# Enriching Document Representation via Translation for Improved Monolingual Information Retrieval

Seung-Hoon Na
Department of Computer Science
National University of Singapore
nash@comp.nus.edu.sg

Hwee Tou Ng
Department of Computer Science
National University of Singapore
nght@comp.nus.edu.sg

## ABSTRACT

Word ambiguity and vocabulary mismatch are critical problems in information retrieval. To deal with these problems, this paper proposes the use of translated words to enrich document representation, going beyond the words in the original source language to represent a document. In our approach, each original document is *automatically translated* into an *auxiliary language*, and the resulting translated document serves as a semantically enhanced representation for supplementing the original bag of words. The core of our translation representation is the *expected term frequency* of a word in a translated document, which is calculated by averaging the term frequencies over all possible translations, rather than focusing on the 1-best translation only. To achieve better efficiency of translation, we do not rely on full-fledged machine translation, but instead use *monotonic translation* by removing the time-consuming reordering component. Experiments carried out on standard TREC test collections show that our proposed translation representation leads to statistically significant improvements over using only the original language of the document collection.

## Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Retrieval models*; I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Machine translation*

## General Terms

Algorithms, Experimentation, Performance, Theory

## 1. INTRODUCTION

*Words*, or stemmed words, are the most popular index terms used in modern information retrieval (IR) systems due to their high simplicity. However, there are critical challenges that retrieval systems face with the use of words, one of which is *word ambiguity* in a query and a document. The meaning of a word changes with its context. For example, the word *interest* may be used to mean "curiosity," or "a charge for borrowing money," depending on the surrounding context. Another challenge is *vocabulary mismatch* between a query and a document. Since there are many alternative ways to represent the same concept, such as using synonyms or paraphrases, a document may not contain a query word but may still be relevant to the query. For example, *find*, *observe*, or *detect* can possibly mean the same thing as *discover*.

To deal with these problems, many studies in IR have investigated *word sense disambiguation* (WSD) on queries and documents [20, 40, 34, 35, 13, 29, 36, 16], or have performed *query expansion* [15, 23, 43, 9, 10, 42, 2, 26] or *document expansion* [3, 24, 21] by appending semantically related words into the original query or document. Some of these approaches (e.g., pseudo-relevance feedback) have shown marked improvements in retrieval performance. However, most existing works in the literature are basically *monolingual approaches* which are restricted to the use of the original source language of the document collection, without taking advantage of potentially rich semantic information drawn from other languages. Through other languages, various ways of adding semantic information to a document could be available, thereby leading to potentially more improvements than using the original source language only.

Taking a step toward using other languages, we propose the use of *translation representation* by alternatively representing the original document content with the words of an additional language (i.e., an *auxiliary language*), one that is different from the language of a given collection (i.e., the source language). In our approach, each original document is "automatically translated" into an auxiliary language, and the resulting translated document serves as a semantically enhanced representation for supplementing the original bag of words. Specifically, the vocabularies of the source and auxiliary languages are connected in a many-to-many relation via translation, which could bring about important benefits in dealing with the two addressed problems. First, there are multiple candidate words in an auxiliary language that are translated to a source word. Therefore, the word ambiguity problem can be resolved, during the process of choosing a correct translation in a given context of a source word. Conversely, various different source words that refer to similar concepts or senses are translated into only a few words or a single word in an auxiliary language. Thus, the vocabulary mismatch problem in the source language is to


*SIGIR'11,* July 24–28, 2011, Beijing, China.

some extent ameliorated by using translated words in the auxiliary language.

To achieve better efficiency in translation, instead of relying on full-fledged machine translation (MT), we propose a simplified method of constructing the translation representation. Since there is no need for a full translation in our ad-hoc retrieval task, we only estimate the *expected term frequency* of a word in the translated document, by taking the average of the term frequencies of the word over all possible translated documents. Our core assumed setting is *monotonic translation* in which the word order in the source sentence is not changed after translation. This assumption enables us to exclude the time-consuming reordering module from the decoding process. In order to extensively apply monotonic translation into any pair of grammatically dissimilar languages (e.g., English and Chinese), we additionally employ a *distorted language model* for an auxiliary language in which the word order follows the grammatical order of the original source language, not of the auxiliary language. Based on these simplified settings for translation, the translation representation could be more efficiently constructed without relying on full-fledged MT.

Experimental results obtained on standard TREC collections show that the use of the proposed translation representation consistently outperforms baseline retrieval methods that use the collection language only. Our comparison is made against three different baselines – a commonly used baseline, query expansion based on pseudo-relevance feedback, and document expansion based on cluster-based retrieval.

# 2. RELATED WORK

## 2.1 Monolingual Retrieval

Word ambiguity has been extensively investigated in information retrieval using WSD. The initial research efforts on WSD for information retrieval were performed using manual sense annotation [20, 13], on artificially created pseudo-words [34], and on automatic sense disambiguation or clustering [40, 35, 29]. More recent work has scaled up the use of WSD to a large test collection [36] and two medium-size test collections [16], reporting improved retrieval performances using WSD compared to baseline word-based indexes.

Query expansion adds new expansion terms into the original query, which has been one of the most effective approaches to resolve the vocabulary mismatch problem. Expansion terms are selected from hand-crafted thesauri such as WordNet [10], co-occurrence based similarity thesauri [15], highly-ranked retrieved documents (i.e., pseudo-relevance feedback) [23, 43], highly-ranked retrieved passages [2, 26], or external collections such as the Web or Wikipedia [9, 42].

Document expansion has a similar motivation as query expansion, but expansion is applied to documents and not to the query [24, 21]. Expansion terms are selected from the cluster that a document belongs to [24], or from documents most similar to the given document [21].

An interesting work related to ours is Berger and Lafferty [3], which suggested the use of a translation model for the information retrieval task. However, no translation was applied between different languages. Instead, expansion was performed by adding words to a document, or reweighting words, so as to better match a given query. Their translation approach is therefore closer to document expansion.

Most approaches for monolingual retrieval tasks have been restricted to the use of the original collection language only, except for a few recent studies [12, 8]. As is the case with our study, they also utilized multilingual information by using an additional language in order to improve monolingual retrieval. However, they used an *external auxiliary language collection*, which is not automatically translated from the originally given collection. Gao et al. [12] expanded an original document by using similar documents from an external auxiliary language collection. Chinnakotla et al. [8] enriched the original expanded query by using additional expanded queries resulting from applying pseudo-relevance feedback on external auxiliary language collections, showing improved performances on CLEF test collections. Compared to our approach, these two methods are closer to monolingual-based approaches using *external* collections [9, 42] of the same language as the original collection. Unlike these approaches, we do not rely on an external collection but instead automatically create new translated documents, and thus our results are obtained *within* the given test collection only.

Recently, Trieschnigg et al. [39] used a *concept*-based representation in order to enrich the original word-based representation, thereby proposing the translation of the original word language to a *concept language*. While their use of concept language in part has a similar motivation to ours, their translation models are based solely on the use of translation at the lexical level (i.e., word-to-concept), and thus their method is very different from our context-dependent style of translation.

## 2.2 Cross-Lingual Retrieval

Cross-lingual information retrieval (CLIR) addresses the problem of retrieving documents written in a language different from the query language [30]. Even though a common approach in CLIR is to perform query translation (QT) using a bilingual dictionary [32], there were studies showing that combining both QT and document translation (DT) improved retrieval performance in CLIR by using bilingual representations in both the source and target language [28, 19, 7, 4]. McCarley [28] trained a statistical MT system from a parallel corpus, applied it to perform QT and DT, and showed that the combination of scores from QT and DT drastically improved either method alone. Similar results have been reported using either a full-fledged MT system [4] or a simple translation algorithm [7]. Kraaij et al. [19] used the translation model of IBM Model 1 [5], obtained from an automatically constructed parallel corpus from the web, and also reported that the combination of QT and DT improved either method alone.

These hybrid approaches in CLIR resemble ours in that we also use translated words of both queries and documents. The major difference, however, is that our goal is to improve monolingual IR and not CLIR. Furthermore, some of these approaches perform translation without taking into account the surrounding context of a source word [19, 7], while our proposed translation model is context-dependent and thus produces different translated words depending on the context of a source word.

