Task Alternation in Parallel Sentence Retrieval for Twitter Translation

Felix Hieber and **Laura Jehl** and **Stefan Riezler**
Department of Computational Linguistics
Heidelberg University
69120 Heidelberg, Germany
{jehl,hieber,riezler}@cl.uni-heidelberg.de

Abstract

We present an approach to mine comparable data for parallel sentences using translation-based cross-lingual information retrieval (CLIR). By iteratively alternating between the tasks of retrieval and translation, an initial general-domain model is allowed to adapt to in-domain data. Adaptation is done by training the translation system on a few thousand sentences retrieved in the step before. Our setup is time- and memory-efficient and of similar quality as CLIR-based adaptation on millions of parallel sentences.

1 Introduction

Statistical Machine Translation (SMT) crucially relies on large amounts of bilingual data (Brown et al., 1993). Unfortunately sentence-parallel bilingual data are not always available. Various approaches have been presented to remedy this problem by mining parallel sentences from comparable data, for example by using cross-lingual information retrieval (CLIR) techniques to retrieve a target language sentence for a source language sentence treated as a query. Most such approaches try to overcome the noise inherent in automatically extracted parallel data by sheer size. However, finding good quality parallel data from noisy resources like Twitter requires sophisticated retrieval methods. Running these methods on millions of queries and documents can take weeks.

Our method aims to achieve improvements similar to large-scale parallel sentence extraction approaches, while requiring only a fraction of the extracted data and considerably less computing resources. Our key idea is to extend a straightforward application of translation-based CLIR to an iterative method: Instead of attempting to retrieve in one step as many parallel sentences as possible, we allow the retrieval model to gradually adapt to new data by using an SMT model trained on the freshly retrieved sentence pairs in the translation-based retrieval step. We alternate between the tasks of translation-based retrieval of target sentences, and the task of SMT, by re-training the SMT model on the data that were retrieved in the previous step. This task alternation is done iteratively until the number of newly added pairs stabilizes at a relatively small value.

In our experiments on Arabic-English Twitter translation, we achieved improvements of over 1 BLEU point over a strong baseline that uses in-domain data for language modeling and parameter tuning. Compared to a CLIR-approach which extracts more than 3 million parallel sentences from a noisy comparable corpus, our system produces similar results in terms of BLEU using only about 40 thousand sentences for training in each of a few iterations, thus being much more time- and resource-efficient.

2 Related Work

In the terminology of semi-supervised learning (Abney, 2008), our method resembles self-training and co-training by training a learning method on its own predictions. It is different in the aspect of task alternation: The SMT model trained on retrieved sentence pairs is not used for generating training data, but for scoring noisy parallel data in a translation-based retrieval setup. Our method also incorporates aspects of transductive learning in that candidate sentences used as queries are filtered for out-of-vocabulary (OOV) words and similarity to sentences in the development set in order to maximize the impact of translation-based retrieval.

Our work most closely resembles approaches that make use of variants of SMT to mine comparable corpora for parallel sentences. Recent work uses word-based translation (Munteanu and

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 323–327,
Sofia, Bulgaria, August 4-9 2013.

Marcu, 2005; Munteanu and Marcu, 2006), full-sentence translation (Abdul-Rauf and Schwenk, 2009; Uszkoreit et al., 2010), or a sophisticated interpolation of word-based and contextual translation of full sentences (Snover et al., 2008; Jehl et al., 2012; Ture and Lin, 2012) to project source language sentences into the target language for retrieval. The novel aspect of task alternation introduced in this paper can be applied to all approaches incorporating SMT for sentence retrieval from comparable data.

For our baseline system we use in-domain language models (Bertoldi and Federico, 2009) and meta-parameter tuning on in-domain development sets (Koehn and Schroeder, 2007).

