Combining Statistical Translation Techniques for Cross-Language Information Retrieval

Ferhan Ture[1] Jimmy Lin[2,3] Douglas W. Oard[2,3]
(1) Department of Computer Science, University of Maryland, College Park
(2) College of Information Studies, University of Maryland, College Park
(3) UMIACS, University of Maryland, College Park
fture@cs.umd.edu, jimmylin@umd.edu, oard@umd.edu

ABSTRACT
Cross-language information retrieval today is dominated by techniques that rely principally on context-independent token-to-token mappings despite the fact that state-of-the-art statistical machine translation systems now have far richer translation models available in their internal representations. This paper explores combination-of-evidence techniques using three types of statistical translation models: context-independent token translation, token translation using phrase-dependent contexts, and token translation using sentence-dependent contexts. Context-independent translation is performed using statistically-aligned tokens in parallel text, phrase-dependent translation is performed using aligned statistical phrases, and sentence-dependent translation is performed using those same aligned phrases together with an *n*-gram language model. Experiments on retrieval of Arabic, Chinese, and French documents using English queries show that no one technique is optimal for all queries, but that statistically significant improvements in mean average precision over strong baselines can be achieved by combining translation evidence from all three techniques. The optimal combination is, however, found to be resource-dependent, indicating a need for future work on robust tuning to the characteristics of individual collections.

KEYWORDS: cross-language information retrieval, machine translation, context.

Proceedings of COLING 2012: Technical Papers, pages 2685–2702,
COLING 2012, Mumbai, December 2012.

1 Introduction

Cross-Language Information Retrieval (CLIR) is the problem of retrieving documents relevant to a query written in a different language. There are two main approaches to tackle this problem: translating the query into the document language, or translating documents into the query language. Query translation has become the more popular approach for experimental work due to the computational feasibility of trying different system variants without repeatedly translating the entire document collection (Oard, 1998; McCarley, 1999).

Query translation approaches for CLIR can be pursued either by applying a Machine Translation (MT) system or by using a token-to-token bilingual mapping, with or without translation probabilities. These approaches have complementary strengths: MT makes good use of context but at the cost of typically producing only one-best results, while token-to-token mappings can produce *n*-best token translations but without leveraging available contextual clues. This has led to a small cottage industry of what we might refer to as "context recovery" in which postprocessing techniques are used to select or reweight translation alternatives, usually based on evidence from term co-occurrence.

We argue that this false choice between MT and *n*-best token-by-token translation results from thinking of MT systems as black boxes. A modern statistical MT system has internally a series of increasingly rich representations that are exploited during the training and decoding processes. First, token alignments are generated for each training sentence pair, a process that also creates context-independent token-to-token translation probabilities. Second, these alignments are generalized to learn a Synchronous Context-Free Grammar (SCFG), in which probabilistic rules describe the translation of larger units of text. Finally, the translation grammar is combined with a language model to produce translations of entire sentences. As the whole process is statistically generated, it is at any point able to produce a ranked list of the highest scoring translations rather than only the one best choice. Although it is desirable to exploit these internal representations when performing retrieval, one possible disadvantage of using such a complex translation model is efficiency. However, modern decoders, e.g., `cdec` (Dyer et al., 2010), use pruning methods to efficiently search for the most likely translations of a given text.

In this paper, we describe two ways to exploit these internal representations and construct context-sensitive term translation probabilities. One method is to extract a context-aware portion of the SCFG by selecting only the grammar rules that apply to a given query. Using token alignments within each rule, a probability distribution can be constructed to represent the translation candidates for each query token, an approach that we refer to as "phrase-based." Another solution is to perform translation in context using the full MT system on the entire query and then to reconstruct context-sensitive token translation probabilities by accumulating translation likelihood evidence from each of the top *n* query translations.

These context-sensitive token translation probabilities can then be used in the same way as context-independent probabilities. In this work we use a technique based on mapping term statistics before computing term weights (Pirkola, 1998; Darwish and Oard, 2003), leading to a representation known as Probabilistic Structured Queries (PSQ). By doing this, we establish a strong context-independent "token-based" baseline that we can then compare directly with our proposed context-sensitive approaches.

Experiments on TREC 2002, NTCIR-8, and CLEF 2006 CLIR tasks, with topics in English and documents in Arabic, Chinese, and French, respectively, show that our approach consistently yields significant improvements over this baseline. The best results are achieved when we

perform a linear interpolation of all three approaches (query-based, phrase-based, and token-based). The remainder of this paper is organized as follows: Related work is described in Section 2, followed by our proposed approaches in Section 3, evaluation on the three collections in Section 4, and conclusions in Section 5. All of our code and test data are available as part of the open-source Ivory retrieval engine, available at `http://ivory.cc/`.

2 Background and Related work

Drawing inspiration from MT, Ponte and Croft (1998) introduced a monolingual information retrieval approach based on language models. This approach, which models the retrieval problem as if some noisy channel had corrupted some document into the query, was later extended by Berger and Lafferty (1999) and others. Combining language models with translation models to perform CLIR was a natural next step, and that approach yielded substantial improvements over earlier dictionary-based baselines, reporting Mean Average Precision (MAP) scores in the range of 90% of monolingual comparison conditions (Xu and Weischedel, 2005; Kraaij et al., 2003; Federico and Bertoldi, 2002). Nie (2010) summarizes this line of work well.

One limitation of applying language and translation models in CLIR is that they have mostly focused on isolated tokens (i.e., unigram models). To address this, there has been a substantial amount of work on exploiting query context in CLIR, dominated by approaches that use term co-occurrence statistics to select the most appropriate set of translation terms, based on some cohesion measure (Gao et al., 2006; Liu et al., 2005; Adriani and Rijsbergen, 2000; Seo et al., 2005). Expressing term dependency relations explicitly has been shown to produce good results in monolingual retrieval (Gao et al., 2004; Metzler and Croft, 2005), but extending that idea to CLIR has proven not to be as straightforward as one might expect. The closest approximation to be widely explored has been translation of multi-word expressions (so-called "phrase translation," although the "phrases" are often statistical rather than linguistic phenomena) in order to limit polysemy effects (Adriani and Rijsbergen, 2000; Arampatzis et al., 1998; Ballesteros and Croft, 1997; Chen et al., 2000; Meng et al., 2004; Zhang et al., 2007). Gao et al. (2012) recently introduced a query expansion approach also inspired by MT, modeling how query tokens are transformed into document tokens based on query context.

