Looking Inside the Box: Context-Sensitive Translation for Cross-Language Information Retrieval

Ferhan Ture[1], Jimmy Lin[2,3], Douglas W. Oard[2,3]

[1]Dept. of Computer Science, [2]College of Information Studies, [3]UMIACS
University of Maryland

fture@cs.umd.edu, jimmylin@umd.edu, oard@umd.edu

ABSTRACT

Cross-language information retrieval (CLIR) today is dominated by techniques that use token-to-token mappings from bilingual dictionaries. Yet, state-of-the-art statistical translation models (e.g., using Synchronous Context-Free Grammars) are far richer, capturing multi-term phrases, term dependencies, and contextual constraints on translation choice. We present a novel CLIR framework that is able to reach inside the translation "black box" and exploit these sources of evidence. Experiments on the TREC-5/6 English-Chinese test collection show this approach to be promising.

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

Keywords: machine translation, context

1. INTRODUCTION

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

We argue that this false choice results from thinking of MT systems as black boxes [5]. Inside an MT system we find not alternative translations for individual tokens, but rather for entire sentences. In state-of-the-art MT systems, these alternative "readings" are based on Synchronous Context-Free Grammars (SCFG) and pieced together from units of varying size with complex hierarchical dependencies. Reducing these to context independent token translation probabilities discards potentially useful contextual constraints. An elegant solution, which we explore in this work, is to perform translation in context using a full SCFG MT system and then to reconstruct context-sensitive n-best token translation probabilities by tokenizing each reading and accumulating translation likelihood evidence, which can then be renormalized as estimates of probabilities. This technique is now routinely used in speech retrieval [7], but we are not aware of its prior use for CLIR.

These context-sensitive token translation probabilities can then be used in the same way as context-independent probabilities. We use a technique based on mapping term statistics before computing term weights [8, 2] to establish a strong context-independent baseline. Experiments on the TREC-5/6 English-Chinese CLIR task show that our new approach yields promising (although not statistically significant) improvements over that baseline.

SIGIR'12, August 12–16, 2012, Portland, Oregon, USA.
ACM 978-1-4503-1472-5/12/08.

2. APPROACH

We consider the technique presented by Darwish and Oard [2] as the baseline. Given a source-language query s, we represent s in the target language as a probabilistic structured query, where weights are derived from word-to-word bilingual translation probabilities that are learned automatically from parallel text ("bitext"):

$$\text{Score}(D|s) = \sum_{j=1}^{\#\text{tokens}} \text{bm25}(\text{tf}(s_j, D), \text{df}(s_j)) \quad (1)$$

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

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

where L is a lower bound on conditional probability. We also impose a cumulative probability threshold, C, so that translation alternatives are added (starting from most probable ones) until the cumulative probability has reached C. We use the Okapi BM25 term weighting function, although in principle any other weighting function can be substituted.

Compared against this baseline, we show how we can take advantage of a full MT decoder to better estimate translation probabilities. A decoder uses a translation model (TM) and language model (LM) to find all possible derivations of a source text, and the corresponding translations in the target language, along with associated derivation scores. We use `cdec` [3], a state-of-the-art MT system which provides fast "decoding" using *Hiero*-style synchronous grammars [1] for representing the translation model in a way that can model distant dependencies within a sentence.

As a point of comparison, we might use only the best translation:

$$\text{Score}(D|s) = \sum_{j=1}^{m} \text{bm25}(\text{TF}(t_i^{(1)}, D), \text{DF}(t_i^{(1)})) \quad (4)$$

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

$$t^{(1)} = \arg\max_t \ell(t|s) = \arg\max_t \text{TM}(s,t)\text{LM}(t) \quad (5)$$

where ℓ is the likelihood function, a mapping learned by the decoder, which scores each derivation using the TM and LM.

Decoders produce a set of candidate sentence translations in the process of computing equation (5), so we can generalize our model

to consider the n candidates with highest likelihood, for some $n > 1$. In this case, the score of document D would be a weighted average of scores with respect to each candidate translation:

$$\text{Score}(D|f) = \sum_{k=1}^{N} \text{Score}(D|e^{(k)}) Pr_{\text{cdec}}(e^{(k)}|f) \qquad (6)$$

where Pr_{cdec} is the normalized likelihood value.

In order to compute tf and df statistics for tokens, we start by tokenizing each candidate sentence translation. For each token s_j of source query s, we use word alignments in the grammar rules to determine which target tokens it is associated with. By doing this, we are constructing a probability distribution of possible translations of s_j based on the n query translations. Specifically, if source token s_j is aligned to (i.e., translated as) t_i in the k^{th} best translation, it receives a weight equal to $Pr_{\text{cdec}}(t^{(k)}|s)$.[1] As a result, we can map tf and df statistics by replacing Pr_{bitext} with Pr_{nbest} (see below, φ is the normalization factor) in Equations 2 and 3.

$$Pr_{\text{nbest}}(t_i|s_j) = \frac{1}{\varphi} \sum_{k=1}^{N} \sum_{\substack{t_i \\ s_j \text{ aligned to } t_i \text{ in } t^{(k)}}} Pr_{\text{cdec}}(t^{(k)}|s) \qquad (7)$$

This new probability distribution (i.e., Pr_{nbest}) is based only on the n translations that the decoder scores highest for the source query. Therefore, the distribution is informed by the query context and its derivation by the translation model. From this we would expect the distribution to be better biased in favor of appropriate translations, but perhaps at the cost of some reduction in variety due to overfitting. Finally, we can combine the two probability estimates to mitigate overfitting using simple linear interpolation: $Pr_c(s_j) = \lambda Pr_{\text{nbest}}(s_j) + (1 - \lambda) Pr_{\text{bitext}}(s_j)$.