To the best of our knowledge, the work most similar to ours is Franz and McCarley [11], who also applied auto-

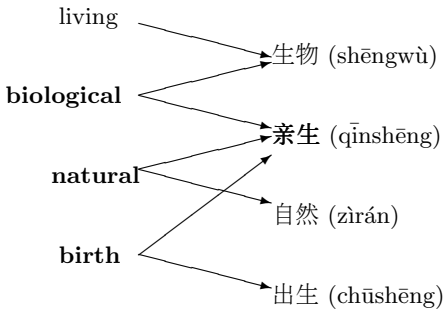


**Figure 1: The word "*biological*" and its paraphrased or synonymous words, and their corresponding translated Chinese words.**

matic translation for monolingual retrieval but using French as the auxiliary language. However, their method did not achieve any statistically significant improvement over a baseline retrieval that used monolingual representation. Moreover, they only considered the 1-best translation, while we use the expected frequency of a word computed from all possible translated representations.

# 3. AN ILLUSTRATING EXAMPLE

To illustrate why translation is helpful in handling the word ambiguity and vocabulary mismatch problems, consider the following TREC query Q335 "*Adoptive **Biological** Parents*", and focus on the ambiguous word "*biological*". The many-to-many translation relations for *biological* and its paraphrased or synonymous words between English and Chinese are depicted in Figure 1.

As shown in Figure 1, *biological* has two different Chinese translations, "生物(shēngwù)" and "亲生(qīnshēng)", which correspond to its two different senses in WordNet, respectively: (1) "pertaining to biology or to life and living things" ("生物": shēngwù), and (2) "of parents and children; related by blood; biological child" ("亲生": qīnshēng). Therefore, word ambiguity in the source language is dealt with during the process of selecting a correct translation between two candidates.

Moreover, the word *biological* in the query context can be equivalently replaced with *natural* or *birth*, (e.g., *natural parents*, *birth parents*) as paraphrased expressions, but the Chinese translation for all of them is only a single word, "亲生(qīnshēng)." This provides a good example to illustrate that the vocabulary mismatch problem in a source language can often be overcome if we use translated words in an auxiliary language.

# 4. OUR APPROACH: OVERVIEW

Our approach of using translation representation for a monolingual retrieval task is summarized in Figure 2. Each document in the source collection is translated using the proposed method of expected frequency estimation, producing bilingual document representations (Section 5). When a new test query is given, the query is translated using the same translation procedure, constituting bilingual query representations (Section 5). Next, initial retrieval on both representations is performed (Sections 6.1 and 6.2), and the two resulting relevance scores are combined to produce a ranked

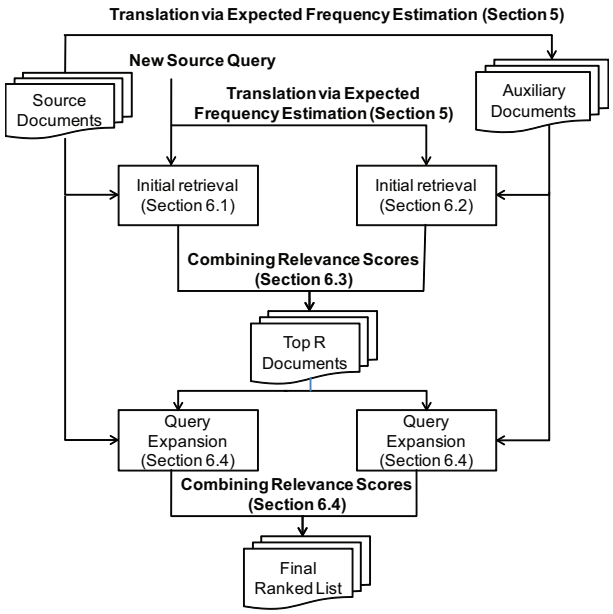


**Figure 2: Overview of our approach of using translation representation.**

list of retrieved documents (Section 6.3). Using the retrieved documents, pseudo-relevance feedback is performed for both representations, and the two resulting relevance scores of documents are again combined (Section 6.4). Furthermore, we also apply document expansion on bilingual representations, as an alternative option for the initial retrieval (Section 6.5).

# 5. CONSTRUCTION OF TRANSLATION REPRESENTATION

## 5.1 Expected Frequency of an Auxiliary Word

In this paper, *source language* refers to the language of a given document collection, and *auxiliary language* refers to an additional language used as the translation representation. Our translation model takes a *phrase* as translation unit, which refers to a contiguous sequence of words, conceptually including a single word. *Auxiliary word*, *auxiliary phrase*, and *auxiliary sentence* refer to a possible translation of a source word, phrase, and sentence, respectively.

Suppose that a source sentence is given by $\mathbf{e} = e_1^I = e_1 \cdots e_I$, where $|\mathbf{e}| = I$ is the length of the source sentence. We do not produce a full translation but instead use the expected frequency of a word in the translated auxiliary sentence for the given source sentence. The expected frequency is defined as follows.

**Definition of expected frequency**: Suppose $\mathbf{F}$ is a random variable, the range of which is $\mathcal{F}$, the set of all sentences of the auxiliary language. Given source sentence $\mathbf{e}$, the *expected frequency* of an auxiliary word $w$ in the translation of $\mathbf{e}$ is the *conditional expectation* of $c(w, \mathbf{F})$ given the event $\mathbf{e}$, as follows:

$$E_{\mathbf{e}}[c(w, \mathbf{F})] = \sum_{\mathbf{f} \in \mathcal{F}} c(w, \mathbf{f}) P(\mathbf{f}|\mathbf{e}) \quad (1)$$

where $c(w, \mathbf{f})$ is the number of word occurrences of $w$ in the auxiliary sentence $\mathbf{f}$, and $P(\mathbf{f}|\mathbf{e})$ is the sentence translation probability of $\mathbf{e}$ translating into $\mathbf{f}$ $\square$.

## 5.2 Word-based Translation Model

Let us first describe a word-based model for computing the expected frequency of Eq. (1). We use simplified monotonic translation allowing only forward translation among possible translations in [38]. The word order is not changed during translation, and a source word at position $i$ (i.e., $e_i$) is translated to an auxiliary word at the same position $i$. In this setting, the translation problem is exactly the same as the tagging problem of a hidden Markov model (HMM), where each state corresponds to an unknown auxiliary word and state translation is modeled by an $n$-gram language model. Thus, the expected frequency of Eq. (1) for the word-based model can be effectively calculated using the EM algorithm, which is based on the forward-backward algorithm of HMM.

## 5.3 Phrase-based Translation Model

We now generalize the word-based model to formulate a phrase-based translation model.

### 5.3.1 Word Lattice

A *word lattice* for a source sentence $\mathbf{e}$ is defined as a connected, directed acyclic graph $\mathcal{G}_\mathbf{e} = (\mathcal{V}_\mathbf{e}, \mathcal{E}_\mathbf{e})$ [27]. Here, $\mathcal{V}_\mathbf{e}$ is the set of vertices $\{0, 1, \cdots, I\}$ consisting of all source word positions, where 0 indicates a specialized *sentence-starting state* #, and $\mathcal{E}_\mathbf{e}$ is the set of edges. Each *edge* $\varepsilon \in \mathcal{E}_\mathbf{e}$ is labeled with a translated auxiliary phrase and it is denoted by $(i, j, \bar{f})$. $i$ and $j$ denote the starting and ending vertices of the source phrase $e_{i+1}^j = e_{i+1} \cdots e_j$, and $\bar{f}$ denotes an auxiliary phrase, that is, a translation of the source phrase $e_{i+1}^j$. The source and auxiliary phrases associated along the edge $\varepsilon$ are referred to as $\bar{e}[\varepsilon]$ and $\bar{f}[\varepsilon]$, respectively. That is, for an edge $\varepsilon = (i, j, \bar{f})$, $\bar{e}[\varepsilon] = e_{i+1}^j$ and $\bar{f}[\varepsilon] = \bar{f}$.