3 CLIR for Parallel Sentence Retrieval

3.1 Context-Sensitive Translation for CLIR

Our CLIR model extends the translation-based retrieval model of Xu et al. (2001). While translation options in this approach are given by a lexical translation table, we also select translation options estimated from the decoder's n-best list for translating a particular query. The central idea is to let the language model choose fluent, context-aware translations for each query term during decoding.

For mapping source language query terms to target language query terms, we follow Ture et al. (2012a; 2012). Given a source language query Q with query terms q_j, we project it into the target language by representing each source token q_j by its probabilistically weighted translations. The score of target document D, given source language query Q, is computed by calculating the Okapi BM25 rank (Robertson et al., 1998) over projected term frequency and document frequency weights as follows:

$$score(D|Q) = \sum_{j=1}^{|Q|} bm25(tf(q_j, D), df(q_j))$$

$$tf(q, D) = \sum_{i=1}^{|T_q|} tf(t_i, D) P(t_i|q)$$

$$df(q) = \sum_{i=1}^{|T_q|} df(t_i) P(t_i|q)$$

where $T_q = \{t|P(t|q) > L\}$ is the set of translation options for query term q with probability greater than L. Following Ture et al. (2012a; 2012) we impose a cumulative threshold C, so that only the most probable options are added until C is reached.

Like Ture et al. (2012a; 2012) we achieved best retrieval performance when translation probabilities are calculated as an interpolation between (context-free) lexical translation probabilities P_{lex} estimated on symmetrized word alignments, and (context-aware) translation probabilities P_{nbest} estimated on the n-best list of an SMT decoder:

$$P(t|q) = \lambda P_{nbest}(t|q) + (1 - \lambda) P_{lex}(t|q) \quad (1)$$

$P_{nbest}(t|q)$ is the decoder's confidence to translate q into t within the context of query Q. Let $a_k(t, q)$ be a function indicating an alignment of target term t to source term q in the k-th derivation of query Q. Then we can estimate $P_{nbest}(t|q)$ as follows:

$$P_{nbest}(t|q) = \frac{\sum_{k=1}^{n} a_k(t, q) \mathcal{D}(k, Q)}{\sum_{k=1}^{n} a_k(\cdot, q) \mathcal{D}(k, Q)} \quad (2)$$

$\mathcal{D}(k, Q)$ is the model score of the k-th derivation in the n-best list for query Q.

In our work, we use hierarchical phrase-based translation (Chiang, 2007), as implemented in the `cdec` framework (Dyer et al., 2010). This allows us to extract word alignments between source and target text for Q from the SCFG rules used in the derivation. The concept of self-translation is covered by the decoder's ability to use pass-through rules if words or phrases cannot be translated.

3.2 Task Alternation in CLIR

The key idea of our approach is to iteratively alternate between the tasks of retrieval and translation for efficient mining of parallel sentences. We allow the initial general-domain CLIR model to adapt to in-domain data over multiple iterations. Since our set of in-domain queries was small (see 4.2), we trained an adapted SMT model on the concatenation of general-domain sentences and in-domain sentences retrieved in the step before, rather than working with separate models.

Algorithm 1 shows the iterative task alternation procedure. In terms of semi-supervised learning, we can view algorithm 1 as non-persistent as we do not keep labels/pairs from previous iterations. We have tried different variations of label persistency but did not find any improvements. A similar effect of preventing the SMT model to "forget" general-domain knowledge across iterations is achieved by mixing models from current and previous iterations. This is accomplished in two ways: First, by linearly interpolating the translation option weights $P(t|q)$ from the current and

