Matching Meaning for Cross-Language Information Retrieval

Jianqiang Wang

Department of Library and Information Studies
University at Buffalo, the State University of New York
Buffalo, NY 14260, U.S.A.

Douglas W. Oard

College of Information Studies and UMIACS
University of Maryland
College Park, MD 20742, U.S.A.

Abstract

This article describes a framework for cross-language information retrieval that efficiently leverages statistical estimation of translation probabilities. The framework provides a unified perspective into which some earlier work on techniques for cross-language information retrieval based on translation probabilities can be cast. Modeling synonymy and filtering translation probabilities using bidirectional evidence are shown to yield a balance between retrieval effectiveness and query-time (or indexing-time) efficiency that seems well suited large-scale applications. Evaluations with six test collections show consistent improvements over strong baselines.

Keywords:
Cross-Language IR, Statistical machine translation

Email addresses: jw254@buffalo.edu (Jianqiang Wang), oard@umd.edu (Douglas W. Oard)

URL: http://www.buffalo.edu/~jw254/ (Jianqiang Wang), http://terpconnect.umd.edu/~oard (Douglas W. Oard)

1. Introduction

Cross-language Information Retrieval (CLIR) is the problem of finding documents that are expressed in a language different from that of the query. For the purpose of this article, we restrict our attention to techniques for ranked retrieval of documents containing terms in one language (which we consistently refer to as f) based on query terms in some other language (which we consistently refer to as e). A broad range of approaches to CLIR involve some sort of direct mapping between terms in each language, either from e to f (query translation) or from f to e (document translation). In this article we argue that these are both ways of asking the more general question "do terms e and f have the same meeting?" Moreover, we argue that this more general question is in some sense the "right" question, for the simple reason that it is the fundamental question that we ask when performing monolingual retrieval. We therefore derive a "meaning matching" framework, first introduced in (Wang and Oard, 2006), but presented here in greater detail.

Instantiating such a model requires that we be specific about what we mean by a "term." In monolingual retrieval we might treat each distinct word as a term, or we might group words with similar meanings (e.g., we might choose to index all words that share a common stem as the same term). But in CLIR there is no escaping the fact that synonymy is central to what we are doing when we seek to match words that have the same meaning. In this article we show through experiments that by modeling synonymy in both languages we can improve efficiency at no cost (and indeed perhaps with some improvement) in retrieval effectiveness. The new experiments in this paper show that this effect is not limited to the three test collections on which we had previously observed this result (Wang, 2005; Wang and Oard, 2006).

When many possible translations are known for a term, a fundamental question is how we should select which translations to use. In our earlier work, we had learned translation probabilities from parallel text and then used however many translations were needed to reach a preset threshold for the Cumulative Distribution Function (CDF) (Wang and Oard, 2006). In this article we extend that work by comparing a CDF threshold to two alternatives: (1) a threshold on the Probability Mass Function (PMF), and (2) a fixed threshold on the number of translations. The results show that thresholding the CDF or the PMF are good choices.

The remainder of this article is organized as follows. Section 2 reviews the salient prior work on CLIR. Section 3 then introduces our "meaning matching" model and explains how some specific earlier CLIR techniques can be viewed as restricted variants of that general model. Section 4 presents new experiment results that demonstrate its utility and that explore which aspects of the model are responsible for the observed improvements in retrieval effectiveness. Section 5 concludes the article with a summary of our findings and a discussion of issues that could be productively explored in future work.

2. Background

Our meaning matching model brings together three key ideas that have previously been shown to work well in more restricted contexts. In this section we focus first on prior work on combining evidence from different document-language terms to estimate useful weights for query terms in individual documents. We then trace the evolution of the idea that neither translation direction may be as informative as using both together. Finally, we look briefly at prior work on the question of which translations to use.

2.1. Estimating Query Term Weights

A broad class of information retrieval models can be thought of as computing a weight for each query term in each document and then combining those query term weights in some way to compute an overall score for each document. This is the so-called "bag of words" model. Notable examples are the vector space model, the Okapi BM25 measure, and some language models.

In early work on CLIR a common approach was to replace each query term with the translations found in a bilingual term list. When only one translation is known, this works as well as anything. But when different numbers of translations are known for different terms this approach suffers from an unhelpful imbalance (because common terms often have many translations, but little discriminating power). Fundamentally this approach is flawed because it fails to structurally distinguish between different query terms (which provide one type of evidence) and different translations for the same query term (which provide a different type of evidence).

Pirkola (1998) was the first to articulate what has become the canonical solution to this problem. Pirkola's method estimates term specificity in essentially the same way as is done when stemming is employed in same-language

retrieval (i.e., any document term that can be mapped to the query term is counted). This has the effect of reducing the term weights for query terms that have at least one translation that is a common term in the document language, which empirically turns out to be a reasonable choice. The year 1998 was also when Nie et al. (1998) and McCarley and Roukos (1998) were the first to try using learned translation probabilities rather than translations found in a dictionary. They, and most researchers since, learned translation probabilities from parallel (i.e., translation-equivalent) texts using techniques that were originally developed for statistical machine translation (Knight, 1999).

The next year, Hiemstra and de Jong (1999) put these two ideas together, suggesting (but not testing) the idea of using translation probabilities as weights on the counts of the known translations (rather than on the Inverse Document Frequency (IDF) values, as Nie et al. (1998) had done, or for selecting a single best translation, as (McCarley and Roukos, 1998) had done). They described this as being "somewhat similar" to Pirkola's structured translation technique, since the unifying idea behind both was that evidence combination across translations should be done before evidence combination across query terms. Xu and Weischedel (2000) were the first to actually run experiments using an elegant variant of this approach in which the Term Frequency (TF) of term e, $tf(e)$, was estimated in the manner that Hiemstra and de Jong (1999) had suggested, but the Collection Frequency (CF) of the term, $cf(e)$, which served a role similar to Hiemstra's document frequency, was computed using a separate query-language corpus rather than being estimated through the translation mapping from the document collection being searched.

Hiemstra and de Jong (1999) and Xu and Weischedel (2000) developed their ideas in the context of language models. It remained for Darwish and Oard (2003) to apply similar ideas to a vector space model. The key turned out to be a computational simplification to Pirkola's method that had been introduced by Kwok (2000) in which the number of documents containing each translation was summed to produce an upper bound on the number of documents that could contain at least one of those translations. Darwish and Oard (2003) showed this bound to be very tight (as measured by the extrinsic effect on Mean Average Precision (MAP)), and from there the extension to using translation probabilities as weights on term counts was straightforward.

Statistical translation models for machine translation are typically trained on strings that represent one or more consecutive tokens, but for informa-

tion retrieval some way of conflating terms with similar meanings can help to alleviate sparsity without adversely affecting retrieval effectiveness. For example, Fraser et al. (2002) trained an Arabic-English translation model on stems (more properly, on the results of what it called "light stemming" for Arabic). Our experiments with aggregation draw on a generalization of this idea.

The idea of using learned translation probabilities as term weights resulted in somewhat of a paradigm shift in CLIR. Earlier "dictionary-based" techniques had rarely yielded MAP values much above 80% of that achieved by a comparable monolingual system. But with translation probabilities available we started seeing routine reports of 100% or more. For example, Xu and Weischedel (2000) reported retrieval results that were 118% of monolingual MAP (when compared using automatically segmented Chinese terms), suggesting that (in the case of their experiments) if you wanted to search Chinese you might actually be better off formulating your queries in English!

2.2. Bidirectional Translation

Throughout these developments, the practice regarding whether to translate f to e or e to f remained somewhat inconsistent. Nie et al. (1998) (and later Darwish and Oard (2003)) thought of the problem as query translation, while McCarley and Roukos (1998), Hiemstra and de Jong (1999) and Xu and Weischedel (2000) thought of it as document translation. In reality, of course, nothing was being "translated." Rather, counts were being mapped.

Indeed, the implications of choosing a direction weren't completely clear at that time. We can now identify three quite different things that have historically been treated monolithically when "query translation" or "document translation" is mentioned: (1) whether the processing is done at query time or at indexing time, (2) which direction is assumed when learning the word alignments from which translation probabilities were estimated (which matters only because widely used efficient alignment techniques are asymmetric), and (3) which direction is assumed when the translation probabilities are normalized. We now recognize these as separable issues, and when effectiveness is our focus it is clear that the latter two should command our attention. Whether computation is done at query time or at indexing time is, of course, an important implementation issue, but if translation probabilities don't change the results will be the same either way.

McCarley (1999) was the first to explore the possibility of using both directions. He did this by building two ranked lists, one based on using the one-best translation by $p(e|f)$ and the other based on using the one-best translation by $p(f|e)$. Combining the two ranked lists yielded better MAP than when either approach was used alone. Similar improvements have since been reported by others using variants of that technique (Braschler, 2004; Kang et al., 2004).

Boughanem et al. (2001) tried one way of pushing this insight inside the retrieval system, simply filtering out potentially problematic translations that were attested in only one direction. They did so without considering translation probabilities, however, working instead with bilingual dictionaries. On that same day, Nie and Simard (2001) introduced a generalization of that approach in which translation probabilities for each direction could be interpreted to as partially attesting the translation pair. The product of those probabilities was (after renormalization) therefore used in lieu of the probability in either direction alone. Our experiments in (Wang and Oard, 2006) suggest that this can be a very effective approach, although the experiments in Nie and Simard (2001) on a different test collection (and with some differences in implementation details) were not as promising. As we show in Section 4.1.3, the relative effectiveness of bidirectional and unidirectional translation does indeed vary between test collections, but aggregation can help to mitigate that effect and, regardless, bidirectional translation offers very substantial efficiency advantages.

2.3. Translation Selection

One challenge introduced by learned translation probabilities is that there can be a very long tail on the distribution (because techniques that rely on automated alignment might in principle try to align any term in one language with any term in the other). This leads to the need for translation selection, one of the most thoroughly researched issues in CLIR. Much of that work has sought to exploit context to inform the choice. For example, Federico and Bertoldi (2002) used an order-independent bigram language model to make choices in a way that would prefer translated words that are often seen together. By relaxing the term independence assumption that is at the heart of all bag-of-words models, these techniques seek to improve retrieval effectiveness, but at some cost in efficiency. In this article, we have chosen to focus on techniques that preserve term independence, all of which are based

on simply choosing the most likely translations. The key question, then, is how far down that list to go.

Perhaps the simplest alternative is to select some fixed number of translations. For example, Davis and Dunning (1995) used 100 translations, Xu and Weischedel (2000) (observing that using large numbers of translations has adverse implications for efficiency) used 20, and Nie et al. (1998) reported results over a range of values. Such approaches are well suited to cases in which a preference order among translations is known, but reliable translation probabilities are not available (as is the case for the order in which translations are listed in some bilingual dictionaries).

Because the translation probability distribution is sharper for some terms than others, it is attractive to consider alternative approaches that can make use of that information. Two straightforward ways have been tried: Xu and Weischedel (2000) used a threshold on the Probability Mass Function (PMF), while Darwish and Oard (2003) used a threshold on the Cumulative Distribution Function (CDF). We are not aware of comparisons between these techniques, a situation we rectify in Section 4.1.3 and Section 4.2.3.

Another approach is to look holistically at the translation model rather than at just the translations of any one term, viewing translation selection as a feature selection problem in which the goal is to select some number of features (i.e., translation pairs) in a way that maximizes some function for the overall translation model between all term pairs. Kraaij et al. (2003) reports that this approach (using an entropy function) yields results that are competitive with using a fixed PMF threshold that is the same for all terms. Our results suggest that the PMF threshold is indeed a suitable reference. Future work to compare effectiveness, efficiency and robustness of approaches based on entropy maximization with those based on a PMF threshold clearly seems called for, although we do not add to the literature on that question in this article.

3. Matching Meaning

In this section, we rederive our overarching framework for matching meanings between queries and documents, presenting a set of computational implementations that incorporate evidence from translation probabilities in different ways.