A *translation path* $\boldsymbol{\pi}$ on $\mathcal{G}_\mathbf{e}$ is the sequence of edges $\boldsymbol{\pi} = \pi_1 \cdots \pi_{|\boldsymbol{\pi}|}$, where $|\boldsymbol{\pi}|$ is the number of edges in the path $\boldsymbol{\pi}$. Each edge $\pi_i$ in the path immediately follows the previous edge $\pi_{i-1}$; the starting position of each edge $\pi_i$ is equal to the ending position of the previous edge $\pi_{i-1}$. A path $\boldsymbol{\pi}$ is called a *complete translation path* on $\mathcal{G}_\mathbf{e}$ if $\boldsymbol{\pi}$ translates all source words $e_1 \cdots e_I$, so that the head of $\pi_1$ is 0 and the tail of $\pi_{|\boldsymbol{\pi}|}$ is $I$. The set of all complete translation paths is referred to as $\Phi$. Given a path $\boldsymbol{\pi}$, the sequence of auxiliary phrases along the path is denoted by $\bar{\mathbf{f}}[\boldsymbol{\pi}]$ which means $\bar{f}[\pi_1], \cdots, \bar{f}[\pi_{|\boldsymbol{\pi}|}]$. Similarly, the sequence of source phrases on the path $\boldsymbol{\pi}$ is denoted by $\bar{\mathbf{e}}[\boldsymbol{\pi}]$ which means $\bar{e}[\pi_1], \cdots, \bar{e}[\pi_{|\boldsymbol{\pi}|}]$.

**Word lattice example**: Figure 3 shows an example of a word lattice, where the source language is English and the auxiliary language is German. There are 7 edges consisting of $\varepsilon_1, \cdots, \varepsilon_7$, and their source and auxiliary phrases are as follows: $\bar{e}[\varepsilon_1]$ = "*he*" and $\bar{f}[\varepsilon_1]$ = "*er*", and $\bar{e}[\varepsilon_2]$ = "*he goes to*" and $\bar{f}[\varepsilon_2]$ = "*er geht nach*", and so on. This lattice contains five complete translation paths: $\Phi = \{\boldsymbol{\pi}_1, \cdots, \boldsymbol{\pi}_5\}$ where $\boldsymbol{\pi}_1 = \varepsilon_1\varepsilon_3\varepsilon_5\varepsilon_7$, $\boldsymbol{\pi}_2 = \varepsilon_1\varepsilon_4\varepsilon_5\varepsilon_7$, $\boldsymbol{\pi}_3 = \varepsilon_1\varepsilon_3\varepsilon_6$, $\boldsymbol{\pi}_4 = \varepsilon_1\varepsilon_4\varepsilon_6$, and $\boldsymbol{\pi}_5 = \varepsilon_2\varepsilon_7$. Among these paths, source phrases and auxiliary phrases of $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_3$ are $\bar{\mathbf{e}}[\boldsymbol{\pi}_1]$ = "*he*" "*goes*" "*to*" "*house*" and $\bar{\mathbf{f}}[\boldsymbol{\pi}_1]$ = "*er*" "*geht*" "*auf*" "*hause*", $\bar{\mathbf{e}}[\boldsymbol{\pi}_3]$ = "*he*" "*goes*" "*to house*" and $\bar{\mathbf{f}}[\boldsymbol{\pi}_3]$ = "*er*" "*geht*" "*zu haus*".

### 5.3.2 Path Probability

The *path probability* $p(\boldsymbol{\pi})$ is defined as the joint probability of the source phrases and auxiliary phrases along the path $\boldsymbol{\pi}$, denoted by $p(\bar{\mathbf{e}}[\boldsymbol{\pi}], \bar{\mathbf{f}}[\boldsymbol{\pi}])$. Applying the chain rule decomposes $p(\boldsymbol{\pi})$ into:

$$p(\boldsymbol{\pi}) \triangleq p(\bar{\mathbf{e}}[\boldsymbol{\pi}], \bar{\mathbf{f}}[\boldsymbol{\pi}]) = p(\bar{\mathbf{e}}[\boldsymbol{\pi}]|\bar{\mathbf{f}}[\boldsymbol{\pi}])p(\bar{\mathbf{f}}[\boldsymbol{\pi}]) \quad (2)$$

Note that $p(\boldsymbol{\pi})$ in Eq. (2) consists of two main parts – the *translation part* $p(\bar{\mathbf{e}}[\boldsymbol{\pi}]\,|\bar{\mathbf{f}}[\boldsymbol{\pi}]\,)$, and the *language model part* $p(\bar{\mathbf{f}}[\boldsymbol{\pi}])$. First, the translation part, $p(\bar{\mathbf{e}}[\boldsymbol{\pi}]\,|\bar{\mathbf{f}}[\boldsymbol{\pi}]\,)$ indicates the translation probability from the auxiliary phrases $\bar{\mathbf{f}}[\boldsymbol{\pi}]$ to the source phrases $\bar{\mathbf{e}}[\boldsymbol{\pi}]$, based on the phrase segmentation underlying the path $\boldsymbol{\pi}$. To estimate the translation probability, we make two simplifying assumptions. (1) Monotonic translation: As mentioned in the introduction, this ensures that the order of the auxiliary phrases in the translated sentence is the same as the original order of the source phrases. (2) Independent translation: Each source phrase is separately translated into an auxiliary phrase independent of the other source phrases. Based on these two assumptions, $p(\bar{\mathbf{e}}[\boldsymbol{\pi}]\,|\bar{\mathbf{f}}[\boldsymbol{\pi}]\,)$ is simply decomposed to:

$$p(\bar{\mathbf{e}}[\boldsymbol{\pi}]|\bar{\mathbf{f}}[\boldsymbol{\pi}]) = \prod_{k=1}^{|\boldsymbol{\pi}|} p(\bar{e}[\pi_k]|\bar{f}[\pi_k])$$

Second, the language model part, $p(\bar{\mathbf{f}}[\boldsymbol{\pi}])$, denotes the probability that the auxiliary sentence $\bar{\mathbf{f}}[\boldsymbol{\pi}]$ along the path $\boldsymbol{\pi}$ is generated from the auxiliary language model. We utilize the trigram language model to estimate $p(\bar{\mathbf{f}}[\boldsymbol{\pi}])$, as it is widely used in statistical machine translation [17]. Let the auxiliary sentence of $\boldsymbol{\pi}$ be given by the word sequence $f_1, \cdots, f_J$. Then, $p(\bar{\mathbf{f}}[\boldsymbol{\pi}])$ is decomposed into trigram probabilities as follows:[1]

$$p(\bar{\mathbf{f}}[\boldsymbol{\pi}]) = \prod_{i=1}^{J} p(f_i|f_{i-2}, f_{i-1}) \quad (3)$$

### 5.3.3 Computing Expected Frequency

The part needed for computing the expected frequency defined by Eq. (1) is the sentence translation probability $p(\mathbf{f}|\mathbf{e})$. In this paper, $p(\mathbf{f}|\mathbf{e})$ is defined as the ratio of the probabilities of the paths generating $\mathbf{f}$ in the word lattice $\mathcal{G}_\mathbf{e}$ to the sum of all path probabilities in $\mathcal{G}_\mathbf{e}$, as follows:

$$p(\mathbf{f}|\mathbf{e}) = \frac{\sum_{\boldsymbol{\pi} \in \Phi(\mathbf{f})} p(\boldsymbol{\pi})}{\sum_{\boldsymbol{\pi} \in \Phi} p(\boldsymbol{\pi})}$$

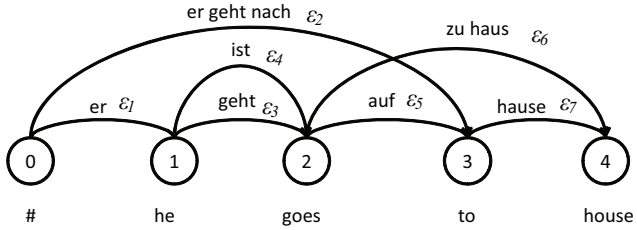


**Figure 3: An example of a word lattice ($I = 4$), where the source language is English and the auxiliary language is German. Here, each $\varepsilon_i$ is an edge, defined as $\varepsilon_1 = (0, 1, \textbf{"}\boldsymbol{er}\textbf{"})$, $\varepsilon_2 = (0, 3, \textbf{"}\boldsymbol{er\ geht\ nach}\textbf{"})$, $\varepsilon_3 = (1, 2, \textbf{"}\boldsymbol{geht}\textbf{'})$, and so on.**

[1] In Eq. (3) , we use the sentence-starting state # to define $f_0$ and $f_{-1}$. For example, $p(f_1|f_{-1}, f_0)$ and $p(f_2|f_0, f_1)$ indicate $p(f_1|\#, \#)$ and $p(f_2|\#, f_1)$, respectively.

where $\Phi(\mathbf{f}) \subseteq \Phi$ denotes the set of the complete translation paths, where each path generates $\mathbf{f}$.

