Combining Bidirectional Translation and Synonymy for Cross-Language Information Retrieval

Jianqiang Wang and Douglas W. Oard
College of Information Studies and UMIACS
University of Maryland, College Park, MD 20742
{wangjq,oard}@glue.umd.edu

ABSTRACT

This paper introduces a general framework for the use of translation probabilities in cross-language information retrieval based on the notion that information retrieval fundamentally requires matching what the searcher means with what the author of a document meant. That perspective yields a computational formulation that provides a natural way of combining what have been known as query and document translation. Two well-recognized techniques are shown to be a special case of this model under restrictive assumptions. Cross-language search results are reported that are statistically indistinguishable from strong monolingual baselines for both French and Chinese documents.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Experimentation, Measurement

Keywords

Cross-Language IR, Statistical translation

1. INTRODUCTION

Information retrieval systems seek to identify documents in which authors chose their words to express the same meanings that the searcher intended as they choose their query terms. Cross-Language Information Retrieval (CLIR) deals with the special case of this problem in which the documents and the queries are expressed using words in different languages. Direct matching of terms between the query and a document would generally fail, so the usual approach has been to translate in one direction or the other so that the query and the document are expressed using terms in the same language; direct term matching techniques can then be employed. Both directions have weaknesses: the limited context available in (typically) short queries adds uncertainty to query translation, and computational costs can limit the extent to which context can be exploited when translating large document collections. Nevertheless, these have proven to be practical approaches; systems that make effective use of translation probabilities learned from parallel corpora can achieve retrieval effectiveness measures similar to those achieved by comparable monolingual systems.

Query translation achieves the information retrieval system's goal by approximating what would have happened if the searcher actually had expressed their query in the document language. Document translation takes the opposite tack, approximating what would have happened if the authors had written in the query language. McCarley found that merging ranked lists generated using query translation and document translation yielded improved mean average precision over that achieved by either approach alone [11], which suggests that bidirectional techniques are worth exploring. In this paper, we return to first principles to derive an approach to CLIR that is motivated by cross-language meaning matching. This framework turns out to be quite flexible, accommodating alternative computational approximations to meaning and subsuming existing approaches to query and document translation as special cases. Moreover, the approach is also effective, repeatedly outperforming the best previously published query translation technique.

The remainder of the paper is organized as follows. In Section 2, we review previous work on CLIR using query translation, document translation, and merged result sets. Section 3 then introduces our meaning matching model and explains how some previously known CLIR techniques can be viewed as restricted implementations of meaning matching. Section 4 then describes the design of an experiment in which three variants of meaning matching are compared to strong monolingual and CLIR baselines. The results presented in section 5 illustrate the effect of exploiting alternative language resources in the meaning matching framework, showing that the use of bidirectional translation knowledge and similarity-based synonymy can yield statistically significant improvements in mean average precision over previously known query translation techniques. Section 6 then concludes the paper with a discussion of the implications of the meaning matching model for future work on CLIR.

2. PREVIOUS WORK

In order to create broadly useful systems that are computationally tractable, it is common in information retrieval

SIGIR'06, August 6–11, 2006, Seattle, Washington, USA

generally, and in CLIR in particular, to treat terms independently. Research on CLIR has therefore focused on three main questions: (1) which terms should be translated?; (2) what possible translations can be found for those terms?; and (3) how should that translation knowledge be used? Our focus in this paper is on the third of those questions. In this section, we review prior work on the question of how a known set of translations should be used.

"Translation" is actually somewhat of a misnomer, since the most effective approaches map term statistics, rather than the terms themselves, from one language to another. Three basic statistics are used in information retrieval systems that use a "bag of words" representation of queries and documents: the number of occurrences of a term in a document (Term Frequency, or TF), the number of terms in the document (Length, or L), and the number of documents in which a term appears (Document Frequency, or DF). Generally, documents in which the query terms have a high TF (after length normalization) are preferred, and highly selective query terms (i.e., those with a low DF) are given extra weight in that computation.

When no translation probabilities are known, Pirkola's "structured queries" have been repeatedly shown to be among the most effective known approaches when several plausible translations are known for some query terms [15]. The basic idea behind Pirkola's method is to treat multiple translation alternatives as if they were all instances of the query term. Specifically, the TF of a query term with regard to a document is computed as the summation of the TF of each of its translation alternatives that are found in that document, and its DF in the collection is computed as the number of documents in which at least one of its translation alternatives appears. Both the TF and DF can be precomputed for each possible query term at indexing time [12], but query-time implementations are more common in experimental settings. The DF computation is expensive at query time, so Kwok later proposed a simplification that upper bounds Pirkola's DF with no noticeable adverse effect on retrieval effectiveness [8]. With the simplified computation, the DF of a query term is estimated as the sum of the DF of each of its translation alternatives.