3.1. IR as Matching Meaning

The basic assumption underlying meaning matching is that some hidden shared meaning space exists for terms in different languages. Meaning matching across languages can thus be achieved by mapping the meanings of individual terms into that meaning space, using it as a "bridge" between terms in different languages. Homography and polysemy (i.e., terms that have multiple distant or close meanings) result in the possibility of several such "bridges" between the same pair of terms. This way of looking at the problem suggests that the probability that two terms share the same meaning can be computed as the summation over some "meaning space" of the probabilities that both terms share each specific meaning.

for a query term e in Language E, we assume that each document-language term f_i $(i = 1, 2, \ldots, n)$ in Language F shares the meaning of e that was intended by the searcher with some probability $p(e \leftrightarrow f_i)$ $(i = 1, 2, \ldots, n)$, respectively. We have coined the notation $p(e \leftrightarrow f_i)$ as a shorthand for this meaning matching probability so as to avoid implying any one translation direction in our basic notation. For a term in Language F that does not share any meaning with e, the meaning matching probability between that term and e will be 0. Any uncertainty about the meaning of e is reflected in these probabilities, the computation of which is described below. If we see a term f_i that matches the meaning of term e one time in document d_k, we can treat this as having seen query term e occurring $p(e \leftrightarrow f_i)$ times in d_k. If term f_i occurs $tf(f_i, d_k)$ times, our estimate of the total "occurrence" of query term e will be $p(e \leftrightarrow f_i)tf(f_i, d_k)$. Applying the usual term independence assumption on the document side and considering all the terms in document d_k that might share a common meaning with query term e, we get:

$$tf(e, d_k) = \sum_{f_i} p(e \leftrightarrow f_i)tf(f_i, d_k) \tag{1}$$

Turning our attention to the *df*, if document d_k contains a term f_i that shares a meaning with e, we can treat the document as if it possibly "contained" e. We adopt a frequentist interpretation and increment the *df* by the sum of the probabilities for each unique term that might share a common meaning with e. We then assume that terms are used independently in different documents and estimate the *df* of query term e in the collection as:

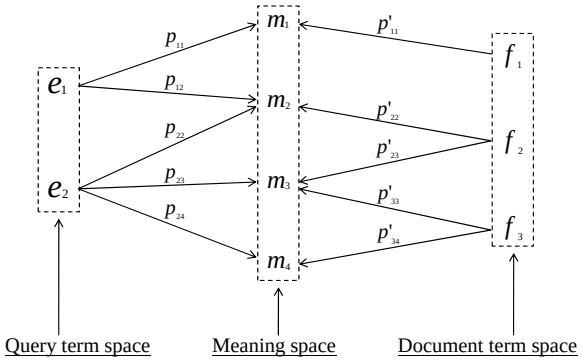


Figure 1: Matching term meanings through a shared meaning space

$$df(e) = \sum_{f_i} p(e \leftrightarrow f_i) df(f_i) \tag{2}$$

Because we are interested only in relative scores when ranking documents, we can (and do) perform document length normalization using the document-language terms rather than the mapping of those terms to the query language.

Equations (1) and (2) show how the meaning matching probability between a query term and a document term can be incorporated into the computation of term weight. The remaining question then becomes how the meaning matching probability $p(e \leftrightarrow f)$ can be modeled and computed for any given pair of query term e and document term f.

3.2. Matching Abstract Term Meanings

Given a shared meaning space, matching term meaning involves mapping terms in different languages into this shared meaning space. Figure 1 illustrates this idea for a case in which two terms in the query language E and three terms in the document language F share subsets of four different meanings. At this point we treat "meaning" as an abstract concept; a computational model of meaning is introduced in the next section. In our example, term e_2 has the same meaning as term f_2 if and only if e_2 and f_2 both express meaning m_2 or e_2 and f_2 both express meaning m_3. If we assume that the searcher's choice of meaning for e_2 is independent of the author's choice of meaning for f_2, we can compute the probabilities of those two events. Generalizing to any pair of terms e and f:

$$p(e \leftrightarrow f) = \sum_{m_i} p(m_i|(e,f)) \tag{3}$$

Applying Bayes' rule, we get:

$$\begin{aligned} p(e \leftrightarrow f) &= \sum_{m_i} \frac{p(m_i, e, f)}{p(e,f)} \\ &= \sum_{m_i} \frac{p((e,f)|m_i)p(m_i)}{p(e,f)} \end{aligned} \tag{4}$$

Assume, given a meaning, that seeing a term in one language is conditionally independent of seeing another term in the other language (i.e., $p((e,f)|m_i) = p(e|m_i)p(f|m_i)$), then:

$$\begin{aligned} p(e \leftrightarrow f) &= \sum_{m_i} \frac{p(e|m_i)p(f|m_i)p(m_i)}{p(e,f)} \\ &= \sum_{m_i} [\frac{p(e,m_i)}{p(m_i)} \frac{p(f,m_i)}{p(m_i)} p(m_i)]/p(e,f) \\ &= \sum_{m_i} \frac{p(e,m_i)p(f,m_i)}{p(m_i)p(e,f)} \\ &= \sum_{m_i} \frac{[p(m_i|e)p(e)][p(m_i|f)p(f)]}{p(m_i)p(e,f)} \\ &= \sum_{m_i} [p(m_i|e)p(m_i|f)] \frac{p(e)p(f)}{p(m_i)p(e,f)} \end{aligned} \tag{5}$$

Furthermore, assuming seeing a term in one language is (unconditionally) independent of seeing another term in the other language (i.e., $p(e,f) = p(e)p(f)$), Equation 5 then becomes:

$$p(e \leftrightarrow f) = \sum_{m_i} [p(m_i|e)p(m_i|f)]p(m_i) \tag{6}$$

Lastly, if we make the somewhat dubious but very useful assumption that every possible shared meaning has an equal chance of being expressed, $p(m_i)$ then becomes a constant. Therefore:

$$p(e \leftrightarrow f) \propto \sum_{m_i} p(m_i|e)p(m_i|f) \tag{7}$$

where:

- $p(e \leftrightarrow f)$: the probability that term e and term f have the same meaning.
- $p(m_i|e)$: the probability that term e has meaning m_i
- $p(m_i|f)$: the probability that term f has meaning m_i

For example (see Figure 1), if all possible meanings of every term were equally likely, then $p_{11} = p_{12} = 0.5$, $p_{22} = p_{23} = p_{24} = 0.33$, $p'_{11} = 1$, $p'_{22} = p'_{23} = 0.5$, and $p'_{33} = p'_{35} = 0.5$; and the meaning matching probability between term e_2 and term f_2 will be: $p(e_2 \leftrightarrow f_2) \propto p_{22} \times p'_{22} + p_{23} \times p'_{23} = 0.33 \times 0.5 + 0.33 \times 0.5 = 0.33$.

3.3. Using Synsets to Represent Meaning

We use "synsets," sets of synonymous terms as a straightforward computational model of meaning. To make this explicit, we denote a synset s_i for each meaning m_i in the shared meaning space, so the meaning matching model described in Equation (7) simply becomes:

$$p(e \leftrightarrow f) \propto \sum_{s_i} p(s_i|e)p(s_i|f) \tag{8}$$

Our problem is now reduced to two subproblems: (1) creating synsets s_i, and (2) computing the probability of any specific term mapping to any specific synset $p(s_i|e)$ and $p(s_i|f)$. For the first task, it is obvious that to be useful synset s_i must contain synonyms in both languages. One way to develop such multilingual synsets is as follows:

1. Create synsets s_{E_j} $(j = 1, 2, \ldots, l)$ in Language E;
2. Create synsets $s_{F_k} (k = 1, 2, \ldots, m)$ in Language F;
3. Align synsets in two languages, resulting in a combined synset (s_{E_i}, s_{F_i}) $(i = 1, 2, \ldots, n)$, which we call s_i.

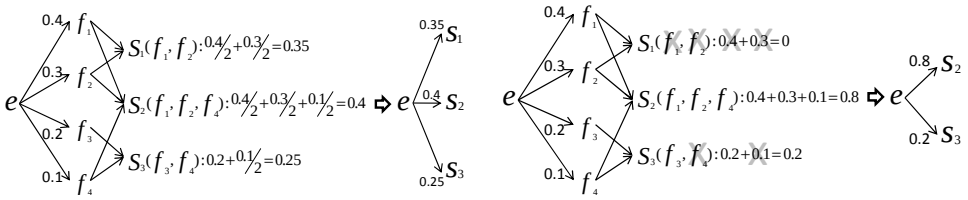
Cross-language synset alignments are available from some sources, most notably lexical resources such as EuroWordNet. However, mapping unrestricted text into WordNet is well known to be error prone (Voorhees, 1993). Our early experiments with EuroWordNet proved to be disappointing (Wang, 2005), so for the experiments in this article we instead adopt the statistical technique for discovering same-language synonymy that we first used in (Wang and Oard, 2006).

Previous work on word sense disambiguation suggests that translation usage can provide a useful basis for identifying terms with similar meaning (Resnik and Yarowsky, 2000; Xu et al., 2002). The key idea is that if term f in language F can translate to a term e_i in language E, which can further back-translate to some term f_j in language F, then f_j *might* be a synonym of f. Furthermore, the more terms e_i exist as bridges between f and f_j, the more confidence we should have that f_j is a synonym of f. Formalizing this notion:

$$p(f_j \in s_f) \approx \sum_{i=1}^{n} p(f_j|e_i)p(e_i|f) \tag{9}$$

where $p(f_j \in s_f)$ is the probability of f_j being a synonym of f (i.e., in a synset s_f of word f), $p(e_i|f)$ is obtained from a statistical translation model from Language F to Language E, and $p(f_j|e_i)$ is obtained from a statistical translation model from Language E to Language F. Probability values generated in this way are usually sharply skewed, with only translations that are strongly attested in both directions retaining much probability mass, so any relatively small threshold on the result of Equation 9 would suffice to suppress unlikely synonyms. We somewhat arbitrarily chose a threshold of 0.1 and have used that value consistently for the experiments reported in this article (and in our previous experiments reported in (Wang, 2005; Wang and Oard, 2006)). Candidate synonyms with a normalized probability larger than 0.1 are therefore retained and, along with f, form synset s_f. The same term can appear in multiple synsets with this method; that fact has consequences for meaning matching, as we describe below.

As an example, applying Equation 9 using the statistical translation probabilities described later in Section 4.2.1, we automatically constructed five synsets that contain the English word "rescue": *(holzmann, rescue), (fund, intervention, ltcm, rescue, hedge), (saving, uses, saved, rescue), (rafts, rescue)*, and *(saving, saved, rescue, salvage)*. As can be seen, many of these



(a) Conservative aggregation. (b) Greedy aggregation.

Figure 2: Two methods of conflating multiple translations into synsets, f_i $(i = 1, 2, 3, 4)$: translations of term e, S_j $(j = 1, 2, 3)$: synsets.

terms are often not actually synonyms in the usual sense, but they do capture useful relationships (e.g., the Holzmann construction company was financially rescued, as was the hedge fund LTCM), and drawing on related terms in information retrieval applications can often be beneficial. So although we refer to what we build as "synsets," in actuality these are simply sets of related terms.

3.4. From Statistical Translation to Word-to-Synset Mapping

Because some translation f_i of term e may appear in multiple synsets, we need some way of deciding how $p(e \leftrightarrow f_i)$ should be allocated across synsets. Figure 2 presents an example of two ways of doing this. Figure 2a illustrates the effect of splitting the translation probability evenly across each synset in which a translation appears, assuming a uniform distribution. For example, since translation f_1 appears in synsets s_1 and s_2 and $p(e \leftrightarrow f_1) = 0.4$, we add $0.4/2 = 0.2$ to both $p(s_1|e)$ and $p(s_2|e)$.

Figure 2b illustrates an alternative in which each translation f_i is assigned only to the synset that results in the sharper translation probability distribution. We call this *greedy aggregation*. We do this by iteratively assigning each translation to the synset that would yield the greatest aggregate probability, as follows:

1. Compute the largest possible aggregate probability that e maps to each s_{F_i}, which is defined as: $p(s_{F_i}|e) = \sum_{f_j \in s_{F_i}} p(f_j|e)$.
2. Rank all s_{i_f} in decreasing order of that largest possible aggregate probability;
3. Select the synset s_{F_i} with the largest aggregate probability, and remove all of its translations f_j from every other synset;
4. Repeat Steps 1–3 until each translation f_j has been assigned to a synset.

Method (b) is *minimalist* in the sense that it seeks to minimize the number of synsets. Moreover, Method (b) does this by rewarding mutually reinforcing evidence: when we have high confidence that e can properly be translated to some synonym of f_j, that might quite reasonably raise our confidence in f_j as a plausible translation. Both of these are desirable properties, so we chose method (b) for the experiments reported in this article.

The two word-to-synset mappings in Figure 3 illustrate the effect of applying Method (b) to the corresponding pre-aggregation translation probabilities. For example, on the left side of that figure each translation (into English) of the French term "sauvetage" is assigned to a single synset, which inherits the sum of the translation probabilities of its members.[1]

At this point, the most natural thing to do would be to index each synset as a term. Doing that would add some implementation complexity, however, since *rescue* and *saving* are together in a synset when translating the French term "sauvetage," but they might wind up in different synsets when translating some other French term. To avoid that complexity, for our experiments we instead constructed ersatz word-to-word translation probabilities by distributing the full translation probability for each synset to each term in that synset and then renormalizing it. The results are shown in the penultimate row in Figure 3.

3.5. Variants of the Meaning Matching Model

Aggregation and bidirectionality are distinguishing characteristics of our full meaning matching model, but restricted variants of the model are also possible. In this section we introduce variants of the basic model, roughly in increasing order of complexity. See Table 1 for a summary and Figure 3 for a worked example.

- *Probabilistic Structured Queries* (PSQ): one of the simplest variants, using only translation probabilities learned and normalized in the query translation direction (Darwish and Oard, 2003).

- *Probabilistic Document Translation* (PDT): an equally simple variant, using only translation probabilities learned and normalized in the document translation direction.

[1]By convention, throughout this article we use a slash to separate a term or a synset from its translation probability.

