extending query translation to cross-language query expansion with markov chain models. In *Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management*. ACM, New York, NY, USA, 351–360.

Carbonell, J. G., Yang, Y., Frederking, R. E., Brown, R., Geng, Y., and Lee, D. 1997. Translingual information retrieval: A comparative evaluation. In *In Proceedings of the 15th International Joint Conference on Artificial Intelligence*. 708–714.

Chen, A. 2002. Cross-language retrieval experiments at CLEF-2002. In *Proceedings of Evaluation of Cross-Language Information Retrieval Systems: Third Workshop of the Cross-Language Evaluation Forum*. Vol. 2785/2003. Springer Berlin / Heidelberg, 28–48.

Chen, J., Chau, R., and Yeh, C.-H. 2004. Discovering parallel text from the World Wide Web. In *Proceedings of the Second Workshop on Australasian Information Security, Data Mining and Web Intelligence, and Software Internationalisation - Volume 32*. Australian Computer Society, Inc., Dunedin, New Zealand, 157–161.

Chen, J. and Nie, J.-Y. 2000. Parallel web text mining for cross-language IR. In *Proceedings of RIAO-2000: Content-Based Multimedia Information Access*. College de France, Paris, France, 188–192.

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

Chew, P. A., Verzi, S. J., Bauer, T. L., and McClain, J. T. 2006. Evaluation of the bible as a resource for cross-language information retrieval. In *Proceedings of the Workshop on Multilingual Language Resources and Interoperability*. Association for Computational Linguistics, Sydney, Australia, 68–74.

Cimiano, P., Schultz, A., Sizov, S., Sorg, P., and Staab, S. 2009. Explicit versus latent concept models for cross-language information retrieval. In *Proceedings of the 21st International Jont Conference on Artifical Intelligence*. Morgan Kaufmann Publishers Inc., Pasadena, California, USA, 1513–1518.

Cleverdon, C. W. 1991. The significance of the Cranfield tests on index languages. In *Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '91. ACM, New York, NY, USA, 3–12.

Darwish, K. and Oard, D. W. 2003. Probabilistic structured query methods. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, New York, NY, USA, 338–344.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. 1990. Indexing by latent semantic analysis. *Journal of the American Society of Information Science 41,* 6, 391–407.

Demner-Fushman, D. and Oard, D. W. 2003. The effect of bilingual term list size on dictionary-based cross-language information retrieval. In *Proceedings of the 36th Annual Hawaii International Conference on System Sciences (HICSS'03) - Track 4 - Volume 4*. HICSS '03. IEEE Computer Society, Washington, DC, USA, 108.2–.

Dolamic, L. and Savoy, J. 2010. Retrieval effectiveness of machine translated queries. *J. Am. Soc. Inf. Sci. Technol. 61*, 2266–2273.

Dumais, S. T. 1993. Latent semantic indexing (LSI) and TREC-2. In *Proceedings of TREC*. 105–115.

Dumais, S. T. 1995. Latent semantic indexing (LSI): TREC-3 report. In *Proceedings of TREC*. 219–230.

Dumais, S. T., Letsche, T. A., Littman, M. L., and Landauer, T. K. 1997. Automatic cross-language retrieval using latent semantic indexing. In *AAAI-97 Spring Symposium Series: Cross-Language Text and Speech Retrieval*. Stanford University, 18–24.

Fautsch, C. and Savoy, J. 2009. Algorithmic stemmers or morphological analysis? An evaluation. *J. Am. Soc. Inf. Sci. Technol. 60*, 1616–1624.

Federico, M. and Bertoldi, N. 2002. Statistical cross-language information retrieval using n-best query translations. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, New York, NY, USA, 167–174.

Ferro, N. and Peters, C. 2009. CLEF 2009 Ad hoc track overview: TEL and Persian tasks. In *Proceedings of the 10th Cross-language Evaluation Forum Conference on Multilingual Information Access Evaluation: Text Retrieval Experiments*. CLEF'09. Springer-Verlag, Corfu, Greece, 13–35.

