Supervised Phrase Table Triangulation with Neural Word Embeddings for Low-Resource Languages

Tomer Levinboim and **David Chiang**
Department of Computer Science and Engineering
University of Notre Dame
{levinboim.1,dchiang}@nd.edu

Abstract

In this paper, we develop a supervised learning technique that improves noisy phrase translation scores obtained by phrase table triangulation. In particular, we extract word translation distributions from small amounts of source-target bilingual data (a dictionary or a parallel corpus) with which we learn to assign better scores to translation candidates obtained by triangulation. Our method is able to gain improvement in translation quality on two tasks: (1) On Malagasy-to-French translation via English, we use only 1k dictionary entries to gain +0.5 B over triangulation. (2) On Spanish-to-French via English we use only 4k sentence pairs to gain +0.7 B over triangulation interpolated with a phrase table extracted from the same 4k sentence pairs.

1 Introduction

Phrase-based statistical machine translation systems require considerable amounts of source-target parallel data to produce good quality translation. However, large amounts of parallel data are available for only a fraction of language pairs, and mostly when one of the languages is English.

Phrase table triangulation (Utiyama and Isahara, 2007; Cohn and Lapata, 2007; Wu and Wang, 2007) is a method for generating source-target phrase tables without having access to any source-target parallel data. The intuition behind triangulation (and pivoting techniques in general) is the transitivity of translation: if a source language phrase **s** translates to a pivot language phrase **p** which in turn translates to a target language phrase **t**, then **s** should likely translate to **t**. Following this intuition, a triangulated source-target phrase table $\hat{T}$ can be composed from a source-pivot and pivot-target phrase table (§2).

However, the resulting triangulated phrase table $\hat{T}$ contains many spurious phrase pairs and noisy probability estimates. Therefore, early triangulation work (Wu and Wang, 2007) already realistically assumed access to a limited source-target parallel data from which a relatively high-quality source-target phrase table T can be directly estimated. The two phrase tables were then combined, resulting in a higher quality phrase table that proposes translations for many source phrases not found in T. Wu and Wang (2007) report that interpolation of the two phrase tables T and $\hat{T}$ leads to higher quality translations. However, the triangulated phrase table $\hat{T}$ is obtained without using the source-target bilingual data, which suggests that the source-target data is not used as fully as it could be.

In this paper, we develop a supervised learning algorithm that corrects triangulated word translation probabilities by relying on word translation distributions w^{sup} derived from the limited source-target data. In particular, we represent source and target words using word embeddings (Mikolov et al., 2013) and learn a transformation between the two embedding spaces in order to approximate w^{sup}, thus down-weighting incorrect translation candidates proposed by triangulation (§3). By representing words as embeddings, our model can generalize the information contained in the source-target data (as encoded in the distributions w^{sup}) to a much larger vocabulary, and can assign lexical-weighting probabilities to most of the phrase pairs in $\hat{T}$.

Fixing English as the pivot language (the most realistic pivot language choice), on a low-resource Spanish-to-French translation task our model gains +0.7 B on top of standard phrase table interpolation. On Malagasy-to-French translation, our model gains +0.5 B on top of triangulation when using only 1k Malagasy-French dictionary entries (§4).

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1079–1083,
Lisbon, Portugal, 17-21 September 2015.

2 Preliminaries

Let s, p, t denote words and $\mathbf{s}, \mathbf{p}, \mathbf{t}$ denote phrases in the source, pivot, and target languages, respectively. Also, let T denote a phrase table estimated over a parallel corpus and $\hat{T}$ denote a triangulated phrase table. We use similar notation for their respective phrase translation features ϕ, lexical-weighting features *lex*, and the word translation probabilities w.

2.1 Triangulation (weak baseline)

In phrase table triangulation, a source-target phrase table T_{st} is constructed by combining a source-pivot and pivot-target phrase table $T_{\text{sp}}, T_{\text{pt}}$, each estimated on its respective parallel data. For each resulting phrase pair $(\mathbf{s}, \mathbf{t})$, we can also compute an alignment $\hat{\mathbf{a}}$ as the most frequent alignment obtained by combining source-pivot and pivot-target alignments $\mathbf{a_{sp}}$ and $\mathbf{a_{pt}}$ across all pivot phrases $\mathbf{p}$ as follows: $\{(s,t) \mid \exists p : (s,p) \in \mathbf{a_{sp}} \wedge (p,t) \in \mathbf{a_{pt}}\}$.