A naive implementation to compute the expected frequency will be extremely inefficient, since the number of all possible translations is exponential with respect to the number of source words. To tractably estimate the expected frequency of Eq. (1), we compute the edge posterior $p(\varepsilon|\mathbf{e})$ for each edge $\varepsilon$ in the lattice $\mathcal{G}_\mathbf{e}$:

$$p(\varepsilon|\mathbf{e}) = \frac{\sum_{\boldsymbol{\pi} \in \Phi(\varepsilon)} p(\boldsymbol{\pi})}{\sum_{\boldsymbol{\pi} \in \Phi} p(\boldsymbol{\pi})}$$

where $\Phi(\varepsilon) \subseteq \Phi$ denotes the subset of complete translation paths through the edge $\varepsilon$. The edge posterior is computed efficiently using a variant of the forward-backward algorithm, as presented in [41].

Finally, $E_\mathbf{e}[c(w, \mathbf{F})]$ is computed in terms of the edge posteriors as follows:

$$E_\mathbf{e}[c(w, \mathbf{F})] = \sum_{\varepsilon \in \mathcal{E}_\mathbf{e}} c(w, \bar{f}[\varepsilon]) p(\varepsilon|\mathbf{e})$$

where $c(w, \bar{f}[\varepsilon])$ is the number of word occurrences of $w$ in the auxiliary phrase $\bar{f}[\varepsilon]$ labeled on the edge $\varepsilon$.

## 5.4 Distorted Language Model

The problem in using our simplified monotonic translation of Section 5.2 and Section 5.3 is that the word order of the trigram $f_{i-2}, f_{i-1}, f_i$ is not consistent with the grammatical order of the auxiliary language. Therefore, our monotonic translation is acceptable only when the auxiliary language is grammatically similar to the source language.

In order to allow monotonic translation for a pair of grammatically dissimilar languages, we use a *distorted language model* for the auxiliary language, in which the auxiliary sentence follows the word order of the source language and not the auxiliary language. To estimate the distorted language model, we constructed a *reordered* auxiliary language corpus, by making the word order of each auxiliary sentence maximally consistent with the word order of the source language.

The details of the reordering procedure are given as follows. Suppose a pair of source and auxiliary sentences is given by $(\mathbf{e}, \mathbf{f})$ (i.e., $(e_1^I, f_1^J)$), a link $(i, j)$ is a word alignment in which $e_i$ is aligned to $f_j$, and $A = \{(i, j)\}$ is the set of word alignments for the given sentence pair. The outcome of the reordering procedure is the *reordered position* for $f_j$, denoted by $b_j$. To determine $b_j$, let $B_j$ be $\{i|(i, j) \in A\}$, the set of all positions of source words that are aligned to $f_j$ in the alignment set $A$. The reordered position $b_j$ is then obtained as follows:

$$b_j = \begin{cases} \max B_j & \text{if } B_j \neq \emptyset \\ b_{a(j)} & \text{otherwise} \end{cases} \quad (4)$$

where $a(j)$ is $\text{argmin}_k \{|j - k| \; | B_k \neq \emptyset\}$.

Once $b_j$ is computed for all auxiliary words, we reorder each auxiliary sentence such that $f_i$ precedes $f_j$ if $b_i < b_j$. To ensure the uniqueness of reordering, when $b_i = b_j$, we make $f_i$ precede $f_j$ for $i < j$.

**Example of reordering**: Figure 4 shows an example of the reordering procedure. Each arrow denotes a word alignment from $e_i$ to $f_j$. For $f_2$, $f_3$, and $f_4$, $b_j$ is simply computed from the set of word alignments $B_j$. For instance, for the auxiliary word $f_4$, $B_4 = \{1, 2\}$, and $b_4$ is 2, according to Eq. (4). For another word $f_1$, however, $B_1$ is the empty set. To handle this case, the definition given by Eq. (4) states that we first find the nearest auxiliary word $f_{a(1)}$ to $f_1$ where $B_{a(1)}$ is not empty, and then use its reordered position $b_{a(1)}$ as $b_1$. In this example, the word $f_{a(1)}$ is $f_2$, resulting in $b_1 = b_2$.

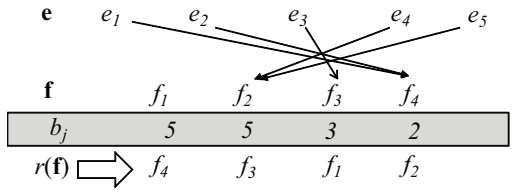


**Figure 4: Example of the reordering procedure. $r(\mathbf{f})$ indicates the reordered auxiliary sentence for the given pair of aligned sentences $(\mathbf{e}, \mathbf{f})$.**

## 5.5 Implementation Detail

We utilized the SRILM toolkit [37] with Kneser-Ney smoothing for estimating the auxiliary language models. Two types of translation probabilities $p(\bar{e}|\bar{f})$ for the word-based and phrase-based models were obtained from the word translation table of GIZA++ [31] and the phrase translation table of Moses [18], respectively. We used the word alignment output of GIZA++ to construct the reordered auxiliary language corpus described in Section 5.4. For the word-based model, we allowed null translation, and thus utilized a slightly modified version of the forward-backward procedure. For the phrase-based model, the maximum length of source phrases was fixed at 7. We only selected the top $M$ auxiliary words (or phrases) for constructing the bilingual dictionaries obtained from GIZA++ and Moses for each source word (or phrase), ranked by the translation probability $P(\bar{f}|\bar{e})$.

We removed the translation candidates with very low probabilities at each word position $j$. To achieve this, we introduced $G(w, j) = \sum_{i<j} Q(w, i, j) / \sum_{\boldsymbol{\pi} \in \Phi} p(\boldsymbol{\pi})$, which is the sum of the frequencies of word $w$, over all paths ending at position $j$, where $Q(w, i, j)$ is defined as follows:

$$Q(w, i, j) = \sum_{\substack{\varepsilon \in \mathcal{E}_\mathbf{e}: \\ \bar{e}[\varepsilon] = e_{i+1}^j}} \sum_{\boldsymbol{\pi} \in \Phi(\varepsilon)} c(w, \bar{f}[\varepsilon]) p(\boldsymbol{\pi}) \quad (5)$$

which conceptually corresponds to the partial expected count of $w$ in the case where the source phrase is restricted to $e_{i+1}^j$. We then applied the cut-off threshold $\theta$ to $G(w, j)$, below which it is excluded (i.e., $G(w, j) < \theta$) when computing the expected frequency of $w$. In addition, we kept only at most the top $T$ values of $G(w, j)$ at each position $j$. In this paper, $M$, $\theta$, and $T$ were fixed at 10, 0.001, and 10, respectively.

# 6. MULTILINGUAL RETRIEVAL METHOD

## 6.1 Retrieval Model

We use the language modeling approach for the retrieval method. The language modeling approach ranks documents according to the likelihood that a query is generated from the document language model [33, 14], or more generally, the negative KL divergence between the query model $P(w|\mathbf{q})$

and document model $P(w|\mathbf{d})$ [22]:

$$Score(\mathbf{q}, \mathbf{d}) = \sum_{w} p(w|\mathbf{q}) log \frac{p(w|\mathbf{d})}{p(w|\mathbf{q})} \quad (6)$$

where $\mathbf{q}$ and $\mathbf{d}$ represent a query and a document, respectively.

We adopt Dirichlet-prior smoothing to estimate $p(w|\mathbf{d})$ as follows [43]:

$$p(w|\mathbf{d}) = \frac{c(w, \mathbf{d}) + \mu p(w|\mathcal{C})}{|\mathbf{d}| + \mu} \quad (7)$$

where $c(w, \mathbf{d})$ is the term frequency of $w$ in document $\mathbf{d}$, $|\mathbf{d}|$ is the total number of words in $\mathbf{d}$, $p(w|\mathcal{C})$ is the background collection model, and $\mu$ is a smoothing parameter. The query model $p(w|\mathbf{q})$ is estimated by using MLE: $c(w, \mathbf{q})/|\mathbf{q}|$.