Fox, C. 1989. A stop list for general text. *SIGIR Forum 24,* 1-2, 19–21.

Franz, M., McCarley, J., and Roukos, S. 1999. Ad hoc and multilingual information retrieval at IBM. In *TREC-7*. 157–168.

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

Gao, J. and Nie, J.-Y. 2006. A study of statistical models for query translation: finding a good unit of translation. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, New York, NY, USA, 194–201.

Gao, J., Nie, J.-Y., Wu, G., and Cao, G. 2004. Dependence language model for information retrieval. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, New York, NY, USA, 170–177.

Gao, J., Nie, J.-Y., Xun, E., Zhang, J., Zhou, M., and Huang, C. 2001. Improving query translation for cross-language information retrieval using statistical models. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, New York, NY, USA, 96–104.

Gao, J., Zhou, M., Nie, J.-Y., He, H., and Chen, W. 2002. Resolving query translation ambiguity using a decaying co-occurrence model and syntactic dependence relations. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, New York, NY, USA, 183–190.

Gao, W., Wong, K.-F., and Lam, W. 2005. Phoneme-based transliteration of foreign names for OOV problem. In *First International Joint Conference in Natural Language Processing IJCNLP 2004*. Vol. 3248/2005. Springer, Hainan Island, China, 110–119.

Gey, F. 2007. Search between Chinese and Japanese text collections. In *Proceedings of the Sixth NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and Cross-Lingual Information Access"*. Tokyo, Japan, 73–76.

Gey, F. and Chen, A. 1998. TREC-9 cross-language information retrieval (English-Chinese) overview. In *Proceedings of the Ninth Text Retrieval Conference (TREC-9)*. 15–23.

Gollins, T. and Sanderson, M. 2001. Improving cross language retrieval with triangulated translation. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, New York, NY, USA, 90–95.

Goto, I., Kato, N., Ehara, T., and Tanaka, H. 2004. Back transliteration from Japanese to English using target English context. In *Proceedings of the 20th International Conference on Computational Linguistics*. COLING '04. Association for Computational Linguistics, Geneva, Switzerland.

Goutte, C., Cancedda, N., Dymetman, M., and Foster, G., Eds. 2009. *Learning Machine Translation*. The MIT Press, Cambridge (MA), USA.

He, D. and Wu, D. 2008. Translation enhancement: a new relevance feedback method for cross-language information retrieval. In *Proceeding of the 17th ACM Conference on Information and Knowledge Management*. ACM, New York, NY, USA, 729–738.

Hedlund, T. 2002. Compounds in dictionary-based cross-language information retrieval. *Information Research 7,* 2.

Hollink, V., Kamps, J., Monz, C., and Rijke, M. D. 2004. Monolingual document retrieval for european languages. *Inf. Retr. 7,* 1-2, 33–52.

Huang, F., Zhang, Y., and Vogel, S. 2005. Mining key phrase translations from Web corpora. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Vancouver, British Columbia, Canada, 483–490.

Hull, D. 1993. Using statistical testing in the evaluation of retrieval experiments. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '93. ACM, New York, NY, USA, 329–338.

Hull, D. A. 1996. Stemming algorithms: a case study for detailed evaluation. *Journal of the American Society for Information Science 47,* 1, 70–84.

Hull, D. A. 1997. Using structured queries for disambiguation in cross-language information retrieval. In *AAAI Symposium on Cross-Language Text and Speech Retrieval*. American Association for Artificial Intelligence, 84–98.

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

Jang, M.-G., Myaeng, S. H., and Park, S. Y. 1999. Using mutual information to resolve query translation ambiguities and query term weighting. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*. Association for Computational Linguistics, College Park, Maryland, 223–229.

Jeong, K., Myaeng, S., Lee, J., and Choi, K.-S. 1999. Automatic identification and back-transliteration of foreign words for information retrieval. *Information Processing and Management 35*, 523–540.