Darwish later extended Kwok's formulation to handle the case in which translation probabilities are available by weighting the TF and DF computations, an approach he called probabilistic structured queries (PSQ) [4].

$$TF(e, D_k) = \sum_{f_i} p(f_i|e) \times TF(f_i, D_k) \quad (1)$$

$$DF(e) = \sum_{f_i} p(f_i|e) \times DF(f_i) \quad (2)$$

where $p(f_i|e)$ is the estimated probability that e would be properly translated to f_i. Similar approaches have also been used in a language modeling framework, often without explicitly modeling DF (e.g., [7, 9, 20]). Translation probabilities can be estimated from corpus statistics (using translation-equivalent "parallel" texts), directly from dictionaries (when presentation order encodes relative likelihood of general usage), or from the distribution of an attested translation in multiple sources of translation knowledge. Darwish found that Pirkola's structured queries yielded declining retrieval effectiveness with increasing numbers of translation alternatives, but that the incorporation of translation probabilities in PSQ tended to mitigate that effect.

McCarley was the first to try bidirectional translation, merging a ranked list generated using query translation with another ranked list generated using document translation [11]. He found that the merged result yielded statistically significant improvements in mean average precision when compared to either query or document translation alone, and similar improvements have since been obtained by others (e.g. [2, 5]). Our "meaning matching" model, introduced in the next section, can be viewed as an effort to build on that insight by more directly incorporating bidirectional translation evidence into the retrieval model. Boughanem et al. took an initial step in the direction that we explore, using bidirectional ("round trip") translation to filter out potentially problematic translations that were attested in only one direction, but without incorporating translation probabilities [1]. In the next section, we derive a general approach to meaning matching and then propose a range of computational implementations.

3. MATCHING MEANING

In this section, we derive an overarching framework for matching meanings between queries and documents and a range of computational implementations that incorporate different sources of evidence.

3.1 IR as Matching Meaning

IR can be viewed as a task of matching the meaning intended in a query with the meaning expressed in each document. The term independence assumption allows us to score each document based on matches between the meaning of each query term with the meaning of each document term. Of course, in human languages different terms may share the same meaning. In monolingual IR it is common to treat words that share a common stem as if they expressed the same meaning, and some automated and interactive query expansion techniques can also be cast in this framework. The key insight between what we call meaning matching is to apply that same perspective directly to CLIR.

The basic formulae are a straightforward generalization of Darwish's PSQ technique with one important difference: no translation direction is specified. Instead, for each word e in query language E, we simply assume that a set of terms f_i $(i = 1, 2, \ldots, n)$ in document language F is known, each of which shares the searcher's intended meaning for term e with some probability $p(e \leftrightarrow f_i)$ $(i = 1, 2, \ldots, n)$ respectively. Any uncertainty about the searcher's meaning for e is reflected in these statistics, the computation of which is described in subsequent parts of this section. If we see a translation f_i appearing one time in document d_k, we can therefore treat this as our having seen query term e occurring $p(e \leftrightarrow f_i)$ times in that document. If term f_i occurs $TF(f_i, d_k)$ times, our estimate of the total "occurrence" of query term e as estimated from the occurrences of document term f_i will be $p(e \leftrightarrow f_i) \times TF(f_i, d_k)$. 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:

$$TF(e, d_k) = \sum_{f_i} p(e \leftrightarrow f_i) \times TF(f_i, d_k) \quad (3)$$

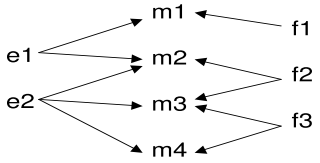

Figure 1: Illustrating the effect of overlapping bidirectional translations.

Turning our attention to the DF, if document d_k contains a term f_i that might share a meaning with e, we can treat the document as possibly "containing" e. Indeed, if every term that shares a meaning with e is found in that document, the meaning of e is sure to have been intended by the author of that document and the contribution of that document to the DF computation should be 1. If only some of the terms that share a common meaning with e appear in a document, 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:

$$DF(e) = \sum_{f_i} p(e \leftrightarrow f_i) \times DF(f_i) \quad (4)$$

Document length normalization is unaffected by this process because it can be performed using only document-language term statistics.

The comparison to Darwish's PSQ (Equations 1 and 2) is direct; PSQ is simply a unidirectional special case of meaning matching. The opposite direction, using $p(e|f_i)$ rather than $p(f_i|e)$ seems at least equally well (and perhaps better) motivated, but the fundamental insight behind meaning matching is that there is no need to commit to one translation direction or the other.

