Toward Statistical Machine Translation without Parallel Corpora

Abstract

The parameters of statistical translation models of are typically estimated from bilingual parallel corpora. In this paper we explore the idea of estimating the parameters of a phrase-based statistical machine translation system from monolingual corpora. Existing research on inducing bilingual dictionaries from monolingual texts has largely focused on learning the translations of individual, high frequency words. We extend it to estimate all the parameters of phrase-based translation: phrasal translation pairs, and their translation and re-ordering probabilities. We begin with a fixed phrase-table and perform lesion experiments that show how much translation performance decreases when model parameters are removed, and how much of that loss can be restored when monolingually-estimated equivalents are added. We then analyze challenges of inducing the phrase table and ...

1 Introduction

Current statistical machine translation (SMT) methods (e.g. (Koehn et al., 2003; Chiang, 2005)) crucially rely on vast amounts of sentence aligned translations in order to achieve state of the art performance. These resources are only available for very few language pairs because producing them in sufficient quantities is an expensive and time consuming endeavor. Moreover, the SMT system performance tends to drop if test data comes from a different domain then the parallel data used in training. One general idea to deal with data sparsity is to attempt to collect (more) parallel data automatically (e.g. (Munteanu and Marcu, 2006; Smith et al., 2010; Uszkoreit et al., 2010)). However, some assumptions are typically made about the comparable bilingual corpus (such as document level alignment) used for mining near parallel text fragments.

Need a good MT adaptation reference

Need a better "however" sentence.

In this work, we approach the problem from an entirely different perspective: we use monolingual resources directly to induce an end-to-end statistical machine translation system. In particular, we extend a long line of work on inducing translation lexicons (e.g. (Rapp, 1995; Fung and Yee, 1998; Koehn and Knight, 2000; Klementiev and Roth, 2006; Haghighi et al., 2008; Mimno et al., 2009)) to induce translation features, and propose a novel algorithm for estimating reordering features for the phrase-based machine translation framework (Koehn et al., 2003). Much of the prior work on lexicon induction is motivated by the resource constrained SMT, however, to the best of our knowledge, this work is the first attempt to extend and apply these ideas in an end-to-end machine translation pipeline.

In this paper we:

- Analyze the challenge of using bilingual lexicon induction for statistical machine translation (performance on low frequency items, moving from words to phrases, and n^2 comparisons).
- Extend bilingual lexicon induction to phrasal translations, and scale it to extract translations for 30,000 phrases (which naively require tens of billions of phrase comparisons).
- Perform a set of lesion experiments where all feature functions are dropped from from a

phrase table, and then replaced with monolingually estimated equivalents.

- Report end-to-end translation quality with a fixed phrase-table with monolingually estimated parameters and for a fully monolingually induced system.

2 Related Work

3 Background

3.1 Parameters of phrase-based SMT

Review phrase-based setup a-la (Koehn et al., 2003).

- Log linear formulation:

$$p(\mathbf{e}|\mathbf{f}) \propto \exp \sum_{i=1}^{n} \lambda_i h_i(\mathbf{e}, \mathbf{f})$$
$$\hat{e} = \arg\max_{e} p(\mathbf{e}|\mathbf{f})$$

- *Phrase extraction*. Size of the phrase table and maximum phrase length: show our plot of performance vs. max phrase length. Mention what we choose max phrase length 3 for our experiments.

- *Phrase features*. Phrase translation probability and lexical translation probability.

- *Lexicalized reordering features*. See Figure 1.

- *Other features*. Language model, penalties.

Figure 1: switch rows and columns to be consistent with Figure 3?

In this work, we will use the same general mathematical formulation, but propose alternative features derived directly from monolingual data.

3.2 Lexicon Induction

Let us now briefly review relevant prior work on lexicon induction. In Section 4 we will build on this this work specifically to propose alternatives to some of the types of features we outlined in Section 3.1.

TODO: Alex: Fill in.