Inside an MT system we find a rich representation of alternative translations of the source "sentence" (which in our case is a query). MT-based CLIR approaches typically use one-best results since it has proven to be convenient to treat MT systems as black boxes (Magdy and Jones, 2011). One early CLIR system did try augmenting MT output using a bilingual dictionary (Kwok, 1999) in order to include alternative translations. More recently, Nikoulina et al. (2012) described techniques to maximize MAP when tuning an MT system and rerank the top n translations. Our approach focuses on combining different sources of evidence within an MT system, and can be considered complementary to these techniques. Combining the n-best derivations is also routinely used in speech retrieval (Olsson and Oard, 2009).

2.1 Context-independent Query Translation

As a baseline, we consider the technique presented by Darwish and Oard (2003). Given a source-language query $s = s_1, s_2, ...$, we represent s in the target language as a Probabilistic Structured Query (PSQ), where each token s_j is represented by its translations in the target language, weighted by the bilingual translation probability. These token-to-token translation probabilities are learned independently from a separate parallel bilingual text using automatic word alignment techniques, and we call this probability distribution Pr_{token}. In this approach,

the score of document d, given source-language query s, is computed by the following equations:

$$\text{Score}(d|s) = \sum_{j=1}^{\#\text{ terms}} \text{BM25}(\text{tf}(s_j, d), \text{df}(s_j)) \quad (1)$$

$$\text{tf}(s_j, d) = \sum_{\{t_i | Pr_{\text{token}}(t_i|s_j) > L\}} \text{tf}(t_i, d) Pr_{\text{token}}(t_i|s_j) \quad (2)$$

$$\text{df}(s_j) = \sum_{\{t_i | Pr_{\text{token}}(t_i|s_j) > L\}} \text{df}(t_i) Pr_{\text{token}}(t_i|s_j) \quad (3)$$

where L is a lower bound on translation probability. We also impose a cumulative probability threshold, C, so that translation alternatives of s_j are added (starting from the most probable ones) until the cumulative probability has reached C. As shown above, we use the Okapi BM25 term weighting function (with parameters $k_1 = 1.2$, $b = 0.75$), although in principle any other weighting function can be substituted.

Let us demonstrate this "token-based" representation model by an example. Following an Indri-like (Metzler and Croft, 2004) notation for query representations, the English query *Maternal leave in Europe* yields the following PSQ under this model for target language French:

```
#comb(#weight(0.74 matern, 0.26 maternel)
      #weight(0.49 laiss, 0.17 quitt, 0.09 cong, ...)
      #weight(0.91 europ, 0.09 européen))
```

Some of the translations are omitted due to space constraints. The `#comb` operator corresponds to the sum operator in equation (1), whereas the `#weight` operator is implemented as the weighted sum in equations (2) and (3). Within the `#weight` structure, terms follow their probabilities, which correspond to the Pr_{token} values in these equations. Notice that the translation distribution for the source token *leave* is uninformed by the context *maternity leave*, therefore the candidates *laisser* (Eng. let go, allow) and *quitter* (Eng. quit) have higher probabilities than *congé* (Eng. vacation, day off) in this model.

2.2 Machine Translation for Cross-Language IR

State-of-the-art statistical MT systems typically use hierarchical phrase-based translation models based on a Synchronous Context-Free Grammar (SCFG) (Chiang, 2005). In an SCFG, the rule `[X] || α || β || 𝒜 || ℓ(α → β)` indicates that the context free expansion $X \rightarrow \alpha$ in the source language occurs synchronously with $X \rightarrow \beta$ in the target language, with a likelihood of $\ell(\alpha \rightarrow \beta)$.[1] In this case, we call α the Left-Hand Side (LHS) of the rule, and β the Right-Hand Side (RHS) of the rule. We use indexed nonterminals (e.g., [X,1]) since in principle more than one nonterminal can appear on the right side. A sequence of token alignments $\mathcal{A}$ indicates which token in α is aligned to which target token in β.

Consider the following four rules from an SCFG:

R_1. `[S] || [S,1] || [S,1]`
R_2. `[S] || [X,1] || [X,1]`
R_3. `[X] || [X,1] leav in europ || cong de [X,1] en europ || 1-0 2-3 3-4 || 1`
R_4. `[X] || matern || matern || 0-0 || 0.69`

In the above notation, S refers to the sentence; therefore, the first two rules are special rules,

[1] The likelihood function ℓ is not a probability density function because it is not normalized.

describing that there is one sentential form, consisting of a single variable. In the third and fourth rules, we see the structure of the English phrase and how it is translated into French.

In contrast to the baseline model, this approach can handle both token and phrase translations. It can consider dependencies between query terms and therefore provide a more context-sensitive and appropriate translation of a given query. On the other hand, it is more dependent on training data and thus may not be as useful when the training set size is limited.

3 Context-Sensitive Query Translation

In this paper, we explore ways to improve the baseline token-translation model discussed above by exploiting the internal representations of the MT system. We describe two ways to construct a context-sensitive probability distribution for each query term, which can then be used directly by a similarly structured PSQ to retrieve ranked documents using equation (1). The first of these techniques (Section 3.1) was described in our previous paper and evaluated on a single collection (Ture et al., 2012); the second approach is new.

3.1 Probabilities from n-best Derivations

In MT, decoding is the process that finds the most probable translation with respect to an SCFG trained on a bilingual parallel corpus and a language model trained on monolingual target-language text. To control computational complexity, most decoders search for the most probable derivations by using pruning strategies.[2] The efficiency of our approaches is discussed in detail in Section 4.