3.2 Matching Abstract Term Meanings

To model how term meaning is matched across languages, consider a case in which two English query terms and three French document terms share subsets of four different meanings (see Figure 1). At this point we treat "meaning" as an abstract concept; a computational model of meaning is introduced in the next section. In this example, the query term e_2 has the same meaning as the document term f_2 if and only if e_2 and f_2 both express meaning m_2 or 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 probability distributions for those two events. Generalizing to any pair of words e and f:

$$p(e \leftrightarrow f) \approx \sum_{s_j} p(s_j|e) \times p(s_j|f) \quad (5)$$

where:

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

Note that despite our notation, $p(e \leftrightarrow f)$ values are not actually probabilities but rather products of probabilities. For example, if all possible meanings of every term were equally likely, then $\sum_i p(e_1 \leftrightarrow f_i) = 0.75$ while $\sum_i p(e_2 \leftrightarrow f_i) = 0.67$. This would have the undesirable effect of giving more weight to some query terms than others, so we renormalize the values so that $\sum_{i=1}^{n} p(e \leftrightarrow f_i)$ is 1 for every query term e. This yields something that we can treat as if it were a probability distribution, although we retain the $\leftrightarrow$ notation throughout as a reminder of the process by which the values were produced.

It can be useful to threshold these probabilities in some way because low probability events are generally not modeled well. We therefore compute the cumulative distribution function for every e and apply a fixed threshold (selected from a grid of values), which we called Cumulative Probability Threshold (CPT), to select the matches that will be used. This is done by ranking the translations in decreasing order of their normalized probabilities, then iteratively selecting translations top-down until the cumulative probability of the selected translations is first reached or exceeds the threshold. A threshold of 0 thus corresponds to using the single most probable translation (a well-studied baseline) and a threshold of 1 corresponds to use of all translation alternatives. The $p(e \leftrightarrow f)$ are again normalized after the threshold is applied.

3.3 Using Synsets to Represent Meaning

Further development of meaning matching requires a computational model of meaning in which meaning representations are aligned across languages. We chose "synsets," sets of synonymous terms, as a simple computational model of meaning. Cross-language syset alignments are available from some sources, most notably EuroWordNet. We call meaning matching implemented in that way *Full Aggregated Meaning Matching* (FAMM). For cases in which aligned synsets do not already exist, we decompose the problem into (1) mapping words across languages, (2) mapping words in each language into monolingual synsets, (3) aggregating the word-to-word mappings to produce word-to-synset mappings, and (4) aligning the resulting synsets.

We could obtain evidence for monolingual synonymy in English from WordNet, but similar resources are available for only a small number of relatively resource-rich languages. We therefore explored one of the several possible sources of statistical evidence for synonymy. Because statistical word-to-word translation models were available for use in our CLIR experiments, we elected to find candidate synonyms by looking for words in the same language that were linked by a common translation. For example, to find document-langauge synonyms, we computed:

$$p(f_j \leftrightarrow f) \approx \sum_{i=1}^{n} p(e_i|f) \times p(f_j|e_i) \quad (6)$$

where $p(f_j \leftrightarrow f)$ refers to the probability of f_j being a synonym of f. Of course, that results in a proliferation of poorly estimated low probability events. We therefore arbitrarily suppressed any candidate synonyms for which $p(f_j \leftrightarrow f) < 0.1$. Alternatively, we could use statistical translation in only one direction (e.g., $\sum_{e_i} p(e_i|f) \times p(e_i|f_j)$) to derive statistical synonyms. Other ways of constructing statistical synonym sets are also possible (e.g., distributional

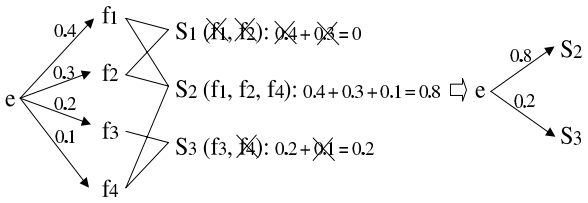

Figure 2: Illustrating the greedy aggregation

similarity in monolingual corpora), but recent work on word sense disambiguation suggests that translation usage can provide a strong basis for identifying synonyms [16].

Statistical word-to-word translation has been well studied, and a number of effective implementations are available (e.g., [13]). To derive a word-to-synset mapping model from a statistical word-to-word translation model, we aggregated multiple translation alternatives based on synsets in the target language. Since some translations might appear in more than one synset, we needed some way of assigning their translation probability across those synsets. We used a simple greedy method, iteratively assigning each translation to the synset that would yield the greatest aggregate probability. Specifically, the algorithm worked as follows:

1. Compute the aggregate probability that e maps to each s_j: $p(s_j|e) = \sum_{f_i \in s_j} p(f_i|e)$, and rank all s_j in decreasing order of aggregate probability;

2. Select Synset s_j with the largest aggregate probability, remove all of its terms from every synset and iterate.

Figure 2 illustrates the greedy method of aggregating synonymous translation alternatives into synsets by an example. In that example, four translations of word e are grouped into two synsets s_2 and s_3: s_2 contains three of the four translation with $p(s_2|e) = 0.8$, while s_3 contains only the other one translation with $p(s_3|e) = 0.2$. Thus, probabilistic mapping of words in one language to synsets in another language is achieved.