Algorithm 1 Task Alternation

Require: source language Tweets Q_{src}, target language Tweets D_{trg}, general-domain parallel sentences S_{gen}, general-domain SMT model M_{gen}, interpolation parameter θ

procedure TASK-ALTERNATION($Q_{src}, D_{trg}, S_{gen}, M_{gen}, \theta$)
  $t \leftarrow 1$
  while true **do**
    $S_{in} \leftarrow \emptyset$ ▷ Start with empty parallel in-domain sentences
    if $t == 1$ **then**
      $M_{clir}^{(t)} \leftarrow M_{gen}$ ▷ Start with general-domain SMT model for CLIR
    else
      $M_{clir}^{(t)} \leftarrow \theta M_{smt}^{(t-1)} + (1-\theta) M_{smt}^{(t)}$ ▷ Use mixture of previous and current SMT model for CLIR
    end if
    $S_{in} \leftarrow \text{CLIR}(Q_{src}, D_{trg}, M_{clir}^{(t)})$ ▷ Retrieve top 1 target language Tweets for each source language query
    $M_{smt}^{(t+1)} \leftarrow \text{TRAIN}(S_{gen} + S_{in})$ ▷ Train SMT model on general-domain and retrieved in-domain data
    $t \leftarrow t + 1$
  end while
end procedure

	BLEU (test)	# of in-domain sents
Standard DA	14.05	-
Full-scale CLIR	14.97	3,198,913
Task alternation	15.31	∼40k

Table 1: Standard Domain Adaptation with in-domain LM and tuning; Full-scale CLIR yielding over 3M in-domain parallel sentences; Task alternation ($\theta = 0.1$, iteration 7) using ∼40k parallel sentences per iteration.

previous model with interpolation parameter θ. Second, by always using $P_{lex}(t|q)$ weights estimated from word alignments on S_{gen}.

We experimented with different ways of using the ranked retrieval results for each query and found that taking just the highest ranked document yielded the best results. This returns one pair of parallel Twitter messages per query, which are then used as additional training data for the SMT model in each iteration.

4 Experiments

4.1 Data

We trained the general domain model M_{gen} on data from the NIST evaluation campaign, including UN reports, newswire, broadcast news and blogs. Since we were interested in relative improvements rather than absolute performance, we sampled 1 million parallel sentences S_{gen} from the originally over 5.8 million parallel sentences.

We used a large corpus of Twitter messages, originally created by Jehl et al. (2012), as comparable in-domain data. Language identification was carried out with an off-the-shelf tool (Lui and Baldwin, 2012). We kept only Tweets classified as Arabic or English with over 95% confidence. After removing duplicates, we obtained 5.5 million Arabic Tweets and 3.7 million English Tweets (D_{trg}). Jehl et al. (2012) also supply a set of 1,022 Arabic Tweets with 3 English translations each for evaluation purposes, which was created by crowdsourcing translation on Amazon Mechanical Turk. We randomly split the parallel sentences into 511 sentences for development and 511 sentences for testing. All URLs and user names in Tweets were replaced by common placeholders. Hashtags were kept, since they might be helpful in the retrieval step. Since the evaluation data do not contain any hashtags, URLs or user names, we apply a post-processing step after decoding in which we remove those tokens.

4.2 Transductive Setup

Our method can be considered transductive in two ways. First, all Twitter data were collected by keyword-based crawling. Therefore, we can expect a topical similarity between development, test and training data. Second, since our setup aims for speed, we created a small set of queries Q_{src}, consisting of the source side of the evaluation data and similar Tweets. Similarity was defined by two criteria: First, we ranked all Arabic Tweets with respect to their term overlap with the development and test Tweets. Smoothed per-sentence BLEU (Lin and Och, 2004) was used as a similarity metric. OOV-coverage served as a second criterion to remedy the problem of unknown words in Twitter translation. We first created a general list of all OOVs in the evaluation data under M_{gen} (3,069 out of 7,641 types). For each of the top 100 BLEU-ranked Tweets, we counted OOV-coverage with respect to the corresponding source Tweet and the general OOV list. We only kept Tweets