When using one-best query translation, equations (1), (2) and (3) simplify to:

$$\text{Score}(d|s) = \sum_{i=1}^{m} \text{BM25}(\text{tf}(t_i^{(1)}, d), \text{df}(t_i^{(1)})) \tag{4}$$

where $t^{(1)}$ is the most probable translation of s, computed by:

$$\begin{aligned} t^{(1)} &= \arg\max_t [\max_{D \in \mathscr{D}(s,t)} \ell(t, D|s)] = \arg\max_t [\max_{D \in \mathscr{D}(s,t)} \text{TM}(t, D|s)\text{LM}(t)] \\ &= \arg\max_t \left[\text{LM}(t) \max_{D \in \mathscr{D}(s,t)} \prod_{r \in D} \ell(r)\right] \end{aligned} \tag{5}$$

where TM and LM correspond to the translation and language model scores, and $\mathscr{D}(s,t)$ is the set of possible derivations that generates the pair of sentences (s,t) (e.g., the sequence of four rules that translate the example query in Section 2.2 is one such derivation). The likelihood of each grammar rule r, $\ell(r)$, is learned as part of the training process of the translation model, by generalizing from token alignments on the training data (Chiang, 2007).

Decoders produce a set of candidate sentence translations in the process of computing equation (5), so we can generalize our model to consider the n candidates with the highest likelihoods, for some $n > 1$. We start by preprocessing the source query s and each candidate translation $t^{(k)}$. For each source token s_j, we use the derivation output to determine which grammar rules were used to produce $t^{(k)}$, and the token alignments in these rules to determine which target tokens are associated with s_j in the derivation. By doing this for each translation candidate $t^{(k)}$, we construct a probability distribution of possible translations of s_j based on the n query

[2] We use *derivation* to make it clear that what results is a rule sequence, not just a translated string.

translations. Specifically, if source token s_j is aligned to (i.e., translated as) t_i in the k^{th} best translation, the value $\ell(t^{(k)}|s)$ is added to its probability mass, producing the formula for Pr_{nbest}:

$$Pr_{\text{nbest}}(t_i|s_j) = \frac{1}{\varphi} \sum_{\substack{k=1 \\ s_j \text{ aligned to } t_i \text{ in } t^{(k)}}}^{n} \ell(t^{(k)}|s) \quad (6)$$

where φ is the normalization factor.[3] We should emphasize that Pr_{nbest} is a well-defined probability distribution for each s_j, so if a source token is translated consistently into the same target token in all n translations, then it will have a single translation with a probability of 1.0. Mapping *tf* and *df* statistics from source to target vocabulary is achieved by replacing Pr_{token} with Pr_{nbest} in equations (2) and (3).

Pr_{token} and Pr_{nbest} are similar in describing the probability of a target-language token given a source-language token, but differ by how the probability values are learned. For both approaches, we start from a large, potentially out-of-domain, sentence-aligned bilingual corpus. This corpus is first token-aligned using a word aligner. From these token alignments, one can directly deduce token translation probabilities, which correspond to Pr_{token}. In order to learn Pr_{nbest}, we add the MT system as an intermediate component, which creates a translation model from the token alignments, and then applies it (along with a language model) to the query text, using a decoder. Therefore, the distribution is informed by the query context and its derivation.

The advantage of Pr_{token} is the ability to model all of the translational varieties existent in the bilingual corpus, although these may be too noisy to properly translate a given query. For Pr_{nbest}, on the other hand, we would expect the distribution to be biased in favor of appropriate translations, but perhaps at the cost of some reduction in variety due to overfitting to the query context. For comparison, below are the translation probabilities for the same example query:

```
#comb(#weight(0.91 matern, 0.09 maternel, ...)
      #weight(1.0 cong) #weight(1.0 europ))
```

The overfitting issue is partially mitigated by using the n-best translation derivations, as opposed to the "1-best translation" approach, which treats the MT system as a black box. However, the lack of textual variety in the n most probable derivations is a known issue, caused by the fact that statistical MT systems identify the most probable derivations (not the most probable strings), many of which can correspond to the same surface form. This phenomenon is called "spurious ambiguity" in the MT literature, and it occurs in both phrase-based (Koehn et al., 2003) and hierarchical phrase-based MT systems (Chiang, 2007). For instance, according to Li et al. (2009), a string has an average of 115 distinct derivations in Chiang's `Hiero` system. Researchers have proposed several ways to cope with this situation, and we plan to integrate some of these in our future work. However, an alternative approach is to exploit grammar rules directly: this allows us to increase variety without introducing noisy translations, and we discuss this approach next.

3.2 Probabilities from the Translation Grammar

An alternative approach to exploit the MT system is to learn context-sensitive translation probabilities directly from the translation grammar. Hierarchical phrase-based MT systems use suffix arrays to extract all rules in an SCFG which apply to a given source text, requiring a

[3] Since a source token may be aligned to multiple target tokens in the same query translation, we still need to normalize the final likelihood values.

smaller memory footprint in the decoding phase (Lopez, 2007). We can use this feature to learn a token translation probability mapping that is a middle point between Pr_{token} and Pr_{nbest} in terms of context-aware choices and providing a varied set of translation alternatives.

We propose the following method to construct a probability distribution from a set of SCFG rules: For each grammar rule, we use the token alignments to determine which source token translates to which target token(s) in the phrase pair. Going over all grammar rules that apply to a given query, we construct a probability distribution for each token that appears on the LHS.