The triangulated source-to-target lexical weights, denoted $\widehat{lex}_{\text{st}}$, are approximated in two steps: First, word translation scores $\hat{w}_{\text{st}}$ are approximated by marginalizing over the pivot words:

$$\hat{w}_{\text{st}}(t \mid s) = \sum_p w_{\text{sp}}(p \mid s) \cdot w_{\text{pt}}(t \mid p). \quad (1)$$

Next, given a (triangulated) phrase pair $(\mathbf{s}, \mathbf{t})$ with alignment $\hat{\mathbf{a}}$, let $\hat{\mathbf{a}}_{s,:} = \{t \mid (s,t) \in \hat{\mathbf{a}}\}$; the lexical-weighting probability is (Koehn et al., 2003):

$$\widehat{lex}_{\text{st}}(\mathbf{t} \mid \mathbf{s}, \hat{\mathbf{a}}) = \prod_{s \in \mathbf{s}} \frac{1}{|\hat{\mathbf{a}}_{s,:}|} \sum_{t \in \hat{\mathbf{a}}_{s,:}} \hat{w}_{\text{st}}(t \mid s). \quad (2)$$

The triangulated phrase translation scores, denoted $\hat{\phi}_{\text{st}}$, are computed by analogy with Eq. 1.

We also compute these scores in the reverse direction by swapping the source and target languages.

2.2 Interpolation (strong baseline)

Given access to source-target data, an ordinary source-target phrase table T_{st} can be estimated directly. Wu and Wang (2007) suggest interpolating phrase pairs entries that occur in both tables:

$$T_{\text{interp}} = \alpha T_{\text{st}} + (1 - \alpha)\hat{T}_{\text{st}}. \quad (3)$$

Phrase pairs appearing in only one phrase table are added as-is. We refer to the resulting table as the interpolated phrase table.

3 Supervised Word Translations

While interpolation (Eq. 3) may help correct some of the noisy triangulated scores, its effect is limited to phrase pairs appearing in both phrase tables. Here, we suggest a discriminative supervised learning method that can affect all phrase pairs.

Our idea is to regard word translation distributions derived from source-target bilingual data (through word alignments or dictionary entries) as the correct translation distributions, and use them to learn discriminately: correct target words should become likely translations, and incorrect ones should be down-weighted. To generalize beyond the vocabulary of the source-target data, we appeal to word embeddings.

We present our formulation in the source-to-target direction. The target-to-source direction is obtained simply by swapping the source and target languages.

3.1 Model

Let c^{sup}_{st} denote the number of times source word s was aligned to target word t (in word alignment, or in the dictionary). We define the word translation distributions $w^{\text{sup}}(t \mid s) = c^{\text{sup}}_{st}/c^{\text{sup}}_{s}$, where $c^{\text{sup}}_{s} = \sum_t c^{\text{sup}}_{st}$. Furthermore, let $q(t \mid s)$ denote the word translation probabilities we wish to learn and consider maximizing the log-likelihood function:

$$\arg\max_q L(q) = \arg\max_q \sum_{(s,t)} c^{\text{sup}}_{st} \log q(t \mid s).$$

Clearly, the solution $q(\cdot \mid s) := w^{\text{sup}}(\cdot \mid s)$ maximizes L. However, we would like a solution that generalizes to source words s beyond those observed in the source-target corpus – in particular, those source words that appear in the triangulated phrase table $\hat{T}$, but not in T.

In order to generalize, we abstract from words to vector representations of words. Specifically, we constrain q to the following parameterization:

$$q(t \mid s) = \frac{1}{Z_s} \exp\left(v_s^T A v_t + f_{st}^T h\right)$$
$$Z_s = \sum_{t \in \mathcal{T}(s)} \exp\left(v_s^T A v_t + f_{st}^T h\right).$$

Here, the vectors v_s and v_t represent monolingual features and the vector f_{st} represents bilingual features. The parameters A and h are to be learned.

In this work, we use monolingual word embeddings for v_s and v_t, and set the vector f_{st} to contain only the value of the triangulated score, such

that $f_{st} := \hat{w}_{\text{st}}$. Therefore, the matrix A is a linear transformation between the source and target embedding spaces, and h (now a scalar) quantifies how the triangulated scores $\hat{w}$ are to be trusted.