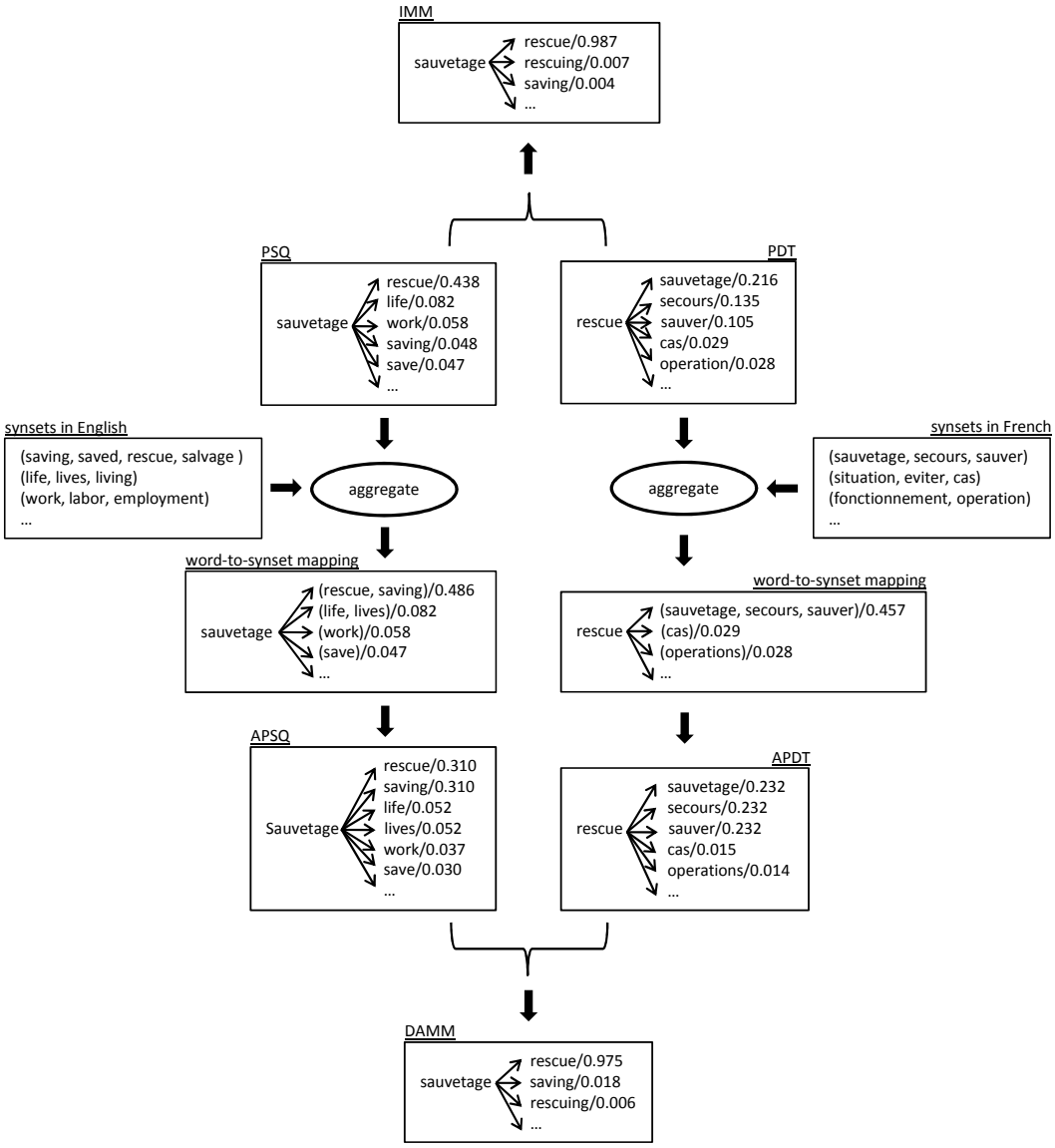


Figure 3: Examples showing how variants of meaning matching model are developed.

- *Individual Meaning Matching* (IMM): translation probabilities for both directions are used without synsets by multiplying the probabilities for PSQ and PDT. Since the result of multiplying probabilities is no longer normalized we renormalize in the query translation direction (so that the sum over each translation f of a query term e is 1). IMM can be thought of as a variant of DAMM (explained below) in which each term encodes a unique meaning.

- *Aggregated Probabilistic Structured Queries* (APSQ): translation probabilities in the query translation direction are aggregated into synsets, replicated, and renormalized as described above.

- *Aggregated Probabilistic Document Translation* (APDT): translation probabilities in the document translation direction are aggregated into synsets, replicated, and renormalized as described above.

- *Derived Aggregated Meaning Matching* (DAMM): translation probabilities are used with synsets for both directions by multiplying the APSQ and APDT probabilities and then renormalizing the result in the query translation direction.

- *Partially Aggregated Meaning Matching* (PAMM): a midpoint between IMM and DAMM, translation probabilities in both directions are used, but with aggregation applied only to one of those directions (to the query translation direction for PAMM-F and the document translation direction for PAMM-E). Specifically, for PAMM-F we multiply APSQ and PDT probabilities, for PAMM-E we multiply PSQ and APDT probabilities; in both cases we then renormalize in the query translation direction. For simplicity, PAMM-F and PAMM-E are not shown in Figure 3.

3.6. Renormalization

Two meaning matching techniques (PSQ and APSQ) are normalized by construction in the query translation direction; two others (PDT and APDT) are normalized in the document translation direction. For the others, probability mass is lost when we multiply and we therefore need to choose a renormalization direction. As specified above, we consistently choose the query translation direction. The "right" choice is, however, far from clear.

Variant acronym	Query trans probs	Doc trans probs	Query lang synsets	Doc lang synsets	$p(e \leftrightarrow f)$
PSQ	$\surd$				$= p(f\|e)$
PDT		$\surd$			$= p(e\|f)$
IMM	$\surd$	$\surd$			$\propto p(f\|e)p(e\|f)$
APSQ	$\surd$			$\surd$	$\propto p(s_f\|e)$
APDT		$\surd$	$\surd$		$\propto p(s_e\|f)$
DAMM	$\surd$	$\surd$	$\surd$	$\surd$	$\propto p(s_f\|e)p(s_e\|f)^*$
PAMM-E	$\surd$	$\surd$	$\surd$		$\propto p(f\|e)p(s_e\|f)$
PAMM-F	$\surd$	$\surd$		$\surd$	$\propto p(s_f\|e)p(e\|f)$

Table 1: Meaning matching variants. D: Derived, P: Partial, A: Aggregated, MM: Meaning Matching; PSQ: Probabilistic Structured Queries; PDT: Probabilistic Document Translation.

* Because we normalize each synonym set and then the product, the proportionality symbols in DAMM and PAMM are useful as a shorthand, but not strictly correct.

The problem arises because what we call Document Frequency (DF) is really a fact about a query term (helping us to weight that term appropriately with respect to other terms in the same query), while Term Frequency (TF) is a fact about a term in a document. This creates some tension, with the query translation direction seeming to be most appropriate for using DF evidence to weight the relative specificity of query terms and the document translation direction seeming to be most appropriate for estimating TF in the query language from the observed TF's in the document language.

To see why this is so, consider first the DF. The question we want to ask is how many documents we believe each query term (effectively) occurs in. For any one query term, that answer will depend on which translation(s) we believe to be appropriate. If query term e can be translated to document language terms f_1 or f_2 with equal probability (0.5 each), then it would be reasonable to estimate the DF of e as the expectation over that distribution of the DF of f_1 and the DF of f_2. This is achieved by normalizing so that $\sum_{f_i} p(f_i|e) = 1$ and then computing $DF(e) = \sum_{f_i} p(f_i|e)DF(f_i)$. Normalizing in the other direction would make less sense, since it could result in DF estimates that exceed the number of documents in the collection.

Now consider instead the TF calculation. The question we want to ask in this case is how many times a query term (effectively) occurred in each document. If we find term f in some document, and if f can be translated as either e_1 or e_2 with equal probability, and if our query term is e_1, then in the absence of any other evidence the best we can reasonably do is to ascribe half the occurrences of f to e_1. This is achieved by normalizing so that $\sum_{f_i} p(e|f_i) = 1$ and then computing $TF(e, d_k) = \sum_{f_i} p(e|f_i)TF(f_i, d_k)$. Normalizing in the other direction would make less sense, since in extreme cases that could result in TF estimates for different query terms that sum to more terms than are actually present in the document.

Our early experience with mismanaging DF effects (Oard and Wang, 1999) and the success of the DF handling in Pirkola's structured queries (Pirkola, 1998) have led us to favor reasonable DF calculations when forced to choose. When probability mass is lost (as it is in IMM, DAMM, PAMM-E, and PAMM-F), we therefore normalize so that $\sum_{f_i} p(f_i|e) = 1$ (i.e., in the query translation direction). This choice maximizes the comparability between those techniques and PSQ and APSQ, which are normalized in that same direction by construction. We do, however, still gain some insight into the other normalization direction from our PDT and APDT experiments (see Section 4 below).

4. Experiments

In our earlier conference paper (Wang and Oard, 2006), we reported on two sets of experiments, one using English queries and French news text, and the second using English queries and Chinese news text. A third set of experiments, again with English queries and Chinese news text, was reported in (Wang, 2005). Table 2 shows the test collection statistics and the best Mean Average Precision (MAP) obtained in those experiments for each Meaning Matching (MM) variant. In each experiment, we swept a CDF threshold to find the peak MAP (usually at a CDF of 0.9 or 0.99).

Several conclusions are evident from these results. First, at the peak CDF threshold DAMM is clearly a good choice, sometimes equaled but never bettered. Second, PSQ and APSQ are at the other end of the spectrum, always statistically significantly below DAMM. The results for IMM, PDT and APDT are more equivocal, with each doing better than the other two in one of the three cases. PAMM-E and PAMM-F turned out to be statistically indistinguishable from DAMM, but perhaps not worthy of as much attention

Collection	CLEF-(01-03)		TREC-5&6		TREC-9	
Queries	English		English		English	
Documents	French news		Chinese news		Chinese news	
Topics	151		54		25	
Documents	87,191		164,789		126,937	
	MAP % of DAMM	MAP % of Mono	MAP % of DAMM	MAP % of Mono	MAP % of DAMM	MAP % of Mono
DAMM	- -	100.3%	- -	97.8%	- -	128.2%
PAMM-F	99.7%	100%	100%	97.8%	96.2%	123.3%
PAMM-E	99.7%	100%	94.9%	92.3%	91.4%	117.1%
IMM	97.2%	97.8%	92.1%	90.1%	**87.9%**	112.7%
PDT	96.3%	96.9%	89.9%	87.9%	98.1%	125.7%
APDT	**92.5%**	**92.7%**	98.7%	96.6%	**88.5%**	113.5%
PSQ	**94.6%**	**94.8%**	**83.7%**	**82.0%**	**90.4%**	115.9%
APSQ	**83.2%**	**83.4%**	**56.6%**	**55.4%**	**49.7%**	**63.7%**

Table 2: Peak retrieval effectiveness for meaning matching variants in three previous experiments ("Mono" is the monolingual baseline, **bold** indicates a statistically significant difference.)

since they occupy a middle ground between IMM and DAMM both in the way they are constructed and (to the extent that the insignificant differences are nevertheless informative) numerically in the results as well.

More broadly, we can conclude that there is clear evidence that bidirectional translation is generally helpful (comparing DAMM to APDT and APSQ, comparing PAMM-F to APDT and PSQ, comparing PAMM-E to APSQ and PDT, and comparing IMM to PSQ and PDT), but not always (PDT yields better MAP than IMM one time out of three, for example). We can also conclude that aggregation results in additional improvement when bidirectional translation is used (comparing DAMM, PAMM-E and PAMM-F to IMM), but that the same effect is not present with unidirectional translation (with APDT below PDT in two cases out of three, and APSQ always below PSQ).

Notably, the three collections on which these experiments were run are relatively small, and all include only news. In this section we therefore extend our earlier work in two important ways. We first present a new set of experiments with a substantially larger test collection than we have used to date. That is followed by another new set of experiments for two content types other than news, using French queries to search English conversational speech or to search English metadata that was manually associated with that

speech. Finally, we look across the results that we have obtained to date to identify commonalities (which help characterize the strengths and weaknesses of our meaning matching model) and differences (which help characterize dependencies on the nature of specific test collections).

4.1. New Chinese Experiments

CLIR results from our previous Chinese experiments in (Wang (2005); Wang and Oard (2006)) were quite good, with DAMM achieving 98% and 128% of monolingual MAP (see Table 2). Many CLIR settings are more challenging, however, so we chose for our third set of English-Chinese experiments a substantially larger English-Chinese test collection from NTCIR-5, for which the best NTCIR-5 system had achieved only 62% of monolingual MAP (Kishida et al., 2005).

4.1.1. Training Statistical Translation Models

For comparability, we re-used the statistical translation models that we had built for our previous experiments with the TREC-5&6 and TREC-9 CLIR collections (Wang, 2005; Wang and Oard, 2006). To briefly recap, we used what was at the time (in 2005) the word alignments from which others in our group were at the time building state-of-the-art hierarchical phrase-based models for statistical machine translation (Chiang et al., 2005). The models were trained using the GIZA++ toolkit (Och and Ney, 2000)[2] on a sentence-aligned English-Chinese parallel corpus that consisted of corpora from multiple sources, including the Foreign Broadcast Information Service (FBIS), Hong Kong News, Hong Kong Laws, the United Nations, and Sinorama. All were written using simplified Chinese characters. A modified version of the Linguistic Data Consortium (LDC) Chinese segmenter was used to segment the Chinese side of the corpus. After removing implausible sentence alignments by eliminating sentence pairs that had a token ratio either smaller than 0.2 or larger than 5, we used the remaining 1,583,807 English-Chinese sentence pairs for MT training. Statistical translation models were built in each direction with 10 IBM Model 1 iterations and 5 HMM iterations. A CDF threshold of 0.99 was applied to the model for each direction before they were used to derive the eight meaning matching variants described in Section 3.