Jin, R., Hauptmann, A. G., and Zhai, C. X. 2002. Title language model for information retrieval. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, New York, NY, USA, 42–48.

Jones, G. J. F., Fantino, F., Newman, E., and Zhang, Y. 2008. Domain-specific query translation for multilingual information access using machine translation augmented with dictionaries mined from wikipedia. In *Proceedings of the 2nd International Workshop on Cross Lingual Information Access - Addressing the Information Need of Multilingual Societies (CLIA-2008)*. Hydrabad, India, 34–41.

Kang, B.-J. and Choi, K.-S. 2000. Two approaches for the resolution of word mismatch problem caused by English words and foreign words in Korean information retrieval. In *Proceedings of the Fifth International Workshop on Information Retrieval with Asian Languages*. IRAL '00. ACM, New York, NY, USA, 133–140.

Kang, I.-H. and Kim, G. 2000. English-to-Korean transliteration using multiple unbounded overlapping phoneme chunks. In *Proceedings of the 18th Conference on Computational Linguistics - Volume 1*. Association for Computational Linguistics, Saarbrcken, Germany, 418–424.

Kang, I.-S., Na, S.-H., and Lee, J.-H. 2004. POSTECH at NTCIR-4: CJKE monolingual and Korean-related cross-language retrieval experiments. In *Proceedings of the Forth NTCIR Workshop*. National Institute of Informatics, Japan.

Kashioka, H., Maruyama, T., and Tanaka, H. 2003. Building a parallel corpus for monologues with clause alignment. In *MT Summit IX*. New Orleans, USA, 216–223.

Keskustalo, H., Pirkola, A., Visala, K., Leppanen, E., and Jarvelin, K. 2003. Non-adjacent digrams improve matching of cross-lingual spelling variants. In *Proceedings of String Processing and Information Retrieval: 10th International Symposium, SPIRE 2003*. Manaus, Brazil. 252-265.

Kishida, K. 2008. Prediction of performance of cross-language information retrieval using automatic evaluation of translation. *Library & Information Science Research 30,* 2, 138–144.

Kishida, K. and Kando, N. 2005. Hybrid approach of query and document translation with pivot language for cross-language information retrieval. In *CLEF 2005: Workshop on Cross-Language Information Retrieval and Evaluation*. Vienna, Austria.

Knight, K. and Graehl, J. 1998. Machine transliteration. *Comput. Linguist. 24,* 4, 599–612.

Korn, M., Schulz, S., Medelyan, O., and Hahn, U. 2005. Bootstrapping dictionaries for cross-language information retrieval. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, New York, NY, USA, 528–535.

Kraaij, W. 2003. Exploring transitive translation methods. In *Proceedings of 4th Dutch-Belgian Information Retrieval Workshop*. CWI in Amsterdam.

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

Kuriyama, K., Kando, N., Nozue, T., and Eguchi, K. 2002. Pooling for a large-scale test collection: An analysis of the search results from the first NTCIR workshop. *Inf. Retr. 5*, 41–59.

Kwok, K. L. 1999. English-Chinese cross-language retrieval based on a translation package. In *In Workshop of Machine Translation for Cross Language Information Retrieval, Machine Translation Summit VII*. 8–13.

Kwok, K. L. 2000. Exploiting a Chinese-English bilingual wordlist for English-Chinese cross language information retrieval. In *Proceedings of the Fifth International Workshop on Information Retrieval with Asian Languages*. ACM Press, New York, NY, USA, 173–179.

Kwok, K. L. and Grunfeld, L. 1996. TREC-5 English and Chinese retrieval experiments using PIRCS. In *TREC-5*. 133–142.

Lancaster, F. and Fayen, E. 1973. *Information Retrieval On-Line*. Melville Publishing Co.,, Los Angeles, California, USA.

Landauer, T. K., Foltz, P. W., and Laham, D. 1998. An introduction to latent semantic analysis. *Discourse Processes 25*, 259–284.

Lavrenko, V., Choquette, M., and Croft, W. B. 2002. Cross-lingual relevance models. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, New York, NY, USA, 175–182.