## 6.2 Retrieval Model for Translation Representation

We now extend the retrieval model described in Section 6.1 in order to support translation representation with expected frequencies. Let $\tau$ be the translation operator, $\tau(\mathbf{d})$ the translated representation resulting from applying the translation operator $\tau$ for a given document $\mathbf{d}$, and $Sent(\mathbf{d})$ the set of sentences in $\mathbf{d}$ obtained by sentence segmentation. The term frequency of word $w$ in $\tau(\mathbf{d})$, denoted by $c(w, \tau(\mathbf{d}))$, is defined as follows:

$$c(w, \tau(\mathbf{d})) = \sum_{\mathbf{e} \in Sent(\mathbf{d})} E_{\mathbf{e}}[c(w, \mathbf{F})]$$

The length of $\tau(\mathbf{d})$, denoted by $|\tau(\mathbf{d})|$, is the expected length of the translated document, which is defined by $\sum_{w \in \mathcal{V}_F} c(w, \tau(\mathbf{d}))$, where $\mathcal{V}_F$ is the vocabulary of the auxiliary language.

By replacing $c(w, \mathbf{d})$ and $|\mathbf{d}|$ in Eq. (7) with $c(w, \tau(\mathbf{d}))$ and $|\tau(\mathbf{d})|$, we obtain the following smoothed model for $\tau(\mathbf{d})$:

$$p(w|\tau(\mathbf{d})) = \frac{c(w, \tau(\mathbf{d})) + \mu p(w|\tau(\mathcal{C}))}{|\tau(\mathbf{d})| + \mu} \quad (8)$$

where $w$ is a word in $\mathcal{V}_F$, and $p(w|\tau(\mathcal{C}))$ is defined by

$$p(w|\tau(\mathcal{C})) = \frac{\sum_{\mathbf{d} \in \mathcal{C}} c(w, \tau(\mathbf{d}))}{\sum_{w \in \mathcal{V}_F} \sum_{\mathbf{d} \in \mathcal{C}} c(w, \tau(\mathbf{d}))} \quad (9)$$

where $\mathcal{C}$ refers to the set of documents in a given collection. Similar to $p(w|\tau(\mathbf{d}))$, we estimate the query model $p(w|\tau(\mathbf{q}))$ for the translation representation of the query $\mathbf{q}$ by using $c(w, \tau(\mathbf{q}))/|\tau(\mathbf{q})|$. Finally, we calculate the relevance score of document $\mathbf{d}$ with respect to query $\mathbf{q}$ using $Score(\tau(\mathbf{q}), \tau(\mathbf{d}))$ based on their translation representations.

## 6.3 Combining Relevance Scores for Multilingual Representation

To produce a single ranked list from two relevance scores using Eq. (6), that is, $Score(\mathbf{q}, \mathbf{d})$ on the source language and $Score(\tau(\mathbf{q}), \tau(\mathbf{d}))$ on the auxiliary language, we use the following linear combination:

$$Score_{E+F}(\mathbf{q}, \mathbf{d}) = \alpha \; Score(\mathbf{q}, \mathbf{d}) + (1 - \alpha) Score(\tau(\mathbf{q}), \tau(\mathbf{d})) \quad (10)$$

where $\alpha$ is an interpolation parameter.

## 6.4 Query Expansion for Multilingual Representation

We further apply query expansion for multilingual representations. We choose pseudo-relevance feedback, because it is one of the most effective query expansion approaches. The procedure is as follows:

1. Obtain an initial set of top $R$ retrieved documents by applying $Score_{E+F}(\mathbf{q}, \mathbf{d})$.
2. Create expanded queries $\mathbf{q}'$ and $\tau(\mathbf{q}')$ by adding expansion terms from the top $R$ documents of the source language and auxiliary language collections, respectively.
3. Re-score documents by $Score(\mathbf{q}', \mathbf{d})$ and $Score(\tau(\mathbf{q}'), \tau(\mathbf{d}))$ based on expanded queries $\mathbf{q}'$ and $\tau(\mathbf{q}')$ using Eq. (6) for both representations, respectively.
4. Combine the two relevance scores by using the linear interpolation $Score_{E+F}(\mathbf{q}', \mathbf{d})$, to obtain the final ranked list of documents as the outcome.

For pseudo-relevance feedback in step 2, we adopt RM3, a variant of the relevance language models of [23], which is one of the most effective and robust pseudo-relevance feedback methods in the language modeling framework [25]. To be more specific, suppose $\mathcal{D}_{init}$ is the set of the initially retrieved documents. Then, the expanded query model $P(w|\mathbf{q}')$ used by RM3 is estimated based on the following formula [25]:

$$P(w|\mathbf{q}') = \beta P(w|\mathbf{q}) + (1 - \beta) \sum_{\mathbf{d} \in \mathcal{D}_{init}} P(w|\mathbf{d}) P(\mathbf{d}|\mathbf{q}) \quad (11)$$

where $\beta$ is an interpolation parameter for combining an original query with an expanded query, and $P(\mathbf{d}|\mathbf{q})$ is the posterior probability of $\mathbf{d}$, conditioned on having observed $\mathbf{q}$. The posterior probability $P(\mathbf{d}|\mathbf{q})$ can be rewritten in terms of $Score(\mathbf{q}, \mathbf{d})$ as follows:

$$P(\mathbf{d}|\mathbf{q}) = \frac{\exp(Score(\mathbf{q}, \mathbf{d}))}{\sum_{\mathbf{d}' \in \mathcal{D}_{init}} \exp(Score(\mathbf{q}, \mathbf{d}'))} \quad (12)$$

For re-scoring the documents in step 3, we re-apply Dirichlet-prior smoothing as described in Eq. (7).

## 6.5 Document Expansion for Multilingual Representation

*Document expansion* (or cluster-based retrieval) can also be applied to multilingual representation, where the representation of each document is enriched with a set of similar documents called a cluster [24, 21]. Suppose cluster $Clu_{\mathbf{d}}$ is a set of documents similar to $\mathbf{d}$, and $\mathbf{d}_{clu}$ is the cluster document representation of $\mathbf{d}$. Then, *cluster-term frequency* of $\mathbf{d}_{clu}$ for word $w$ is defined as follows:

$$c(w, \mathbf{d}_{clu}) = \sum_{\mathbf{d}' \in Clu_{\mathbf{d}}} c(w, \mathbf{d}')$$

We now additionally introduce $\mathbf{d}'_{clu}$ to indicate the *cluster-enhanced document representation* of $\mathbf{d}$, which is the weighted representation of the original source document $\mathbf{d}$ and its cluster $\mathbf{d}_{clu}$ as follows:

$$c(w, \mathbf{d}'_{clu}) = c(w, \mathbf{d}) + \lambda_{clu} \cdot c(w, \mathbf{d}_{clu})$$

where $\lambda_{clu}$ is the weight of document representation to cluster representation. To estimate the smoothed cluster language model, we adopt two-stage smoothing [24]: (1) The cluster-based model is first smoothed with the background

**Table 1: Statistics for each test collection.**

| Statistic | ROBUST | WT2G | WT10G |
|---|---|---|---|
| *NumDocs* | 528,156 | 247,491 | 1,692,096 |
| *NumWords* | 572,180 | 1,407,283 | 6,346,858 |
| *TopicSet* | Q301–450 Q601–700 | Q401–450 | Q451–550 |

collection model. (2) The original document model is further smoothed by the smoothed cluster model. Starting from the basic form given by Eq. (7), two-stage smoothing is derived by replacing term frequencies $c(w, \mathbf{d})$ by the cluster-enhanced frequencies $c(w, \mathbf{d}'_{clu})$ as follows:

$$p(w|\mathbf{d}'_{clu}) = \frac{c(w, \mathbf{d}'_{clu}) + \mu p(w|\mathcal{C})}{|\mathbf{d}'_{clu}| + \mu} = \frac{c(w, \mathbf{d}) + \lambda_{clu} c(w, \mathbf{d}_{clu}) + \mu p(w|\mathcal{C})}{|\mathbf{d}| + \lambda_{clu}|\mathbf{d}_{clu}| + \mu} \quad (13)$$

where $|\mathbf{d}_{clu}|$ and $|\mathbf{d}'_{clu}|$ are the length of $\mathbf{d}_{clu}$ and $\mathbf{d}'_{clu}$, respectively.