The selected synsets then form a word-to-synset mapping for e. The same computation can be performed in the other direction. Because greedy aggregation results in unique mappings, at most one alignment can exist in which a query term e maps to a document-language synset s_d that contains f and document term f maps to a query-language synset s_q that contains e. As a result, the summation in Equation 5 will be unused. We call the resulting technique *Derived Aggregated Meaning Matching* (DAMM).

The incorporation of aggregation is a distinguishing characteristic of the meaning matching model, so we wanted to isolate the effect of aggregation for a contrastive analysis. If we simply assume that each term encodes a unique meaning, we get $p(e \leftrightarrow f) = p(e|f) \times p(f|e)$. We call this *Individual Meaning Matching* (IMM). Similarly, we can isolate the effect of bidirectional translation knowledge by further assuming uniform translation probabilities in one direction. For example, assuming a uniform distribution for $p(e|f)$ across all f yields (after normalization) $p(e \leftrightarrow f) = p(f|e)$, which is exactly the formulation of PSQ. If uniform translation probabilities are assumed in both directions $p(e \leftrightarrow f)$ becomes a constant factor. In this case, PSQ is simplified as Pirkola's structured queries. In the next section we describe experiments to compare the relative effectiveness of PSQ, IMM, DAMM, and FAMM.

Test collection from	CLEF'01-03	TREC-5,6
Query language	English	English
Document language	French	Chinese
# of search topics	151	54
# of documents	87,191	139,801
Avg. # of rel docs per topic	23	95

Table 1: Test collection statistics

4. EXPERIMENT DESIGN

To evaluate the effectiveness of the proposed meaning matching model for CLIR, we conducted two sets of experiments: one on retrieving French news stories with English queries and the other on retrieving Chinese news stories with English queries. This section describes the experiment setup for the study, including the selection of the test collection and IR system, and training translation models, inducing statistical synonyms

4.1 Test collection and IR system

Table 1 shows the statistics of the two test collections used in our experiments. For English-French CLIR, we accumulated the French test collections created by the Cross-Language Evaluation Forum (CLEF) in 2001, 2002 and 2003 into a single collection.[1] We stripped accents from the document collection and removed French terms contained on the stopword list provided with the open source Snowball stemmer.[2] We then created a document index based on stemmed French terms. We formulated TD queries with words from the title and description filed in the search topics. For English queries, we performed pre-translation stopword-removal using an English stopword list provided with Inquery. For French queries, we performed accent-removal, stopword-removal, and stemming using the same tools that we used for processing the document collection. The French queries serve to establish a useful upper baseline for CLIR effectiveness.

For English-Chinese CLIR, we accumulated search topics from TREC-5 and TREC-6, which used the same Chinese document collection. That gives us a total of 54 topics. The Chinese documents, originally encoded in GB code, were converted into UTF-8 using the uconv codeset conversion tool and then segmented into individual words using the LDC Chinese segmenter.[3] The resulting document collection was then converted into hexadecimal format that guards against character handling problems [10]. We also formulated TD queries. For Chinese queries, we performed codeset conversion and segmentation in the same way that the Chinese documents were processed. For English queries, we again removed stopwords using the Inquery stopword list.

All our experiments were run using the Perl Search Engine (PSE), a document retrieval system based on Okapi BM25 weights that already implements PSQ. We obtained PSE from the University of Maryland and modified it to implement other variants of cross-language meaning matching. In

[1]The 9 of the 160 accumulated topics that do not have relevant French documents were removed from the collection.
[2]http://snowball.tartarus.org/
[3]http://www.ldc.upenn.edu/Projects/Chinese/segmenter/mansegment.perl

Parallel corpus	EUROPARL	Multiple sources
Language	English-French	English-Chinese
Sentence pairs	672,247	1,583,807
Model iterations	10 Model 1 5 HMM 5 Model 4	10 Model 1

Table 2: Corpus statistics and model iterations for training translation models.

the Okapi BM25 formula [17], We used $k_1 = 1.2$, $b = 0.75$, and $k_3 = 7$ as has been commonly used.

4.2 Training statistical translation models

Table 2 describes the process that we used to train our statistical translation models. For both language pairs, we derived word-to-word translation models in both directions using the freely available GIZA++ toolkit [13].[4] For French, we trained the translation models with the Europarl parallel corpus [6]. For Chinese, we combined corpora from multiple sources including the Foreign Broadcast Information Service (FBIS) corpus, HK News and HK Law, UN corpus, and Sinorama, the same corpora also used by Chiang et al [3]. We stripped accents from the French documents, segmented the Chinese documents with the same version of LDC segmenter that was used for indexing, and filtered out implausible sentence alignments by eliminating sentence pairs with a token ratio either smaller than 0.2 or larger than 5.

For both language pairs, we ran GIZA++ twice, with either of the two languages as the source language respectively. When training translation models for the English-French pair, we started with 5 HMM iterations, followed by 10 IBM Model 1 iterations, and ending with 5 IBM Model 4 iterations. The net result of this process was two translation tables, one from English words to French words and the other from French words to English words. All nonzero values produced by GIZA++ were retained in each table.