Lavrenko, V. and Croft, W. B. 2001. Relevance based language models. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, New York, NY, USA, 120–127.

Lee, C.-J., Chen, C.-H., Kao, S.-H., and Cheng, P.-J. 2010. To translate or not to translate? In *Proceeding of the 33rd international ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, NY, USA, 651–658.

Leek, T., Jin, H., Sista, S., and Schwartz, R. 2000. The BBN cross-lingual topic detection and tracking system. In *Working Notes of the Third Topic Detection and Tracking Workshop*. National Institutes of Standards and Technology, National Institutes of Standards and Technology.

Lehtokangas, R., Airio, E., J, K., and rvelin. 2004. Transitive dictionary translation challenges direct dictionary translation in CLIR. *Inf. Process. Manage. 40,* 6, 973–988.

Lehtokangas, R., Keskustalo, H., and Järvelin, K. 2008. Experiments with transitive dictionary translation and pseudo-relevance feedback using graded relevance assessments. *J. Am. Soc. Inf. Sci. Technol. 59*, 476–488.

Leveling, J., Zhou, D., Jones, G. J. F., and Wade, V. 2009. Document expansion, query translation and language modeling for ad-hoc IR. In *Proceedings of the 10th Cross-language Evaluation Forum Conference on Multilingual Information Access Evaluation: Text Retrieval Experiments*. CLEF'09. Springer-Verlag, Corfu, Greece, 58–61.

Levow, G.-A. and Oard, D. W. 2000. Translingual topic tracking with PRISE. In *In Working Notes of the Third Topic Detection and Tracking Workshop*. National Institutes of Standards and Technology.

Levow, G.-A., Oard, D. W., and Resnik, P. 2005. Dictionary-based techniques for cross-language information retrieval. *Inf. Process. Manage. 41,* 3, 523–547.

Lin, M.-C., Li, M.-X., Hsu, C.-C., and Wu, S.-H. 2010. Query expansion from Wikipedia and topic Web crawler on CLIR. In *Proceedings of the Eighth NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and Cross-Lingual Information Access"*. Tokyo, Japan, 101–106.

Liu, X. and Croft, W. B. 2005. Statistical language modeling for information retrieval. *Annual Review of Information Science and Technology 39,* 1, 1–31.

Liu, Y., Jin, R., and Chai, J. Y. 2005. A maximum coherence model for dictionary-based cross-language information retrieval. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, New York, NY, USA, 536–543.

Lopez, A. 2008. Statistical machine translation. *ACM Comput. Surv. 40,* 3, 1–49.

Loponen, A. and Järvelin, K. 2010. A dictionary- and corpus-independent statistical lemmatizer for information retrieval in low resource languages. In *Proceedings of the 2010 International Conference on Multilingual and Multimodal Information Access Evaluation: Cross-language Evaluation Forum*. CLEF'10. Springer-Verlag, New York, NY, USA, 3–14.

Lu, W.-H., Chien, L.-F., and Lee, H.-J. 2002. Translation of web queries using anchor text mining. *ACM Transactions on Asian Language Information Processing (TALIP) 1,* 2, 159–172.

Lu, W.-H., Chien, L.-F., and Lee, H.-J. 2004. Anchor text mining for translation of web queries: a transitive translation approach. *ACM Trans. Inf. Syst. 22,* 2, 242–269.

Maeda, A., Sadat, F., Yoshikawa, M., and Uemura, S. 2000. Query term disambiguation for web cross-language information retrieval using a search engine. In *Proceedings of the Fifth International Workshop on Information Retrieval with Asian Languages*. ACM Press, Hong Kong, China, 25–32.

Majumder, P., Mitra, M., Parui, S. K., Kole, G., Mitra, P., and Datta, K. 2007. YASS: Yet Another Suffix Stripper. *ACM Transactions on Information Systems (TOIS) 25,* 4, 18:1–3:20.