[2]http://www-i6.informatik.rwth-aachen.de/Colleagues/och/software/GIZA++.html

4.1.2. Preprocessing the Test Collection

The NTCIR-5 English-to-Chinese CLIR test collection (formally, CIRB040r), contains 901,446 documents from United Daily News, United Express, Ming Hseng News, and Economic Daily News. All of the documents were written using traditional Chinese characters. Relevance judgments for total of 50 topics are available. These 50 topics were originally authored in Chinese (using traditional characters), Korean or Japanese (18, 18 and 14 topics, respectively) and then manually translated into English, and then translated from English into each of the two other languages. For our study, the English version of each topic was used as a basis for forming the corresponding CLIR query; the Chinese version was used as a basis for forming the corresponding monolingual query. Specifically, we used the TITLE field from each topic to form its query. Four degrees of relevance are available in this test collection. We treated "highly relevant" and "relevant" as relevant, and "partially relevant" and "irrelevant" as not relevant; in NTCIR this choice is called *rigid* relevance.

With our translation models set up for simplified Chinese characters and the documents and queries written using traditional Chinese characters, some approach to character conversion was required. We elected to leave the queries and documents in traditional characters and to convert the translation lexicons (i.e., the Chinese sides of the indexes into the two translation probability matrices) from simplified Chinese characters to traditional Chinese characters. Because the LDC segmenter is lexicon driven and can only generate words in its lexicon, it suffices for our purposes to convert the LDC segmenter's lexicon from simplified to traditional characters. We used an online character conversion tool[3] to perform that conversion. As a side effect, this yielded a one-to-one character conversion table, which we then used to convert each character in the Chinese indexes to our two translation matrices. Of course, in reality a simplified Chinese character might be mapped to different traditional characters in different contexts, but (as is common) the conversion software that we used is not context-sensitive. As a result, this character mapping process is lossy in the sense that it might introduce some infelicitous mismatches. Spot checks indicated the results to be generally reasonable in our opinion, however.

For document processing, we first converted all documents from BIG5

[3]http://www.mandarintools.com/zhcode.html

(their original encoding) to UTF8 (which we used consistently when processing Chinese). We then ran our modified LDC segmenter to identify the terms to be indexed. The TITLE field of each topic was first converted to UTF8 and then segmented in the same way. The retrieval system used for our experiments, the Perl Search Engine (PSE), is a local Perl implementation of the Okapi BM25 ranking function (Robertson and Sparck-Jones, 1997) with provisions for flexible CLIR experiments in a meaning matching framework. For the Okapi parameter settings, we used $k_1 = 1.2$, $b = 0.75$, and $k_3 = 7$, as is common. To guard against incorrect character handling for multi-byte characters by PSE, we rendered each segmented Chinese word (in the documents, in the index to the translation probability tables, and in the queries) as a space-delimited hexadecimal token using ASCII characters.

4.1.3. Retrieval Effectiveness Results

To establish a monolingual baseline for comparison, we first used TITLE queries built from the Chinese topics to perform a monolingual search. The MAP for our monolingual baseline was 0.3077 (which compares favorably to the median MAP for title queries with Chinese documents at NTCIR-5, 0.3069, but which is well below the maximum reported MAP of 0.5047, obtained using overlapping character n-grams rather than word segmentation). We then performed CLIR using each MM variant, sweeping a CDF threshold from 0 to 0.9 in steps of 0.1 and then further incrementing the threshold to 0.99 and (for variants for which MAP values did not decrease by a CDF of 0.99) to 0.999. A CDF threshold of 0 selects only the most probable translation, whereas a CDF threshold of 1 would select all possible translations.

Figure 4 shows the MAP values relative to the monolingual baseline for each MM variant at a set of CDF thresholds selected between 0 and 1. The peak MAP values are between 50% and 73% of the monolingual baseline for all MM variants; all are statistically significantly below the monolingual baseline (by a Wilcoxon signed rank test for paired samples at $p < 0.05$). For the most part the eight results are statistically indistinguishable, although APSQ is statistically significantly below PDT, DAMM, APDT and PAMM-F at each variant's peak MAP. For comparison, the best official English-to-Chinese CLIR runs under comparable conditions achieved 62% of the same team's monolingual baseline (Kishida et al., 2005; Kwok et al., 2005). All four bidirectional MM variants (DAMM, PAMM-E, PAMM-F, and IMM) achieved their peak MAP at a CDF of 0.99, consistent with the optimal

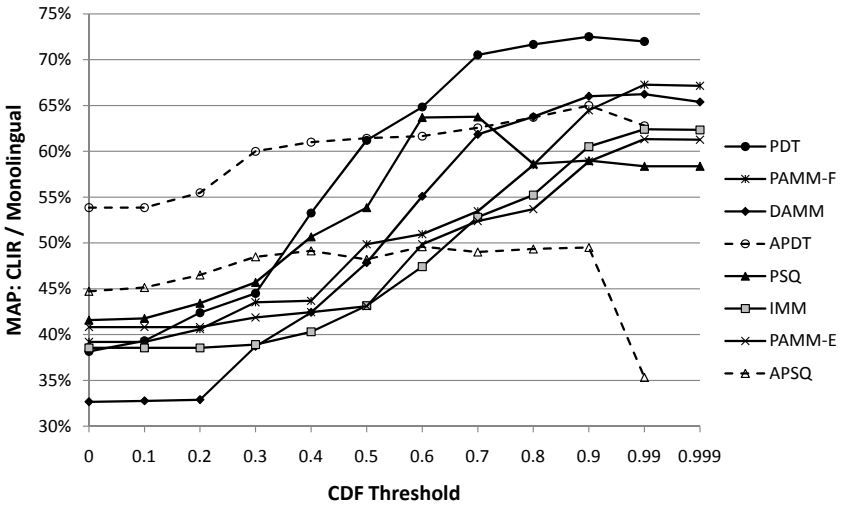


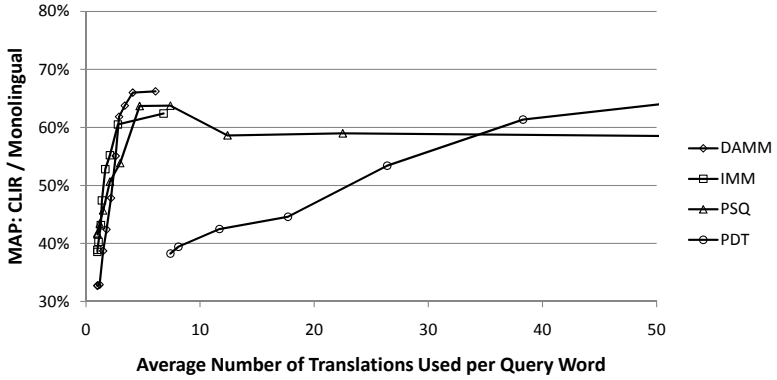
Figure 4: MAP fraction of monolingual baseline, NTCIR-5 English-Chinese collection.

CDF threshold learned in our earlier experiments (Wang, 2005; Wang and Oard, 2006).

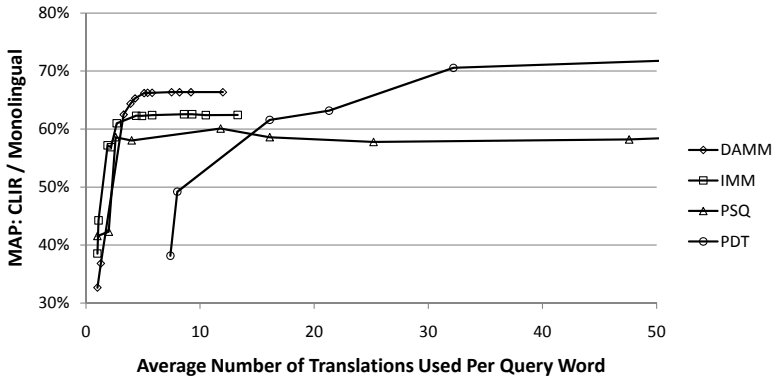
Overall, adding aggregation on the document-language (Chinese) side to bidirectional translation seems to help, as indicated by the substantial increase in peak MAP from IMM to PAMM-F and from PAMM-E to DAMM. By contrast, adding aggregation on the query-language (English) side to bidirectional translation did not help, as shown by the decrease of the best MAP from IMM to PAMM-E and from PAMM-F to DAMM. Comparing PDT with APDT and PSQ with APSQ indicates that applying aggregation with unidirectional translation hurts CLIR effectiveness (at peak thresholds), which is consistent with our previous results on other collections. Surprisingly, PDT yielded substantially (nearly 10%) better MAP than DAMM (although the difference is not statistically significant). As explained below, this seems to be largely due to the fact that PDT does better at retaining some correct (but rare) translations of some important English terms.

4.1.4. Retrieval Efficiency Results

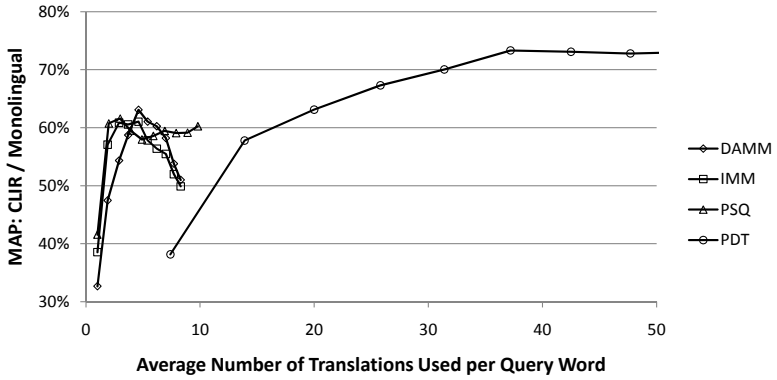
One fact about CLIR that is not remarked on as often as it should be is that increasing the number of translations for a term adversely affects efficiency. If translation is performed at indexing time, the number of disk



(a) Sweeping a CDF threshold.



(b) Sweeping a PMF threshold.



(c) Sweeping a top-n threshold.

Figure 5: MAP fraction of monolingual baseline by the average number of translations used per query term, NTCIR English-Chinese collection.

operations (which dominates the indexing cost) rises with the number of unique terms that must be indexed (Oard and Ertunc, 2002). If translation is instead performed at query time, then the number of disk operations rises with the number of unique terms for which the postings file must be retrieved. Moreover, when some translations are common (i.e., frequently used) terms in the document collection, the postings files can become quite large. As a result, builders of operational systems must balance considerations of effectiveness and efficiency.[4]

Figure 5 shows the effectiveness (vertical axis) vs. efficiency (horizontal axis) tradeoff for four MM variants and three ways of choosing how many translations to include. Figure 5a was created from the same data as Figure 4, sweeping a CDF threshold, but in this case plotting the resulting average number of translations (over all query terms, over all 50 topics) rather than the threshold value. Results for FAMM-F and FAMM-E (not shown) are similar to those for IMM; APSQ and APDT are not included because each yields lower effectiveness than its unaggregated counterpart (PSQ and PDT, respectively).

Three points are immediately apparent from inspection of the figure. First, PSQ seems to be a good choice when only the single most likely translation of each query term is selected (i.e., at a CDF threshold of 0). Second, by the time we get to a CDF threshold that yields an average of three translations DAMM becomes the better choice. This comports well with our intuition, since we would expect that synonymy might initially adversely impact precision, but that our greedy aggregation method's ability to leverage reinforcement could give it a recall advantage as additional translations are added. Third, although PDT does eventually achieve better MAP than DAMM, the consequences for efficiency are very substantial, with PDT first yielding better MAP than DAMM somewhere beyond an average of 40 translations per query term (and, not shown, peaking at an average of 100 translations per query term).

One notable aspect of the PDT results is that, unlike the other cases, the PDT results begin at an average of 8 translations per query term. For DAMM, IMM and PSQ, a CDF threshold of 0 selects only the one most likely

[4]The time required to initially learn translation models from parallel text is also an important efficiency issue, but that cost is independent of the number of terms that require translation.

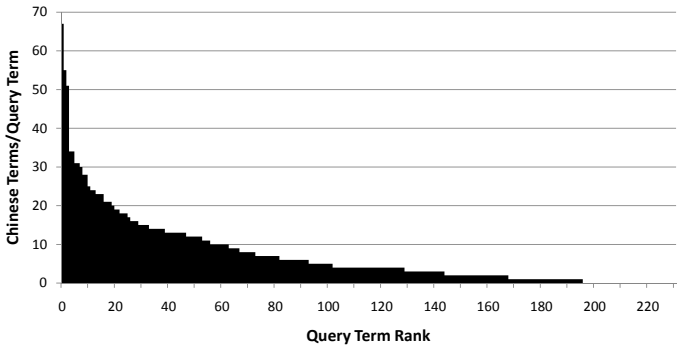


Figure 6: Number of Chinese translations for each English query term, PDT with the most probable translation (at a CDF threshold of 0), NTCIR-5 collection.

translation for each query term. For PDT, by contrast, a CDF threshold of 0 selects only the most likely translation for each *document* term. Because of the relative lack of morphological variation in Chinese, and because our Chinese segmentation cannot generate words that are outside its lexicon, it turns out that the English side of our parallel corpus has roughly twice as many unique terms as the Chinese side. It must, therefore, be the case that half of all English terms have no Chinese translation when we run PDT with a CDF threshold of zero. This turns out not to be the case for query terms, however. As Figure 6 illustrates, 72 of the 130 English query terms have 8 or more Chinese translations for PDT when only the most likely English translation of each Chinese document term is selected. The most extreme of these is the term "time" in the query "time warner american online aol merger impact", which has 67 different Chinese translations with PDT at a CDF threshold of 0. Of course, PDT yield no Chinese translations at all with that threshold for 35 English terms across the 50 queries, notably including "warner". Thus we are searching for "time warner", finding 75 Chinese translations for "time" (many of which have a plausible relation to some meaning of "time") and nothing at all for "warner". Normalizing in the query translation direction (as is the case for PSQ, IMM and DAMM) avoids both problems, and thus is the better choice when seeking to optimize retrieval effectiveness without using very many translations.