Most of the previous work evaluates results on a small set of hand selected words (e.g. 100 nouns in (Rapp, 1995)). However, if the objective is to induce large translation tables, as it is in our work, the reported results can be misleading. TODO: Describe Figure 2.

Diss prev work, but make sure that it comes across that these features are informative.

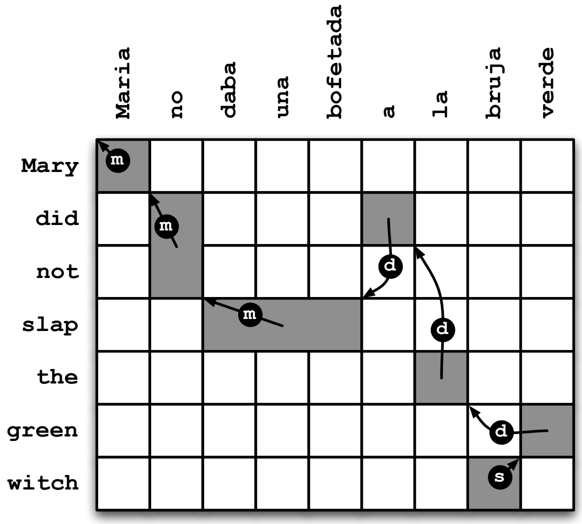

Figure 1: Example alignment along with three kinds of orientations: monotone (m), swap (s), and discontinuous (d).

4 Reducing the Parallel Data requirement / Estimating Parameters from Monolingual Data

4.1 Phrase extraction

Building on the many successful efforts in bilingual lexicon induction from monolingual corpora, we compile a table of *phrasal* translations from monolingual corpora. Current methods for inducing a bilingual lexicon, or, equivalently, a unigram phrase translation table, are computationally expensive as each source and target word pair must be scored. That is, the models must compute $|V_s| * |V_t|$ similarity scores, where $|V_s|$ and $|V_t|$ are the sizes of the source and target vocabularies, respectively. As we search for longer phrase pairs, this search space increases exponentially with the size of the n-grams. Exhaustively scoring all phrase pairs up to length three requires $|V_s|^3 * |V_t|^3$ score computations. We prune the phrase pair search space using methods from Information Retrieval, like that used in Uszkoreit et al. (2010).

First, we compose a set of all n-grams up to length three in the source side monolingual corpus and in the target side monolingual corpus. We store each source side n-gram's frequency along with the

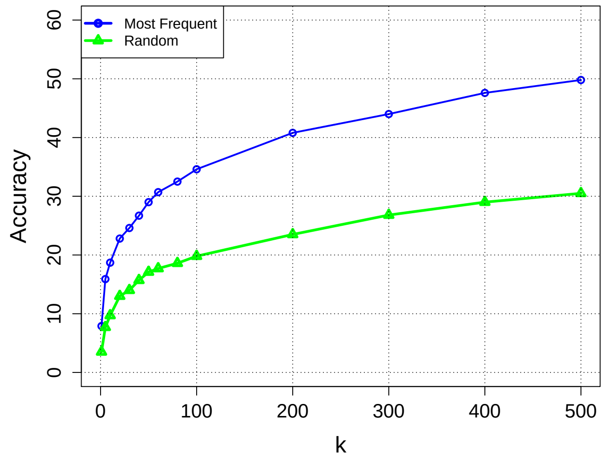

Figure 2: Accuracy on wikipedia for most frequent and random 1000 source words.

phrase. We also store each target side n-gram in a frequency inverted index. We index target side phrases by frequencies within a band of their actual observed frequency in the monolingual corpus. Additionally, for the target side n-grams, we look up each word in the phrase in our bilingual dictionary, and we store a set of source side words that translate into any word in the phrase.

discuss dictionary(ies)

These sets of observed phrases, frequencies, and the bilingual dictionary together allow us to effectively filter our search space using combinations of inverted indices. Given a source side test set, we collect n-grams up to length three (in our Urdu test set, there are 36,423 phrases), as we did for the monolingual corpora. We then look up each source side phrase's monolingual corpus frequency and consult the target side inverted index to find a list of target side phrases that occur in the same frequency band. Of those target side phrases, we keep only the ones with at least one word translating into a word in the source side phrase.

discuss dictionary parameters

This method of pruning the phrase pair search space involves two manually tuned parameters. In order to evaluate the filtered phrase tables and tune the parameters, we used the Moses decoder's trace function to find the set of phrases used in decoding an Urdu test set. 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.

data details?

