Exploiting Representations from Statistical Machine Translation for Cross-Language Information Retrieval

FERHAN TURE, Raytheon BBN Technologies
JIMMY LIN, University of Maryland at College Park

This work explores how internal representations of modern statistical machine translation systems can be exploited for cross-language information retrieval. We tackle two core issues that are central to query translation: how to exploit context to generate more accurate translations and how to preserve ambiguity that may be present in the original query, thereby retaining a diverse set of translation alternatives. These two considerations are often in tension since ambiguity in natural language is typically resolved by exploiting context, but effective retrieval requires striking the right balance. We propose two novel query translation approaches: the *grammar-based* approach extracts translation probabilities from translation grammars, while the *decoder-based* approach takes advantage of *n*-best translation hypotheses. Both are *context-sensitive*, in contrast to a baseline *context-insensitive* approach that uses bilingual dictionaries for word-by-word translation. Experimental results show that by "opening up" modern statistical machine translation systems, we can access intermediate representations that yield high retrieval effectiveness. By combining evidence from multiple sources, we demonstrate significant improvements over competitive baselines on standard cross-language information retrieval test collections. In addition to effectiveness, the efficiency of our techniques are explored as well.

Categories and Subject Descriptors: H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Retrieval models*

General Terms: Algorithms, Experimentation

ACM Reference Format:
Ferhan Ture and Jimmy Lin. 2014. Exploiting representations from statistical machine translation for cross-language information retrieval. ACM Trans. Inf. Syst. 32, 4, Article 19 (October 2014), 32 pages.
DOI: http://dx.doi.org/10.1145/2644807

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 tackling this challenge: translating the query into the document language or translating documents into the query language. Query translation has become the more popular approach for experimental studies due to the computational feasibility of trying different system variants without repeatedly translating the entire document collection [Oard 1998; McCarley 1999]; it is also the approach we adopt in this work.

This research was supported in part by the BOLT program of the Defense Advanced Research Projects Agency (Contract No. HR0011-12-C-0015), and by the NSF under awards IIS-0916043 and IIS-1144034. Any opinions, findings, conclusions, or recommendations expressed are those of the authors and do not necessarily reflect views of the sponsors.
F. Ture was at the University of Maryland at the time of this research.
Authors' addresses: F. Ture (corresponding author), Raytheon BBN Technologies, 10 Moulton St., Cambridge, MA 02138; email: fture@bbn.com; J. Lin, The iSchool, College of Information Studies, University of Maryland at College Park; email: jimmylin@umd.edu.

DOI: http://dx.doi.org/10.1145/2644807

There are currently two popular approaches to query translation for CLIR: one could translate the query by taking advantage of a machine translation (MT) system, or alternatively, one could perform word-by-word translation, most often using a bilingual dictionary induced automatically from parallel text. These approaches have complementary strengths: MT makes good use of context but at the cost of typically producing only one-best results—in other words, it "collapses" ambiguity during translation and eliminates diversity in translation choices. On the other hand, bilingual dictionaries can easily produce multiple translations, thus providing diverse translation alternatives that preserve ambiguities present in the original source—but such techniques often have difficulty leveraging available contextual clues. There is a tension between these two considerations, since ambiguity in natural language is usually resolved by taking context into account. Reducing ambiguity allows for more focused retrieval, but there are downsides to trying to eliminate ambiguity completely: first, there is the danger of brittleness, if the system's single interpretation is incorrect; second, information needs are often vague and ill-defined, thus rendering disambiguation a futile endeavor. These are themes we will return to throughout this article, and a high-level summary of our contribution is an exploration of how to best balance context and ambiguity/diversity given the range of modern techniques in statistical machine translation.

We argue that query translation using either single-best MT output or bilingual dictionaries is a false choice that stems from thinking of MT systems as black boxes. Internally, a modern statistical machine translation system builds a series of increasingly-rich intermediate representations that can be exploited for cross-language information retrieval. The contribution of this work is "opening up" such a system and exploring the extent to which its internal representations can be leveraged for CLIR. In particular, the following.

(1) Modern statistical machine translation systems begin by performing word alignment on parallel text, the output of which is context-independent word-to-word translation probabilities. These distributions serve as a query translation baseline in CLIR using what we call the *word-based* approach.
(2) From word alignments we can generate translation grammars in the form of rules that describe the translation of larger textual units. Two popular types of grammars are phrase-based grammars [Koehn et al. 2003; Och and Ney 2004; Marcu and Wong 2002] and hierarchical phrase-based grammars [Chiang 2005, 2007]. These translation grammars capture context beyond individual words, but differ in complexity, expressivity, and translation quality. We present novel techniques for exploiting translation grammars for CLIR using what we refer to as the *grammar-based* approach.
(3) The translation grammar (flat or hierarchical) can be combined with a language model to produce complete translations (a process called *decoding*). Although many applications only use the single-best hypothesis, MT systems can generate n-best hypotheses with equal ease. We present novel techniques for renormalizing n-best MT output for CLIR using what we refer to as the *decoder-based* approach.
(4) Finally, all three approaches can be integrated in an interpolated model that combines evidence from different sources.

We characterize the second and third approaches as *context-sensitive* translation, in contrast to the first approach, which is *context-independent*. Successive stages in the standard MT pipeline capture different trade-offs between context and ambiguity (and by extension, diversity). Translation grammars go beyond single words to model multiword phrases and long-distance dependencies; decoding adds contextual information from the language model. However, at each stage, the diversity of translation alternatives is reduced, such that at the end of the MT pipeline, we are left with only n-best

translations. Our study explores the trade-off between these two factors in the context of CLIR effectiveness as well as efficiency. This work pulls together results that have been incrementally reported in a SIGIR poster [Ture et al. 2012b], a COLING conference paper [Ture et al. 2012a], a SIGIR short paper [Ture and Lin 2013], and a Ph.D. dissertation [Ture 2013]. We seek to synthesize results from these previous publications in a manner that has not yet been accomplished. We summarize our findings as follows.

—Experiments on three test collections in different languages consistently show that context-sensitive models are more effective than the context-independent baseline.
—When comparing the grammar-based approach and the decoder-based approach, the former is more effective overall for two of the three collections. However, the interpolated model is the most effective but requires training data for parameter tuning.
—With the grammar-based approach, hierarchical translation grammars yield higher retrieval effectiveness than flat translation grammars, but this advantage appears to disappear with the decoder-based approach. The greater expressivity of hierarchical grammars and their ability to capture long-distance dependencies does not appear to be important once a language model is introduced due to the short length of queries.
—The grammar-based approach exhibits less per-topic variation than the decoder-based approach—for some topics, the decoder-based approach fails spectacularly. Here, again, the interpolated model is the most effective overall—it is almost always better than the worst individual approach and often close to the best individual approach.
—In terms of striking the right balance between effectiveness and efficiency, both the grammar-based approach and decoder-based approach represent good options. While the interpolated model is more effective, it is also much slower due to the complexity of the queries generated.

The remainder of the article proceeds in the following manner: First, we begin with an overview of how modern statistical machine translation systems work in Section 2, followed by a discussion of related work in Section 3. Next, we describe the context-independent word-based approach for query translation in CLIR: the technique of Darwish and Oard [2003] provides an experimental baseline (Section 4). We continue in Section 5 by detailing our two novel context-sensitive query translation approaches: one that leverages the translation grammar and another that takes advantage of the decoder. Our experimental setup is described in Section 6, and results are presented in Section 7, where we analyze not only the effectiveness but also the efficiency of our techniques. We conclude by discussing limitations and future work.

2. A PRIMER ON STATISTICAL MACHINE TRANSLATION

In modern statistical machine translation, the translation process is modeled as a noisy channel: we assume some sentence t was generated in the target language but then got "garbled" into the source language, so that the task is to recover the most probable sentence t^* that explains the source sentence s. The process consists of two components: (i) generation of the target-language sentence, that is, $P(t)$, and (ii) translation from target to source language, that is, $P(s|t)$. By Bayes' Theorem,

$$P(t|s) = \frac{P(s|t)P(t)}{P(s)}. \tag{1}$$

Given s, we want to compare possible translation hypotheses, and thus we can safely drop the denominator $P(s)$. The remaining two processes $P(s|t)$ and $P(t)$ are usually modeled separately, commonly referred to as the *translation model* and *language model*,

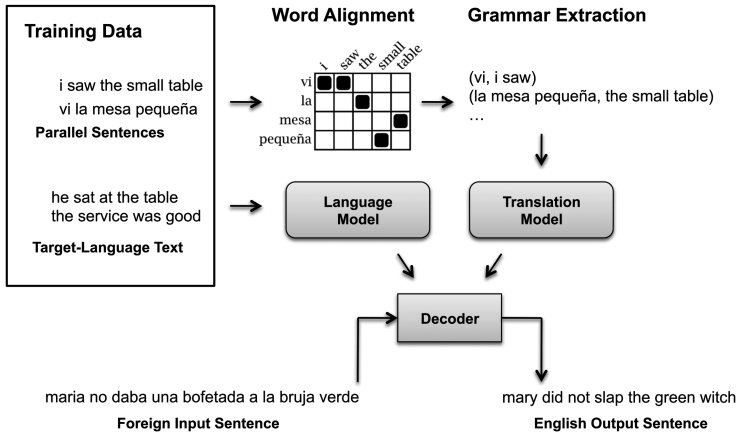

Fig. 1. Overview of a modern statistical machine translation system.

respectively. Typically, they are combined into a log-linear model:

$$\log P(t|s) = \log[P_{\text{TM}}(s|t) \times P_{\text{LM}}(t)] \tag{2}$$

$$= \log P_{\text{TM}}(s|t) + \log P_{\text{LM}}(t). \tag{3}$$

In this work, we leave aside research issues in language modeling and simply adopted best practices. Language modeling in the retrieval context is less interesting for two reasons: (i) we are translating queries, which are linguistically-impoverished compared to typical natural language sentences, and (ii) proper language modeling of queries requires access to query logs, which are difficult for academic researchers to obtain.

Training a translation model refers to the process of learning the parameters of P_{TM} in Equation (2). In most modern statistical MT systems, this involves *word alignment* and *grammar extraction*. Once parameters of the translation and language models are learned, one can apply a *decoder* to search for the best target-language text. Figure 1 illustrates such an architecture. A brief overview of each component is provided next, but for more details, we refer the reader to a survey by Lopez [2008] or a textbook by Koehn [2010].

2.1. Word Alignment

Most statistical models of translation descend from the IBM Models [Brown et al. 1990, 1993], which encode translation processes at the word level. These models assume a mapping between source and target words (i.e., a word alignment) that explains how words are transformed from the target language into the source language.

The word alignment process can be described probabilistically in terms of possible mappings between source and target words:

$$P_{\text{TM}}(s|t) = \sum_{\text{alignment } a} P(s, a|t) = \sum_{\text{alignment } a} P(a|t)P(s|t, a). \tag{4}$$

Each of the IBM Models 1–4 describes such a statistical model, differing only in the independence assumptions each makes. For instance, IBM Model 1 naïvely assumes that (i) each possible alignment is equally probable for a given target sentence, and (ii) each source word is determined only by the target word it is aligned to. Given a target sentence t, the generative story is as follows.

(1) Pick the number of source words to generate m with constant probability.
(2) Pick some alignment between $s = s_1 \dots s_m$ and $t = t_1 \dots t_l$ from a uniform distribution. Each alignment has equal probability: $\frac{1}{(l+1)^m}$.
(3) Generate each source word s_j with probability $P(s_i|t_{a(i)})$.

Thus, under IBM Model 1, the alignment probability is the following:

$$P_{\text{Model1}}(s, a|t) = C\frac{1}{(l+1)^m}\prod_{i=1}^{m} P(s_i|t_{a(i)}). \tag{5}$$

The parameters of IBM models can be learned in an unsupervised fashion by applying the expectation-maximization algorithm (EM) [Dempster et al. 1977] on a sentence-aligned bilingual corpus (also referred to as a *parallel corpus* or *bitext* in MT parlance).

While illustrative, IBM Model 1 performs poorly in practice and is used today only as an initialization step to more sophisticated models, for example, based on hidden Markov models (HMMs). Nevertheless, all word alignment models are similar in that they induce word translation probabilities in a completely unsupervised manner from bilingual text. These translation probabilities can be used as a bilingual dictionary for query translation in CLIR—this word-based approach provides a baseline, which we describe in Section 4.