Figure 5b shows comparative results for sweeping a PMF threshold. As with the CDF threshold, PSQ is a good choice when 1 translation per query term is desired, DAMM is the better choice by 3 translations per query term

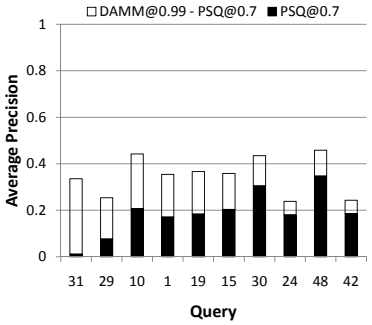
(peaking around 5), and PDT becomes better somewhere much further out (in this case, somewhere after 20, peaking at an average of 57 translations per query term). Notably, PDT exhibits a markedly better effectiveness-efficiency tradeoff with a PMF threshold than with a CDF threshold (PSQ shows the opposite effect; IMM and DAMM are for the most part unaffected).

As Figure 5c shows, DAMM, IMM and PSQ are adversely affected when a fixed top-n threshold is applied to the number of translations, both because of lower peaks (than achieved by the same technique with a CDF threshold) and because of sharper peaks (thus making actual results in operational settings more sensitive to parameter tuning). PDT, by contrast, does about as well with a top-n threshold as with a PMF threshold (rising about as rapidly to peak at 73% of monolingual MAP, compared with 75% for a PMF threshold), and using of a constant number of translations may have some modest benefits for storage management.

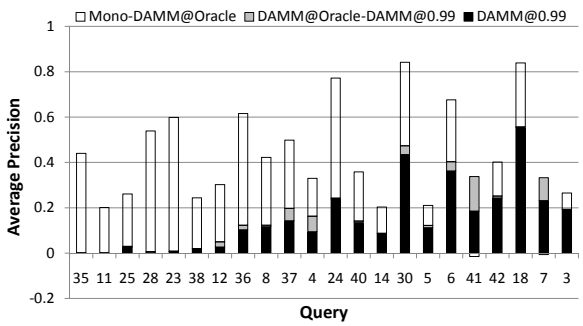
It is important to recognize that the differences we are seeing here are not statistically significant, so it is the broader trends on which we must focus. From these results we can reasonably conclude that CDF and PMF thresholds are both good choices over a broad range of effectiveness-efficiency tradeoffs, and that a PMF or a top-n threshold may be a reasonable choice if optimizing effectiveness regardless of the computational cost is the goal. This comports well with our intuition that the information provided by the probability distributions can provide a useful degree of query-specific tuning when the choice involves relatively common events, but that empirical statistics for uncommon events (which is what we must work with when the number of translations becomes very large) are generally not as reliable.

4.1.5. Query-Specific Analysis

Taking the mean of the AP for the 50 queries in our test collection is useful when seeking to characterize expected retrieval effectiveness for the as-yet-unseen 51st query that what we really care about, but as we saw with "time warner" when we are seeking to understand why one technique works better than another it can also be useful to look at what happens in specific cases. For the analysis in this section we chose the run with the peak MAP (obtained by sweeping a CDF threshold) for each of two techniques, which we generically refer to as Run A and Run B. We remove any topics for which neither Run A nor Run B achieved an AP of 0.2 or better (in order to avoid focusing on small differences between bad and worse), we then compute a *relative AP difference* (rAPd), defined as $(AP_A - AP_B)/AP_B$ on a topic-by-



(a) Comparing DAMM to PSQ. Black: MAP for PSQ at a CDF threshold of 0.7, White: Additional MAP for DAMM at a CDF threshold of 0.99



(b) Comparing monolingual to DAMM. Black: MAP for DAMM at a CDF threshold of 0.99, Gray: Additional MAP for DAMM computed with the best AP of each query, White: Further additional MAP for monolingual.

Figure 7: Queries with marked improvements, NTCIR-5 English-Chinese collection.

topic basis, and we finally divide the topics into three groups: (1) the rAPd is markedly better (at least +20%), (2) the rAPd is markedly worse (at least -20%), and the remainder (which we normally ignore) in which the AP is little changed.

As Figure 7a shows, there were 10 topics for which DAMM was markedly better than PSQ and 5 topics (not shown) for which DAMM was markedly worse. Table 3 shows query terms from several queries that (in our opinion) have substantial probability mass assigned to incorrect translations (for space reasons, translations with very low probabilities are not shown). As expected, DAMM (the upper row of translations for each term) often produces sharper distributions that emphasize better translations. Of course, a translation need not be *correct* to be *useful* for CLIR, and several of the "translations" shown in Table 3a are simply words that co-occur frequently with correct translations. Nonetheless, many of the additional translations to which PSQ (the lower row) assigns substantial probability mass are both incorrect and (in our opinion) unlikely to be helpful. We have indicated those that we feel could be useful using bold, and those that we expect would adversely affect retrieval effectiveness using italics.

For three of the five cases in which DAMM was markedly below PSQ (Queries 23, 47, and 7 in Table 3b), DAMM seems to be *overtuned* to the domain of the parallel text collection. For example, it may be the case that

Query 31: fine dust particles heart disease	
particles	**粒子/0.5207 微粒/0.4015 顆粒/0.0725**
	粒子/0.1669 顆粒/0.0557 微粒/0.0524 粒/0.0453 物質/0.0438 ...
heart	**心臟病/0.3749 心臟/0.2465 心中/0.231 心/0.0834 心裡/0.0592**
	心臟病/0.2887 患/0.2642 有/0.246
Query 29: alternative energy air pollution electricity	
alternative	**替代/0.5539** 另/0.1813 無如/0.0675 其他/0.0461 不得已/0.0279 ...
	另/0.2468 類/0.1388 其他/0.122 選擇/0.0621 **替代/0.039** ...
Query 10: anthrax bacteria war terrorist attack	
anthrax	疽/0.5947 **炭疽病/0.4053**
	炭/0.1539 疽/0.1532 **炭疽病/0.1251** 菌/0.0866 熱/0.0577...
terrorist	**恐怖/0.6445** 恐/0.3502
	恐怖/0.5684 分子/0.1015 份子/0.1005
Query 1: time warner american online aol merger impact	
online	**網絡/0.6654 網/0.3338**
	網/0.416 上/0.3703 ...
Query 19: supersonic airliner concord airplane crash	
concord	**和諧/0.5097** 和睦/0.218 康/0.2165 式/0.0268 苑/0.0118 ...
	式/0.2127 康/0.1788 和/0.1669 **和諧/0.065** 苑/0.0236 屋/0.0219 ...
Query 15: ep surveillance aircraft fighter aircraft collision	
surveillance	**偵察機/0.2279** 監測/0.2138 監察/0.1965 監管/0.171 監視/0.1373 ...
	監察/0.3225 監視/0.1336 **偵察機/0.1034 偵察/0.0828** 監/0.0415 ...
collision	事件/0.2649 **撞/0.2331 相撞/0.1837** 軍機/0.1479 **碰撞/0.104** ...
	撞/0.301 機/0.2801 事件/0.1118 **相撞/0.0638**

(a) DAMM is better than PSQ.

Query 23: space station mir disposal storage waste	
mir	*中大/0.4837 退休/0.3224 壽命/0.1939*
	"/0.1819 號/0.1727 **和平/0.1445** *退休/0.0304 補償/0.0279* ...
Query 47: korean general election 2000 han nara party	
korean	*半島/0.1612* **南韓/0.1604 韓國/0.1559** 韓/0.1509 **朝鮮/0.1478** ...
	朝鮮/0.5192 韓國/0.2003 ...
general	*秘書長/0.1361 總書記/0.1322* 一般/0.1287 *市民/0.1275* 綜合/0.123 ...
	一般/0.5161 總/0.1442 *秘書長/0.0313* 綜合/0.0294
Query 7: wen ho lee case classified information national security	
wen	溫/0.5455 **文/0.4361** *家寶/0.0138*
	文/0.5993 溫/0.253

(b) PSQ is better than DAMM

Table 3: Translations of selected query terms for DAMM (upper) and PSQ (lower) at the optimal CDF threshold for each. **Bold** indicates useful (in our opinion); *italics* indicate harmful (in our opinion).

many of the sentences in the parallel text collection that contain the word "general" are talking about attorneys general, military generals, or the U.N. Secretary General. If so, bidirectional translation would tend to reinforce those meanings of "general" to the detriment of the sense needed for "Korean general election" in Query 47. In future work we might be able to mitigate this overtuning to some extent either by using factored translation models (e.g., incorporating part-of-speech evidence) (Koehn and Hoang, 2007) or by using broader context in some way (e.g., with phrase translation models).

With a similar analysis (not shown), we found that PDT did markedly better than IMM for 10 queries and markedly worse for 2 queries. Explaining why that happened is harder in this case, however, since (at peak) PDT queries typically use an enormous number of translations. Examining only the most likely translations for each query term in the 10 queries for which PDT achieved a marked improvement, we found just two English query terms for which PDT assigned a much higher probability to a good translation than IMM did: "operation" in Query 36 ("remote operation robot"), and "class" in Query 24 ("economy class syndrome flight"). This paucity of evidence suggests that other as-yet-uncharacterized effects must be responsible for the majority of the benefit that we see from PDT at high CDF thresholds.

Figure 7b shows the per-topic AP for the 22 queries in which the monolingual condition (i.e., Chinese queries) yielded a markedly higher AP than DAMM (there were also 8 cases in which DAMM was markedly better than the monolingual baseline). In aggregate, the MAP over those 22 queries is only 34% of the monolingual baseline, so those 22 queries account for almost all of the observed MAP difference between the monolingual baseline and DAMM over the full 50-query set. For each of those 22 queries we inspected the DAMM translations for each query term and identified the following factors that in our opinion had likely degraded CLIR effectiveness:

- *Incompatible tokenization.* Alternative ways of tokenizing text are sometimes plausible, and this effect is particularly notable for languages such as Chinese in which word boundaries are not marked when writing. For example, both "Kim Dae Jun" and "Kim Jong Il" in Query 3 correspond to three-character person names on the Chinese side of the parallel corpus. These names were both correctly segmented as three-character terms. On the English side, each space-delimited word was tokenized, also correctly, resulting in three terms in each case (one for each token in the name). This created some problems because "Jun" is

Query 35: capital punishment survey data	
capital	資本/0.1673 基本/0.1671 資產/0.1636 資金/0.1508 首都/0.1282 外資/0.121 資/0.0846 經常/0.0059 投資/0.0033
Query 11: ichiro rookie major league	
ichiro	郎/1
Query 12: jennifer capriati tennis	
capriati	莉/0.7869 休止/0.1574 樂於/0.0542
Query 25: tiger woods sports star	
woods	林子/0.3539 樹叢/0.297 樹林/0.2961 茲/0.025 森林/0.0193
Query 36: remote operation robot	
operation	作戰/0.2629 經營/0.2208 行動/0.2163 運作/0.2001 運行/0.0294 操作 /0.0245 營/0.0185 運/0.0139 開刀/0.0019 投入/0.0019
Query 24: economy class syndrome flight	
class	級/0.1937 階級/0.1882 課程/0.1787 中產階級/0.0889 班/0.0875 上課/0.0704 甲級/0.0617 課/0.0545 一流/0.0429 類別/0.0065 階層/0.0043 教室/0.0038 同學/0.0034 課堂/0.003 班上/0.0021 工人/0.0016
Query 5: g8 okinawa summit	
g8	工業國/0.8295 g/0.1547 丨/0.0107

Table 4: Examples of the domain difference effect with DAMM.

a common abbreviation in English that was aligned with the Chinese term for the month "June" more often than with the Chinese term for "Kim Dae Jun" and because "Il" was more often aligned with the Chinese term for the name "Kim Il Sung" (the father of Kim Jong Il). The resulting translation probability distributions were clearly suboptimal for this query. Similar problems arose in Queries 7, 11, 14, and 38.

- *Differing transliteration conventions.* Terms that are transliterated differently in the collection and in the parallel text can cause problems. For this Chinese test collection, all four cases involved proper names: "Kursk," "Greenspan," "Dennis" and "Tito" in Queries 6, 42, 30 and 30, respectively.

- *English vocabulary gaps*: Some words not covered by the lexicon because they are not present sufficiently often in the parallel corpus. In all three cases these were proper names "Bubka," "Maru," and "iloveyou" (the name of a virus) in Queries 23, 8, and 37, respectively. The effect on these queries was rather severe (all are on the left side of the Figure 7b, which is ordered by decreasing relative advantage of monolingual).

- *Domain differences.* All of the other English query terms have one or more translations, but in some cases one or more appropriate Chinese translations of an English query term are simply not common enough in the parallel text to result in the right translations receiving much probability mass. Table 4 shows the queries containing these words and their translations. In our opinion, very few of the translations that are receiving probability mass from DAMM in these cases would be helpful.