3. EVALUATION

We evaluated our system on the TREC-5/6 CLIR task, using a corpus of 164,778 Chinese documents and titles of the 54 English topics as queries. The evaluation metric is Mean Average Precision (MAP). The English-to-Chinese translation model was trained using the FBIS parallel text collection, which contains 1.6 million parallel sentences. The Chinese collection was tokenized using the Stanford segmenter for Chinese, the Porter stemmer was used for English, and alignment was performed using GIZA++ [6]. A SCFG was extracted from these alignments using a suffix array [4]. A Chinese token 3-gram model serves as the LM.

Results are summarized in Figure 1. At the left edge of the graph, at $\lambda = 0$, we have the approach in equation (2) with context-independent translation probabilities (call this A).[2] At the right edge of the graph, at $\lambda = 1.0$, we rely exclusively on context-sensitive probabilities (call this B). Effectiveness peaks at $\lambda = 0.78$ (call this C).[3] For reference, the horizontal line represents simply taking the one-best translation from the MT system (call this D). We also tried one-best context-independent translation for each token, and the MAP score was 0.2431. A randomization test shows that this is significantly below A, B, C and D ($p < 0.05$). On the other hand, A, B, C, and D are statistically indistinguishable on this test collection. Yet, results are still promising: when C is compared to A and D, the p-value is approximately 0.15 and 0.18, respectively. Also, a topic-specific analysis shows that C yields better average precision than D on 36 of the 54 topics (which actually is significant by a two-tailed sign test at $p < 0.05$).

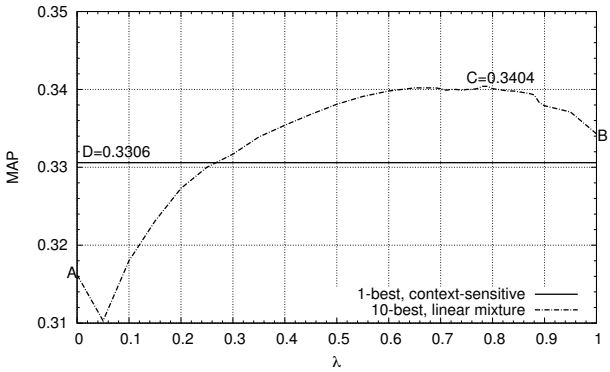

Figure 1: Evaluation on TREC-5/6 English-Chinese CLIR task.

[1] Since a source term may be aligned to multiple target terms in the same query translation, we still need to normalize the final weights.

[2] We selected $C = 0.95$ and $L = 0.005$ for baseline model parameters after manually trying a range of values.

[3] We selected $N = 10$ for conditions B and C, but $N = 5$ also yields similar results, peaking at a MAP score of 0.3387.

4. CONCLUSIONS AND FUTURE WORK

In this work, we have introduced an approach that combines the representational advantage of probabilistic structured queries with the richness of the internal representation of a translation model. We introduced a novel way to learn term translation probabilities from the top scoring "readings" of alternative query translations, as generated by the decoder. We evaluated our approach on the English-Chinese CLIR task of TREC-5/6: although we did not observe significant improvements, we feel that this approach is nevertheless promising. In future work we plan to try this approach for document translation (where we would expect greater benefit from context, although with higher computational cost, at least in experimental settings). Replications on test collections with larger numbers of topics, and with a greater variety of query and topic languages, can also be expected to yield additional insights.

Acknowledgements. This work was supported in part by DARPA BOLT under contract HR0011-12-C-0015. Any opinions, findings, conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the view of DARPA.

5. REFERENCES

[1] D. Chiang. Hierarchical phrase-based translation. *Computational Linguistics*, 33:201–228, 2007.

[2] K. Darwish and D. W. Oard. Probabilistic structured query methods. In *SIGIR*, 2003.

[3] C. Dyer, J. Weese, H. Setiawan, A. Lopez, F. Ture, V. Eidelman, J. Ganitkevitch, P. Blunsom, and P. Resnik. cdec: a decoder, alignment, and learning framework for finite-state and context-free translation models. In *ACL Demos*, 2010.

[4] A. Lopez. Hierarchical phrase-based translation with suffix arrays. In *EMNLP*, 2007.

[5] W. Magdy and G. Jones. Should MT systems be used as black boxes in CLIR? In *ECIR*, 2011.

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

[7] J. Olsson and D. Oard. Combining LVCSR and vocabulary-independent ranked utterance retrieval for robust speech search. In *SIGIR*, 2009.

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