More specifically, given a translation grammar $\mathcal{G}$ and query s, we first use a suffix array extractor (Lopez, 2007) to obtain the subset of rules $\mathcal{G}(s)$ for which the source side pattern matches s. For each rule r in $\mathcal{G}(s)$, we identify each source token s_j on the LHS of r, ignoring any non-terminal symbols. From the token alignment information included in the rule structure, we can find all target tokens that s_j is aligned to. For each such target token t_i, the likelihood value of s_j being translated as t_i is increased by the likelihood score of r. If there are multiple target tokens, we increase the likelihood of each one equally. After repeating this process for all rules in the subset, we have constructed a list of possible translations and associated likelihood values for each source token that has appeared in any of the rules. We can then convert each list into a probability distribution, Pr_{phrase}, by normalizing the likelihood scores:

$$Pr_{\text{phrase}}(t_i|s_j) = \frac{1}{\psi} \sum_{\substack{r \in \mathcal{G}(s) \\ s_j \leftrightarrow t_i \text{ in } r}} \ell(r) \tag{7}$$

where ψ is the normalization factor and $s_j \leftrightarrow t_i$ represents an alignment between tokens s_j and t_i. Pr_{phrase} is different than Pr_{token} because it takes query context into account. Basically, we only look at the part of the grammar that applies to phrases in the source query, therefore create a bias in the probability distribution based on this context. It is also different than Pr_{nbest} because we do not perform any search, and there is no use of a language model. Thinking in terms of the MT pipeline, the representation we are exploiting in this approach is the extracted grammar, right before decoding has taken place, after token alignments have been generated. In order to illustrate the intuition behind this approach, the same example query is represented as follows using Pr_{phrase}:

```
#comb(#weight(0.68 matern, 0.06 maternel, ...)
      #weight(0.35 cong, 0.24 laiss, 0.13 quitt, ...)
      #weight(0.90 europ, 0.07 européen, ...))
```

When compared to Pr_{token}, notice that the translation distribution of *leave* shifts towards the more appropriate translation *congé* as a result of this approach.

3.3 Combining Sources of Evidence

All three approaches for query translation (i.e., token-based, phrase-based, and query-based) have complementary strengths, so we introduce a unified CLIR model by performing a linear interpolation of the three probability distributions:

$$Pr_{\text{c}}(t_i|s_j;\lambda_1,\lambda_2) = \lambda_1 Pr_{\text{nbest}}(t_i|s_j) + \lambda_2 Pr_{\text{phrase}}(t_i|s_j) + (1-\lambda_1-\lambda_2) Pr_{\text{token}}(t_i|s_j) \tag{8}$$

Replacing Pr_{token} with Pr_{c} in equation (1) gives us the document scoring formula for the combined model (call Score_{c}).

Until now, we focused on the translation of single-token terms, but we can also use the n-best derivation list to identify how multi-token "phrases" are translated in context. The right hand side of the rules in the n most probable derivations provides us with statistically meaningful target-language phrases, along with their associated probabilities (described by Pr_{multi} below). With this addition, we score each document by a weighted average of the single-token approach (i.e., Score_c) and the sum of document scores for the multi-token terms:

$$\text{Score}(d|s;\gamma) = \gamma \text{Score}_{\text{c}}(d|s;\lambda_1,\lambda_2) + (1-\gamma) \sum_{\text{phrase } p} \text{BM25}(\text{tf}(p,d), \text{df}(p)) Pr_{\text{multi}}(p) \quad (9)$$

$$Pr_{\text{multi}}(p) = \frac{1}{\psi} \sum_{k=1}^{n} \sum_{\substack{\text{rule } r \in D^{(k)} \\ p \in RHS(r)}} \ell(r) \quad (10)$$

where ψ is the normalization factor and $D^{(k)}$ is the derivation of the k^{th} best translation. Below is the representation of the example query under this model, with γ set to 0.8:

```
#combweight(0.8 #comb(#weight(0.81 matern, 0.12 maternel, ...)
                      #weight(0.45 cong, 0.25 laiss, 0.10 quitt, ...)
                      #weight(0.95 europ, 0.04 européen, ...))
            0.1 ''en europ'', 0.08 ''cong de'', 0.01 ''cong matern'', ...)
```

The `#comb` structure represents Pr_{c} and the remaining multi-token terms represent Pr_{multi}, all extracted from the top n derivations. The `#combweight` operator corresponds to equation (9).

4 Evaluation

We evaluated our system on the latest available CLIR test collections for three languages: TREC 2002 English-Arabic CLIR, NTCIR-8 English-Chinese Advanced Cross-Lingual Information Access (ACLIA), and CLEF 2006 English-French CLIR. For the Arabic and French collections, we used title queries because they are most representative of the short queries that searchers frequently pose to web search engines. Chinese queries in the NTCIR-8 ACLIA test collection are in the form of complete syntactically correct questions, but for consistency we treated them as bag-of-words queries in our experiments with no special processing. The collections contain 383,872, 388,589 and 177,452 documents, and 50, 50, and 73 topics, respectively.

We learned our English-to-Arabic translation model using 3.4 million aligned sentence pairs from the GALE 2010 evaluation. Our English-to-Chinese translation model was trained on 302,996 aligned sentence pairs from the FBIS parallel text collection. We trained an English-to-French translation model using 2.2 million aligned sentence pairs from the latest Europarl corpus (version 7) that was built from the European parliament proceedings.[4]

Token alignments were learned with GIZA++ (Och and Ney, 2003), using 5 Model 1 and 5 HMM iterations. An SCFG serves as the basis for the translation model (Chiang, 2007), which was extracted from these token alignments using a suffix array (Lopez, 2007). We used `cdec` for decoding, due to its support for SCFG-based models and its efficient C-based implementation, making it faster than most of the other state-of-the-art systems (Dyer et al., 2010). A 3-gram language model was trained from the target side of the training data for Chinese and Arabic, using the SRILM toolkit (Stolcke, 2002). For French, we trained a 5-gram LM from the monolingual dataset provided for WMT-12. The Chinese collection was segmented using the Stanford segmenter (Tseng et al., 2005), English topics and the French collection

[4] http://www.statmt.org/europarl

were tokenized using the OpenNLP tokenizer,[5] and Arabic was tokenized and stemmed using the Lucene package.[6] For English and French, we also lowercased text, stemmed using the Snowball stemmer, and removed stopwords.

