Machine Translation Phrase Table Induction from Monolingual Data

Abstract

We explore deriving a machine translation phrase table from only a seed dictionary and monolingual data, rather than using a word-aligned parallel corpus. Our motivation comes from the lack of large parallel corpora for most low resource languages and from the extensive body of research on bilingual lexicon induction. We describe in detail ways to prune the very large search space and report results on an end-to-end machine translation task for several languages.

1 Introduction

In the standard phrase-based statistical machine translation (SMT) pipeline, the table of phrase pairs used to translate, or the grammar, is derived from a word-aligned parallel corpus. Researchers have found that, in general, the larger the parallel corpus from which the grammar is derived, the better translation performance is. However, for nearly all of the world's languages, there is no corpus of text that has a parallel counterpart in any other language that is big enough to be used to derive a grammar and translate from the given language. Following the work of (Klementiev et al., 2011), we consider the translation phrase table itself a parameter of SMT, and we explore using a seed dictionary and large monolingual corpora to derive it.

In this paper we:

- Motivate the need for generating translation phrase tables from non-parallel sources
- Describe the challenges involved, in particular the large computational cost of comparing all source language n-grams with all target language n-grams
- Present several novel methods for pruning the phrase pair search space
- Propose an intrinsic metric for measuring the quality of the phrase table
- Report end-to-end translation performance for several language pairs

2 Related Work

Statistical machine translation (SMT) was first formulated by Brown et al. (1993) at IBM, who proposed a series of probabilistic models based on word-to-word correspondences. The best performing current methods, so-called *phrase-based* and *hierarchical* models (Och, 2002; Koehn et al., 2003; Chiang, 2005) build on the IBM models, treating multi-token phrases as building blocks for producing new translations. Their parameters are estimated from large quantities of sentence aligned translations. These parallel resources are only available for a small set of language pairs and are very expensive to compile in sufficient quantities. In this work, we estimate the parameters of the phrase-based formalism (reviewed in Section 3.1) from *monolingual* texts, which are more plentiful.

Our methods for estimating model parameters build on the long line of work in bilingual lexicon induction from monolingual corpora. Rapp (1995) was the first to propose using the context of a given word as a clue to its translation. The model populates a bilingual lexicon by using a small seed dictionary to project context vectors across two languages and score their similarity. Rapp (1999), and Fung and Yee (1998) build on the original idea proposing alternative similarity metrics. Schafer and Yarowsky (2002) exploited the idea that word translations tend to co-occur in time across languages. Klementiev and Roth (2006) used this temporal cue to train a phonetic similarity model for associating Named Entities across languages. Koehn and Knight (2002) used similarity in spelling as another kind of cue that a pair of words may be translations of one another.

In this work, we extend these ideas to estimate three independent kinds of metrics to score *phrasal* translation candidates. More recent work on lexical induction includes Haghighi et al. (2008) who made use of contextual and orthographic clues to learn a generative model from monolingual texts and a seed lexicon, and Mimno et al. (2009) who proposed a polylingual topic model and matched high probability words in each topic across languages. While effective for a small number of topics, it is unlikely to scale well for our purposes.

Alternative research focuses on mining parallel or comparable corpora from the web (Munteanu and Marcu, 2006; Smith et al., 2010; Uszkoreit et al., 2010).

3 Background

3.1 Monolingual estimation of phrase-based SMT parameters

Brief recap of accepted ACL '11 paper

3.2 Bilingual lexicon induction for SMT

There are inherent challenges in extending lexicon induction methods to statistical machine translation.

First, scaling these methods up substantially degrades their performance due to sparsity issues. High accuracy results reported in most of the previous literature on a small set of source language words (e.g. 100 nouns in Rapp (1995), or 1,000 most frequent words in (Koehn and Knight, 2002)) tend to be misleading when the goal is to induce large translation dictionaries. Figure **??** shows an experiment using English and Spanish Wikipedia articles and contextual similarity to rank candidate target words for 1,000 most frequent and 1,000 random source language tokens[1]. Unsurprisingly, frequent terms have a substantially better chance of being paired with a correct translation. These concerns are exacerbated when we move to multi-token phrases. As with phrase translation features estimated from parallel data, longer phrases are more sparse, making similarity scores less reliable than for single words. In Section **??**, we introduce additional similarity estimates less prone to sparsity issues. In Section **??** we show that monolingual features are very informative and reliable in our setting; even when added alongside the feature functions derived from parallel data, they account for over a BLEU point gain in performance.