Manning, C. D., Raghavan, P., and Schtze, H. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Mayfield, J. and McNamee, P. 2004. Triangulation without translation. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, New York, NY, USA, 490–491.

McCarley, J. S. 1999. Should we translate the documents or the queries in cross-language information retrieval? In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*. Association for Computational Linguistics, College Park, Maryland, 208–214.

McEwan, C. J. A., Ounis, I., and Ruthven, I. 2002. Building bilingual dictionaries from parallel web documents. In *Proceedings of the 24th BCS-IRSG European Colloquium on IR Research: Advances in Information Retrieval*. Springer-Verlag, 303–323.

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

McNamee, P. and Mayfield, J. 2004a. Character n-gram tokenization for European language text retrieval. *Inf. Retr. 7,* 1-2, 73–97.

McNamee, P. and Mayfield, J. 2004b. Cross-language retrieval using HAIRCUT at CLEF 2004. In *CLEF 2004: Workshop on Cross-Language Information Retrieval and Evaluation*. Bath, UK.

McNamee, P., Mayfield, J., and Piatko, C. 2002. HAIRCUT: a system for multilingual text retrieval in java. *J. Comput. Small Coll. 17*, 8–22.

Melamed, I. D. 2000. Models of translational equivalence among words. *Comput. Linguist. 26*, 221–249.

Melucci, M. and Orio, N. 2003. A novel method for stemmer generation based on hidden markov models. In *Proceedings of the Twelfth International Conference on Information and Knowledge Management*. ACM Press, New York, NY, USA, 131–138.

Meng, H., Chen, B., Khudanpur, S., Levow, G.-A., Lo, W.-K., Oard, D., Schone, P., Tang, K., Wang, H.-M., and Wang, J. 2001. Mandarin-English information (MEI): investigating translingual speech retrieval. In *Proceedings of the First International Conference on Human Language Technology Research*. Association for Computational Linguistics, San Diego, 1–7.

Meng, H., Khudanpur, S., Levow, G., Oard, D. W., and Wang, H.-M. 2000. Mandarin-English information (MEI): investigating translingual speech retrieval. In *NAACL-ANLP 2000 Workshop on Embedded Machine Translation Systems - Volume 5*. Association for Computational Linguistics, Seattle, Washington, 23–30.

Miller, D. R. H., Leek, T., and Schwartz, R. M. 1999. A hidden markov model information retrieval system. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, New York, NY, USA, 214–221.

Monz, C. and Dorr, B. J. 2005. Iterative translation disambiguation for cross-language information retrieval. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, New York, NY, USA, 520–527.

Moreau, F., Claveau, V., and Sebillot, P. 2007. Automatic morphological query expansion using analogy-based machine learning. In *Proceedings of the 29th European conference on IR Research*. Springer-Verlag, Rome, Italy, 222–233.

Mori, T., Kokubu, T., and Tanaka, T. 2001. Cross-lingual information retrieval based on LSI with multiple word spaces. In *Proceedings of the Second NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and Cross-Lingual Information Access"*. Tokyo, Japan.

Nie, J.-y. 1998. Using a probabilistic translation model for cross-language information retrieval. In *Sixth workshop on Very Large Corpora*. Morgan Kaufmann Publishers, Montreal, Canada.

Nie, J.-y. 2010. *Cross-Language Information Retrieval*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Nie, J.-Y. and Ren, F. 1999. Chinese information retrieval: using characters or words? *Information Processing & Management 35,* 4, 443–462.

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

Oard, D. W. 1998. A comparative study of query and document translation for cross-language information retrieval. In *Proceedings of the Third Conference of the Association for Machine Translation in the Americas on Machine Translation and the Information Soup*. Springer-Verlag, 472–483.

Oard, D. W. 1999. Topic tracking with the PRISE information retrieval system. In *In Proceedings of the DARPA Broadcast News Workshop*. 209–211.

Oard, D. W. and Dorr, B. J. 1996. A survey of multilingual text retrieval. Tech. rep., Univ. of Maryland Institute for Advanced Computer Studies Report No. UMIACS-TR-96-19, University of Maryland at College Park, MD, USA.