Our baseline phrase table is generated using a bilingual dictionary. For each Urdu test set phrase up to length three, we generated English phrases from all combinations of dictionary translations and all possible reorderings. For the baseline and our pruning methods, the number of filtered phrase pairs and the percent of phrases used by the Moses decoder not pruned away are given in Table 4.1.

Second round of pruning: after monolingual feature extraction, before re-ordering estimation. Needs to be discussed after explanation of those methods?

4.2 Phrase scoring

In place of phrase translation probabilities estimated from bilingual alignments, we propose to compute similarity scores computed (almost) solely from monolingual resources.

Contextual similarity. We extend the vector space approach of (Rapp, 1999) to compute similarity between *phrases* in source and target language. More formally, assume that $(f_1, f_2, \ldots f_N)$ and $(e_1, e_2, \ldots e_M)$ are (arbitrarily indexed) source and target vocabularies, respectively. A source phrase f (target phrase e) is represented with an N (M) dimensional vector. Only the components corresponding to words that appear in the context of f (e) in data take on non-zero values, which typically measure how "unique" a word is to the context in the dataset. Next, f's contextual vector is projected by mapping each component to a component in the target space corresponding to its translation (taken from a small seed dictionary), but retaining the source component value. Finally, the pair (f, e) is scored by computing similarity between the (projected) source and target vectors. Various means of computing the component values and vector similarity measures have been proposed in literature (e.g. (Rapp, 1999; Fung and Yee, 1998)). While the quality of the resulting induced lexicon depends on the data, we found the following to work best in our experiments. We compute the value of the k-th component of f's contextual vector as follows:

$$w_k^{(i)} = n_{i,k} \times (log(n/n_i) + 1)$$

Pruning filters	Phrase Pairs	Percent of total search space	Findable types	Findable tokens
Unpruned phrase table	37,322,465,985	100%	100%	100%
Baseline phrase pairs	29245036	0.08%	15.37	25.07
Frequency-based pruning	4,450,429,494	11.92%	85.21%	87.48%
Frequency and Dictionary pruning	1,436,823,109	3.85%	57.79%	58.76%

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.

where $n_{i,k}$ and f_k are the number of times f_k appears in the context of f_i and in the entire corpus, and n is the maximum number of occurrences of any word in the data. Intuitively, the more frequently f_k appears with f_i and the less common it is in the corpus in general, the higher its component value. Similarity between two resulting vectors is measured as a cosine of the angle between them.

Temporal similarity. Online content is often published along with temporal information: news feeds, for example, are comprised of news stories annotated with date and time of publication. The feeds are specialized for the target geographical locations and vary in content across languages. Still, many events are deemed relevant to multiple audiences and the news stories related to them appear in several languages, although rarely as direct translations of one another. Phrases associated with these events will appear with increased frequency in multiple languages around the dates when these events are reported. Such weak synchronicity provides a cue about the relatedness of phrases across the two languages. In order to score a pair of phrases across languages, we can compute the similarity of their temporal signatures. To generate a time sequence for a given word, we first sort the set of (time-stamped) documents of our corpus into a sequence of equally sized temporal bins. We then count the number of occurrences of a phrase in each bin. Changing the size of the bin or computing counts in a sliding window instead can recover some accuracy if the temporal alignment between two languages in our dataset is poor (Klementiev and Roth, 2006). Finally, we normalize the sequence and use either the cosine measure to score similarity.