2.2. Grammar Extraction

From word alignments, we can obtain word translation probabilities. However, using these distributions directly for machine translation yields poor-quality output, so modern MT systems use word alignments to build richer translation models. One successful approach has been to model "phrases," which are simply contiguous sequences of words (and do not necessarily have any linguistic basis). In this work, we focus on two popular phrase-based MT models, a *flat* phrase-based MT (PBMT) model [Koehn et al. 2003; Och and Ney 2004; Marcu and Wong 2002] and a *hierarchical* phrased-based model [Chiang 2005, 2007]. We selected these two methods for a few reasons: (i) they represent mature and well-understood techniques that form the foundation of the current state of the art (and in their basic forms still provide competitive translation quality); (ii) both are implemented in open-source tools that facilitate rapid experimentation and reproduction of results; (iii) phrases represent textual units that are easy to work with from a retrieval perspective.

In both flat and hierarchical phrase-based MT, the translation model is represented as a *translation grammar*, which can be extracted by inducing all bilingual phrase pairs that are consistent with word alignments [Och et al. 1999]. In a phrase-based MT system, the grammar consists of rules of the following form:

$$\text{rule r}: \ \alpha \ || \ \beta \ || \ \mathcal{A} \ || \ \ell(r).$$

Rule r states that source-language text α can be translated into target-language text β, with an associated likelihood value $\ell(r)$, an unnormalized estimate of the event probability.[1] We call α the left-hand side (LHS) of the rule, and β the right-hand side (RHS) of the rule. $\mathcal{A}$ represents the word alignments, which is a many-to-many mapping between words on the LHS and RHS of the rule.

In flat phrase-based MT systems, the LHS and RHS of a rule contain only text, or multiword expressions known as *phrases*. Thus, such rules are also referred to as *phrase pairs*, and the set of rules in the translation model form the *phrase translation table*.

[1] In practice, there are usually additional features.

R_1: [S] || [S,1] || [S,1]
R_2: [S] || [X,1] || [X,1]
R_3: [X] || [X,1] leave in europe || congé de [X,1] en europe || 1-0 2-3 3-4 || 1.0
R_4: [X] || maternal || maternité || 0-0 || 0.69

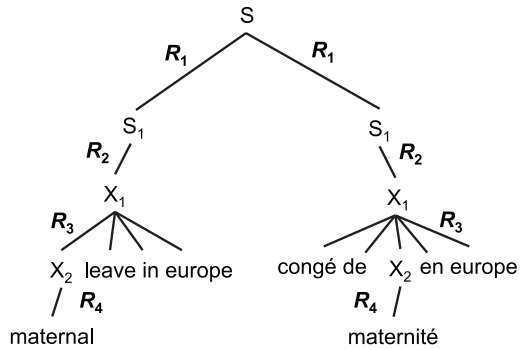

Fig. 2. A toy hierarchical grammar with four rules (top) and an illustration of how *maternal leave in Europe* is translated using these rules (bottom).

To avoid confusion and to emphasize the contrast with the hierarchical phrase-based approach (see the following), we refer to such a translation grammar as a *flat grammar*.

In hierarchical phrase-based MT systems, rules take a slightly different form:

$$\text{hierarchical rule } r: \ \texttt{[X]} \,||\, \alpha \,||\, \beta \,||\, \mathcal{A} \,||\, \ell(r).$$

This indicates that the context-free expansion $X \rightarrow \alpha$ in the source language occurs synchronously with $X \rightarrow \beta$ in the target language. These rules form a formal model of translation based on synchronous context-free grammars (SCFGs). This *hierarchical* grammar differs from a flat grammar in terms of rule expressivity: the LHS and RHS are allowed to contain one or more non-terminals, each acting as a variable that can be expanded into other expressions using other rules. In other words, the rules describe a context-free expansion on both source and target sides, carried out recursively to generate translation hypotheses.

Consider four rules from a toy hierarchical grammar in Figure 2. The first two rules are special rules without any lexical items, stating that there is one sentential form S 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 R_3, *leave* is aligned to *congé*, *in* is aligned to *en*, and *europe* is aligned to *europe*. In this case, the likelihood of the rule is 1.0; in other words, whenever the LHS is observed in English, we are very confident that the corresponding RHS should be generated in French. The bottom of Figure 2 shows the derivation tree that synchronously parses and translates the source text into *congé de maternité en europe* using the grammar. The left and right trees correspond to the parse of the source text and its translation, respectively. Each line between symbols in the tree indicates a rule application, annotated with the corresponding rule ID.

Having variables in the rules provides a representational advantage for hierarchical grammars in terms of conciseness and expressivity. For example, this grammar can accommodate the translation of *paternal leave in Europe* by adding the rule R_5 for the lexical translation.

$$R_5: \ \texttt{[X]} \,||\, \texttt{paternal} \,||\, \texttt{paternité} \,||\, \texttt{0-0} \,||\, 0.72$$

We can apply R_5 to produce *congé de paternité en europe* as the translation. This derivation is exactly the same as the derivation in Figure 2, except that *maternal* is

replaced with *paternal* on the English side (and corresponding changes on the French side). In this case, rules R_4 and R_5 provide a way to lexicalize the transformation represented in R_3. In order to obtain a corresponding flat grammar, we would need to include a separate rule to specify each lexicalized form of every transformation, resulting in a much larger translation model.

Another benefit of hierarchical grammars is higher expressivity through "gaps," encoded in non-terminal variables such as `[X]` in R_3. This allows an arbitrarily long part of the sentence to be "moved" from the left of the sentence in English to the middle of the sentence in French. Using such rules, a hierarchical grammar can capture distant dependencies in a sentence that cannot be easily expressed in flat grammars.

Although the complete translation grammars for real-world MT systems can be quite large, for any given input sentence, only a modest number of rules are applicable—modern systems take advantage of this property to increase decoding speed (see next section) and reduce memory footprint. For flat grammars, a binarized representation can be used to filter rules with respect to an input sentence [Koehn et al. 2007]. For hierarchical grammars, a suffix array built over the parallel text can be used to efficiently extract all applicable rules given an input sentence [Lopez 2007]. For CLIR, we can take advantage of these techniques to learn translation probabilities at query time, using what we call the grammar-based approach—this is detailed in Section 5.1.

2.3. Decoding

The third major component of the MT pipeline is the decoder, which performs a search through the hypothesis space to find top-scoring translations. By combining the translation and language models, the decoder attempts to ensure (i) that the hypothesis represents an accurate translation and (ii) that the hypothesis represents fluent target-language text.

In statistical MT, each sequence of rules that covers the input is called a *derivation* D and produces a translation candidate t, typically scored by a log-linear combination of features. One can include arbitrarily many features, but two are essential: the translation model score $\mathrm{TM}(t, D|s)$, which is the product of rule likelihood values and indicates how well the candidate preserves the original meaning, and the language model score $\mathrm{LM}(t)$, which captures the fluency of the translation. To control computational complexity, most decoders search for the most probable derivation (as opposed to the most probable string):

$$t^{(1)} = \arg\max_t \left[\max_{D \in \mathcal{D}(s,t)} \ell(t, D|s) \right] \tag{6}$$

$$= \arg\max_t \left[\max_{D \in \mathcal{D}(s,t)} \log \ell(t, D|s) \right] \tag{7}$$

$$= \arg\max_t \left[\max_{D \in \mathcal{D}(s,t)} (\log \mathrm{TM}(t, D|s) + \log \mathrm{LM}(t)) \right] \tag{8}$$

$$= \arg\max_t \left[\max_{D \in \mathcal{D}(s,t)} \sum_{r \in D} \log \ell(r) + \log \mathrm{LM}(t) \right], \tag{9}$$

where $\mathcal{D}(s, t)$ is the set of possible derivations that generate the pair (s, t). The sequence of four rules that translate the example query in Figure 2 forms one such derivation.

For CLIR, we can take advantage of the n-best hypotheses generated during decoding—this is referred to as the decoder-based approach and will be detailed in Section 5.2.

3. RELATED WORK

3.1. Query Translation and Context Recovery

The earliest approaches to query translation for cross-language information retrieval used machine-readable bilingual dictionaries [Hull and Grefenstette 1996; Ballesteros and Croft 1996]. For words that have multiple translations, these approaches either used all entries or restricted the translation to the first entry (assuming entries are sorted by decreasing frequency). The first approach suffers from noisy translation candidates and over-representation of highly polysemous query terms, whereas the second leaves out many good alternative translations. Generally speaking, these techniques achieve 40–60% of monolingual IR effectiveness. Xu and Weischedel [2005] showed that this can be increased to 70–80% simply by weighting each translation by $1/n$, where n is the entry order in the dictionary.

Pirkola [1998] was the first to separate the estimation of term frequency (tf) and document frequency (df) for query terms in CLIR, introducing the notion of "structured queries." In his work, the tf and df of each query term w is computed from its translation alternatives. The idea is to consider each translation as a synonym, so that the tf of w is the sum of the tf of its translations and the df is the cardinality of the union of documents its translations occur in. In a slight variant introduced by Kwok [1999], df(w) is computed as the maximum of the df values of its translations, mainly to simplify implementation. Both of these approaches have the same weakness: since all translation alternatives are treated equally, any rare translation that is a common word will dominate the df and disproportionately influence the overall scoring behavior. The model of Darwish and Oard [2003], which we detail in Section 4, overcomes this weakness.

There is a rich literature on post-processing translation alternatives to discard infelicitous candidates based on different notions of context. Many proposed methods are based on cohesion between the translated terms: alternatives include pointwise mutual information [Gao et al. 2001], dice similarity [Adriani and Rijsbergen 2000], and mutual information [Liu et al. 2005]. These approaches select terms greedily (i.e., pick the translation of the first word that maximizes cohesion, then move to the second word, etc.), but Seo et al. [2005] showed further improvements when all possibilities are considered.

Explicitly modeling term dependencies has been shown to increase effectiveness in monolingual retrieval [Gao et al. 2004; Metzler and Croft 2005], but extending these ideas to the cross-language case has proven not to be as straightforward as one might expect. Cao et al. [2006] attempted this by modeling the translation space as a graph structure and showed that a random walk algorithm can be applied to compute probabilities of different paths. Their approach considered synonyms for query expansion as well as translations. Wu and He [2010] also integrated expansion and translation using an MT-based translation model.

Another way to incorporate dependencies between query terms is to consider multiword expressions (i.e., "phrases") in order to limit polysemy effects [Adriani and Rijsbergen; Arampatzis et al. 1998; Ballesteros and Croft 1997; Chen 2000; Meng et al. 2004]. The common approach is to identify possible phrases in the query and search for translations in dictionaries with multiword expressions. Such dictionaries are not easy to build from scratch, so researchers have relied on existing resources. One challenge with this approach is low coverage: for example, Chen [2000] reported that only 45 of 367 identified phrases were in their dictionary. Building a domain-specific dictionary might address this issue, but even for matching phrases, authors reported that unreliable translations hurt overall effectiveness, mainly due to the lack of contextual clues to disambiguate among many translation alternatives. We argue that phrase-based MT systems possess strengths that are missing from these earlier approaches.

3.2. Noisy Channel Models

Application of the noisy channel model to information retrieval has a long history. For monolingual retrieval, an initial formulation can be traced back to Ponte and Croft's [1998] language modeling approach. A subsequent extension by Berger and Lafferty [1999] explicitly views retrieval as "statistical translation": it is imagined that a noisy channel "corrupted" documents into the query, and thus the retrieval task is to recover the (relevant) documents that "generated" the query.

Applying these ideas to the cross-language case was a natural extension: Federico and Bertoldi [2002] presented an approach based on hidden Markov models (HMMs), which models sequential dependencies using a bigram language model for the transition probabilities. Kraaij et al. [2003] presented two probabilistic models of how text gets "corrupted" in the retrieval process, originating from the document in one case and the query in the other. Each model has components that correspond to translation and language modeling. Both approaches exhibit improvements over earlier dictionary-based methods, reporting mean average precision of around 90% of monolingual comparison conditions. Others have presented similar approaches with different smoothing techniques, also yielding strong empirical results [Xu et al. 2001; Xu and Weischedel 2005]. In general, the translation models used in these techniques are less sophisticated than the phrase-based and hierarchical models in our work. Finally, another related thread of work explores relevance models in the language modeling framework, which was applied to both monolingual [Lavrenko and Croft 2001] and cross-language [Lavrenko et al. 2002] retrieval.