We ran our Chinese-English experiments after the English-French experiments with the goal of confirming our results using a different language pair, so we made a few changes to reduce computational costs. Model 4 seeks to achieve better alignments by modeling systematic position variations; that is an expensive step not commonly done for CLIR experiments. We therefore omitted Model 4 for the English-Chinese pair. We ran 10 IBM Model 1 iterations followed by 5 HMM iterations. A comparison of results using lexicons from before and after the 5 HMM iterations indicated no noticeable difference between the two conditions, so in this paper we report Chinese-English results only for the 10 IBM Model 1 iterations. Finally, we observed in our English-French experiments that working with a large number of low probability translations yielded both lower effectiveness and greater computational costs, so we imposed a cumulative probability threshold of 0.99 on the model for each translation direction before creating bidirectional models for our English-Chinese experiments.

5. RESULTS

In this section, we report our experiment results for both English-French CLIR and English-Chinese CLIR. We present the results in three parts: (1) establishing a strong upper baseline using French queries, (2) establishing a strong lower baseline using known CLIR techniques with English queries, and (3) comparing the retrieval effectiveness of the meaning matching model with those baselines. We show that meaning matching that combines bidirectional translation and statistical synonymy knowledge achieved results that were statistically indistinguishable from the upper (monolingual) baseline and significantly better than the lower (CLIR) baseline for CLIR with both language pairs.

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

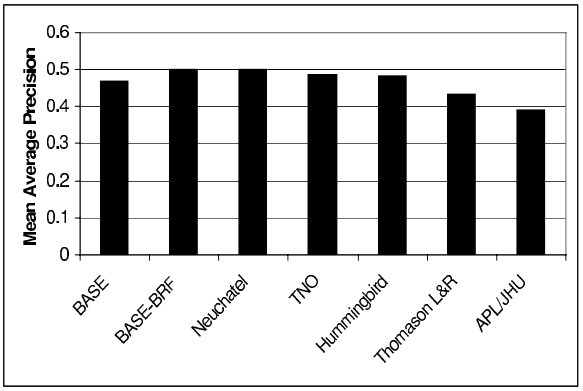

Figure 3: Comparison with the top 5 official CLEF runs.

5.1 Upper (monolingual) baseline

Although not strictly an upper bound (because of expansion effects), it is quite common in CLIR evaluation to compare the effectiveness of a CLIR system with a monolingual baseline. We obtained monolingual baselines for each language pair by retrieving documents with TD queries formulated from search topics that are expressed in the same language as the documents.

To get a better idea of the effectiveness of our monolingual baselines, we compared them with published top results gained from experiments with the same test collections. For the English-French CLIR experiments, we computed the mean average precision (MAP) over 50 queries formulated from the CLEF 2001 topic set (Topics 41-90). Figure 3 shows the MAP of the top five official monolingual French runs from CLEF 2001. Our baseline (BASE in the figure) achieved a MAP of 0.470, which is above the average (0.460) of those top five runs but lower than the top three runs. We noticed the best CLEF 2001 run tweaked the stopword list and stemming, and, in particular, used query expansion based on blind relevance feedback [18]. To facilitate comparison, we also expanded our original French queries with the top 20 words selected from the top 10 retrieved documents based on Okapi weights, weighting the added words with a coefficient of 0.1. This resulted in a monolingual MAP of 0.501 (BASE-BRF in Figure 3) that closely matched the best official run in CLEF 2001 monolingual French retrieval. This suggests that our monolingual baseline is strong. With a goal to study the relative effectiveness of the meaning matching model, we want to avoid masking those effects by other factors. Therefore, blind relevance feedback was not used in the remaining runs.

For the monolingual baseline in the English-Chinese CLIR experiments, we computed results for the same 19 TREC-5

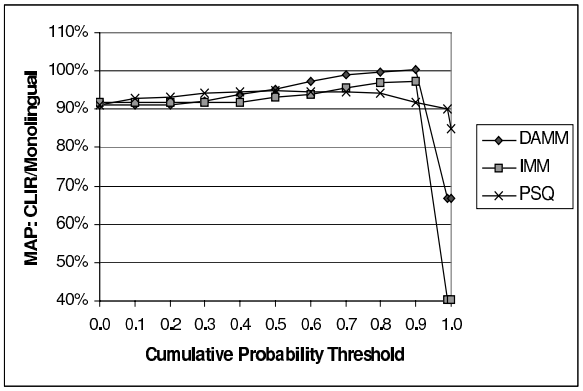

Figure 4: Comparison of meaning matching with monolingual baseline and PSQ for English-French CLIR