Orthographic / phonetic similarity. Etymologically related words often retain similar spelling across languages with the same writing system, and the edit distance can be used to measure their orthographic similarity. We extend this idea to phrases by using word alignments within a phrase pair (see Section 4.1): we score pairs of aligned words and normalize by their average length.

Make sure it is correct.

We can further extend this idea to language pairs not sharing the same writing system, since many cognates and transliterated words are phonetically similar. Following (Virga and Khudanpur, 2003; Irvine et al., 2010), we treat transliteration as a monotone character translation task and use a generative model to propose a transliteration of tokens in a source phrase. Once the source words are mapped to the target writing system, the phrase similarity is computed as before.

Argue that enough training data is easy to get

Depending on the available monolingual data (and its associated metadata), various other similarity scores can be computed and added to the list (see, e.g. (Schafer and Yarowsky, 2002)).

4.3 Reordering

In the phrase-based SMT pipeline we reviewed in Section 3.1, phrase pair orientation statistics were collected from induced word alignments. We keep a similar lexicalized reordering model formulation, but infer its parameters from monolingual data instead. The orientation information for a phrase pair is collected from source and target sentences containing the two phrases as well as other hypothesized translation pairs. Given a phrase pair (f, e), the idea is to estimate the probability that other phrases preceding f will precede, follow, or become discontinuous with e in target sentences when translated. Contextual phrases used to estimate these orientation features for (f, e) will be all entries in the phrase

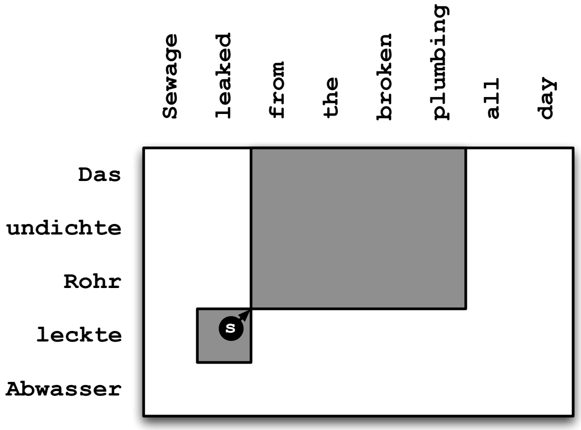

Figure 3: Collecting phrase ordering statistics for a de-en phrase pair (*“leckte”*, *“leaked”*). The longest preceding phrase *“Das undichte Rohr”* in source, has a phrase table translation *“from the broken plumbing”* appearing after the target phrase.

table. Consider a simple example on Figure 3: the phrase pair is (f = *“leckte”*, e = *“leaked”*), and a given pair of unaligned sentences also contains a phrase table entry (f_b = *“Das undichte Rohr”*, *“from the broken plumbing”*). In this example, the phrase f_b preceding f in the source sentence swaps order with e in the target. When collected a over large unaligned bilingual corpora, we expect the swap, monotone, and discontinuous counts to provide good estimates for the orientation features. Note that multiple phrases may immediately precede f and appear in the phrase table; however, we only use the longest one to collect reordering counts.

Explain why?

The algorithm on Figure 4 estimates monotone, swap, and discontinuous orientation features (p_m, p_s, p_d) for a phrase pair (f, e). It begins by calling `CollectOccurs()` to collect counts for the longest phrase table phrases preceding f in source monolingual data, and preceding, following, and discontinuous with e in target data. It then proceeds to