Similarly, we could define the translation representation $\tau(\mathbf{d}'_{clu})$ for the cluster-enhanced document $\mathbf{d}'_{clu}$ by setting the term frequency $c(w, \tau(\mathbf{d}'_{clu}))$ of word $w$ as follows:

$$c(w, \tau(\mathbf{d}'_{clu})) = c(w, \tau(\mathbf{d})) + \lambda_{clu} \cdot c(w, \tau(\mathbf{d}_{clu}))$$

with the following definition of $c(w, \tau(\mathbf{d}_{clu}))$:

$$c(w, \tau(\mathbf{d}_{clu})) = \sum_{\mathbf{d}' \in Clu_{\mathbf{d}}} c(w, \tau(\mathbf{d}'))$$

Given bilingual cluster-enhanced representations, we again use $Score_{E+F}(\mathbf{q}, \mathbf{d}'_{clu})$ for combining the two relevance scores obtained from the bilingual representations.

To define $Clu_{\mathbf{d}}$, we use the method suggested by [21], where $Clu_{\mathbf{d}}$ is a large virtual document comprising a concatenation of the $k$ most similar documents to $\mathbf{d}$ ($\mathbf{d}$ itself can be included among the $k$ documents). In this paper, $k$ is fixed to 50. We apply Eq. (6) to find similar documents, with $\mu$ fixed to 1,500 and by taking the document as a query. Note that finding similar documents is only based on the source language, and thus $Clu_{\mathbf{d}}$ is shared by both the source and auxiliary language.

# 7. EXPERIMENTAL SETTING

For evaluation, we used three different standard TREC collections – ROBUST, WT2G, and WT10G. Table 1 shows the basic statistics for each test collection, where *NumDocs* is the number of documents, *NumWords* is the total number of word occurrences in each collection, and *TopicSet* is the range of topic numbers used for training and testing.

All experiments were based on the Lemur toolkit (version 4.12)[2]. When indexing English documents, we performed standard preprocessing on queries and documents by applying Porter's stemmer and removing stopwords using the standard INQUERY stoplist [1]. We used only the words in the "title field" of a query topic for all our evaluations. For translating English documents and queries, we used the Penn Treebank tokenizer to preprocess them[3].

MAP (mean average precision) was used as the evaluation measure. For each query, our evaluation was based on the top 1,000 retrieved documents. We reported statistical significance using paired $t$-test at 0.95 confidence level.

There are several parameters for each retrieval method: $\mu$, $\lambda_{clu}$, $\alpha$, and $\beta$. For ROBUST and WT10G, given a test set consisting of 50 queries, each parameter was selected by tuning on the other queries in the same test collection. For example, parameters for Q301–350 in ROBUST were tuned using Q351–450 and Q601–700 in the same ROBUST collection. For WT2G, we applied 5-fold cross validation, by dividing the 50 queries into 5 folds consisting of 10 queries each. For cluster-based retrieval, instead of directly tuning $\mu$, we tuned $\mu'$ by setting $\mu = \mu'(\lambda_{clu} + 1)$, so that the final value for $\mu'$ is more similar to the value of $\mu$ used in the baseline retrieval.

For preparing the translation representation, we considered Chinese as the auxiliary language, and applied the proposed translation models from English to Chinese over the collection. To obtain the translation probabilities $p(\bar{e}|\bar{f})$, we used a subset of the parallel corpora used in [6], containing approximately 2.5M sentence pairs, 72M English tokens, and 65M Chinese characters. We removed long sentences containing more than 40 tokens when applying GIZA++ and Moses. Using the parallel corpus, we created two types of Chinese translation representations:

- Word: using the proposed word-based translation model (Section 5.2),
- Phrase: using the proposed phrase-based translation model (Section 5.3).

# 8. EXPERIMENTAL RESULTS

The major goal of our experiments is to examine whether the use of translation-enhanced document representation leads to improvements in retrieval performance, compared to using only the given collection language. Comparisons are made on three retrieval methods using the following language models:

- **LM** [43]: The commonly used baseline described in Section 6.1, which uses Dirichlet-prior smoothing described by Eq. (7) for computing $Score(\mathbf{q}, \mathbf{d})$;
- **RM3** [25]: The query expansion method described in Section 6.4, which is based on pseudo-relevance feedback RM3;
- **CLM** [21]: The document expansion method described in Section 6.5, which uses the cluster-based language model $P(w|\mathbf{d}'_{clu})$ described in Eq. (13) for the document model $P(w|\mathbf{d})$ of Eq. (6).

Throughout this section, we refer to the original source representation by E, and refer to the Chinese translation representation obtained from Word and Phrase by $\mathsf{C}_{\mathsf{Word}}$ and $\mathsf{C}_{\mathsf{Phrase}}$, respectively.

## 8.1 Results

Table 2 shows a comparison of the results obtained using monolingual and bilingual representations on the setting of LM without query expansion and document expansion across three different collections. In Table 2, E denotes the baseline LM performed using Eq. (7), all of which used only the English queries and documents; $\mathsf{C}_\mathsf{X}$ denotes the run of LM using Chinese translation representation, where X could be Word or Phrase; and $\mathsf{E}+\mathsf{C}_\mathsf{X}$ denotes the run of LM with the combination of monolingual and bilingual representation E and $\mathsf{C}_\mathsf{X}$.

[2]http://www.lemurproject.org

[3]http://www.cis.upenn.edu/∼treebank/tokenizer.sed

**Table 2: Comparison of bilingual representation with monolingual representation on the setting of LM. The mark * indicates statistical significance over E.**

| | ROBUST | WT2G | WT10G |
|---|---|---|---|
| E | 0.2410 | 0.3067 | 0.1963 |
| $C_{Word}$ | 0.2454 | 0.3021 | 0.1871 |
| E+$C_{Word}$ | 0.2591* | 0.3149 | 0.2036* |
| $C_{Phrase}$ | 0.2448 | 0.3164 | 0.1833 |
| E+$C_{Phrase}$ | **0.2684*** | **0.3294*** | **0.2054*** |

**Table 3: Comparison of bilingual representation with monolingual representation on the setting of RM3. The symbols * and + indicate statistical significance over two baselines LM and RM3 using only English representation, respectively.**

| $R$ | | ROBUST | WT2G | WT10G |
|---|---|---|---|---|
| *5* | E | 0.2788* | 0.3322* | 0.2124* |
| | E+$C_{Word}$ | 0.2910*+ | 0.3457*+ | 0.2252*+ |
| | E+$C_{Phrase}$ | **0.2998***+ | **0.3586***+ | **0.2294***+ |
| *10* | E | 0.2750* | 0.3314* | 0.2150* |
| | E+$C_{Word}$ | 0.2896*+ | 0.3493*+ | 0.2209* |
| | E+$C_{Phrase}$ | **0.3012***+ | **0.3594***+ | **0.2287***+ |
| *15* | E | 0.2794* | 0.3441* | 0.2161* |
| | E+$C_{Word}$ | 0.2899*+ | 0.3556*+ | 0.2194* |
| | E+$C_{Phrase}$ | **0.2982***+ | **0.3643*** | **0.2247*** |
| *20* | E | 0.2780* | 0.3328* | 0.2063* |
| | E+$C_{Word}$ | 0.2903*+ | 0.3526*+ | 0.2168*+ |
| | E+$C_{Phrase}$ | **0.2972***+ | **0.3669***+ | **0.2208***+ |
| *30* | E | 0.2735* | 0.3270* | 0.2043 |
| | E+$C_{Word}$ | 0.2843*+ | 0.3485*+ | 0.2126*+ |
| | E+$C_{Phrase}$ | **0.2934***+ | **0.3598***+ | **0.2206***+ |

Our translation models (E+$C_{Word}$ and E+$C_{Phrase}$) significantly improve the baseline E for most test collections. Comparing the two translation types Word and Phrase, we see that the phrase-based model (E+$C_{Phrase}$) gives the best results in combination for all test collections, and its improvements over E are statistically significant for all three test collections, especially achieving an increase of more than 2.5% in MAP on the ROBUST test collection. Interestingly, even our models using only the auxiliary language ($C_{Phrase}$ and $C_{Word}$) often show better performances than E in ROBUST and WT2G for $C_{Phrase}$, and in ROBUST for $C_{Word}$.