3.3. Document Translation and Language-Independent Approaches

Document translation, which involves translating the collection to be searched into the query language, also has a long history. One of the first attempts was by Oard and Hackett [1997], followed by the introduction of a technique for faster document translation [McCarley and Roukos 1998]. Researchers have also compared query and document translation, concluding that while the differences between the approaches were negligible, combining both into a hybrid approach yielded significant improvements [Oard 1998; McCarley 1999]. These results, however, depended on the specifics of the translation technique, the language pair, as well as the quality of the underlying translation system. With different setups, later papers contradicted earlier conclusions, instead claiming that query translation outperforms document translation [Hayurani et al. 2007], and that there is no benefit from translating in both directions [Kishida and Kando 2006].

In addition to query translation and document translation, there have also been attempts to map both into a language-independent semantic space, modeled as latent variables [Littman et al. 1998]. The major drawback of these approaches is the computationally-intensive nature of the underlying transformations (e.g., latent semantic indexing [Furnas et al. 1988]). The other downside is that the transformations are often difficult to understand and interpret. One attempt to address both issues is the work of Wang and Oard [2006], which leverages bidirectional translations to match the meaning of queries and documents, where "meanings" are represented by synsets in WordNet.

3.4. Integration of Machine Translation and Information Retrieval

With the development of sophisticated statistical translation models, modern MT systems construct rich internal representations that typically include translation alternatives encoded in lattices. However, MT-based CLIR approaches often still use one-best results, since it is more convenient to treat MT systems as black boxes. Nevertheless,

Magdy and Jones [2011] showed that it is beneficial to adapt MT specially for query translation. The authors reported improvements in retrieval effectiveness by simply applying standard IR text preprocessing prior to MT training. More recently, Nikoulina et al. [2012] built an MT model tailored to query translation by (i) tuning model weights on a set of queries and reference translations, and (ii) reranking the top-n translations to maximize effectiveness on a held-out query set. While improvements were more substantial using the second method, an interesting finding was the low correlation between translation and retrieval quality. Ma et al. [2012] also exploited n-best output from an MT system, using word alignments to produce a separate query from each translation. These translations were combined post-retrieval without using the MT probabilities, which is complementary to the pre-retrieval evidence combination techniques we propose. The authors reported improvements over one-best MT, which is consistent with our argument that diversity is beneficial in query translation. In speech retrieval, combining n-best derivations is also routinely used [Olsson and Oard 2009].

For a summary of trends in CLIR research, we refer the reader to a book by Nie [2010], in which he points out the need for better integration of MT and IR, with some ideas of how this can be accomplished. Our work can be seen as a realization of this integration.

4. CONTEXT-INDEPENDENT QUERY TRANSLATION

As a baseline, we consider a CLIR approach based on translation probabilities derived automatically from word alignments (see Section 2.1), which can be exploited for retrieval using the technique of Darwish and Oard [2003] for "projecting" vector representations of text from one language into another. Using this technique, we represent a source-language query $s = s_1, s_2, \ldots$ in the target language (i.e., the document language) as a probabilistic structured query (PSQ), where each word s_j is represented by its translations in the target language, weighted by the translation probability $P(t_i|s_j)$.

In order to build a translation probability distribution suitable for CLIR, we need to perform some "cleaning" on the raw output of the word aligner. For each source language term s_j, we sort its translations by decreasing probability into a list $[t_{i_1}, t_{i_2}, \ldots]$. In sorted order, these terms are added to a new probability distribution, called P_{word}, until (i) the probability falls below a threshold L, or (ii) the cumulative sum of probabilities reaches C, or (iii) the number of translations in the distribution exceeds H. Finally, to generate a valid probability distribution, we renormalize the probabilities of the selected terms:

$$P_{\text{word}}(t_{i_k}|s_j) = \begin{cases} 0 \text{ if filter}(k, [t_{i_1}, \ldots, t_{i_k}]) \\ \frac{1}{\xi_j} P(t_{i_k}|s_j) \text{ otherwise,} \end{cases} \tag{10}$$

$$\text{filter}(k, [t_{i_1}, \ldots, t_{i_k}]) = (k > H) \vee (P(t_{i_k}|s_j) \leq L) \vee \left(\sum_{l=1}^{k-1} P(t_{i_l}|s_j) \right) > C, \tag{11}$$

where ξ_j is the normalization factor, given by the sum of all the probabilities for each of the alternative translations (from the target-language vocabulary V_t) added to the distribution:

$$\xi_j = \sum_{\substack{t_{i_k} \in V_t \\ \neg\text{filter}(k, [t_{i_1}, \ldots, t_{i_k}])}} P(t_{i_k}|s_j). \tag{12}$$

Most retrieval models require term statistics such as term frequency (tf) and document frequency (df) to compute query-document scores. In CLIR, these statistics are available for target-language terms only. To address this problem, Darwish and Oard introduced a mechanism to translate these term statistics into the source-language

```
#comb(#wsyn(0.74 matern, 0.26 maternel)
      #wsyn(0.49 laiss, 0.17 quitt, 0.09 cong,
            0.08 part, 0.04 abandon, 0.04 voyag, ...)
      #wsyn(0.91 europ, 0.09 européen))
```

Fig. 3. The translation of the query *maternal leave in Europe* using P_{word}.

vocabulary space using term translation probabilities P_{word}. In this approach, the score of document d, given source query s, is computed as follows:

$$\text{Score}(d|s) = \sum_{s_j \in s} \text{BM25}(\text{tf}(s_j, d), \text{df}(s_j)), \tag{13}$$

$$\text{tf}(s_j, d) = \sum_{t_i} \text{tf}(t_i, d) P_{\text{word}}(t_i|s_j), \tag{14}$$

$$\text{df}(s_j) = \sum_{t_i} \text{df}(t_i) P_{\text{word}}(t_i|s_j). \tag{15}$$

As shown, we use the Okapi BM25 term weighting function [Robertson et al. 1994] due to its effectiveness in a wide range of empirical evaluations, but any other weighting function can be substituted into Equation (13).

Example. Using an Indri-like notation [Metzler and Croft 2004], Figure 3 shows how the translation of the English query *maternal leave in Europe* is represented as a PSQ for target language French. We assume English and French words have been stemmed and stopwords removed. The `#comb` operator corresponds to the sum operation in Equation (13), and the `#wsyn` operator represents the weighted sum in Equations (14) and (15). Each of the three `#wsyn` clauses represents the translation of a query term: *maternal*, *leave*, and *Europe*. Within each `#wsyn` clause, translation alternatives are weighted according to the P_{word} distribution. Since the translation distribution for the source term *leave* is unaware of the context *maternal leave*, candidates that occur most frequently in the training text (from which the alignments were induced), such as *laisser* (Eng. let go, allow) and *quitter* (Eng. quit), are assigned higher weights than more appropriate candidates, such as *congé* (Eng. vacation, day off). Due to the many senses of *leave*, there are ten translation candidates, some of which are omitted for brevity.

Discussion. One drawback of this query translation approach (hereafter referred to as "word-based") is its *context-independent* nature, since the model assumes that the translation of query terms is independent. In contrast, the *context-sensitive* models in the next section consider context when producing translations. This information allows a model to disambiguate translation alternatives and discard infelicitous candidates.

On the other hand, it may not be possible to disambiguate translation alternatives in some cases due to sense ambiguity in the original query that cannot be resolved given the available context. Furthermore, since information retrieval is fundamentally about bridging the mismatch between query terms that an information seeker uses and document terms that a writer selects (even if they are in different languages), multiple translations are desirable to increase conceptual coverage. Hence, the ability to represent multiple plausible translations for a single meaning in a probabilistic manner can be beneficial—in contrast to sense ambiguity, we might refer to this as retaining diversity via "representational ambiguity." We argue that translation models should be *ambiguity-preserving* in both of the ways previously discussed. However, we should take advantage of contextual information to prune or down-weight translations that are less plausible—which means that the amount of disambiguation must be

carefully balanced. Experiments show that the context-sensitive methods discussed in the next section are able to achieve this balance and are more effective than the word-based approach.

5. CONTEXT-SENSITIVE QUERY TRANSLATION

In this section, we present two ways to exploit the internal representations of a modern statistical MT system for query translation in CLIR. The grammar-based approach extracts translation probabilities from either flat or hierarchical grammars. The decoder-based approach extracts translation probabilities from n-best translation hypotheses from the decoder. The grammar-based approach exploits the translation model, while the decoder-based approach exploits both the translation and language models. Finally, different sources of evidence can be combined into an interpolated model. To our knowledge, these techniques are novel and represent our contribution to query translation approaches to CLIR.

5.1. The Grammar-Based Approach: Learning Probabilities from the Translation Model

Given an input sentence to translate, a modern statistical MT system extracts a subset of the grammar that is applicable to the input. Each translation rule describes one possible way of translating a portion of the query (along with associated features and likelihoods). In other words, the grammar encapsulates all the possible ways that the translation model can be applied to the input text. Since the rules contain multiword expressions, they capture contextual cues that can be exploited for query translation.

We propose the following method to exploit either a flat or hierarchical grammar for query translation: Given a grammar $\mathcal{G}$ and a query s, we first obtain the subset of rules $\mathcal{G}(s)$ for which the source side pattern matches s by using either the suffix array extraction or filtering techniques described in Section 2.2. Once $\mathcal{G}(s)$ is obtained, for each rule r in $\mathcal{G}(s)$, we identify each source word s_j on the LHS, ignoring any nonterminals. From the word alignment information included in the rule, we can find all target words that s_j is aligned to. By processing all the rules to accumulate likelihoods, we can construct translation probabilities for each source word (more details to follow). This procedure can be applied to both flat and hierarchical grammars.

When s_j is aligned to multiple target words in a rule, it is not obvious how to distribute the probability mass. One obvious approach is to treat each alignment as an independent event with the same probability (equal to the likelihood of rule r): we call this the *one-to-one* heuristic. This heuristic ignores the fact that target words aligned to s are not usually independent. To illustrate, consider the examples in Figure 4, which shows three grammar rules on the left and, for each, three different heuristics for learning translation pairs. In the first example, the English word *after* is aligned to two French words *après*, *avoir* (Eng. *after, have*), forming the phrase *après avoir*, which is a valid translation. However, according to the one-to-one heuristic, *avoir* would be incorrectly considered a valid translation of *after*. Similarly, in the second example, the English word *brand* is aligned to three French words *marque*, *de*, *fabrique* (Eng. *brand, of, factory*), forming the phrase *marque de fabrique* (Eng. *trademark*). Even if *de* is discarded as a stopword, the one-to-one heuristic would extract the translation pair (*brand*, *fabrique*) incorrectly. Note that in the third example, we do not learn the pair (*anti*, *aux*) because the latter is a stopword in French.

An alternative is to ignore these cases altogether and assume that good translation pairs will appear in other rules, so that discarding multiword alignments will not hurt: we call this the *one-to-none* heuristic. In the third example, the association between *drug* and *médic* is still learned, since *drug* is aligned to a single target word. All other possible associations are ignored with the one-to-none heuristic.

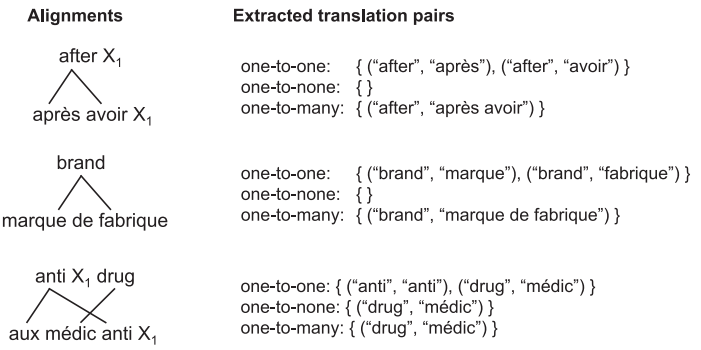

Fig. 4. Examples illustrating the three different heuristics for handling multiple word alignments when extracting translation pairs from grammar rules (one-to-one, one-to-none, and one-to-many). Hierarchical grammar rules are shown here but the heuristics apply in the same way to flat grammars.

A third approach is to combine the target words into a multiword expression. In the first two examples, we learn the translation of *after* as *après avoir* and *brand* as *marque de fabrique*. In contrast to the other two heuristics, this creates multiword translations, which are realized as phrase queries at retrieval time.[2] We combine the target words if they are consecutive or if they are separated by one or more stopwords. The latter case applies in the second example: *marque* and *fabrique* are not consecutive, but the only intervening word *de* is a stopword in French. In the third example, we do not learn an association between *anti* and *aux médic anti*, because the non-aligned middle word *médic* is not a stopword. We call this approach *one-to-many*, and compare the three heuristics experimentally.