Nearly half (10 of 22) of the queries in Figure 7b contain names of persons, and most of the rest (7 of 22) contain proper names of objects or organizations. Together, these queries that were difficult for DAMM account for most of the queries in the 50-query NTCIR-5 Chinese test collection that contain proper names (10 of 13 person names, 17 of 24 total proper names). For comparison, only 4 of of the 25 topics in the TREC-9 Chinese collection contain organization or location names, and no person's name appears anywhere in those 25 topics. Similarly, the 54 topics of the TREC-5&6 Chinese test collection include only 2 topics that contain organization or location names, and none that contains person names. We often think of the purpose of replicating experiments on a different test collection as seeing what will happen with a new set of documents. As these statistics clearly indicate, however, seeing what happens with queries that are constructed in a different way can be equally important.

We can also use Figure 7b to see whether topic-specific CDF thresholds might yield further improvement. Although it is not clear how topic-specific thresholds would best be chosen, it is straightforward to bound the potential improvement by using post-hoc optimal topic-specific thresholds as an oracle. Focusing on the same 22 queries on which the monolingual baseline obtained a markedly higher AP than DAMM, our oracle found that higher AP could be obtained at some CDF threshold below 0.99 for 19 of those 22 queries and that a CDF threshold above 0.99 was optimal for the remaining 3. In other words, although a CDF threshold of 0.99 was optimal when averaging over all 50 queries, it was not optimal for *any* of the 22 queries that we have identified as having the greatest potential for improvement! As the few middle (gray) bars in Figure 7b indicate, the magnitude of the potential gain in AP is in most cases very small. For queries 7 and 41 the relative improvement in DAMM is quite large, however (even slightly exceeding monolingual MAP for topic 41). On average over all 50 topics, AP from DAMM rises from 66% of

monolingual AP with a constant (0.99) threshold to 72% of monolingual AP with oracle topic-specific thresholds. From this we conclude that although CDF (and PMF) thresholds provide some degree of topic-specific behavior, further work on topic-tuned thresholds might be worthwhile.

Turning our attention to the 8 queries for which DAMM achieved a markedly higher AP than the monolingual condition (not shown) we see that as we would expect the right translations are being used. Specifically, for every English term in all 8 of those queries the corresponding term from the Chinese query appeared among the DAMM translations, often with the highest translation probability. The additional benefit comes from the frequent presence of synonyms among the Chinese translations that (in our opinion) would provide a useful query expansion effect. In some sense this makes the MAP obtained using unexpanded monolingual queries an artificially low point of comparison, but we prefer an unexpanded monolingual reference because monolingual query expansion would introduce an additional source of variance (potentially harming some queries and helping others).

We expected that our decision in these experiments to threshold the two translation probability tables at a CDF of 0.99 before the DAMM computation would result in less sensitivity to specific threshold choices near the optimal value of 0.99. To check this, we looked across all 50 queries, finding only 12 that achieved their maximum AP at a CDF threshold at or above 0.99. Omitting 3 that yielded very low DAMM AP (peaking at 0.0003) and 1 that yielded very low monolingual AP (0.0002), Figure 8 shows the ratio of DAMM AP to monolingual AP for the remaining 8 queries. Except for Query 41, the AP of these 8 queries is generally quite stable above a CDF threshold of 0.9. We therefore would not expect to see further improvements from exploring the region between 0.9 and 1.0 with greater granularity.

These English-Chinese experiments confirm what we have seen before: DAMM is generally a good choice. Importantly, we have now seen that on a larger test collection (although we should note that the NTCIR-5 collection that we used is still far smaller than the Web-scale test collections that are now commonly used in monolingual experiments). Among our new contributions are detailed analysis of effectiveness-efficiency tradeoffs, better understanding of why DAMM, IMM, PSQ and PDF behave as they do, and obtaining some indication that some further gain might be obtained from query-specific CDF thresholds. Along the way we have seen that the same CDF threshold (0.99) is optimal across for DAMM across several test collections, at least in part because initial pruning of the raw translation probabil-

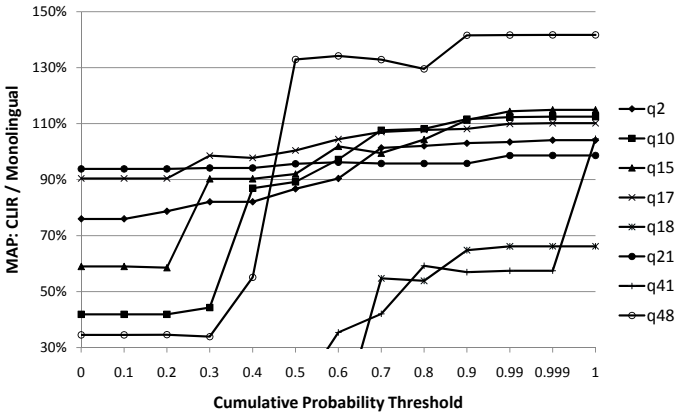


Figure 8: Queries with peak DAMM AP at CDF threshold of 0.99 or greater, NTCIR-5 English-Chinese collection.

ity tables yields stable performance for individual topics at higher threshold values. Our analysis also illuminates some issues that are of broader interest in CLIR research generally, most notably consequential differences in the prevalence of named entities among the queries when the two languages use different character sets.

4.2. New French Experiments

Our Chinese experiments have told us what happens with larger test collections, but we don't yet know what we will see when we look beyond news to other types of content. In this section we therefore report on another new set of experiments, this time with the Cross-language Speech Retrieval (CL-SR) collection of the 2006 Cross-language Evaluation Forum (CLEF). Queries written in French were used to retrieve manually partitioned segments of English interviews. The collection includes two parallel representations of the same content, one which we call the "manual" condition that was prepared entirely by subject matter experts, and the other that we call "automatic" condition that was produced with an Automatic Speech Recognition (ASR) system that had been tuned (on held-out data) to optimize recognition accuracy for the accented, elderly, domain-specific speech content of the interviews (which were conducted with survivors of, witnesses to, or rescuers during the Holocaust).

4.2.1. Training Statistical Translation Models

For comparability with our earlier work, we used the same models for the French-English language pair as in (Wang and Oard, 2006), although in this case the query and document languages are reversed (in our earlier experiment, we had used English queries and French documents). Recapping briefly, we derived word-to-word translation models from the European Parliament Proceedings Parallel Corpus, known as *Europarl Corpus* (Koehn, 2005),[5] using the freely available GIZA++ toolkit. Before word alignment we stripped accents from the French documents and we filtered implausible sentence alignments by eliminating sentence pairs that had a token ratio either smaller than 0.2 or larger than 5. GIZA++ was then run twice on the remaining 672,247 sentence pairs, first with English as the source language and subsequently with French as the source language. When training translation models, we started with five Hidden Markov Model (HMM) iterations, followed by ten IBM Model 1 iterations, and ended with five IBM Model 4 iterations. The result of this process was two translation tables, one from English words to French words and the other from French words to English words. In contrast with our Chinese experiments, all nonzero values produced by GIZA++ were retained in each table.

4.2.2. Test Collection

The "documents" in the CLEF 2006 CL-SR test collection correspond to 8,104 manually partitioned intervals (called "segments," which average about 4 minutes duration) from 272 interviews that were collected and indexed by the Survivors of the Shoah Visual History Foundation (now the University of Southern California Shoah Foundation Institute for Visual History and Education) (Oard et al., 2004, 2007). Three types of metadata were created for each segment as a part of the indexing process: the names of mentioned people were recorded (regardless of whether the person was actually mentioned by name), some thesaurus keywords were assigned, and a somewhat stylized three-sentence summary was written that focused on providing "who, what, when, where" information (and that was originally intended to be displayed with search results to support segment-level selection). We used the terms in these fields (formally, the NAME, MANUALKEYWORD and SUMMARY fields of the CLEF 2006 CL-SR collection, respectively) for the document rep-

[5]The Europarl corpus is available at: http://www.statmt.org/europarl/.

resentation that we call "manual" in this article. No distinction was made between the fields for this purpose; all were tokenized in the same way, and indexed together.

For our "automatic" representation, ASR was used to generate a (potentially erroneous) one-best word transcript of what had been said in the interview. The ASR process (with a 38% measured word error rate on held out data) was optimized for the interviewee rather than the interviewer (by automatically detecting and then consistently using only the interviewee's microphone) and was trained using 200 hours of in-domain held out data along with other other standard ASR training resources (Byrne et al., 2004). This resulted in the text contained in the ASRTEXT2004A field of each segment in the test collection. In an effort to automatically capture some knowledge from the manual indexing process, these words were used as the feature set for two k-Nearest-Neighbor (kNN) classifiers trained (using cross-validation) to approximate human assignment of thesaurus terms. The resulting kNN keywords were contained in the AUTOKEYWORD2004A1 field and the AUTOKEYWORD2004A2 field, respectively. A second ASR transcript (with a measured word error rate of 25% on held out data) that had been trained similarly to the first system, but using a later generation of ASR models, was also used to create the ASRTEXT2006A1 field (although that system was not used by the classifiers that automatically assigned thesaurus terms). We tokenized these four fields consistently and indexed the resulting terms together.

The Porter stemmer was used to stem the English collections and the English side of the translation probability matrix. The 33 French evaluation topics in the CLEF CL-SR 2006 test collection were created initially in English and then translated into French by bilingual speakers. We used the TITLE field (2-3 words) and the DESCRIPTION field (typically, one sentence) together as unstructured (i.e., bag of words) queries. Binary relevance judgments (relevant or not relevant) for segments that have previously been judged by subject matter experts are distributed with the test collection. The Perl Search Engine (PSE) with the same settings as in our NTCIR-5 experiments was used for retrieval.

4.2.3. Results

To facilitate cross-collection comparisons, we again report the fraction of monolingual MAP achieved by each CLIR system (monolingual MAP is 0.0466 for the "automatic" condition and 0.2300 for the "manual" condition,

Collection	NTCIR-5		CL-SR Automatic		CL-SR Manual	
Queries in	English		French		French	
Documents in	Chinese news		French ASR		French metadata	
Total topics	50		33		33	
Total docs.	901,446		8,104		8,104	
	MAP % of DAMM	MAP % of Mono	MAP % of DAMM	MAP % of Mono	MAP % of DAMM	MAP % of Mono
DAMM	- -	**66.2%**	- -	83.5%	- -	**75.6%**
PAMM-F	101.6%	**67.3%**	85.9%	71.7%	94.1%	**71.1%**
PAMM-E	92.6%	**61.3%**	89.9%	74.2%	96.3%	**72.7%**
IMM	94.1%	**62.3%**	92.5%	77.3%	94.5%	**71.6%**
PDT	109.5%	**72.5%**	102.1%	85.2%	106.2%	**80.1%**
APDT	98.1%	**65.0%**	89.2%	74.5%	99.9%	**75.6%**
PSQ	96.3%	**63.8%**	85.7%	71.7%	96.0%	**72.5%**
APSQ	**74.9%**	**49.6%**	98.7%	82.4%	98.1%	**74.1%**

Table 5: The best retrieval effectiveness of meaning matching variants in new experiments ("Mono" is the monolingual baseline, **bold** indicates a statistically significant difference.)

respectively).[6] Table 5 shows the MAP of each MM variant at the best CDF threshold for the "automatic" and "manual" coditions. For comparison, we also include in the table the results reported in Section 4.1. For the automatic condition, the MAP of these MM variants ranges between 72% and 85% of monolingual MAP, all at a CDF threshold of 0.99. Wilcoxon signed rank tests for paired samples (at $p < 0.05$) found no significant differences among these results, and moreover found that at the peak CDF threshold each technique is statistically indistinguishable from the monolingual baseline. For the manual condition, the best MAP of each MM variant ranges between 71% and 80% of monolingual MAP, all except PSQ at a CDF threshold of 0.99 (PSQ peaked at 0.9). No statistically significant differences were found between the cross-language results at peak, but at peak each was statistically significantly below the monolingual baseline. The relatively small number of queries (33) may have contributed to this failure to observe significant differences among MM variants. Studies of text retrieval (Voorhees, 2000; Sanderson and Zobel,

[6]For comparison, the best reported monolingual MAP at CLEF-2006 with the same queries and indexed fields was 0.0565 for our automatic condition and 0.2350 for our manual condition.