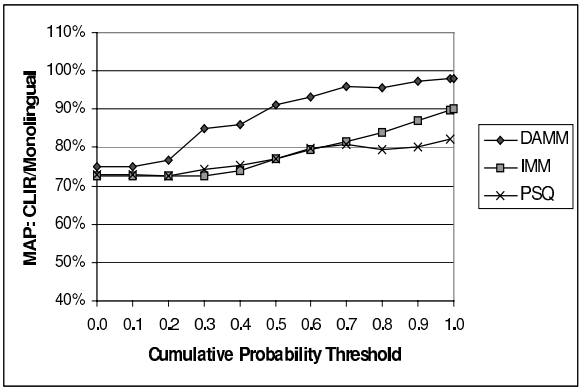

Figure 5: Comparison of meaning matching with monolingual baseline and PSQ for English-Chinese CLIR

queries for which results had been reported at the TREC-5 conference. We obtained a MAP of 0.280, which was at the median of the 15 automatic official runs submitted to TREC-5.[5] Most of those runs across which the median was computed used longer queries (all words from the title, description, and narrative field), however, whereas we used only the title and description fields for all of our experiments (in both language pairs). Moreover, as had been the case for French we did no automatic query expansion. We therefore feel that our monolingual baseline for Chinese is a reasonable one.

5.2 Lower (CLIR) baseline

A major motivation for us to develop the cross-language meaning matching model is to improve CLIR effectiveness over a strong CLIR baseline. We chose probabilistic structured queries (PSQ) as our CLIR baseline because among vector space techniques for CLIR it presently yields the best retrieval effectiveness. Direct comparison to techniques based on language modeling would be more difficult to interpret because vector space and language modeling handle issues such as smoothing and DF differently.

Figure 4 shows the relative English-French CLIR effectiveness as compared to the monolingual French baseline. We ran CLIR and computed MAP at different Cumulative Probability Thresholds (CPT). What is shown at each point in the figure is the monolingual percentage of the CLIR MAP. Overall, English-French CLIR was very effective, achieving at least 90% of monolingual MAP when translation alternatives with very low probability were excluded. In addition, the baseline PSQ technique exhibited the same decline in MAP near the tail of the translation probability distribution (i.e., at high cumulative probability thresholds) that Darwish and Oard reported [4]. The best MAP of PSQ was obtained at a CPT of 0.5, which is near 95% of monolingual effectiveness. However, the difference is still statistically significant by a Wilcoxon signed rank test (at $p < 0.05$).

In the English-Chinese case, PSQ with multiple translations was always better than with the one-best translation (corresponding to the CPT of 0) before the cumulative probability reached 0.99, which is where the best PSQ was obtained. However, MAP of the best PSQ was just about 82% of monolingual MAP, and was significantly lower.

In the English-Chinese CLIR experiments, CLIR MAP did not tail off because we excluded translations after the cumulative probability reached 0.99.

5.3 Cross-language meaning matching

Also shown in Figure 4 and Figure 5 are cross-language meaning matching based on bidirectional translation and synonym aggregation. The effectiveness of English-French CLIR based on IMM, which uses bidirectional translation but without synonymy knowledge, showed monotonic increase before CPT reaches 0.9. The highest MAP (0.376 at a CPT of 0.9) is about 97% of monolingual MAP, which is statistically indistinguishable from either the best PSQ or the monolingual baseline. For English-Chinese CLIR, the effectiveness of IMM showed similar pattern of changes. As far as comparison is concerned, the best IMM (at a CPT of 0.99) is about 90% of monolingual MAP, which is significantly better than the best PSQ while still worse than monolingual baseline.

The monotonic increase of MAP at low and medium CPT regions seems to indicate some advantage of using bidirectional translation knowledge over unidirectional translation knowledge. Essentially this is because using bidirectional translation knowledge can both eliminate some spurious translation alternatives that are otherwise included in unidirectional translation and gives better estimation of meaning matching probability. However, such effects are limited, especially when many low probability translations are included. In fact, after a CPT of 0.9 in English-French CLIR, IMM decreased faster than PSQ, showing combining bidirectional translation knowledge may have included more *low-probability* translations than using unidirectional translation knowledge. A statistical translation model can in principle translate any word into any other word appearing in any aligned sentence, and low probability events are naturally not very well modeled. We show below that synonymy knowledge can partially offset the negative effect due to the inclusion of too many low-probability translations.

When bidirectional translation knowledge is combined with statistical synonymy knowledge, which is the case of derived aggregated meaning matching (DAMM), the best DAMM was significantly better than the best PSQ for both English-French CLIR (with 6% relative improvement) and English-

[5]http://trec.nist.gov/pubs/trec5/t5_proceedings.html

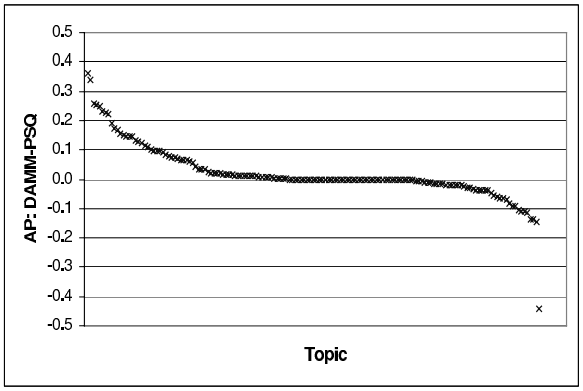

Figure 6: Query-by-query comparison of the best DAMM and the best PSQ for English-French CLIR.