4.1 Effectiveness

We used Mean Average Precision (MAP) as the evaluation metric. The baseline token-based model yields a MAP of 0.2712 for Arabic, 0.1507 for Chinese, and 0.2617 for French. Direct comparisons to results reported at TREC, NTCIR, and CLEF (respectively) are hard to make because of differences in experimental conditions, but the comparisons we are able to make suggest that these baseline MAP values are reasonable.[7] For Arabic, the best reported results from TREC-2002 were close to 0.40 MAP (Fraser et al., 2002), but those results were achieved by performing query expansion and learning stem-to-stem mappings; our experiment design requires token-to-token mappings (which result in sparser alignments). For Chinese, the NTCIR-8 topics are in the form of questions, and systems that applied question rewriting performed better than those that did not. Also, 15 of the questions are about people, for which our vocabulary coverage was not tuned. If we disregard these 15 topics, our baseline system achieves 0.1778, close to the best reported results with comparable settings, with a MAP of 0.181 (Zhou and Wade, 2010). For French, our baseline achieves close to the same score as the one reported result at CLEF-2006 that did not incorporate blind relevance feedback (0.2606 MAP) (Savoy and Abdou, 2006).

As discussed in Section 3, we implemented three techniques to construct a term translation probability distribution: Pr_{nbest}, Pr_{phrase} and Pr_{token}, described by equations (6), (7) and (1) above.[8] We assessed these three approaches by (i) comparing them against each other, and (ii) measuring the benefit of a linear combination, i.e., Pr_{c}, described by equation (8).

Experiment results are summarized in Table 1 and illustrated in Figure 1. In that figure, we provide three connected scatterplots of MAP scores within a range of values for λ_1 and λ_2. In order to see the effectiveness of the interpolated model with respect to parameters λ_1 and λ_2, we performed a grid search by applying values in increments of 0.1 (ranging from 0 to 1) to the interpolated model Pr_{c}. For readability, figures only include a representative subset of λ_2 settings, where different lines represent different values for λ_2. To distinguish the extreme settings of $\lambda_2 = 0$ and $\lambda_2 = 1$, we use a filled circle or square, respectively.

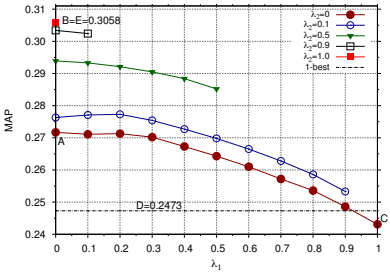
The left edge represents $\lambda_1 = 0$, meaning that we do not use probabilities learned from the n-best derivations (i.e., Pr_{nbest}) in our interpolation. Along the y-axis on the left edge, we see results for various settings of λ_2, which controls how much weight is put on Pr_{phrase} and Pr_{token}. Within these settings, a particularly interesting one is when λ_2 is set to 0. In this case, the approach is solely based on context-independent translation probabilities (i.e., Pr_{token}), which is the baseline model (call this condition A). When λ_2 is set to 1, we rely on phrase-based term translation probabilities (i.e., Pr_{phrase}, call this condition B). By contrast, at the right edge, $\lambda_1 = 1$, so we rely only on Pr_{nbest} when translating query terms (call this condition C). For

[5] http://opennlp.apache.org

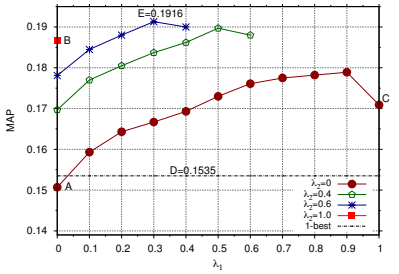
[6] http://lucene.apache.org

[7] The best results often employ blind relevance feedback, multiple lexical resources and/or very long queries. While these techniques can be useful in deployed applications, we have chosen not to run such conditions in order to avoid masking the effects that we wish to study.

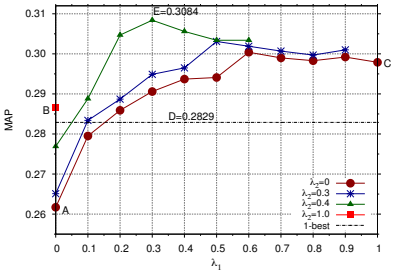
[8] We fixed $C = 0.95, L = 0.005, n = 10$ for all models after manually trying a range of values.



(a) TREC 2002 English-Arabic CLIR task



(b) NTCIR-8 English-Chinese CLIR task



(c) CLEF 2006 English-French CLIR task

Figure 1: Effectiveness results.

reference, the dotted horizontal line represents simply taking the one-best translation from the MT system (i.e., described by equation (4), call this condition D).

In the case of the Arabic collection, we observe a strictly decreasing trend for the MAP scores as λ_2 decreases, and the best results are obtained when λ_1 is 0 and λ_2 is 1.0 (call the condition with the best MAP score E). In other words, the interpolation yields a maximum 0.3058 MAP when it was based entirely on Pr_{phrase}, ignoring distributions Pr_{token} and Pr_{nbest}. For the Chinese collection, λ_1=0.2 and λ_2=0.7 yields the best result (MAP=0.1916), whereas effectiveness peaks at λ_1=0.3 and λ_2=0.4 for the French collection, with a MAP score of 0.3084.

Based on the randomized significance test proposed by Smucker et al. (2007), the combined approach (E) outperforms all models (except for the phrase-based approach) in the Arabic collection with 95% confidence. When we ran the same test on the other two collections, we found that the combined approach is significantly better than the baseline (A) and 1-best (D) approaches for Chinese, whereas MAP is significantly higher than baseline A for French. These results confirm that the complementary advantages of each model can be combined into a single superior model using our approach.

We also experimented with the multi-token term representation in equation (9), by varying the γ parameter. With γ set empirically to 0.8, the MAP increased by 0.005 for French, and remained about the same for Arabic and Chinese.