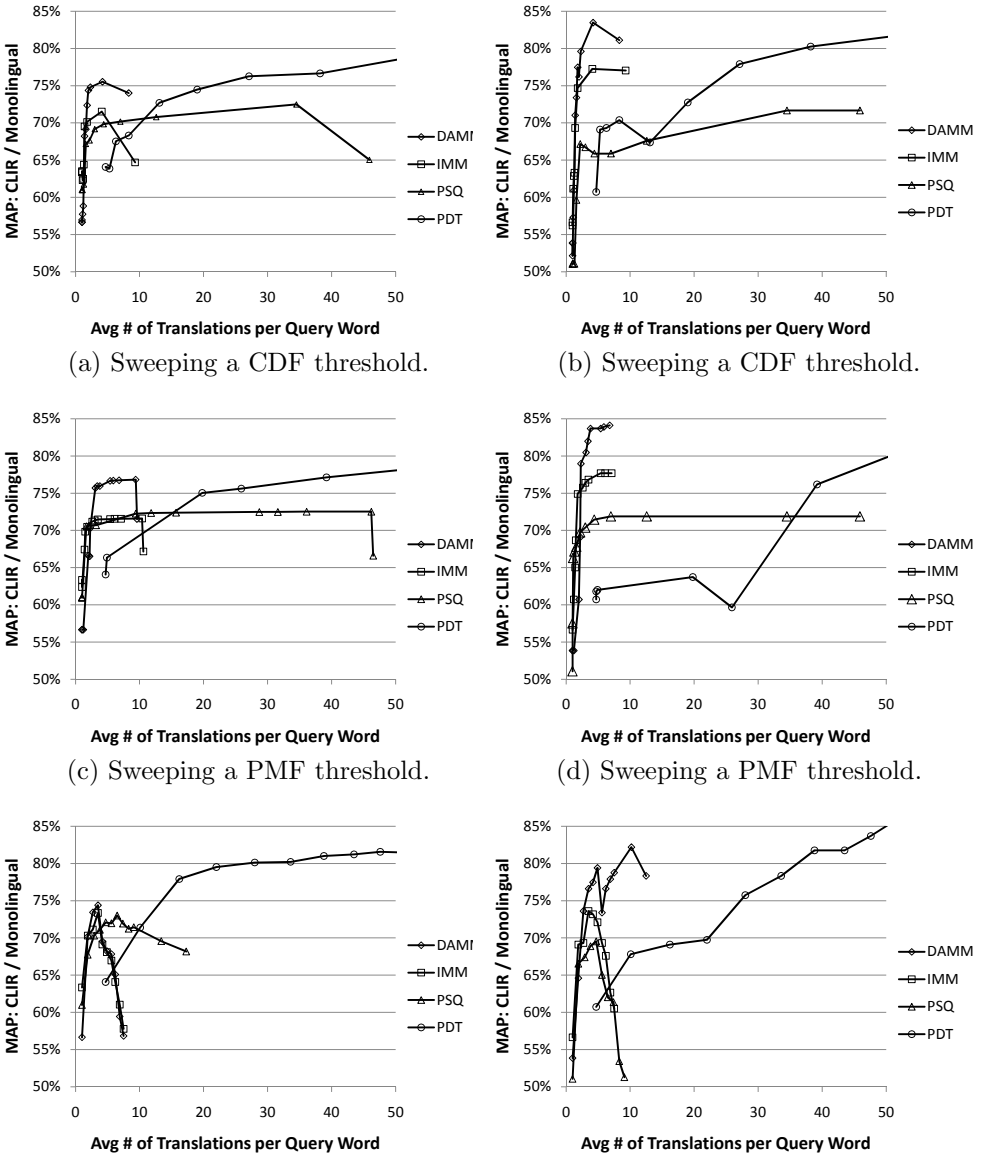
2005, e.g.) have shown that relative effectiveness comparisons become more stable as more queries are added, and that more than 40 queries are typically needed to reliably make relatively fine-grained comparisons.

Figure 9 shows the effectiveness-efficiency tradeoff for sweeping CDF, PMF, and top-n thresholds, focusing again on DAMM, IMM, PDT, and PSQ. For both the "manual" condition (Figures 9a, 9c, and 9e) and the "automatic" condition (Figures 9b, 9d, and 9f), the PMF threshold seems to be a good choice for DAMM, both because DAMM seems to peak at a somewhat higher MAP and because that peak is more robust to differences in the resulting average number of translations. A top-n threshold seems like a good choice for PDT, which in this case starts outperforming the best DAMM results (DAMM with a PMF threshold) at an average of around 15 translations per query term in the "manual" condition and around 45 translations per query term in the "automatic" condition, respectively. In contrast with the pattern seen for Chinese, when only one translation is used IMM now looks to be a better choice than PSQ for both the manual and the automatic conditions. For Chinese, IMM had not been far below PSQ for top-1 translation, so looking back over both sets of experiments, we can therefore recommend IMM (or perhaps PSQ) for one-best translation, DAMM when it is possible to use 3-5 translations, and PDT when efficiency is not a limiting factor.

4.2.4. Query-Specific Analysis

As in our Chinese results above, we identified the topics whose peak DAMM Average Precision (AP) in the manual condition was substantially lower than the corresponding monolingual AP for the same topic, and then looked for factors that might explain the difference. There were four such topics, from which we identified the following factors:

- *Normalization error: Topic 3005* (English: *Death marches; Experiences on the death marches conducted by the SS to evacuate the concentration camps as the allied armies approached.* French: *Les marches de la mort; Expériences concernant ces marches de la mort conduites par les SS afin d'évacuer les camps de concentration pour chapper aux armées alliées qui approchaient.*): According to our meaning matching lexicon, the French word "marches" has only one English synonym "markets" (with a probability of 1). Clearly this is wrong, since a better translation would be "marches." This error was almost certainly caused



(a) Sweeping a CDF threshold.

(b) Sweeping a CDF threshold.

(c) Sweeping a PMF threshold.

(d) Sweeping a PMF threshold.

(e) Sweeping a top-n threshold.

(f) Sweeping a top-n threshold.

Figure 9: MAP fraction of monolingual baseline by the average number of translations used per query term, CL-SR French-English collection. The three figures on the left side correspond to the "manual" condition; the three on the right side correspond to the "automatic" condition.

by stripping French accents before training the statistical translation models, since the French word "marché" can properly be translated as "market").

- *Incompatible tokenization: Topic 1325* (English: *Liberators of Concentration Camps; African-American Liberators of Concentration Camps.* French: *Les libérateurs des Camps de Concentration; La libération des camps de concentration par les afro-américains.*): The English term "African-American" was expressed as "afro-américans" in the French query. After tokenization this became "afro" and "american" which wound up in different synonym sets in English, yielding poor results for DAMM.

- *Differing transliteration conventions: Topic 3013* (English: *Yeshiva in Poland; Accounts regarding Poland's Pre-war Yeshiva and its influence on its graduates and their descendants.* French: *Yéshiva en Pologne; Récits évoquant la Yéshiva de la Pologne d'avant-guerre, ainsi que son influence sur ses diplômés et leurs descendants.*): Our meaning matching lexicon mapped French word "yéshiva" to the English word "jeshiwa" with a probability of 1, presumably because "Yeshiva" rather than "Jeshiwa" was used by convention in the European Parliament parallel corpus from which we learned our translation model.

- *English Vocabulary gap: Topic 3031* (English: *Activities in US DP camps; Religious and cultural activities and observances in the American DP camps, especially involving US Jewish chaplains.* French: *Activités au sein des camps de déportés américains; Activités et pratiques culturelles et religieuses dans les camps de déportés américains, plus particuliérement, impliquant les aumôneries juives américaines.*): The French word "déportés" was (correctly) translated by DAMM to the English stem "deport," in part because the term "DP" does not occur on the English side of the parallel text collection. However, the interviewers and interviewees consistently spoke of "DP camps" in this context, and used "deport" and "departation" mostly in other contexts. As a result, the English query did quite well with the manually created metadata (reliably finding the term "DP"), while the French query yielded poor results.

That these are similar to the types of errors seen in Chinese suggests that error patterns do vary from one test collection to another, but that the

Collection	CLEF (01-03)	CL-SR manual	CL-SR auto	Avg. E-F	TREC (5&6)	TREC (9)	NTCIR (5)	Avg. E-C
Queries	English	French	French		English	English	English	
Documents	French news	English metadata	English ASR		Chinese news	Chinese news	Chinese news	
Topics	151	33	33		54	25	50	
Documents	87,191	8,104	8,104		164,789	126,937	901,446	
PDT/DAMM	96.3%	106.2%	102.1%	102%	89.9%	98.1%	109.5%	99%
PAMM-F/DAMM	99.7%	94.1%	85.9%	96%	100.0%	96.2%	101.6%	99%
APDT/DAMM	**92.5%**	99.9%	89.2%	94%	98.7%	**88.5%**	98.1%	95%
PAMM-E/DAMM	99.7%	96.3%	89.9%	95%	94.9%	91.4%	92.6%	93%
IMM/DAMM	97.2%	94.5%	92.5%	95%	92.1%	**87.9%**	94.1%	91%
PSQ/DAMM	**94.6%**	96.0%	85.7%	92%	**83.7%**	**90.4%**	96.3%	90%
APSQ/DAMM	**82.3%**	98.1%	98.7%	93%	**56.6%**	**49.7%**	**74.9%**	60%
DAMM/Mono	100.3%	**75.6%**	82.4%	86%	97.8%	128.2%	**66.2%**	97%

Table 6: Summarizing six sets of meaning matching experiments ("Mono" is the monolingual baseline, **bold** indicates a statistically significant difference.)

nature of the content is (at least in this case) no greater a factor than the nature of the queries (as we saw with the names in the NTCIR-5 Chinese test collection) or the nature of the languages involved (e.g., as we have seen with the differing transliterating conventions in both experiments). Moreover, as we would expect, none of these error types are specific to meaning matching; all are well known in CLIR research generally. From this we can conclude that our implementations of bidirectional translation and synonymy have performed as we would expect given the design decisions that we have made.

4.3. Comparing Six Experiments

Table 6 compares the results from the six experiments that we have conducted to date, including our three previous experiments from Table 2 (Wang, 2005; Wang and Oard, 2006) and our three new experiments from Table 5.[7] Each of the three experiments involving French used the same translation models (i.e., trained in the same way on the same training data), as did each of the three experiments involving Chinese experiments.

The rightmost column in Table 6 provides an easy way of discerning general trends, so Table 6 is sorted by those macroaveraged values. Averaged over the six experiments, DAMM and PDT yield quite similar results at peak, with none of the observed differences in individual experiments being

[7]One other very preliminary experiment in Xu and Oard (2008) using English queries and Hindi documents is not included in this analysis.

statistically significant; the choice between these two techniques thus turns more on efficiency (where DAMM has a decided advantage) than effectiveness. Other meaning matching methods yield reasonable results on both languages, with the exception of APSQ which does very poorly on all three Chinese test collections. Indeed, with the exception of APSQ the average results are strikingly similar, both across languages and across techniques. Differences in specific cases should of course be interpreted with caution, but the consistent pattern of improvement over a broad range of conditions (two quite different language pairs, two types of sources (news and conversations), six collections, and over 300 topics) gives credibility to a broad conclusion that the choice to be made is between DAMM and PDT.

5. Conclusion

Excellent results have been reported for meaning matching in prior work (Wang, 2005; Wang and Oard, 2006), but the evaluation framework that we use for information retrieval experimentation relies on replication to achieve high confidence in the results. In "batch" experiments of the type reported in this article we always replicate across queries, but it is also important (and far more time consuming) to replicate across collections. Indeed, our experiments in this paper showed that PDT can be quite effective at peak, which had not been apparent in our earlier work.[8] The analysis of batch experiments is often sharply focused on effectiveness; we have sought to balance that with some attention to efficiency issues as well, for which the average number of translations per query term provides a useful proxy. Finally, we have looked in some detail at what is happening with individual queries, thus better understanding the effects of specific design decisions, and thus which parts of what we are seeing are fundamental to the methods we have tested and which are incidental to the way we have implemented our experiments.

As for what we might grandly call "theory," we have presented the derivation of meaning matching in greater detail, thus highlighting the key role of normalization. This led to a new discussion of the tension between normalizing in the document translation direction, which we see as favorable for principled mapping of TF and normalizing in the query translation direction, which we see as favorable for principled mapping of DF. For the experiments

[8] We have carefully checked and repeated our earlier experiments to confirm that we had not inadvertently misreported PDT results earlier.

in this paper we simply chose one direction and used it consistently (although in some side experiments we did confirm that choosing the other direction was not clearly better, at least in the cases we looked at). But this tension clearly points to a need for further work, as we describe below.

Our experiments have led us to formulate some guidelines for practice as well. The use of top-1 translation is today quite common (often implemented by simply using Google Translate), and our results suggest that under those conditions translating the queries is a better choice than translating the documents because that choice ensures that every query term will have an opportunity to influence the results. This comports well with present practice, which (at least in experimental settings) favors query translation. When access to the internals of the translation system is possible, our more complex DAMM approach offers some potential to leverage translation probability tables in ways that can yield better retrieval effectiveness than one-best translation, although at the expense of somewhat greater disk activity (because of using 3-5 times as many translations per query term). When efficiency is not a factor, we have also now seen cases in which the probability-weighted translation of document term counts that has been widely used with language modeling techniques for information retrieval can yield excellent effectiveness results. By also characterizing the implications of that approach for efficiency, we have illuminated a tradeoff that must be considered by designers of operational systems. Future work on effectiveness-efficiency tradeoffs for specific settings should, of course, also consider the amortized costs of constructing the translation model.

Our analysis results may also be of broader interest to CLIR researchers. For example, our observations on the relative predominance of named entities in different Chinese test collections may help others to identify which test collections are most suitable to the research questions they wish to explore. We also saw some evidence of the domain difference effects between the parallel texts on which our translation models were trained and the texts in our information retrieval test collections. Although this is a well known effect among machine translation researchers, we believe that it is an understudied issue in cross-language information retrieval research. Some previous studies have demonstrated substantial effects from lexicon size on retrieval effectiveness (McNamee et al., 2009; Demner-Fushman and Oard, 2003, e.g.), but we are not aware of comparable studies in which the effect of parallel corpus domain has been well characterized in CLIR experiments. There is now a considerable amount of work underway on what is broadly called "domain

adaptation" in which the statistics learned from corpora that are not quite "close enough" are adjusted in some way to make them "closer"—as that work matures, it may be possible to begin to apply the resulting techniques in a CLIR setting (Daumé, 2007).