Second, extending the bilingual lexicon induction methods to the induction of multi-word translation pairs greatly increases computational cost. Since each source-target pair must be scored, we would need to compute $|V_f| * |V_e|$ similarity scores, where $|V_f|$ and $|V_e|$ are the number of unique items in the source and target vocabularies. When we move from tokens to multi word phrases, $|V|$ grows to the number of unique *n*-grams. Exhaustively scoring all phrase pairs up to length three, for example, requires computing scores between the set of unique unigrams, bigrams and trigrams in the source and target languages. In Section 4.1, we propose methods to substantially reduce the number of comparisons, while keeping good translation candidates.

4 Monolingual Parameter Estimation

4.1 Extracting a phrase table without a bitext

We propose three methods for assembling a table of phrase pairs without using parallel data: (1) We extend past efforts in bilingual lexicon induction from monolingual corpora to the induction of multi-word phrases. (2) We build a table of phrase pair translations from an existing bilingual seed dictionary. (3) We identify *unigrams* in our development and test sets that are OOV in our existing bilingual dictionary and then use a bilingual lexicon induction framework to induce new translations for each. In Section 5.1 we perform end-to-end MT experiments to test the quality of these phrase tables both individually and combined.

Phrase table induction. As we argued in Section 3.2, we cannot directly extend translation lexicon induction techniques to induce phrasal translations. Because this search space is very large, we must aggressively prune it. Specifically, we compare phrases in the same frequency bands (Uszkoreit et al., 2010), require that the stemmed version of each phrase-pair contain at least one word-pair from a stemmed version of our bilingual dictionary[2], and

[1] These accuracy estimates are conservative, since only the exact translation matches against our seed dictionary were counted.

[2] Our stemmer truncates words to a max. of six characters.

Pruning filters	Phrase Pairs	Search Space	Findable Types	Findable Tokens
Unpruned	1.6 T	100%	100%	100%
Dict combs.	74 M	<.01%	38%	65%
Freq Filter	31 B	2%	75%	81%
Freq + Dict	566 M	<.04%	49%	55%

Table 1: This shows the tradeoff between pruning the phrase pair search space and the accuracy of the final set of phrase pairs. The findable types and tokens measures refer to the percent of phrase types and tokens used by Moses to decode a test set that are not pruned away.

prune all very low frequency target phrases.

To determine whether our pruning parameters are reasonable, we evaluate what percent of the items from a bilingually induced phrase table are retained. Rather than consider all items from the phrase table (since many of them are bad), we consider only those phrase-pairs that are used by the decoder to translate a test set.[3] We compare our filtered phrase tables to this set of phrase translation rules and attempt to maintain as many of them as possible, while pruning the set of phrase pairs down to manageable size.

Dictionary-based phrase table. We build a second phrase table by enumerating all combinations and permutations of up to three dictionary translation entries (Garera and Yarowsky, 2008). Our Spanish–English dictionary has over 300,000 entries and the resulting phrase table has a large amount of overlap with the decoder table, as shown in Table 1.

Induced lexical translations. We use a bilingual lexical induction framework that works by comparing context, time, and edit distance source and target word vectors to build a table of candidate translations for the approximately ten thousand word types in our development and test sets that do not have entries in our bilingual dictionary. In Section 5.1 we use this relatively small set of unigram translation pairs to supplement the above tables.

5 Experiments

We used two Spanish-English corpora to estimate monolingual features: (1) in most of our experiments, we use the Spanish-English parts of the Europarl version 5 corpus[4] (Koehn, 2005); however, we treat the two sides of the corpus as two *independent monolingual* corpora, and (2) Spanish and English texts (106 and 187 million tokens, respectively) annotated with publication dates we collected by crawling news sites[5]. These are genuinely independent (but used only in one experiment). Our development data was comprised of *news-test2008* and *news-syscomb2009*, and are test set was *news-test2009* (Callison-Burch et al., 2010).

We use the Moses SMT toolkit (Koehn et al., 2007) and a language model trained on the English side of Europarl. With the exception of maximum phrase length (set to 3 in our experiments), default values were used for all of the parameters.

5.1 Phrase table extraction

As alternatives to using the Moses phrase table, which is based on the word-aligned parallel corpus, we perform several experiments using phrase tables generated in other ways. As explained in Section 4.1, we generate phrase translation pairs in three ways: combining and permuting dictionary translations (henceforth, dictionary-based); using lexicon induction methods to identify candidate word translations for OOV unigrams (henceforth, lexically induced); and inducing multi-word translations from the bilingual text treated as monolingual corpora (henceforth, induced phrase pairs). We compute both lexical and phrasal monolingual feature scores for all three types of phrase translations. To score lexical monolingual features, we use the word alignments inherent to both the dictionary-based translations and the lexically induced translations. We use an aligner (DeNero and Klein, 2007) to generate word alignments for the induced phrase pairs. In all of the experiments, we used random reordering features.