When the three individual models (conditions A, B and C) are compared (i.e., ignoring the

Condition	MAP		
	Arabic	Chinese	French
A: λ_1=0, λ_2=0 (Pr_{token})	0.2712	0.1507	0.2617
B: λ_1=0, λ_2=1 (Pr_{phrase})	0.3058	0.1867	0.2868
C: λ_1=1, λ_2=0 (Pr_{nbest})	0.2431	0.1709	0.2979
D: 1-best	0.2473	0.1535	0.2829
E: best $\{\lambda_1, \lambda_2\}$	$0.3058^{a,c,d}$	$0.1916^{a,d}$	0.3084^{a}

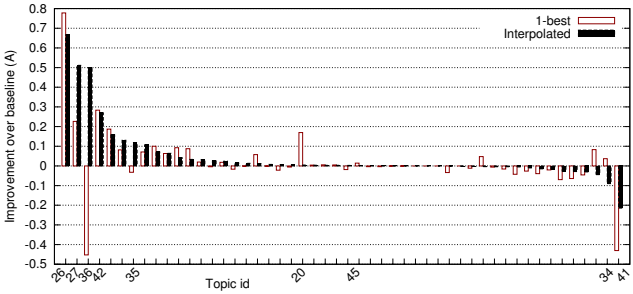
Table 1: A summary of experimental results under different conditions, for all three CLIR tasks. Superscripts indicate if the best result is significantly better than conditions A, B, C, and D.

interpolated results), the phrase-based model (B) is significantly better than the token-based baseline (A) for Arabic and Chinese, but statistically indistinguishable from the same baseline model in the case of French. For French, the best retrieval effectiveness results from the *n*-best full query translation model (C), significantly better than the baseline model (A). This shows that there is no individual model that outperforms the rest in all three collections. The real strength of our approach is therefore to introduce a unified probabilistic model that can combine all of these different approaches in a principled manner.

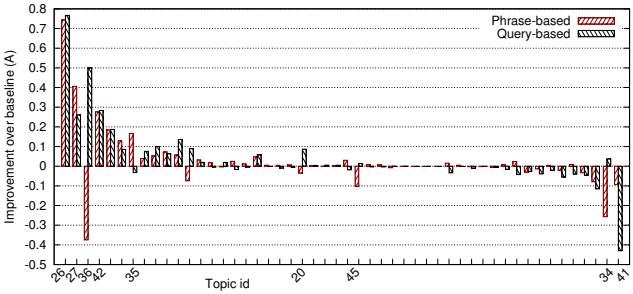
For topic-specific analysis, we looked at the distribution of the average precision (AP) differences between the various models. We observe that our best interpolated model (E) yields better AP than the token-based baseline model (A) for 36 of the 43 Arabic topics in which there was a noticeable difference (7 of the 50 Arabic topics exhibited differences of 0.001 or less). For the Chinese collection, the same was true for 41 of 57 topics (with 16 exhibiting a negligible difference), whereas the comparable statistic is 29 of 44 for French.

For space reasons, we illustrate these differences only for French. Figure 2(a) plots the AP improvement of the best interpolated model (E) and the one-best MT approach (D) over (or the average degradation below) the token-based baseline (A), sorted left to right by decreasing AP improvement for the interpolated model (E); Figure 2(b) similarly plots the same AP differences for the phrase-based (B) and *n*-best full query translation approaches (C), again with reference to the token-based baseline (A), with topics sorted in the same order to facilitate comparison. These plots make it quite clear that the three approaches vary in their per-topic effectiveness. Rather than slight variations in all of the topics, we see several cases in which one of the models is superior to the others. For instance, the *n*-best full query translation approach is a clear winner for topics 26, 36 and 42, whereas the phrase-based approach outperforms for topics 27 and 35. Despite its drawbacks, there are topics in which the token-based model is superior to our more sophisticated approaches, notably topics 34 and 41. Once again, this analysis supports our argument that combining these three probabilistic models into one unified approach can capture some of the best of each. In general, we expect the interpolated model to be more robust, since it has access to more evidence than the individual models.

In practice, we would like to select model parameters without observing all the test topics. Therefore, we ran 10-fold cross-validation experiments on each collection, selecting parameters that maximize MAP on nine folds and evaluating on the remaining one. This method yields a MAP of 0.2979 for Arabic, 0.1733 for Chinese, and 0.2872 for French, all significantly better than the token-based baseline (A). We also explored if we could use two of the collections to tune parameters for the third. For this, we first ranked each (λ_1, λ_2) pair by MAP on each collection. In order to select the parameters for a particular collection, we added the ranks



(a) Interpolated and 1-best models vs. the token-based baseline approach.



(b) Phrase- and query-based (i.e., n-best) models vs. the token-based baseline approach.

Figure 2: Per-topic AP improvement over token-based baseline (condition A) for French.

from the other two collections and picked the one with the lowest sum. Using this method, the selected parameters were $(0.1, 0.1)$ for Arabic, $(0.3, 0.5)$ for Chinese, and $(0.1, 0.1)$ for French. When compared to the token-based baseline (A), this approach showed significant improvements only for Chinese. We conclude from this analysis that the optimal combination of models depends on the collection, language, and resources. Once these are fixed, we can use a subset of the topics to appropriately tune parameters for the rest. However, better tuning methods need to be devised for a truly robust approach to combining these CLIR models.

4.2 Efficiency

We compared the various CLIR approaches in terms of efficiency (query evaluation time), performing experiments on a machine running Red Hat Linux on a 2.4 GHz processor. We processed the Arabic topics using each model and measured running time per query in milliseconds. Averages over three repeated runs are reported in Table 2 (with 95% confidence intervals).