After processing all rules in the manner just described, (using one of the three alignment heuristics), we can construct a distribution for each query term by renormalizing the likelihood scores. We call this distribution P_{PBMT} if the underlying rules are from a flat grammar or P_{SCFG} if the underlying rules are from a hierarchical grammar. Formally, this is described as follows.

$$P_{\text{SCFG/PBMT}}(t_i|s_j) = \frac{1}{\psi} \sum_{\substack{r \in \mathcal{G}(s) \\ s_j \leftrightarrow t_i \text{ in } r}} \ell(r) \tag{16}$$

$$\text{tf}(s_j, d) = \sum_{\{t_i | s_j \leftrightarrow t_i \in \mathcal{G}(s)\}} \text{tf}(t_i, D) P_{\text{SCFG/PBMT}}(t_i|s_j) \tag{17}$$

$$\text{df}(s_j) = \sum_{\{t_i | s_j \leftrightarrow t_i \in \mathcal{G}(s)\}} \text{df}(t_i) P_{\text{SCFG/PBMT}}(t_i|s_j), \tag{18}$$

where ψ is the normalization factor, and $s_j \leftrightarrow t_i$ represents an alignment between words s_j and t_i. Mapping tf and df statistics from source to target vocabulary is achieved by replacing P_{word} with $P_{\text{SCFG/PBMT}}$ in Equations (14) and (15).

Example. Let us compute P_{SCFG} for the second term in our example query, *maternal leave in Europe* (which is preprocessed into *matern leav europ*). In Figure 5, we show the (abbreviated) set of hierarchical rules that contain the word *leav* extracted from the translation model and also how the distributions are computed with each alignment heuristic.

[2]At query time, stopwords in phrase queries are ignored since we remove stopwords prior to indexing.

```
[X] || leav || cong || 0-0 || 0.38
[X] || leav || quitt || 0-0 || 0.08
[X] || leav || laiss || 0-0 || 0.27
[X] || [X] leav || [X] laiss || 1-1 || 0.22
[X] || [X] leav || cong [X] || 1-0 || 0.07
[X] || [X] leav || [X] en laiss || 1-2 || 0.02
[X] || [X] leav || [X] elle quitt || 1-2 || 0.01
[X] || [X] leav || [X] en train de quitt || 1-4 || 0.01
[X] || leav || quitt cet assembl || 0-0 0-2 || 0.01
```

One-to-One Alignment Heuristic:

$$\psi = (0.38 + 0.07) + (0.08 + 0.01 + 0.01 + 0.01) + (0.27 + 0.02) + 0.01 = 0.86$$
$$P_{\text{SCFG}}(\texttt{cong}|\texttt{leav}) = (0.38 + 0.07)/0.86 \sim 0.52$$
$$P_{\text{SCFG}}(\texttt{quitt}|\texttt{leav}) = (0.08 + 0.01 + 0.01 + 0.01)/0.86 \sim 0.13$$
$$P_{\text{SCFG}}(\texttt{laiss}|\texttt{leav}) = (0.27 + 0.02)/0.86 \sim 0.34$$
$$P_{\text{SCFG}}(\texttt{assembl}|\texttt{leav}) = 0.01/0.86 \sim 0.01$$

One-to-Many Alignment Heuristic:

$$\psi = (0.38 + 0.07) + (0.08 + 0.01 + 0.01) + (0.27 + 0.02) + 0.01 = 0.85$$
$$P_{\text{SCFG}}(\texttt{cong}|\texttt{leav}) = (0.38 + 0.07)/0.85 \sim 0.53$$
$$P_{\text{SCFG}}(\texttt{quitt}|\texttt{leav}) = (0.08 + 0.01 + 0.01)/0.85 \sim 0.12$$
$$P_{\text{SCFG}}(\texttt{laiss}|\texttt{leav}) = (0.27 + 0.02)/0.85 \sim 0.34$$
$$P_{\text{SCFG}}(\texttt{quitt cet assembl}|\texttt{leav}) = 0.01/0.85 \sim 0.01$$

One-to-None Alignment Heuristic:

$$\psi = (0.38 + 0.07) + (0.08 + 0.01 + 0.01) + (0.27 + 0.02) = 0.84$$
$$P_{\text{SCFG}}(\texttt{cong}|\texttt{leav}) = (0.38 + 0.07)/0.84 \sim 0.54$$
$$P_{\text{SCFG}}(\texttt{quitt}|\texttt{leav}) = (0.08 + 0.01 + 0.01)/0.84 \sim 0.12$$
$$P_{\text{SCFG}}(\texttt{laiss}|\texttt{leav}) = (0.27 + 0.02)/0.84 \sim 0.35$$

Fig. 5. Sample hierarchical rules that contain the term *leav* and the computation of P_{SCFG} using different alignment heuristics.

In order to construct a translation distribution for the term *leav*, we iterate over rules and accumulate likelihood values for each translation alternative. The candidate in the first rule is *cong* (Eng. vacation, day off), thus we accumulate a value of 0.38 for the distribution $P_{\text{SCFG}}(cong|leav)$. Similarly, we process the remaining rules and add values for two other translation candidates: *laiss* (Eng. let go, allow) and *quitt* (Eng. quit). In the final rule, the approach depends on the heuristic choice: If we apply the one-to-one strategy, we add 0.01 for each candidate, *quitt* and *assembl* (*cet* is a stopword in French). In this case, the computation of the final distribution is shown right below the grammar in Figure 5. If the one-to-many heuristic is applied, we add 0.01 to the multiword expression *quitt cet assembl*, yielding a different computation, shown in the middle of Figure 5. Finally, with the one-to-none heuristic, we ignore the last rule, producing the distribution shown in the bottom of Figure 5.

In Figure 6, we show the structured query that captures translation probabilities as learned from the complete translation grammar using a hierarchical MT system and the one-to-one alignment heuristic. The query looks similar using a flat grammar and therefore is not shown. Note that the probabilities here do not match the probabilities in Figure 5 since they were computed from the entire translation grammar (and not just a toy example).

```
#comb(#wsyn(0.68 matern, 0.06 maternel, ... )
      #wsyn(0.36 cong, 0.25 laiss, 0.11 quitt, ... )
      #wsyn(0.90 europ, 0.07 européen, ... ))
```

Fig. 6. The translation of the query *maternal leave in Europe* using P_{SCFG}.

Discussion. P_{word} and $P_{\text{SCFG/PBMT}}$ both capture the probability of a target-language word given a source-language word, but differ in how the probabilities are computed. For both approaches, we start from word alignments. P_{word} can be directly computed from the word alignments, but to compute $P_{\text{SCFG/PBMT}}$, we leverage the MT system's translation model (flat or hierarchical translation grammar) and filter the grammar based on the input query. Thus, $P_{\text{SCFG/PBMT}}$ takes context into account: instead of an unconditioned word probability distribution, we generate a distribution that is conditioned on multiword expressions that each source word appears in.

The context-sensitive nature of the grammar-based approach produces a structured query that is different from the one derived from P_{word}. The distribution of *leave* in Figure 6 shifts toward the more appropriate translation *congé* by incorporating context; this is apparent if we compare against the context-independent distribution in Figure 3. Moreover, the number of translation candidates decreases from ten to five with the grammar-based approach.

Note that once we perform grammar extraction, the rules are processed in exactly the same way; non-terminals in the hierarchical rules are ignored. This naturally begs the question: What's the difference between flat and hierarchical grammars? These two types of grammars represent different trade-offs in the design of MT systems. Flat grammars are simpler and less expressive, while hierarchical grammars are more expressive since they are able to model long-distance dependencies. In terms of translation quality on standard MT tasks, systems using hierarchical grammars are generally better, but details vary by language, genre, amount of training data, etc. Since application of hierarchical grammars in decoding requires synchronous parsing, they are usually slower [Lopez 2008]. On the other hand, because flat grammars are less expressive and more verbose, the translation models usually contain more rules; this verbosity often leads to more noisy translation alternatives.

The trade-offs between flat and hierarchical grammars are usually discussed in terms of translation tasks, not cross-language information retrieval, and thus their impact for retrieval applications remains an open question. For example, is the greater expressivity offered by hierarchical grammars useful for translating queries, given that most queries are short? Or, can simpler flat grammars work just as well in terms of retrieval effectiveness? Does the verbosity of flat grammars have any impact on efficiency? In an MT system using flat grammars, the decoder may ignore noisy translations since it is guided by a language model, but here we are using the rules directly, which results in complex structured queries. We explore these and related questions in our experiments.

5.2. The Decoder-Based Approach: Learning Probabilities from Multiple Derivations

In the grammar-based approach to query translation, we take advantage of context that is encapsulated in the translation grammar. A natural next step would be to exploit the language model as well and work with representations that are generated as part of the decoding process. We refer to this as the decoder-based approach.

The most obvious way to use the MT decoder for CLIR is to replace the source query with its most probable translation. In this one-best query translation approach, Equations (13)–(15) simplify to the following, where $t_i^{(1)}$ is the ith word of the best

translation of s:

$$\text{Score}(d|s) = \sum_{i=1}^{m} \text{BM25}\big(\text{tf}(t_i^{(1)}, d), \text{df}(t_i^{(1)})\big). \tag{19}$$

Since modern statistical MT systems generate high-quality translations for many language pairs, this one-best strategy works reasonably well for retrieval and provides a competitive baseline. However, in the context of our ongoing discussion, we argue that this approach "collapses" too much ambiguity, resulting in a brittle system. In cases where the translation is incorrect or even "slightly off" (i.e., an acceptable but awkward phrasing), effectiveness can suffer substantially. Furthermore, as we have discussed, since the same concept can be expressed using multiple terms, it would be desirable to retain representational diversity in our queries.

We can address this issue by considering the n-best hypotheses. Decoders produce many candidate translations in the process of computing Equation (6); most of the time, all but the best hypothesis are discarded. However, by "opening up" the MT system, we can take advantage of the n most probable candidates.

To learn term translation probabilities from the n-best translations, we start by preprocessing the source query s and each candidate translation $t^{(k)}, k = 1 \ldots n$. Here, $t^{(k)}$ denotes the kth most likely translation of s according to the log-linear MT model. For each source word s_j, we use the derivation information to determine which grammar rules were used to produce $t^{(k)}$ and the word alignments within the rules to determine which target terms 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 hypotheses. Specifically, if source word s_j is aligned to (i.e., translated as) t_i in the kth best translation, the value $\ell(t^{(k)}|s)$ is added to its probability mass. Similar to $P_{\text{SCFG/PBMT}}$, we apply one of the three heuristics when a source word is aligned to multiple target words in a rule. The following specifies how this new probability distribution (called P_{nbest}) is constructed:

$$P_{\text{nbest}}(t_i|s_j) = \frac{1}{\varphi} \sum_{\substack{k=1 \\ s_j \leftrightarrow t_i \text{ in } t^{(k)}}}^{n} \ell(t^{(k)}|s), \tag{20}$$

$$\text{tf}(s_j, d) = \sum_{t_i} \text{tf}(t_i, d) P_{\text{nbest}}(t_i|s_j), \tag{21}$$

$$\text{df}(s_j) = \sum_{t_i} \text{df}(t_i) P_{\text{nbest}}(t_i|s_j), \tag{22}$$

where φ is the normalization factor. If a source word is translated consistently into the same target word in all n hypotheses, it will have a single translation with a probability of 1.0.

Example. Construction of P_{nbest} is similar to $P_{\text{SCFG/PBMT}}$, but we consider only rules that participate in the derivations of the top n translations. Figure 7 shows the resulting structured query for our running example using P_{nbest}. In this example, *leave* is consistently translated to *congé* in all top n hypotheses, so it receives a weight of 1.0 in the `#wsyn` clause (and similarly for *europe*).

Discussion. The grammar-based and decoder-based approaches take advantage of different stages in the standard MT pipeline. Whereas $P_{\text{PBMT/SCFG}}$ exploits only the translation model to compute translation probabilities (either flat or hierarchical

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

Fig. 7. The translation of the query *maternal leave in Europe* using P_{nbest} ($n = 10$).

grammars), P_{nbest} additionally benefits from the language model, which guides the decoder to hypotheses that are more fluent in the target language. In the context of our ongoing discussion, the decoder-based approach applies more context to disambiguate translation alternatives, but correspondingly sacrifices diversity.