Chinese CLIR (with 19% relative improvement), achieving cross-language MAP comparable to monolingual baselines in both cases. However, in both cases, the best DAMM was statistically indistinguishable from the best IMM. Putting these findings together with the above comparisons of IMM with PSQ and monolingual retrieval, it is reasonable to say that both bidirectional translation knowledge and synonymy knowledge can help, and combining them can help more.

For English-French CLIR, full aggregated meaning matching (FAMM) with aligned synsets obtained from EuroWordNet reached only about 30% of monolingual MAP, which is significantly worse than any of the meaning matching techniques we tried. We found that many high-probability translations contained in the GIZA++ translation tables were not covered by the aligned synsets, and our implementation of FAMM therefore treated their probabilities as zero. This is clearly undesirable, and future work on compensating for limited word coverage of aligned synsets is needed.

Overall, aggregation had little effect at low CPT values. This is mainly because the number of translation alternatives included at low CPT values was very small (in most cases there was just one translation selected). Generally, the more translations involved, the larger effect aggregation is likely to have. Therefore, at high CPT values where more translations are included, aggregation tends to have more effect on meaning matching.

Although a Wilcoxon signed rank test shows DAMM significantly outperformed PSQ when the CPT threshold was adjusted most favorably for each, we want to further investigate what actually happened through query-by-query comparison. We plot the non-interpolated average precision (AP) difference for each query between the best DAMM and the best PSQ in the English-French CLIR experiments (see Figure 6). Among the 151 queries, 67 had higher AP with DAMM, 48 had higher AP with PSQ, and the remaining 36 were the same — revealing the difference between them was not due to a small set of topics. Same comparison of the best DAMM and the best PSQ in the English-Chinese CLIR experiments confirmed this finding. There are other variants of cross-language meaning matching, depending on translation in which direction is used and synonymy knowledge in which language is used. For example, a "Probabilistic Document Translation" (PDT) technique which uses document translation knowledge in a similar way as PSQ can be developed; synonymy knowledge in the target language can also be used when only unidirectional translation is considered. We did run experiments for both language pairs and found PDT was at least as effective as PSQ, but adding statistical synonymy knowledge to unidirectional translation could hurt CLIR performance. The latter finding suggests the necessity of combining bidirectional translation with synonymy knowledge. We also compared our meaning matching technique, which basically multiplies translation probabilities, with an earlier approach in which an arithmetic mean was used [20]. Both techniques used bidirectional translation statistics more effectively than unidirectional probabilities. We found, however, when synonym aggregation was used, meaning matching was the more effective technique. Detailed cross-language meaning matching variants and their experimental evaluation can be found in [19].

We want to point out that the interpretation of the statistical significance tests in our study should be cautious. We compared the optimal effectiveness of different meaning matching variants, which is usually achieved at different CPT levels. In an operational system, however, it is hard to tune the parameter without pre-existing knowledge of relevance. Therefore, our findings should only be interpreted as the meaning matching technique *could* potentially outperform one of the best known query translation techniques.

6. CONCLUSIONS AND FUTURE WORK

This paper introduced a general framework for the use of translation probabilities in CLIR. We started with one of the most fundamental issues in IR, the question of how to match what the searcher means with what the document author meant. That naturally pointed us to the direction of translating both queries and documents, or more precisely, using translation knowledge in both directions. Differential polysemy makes statistical translation models by nature asymmetric, and selection of either direction alone would be counterintuitive when matching meanings is the goal. From that key insight, we developed a computational formalism that integrated knowledge about translation and synonymy into a unified model using techniques similar to those previously developed for the probabilistic structured query technique. We then showed that the probabilistic structured query method is a special case of our meaning matching model when only query translation knowledge is used.

Our experiments with an English-French test collection for which a large number of topics are available showed that CLIR using bidirectional translation knowledge together with statistical synonymy significantly outperformed CLIR in which only unidirectional translation knowledge was exploited, achieving CLIR effectiveness comparable to monolingual effectiveness under similar conditions. Despite the big differences between the two language pairs, our experiments on English-Chinese CLIR consistently confirmed these findings, showing the proposed cross-language meaning matching technique is not only effective, but also robust. The importance of the technique and the study lies in it introduces a novel and effective way of using statistical translation knowledge for searching information across language boundaries.