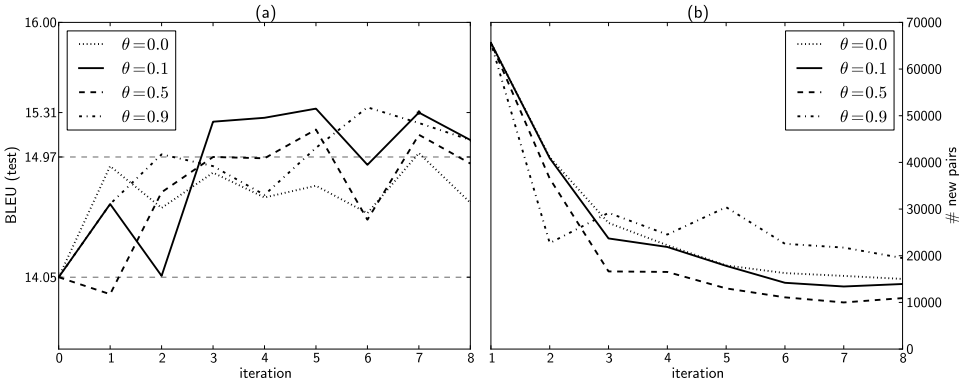

Figure 1: Learning curves for varying θ parameters. (a) BLEU scores and (b) number of new pairs added per iteration.

containing at least one OOV term from the corresponding source Tweet and two OOV terms from the general list, resulting in 65,643 Arabic queries covering 86% of all OOVs. Our query set Q_{src} performed better (14.76 BLEU) after one iteration than a similar-sized set of random queries (13.39).

4.3 Experimental Results

We simulated the full-scale retrieval approach by Jehl et al. (2012) with the CLIR model described in section 3. It took 14 days to run 5.5M Arabic queries on 3.7M English documents. In contrast, our iterative approach completed a single iteration in less than 24 hours.[1]

In the absence of a Twitter data set for retrieval, we selected the parameters $\lambda = 0.6$ (eq.1), $L = 0.005$ and $C = 0.95$ in a mate-finding task on Wikipedia data. The n-best list size for $P_{nbest}(t|q)$ was 1000. All SMT models included a 5-gram language model built from the English side of the NIST data plus the English side of the Twitter corpus D_{trg}. Word alignments were created using GIZA++ (Och and Ney, 2003). Rule extraction and parameter tuning (MERT) was carried out with `cdec`, using standard features. We ran MERT 5 times per iteration, carrying over the weights which achieved median performance on the development set to the next iteration.

Table 1 reports median BLEU scores on test of our standard adaptation baseline, the full-scale retrieval approach and the best result from our task alternation systems. Approximate randomization tests (Noreen, 1989; Riezler and Maxwell, 2005) showed that improvements of full-scale retrieval and task alternation over the baseline were statistically significant. Differences between full-scale retrieval and task alternation were not significant.[2]

Figure 1 illustrates the impact of θ, which controls the importance of the previous model compared to the current one, on median BLEU (a) and change of S_{in} (b) over iterations. For all θ, few iterations suffice to reach or surpass full-scale retrieval performance. Yet, no run achieved good performance after one iteration, showing that the transductive setup must be combined with task alternation to be effective. While we see fluctuations in BLEU for all θ-values, $\theta = 0.1$ achieves high scores faster and more consistently, pointing towards selecting a bolder updating strategy. This is also supported by plot (b), which indicates that choosing $\theta = 0.1$ leads to faster stabilization in the pairs added per iteration (S_{in}). We used this stabilization as a stopping criterion.

5 Conclusion

We presented a method that makes translation-based CLIR feasible for mining parallel sentences from large amounts of comparable data. The key of our approach is a translation-based high-quality retrieval model which gradually adapts to the target domain by iteratively re-training the underlying SMT model on a few thousand parallel sentences retrieved in the step before. The number of new pairs added per iteration stabilizes to a few thousand after 7 iterations, yielding an SMT model that improves 0.35 BLEU points over a model trained on millions of retrieved pairs.