The most important new research question that we now need to grapple with is how best to address normalization. The issue arises from the assumption in our derivation of meaning matching that every possible shared meaning has an equal chance of being expressed (i.e., that $p(m_i)$ in Equation 6 is a constant). In reality, of course, some meanings will be more often expressed than others. One approach to avoiding the normalization issue would be to find some reasonable way of estimating $p(m_i)$. Alternatively, we might try the rather obvious expedient of simply normalizing in different directions for TF and DF. The competing advantages of DAMM and PDT (at least on some test collections) suggest that these could be productive lines of inquiry.

The other place where normalization arises in our work is in our implementation of aggregation. The whole point of aggregation is to treat equivalent translations equivalently, and our term-oriented normalization indeed accomplishes that while permitting us to retain a term-oriented architecture. But the cost of that design is some reduction in fidelity when compared to what we actually wish to model, which is that sets of translations are (for our purposes) interchangeable. Modeling that situation more faithfully would require an architecture in which we index synonym sets that have potentially overlapping elements. That can be done, but at some cost in implementation complexity that would have added complexity to the comparisons between techniques that we have been able to make using a simpler term-oriented model.

Although not a focus of our work in this article, one clear implication of the meaning matching framework is an interaction between segmentation granularity, translation ambiguity, and recall-preserving generality that has not yet been well characterized. Modern statistical machine translations can learn translation statistics on fixed collocations (so-called "statistical phrases") that result in more accurate translation (Och and Ney, 2004). Information retrieval, however, requires that we optimize an objective function that balances specificity (to enhance precision) with generality (to enhance recall). In particular, it is well known that it is more effective to index both phrases and their component words than it would be to index phrases alone. Our present meaning matching framework assumes, however, that we have

some specific tokenization process for the queries and for the documents. Future work on integrating evidence from multiple plausible tokenizations might be productive. The idea of bidirectional translation is also not unique to CLIR. Machine translation researchers leverage a comparable idea ("alignment by agreement"), which is now available as a replacement for GIZA++ in the Berkeley Aligner (Liang et al., 2006)). Comparison of our implementations of IMM and DAMM with variants based on Berkeley alignment results would be a logical first next step towards understanding the potential of these alignments in CLIR applications

Of course, there are many other ways in which our work could be extended. For example, we chose not to use blind relevance feedback (Ballesteros and Croft, 1997), not to perform context-based reweighting of translation probabilities (Gao et al., 2001, e.g.), and not to use term proximity features (Lv and Zhai, 2009). These choices were motivated by our desire to keep our experiment designs straightforward enough for the analysis that we wished to perform, but in future work all of those techniques might productively be tried.

There are also many sources of evidence for synonymy that might be explored, including traditional sources such as stemming, WordNets and thesauri, and more recent developments such as learning term relationships from distributional statistics in a monolingual corpus, or extracting emergent term relationships from crowdsourced resources such as Wikipedia. In interactive settings it might also be possible to leverage the user's understanding of term meaning in some way (He and Wu, 2008, e.g.). It is also worth noting that for implementation convenience we leveraged synonymy after translation probabilities had been generated, but architectures in which putative synonyms are conflated before the translation probabilities are learned also deserve study. Our greedy approach to conflation might also be improved upon.

Notably, we have to date relied on just one kind of evidence (term counts) and just one way of using that evidence (Okapi BM25 term weights), but of course real applications can and should rely on a broader range of evidence. Indeed, experience with learning to rank for Web search has shown that non-content evidence (e.g., associating queries and clicks, learning authority measures from link graphs, and a bias in favor of shorter URL's) can have substantial effects on retrieval effectiveness. Investigating the relative contribution of meaning matching in such settings should therefore command the attention of those researchers who have access to more comprehensive feature sets.

Finally, the way in which we have modeled uncertainty in meaning matching would seem to have a clear applicability to other cases such as speech retrieval and retrieval of scanned documents based on Optical Character Recognition (OCR); in both settings, uncertainty naturally arises in the query-document matching process. The generative techniques that have been applied to those problems to date (Olsson and Oard, 2009; Darwish and Oard, 2003, e.g.) suffer from serious efficiency issues that could likely be mitigated by modeling uncertainty as a translation process (even when the query and document languages are the same). Indeed, this perspective bring us full circle. In information retrieval our goal has always been to match meanings. CLIR evolved in its early days through a stage in which the translation and retrieval components were seen as modules to be coupled. It was only when information researchers who were working with language models turned their attention to CLIR that closer integration began to be explored. We have now formalized that closer integration in a way that extends the technique to include ranking with Okapi BM25 weights, and further extension to other ranking techniques would be quite straightforward. But the story does not end there—the key idea is that representations of uncertainty have a natural place in the retrieval process, and with that in mind we are well positioned to think broadly about how best these techniques can be further extended, and applied in new ways.

6. Acknowledgments

The authors would like to thank Vedat Diker, Jimmy Lin, Jim Mayfield, Philip Resnik, Dagobert Soergel and the anonymous reviewers for their valuable comments. This work was supported in part by DARPA contracts N661010028910 (TIDES) and HR0011-06- 2-0001 (GALE).

References

Ballesteros, L., 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. pp. 84–91.

Boughanem, M., Chrisment, C., Nassr, N., 2001. Investigation on disambiguation in CLIR: Aligned corpus and bi-directional translation-based

strategies. In: Peters, C., Braschler, M., Gonzalo, J., Kluck, M. (Eds.), Evaluation of Cross-Language Information Retrieval Systems: Second Workshop of the Cross-Language Evaluation Forum. pp. 158–167.

Braschler, M., 2004. Combination approaches for multilingual text retrieval. Information Retrieval 7 (1-2), 183–204.

Byrne, W., Doermann, D. S., Franz, M., Gustman, S., Hajic, J., Oard, D. W., Picheny, M., Psutka, J., Ramabhadran, B., Soergel, D., Ward, T., Zhu, W.-J., 2004. Automatic recognition of spontaneous speech for access to multilingual oral history archives. IEEE Transactions on Speech and Audio Processing 12 (4), 420–435.

Chiang, D., Lopez, A., Madnani, N., Monz, C., Resnik, P., Subotin, M., 2005. The Hiero machine translation system: Extensions, evaluation, and analysis. In: Proceedings of HLT/EMNLP 2005. pp. 779–786.

Darwish, K., 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. pp. 338–344.

Daumé, H., 2007. Frustratingly easy domain adaptation. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics. pp. 256–263.

Davis, M., Dunning, T., Nov. 1995. A TREC evaluation of query translation methods for multilingual text retrieval. In: The Fourth Text Retrieval Conference (TREC-4). National Institute of Standards and Technology, http://trec.nist.gov/.

Demner-Fushman, D., 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. p. 108.2.

Federico, M., Bertoldi, N., Aug. 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, pp. 167–174.

Fraser, A., Xu, J., Weischedel, R., 2002. TREC 2002 cross-lingual retrieval at BBN. In: The 11th Text REtrieval Conference.

Gao, J., Xun, E., Zhou, M., Huang, C., Nie, J.-Y., Zhang, J., 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, pp. 96–104.

He, D., Wu, D., 2008. Translation enhancement: a new relevance feedback method for cross-language information retrieval. In: CIKM '08: Proceeding of the 17th ACM Conference on Information and Knowledge Management. pp. 729–738.

Hiemstra, D., de Jong, F., 1999. Disambiguation strategies for cross-language information retrieval. In: Proceedings of the third European Conference on Research and Advanced Technology for Digital Libraries. pp. 274–293.

Kang, I.-S., Na, S.-H., Lee, J.-H., 2004. Combination approaches in information retrieval: Words vs. n-grams and query translation vs. document translation. In: Proceedings of the 4th NTCIR Workshop. National Institute of Informatics.

Kishida, K., Chen, K., Lee, S., Kuriyama, K., Kando, N., Chen, H.-H., Myaeng, S. H., 2005. Overview of CLIR task at the fifth NTCIR workshop. In: Proceedings of the 5th NTCIR Workshop. National Institute of Informatics.

Knight, K., 1999. A statistical MT tutorial workbook. Http://www.clsp.jhu.edu/ws99/projects/mt/wkbk.rtf.

Koehn, P., 2005. Europarl: A parallel corpus for statistical machine translation. In: MT Summit 2005.

Koehn, P., Hoang, H., 2007. Factored translation models. In: EMNLP-CoNLL. pp. 868–876.

Kraaij, W., Nie, J.-Y., Simard, M., 2003. Embedding web-based statistical translation models in cross-language information retrieval. Computational Linguistics 29 (3), 381–419.

Kwok, K. L., 2000. Exploiting a Chinese-English bilingual wordlist for english-chinese cross language information retrieval. In: Proceedings of the 5th International Workshop on Information Retrieval with Asian languages. pp. 173–179.

Kwok, K.-L., Choi, S., Dinstl, N., Deng, P., 2005. NTCIR-5 chinese, english, korean cross language retrieval experiments using PIRCS. In: Proceedings of the Second NTCIR Workshop on Research in Chinese & Japanese Text Retrieval and Text Summarization. National Institute of Informatics, http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings5/index.html.

Liang, P., Taskar, B., Klein, D., 2006. Alignment by agreement. In: Proceedings of Human Language Technologies: The 7th Annual Conference of the North American Chapter of the Association for Computational Linguistics. pp. 104–111.

Lv, Y., Zhai, C., 2009. Positional language models for information retrieval. In: Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR. pp. 299–306.

McCarley, J. S., 1999. Should we translate the documents or the queries in cross-language information retrieval? In: Proceedings of the 37th Annual Conference of the Association for Computational Linguistics. pp. 208–214.

McCarley, J. S., Roukos, S., 1998. Fast document translation for cross-language information retrieval. In: AMTA '98: Proceedings of the Third Conference of the Association for Machine Translation in the Americas on Machine Translation and the Information Soup. pp. 150–157.

McNamee, P., Mayfield, J., Nicholas, C., 2009. Translation corpus source and size in bilingual retrieval. In: NAACL '09: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers. pp. 25–28.

Nie, J.-Y., Isabelle, P., Plamondon, P., Foster, G., 1998. Using a probabilistic translation model for cross-language information retrieval. In: 6th Workshop on Very Large Corpora. pp. 18–27.

Nie, J.-Y., Simard, M., 2001. Using statistical models for bilingual IR. In: Proceedings of Evaluation of Cross-Language Information Retrieval Systems: Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001.

Oard, D. W., Ertunc, F., 2002. Translation-based indexing for cross-language retrieval. In: 24th BCS-IRSG European Colloquium on IR Research.

Oard, D. W., Soergel, D., Doermann, D., Huang, X., Murray, G. C., Wang, J., Ramabhadran, B., Franz, M., Gustman, S., Mayfield, J., Kharevych, L., Strassel, S., 2004. Building an information retrieval test collection for spontaneous conversational speech. In: Proceedings of the 20th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 41–38.

Oard, D. W., Wang, J., Sep. 1999. NTCIR CLIR experiments at the University of Maryland. In: Proceedings of the first NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition. National Institute of Informatics.

Oard, D. W., Wang, J., Jones, G. J. F., White, R. W., Pecina, P., Shafran, I., 2007. Overview of the CLEF-2006 cross-language speech retrieval track. In Carol Peters et al. (Eds) Evaluation of Multilingual and Multi-modal Information Retrieval., Lecture Notes in Computer Science 4730, 786–793.

Och, F. J., Ney, H., October 2000. Improved statistical alignment models. In: Proceedings of the 38th Annual Conference of the Association for Computational Linguistics. pp. 440–447.

Och, F. J., Ney, H., 2004. The alignment template approach to statistical machine translation. Computational Linguistics 30 (4).

Olsson, J. S., Oard, D. W., 2009. Combining LVCSR and vocabulary-independent ranked utterance retrieval for robust speech search. In: Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 91–98.

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. pp. 55–63.

Resnik, P., Yarowsky, D., 2000. Distinguishing systems and distinguishing senses: New evaluation methods for word sense disambiguation. Natural Language Engineering 5(2), 113–133.

Robertson, S. E., Sparck-Jones, K., 1997. Simple proven approaches to text retrieval. Technical Report, Cambridge University Computer Laboratory.

Sanderson, M., Zobel, J., Aug. 2005. Information retrieval system evaluation: Effort, sensitivity and reliability. In: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 162–169.

Voorhees, E. M., 1993. Using WordNet to disambiguate word senses for text retrieval. In: Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 171–180.

Voorhees, E. M., 2000. Variations in relevance judgments and the measurement of retrieval effectiveness. Information Processing & Management 36 (5), 697–716.

Wang, J., 2005. Matching meaning for cross-language information retrieval. Ph.D. thesis, University of Maryland, College Park.

Wang, J., 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. pp. 202–209.

Xu, J., Fraser, A., Weischedel, R., 2002. Empirical studies in strategies for arabic retrieval. In: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 269–274.

Xu, J., Weischedel, R., Nov. 2000. TREC-9 cross-lingual retrieval at BBN. In: The Ninth Text REtrieval Conference. National Institute of Standards and Technology.

Xu, T., Oard, D. W., 2008. FIRE 2008 at Maryland: English-Hindi CLIR. In: Working notes of the 2008 Forum for Information Retrieval Evaluation.