## 8.2 Results with Query Expansion

Table 3 shows the comparison results of monolingual and bilingual representation on the setting of RM3 described in Section 6.4. In Table 3, E denotes the baseline RM3 using only the original English representation, and E+$C_X$ denotes the run of RM3 based on the bilingual representation of E and $C_X$. The number of expanded terms for pseudo-relevance feedback in Section 6.4 was fixed to a maximum of 100. For the original English representation, the smoothing parameter $\mu$ used for computing Eq. (11) was fixed to 1,500. For Chinese translation representation, different values of $\mu$ were used for $P(w|\mathbf{d})$ and $P(\mathbf{d}|\mathbf{q})$ of Eq. (11), by fixing to 0 and 1,500, respectively.[4] The smoothing parameter $\mu$ at the second retrieval was fixed to 1,500 for both representations.

Applying RM3 alone without bilingual representation significantly improves the baseline E, which is also known from previous results on RM3 [25]. Importantly, further improvements over RM3 are obtained by utilizing the translation words, in both word and phrase translation types, and these improvements are statistically significant especially on ROBUST and often on WT2G and WT10G. Comparing the two translation types, the phrase-based model (E+$C_{Phrase}$) gives better retrieval performances than the word-based model (E+$C_{Word}$).

**Table 4: Comparison of bilingual representation with monolingual representation on the setting of CLM. The symbols * and + indicate statistical significance over two baselines LM and CLM using only English representation, respectively.**

| | ROBUST | WT2G | WT10G |
|---|---|---|---|
| E | 0.2699* | 0.3091 | 0.2007 |
| $C_{Word}$ | 0.2700* | 0.3034 | 0.1866 |
| E+$C_{Word}$ | 0.2783*+ | 0.3225*+ | 0.2068 |
| $C_{Phrase}$ | 0.2743* | 0.3166 | 0.1865 |
| E+$C_{Phrase}$ | **0.2909***+ | **0.3321***+ | **0.2119***+ |

## 8.3 Results with Document Expansion

Table 4 shows the comparison results of monolingual and bilingual representation on the setting of CLM using the document expansion method of Section 6.5. In Table 4, E denotes the baseline CLM using the original English representation only, and E+$C_X$ denotes the run of CLM based on the bilingual representation of E and $C_X$.

Document expansion without translation representation (i.e., E in Table 4) is highly effective on the ROBUST collection, achieving more than 2.5% MAP increase over the baseline LM (i.e., E in Table 2), with statistical significance. However, its improvements on the other web test collections of WT2G and WT10G are insignificant. Additionally using the Chinese translation representation (E+$C_X$) achieves further improvements over CLM. Specifically, our phrase-based model (E+$C_{Phrase}$) achieves about 2% further increase of MAP over the baseline CLM (E) on ROBUST, finally leading to a noticeable increase of 4.5% MAP over the baseline LM. Even on the web collections of WT2G and WT10G, which are not improved by CLM, our phrase-based model (E+$C_{Phrase}$) leads to statistically significant improvements, achieving about 2% increase in MAP on WT2G and about 1.5% increase in MAP on WT10G.

## 8.4 Parameter Sensitivity of Combination

Figure 5 shows the curves of retrieval performances using our translation-enriched representations (E+$C_{Word}$ and E+$C_{Phrase}$) with respect to the parameter $\alpha$ used in Eq. (10), across LM, CLM, and RM3. In Figure 5, we chose 1,500 for the smoothing parameter $\mu$, fixed $R$ to 10 for RM3, and used the best values for the other additional parameters for each test collection ($\lambda_{clu}$ in CLM and the interpolation pa-

[4]Without this heuristic modification, since we do not remove any stopword for Chinese translation representation, the original RM3 of Eq. (11) gives common words unnecessarily high probabilities.

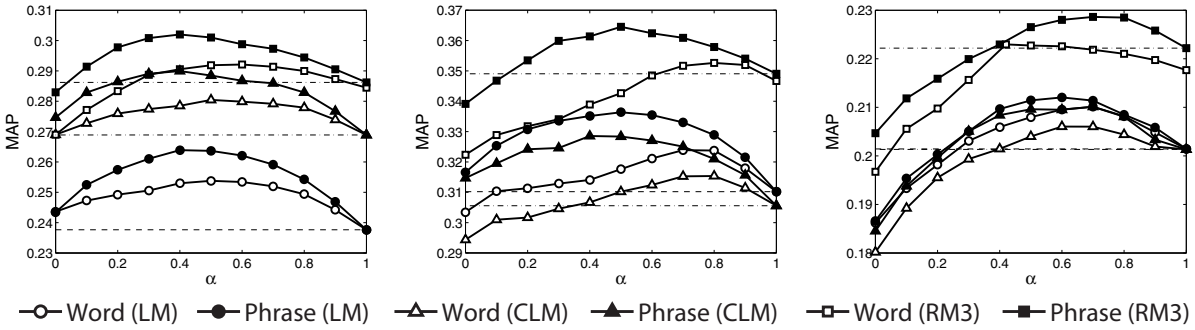


**Figure 5: Performance curves of bilingual-based retrieval by using $\mu = 1,500$ (but using $\mu = 1,500 \times (\lambda_{clu} + 1)$ for CLM), varying $\alpha$, on ROBUST (left), WT2G (middle), and WT10G (right). Performances of two single monolingual representations $\mathsf{E}$ and $\mathsf{C_X}$ are plotted at the points of $\alpha$=1 and $\alpha$=0, respectively.**

**Table 5: Comparison of our translation models ($\mathsf{C_{Word}}$ and $\mathsf{C_{Phrase}}$) to full-fledged MT ($\mathsf{C_{MT}}$) on the setting of LM, CLM, and RM3 on the ROBUST test collection. The symbols * and + indicate statistical significance over the baseline LM and the expansion-based baseline (CLM or RM3) using only English representation, respectively.**

| | $\mathsf{E}$+$\mathsf{C_{MT}}$ | $\mathsf{E}$+$\mathsf{C_{Word}}$ | $\mathsf{E}$+$\mathsf{C_{Phrase}}$ |
|---|---|---|---|
| *LM* | 0.2564* | 0.2591* | **0.2684*** |
| *CLM* | 0.2814*+ | 0.2783*+ | **0.2909***+ |
| *RM3(R=5)* | 0.2964*+ | 0.2910*+ | **0.2998***+ |
| *RM3(R=10)* | 0.2952*+ | 0.2896*+ | **0.3012***+ |
| *RM3(R=15)* | 0.2956*+ | 0.2899*+ | **0.2982***+ |
| *RM3(R=20)* | 0.2939*+ | 0.2903*+ | **0.2972***+ |
| *RM3(R=30)* | 0.2896*+ | 0.2843*+ | **0.2934***+ |

rameter between the original query and in RM3, etc.). Two single monolingual representations $\mathsf{E}$ and $\mathsf{C_X}$ correspond to the case of $\alpha$=1 and $\alpha$=0, respectively.

We see that the best value of $\alpha$ depends on the performance difference between $\mathsf{E}$ and $\mathsf{C_X}$ for each retrieval method. The best value of $\alpha$ is larger when $\mathsf{E}$ produces a better performance than $\mathsf{C_X}$. Despite the variations across different retrieval methods, the common range of the best $\alpha$ is between 0.4 and 0.8. In particular, the phrase-based model ($\mathsf{E}$+$\mathsf{C_{Phrase}}$) achieves in most cases the best improvements (at least the significant improvements) over the baseline ($\mathsf{E}$) when $\alpha$ is around 0.4–0.6.

## 8.5 Comparison with Full-Fledged MT System

We now evaluate how different the results of our translation models are, compared to the results from a full-fledged MT system. To build a full-fledged MT system, we used Moses on the same parallel corpus in Section 7 based on the default feature weights without any development data set. For our evaluation, since applying our MT system to TREC collections requires substantial time, we only considered the ROBUST collection. We used the 1-best translation generated by Moses.