Oard, D. W. and Ertunc, F. 2002. Translation-based indexing for cross-language retrieval. In *Proceedings of the 24th BCS-IRSG European Colloquium on IR Research: Advances in Information Retrieval*. Springer-Verlag, London, UK, UK, 324–333.

Oard, D. W. and Hackett, P. 1997. Document translation for cross-langauge text retrieval at the university of Maryland. In *The Sixth Text Retrieval Conference (TREC-6)*. NIST, 687–696.

Oard, D. W., Levow, G.-A., and Cabezas, C. I. 2000. CLEF experiments at Maryland: statistical stemming and backoff translation. In *Proceedings of Evaluation of Cross-Language Information Retrieval Systems: Third Workshop of the Cross-Language Evaluation Forum*.

Oard, D. W. and Wang, J. 2001. NTCIR-2 ECIR experiments at Maryland: comparing pirkola's structured queries and balanced translation. In *Proceedings of the Second NTCIR Workshop on Research in Chinese & Japanese, Text Retrieval and Text Summarization*. National Institute of Informatics, Japan.

Och, F. J. and Ney, H. 2003. A systematic comparison of various statistical alignment models. *Comput. Linguist. 29,* 1, 19–51.

Parton, K., McKeown, K. R., Allan, J., and Henestroza, E. 2008. Simultaneous multilingual search for translingual information retrieval. In *Proceeding of the 17th ACM Conference on Information and Knowledge Management*. ACM, New York, NY, USA, 719–728.

Peters, C. and Picchi, E. 1996. A system for cross-language information retrieval. *ERCIM News* 27.

Peters, C. and Sheridan, P. 2001. Lectures on information retrieval. Springer-Verlag New York, Inc., New York, NY, USA, Chapter Multilingual information access, 51–80.

Pirkola, A. 1998. The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, New York, NY, USA, 55–63.

Pirkola, A., Keskustalo, H., Leppanen, E., Kansala, A.-P., and Jarvelin, K. 2002. Targeted s-gram matching: a novel n-gram matching technique for cross- and monolingual word form variants. *Information Research 7,* 2.

Pirkola, A., Puolamäki, D., and Järvelin, K. 2003. Applying query structuring in cross-language retrieval. *Inf. Process. Manage. 39*, 391–402.

Pirkola, A., Toivonen, J., Keskustalo, H., Visala, K., J, K., and rvelin. 2003. Fuzzy translation of cross-lingual spelling variants. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, Toronto, Canada, 345–352.

Ponte, J. M. and Croft, W. B. 1998. A language modeling approach to information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, New York, NY, USA, 275–281.

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

Potthast, M., Stein, B., and Anderka, M. 2008. A Wikipedia-based multilingual retrieval model. In *Proceedings of 30th European Conference on Information Retrieval*. Springer, Glasgow, Scotland, 522–530.

Qu, Y., Grefenstette, G., and Evans, D. A. 2003. Automatic transliteration for Japanese-to-English text retrieval. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, New York, NY, USA, 353–360.

Resnik, P. 1998. Parallel strands: a preliminary investigation into mining the web for bilingual text. In *Proceedings of the Third Conference of the Association for Machine Translation in the Americas on Machine Translation and the Information Soup*. Springer-Verlag, 72–82.

Resnik, P. 1999. Mining the web for bilingual text. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*. Association for Computational Linguistics, College Park, Maryland, 527–534.

Resnik, P., Oard, D., and Levow, G. 2001. Improved cross-language retrieval using backoff translation. In *Proceedings of the First International Conference on Human Language Technology Research*. HLT '01. Association for Computational Linguistics, San Diego, 1–3.

Resnik, P. and Smith, N. A. 2003. The Web as a parallel corpus. *Comput. Linguist. 29,* 3, 349–380.

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

Rocchio, J. 1971. Relevance feedback in information retrieval. *The SMART Retrieval System: Experiments in Automatic Document Processing*, 313–323.