However, it remains an open question whether using a language model for query translation is beneficial. Most often, language models are trained on well-formed text, typically from sources such as newswire articles. Thus, the decoder is guided to hypotheses that "look like" newswire text, which is different in nature from queries. Of course, this issue could be addressed if we built language models from target language queries, but this requires query logs in foreign languages, which are difficult to obtain for academic researchers. Also, bringing language models to bear for query translation has efficiency implications in terms of the time needed for the decoder to explore the hypothesis space and a larger memory footprint for loading the language model. These issues are worth considering as part of an overall evaluation.

Another implication of learning translation probabilities after the decoding process is the need for MT system tuning: this is a nontrivial and time-intensive procedure. Typically, system tuning is accomplished using a validation dataset that is similar in nature to the types of input that the system will encounter. Since we did not have access to appropriate query logs (and their translations), tuning was performed using standard data (from the newswire domain). To our knowledge, other than a recent paper by Nikoulina et al. [2012], tuning MT systems specifically for CLIR has received little attention by researchers.

We conclude with a subtle point about MT decoders: to preserve translation diversity, we take advantage of the *n*-best hypotheses as opposed to the single best translation. However, the decoder scores most likely *derivations*, that is, the application of rules that spans the input to generate the output—not the most likely surface strings (i.e., text). In fact, many derivations share the same surface string. This lack of textual variety is a known issue, called "spurious ambiguity" in the MT literature, and it occurs in both flat and hierarchical phrase-based MT systems. For instance, according to Li et al. [2009], a string has an average of 115 distinct derivations in Chiang's hierarchical system [2007]. Researchers have proposed several ways to cope with this issue, and integrating some of these ideas into our CLIR approach might be worthwhile to explore in the future.

Our running example of *maternal leave in Europe* suffers from this spurious ambiguity phenomenon, since two of the three query terms are consistently translated in all top-10 MT output. To provide the reader with a better sense of the diversity that is captured in *n*-best output, we show the top-10 translations of the query *earthquakes in Mexico City* in Figure 8. There are two common ways to translate *earthquake* into French: *tremblement de terre* and *séisme*; these alternative are reflected in the translation output.

5.3. Combining Sources of Evidence

The three approaches that we have discussed for query translation (word-based, grammar-based, and decoder-based) represent different trade-offs in applying context for disambiguation and preserving translation diversity. It seems natural that we would want to combine multiple sources of evidence, and a simple way to accomplish

trembl de terr à mexico	0.44
trembl de terr et de la vill de mexico	0.14
trembl à mexico	0.12
trembl de terr à mexico	0.08
trembl de terr à mexico	0.07
séism de mexico	0.05
séism à mexico	0.05
trembl à mexico	0.03
séism de mexico	0.01
séism à mexico	0.01

Fig. 8. The 10 best translations (and their respective probabilities) for the query *earthquakes in Mexico City* by our English-French MT system.

Table I. Summary of the Data Used in Our Evaluations

Language	Collection		# topics	MT Training data	
	Source	Size (docs)		Source	Size (sent)
Arabic	TREC 2002	383,872	50	GALE	3.4m
Chinese	NTCIR-8	388,589	73	FBIS	0.3m
French	CLEF 2006	177,452	50	Europarl	2.2m

this would be via a linear interpolation of the three probability distributions:

$$\begin{aligned} P_c(t_i|s_j;\lambda_1,\lambda_2) \;=\; & \lambda_1 P_{\text{nbest}}(t_i|s_j) \\ & + \lambda_2 P_{\text{SCFG/PBMT}}(t_i|s_j) \\ & + (1-\lambda_1-\lambda_2) P_{\text{word}}(t_i|s_j), \end{aligned} \tag{23}$$

where λ_1 and λ_2 define how much weight is assigned to the decoder-based and grammar-based models, respectively. Replacing P_{word} with P_c in Equations (14) and (15) gives us the document scoring formula for the interpolated model.

6. EVALUATION SETUP

We evaluated the query translation approaches described in this article using 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. In all three cases, the source language is English and we are searching for documents in the foreign language. For the Arabic and French collections, we used title queries because they are most representative of the short queries that users pose to search engines. Queries in the NTCIR-8 ACLIA test collection are in the form of well-formed questions, but for consistency, we treated them as bag-of-words queries with no special processing. Statistics for each collection are summarized in Table I.

Bilingual training data used for translation modeling were as follows:

—*Arabic*, 3.4 million aligned sentence pairs from the DARPA GALE evaluation [Olive et al. 2011], which consists of NIST and LDC releases;
—*Chinese*, 302,996 aligned sentence pairs from the Foreign Broadcast Information Service (FBIS) corpus, which is a collection of radio newscasts, provided by LDC (catalog number LDC2003E14);[3]
—*French*, 2.2 million aligned sentence pairs from the Europarl corpus (version 7) built from the proceedings of the European Parliament.[4]

[3]http://projects.ldc.upenn.edu/TIDES/mt2003.html.
[4]http://www.statmt.org/europarl.

From these data, word alignments were learned with GIZA++ [Och and Ney 2003], using five Model 1 and five HMM iterations. Flat grammars were extracted using Moses [Koehn et al. 2007] and hierarchical grammars were extracted using `cdec` [Dyer et al. 2010], which implements the suffix array approach of Lopez [2007] for grammar extraction.

To complete the components required for translation, a 3-gram language model was trained from the non-English side of the training text 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 were tokenized using the OpenNLP tokenizer.[5] Arabic was tokenized and stemmed using Lucene.[6] For English and French, we also lowercased text, stemmed words using the Snowball stemmer, and removed stopwords.

We evaluated the following approaches to query translation, detailed in Sections 4 and 5:

—word-based, using P_{word}, as described by Equation (10);
—grammar-based, using P_{SCFG} or P_{PBMT}, as described by Equation (16);
—decoder-based, using P_{nbest}, as described by Equation (20);
—interpolated, using P_{c}, as described in Equation (23).

We fixed $C = 0.95$, $L = 0.005$, $H = 15$ for the word-based approach (based on previous work) and $n = 10$ for the decoder-based approach. For the grammar-based, decoder-based, and interpolated approaches, we compared the effectiveness of different heuristics for multiple word alignments (one-to-one, one-to-many, one-to-none), as described in Section 5.1. For the interpolated approach, in our initial experiments we performed a grid search on the weights λ_1 and λ_2 in increments of 0.1, ranging from 0 to 1, and report the setting with the highest effectiveness. Note that this represents the upper bound on model effectiveness, since the weights are optimized on the same set of topics used for testing. However, in a later set of experiments we report results with weights learned from cross-validation. We employed this experimental setup because of the small numbers of topics in our test collections.

All query translation approaches were implemented in our open-source Java retrieval toolkit called Ivory.[7] Code and data (grammars, vocabularies, etc.) necessary to replicate these experiments are available on the Web. As previously described, ranking is performed using the Okapi BM25 scoring function [Robertson et al. 1994], with parameters set to $k_1 = 1.2$, $b = 0.75$. In all our experiments, we retrieved up to 1,000 hits for each topic and used mean average precision (MAP) as the evaluation metric.

7. EXPERIMENTAL RESULTS

Our experimental results are organized as follows: First, we provide a basic comparison of different query translation approaches. Next, we examine in detail the impact of parameter settings, followed by a per-topic analysis. Finally, we explore the efficiency implications of various query translation techniques.

7.1. Comparison of Query Translation Approaches

The context-independent word-based approach of Darwish and Oard [2003] detailed in Section 4 provides a strong baseline for query translation in CLIR. This model achieves a MAP of 0.271 for Arabic, 0.150 for Chinese, and 0.262 for French. Direct comparisons

[5] `http://opennlp.apache.org`.
[6] `http://lucene.apache.org`.
[7] `http://ivory.cc`.

Table II. Summary of MAP Values under Different Conditions for All Three CLIR Tasks

Method	Heuristic	Arabic		Chinese		French	
		SCFG	PBMT	SCFG	PBMT	SCFG	PBMT
word	-	0.271		0.150		0.262	
grammar	1-to-M	0.293	0.274	0.182	0.156	0.297	0.264
	1-to-0	0.302	0.273	0.188	0.167	0.292	0.262
	1-to-1	0.282	0.266	0.170	0.151	0.288	0.257
1-best	1-to-M	0.242	0.246	0.156	0.150	0.276	0.297
	1-to-0	0.250	0.230	0.155	0.146	0.235	0.242
	1-to-1	0.249	0.249	0.155	0.155	0.276	0.297
10-best	1-to-M	0.255	0.264	0.159	0.169	0.307	0.289
	1-to-0	0.248	0.249	0.159	0.163	0.295	0.282
	1-to-1	0.249	0.254	0.159	0.163	0.304	0.300
interpolated	1-to-M	$0.293^{*\dagger}$	$0.280^{\dagger}$	$0.192^{*\dagger}$	$0.183^{*\dagger}$	$0.318^{*\dagger}$	0.307^{*}
	1-to-0	$0.302^{*\dagger}$	$0.276^{\dagger}$	$0.193^{*\dagger}$	$0.188^{*\dagger}$	$0.315^{*\dagger}$	$0.300^{\dagger}$
	1-to-1	$0.282^{\dagger}$	$0.274^{\dagger}$	$0.182^{*\dagger}$	$0.177^{*\dagger}$	$0.314^{*\dagger}$	0.301

Note: 1-to-M, 1-to-0, and 1-to-1 indicate alignment heuristics: one-to-many, one-to-none, and one-to-one. Superscripts $*$ and † indicate that the interpolated result is significantly better than the word-based and corresponding one-best approaches, respectively.

to results reported at TREC, NTCIR, and CLEF (respectively) are difficult because of differences in experimental conditions, but the comparisons we are able to make suggest that these scores are competitive. The best results at those evaluations took advantage of blind relevance feedback, multiple lexical resources, and long queries. While these techniques can be useful in deployed applications, we decided not to run experiments that include them to avoid masking the effects we wish to study. For Arabic, the best reported results from TREC 2002 were close to 0.400 MAP [Fraser et al. 2002], achieved by performing query expansion and learning stem-to-stem mappings. For Chinese, the NTCIR-8 topics are in the form of well-formed questions, and systems that applied question rewriting performed better than those that did not—this is not germane to the focus of our study. 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 a MAP of 0.178, close to the best reported results with comparable settings, with a MAP of 0.181 [Zhou and Wade 2010]. For French, our baseline achieves a score close to the single reported result at CLEF 2006 that did not incorporate blind relevance feedback (0.261 MAP) [Savoy and Abdou 2007].

Results comparing the various query translation approaches are presented in Table II. Blocks of rows represent the word-based approach, the grammar-based approach, the decoder-based approach (using either the one-best translation or the ten-best translations), and the interpolated model (based on a grid search of the interpolation parameters λ_1 and λ_2 in increments of 0.1 ranging from 0 to 1). Rows within each block indicate which alignment heuristic was used: "1-to-M" for one-to-many, "1-to-0" for one-to-none, and "1-to-1" for one-to-one. Each group of columns represents a test collection: Arabic, Chinese, and French, and each individual column shows results with either hierarchical grammars (SCFG) or flat grammars (PBMT). Note that for the one-best, the one-to-none heuristic has the effect of discarding query terms that are aligned to multiple target-language words.

For the grammar-based approach, we see consistent differences between flat and hierarchical grammars. According to a standard randomized significance test used for IR [Smucker et al. 2007], `cdec`-based hierarchical grammars significantly outperform

Moses-based flat grammars ($p < 0.05$) for all nine settings from {ar, zh, fr}×{one-to-many, one-to-one, one-to-none}. Furthermore, when we compare the best out of the three for each MT approach separately (i.e., we pick the best-performing heuristic for SCFG and compare it to the best heuristic for PBMT), the p-value is still under 0.1. This suggests that the SCFG-based translation model is better at capturing query translation alternatives for CLIR. Hierarchical grammars generally produce higher quality output than flat grammars in standard translation tasks and this appears to carry over to retrieval effectiveness as well.

One important implication of using different MT models is grammar size: As discussed in Section 5.1, flat grammars are much larger than hierarchical grammars because they are less expressive and require many lexicalized rules to enumerate transformations that can be succinctly captured in a few hierarchical rules. For the grammar-based approach, the processing time for P_{PBMT} is an order of magnitude higher than for P_{SCFG}. Note that the grammar-based approach does not use the decoder, and thus grammar size dominates query processing since we need to analyze all applicable rules to generate the translation probabilities. In contrast, with the decoder-based approach, using flat grammars is usually faster than using hierarchical grammars since the latter requires synchronous parsing.