[1] Retrieval was done in 4 batches on a Hadoop cluster using 190 mappers at once.

[2] Note that our full-scale results are not directly comparable to those of Jehl et al. (2012) since our setup uses less than one fifth of the NIST data, a different decoder, a new CLIR approach, and a different development and test split.

References

Sadaf Abdul-Rauf and Holger Schwenk. 2009. On the use of comparable corpora to improve SMT performance. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL'09)*, Athens, Greece.

Steven Abney. 2008. *Semisupervised Learning for Computational Linguistics*. Chapman and Hall.

Nicola Bertoldi and Marcello Federico. 2009. Domain adaptation for statistical machine translation with monolingual resources. In *Proceedings of the 4th EACL Workshop on Statistical Machine Translation (WMT'09)*, Athens, Greece.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2).

David Chiang. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2).

Chris Dyer, Adam Lopez, Juri Ganitkevitch, Jonathan Weese, Ferhan Ture, Phil Blunsom, Hendra Setiawan, Vladimir Eidelman, and Philip Resnik. 2010. cdec: A decoder, alignment, and learning framework for finite-state and context-free translation models. In *Proceedings of the ACL 2010 System Demonstrations (ACL'10)*, Uppsala, Sweden.

Laura Jehl, Felix Hieber, and Stefan Riezler. 2012. Twitter translation using translation-based cross-lingual retrieval. In *Proceedings of the Seventh Workshop on Statistical Machine Translation (WMT'12)*, Montreal, Quebec, Canada.

Philipp Koehn and Josh Schroeder. 2007. Experiments in domain adaptation for statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, Prague, Czech Republic.

Chin-Yew Lin and Franz Josef Och. 2004. Orange: a method for evaluating automatic evaluation metrics for machine translation. In *Proceedings the 20th International Conference on Computational Linguistics (COLING'04)*.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, Demo Session (ACL'12)*, Jeju, Republic of Korea.

Dragos Stefan Munteanu and Daniel Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31(4).

Dragos Stefan Munteanu and Daniel Marcu. 2006. Extracting parallel sub-sentential fragments from non-parallel corpora. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics (COLING-ACL'06)*, Sydney, Australia.

Eric W. Noreen. 1989. *Computer Intensive Methods for Testing Hypotheses. An Introduction*. Wiley, New York.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1).

Stefan Riezler and John Maxwell. 2005. On some pitfalls in automatic evaluation and significance testing for MT. In *Proceedings of the ACL-05 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, Ann Arbor, MI.

Stephen E. Robertson, Steve Walker, and Micheline Hancock-Beaulieu. 1998. Okapi at TREC-7. In *Proceedings of the Seventh Text REtrieval Conference (TREC-7)*, Gaithersburg, MD.

Matthew Snover, Bonnie Dorr, and Richard Schwartz. 2008. Language and translation model adaptation using comparable corpora. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'08)*, Honolulu, Hawaii.

Ferhan Ture and Jimmy Lin. 2012. Why not grab a free lunch? mining large corpora for parallel sentences to improve translation modeling. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'12)*, Montreal, Canada.

Ferhan Ture, Jimmy Lin, and Douglas W. Oard. 2012. Combining statistical translation techniques for cross-language information retrieval. In *Proceedings of the International Conference on Computational Linguistics (COLING'12)*, Mumbai, India.

Ferhan Ture, Jimmy Lin, and Douglas W. Oard. 2012a. Looking inside the box: Context-sensitive translation for cross-language information retrieval. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'12)*, Portland, OR.

Jakob Uszkoreit, Jay M. Ponte, Ashok C. Popat, and Moshe Dubiner. 2010. Large scale parallel document mining for machine translation. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING'10)*, Beijing, China.

Jinxi Xu, Ralph Weischedel, and Chanh Nguyen. 2001. Evaluating a probabilistic model for cross-lingual information retrieval. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'01)*, New York, NY.