As described before, there are three processes in the MT pipeline: token alignment, grammar extraction, and decoding. Token alignment is query-independent and required for all three approaches, so we did not include it in our running time comparison of running times. For the construction of Pr_{phrase}, we only need to extract grammar rules that apply to each given query,

	Process	Pr_{token}	Pr_{phrase}	Pr_{nbest}			Pr_{c}
				1-best	**5-best**	**10-best**	
MT	Grammar extraction	-	7.57				
	Decoding	-	-	134.94			134.94
IR	Initialization	negligible	64.38	negligible			64.38
	Generation	48.12	negligible	5.80	59.47	62.25	49.11
	Ranking	545.64	514.17	97.64	158.81	179.07	601.95
Total time (in *ms*)		594±22	586±13	246±15	361±28	383±22	858±20

Table 2: Average running times for processes in the CLIR pipeline (in *ms*).

whereas Pr_{nbest} also requires decoding.[9]

The remaining processes that we need to consider are part of the IR pipeline: initialization of the CLIR model, generation of query representations in the target language, and ranking of the most relevant documents in the collection. We only count query-dependent initialization costs, since other costs such as loading the bilingual dictionary need to be done only once, even with many queries. The input of the generation step is the source-language query, and the output is a PSQ that represents that query in the target language. In the phrase-based method, this step takes a negligible amount of time, because the probability distribution is already in memory at the beginning of this step, and it is very small (i.e., probabilities for a few query terms only). For Pr_{nbest}, generation time rises linearly as n is increased.

Ranking time depends on the complexity of the query representation. With more complex representations, it is possible to increase effectiveness, but at the cost of efficiency. Therefore, a desirable CLIR approach would express all the relevant information and nothing more. The distributions Pr_{token} and Pr_{phrase} tend to include more translation alternatives per query term, resulting in a more complex representation and longer ranking time. As a result, interpolating all three distributions generates a complex representation as well.

When we look at the total running times in Table 2, we observe that the n-best approach is significantly more efficient than the token-based baseline, even though it requires additional MT processes to fully translate the queries. When $n = 1$, the reduction in total running time is nearly 60%. The savings become more modest as n increases, approximately 39% and 35% for 5-best and 10-best MT approaches. Increasing n also improves effectiveness, thus there is a tradeoff to consider when deciding on the value for n. There is a similar tradeoff for the token-based approach: the representation can be simplified if more aggressive thresholding is used, e.g., if C increased in equation (1); however, this may result in a less effective model.

We do not see the same efficiency improvements from reduction in query complexity with the phrase-based model; the query complexity is similar to the baseline approach. As a result, the phrase-based approach runs in about the same total time. However, the MAP score improves considerably for all of the collections, so we can say that Pr_{phrase} is superior to Pr_{token}.

The combined model Pr_{c} yields the highest MAP scores but also takes the longest time to complete. When compared to the baseline model, running time increases by 44%, which seems acceptable given the consistently significant improvements. We should note that our implementation is not fully optimized, and is open to further improvements in the future.

As a summary of our evaluation, we believe that the best choice depends on user expectations. For a faster and possibly more effective model, Pr_{nbest} and Pr_{phrase} seem to be good alternatives

[9] It is reasonable to assume that the decoder time to find top n translations is the same as finding the one-best result.

to Pr_{token}. For best effectiveness, the interpolation of the three probability distributions is a good choice, providing significantly better results at the cost of additional complexity.

5 Conclusions and Future Work

In this paper, we introduced a theoretical framework that uses a statistical translation model for cross-language information retrieval. Our approach combines the representational advantage of probabilistic structured queries with the richness of the intermediate information produced by translation models. We proposed two ways of exploiting the internal representation of translation models to learn context-sensitive term translation probabilities: (1) aggregate information from the *n*-best translation outputs by an MT decoder, or (2) extract the subset of the translation grammar that applies to a given query, and use the token alignments in each rule to construct a probability distribution. Although using translation models for CLIR is not a novel approach, we have introduced novel ideas on *how* one can utilize the rich internal representation of MT systems for this task.

We evaluated our models on an English-Arabic task from TREC 2002, an English-Chinese task from NTCIR-8, and an English-French task from CLEF 2006, finding in all three cases that an optimal linear combination of the three approaches can significantly improve MAP, but that the optimal parameters vary by collection. We also compared approaches in terms of efficiency and showed that our framework provides a set of choices, allowing a beneficial tradeoff between improving efficiency and effectiveness. Because we used only one collection per language, experiments with multiple collections for the same language will be needed before we can begin to speculate on whether these differences are language-dependent, collection-dependent, or some combination of the two. Additionally, we would like to try this approach on more languages to further study the consistency in improvements, and also with different parallel corpora and monolingual language modeling collections, in order to tease out whether the differences we are seeing in the optimal combination weights are resource dependent (varying principally with different parallel corpora and/or language models).

In terms of modeling, we plan to revisit the rather ad hoc way we have incorporated multi-word expressions, exploring ways of leveraging them in each model separately rather than at the final evidence combination stage. Also, since the benefit of performing full machine translation would be expected to increase as available context increases, we would like to explore the potential for translating documents in addition to queries. Following the same methods described in this paper, we could learn a new set of probability distributions from the document translations, which could be combined with the current three approaches to construct an even richer and possibly more accurate CLIR model. We also plan to explore the effect of using a phrase-based MT system as an alternative to the SCFG-based model in our experiments.

In conclusion, we have introduced ways of using statistical translation models for CLIR that take greater advantage of the capabilities of current statistical MT systems, and we hope that the promising results we have reported will spur the community to further explore this space.

Acknowledgments

This research was supported in part by the BOLT program of the Defense Advanced Research Projects Agency, Contract No. HR0011-12-C-0015; NSF under awards IIS-0916043 and IIS-1144034. Any opinions, findings, conclusions, or recommendations expressed in this paper are those of the authors and do not necessarily reflect views of the sponsors. The second author is grateful to Esther and Kiri for their loving support and dedicates this work to Joshua and Jacob.

References

Adriani, M. and Rijsbergen, C. J. V. (2000). Phrase identification in cross-language information retrieval. In *Proceedings of RIAO 2000*.

Arampatzis, A. T., Tsoris, T., Koster, C. H. A., and Weide, P. V. D. (1998). Phrase-based information retrieval. *Information Processing & Management*, 34(6):693–707.

Ballesteros, L. and Croft, W. B. (1997). Phrasal translation and query expansion techniques for cross-language information retrieval. *SIGIR Forum*, 31:84–91.