Turning to the decoder-based approach, examination of the output shows that the final translated queries are quite similar for flat and hierarchical grammars. As a result, there are few effectiveness differences between the one-best and 10-best settings when comparing PBMT and SCFG. In contrast to the previous findings with the grammar-based approach, the decoder-based approach appears to be insensitive to the underlying MT model. We believe that the language model introduces an additional source of constraints that helps to suppress infelicitous translations, thus reducing the quality gap between flat and hierarchical grammars. Furthermore, since queries are relatively short, the language model can do more "heavy lifting" in modeling translation fluency.

When comparing the grammar-based approach and the decoder-based approach, the former seems to be more effective overall for Arabic and Chinese, but not French. We will next discuss the relative effectiveness of the interpolated approach. In considering the one-to-one, one-to-none, and one-to-many alignment heuristics, experimental results suggest that the one-to-many method is the most effective overall, with the best MAP score in 18 out of 24 cases. Four of the the six cases where one-to-many is not the best are from Chinese, most likely due to word segmentation issues. Based on these findings, we decided to use hierarchical grammars (`cdec`) and the one-to-many heuristic for our remaining experiments.

7.2. Impact of Interpolation Parameters

Next, we take a closer look at parameter settings in the interpolated model, which combines evidence from the word-based, grammar-based, and decoder-based approaches. In order to examine the effectiveness of the interpolated model P_{c} with respect to parameters λ_1 and λ_2, we performed a grid search in increments of 0.1 ranging from 0 to 1 (same as in the previous section). These experimental results are summarized in Table III. Note that this table contains results selected from Table II for comparison purposes, since we focus only on the one-to-many heuristic using hierarchical translation grammars. In Figure 9, for each collection, we provide a scatterplot of MAP scores for different values of λ_1 and λ_2. These plots also provide an alternative method for visualizing the different conditions presented in Table III. For readability, the plots only include a representative subset of λ_2 settings (represented by different lines).

The left edge of each plot represents $\lambda_1 = 0$, where we do not use P_{nbest}. Along the y-axis, we see results for different settings of λ_2, which controls the balance between

Table III. Summary of Experimental Results Using the One-to-Many Heuristic and Hierarchical Grammars for All Three CLIR Collections

	Condition	Parameters	MAP		
			Arabic	Chinese	French
A	word-based (P_{word})	λ_1=0, λ_2=0	0.271	0.150	0.262
B	grammar-based (P_{SCFG})	λ_1=0, λ_2=1	0.293	0.182	0.297
C	decoder-based (P_{nbest})	λ_1=1, λ_2=0	0.255	0.159	0.307
D	1-best	-	0.242	0.156	0.276
E	interpolated (P_{c})	best $\{\lambda_1, \lambda_2\}$	$0.293^{a,c,d}$	$0.192^{a,b,c,d}$	$0.318^{a,d}$
	interpolated (P_{c})	10-fold CV	$0.293^{a,c,d}$	$0.190^{a,b,c,d}$	0.311^{a}
	interpolated (P_{c})	transfer	0.288	0.188^{a}	0.287

Note: Superscripts indicate that the interpolated model (E) is significantly more effective than conditions A, B, C, and D as detailed in the text.

P_{SCFG} and P_{word}. Within these settings, $\lambda_2 = 0$ corresponds to using only P_{word}; let us call this condition A. When λ_2 is set to 1, we rely only on P_{SCFG}; call this condition B. At the right edge of each scatterplot, $\lambda_1 = 1$, we use only P_{nbest}; call this condition C. For reference, the dotted horizontal line represents the one-best MT translation; call this condition D.

Let us label the interpolation setting with the best MAP score condition E. For the Chinese collection, this occurs with $\lambda_1 = 0.1$ and $\lambda_2 = 0.8$ (0.192 MAP), which means most of the weight is placed on the grammar-based approach P_{SCFG}. With the French collection, the most effective setting is $\lambda_1 = 0.5$ and $\lambda_2 = 0.3$, resulting in a MAP score of 0.318. Interestingly, for the Arabic collection, the best result is obtained when $\lambda_1 = 0$ and $\lambda_2 = 1.0$, with 0.293 MAP. In other words, the highest effectiveness is obtained based entirely on P_{SCFG}, ignoring the distributions P_{word} and P_{nbest}. There is no single interpolation setting that simultaneously maximizes effectiveness for all three collections.

Based on randomized significance testing [Smucker et al. 2007], the interpolated approach (E) outperforms all other conditions with 95% confidence in the Arabic collection, except for the grammar-based approach (B), since they are identical. For French, we found that the interpolated approach (E) is significantly better than the word-based (A) and one-best (D) approaches. For Chinese, condition E is significantly better than all of the other conditions (A, B, C, and D). These results suggest that the individual models are complementary, and a combination of evidence yields higher effectiveness overall.

Leaving out the interpolated condition, we find that the grammar-based approach (B) is significantly better than the word-based approach (A) for Arabic and Chinese, but statistically indistinguishable for French. The 10-best decoder-based approach (C) is significantly better than the word-based approach (A) for French, but this does not hold for the other two collections. The lack of consistent findings across test collections may be due to a variety of reasons: overall quality of the translation model (in general, MT quality on European languages is higher), mismatch in language models, or characteristics of the information needs. In any case, results indicate that there is no single approach that outperforms the rest in all three collections.

The experiments thus far involved tuning interpolation parameters on the same topics that are used for testing. This is of course not reflective of a real-world setting, but because of the relatively small test collections available to us, we felt it was reasonable to first establish upper-bound effectiveness given oracle parameter settings. Separately, we ran tenfold cross-validation experiments on each collection, by selecting the parameters that maximizes MAP on nine folds and evaluating on the remaining one. This method yields a MAP of 0.293 for Arabic, 0.190 for Chinese, and 0.311 for

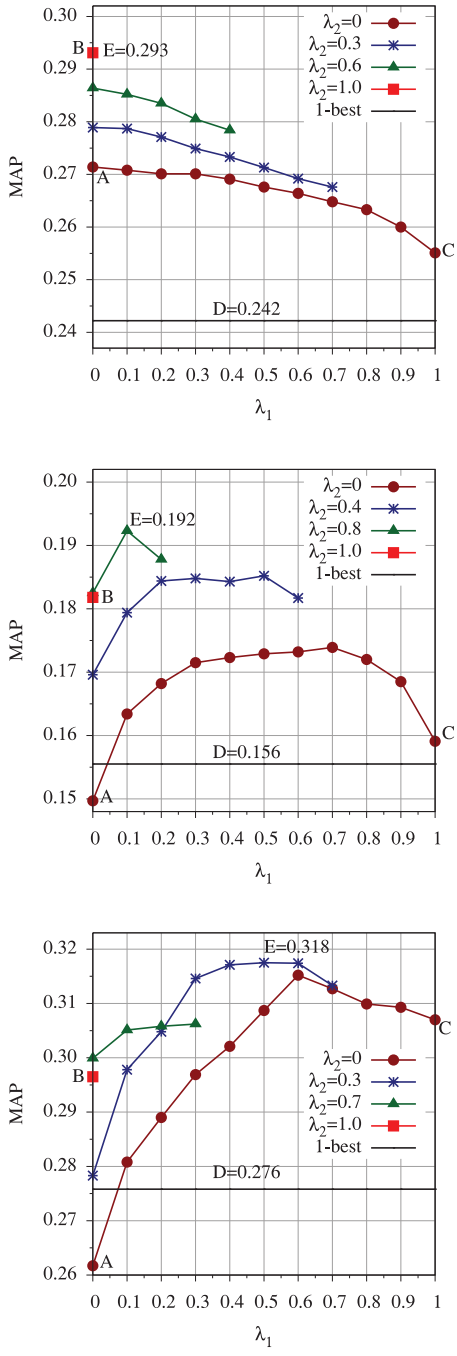

Fig. 9. Results of grid search on the interpolated model parameters for the TREC 2002 English-Arabic, NTCIR-8 English-Chinese, and CLEF 2006 English-French collections.

French, all significantly better than the word-based approach (A). These figures are reported in Table III under the condition "10-fold CV". In the case of Arabic, the cross-validation run is significantly better than the one-best (D) and 10-best decoder-based (C) approaches as well. For the Chinese collection, the cross-validation run significantly outperforms all other models. For the French collection, cross-validation results are significantly better than the word-based approach (A). These results show that with training data, we can achieve significant improvements in effectiveness using the interpolated model. However, it appears that the training topics must be similar to the test topics, since the parameters learned for each collection are different.

To further examine this issue, we also explored using two of the collections to tune parameters for the third collection—simulating the scenario where we have heterogeneous training data (a crude form of "transfer learning"). 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 from the other two collections and picked the one with the lowest sum. Using this method, the selected parameters were $(0.1, 0.8)$ for Arabic, $(0.3, 0.5)$ for Chinese, and $(0.1, 0.1)$ for French, yielding MAP scores of 0.288, 0.188, and 0.287, respectively. These results are reported in Table III under the condition "transfer." When compared to the word-based approach (A), these settings showed significant improvements only for Chinese. This analysis shows 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. It does not appear, however, that we can generalize the interpolation model in a robust, collection-independent manner.

7.3. Per-Topic Analysis

It is well known that comparing mean effectiveness across many topics hides topical variations, and thus for a detailed analysis, we examined the distribution of the average precision (AP) differences between the various approaches for each topic. The interpolated model achieved better AP than the word-based approach for 36 of 46 topics (78%) in the Arabic collection, ignoring 4 of the 50 topics which exhibited differences of 0.001 or less. For the Chinese collection, the same was true for 42 of 57 topics (74%), with 16 exhibiting negligible differences. For the French collection, the comparable statistic is 30 of 46 (65%), with 4 topics exhibiting negligible differences. Per-topic AP differences are plotted in Figure 10 for the grammar-based (B), 10-best decoder-based (C), one-best (D), and interpolated (E) approaches with respect to the word-based approach (A). Points above the x-axis denote higher effectiveness and points below denote lower effectiveness. The topics are sorted left to right by decreasing AP difference for the interpolated model (E).

These plots clearly show that the approaches behave very differently for many topics; instead of small variations across topics, we see instances where one approach really helped or hurt. For instance, in the Arabic collection, the 10-best decoder-based approach (C) is a clear winner for topic 66, where the interpolated model underperforms, but for topic 32, the decoder-based approach really hurt. In the Chinese collection, we see a few topics where the decoder (C) or one-best (D) beats the interpolated model, but far more topics where effectiveness is substantially worse. In the French collection, we see less per-topic variability, but for a few topics the one-best approach (D) is terrible. In all three collections, we note that the grammar-based approach exhibits less per-topic variation than the decoder-based approach (with respect to the interpolated model)—for some topics, the decoder-based approach fails spectacularly (points far below the x-axis), but this is less often the case for the grammar-based approach. Overall, this analysis supports our argument that a combination-of-evidence approach captures the strengths of each individual model and "moderates" the large negative