Table 5 shows comparison results of three bilingual representations on the setting of the bilingual-based retrieval ($\mathsf{E}$+$\mathsf{C_X}$) for LM, CLM, and RM3 on the ROBUST test collection. In Table 5, Chinese translation from the full-fledged MT system is referred to by $\mathsf{C_{MT}}$. Full-fledged MT shows almost similar performances as the word-based model ($\mathsf{E}$+$\mathsf{C_{Word}}$) for all three retrieval methods of LM, CLM, and RM3. The phrase-based model ($\mathsf{E}$+$\mathsf{C_{Phrase}}$) achieves slightly better performances than the full-fledged MT model ($\mathsf{E}$+$\mathsf{C_{MT}}$) on the settings of LM and CLM. The general tendency is that as some expansion method (i.e., query or document expansion) is performed, the full-fledged MT model shows closer performance to that of the phrase-based translation models. This is because $\mathsf{MT}$ adopts the 1-best translation, in contrast to our translation models exploiting the N-best translations in calculating expected frequencies. That is, our models internally have the default expansion effect, whereas the full-fledged MT model does not, before performing query or document expansion. Without combining with the original representation, we also applied MT in the setting of the non-combined model ($\mathsf{C_X}$) by setting $\alpha = 0$ in Eq. (10), and observed that the full-fledged MT model only achieved MAP of 0.1905 and 0.2305 in the case of LM and CLM, respectively. The performance is worse than that of our translation models in Table 2 and Table 4.

As a result, our proposed translation models perform at least as well as or better than the full-fledged MT system. However, there is one issue to be handled in estimating the distorted language model: the corpus used for learning the distorted language model is currently limited to only the auxiliary part of the available parallel corpus, and thus its size is far smaller than that for traditional MT. Resolving the corpus size limitation for estimating the distorted language model would be a subject worthy of further investigation.

# 9. CONCLUSION AND FUTURE WORK

In this paper, we proposed the use of translation representation, encouraged by the fact that a translated word in an auxiliary language can be taken in a disambiguated sense, or can act as a concept to capture various different expressions in the source language. We used a simplified translation model with monotonic translation to automatically translate all documents in the test collection, producing multilingual representations. Then, the relevance score of a document was calculated by combining multiple evidences derived from multilingual representations. Experimental results on standard TREC English test collections showed that by using English-to-Chinese translation, our approach achieves im-

provements over baseline monolingual retrieval, and the improvements are in many cases statistically significant.

For future work, we would like to extend the current experiments by considering other Western languages, for example, English-French, English-German, etc. We want to see how strongly the linguistic diversity between source and auxiliary languages affects retrieval performance.

**Acknowledgments** This research was done for CSIDM Project No. CSIDM-200804 partially funded by a grant from the National Research Foundation (NRF) administered by the Media Development Authority (MDA) of Singapore. We would like to thank the anonymous reviewers for their valuable comments.

# 10. REFERENCES

[1] J. Allan, M. E. Connell, W. B. Croft, F. Feng, D. Fisher, and X. Li. INQUERY and TREC-9. In *TREC-9*, 2000.

[2] M. Bendersky and O. Kurland. Utilizing passage-based language models for ad hoc document retrieval. *Information Retrieval*, 13, 2010.

[3] A. Berger and J. Lafferty. Information retrieval as statistical translation. In *SIGIR '99*, 1999.

[4] M. Braschler. Combination approaches for multilingual text retrieval. *Information Retrieval*, 7(1-2), 2004.

[5] P. F. Brown, V. J. D. Pietra, S. A. D. Pietra, and R. L. Mercer. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19, 1993.

[6] Y. S. Chan and H. T. Ng. Scaling up word sense disambiguation via parallel texts. In *AAAI'05*, 2005.

[7] A. Chen and F. C. Gey. Multilingual information retrieval using machine translation, relevance feedback and decompounding. *Information Retrieval*, 7(1-2), 2004.

[8] M. K. Chinnakotla, K. Raman, and P. Bhattacharyya. Multilingual PRF: English lends a helping hand. In *SIGIR '10*, 2010.

[9] F. Diaz and D. Metzler. Improving the estimation of relevance models using large external corpora. In *SIGIR '06*, 2006.

[10] H. Fang. A re-examination of query expansion using lexical resources. In *ACL-08: HLT*, 2008.

[11] M. Franz and J. S. McCarley. Machine translation and monolingual information retrieval. In *SIGIR '99*, 1999.

[12] W. Gao, J. Blitzer, M. Zhou, and K.-F. Wong. Exploiting bilingual information to improve web search. In *ACL-IJCNLP '09*, 2009.

[13] J. Gonzalo, F. Verdejo, I. Chugur, and J. Cigarran. Indexing with WordNet synsets can improve text retrieval. In *Proceedings of COLING-ACL '98 Workshop on Usage of WordNet in Natural Language Processing Systems*, 1998.

[14] D. Hiemstra. A linguistically motivated probabilistic model of information retrieval. In *ECDL '98*, 1998.

[15] Y. Jing and W. B. Croft. An association thesaurus for information retrieval. In *RIAO '94*, 1994.

[16] S.-B. Kim, H.-C. Seo, and H.-C. Rim. Information retrieval using word senses: root sense tagging approach. In *SIGIR '04*, 2004.

[17] P. Koehn. *Statistical Machine Translation*. Cambridge University Press, 2010.

[18] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. Moses: Open source toolkit for statistical machine translation. In *ACL '07. Demonstration Session*, 2007.

[19] W. Kraaij, J.-Y. Nie, and M. Simard. Embedding web-based statistical translation models in cross-language information retrieval. *Computational Linguistics*, 29, 2003.

[20] R. Krovetz and W. B. Croft. Lexical ambiguity and information retrieval. *ACM Transactions on Information Systems*, 10(2), 1992.

[21] O. Kurland and L. Lee. Corpus structure, language models, and ad hoc information retrieval. In *SIGIR '04*, 2004.

[22] J. Lafferty and C. Zhai. Document language models, query models, and risk minimization for information retrieval. In *SIGIR '01*, 2001.

[23] V. Lavrenko and W. B. Croft. Relevance based language models. In *SIGIR '01*, 2001.

[24] X. Liu and W. B. Croft. Cluster-based retrieval using language models. In *SIGIR '04*, 2004.

[25] Y. Lv and C. Zhai. A comparative study of methods for estimating query language models with pseudo feedback. In *CIKM '09*, 2009.

[26] Y. Lv and C. Zhai. Positional relevance model for pseudo-relevance feedback. In *SIGIR '10*, 2010.

[27] W. Macherey, F. J. Och, I. Thayer, and J. Uszkoreit. Lattice-based minimum error rate training for statistical machine translation. In *EMNLP '08*, 2008.

[28] J. S. McCarley. Should we translate the documents or the queries in cross-language information retrieval? In *ACL '99*, 1999.

[29] R. Mihalcea and D. Moldovan. Semantic indexing using WordNet senses. In *Proceedings of the ACL Workshop on Recent Advances in Natural Language Processing and Information Retrieval*, 2000.

[30] D. W. Oard and B. J. Dorr. A survey of multilingual text retrieval. Technical report, 1996.

[31] F. J. Och and H. Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1), 2003.

[32] A. Pirkola. The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. In *SIGIR '98*, 1998.

[33] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *SIGIR '98*, 1998.

[34] M. Sanderson. Word sense disambiguation and information retrieval. In *SIGIR '94*, 1994.

[35] H. Schütze and J. Pedersen. Information retrieval based on word senses. In *Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval*, 1995.

[36] C. Stokoe, M. P. Oakes, and J. Tait. Word sense disambiguation in information retrieval revisited. In *SIGIR '03*, 2003.

[37] A. Stolcke. SRILM - An extensible language modeling toolkit. In *ICSLP*, 2002.

[38] C. Tillmann, S. Vogel, H. Ney, and A. Zubiaga. A DP based search using monotone alignments in statistical translation. In *EACL '97*, 1997.

[39] D. Trieschnigg, D. Hiemstra, F. de Jong, and W. Kraaij. A cross-lingual framework for monolingual biomedical information retrieval. In *CIKM '10*, 2010.

[40] E. M. Voorhees. Using WordNet to disambiguate word senses for text retrieval. In *SIGIR '93*, 1993.

[41] F. Wessel, R. Schlüter, K. Macherey, and H. Ney. Confidence measures for large vocabulary continuous speech recognition. *IEEE Transactions on Speech and Audio Processing*, 9, 2001.

[42] Y. Xu, G. J. Jones, and B. Wang. Query dependent pseudo-relevance feedback based on Wikipedia. In *SIGIR '09*, 2009.

[43] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *SIGIR '01*, 2001.