Ruiz, M., Diekema, A., and Sheridan, P. 1999. CINDOR conceptual interlingua document retrieval: TREC-8 evaluation. In *In Proceedings of the Eighth Text Retrieval Conference (TREC-8)*.

Sakai, T., Kando, N., Lin, C.-J., Mitamura, T., Shima, H., Ji, D., Chen, K.-H., and Nyberg, E. 2008. Overview of the NTCIR-7 ACLIA IR4QA task. In *Proceedings of the Seventh NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and Cross-Lingual Information Access*. Tokyo, Japan, 77–114.

Sakai, T., Shima, H., Kando, N., Song, R., Lin, C.-J., Mitamura, T., Sugimito, M., and Lee, C.-W. 2010. Overview of NTCIR-8 ACLIA IR4QA. In *Proceedings of the Eighth NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and Cross-Lingual Information Access*. Tokyo, Japan, 63–93.

Salton, G. 1971. *The SMART Retrieval System & Experiments in Automatic Document Processing*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

Salton, G., Fox, E. A., and Wu, H. 1983. Extended Boolean information retrieval. *Commun. ACM 26*, 1022–1036.

Salton, G., Wong, A., and Yang, C. S. 1975. A vector space model for automatic indexing. *Commun. ACM 18*, 613–620.

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

Savoy, J. 2005. Comparative study of monolingual and multilingual search models for use with asian languages. *ACM Trans. Asian Lang. Inf. Process. 4,* 2, 163–189.

Savoy, J. 2007. Why do successful search systems fail for some topics. In *Proceedings of the 2007 ACM Symposium on Applied computing*. SAC '07. ACM, New York, NY, USA, 872–877.

Savoy, J. and Dolamic, L. 2009. How effective is google's translation service in search? *Commun. ACM 52*, 139–143.

Schäuble, P. 1993. SPIDER: a multiuser information retrieval system for semistructured and dynamic data. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '93. ACM, New York, NY, USA, 318–327.

Schönhofen, P., Benczúr, A., Bíró, I., and Csalogány, K. 2008. Cross-language retrieval with Wikipedia. In *Advances in Multilingual and Multimodal Information Retrieval, 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers*. Springer, Budapest, Hungary, 72–79.

Shannon, C. E. and Weaver, W. 1963. *A Mathematical Theory of Communication*. University of Illinois Press.

Shaw, J. A. and Fox, E. A. 1994. Combination of multiple searches. In *The Second Text REtrieval Conference (TREC-2*. 243–252.

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

Shi, L. 2010. Mining OOV translations from mixed-language Web pages for cross language information retrieval. In *32nd European Conference on Information Retrieval, ECIR 2010.* Milton Keynes, UK, 471–482.

Shi, L., Nie, J.-Y., and Bai, J. 2007. Comparing different units for query translation in Chinese cross-language information retrieval. In *Proceedings of the 2nd International Conference on Scalable Information Systems*. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), Suzhou, China, 1–9.

Singhal, A. and Pereira, F. 1999. Document expansion for speech retrieval. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '99. ACM, New York, NY, USA, 34–41.

Snajder, J., Basic, B. D., and Tadic, M. 2008. Automatic acquisition of inflectional lexica for morphological normalisation. *Inf. Process. Manage. 44,* 5, 1720–1731.

Song, F. and Croft, W. B. 1999. A general language model for information retrieval. In *Proceedings of the Eighth International Conference on Information and Knowledge Management*. ACM Press, New York, NY, USA, 316–321.

Sorg, P. and Cimiano, P. 2008. Cross-language information retrieval with explicit semantic analysis. In *the Working Notes of the CLEF 2008 Workshop*. Aarhus, Denmark.

Sparck Jones, K. 1988. *A statistical interpretation of term specificity and its application in retrieval*. Taylor Graham Publishing, London, UK, UK, 132–142.