In the normalization factor Z_s, we let t range only over possible translations of s suggested by either w^{sup} or the triangulated word probabilities. That is:

$$\mathcal{T}(s) = \{t \mid w^{\text{sup}}(t \mid s) > 0 \ \vee \ \hat{w}(t \mid s) > 0\}.$$

This restriction makes efficient computation possible, as otherwise the normalization term would have to be computed over the entire target vocabulary.

Under this parameterization, our goal is to solve the following maximization problem:

$$\max_{A,h} L(A, h) = \max_{A,h} \sum_{s,t} c_{st}^{\text{sup}} \log q(t \mid s). \quad (4)$$

3.2 Optimization

The objective function in Eq. 4 is concave in both A and h. This is because after taking the log, we are left with a weighted sum of linear and concave (negative log-sum-exp) terms in A and h. We can therefore reach the global solution of the problem using gradient descent.

Taking derivatives, the gradient is

$$\frac{\partial L}{\partial A} = \sum_{s,t} m_{st} v_s v_t^{\text{T}} \qquad \frac{\partial L}{\partial h} = \sum_{s,t} m_{st} f_{st}$$

where the scalar $m_{st} = c_{st}^{\text{sup}} - c_s^{\text{sup}} q(t \mid s)$ for the current value of q.

For quick results, we limited the number of gradient steps to 200 and selected the iteration that minimized the total variation distance to w^{sup} over a held out dev set:

$$\sum_s \|q(\cdot \mid s) - w^{\text{sup}}(\cdot \mid s)\|_1. \quad (5)$$

We obtained better convergence rate by using a batch version of the effective and easy-to-implement Adagrad technique (Duchi et al., 2011). See Figure 1.

3.3 Re-estimating lexical weights

Having learned the model (A and h), we can now use $q(t \mid s)$ to estimate the lexical weights (Eq. 2) of any aligned phrase pairs $(\mathbf{s}, \mathbf{t}, \hat{\mathbf{a}})$, assuming it is composed of embeddable words.

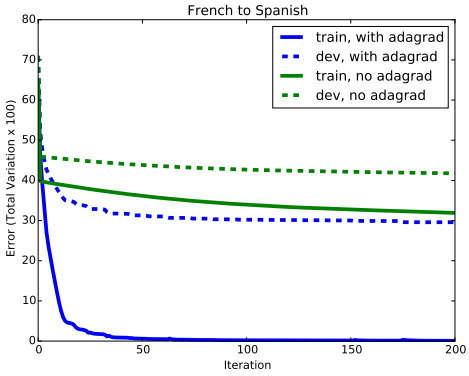

Figure 1: The (target-to-source) objective function per iteration. Applying batch Adagrad (blue) significantly accelerates convergence.

However, we found the supervised word translation scores q to be too sharp, sometimes assigning all probability mass to a single target word. We therefore interpolated q with the triangulated word translation scores $\hat{w}$:

$$q_\beta = \beta q + (1 - \beta)\hat{w}. \quad (6)$$

To integrate the lexical weights induced by q_β (Eq. 2), we simply appended them as new features in the phrase table in addition to the existing lexical weights. Following this, we can search for a β value that maximizes B on a tuning set.

3.4 Summary of method

In summary, to improve upon a triangulated or interpolated phrase table, we:

1. Learn word translation distributions q by supervision against distributions w^{sup} derived from the source-target bilingual data (§3.1).

2. Smooth the learned distributions q by interpolating with triangulated word translation scores $\hat{w}$ (§3.3).

3. Compute new lexical weights and append them to the phrase table (§3.3).

4 Experiments

To test our method, we conducted two low-resource translation experiments using the phrase-based MT system Moses (Koehn et al., 2007).

4.1 Data

Fixing the pivot language to English, we applied our method on two data scenarios:

1. **Spanish-to-French:** two related languages used to simulate a low-resource setting. The baseline is phrase table interpolation (Eq. 3).

2. **Malagasy-to-French:** two unrelated languages for which we have a small dictionary, but no parallel corpus (aside from tuning and testing data). The baseline is triangulation alone (there is no source-target model to interpolate with).