Describe the algorithm (Figure 4) for estimating orientation probabilities[1]. Talk about the issue of too much weight on out-of-order orientation. Talk about where w comes from. Mention that it is not as expensive as it looks. The intuition is simple - see if translated phrases tend to keep the order.

```
**Input**: Source and target phrases $f$ and $e$,
source and target monolingual corpora $C_f$ and $C_e$,
phrase table pairs $T = \{(f^{(i)}, e^{(i)})\}_{i=1}^{N}$.
**Output**: Orientation features $(p_m, p_s, p_d)$.

$S_f \leftarrow$ sentences containing $f$ in $C_f$;
$S_e \leftarrow$ sentences containing $e$ in $C_e$;
$(B_f, -, -) \leftarrow$ CollectOccurs$(f, \cup_{i=1}^{N} f^{(i)}, S_f)$;
$(B_e, A_e, D_e) \leftarrow$ CollectOccurs$(e, \cup_{i=1}^{N} e^{(i)}, S_e)$;
$c_m = c_s = c_d = 0$;
**foreach** $f_b \in B_f$ **do**
    **foreach** *translation* $e^*$ *of* $f_b$ *in* T **do**
        $c_m = c_m + \#_{B_e}(e^*)$;
        $c_s = c_s + \#_{A_e}(e^*)$;
        $c_d = c_d + \#_{D_e}(e^*)$;
$c \leftarrow c_m + c_s + c_d$;
**return** $(\frac{c_m}{c}, \frac{c_s}{c}, \frac{c_d}{c})$

CollectOccurs(*r*, *R*, *S*)
    $B \leftarrow ()$; $A \leftarrow ()$; $D \leftarrow ()$;
    **foreach** *sentence* $s \in S$ **do**
        **foreach** *occurrence of phrase* $r$ *in* $s$ **do**
            $B \leftarrow B$ + (longest preceding and in $R$);
            $A \leftarrow A$ + (longest following and in $R$);
            $D \leftarrow D$ + (longest discontinuous and in $R$);
    **return** *(B, A, D)*;
```

Figure 4: Estimating reordering probabilities from monolingual data.

5 Experiments

5.1 Data

Describe data we use in the experiments: Europarl (Koehn, 2005), Gigaword, our own crawls[2].

5.2 Single language

1. *Phrase features.* (a) Augment phrase scores with mono features. If we see better performance, reduce the amount of parallel data until it matches the performance of the original system. Make the tradeoff argument. (b) (**lesion experiments**) See how well we do with mono features alone.

2. *Orientation features.* Use mono orientation features.

3. *Induce phrase table.*

4. *Put everything together.* Run the entire pipeline.

[1] $\#_S(x)$ returns the count of object x in multiset S.

[2] Promise to distribute after publication.

5.3 Big experiment

Now, run the entire pipeline on a handful of languages extracting monolingual features from the Gigaword and our crawls.

6 Discussion

7 Conclusions and Future Work

First to make use of plentiful monolingual data to reduce the dependence on expensive parallel data. In particular:

- Showed that augmenting standard pipeline with monolingual features helps.
- Demonstrated that monolingual features are informative enough on their own for a competitive system.
- Proposed an algorithm for estimating orientation probabilities from monolingual data alone.
- Build complete systems for X low-resource languages.

References

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)*.

Pascale Fung and Lo Yuen Yee. 1998. An ir approach for translating new words from nonparallel, comparable texts. In *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 414–420.

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)*.

Ann Irvine, Chris Callison-Burch, and Alexandre Klementiev. 2010. Transliterating from all languages. In *Proc. of the Conference of the Association for Machine Translation in the Americas (AMTA)*.

Alexandre Klementiev and Dan Roth. 2006. Weakly supervised named entity transliteration and discovery from multilingual comparable corpora. In *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Philipp Koehn and Kevin Knight. 2000. Estimating word translation probabilities from unrelated monolingual corpora using the EM algorithm. In *Proc. of the National Conference on Artificial Intelligence (AAAI)*.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proc. of the Annual Meeting of the North American Association of Computational Linguistics (NAACL)*, pages 48–54.

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

Adam Lopez. 2008. Statistical machine translation. *ACM Computing Surveys*, 40(3):1–49, August.

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 nonparallel corpora. In *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 81–88.

Reinhard Rapp. 1995. Identifying word translations in non-parallel texts. In *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 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)*.

Paola Virga and Sanjeev Khudanpur. 2003. Transliteration of proper names in cross-lingual information retrieval. In *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*.