Berger, A. and Lafferty, J. (1999). Information retrieval as statistical translation. In *Proceedings of SIGIR 1999*, pages 222–229.

Chen, A., Jiang, H., and Gey, F. (2000). English-Chinese cross-language IR using bilingual dictionaries. In *Proceedings of TREC-9*.

Chiang, D. (2005). A hierarchical phrase-based model for statistical machine translation. In *Proceedings of ACL 2005*, pages 263–270.

Chiang, D. (2007). Hierarchical phrase-based translation. *Computational Linguistics*, 33:201–228.

Darwish, K. and Oard, D. W. (2003). Probabilistic structured query methods. In *Proceedings of SIGIR 2003*, pages 338–344.

Dyer, C., Weese, J., Setiawan, H., Lopez, A., Ture, F., Eidelman, V., Ganitkevitch, J., Blunsom, P., and Resnik, P. (2010). cdec: a decoder, alignment, and learning framework for finite-state and context-free translation models. In *Proceedings of the ACL 2010 System Demonstrations*, pages 7–12.

Federico, M. and Bertoldi, N. (2002). Statistical cross-language information retrieval using n-best query translations. In *Proceedings of SIGIR 2002*, pages 167–174.

Fraser, A., Xu, J., and Weischedel, R. (2002). TREC 2002 cross-lingual retrieval at BBN. In *Proceedings of TREC-11*.

Gao, J., Nie, J.-Y., Wu, G., and Cao, G. (2004). Dependence language model for information retrieval. In *Proceedings of SIGIR 2004*, pages 170–177.

Gao, J., Nie, J.-Y., and Zhou, M. (2006). Statistical query translation models for cross-language information retrieval. *ACM Transactions on Asian Language Information Processing (TALIP)*, 5:323–359.

Gao, J., Xie, S., He, X., and Ali, A. (2012). Learning lexicon models from search logs for query expansion. In *Proceedings of EMNLP 2012*, pages 666–676.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of NAACL-HLT 2003*, pages 48–54.

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

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

Li, Z., Eisner, J., and Khudanpur, S. (2009). Variational decoding for statistical machine translation. In *Proceedings of ACL 2009*, pages 593–601.

Liu, Y., Jin, R., and Chai, J. Y. (2005). A maximum coherence model for dictionary-based cross-language information retrieval. In *Proceedings of SIGIR 2005*, pages 536–543.

Lopez, A. (2007). Hierarchical phrase-based translation with suffix arrays. In *Proceedings of EMNLP-CoNLL 2007*, pages 976–985.

Magdy, W. and Jones, G. J. F. (2011). Should MT systems be used as black boxes in CLIR? In *Proceedings of ECIR 2011*, pages 683–686.

McCarley, J. S. (1999). Should we translate the documents or the queries in cross-language information retrieval? In *Proceedings of ACL 1999*, pages 208–214.

Meng, H. M., Chen, B., Khudanpur, S., Levow, G.-A., Lo, W. K., Oard, D. W., Schone, P., Tang, K., Wang, H.-M., and Wang, J. (2004). Mandarin-English information (MEI): Investigating translingual speech retrieval. *Computer Speech & Language*, 18(2):163–179.

Metzler, D. and Croft, W. B. (2004). Combining the language model and inference network approaches to retrieval. *Information Processing & Management*, 40:735–750.

Metzler, D. and Croft, W. B. (2005). A Markov random field model for term dependencies. In *Proceedings of SIGIR 2005*, pages 472–479.

Nie, J.-Y. (2010). *Cross-Language Information Retrieval*. Morgan & Claypool Publishers.

Nikoulina, V., Kovachev, B., Lagos, N., and Monz, C. (2012). Adaptation of statistical machine translation model for cross-language information retrieval in a service context. In *Proceedings of EACL 2012*.

Oard, D. W. (1998). A comparative study of query and document translation for cross-language information retrieval. In *Proceedings of AMTA 1998*, pages 472–483.

Och, F. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Olsson, J. S. and Oard, D. W. (2009). Combining LVCSR and vocabulary-independent ranked utterance retrieval for robust speech search. In *Proceedings of SIGIR 2009*, pages 91–98.

Pirkola, A. (1998). The effects of query structure and dictionary-setups in dictionary-based cross-language information retrieval. In *Proceedings of SIGIR 1998*, pages 55–63.

Ponte, J. M. and Croft, W. B. (1998). A language modeling approach to information retrieval. In *Proceedings of SIGIR 1998*, pages 275–281.

Savoy, J. and Abdou, S. (2006). Experiments with monolingual, bilingual, and robust retrieval. In *Proceedings of CLEF 2006*, pages 137–144.

Seo, H.-C., Kim, S.-B., Rim, H.-C., and Myaeng, S.-H. (2005). Improving query translation in English-Korean cross-language information retrieval. *Information Processing & Management*, 41(3):507–522.

Smucker, M. D., Allan, J., and Carterette, B. (2007). A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of CIKM 2007*, pages 623–632.

Stolcke, A. (2002). SRILM - an extensible language modeling toolkit. In *Proceedings of ICSLP 2002*, pages 901–904.

Tseng, H., Chang, P.-C., Andrew, G., Jurafsky, D., and Manning, C. (2005). A conditional random field word segmenter. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*.

Ture, F., Lin, J., and Oard, D. W. (2012). Looking inside the box: context-sensitive translation for cross-language information retrieval. In *Proceedings of SIGIR 2012*, pages 1105–1106.

Xu, J. and Weischedel, R. (2005). Empirical studies on the impact of lexical resources on CLIR performance. *Information Processing & Management*, 41:475–487.

Zhang, W., Liu, S., Yu, C., Sun, C., Liu, F., and Meng, W. (2007). Recognition and classification of noun phrases in queries for effective retrieval. In *Proceedings of CIKM 2007*, pages 711–720.

Zhou, D. and Wade, V. (2010). The effectiveness of results re-ranking and query expansion in cross-language information retrieval. In *Proceedings of NTCIR-8 Workshop Meeting*.