We use the monolingual feature scores to prune the induced phrase pairs from over 500 million to 1.1 million and to prune the dictionary-based table from 74 million to about 500 thousand entries. We

[3] We use the Moses decoder's trace function to find the set of phrases it used in decoding our test set.

[4] Europarl filenames include dates, making temporal signatures easy to compute.

[5] We have been collecting similar resources for a number of low density languages, and we plan to make them available to the community.

Phrase Table	Monolinual Features	BLEU
Dict-B.	None	9.09
Dict-B.	Ph-Mono	11.52
Induc.	Ph-Mono	8.97
Dict-B.+Induc.	Ph+Lx-Mono	12.25
Dict-B.+Induc.+Lex-I	Ph+Lx-Mono	12.65

Table 2: Decoding with the dictionary-based phrase table (Dict-B.), induced phrase pairs (Induc.), and lexically induced OOVs (Lex-I) scored with lexical monolingual features (Lx-Mono) and phrasal monolingual features (Ph-Mono). In all of the experiments shown in this table, we used random reordering features.

include the top five lexically induced translations for each OOV unigram.

BLEU scores resulting from using different combinations of phrase pair sets and features are shown in Table 2. In all experiments using phrase pairs from multiple sources, we add a feature to indicate whether a given phrase pair came from the dictionary-based or lexically induced tables or from the large table of induced phrase pairs. Adding monolingual features to the dictionary-based table improves the BLEU score to 11.52, and combining the dictionary-based table with the induced phrase pairs further improves the BLEU score to 12.25. Finally, combining all three alternative phrase pair sources results in a surprisingly competitive BLEU score of 12.65.

Although our alternative phrase pair sources do not reach the performance level of the Moses phrase table, they are promising because they show that it is possible to perform end-to-end MT without a sentence-aligned parallel corpus.

6 Conclusions and Future Work

References

Peter Brown, Stephen Della Pietra, Vincent Della Pietra, and Robert Mercer. 1993. The mathematics of machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311, June.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki, and Omar Zaidan. 2010. Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 17–53, Uppsala, Sweden, July. Association for Computational Linguistics.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

John DeNero and Dan Klein. 2007. Tailoring word alignments to syntactic machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL-2007)*, Prague, Czech Republic.

Pascale Fung and Lo Yuen Yee. 1998. An IR approach for translating new words from nonparallel, comparable texts. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (ACL-CoLing-1998)*, pages 414–420.

Nikesh Garera and David Yarowsky. 2008. Translating compounds by learning component gloss translation models via multiple languages. In *Proceedings of the International Joint Conference on Natural Language Processing*, pages 403–410.

Aria Haghighi, Percy Liang, Taylor Berg-Kirkpatrick, and Dan Klein. 2008. Learning bilingual lexicons from monolingual corpora. In *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Alexandre Klementiev and Dan Roth. 2006. Weakly supervised named entity transliteration and discovery from multilingual comparable corpora. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (ACL-CoLing-2006)*.

Alexandre Klementiev, Ann Irvine, and Chris Callison-Burch. 2011. Toward Statistical Machine Translation without Parallel Corpora. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, June.

Philipp Koehn and Kevin Knight. 2002. Learning a

translation lexicon from monolingual corpora. In *ACL Workshop on Unsupervised Lexical Acquisition*.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the Human Language Technology Conference of the North American chapter of the Association for Computational Linguistics (HLT/NAACL-2003)*, pages 48–54.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the ACL-2007 Demo and Poster Sessions*, Prague, Czech Republic.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proc. of the Machine Translation Summit*.

David Mimno, Hanna Wallach, Jason Naradowsky, David Smith, and Andrew McCallum. 2009. Polylingual topic models. In *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Dragos Stefan Munteanu and Daniel Marcu. 2006. Extracting parallel sub-sentential fragments from non-parallel corpora. In *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 81–88.

Franz Joseph Och. 2002. *Statistical Machine Translation: From Single-Word Models to Alignment Templates.* Ph.D. thesis, RWTH Aachen.

Reinhard Rapp. 1995. Identifying word translations in non-parallel texts. In *33rd Annual Meeting of the Association for Computational Linguistics (ACL-1995)*, pages 320–322.

Reinhard Rapp. 1999. Automatic identification of word translations from unrelated English and German corpora. In *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 519–526.

Charles Schafer and David Yarowsky. 2002. Inducing translation lexicons via diverse similarity measures and bridge languages. In *Proc. of the Annual Conference on Computational Natural Language Learning (CoNLL)*, pages 146–152.

Jason R. Smith, Chris Quirk, and Kristina Toutanova. 2010. Extracting parallel sentences from comparable corpora using document level alignment. In *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 403–411.

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