Su, C.-Y., Lin, T.-C., and Wu, S.-H. 2007. Using wikipedia to translate OOV terms on MLIR. In *Proceedings of the Sixth NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and Cross-Lingual Information Access*. Tokyo, Japan, 109–115.

Sun, L., Xue, S., Qu, W., Wang, X., and Sun, Y. 2002. Constructing of a large-scale Chinese-English parallel corpus. In *Proceedings of the 3rd Workshop on Asian Language Resources and International Standardization - Volume 12*. Association for Computational Linguistics, 1–8.

Virga, P. and Khudanpur, S. 2003. Transliteration of proper names in cross-lingual information retrieval. In *Proceedings of the ACL 2003 Workshop on Multilingual and Mixed-Language Named Entity Recognition - Volume 15*. Association for Computational Linguistics, 57–64.

Voorhees, E. M. and Harman, D. 2000. Overview of the ninth text retrieval conference (trec-9). In *In Proceedings of the Ninth Text REtrieval Conference (TREC-9*. 1–14.

Wang, J. and Oard, D. W. 2006. Combining bidirectional translation and synonymy for cross-language information retrieval. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, New York, NY, USA, 202–209.

Wong, S. K. M., Ziarko, W., and Wong, P. C. N. 1985. Generalized vector spaces model in information retrieval. In *Proceedings of the 8th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '85. ACM, New York, NY, USA, 18–25.

Xu, J. and Weischedel, R. 2000. Cross-lingual information retrieval using hidden markov models. In *Proceedings of the 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora: Held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics - Volume 13*. Association for Computational Linguistics, Hong Kong, 95–103.

Xu, J. and Weischedel, R. 2005. Empirical studies on the impact of lexical resources on CLIR performance. *Inf. Process. Manage. 41,* 3, 475–487.

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

Yang, C. C. and Li, K. W. 2002. Mining English/Chinese parallel documents from the World Wide Web. In *Proceedings of the 11th International World Wide Web Conference*. ACM Press, New York, NY, USA, 188–192.

Zhai, C. 2009. *Statistical Language Models for Information Retrieval*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Zhai, C. and Lafferty, J. 2001. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '01. ACM, New York, NY, USA, 334–342.

Zhang, Y., Uchimoto, K., Ma, Q., and Isahara, H. 2005. Building an annotated Japanese-Chinese parallel corpus - a part of NICT multilingual corpora. In *Proceedings of the Tenth Machine Translation Summit MT Summit X*. Phuket, Thailand, 71–78.

Zhang, Y. and Vines, P. 2004. Using the web for automated translation extraction in cross-language information retrieval. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, New York, NY, USA, 162–169.

Zhang, Y., Vines, P., and Zobel, J. 2005. Chinese OOV translation and post-translation query expansion in Chinese-English cross-lingual information retrieval. *ACM Transactions on Asian Language Information Processing (TALIP) 4,* 2, 57–77.

Zhou, D., Truran, M., Brailsford, T., and Ashman, H. 2007. NTCIR-6 experiments using pattern matched translation extraction. In *Proceedings of the Sixth NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and Cross-Lingual Information Access*. Tokyo, Japan, 145–151.

Zhou, D., Truran, M., Brailsford, T., and Ashman, H. 2008. A hybrid technique for English-Chinese cross language information retrieval. *ACM Transactions on Asian Language Information Processing (TALIP) 7*, 5:1–5:35.

Zhou, D., Truran, M., Brailsford, T., Ashman, H., and Goulding, J. 2008. Gcon: A graph-based technique for resolving ambiguity in query translation candidates. In *Proceedings of the 23rd Annual ACM Symposium on Applied Computing*. New York, NY, USA, 1566–1573.

Zhu, J. and Wang, H. 2006. The effect of translation quality in MT-based cross-language information retrieval. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Sydney, Australia, 593–600.

Zobel, J. and Dart, P. 1995. Finding approximate matches in large lexicons. *Software - Practice and Experience 25,* 3, 331–345.

Received October 2010; Revised March 2011; Accepted June 2011