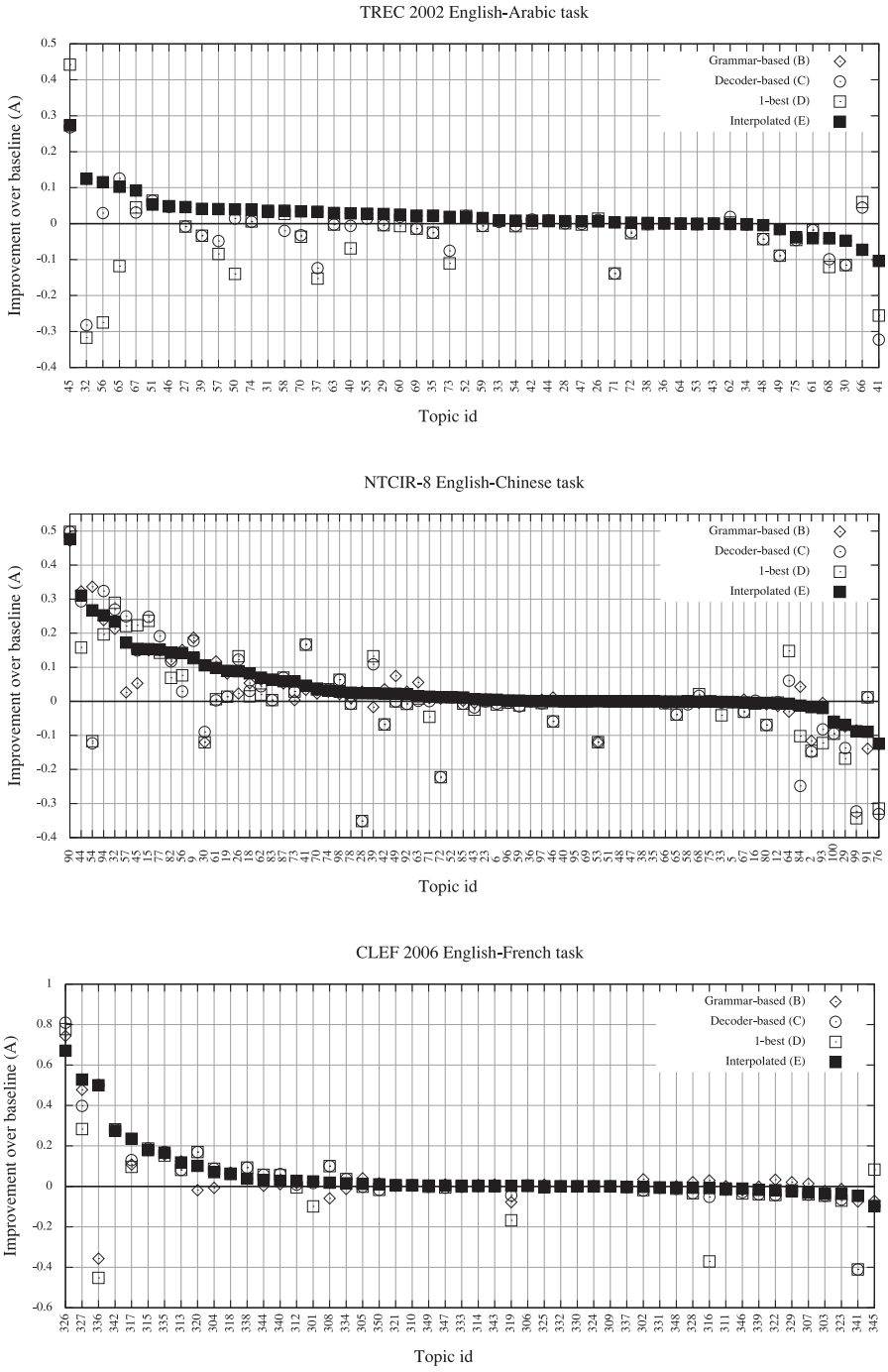

Fig. 10. Per-topic AP differences compared to the word-based approach (condition A) for the TREC 2002 English-Arabic, NTCIR-8 English-Chinese, and CLEF 2006 English-French collections. Topics are sorted by the AP difference for the interpolated model (condition E).

effects with individual models for some topics. The interpolated model is almost always better than the worst individual approach and often close to the best individual approach. The only topic in which the interpolated model (E) is worse than all other approaches (B, C, and D) is CLEF topic 326.

We analyzed a few topics in more detail to gain further insight. As expected, the decoder-based approach (C) is highly effective when appropriate translations are found. As an example, CLEF topic 336 *NBA labor conflict* is translated into the following query:

```
#comb(#wsyn(1.0 nba)
      #wsyn(1.0 travail)
      #wsyn(0.96 confl, 0.04 conflit))
```

In contrast, with the grammar-based (B) and word-based (A) approaches, the structured query contains other translation alternatives such as *contradiction* instead of *conflit* or words related to *labor* such as *social*. These alternative translations introduce noise and hurt retrieval effectiveness. On the other hand, the one-best MT approach (D) works poorly because the top scoring hypothesis omits the translation of *labor* altogether. The second-best translation does not suffer from this issue, serving as an excellent example of the benefits of using *n*-best translations instead of only relying on the single best.

Another interesting case is CLEF topic 313, *centenary celebrations*, which is translated correctly into French as *centenaire* by the decoder. However, the task of CLIR is not only about finding the best translation, but retrieving relevant documents. The grammar-based approach (B) includes *célébrations* (Eng. *celebration*) in addition to *centenaire*, which increases recall and improves the average precision. This is an example where translation diversity is beneficial—as we have previously discussed, alternative translation better alleviates mismatches between vocabularies used by searchers and document writers.

A case that illustrates the downside of relying on language models is CLEF topic 341 *theft of "the scream."* In this case, the translation candidate *vol du "cri"* is downweighted by the language model since it is a sequence of words never seen before (in the training data). Instead, the decoder picks *vol de "scream"* as the top translation, which results in lower retrieval effectiveness.[8] When the MT system fails to find an appropriate translation, the word-based approach (A) is superior because it does not discard translation alternatives. In this case, the grammar-based approach (B) performs slightly worse than the word-based approach (A), but much better than the decoder-based approach (C).

7.4. Efficiency

An important, but often neglected, aspect of cross-language information retrieval is efficiency. In general, query translation approaches generate complex structured queries, which are substantially slower to evaluate than simple bag-of-words queries. An approach that is significantly more effective may be substantially slower—whether the trade-off is worthwhile depends on the application, but effectiveness/efficiency trade-offs should be explicitly explored to better inform the system designer.

In this section, we present experimental results that quantify the efficiency of the previously-explored techniques in terms of per-topic query latency. Experiments were performed on a server running Red Hat Linux, with dual Intel Xeon "Westmere" quadcore processors (E5620 2.4GHz) and 128GB RAM. Note that none of the inverted

[8]"Scream" also refers to a popular horror movie.

Table IV. Average Query Latency (in *ms*) for the TREC 2002 English-Arabic Collection, Broken Down into the Various Processing Stages

	Process	P_{word}	P_{SCFG}	P_{nbest}		P_{c}
				1-best	10-best	
MT	Extraction	-	7.6			
	Decoding	-	-	134.9		
IR	Generation	48.1	64.4	5.8	62.3	113.5
	Ranking	545.6	514.2	97.6	179.0	602.0
Total time (in *ms*)		594±22	586±13	246±15	383±22	858±20

indexes are very big, and experiments were conducted with a "warm" cache, which likely meant that postings lists were memory resident (based on OS-level caching). All experiments ran in a single thread, one topic after another sequentially. We focused on the TREC 2002 Arabic-English topics, although results from the other collections are qualitatively similar. End-to-end average query evaluation latency (measured in milliseconds) is shown in the bottom row of Table IV for each of our approaches. The values reported are averaged across three trials with 95% confidence intervals shown.

As outlined in Section 2, the three main stages in the MT pipeline are word alignment, grammar extraction, and decoding. Word alignment is treated as a preprocessing step since it is query-independent and required for all three approaches; once the alignments are generated, they can be stored on disk and loaded when needed. Thus, the time for word alignment is not included in our evaluation. For the grammar-based approach P_{SCFG} we need to extract the rules that apply to each query, whereas the decoder-based P_{nbest} and interpolation P_{c} approaches require decoding. Decoding is relatively expensive since it involves a search through the hypothesis space, but there is no measurable difference between generating one-best and 10-best hypotheses. Since queries are short, MT processing times are much lower than for typical translation tasks that involve complete sentences.

The remaining two processing stages are part of retrieval: query generation and document ranking. For query generation, we only measure query-dependent costs, since other costs such as loading the bilingual dictionary need to be performed only once at startup. For the word-based approach, P_{word} can be computed for all words in the vocabulary before query time, so we only need to preprocess the query and load precomputed translation probabilities. For the grammar-based approach, query generation involves loading the extracted translation grammar and processing the rules to compute P_{SCFG}. For the decoder-based approach, each translation hypothesis needs to be processed to compute P_{nbest}, so query generation time increases (roughly) linearly with n.

The speed of the final step, document ranking (i.e., query evaluation), depends on the complexity of the structured query—in particular, the number of clauses and the number of translation alternatives. All our approaches generate the same number of clauses but differ in the number of query terms in each clause. The distributions P_{word} and P_{SCFG} usually contain more translation alternatives than P_{nbest}, thus resulting in slower query evaluation. For this reason, query evaluation with the interpolated model is slower as well.

From Table IV, we see that the n-best approach is significantly faster than the word-based approach, even though it requires additional MT processing to translate the queries. When $n = 1$, the reduction in total running time is nearly 60%, and with 10-best, 35%. A similar trade-off exists for the word-based approach: the structured queries can be simplified if more aggressive thresholding is applied, for example, if C or H increased in Equation (10), but at the cost of effectiveness. In terms of efficiency, there is not much difference between the grammar-based approach and the

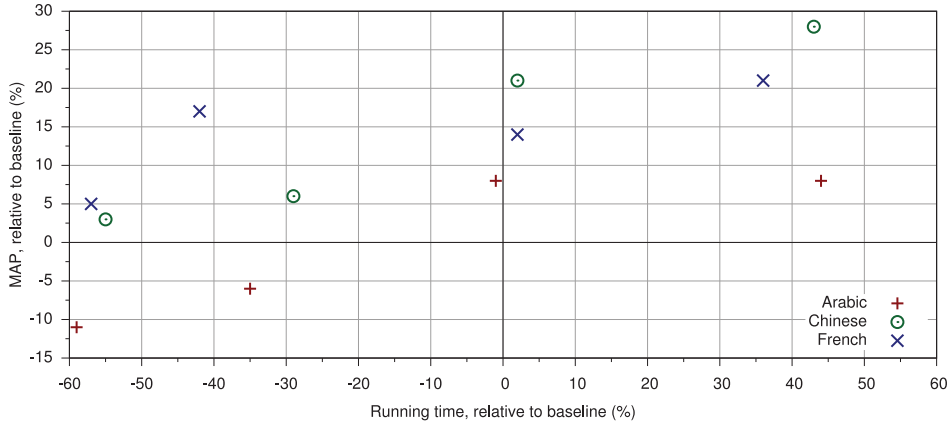

Fig. 11. Visualization of the effectiveness/efficiency trade-offs for our CLIR models.

word-based approach, but since the grammar-based approach is more effective, it should be preferred overall. The interpolated model P_c is the most effective but also the slowest. Compared to the word-based approach, running time is 44% longer, which might be acceptable given the significant improvements, but this ultimately depends on the application.

For a better understanding of the trade-off between effectiveness and efficiency, Figure 11 shows MAP with respect to total running time for all three collections. Instead of absolute values, we plot relative differences with respect to the word-based approach. As expected, we see a general trend of more effective approaches being slower for all three collections. The lower-right quadrant contains settings that are slower and worse (fortunately, no settings fall into this category). The upper-left quadrant contains settings that are faster and better: the one-best and 10-best approaches appear here for the Chinese and French collections as well as the grammar-based approach for the Arabic collection. The other two quadrants represent a trade-off of efficiency for effectiveness and vice versa.

Efficiency results reported in Table IV and Figure 11 were all computed using a hierarchical MT system (`cdec`). Although flat MT approaches are considered faster than hierarchical MT, we did not observe much difference in the MT running time when Moses was used instead of `cdec` for our queries. Thus, the choice of grammars does not appear to have much of an efficiency impact for the decoder-based approach. However, for the grammar-based approach, using flat grammars is a poor choice—the generation step takes more than an order of magnitude longer because of the verbosity of flat grammars.

Ultimately, the best balance between effectiveness and efficiency depends on the end application, but we can offer some guidance. For a faster and possibly more effective model, P_{nbest} and P_{SCFG} seem to be good alternatives to P_{word}. The interpolated model is significantly more effective, but substantially slower.

8. FUTURE WORK AND CONCLUSIONS

In this article, we introduced a framework for exploiting internal representations of modern statistical machine translation systems for cross-language information retrieval. Effective use of these representations requires balancing the use of context to disambiguate translation candidates with the need to preserve diversity in translation alternatives. We proposed two specific stages in the MT pipeline that are particularly amenable to integration with cross-language retrieval: with the grammar-based

approach, we construct translation probabilities from the translation grammar extracted for a specific query, thereby incorporating context as captured in the translation rules; with the decoder-based approach, we reconstruct translation probabilities from the *n*-best translations, thereby taking advantage of both the translation model and the language model. Within this framework, we explored design alternatives regarding the choice of flat vs. hierarchical grammars and different heuristics for handling multiple word alignments. Experiments show that an interpolated model which combines evidence from the word-, grammar-, and decoder-based approaches is significantly more effective than competitive baselines across multiple collections. These results advance the state of the art in cross-language information retrieval, but there are two limitations we hope to address in future work.