Several things should be considered for improving the proposed model. First, studies in statistical MT have showed that translation based on learned phrases (or "alignment templates") can be more accurate than translation based solely on individual words [14]. A natural next step would

therefore be to integrate phrase translation into our meaning matching model. Second, we only tried the greedy method of aggregation. The method assigns each translation alternative to only one synset. It may also be worth testing other techniques that assign each translation alternative to multiple synsets with some weighting factor, e.g., based on information such as orthographic similarity between the translation and words in each synset. Next, an obvious limitation of our current implementation of meaning matching is its reliance on sentence-aligned parallel corpus, which is necessary for training statistical translation models. Now that our experiments have shown that meaning matching based on bidirectional translation knowledge is quite robust with respect to noisy translations, it might be interesting to see how it performs with translation knowledge obtained from comparable corpora. Finally, decisions for some parameter settings in our study were somewhat arbitrary, e.g., synonyms were cut off at the probability of 0.1, and selections and iterations of IBM Models in statistical MT training were also quite limited. In the future, we plan to explore a broader spectrum of parameter settings, which will hopefully provide us a better and more complete understanding of the cross-language meaning matching framework.

7. ACKNOWLEDGMENTS

The authors would like to thank James Mayfield, Philip Resnik, Vedat Diker, Dagobert Soergel, Jimmy Lin, and all the members of the Computational Linguistics and Information Processing Laboratory at the University of Maryland Institute for Advanced Computer Studies for their valuable comments. This work has been supported in part by DARPA contract N661010028910 (TIDES) and HR0011-06-2-0001 (GALE).

8. REFERENCES

[1] M. Boughanem, C. Chrisment, and N. Nassr. Investigation on disambiguation in CLIR: Aligned corpus and bi-directional translation-based strategies. In *Evaluation of Cross-Language Information Retrieval Systems: Second Workshop of the Cross-Language Evaluation Forum*, pages 158–167. Springer-Verlag GmbH, 2001.

[2] Martin Braschler. Combination approaches for multilingual text retrieval. *Information Retrieval*, 7(1-2):183–204, 2004.

[3] David Chiang, Adam Lopez, Nitin Madnani, Christof Monz, Philip Resnik, and Michael Subotin. The hiero machine translation system: Extensions, evaluation, and analysis. In *Proceedings of HLT/EMNLP 2005*, pages 779–786, 2005.

[4] Kareem Darwish and Douglas W. Oard. Probabilistic structured query methods. In *Proceedings of the 21st Annual 26th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 338–344. ACM Press, July 2003.

[5] In-Su Kang, Seung-Hoon Na, and Jong-Hyeok Lee. POSTECH at NTCIR-4: CJKE monolingual and Korean-related cross-language retrieval experiments. In *Working Notes of the 4th NTCIR Workshop*. National Institute of Informatics, 2004. http://research.nii.ac.jp/ntcir/index–en.html.

[6] Philipp Koehn. Europarl: A multilingual corpus for evaluation of machine translation. unpublished draft. 2002.

[7] Wessel Kraaij. *Variations on Language Modeling on Information Retrieval*. Ph.D. thesis, University of Twente, 2004.

[8] K. L. Kwok. 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*, pages 173–179, 2000.

[9] Victor Lavrenko and W. Bruce Croft. Relevance-based language models. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 120–127. ACM Press, 2001.

[10] Gina-Anne Levow and Douglas W. Oard. Evaluating lexical coverage for cross-language information retrieval. In *Workshop on Multilingual Information Processing and Asian Language Processing*, pages 69–74, February 2000.

[11] J. Scott McCarley. 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*, pages 208–214, 1999.

[12] Douglas W. Oard and Funda Ertunc. Translation-based indexing for cross-language retrieval. In *Proceedings of ECIR'02*, 2002.

[13] F. J. Och and H. Ney. Improved statistical alignment models. In *Proceedings of the 38th Annual Conference of the Association for Computational Linguistics*, pages 440–447, October 2000.

[14] Franz Josef Och and Hermann Ney. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4), 2004.

[15] Ari Pirkola. 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*, pages 55–63. ACM Press, August 1998.

[16] Philip Resnik and David Yarowsky. Distinguishing systems and distinguishing senses: New evaluation methods for word sense disambiguation. *Natural Language Engineering*, 5(2):113–133, 2000.

[17] S. E. Robertson and Karen Sparck-Jones. Simple proven approaches to text retrieval. Cambridge University Computer Laboratory, 1997.

[18] Jacques Savoy. Report on CLEF-2001 experiments: Effective combined query-translation approach. In *Evaluation of Cross-Language Information Retrieval Systems : Second Workshop of the Cross-Language Evaluation Forum*. Springer-Verlag GmbH, 2001.

[19] Jianqiang Wang. *Matching Meaning for Cross-Language Information Retrieval*. Ph.D. thesis, University of Maryland, 2005.

[20] Jinxi Xu and Ralph Weischedel. TREC-9 cross-lingual retrieval at BBN. In *The Nineth Text REtrieval Conference*. National Institutes of Standards and Technology, November 2000. http://trec.nist.gov.