The first limitation is that the optimal setting of parameters in our interpolated model requires a tuning set. Our cross-validation experiments show that this is possible as long as we have topics and relevance judgments for the same collection. The parameter settings do not generalize across collections, and our simple attempt at "transfer learning" (tuning with heterogeneous collections) was not successful. At present, it is unclear if the specificity of the parameter settings is a result of topic and language effects, or other characteristics of the test collections. It is also possible that data used to train the translation and language models play an important role. For future work, we are attempting to better understand these issues to increase the generalizability of our techniques.

The second limitation is that our techniques currently handle phrases on the target language side but not the source language side. With the one-to-many multiple alignment heuristic, we are able to learn correspondences such as (*brand*, *marque de fabrique*), which are realized as phrase queries against the target collection. However, the source side of these pairs are individual query terms in our present formulation. Put another way, our structured queries always have one clause per query term on the source side, even though the target side translations may contain phrases. The challenge of handling source phrases concerns the proper treatment of ambiguous phrase segmentation. For example, suppose a query with three terms, *A*, *B*, and *C*, can be segmented as (*A B*, *C*) or (*A*, *B C*): What's the form of the structured query that captures these alternatives, and how would documents be scored accordingly? Solving this problem in the general case requires a technique that will handle queries in the form of arbitrary lattices, since alternative segmentations will necessarily overlap. We are not aware of any current solutions, but this would be an interesting future direction to explore.

In conclusion, we believe this work represents a good example of the synergies that are possible from closer integration between disparate disciplines. There is little intersection between researchers who work on information retrieval and those who work on machine translation, but clearly there is much to gain from cross-language retrieval techniques that treat MT systems as more than black boxes. We hope that our work will foster greater collaboration between the two communities who are ultimately working on the same goal of improving multilingual information access.

ACKNOWLEDGMENTS

We'd like to thank the anonymous reviewers for comments that have helped improve this work. J. Lin is grateful to Esther for her loving support and dedicates this work to Joshua and Jacob.

REFERENCES

M. Adriani and C. J. V. Rijsbergen. 2000. Phrase identification in cross-language information retrieval. In *Proceedings of RIAO: Content-Based Multimedia Information Access*.

A. T. Arampatzis, T. Tsoris, C. H. A. Koster, and P. V. D. Weide. 1998. Phrase-based information retrieval. *Inf. Process. Manag.* 34, 6, 693–707.

L. Ballesteros and B. Croft. 1996. Dictionary methods for cross-lingual information retrieval. In *Proceedings of the 7th International DEXA Conference on Database and Expert Systems Applications*. 791–801.

L. Ballesteros and W. B. Croft. 1997. Phrasal translation and query expansion techniques for cross-language information retrieval. In *Proceedings of the 20th International ACM Conference on Research and Development in Information Retrieval (SIGIR'97)*. 84–91.

A. Berger and J. Lafferty. 1999. Information retrieval as statistical translation. In *Proceedings of the 22nd Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR'99)*. 222–229.

P. F. Brown, J. Cocke, S. A. D. Pietra, V. J. D. Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S. Roossin. 1990. A statistical approach to machine translation. *Comput. Ling.* 16, 2, 79–85.

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

G. Cao, J.-Y. Nie, and J. Bai. 2006. Constructing better document and query models with Markov chains. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management (CIKM'06)*. 800–801.

A. Chen. 2000. Phrasal translation for English-Chinese cross language information retrieval. In *Proceedings of the Workshop on English-Chinese Cross Language Information Retrieval at the International Conference on Chinese Language Computing*. 195–202.

D. Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*. 263–270.

D. Chiang. 2007. Hierarchical phrase-based translation. *Comput. Ling.* 33, 2, 201–228.

K. Darwish and D. W. Oard. 2003. Probabilistic structured query methods. In *Proceedings of the 26th Annual International ACM Conference on Research and Development in Informaion Retrieval (SIGIR'03)*. 338–344.

A. P. Dempster, N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Stat. Soc. Series B* 39, 1, 1–38.

C. Dyer, J. Weese, H. Setiawan, A. Lopez, F. Ture, V. Eidelman, J. Ganitkevitch, P. Blunsom, and P. Resnik. 2010. cdec: A decoder, alignment, and learning framework for finite-state and context-free translation models. In *Proceedings of the ACL System Demonstrations (ACL'10)*. 7–12.

M. Federico and N. Bertoldi. 2002. Statistical cross-language information retrieval using n-best query translations. In *Proceedings of the 25th Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR'02)*. 167–174.

A. Fraser, J. Xu, and R. Weischedel. 2002. TREC 2002 cross-lingual retrieval at BBN. In *Proceedings of the 11th Text REtrieval Conference (TREC'02)*.

G. W. Furnas, S. Deerwester, S. T. Dumais, T. K. Landauer, R. A. Harshman, L. A. Streeter, and K. E. Lochbaum. 1988. Information retrieval using a singular value decomposition model of latent semantic structure. In *Proceedings of the 11th Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR'88)*. 465–480.

J. Gao, J.-Y. Nie, G. Wu, and G. Cao. 2004. Dependence language model for information retrieval. In *Proceedings of the 27th Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR'04)*. 170–177.

J. Gao, J.-Y. Nie, E. Xun, J. Zhang, M. Zhou, and C. Huang. 2001. Improving query translation for cross-language information retrieval using statistical models. In *Proceedings of the 24th Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR'01)*. 96–104.

H. Hayurani, S. Sari, and M. Adriani. 2007. Query and document translation for English-Indonesian cross language IR. In *Proceedings of the 7th International Conference on Cross-Language Evaluation Forum (CLEF'06)*. 57–61.

D. A. Hull and G. Grefenstette. 1996. Querying across languages: A dictionary-based approach to multilingual information retrieval. In *Proceedings of the 19th Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR'96)*. 49–57.

K. Kishida and N. Kando. 2006. A hybrid approach to query and document translation using a pivot language for cross-language information retrieval. In *Proceedings of the 6th International Conference on Cross-Language Evalution Forum (CLEF'05)*. 93–101.

P. Koehn. 2010. *Statistical Machine Translation*. Cambridge University Press.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the ACL Demo and Poster Sessions (ACL'07)*. 177–180.

P. Koehn, F. J. Och, and D. Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL'03)*. 48–54.

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

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

V. Lavrenko, M. Choquette, and W. B. Croft. 2002. Cross-lingual relevance models. In *Proceedings of the 25th Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR'02)*. 175–182.

V. Lavrenko and W. B. Croft. 2001. Relevance-based language models. In *Proceedings of the 24th Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR'01)*. 120–127.

Z. Li, J. Eisner, and S. Khudanpur. 2009. Variational decoding for statistical machine translation. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. 593–601.

M. Littman, S. T. Dumais, and T. K. Landauer. 1998. Automatic cross-language information retrieval using latent semantic indexing. In *Cross-Language Information Retrieval*. Kluwer Academic Publishers, 51–62.

Y. Liu, R. Jin, and J. Y. Chai. 2005. A maximum coherence model for dictionary-based cross-language information retrieval. In *Proceedings of the 28th Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR'05)*. 536–543.

A. Lopez. 2007. Hierarchical phrase-based translation with suffix arrays. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. 976–985.

A. Lopez. 2008. Statistical Machine Translation. *ACM Comput. Surv.* 40, 3, 8:1–8:49.

Y. Ma, J.-Y. Nie, H. Wu, and H. Wang. 2012. Opening machine translation black box for cross-language information retrieval. In *Information Retrieval Technology*. Lecture Notes in Computer Science, Vol. 7675, Springer, Berlin, 467–476.

W. Magdy and G. J. F. Jones. 2011. Should MT systems be used as black boxes in CLIR? In *Proceedings of the 33rd European Conference on Information Retrieval (ECIR'11)*. 683–686.

D. Marcu and W. Wong. 2002. A phrase-based, joint probability model for statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'02)*. 133–139.

J. S. McCarley. 1999. Should we translate the documents or the queries in cross-language information retrieval? In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*. 208–214.

J. S. McCarley and S. Roukos. 1998. Fast document translation for cross-language information retrieval. In *Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas (AMTA'98)*. 150–157.

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

D. Metzler and W. B. Croft. 2004. Combining the language model and inference network approaches to retrieval. *Inf. Process. Manag.* 40, 5, 735–750.

D. Metzler and W. B. Croft. 2005. A Markov random field model for term dependencies. In *Proceedings of the 28th Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR'05)*. 472–479.

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

V. Nikoulina, B. Kovachev, N. Lagos, and C. Monz. 2012. Adaptation Of statistical machine translation model for cross-language information retrieval in a service context. In *Proceedings of 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL'12)*. 109–119.

D. W. Oard. 1998. A comparative study of query and document translation for cross-language information retrieval. In *Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas (AMTA'98)*.

D. W. Oard and P. Hackett. 1997. Document translation for cross-language text retrieval at the University of Maryland. In *Proceedings of the 7th Text REtrieval Conference (TREC-7)*.

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

F. J. Och and H. Ney. 2004. The alignment template approach to statistical machine translation. *Comput. Ling.* 30, 4, 417–449.

F. J. Och, C. Tillmann, and H. Ney. 1999. Improved alignment models for statistical machine translation. In *Proceedings of the 3rd Conference on Empirical Methods for Natural Language Processing (EMNLP'99)*. 20–28.

J. Olive, C. Christianson, and J. McCary. 2011. *Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation*. Springer.

J. S. Olsson and D. W. Oard. 2009. Combining LVCSR and vocabulary-independent ranked utterance retrieval for robust speech search. In *Proceedings of the 32nd Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR'09)*. 91–98.

A. Pirkola. 1998. The effects of query structure and dictionary-setups in dictionary-based cross-language information retrieval. In *Proceedings of the 21st Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR'98)*. 55–63.

J. M. Ponte and W. B. Croft. 1998. A language modeling approach to information retrieval. In *Proceedings of the 21st Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR'98)*. 275–281.

S. E. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford. 1994. Okapi at TREC-3. In *Proceedings of the 3rd Text REtrieval Conference (TREC-3)*. 109–126.

J. Savoy and S. Abdou. 2007. Experiments with monolingual, bilingual, and robust retrieval. In *Proceedings of the 7th International Conference on Cross-Language Evaluation Forum (CLEF'06)*. 137–144.

H.-C. Seo, S.-B. Kim, H.-C. Rim, and S.-H. Myaeng. 2005. Improving query translation in English-Korean cross-language information retrieval. *Inf. Process. Manag.* 41, 3, 507–522.

M. D. Smucker, J. Allan, and B. Carterette. 2007. A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of the 16th ACM conference on Conference on Information and Knowledge Management (CIKM'07)*. 623–632.

A. Stolcke. 2002. SRILM – An extensible language modeling toolkit. In *Proceedings of the 7th International Conference on Spoken Language Processing*. 901–904.

H. Tseng, P.-C. Chang, G. Andrew, D. Jurafsky, and C. Manning. 2005. A conditional random field word segmenter. In *Proceedings of the 4th SIGHAN Workshop on Chinese Language Processing*.

F. Ture. 2013. Searching to translate and translating to search: When information retrieval meets machine translation. Ph.D. Dissertation, University of Maryland.

F. Ture and J. Lin. 2013. Flat vs. hierarchical phrase-based translation models for cross-language information retrieval. In *Proceedings of the 36th International ACM Conference on Research and Development in Information Retrieval (SIGIR'13)*. 813–816.

F. Ture, J. Lin, and D. W. Oard. 2012a. Combining statistical translation techniques for cross-language information retrieval. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING'12)*. 2685–2702.

F. Ture, J. Lin, and D. W. Oard. 2012b. Looking inside the box: Context-sensitive translation for cross-language information retrieval. In *Proceedings of the 35th International ACM Conference on Research and Development in Information Retrieval (SIGIR'12)*. 1105–1106.

J. Wang and D. W. Oard. 2006. Combining bidirectional translation and synonymy for cross-language information retrieval. In *Proceedings of the 29th Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR'06)*. 202–209.

D. Wu and D. He. 2010. Exploring the further integration of machine translation in multilingual information access. In *Proceedings of the iConference*.

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

J. Xu, R. Weischedel, and C. Nguyen. 2001. Evaluating a probabilistic model for cross-lingual information retrieval. In *Proceedings of the 24th Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR'01)*. 105–110.

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

Received December 2013; revised May 2014; accepted July 2014