Table 1 lists some statistics of the bilingual data we used. European-language bitexts were extracted from Europarl (Koehn, 2005). For Malagasy-English, we used the Global Voices parallel data available online.[1] The Malagasy-French dictionary was extracted from online resources[2] and the small Malagasy-French tune/test sets were extracted[3] from Global Voices.

	lines of data		
language pair	train	tune	test
sp-fr	4k	1.5k	1.5k
mg-fr	1.1k	1.2k	1.2k
sp-en	50k	–	–
mg-en	100k	–	–
en-fr	50k	–	–

Table 1: Bilingual datasets. Legend: sp=Spanish, fr=French, en=English, mg=Malagasy.

Table 2 lists token statistics of the monolingual data used. We used word2vec[4] to generate French, Spanish and Malagasy word embeddings. The French and Spanish embeddings were (independently) estimated over their combined tokenized and lowercased Gigaword[5] and Leipzig news corpora.[6] The Malagasy embeddings were similarly estimated over data form Global Voices,[7] the Malagasy Wikipedia and the Malagasy Common Crawl.[8] In addition, we estimated a 5-gram French language model over the French monolingual data.

[1]http://www.ark.cs.cmu.edu/global-voices
[2]http://motmalgache.org/bins/homePage
[3]https://github.com/vchahun/gv-crawl
[4]https://radimrehurek.com/gensim/models/word2vec.html
[5]http://catalog.ldc.upenn.edu
[6]http://corpora.uni-leipzig.de/download.html
[7]http://www.isi.edu/~qdou/downloads.html
[8]https://commoncrawl.org/the-data/

language	words
French	1.5G
Spanish	1.4G
Malagasy	58M

Table 2: Size of monolingual corpus per language as measured in number of tokens.

4.2 Spanish-French Results

To produce w^{sup}, we aligned the small Spanish-French parallel corpus in both directions, and symmetrized using the intersection heuristic. This was done to obtain high precision alignments (the often-used *grow-diag-final-and* heuristic is optimized for phrase extraction, not precision).

We used the skip-gram model to estimate the Spanish and French word embeddings and set the dimension to $d = 200$ and context window to $w = 5$ (default). Subsequently, to run our method, we filtered out source and target words that either did not appear in the triangulation, or, did not have an embedding. We took words that appeared more than 10 times in the parallel corpus for the training set (~690 words), and between 5–9 times for the held out dev set (~530 words). This was done in both source-target and target-source directions.

In Table 3 we show that the distributions learned by our method are much better approximations of w^{sup} compared to those obtained by triangulation.

Method	source→target	target→source
triangulation	71.6%	72.0%
our scores	30.2%	33.8%

Table 3: Average total variation distance (Eq. 5) to the dev set portion of w^{sup} (computed only over words whose translations in w^{sup} appear in the triangulation). Using word embeddings, our method is able to better generalize on the dev set.

We then examined the effect of appending our supervised lexical weights. We fixed the word level interpolation $\beta := 0.95$ (effectively assigning very little mass to triangulated word translations $\hat{w}$) and searched for $\alpha \in \{0.9, 0.8, 0.7, 0.6\}$ in Eq. 3 to maximize B on the tuning set.

Our MT results are reported in Table 4. While interpolation improves over triangulation alone by +0.8 B , our method adds another +0.7 B on top of interpolation, a statistically significant gain ($p < 0.01$) according to a bootstrap resampling significance test (Koehn, 2004).

Method	α	tune	test
source-target	–	26.8	25.3
triangulation	–	29.2	28.4
interpolation	0.7	30.2	29.2
interpolation+our scores	0.6	30.8	**29.9**

Table 4: Spanish-French B scores. Appending lexical weights obtained by supervision over a small source-target corpus significantly outperforms phrase table interpolation (Eq. 3) by +0.7 B .

4.3 Malagasy-French Results

For Malagasy-French, the w^{sup} distributions used for supervision were taken to be uniform distributions over the dictionary translations. For each training direction, we used a 70%/30% split of the dictionary to form the train and dev sets.

Having significantly less Malagasy monolingual data, we used $d = 100$ dimensional embeddings and a $w = 3$ context window to estimate both Malagasy and French words.

As before, we added our supervised lexical weights as new features in the phrase table. However, instead of fixing $\beta = 0.95$ as above, we searched for $\beta \in \{0.9, 0.8, 0.7, 0.6\}$ in Eq. 6 to maximize B on a small tune set. We report our results in Table 5. Using only a dictionary, we are able to improve over triangulation by +0.5 B , a statistically significant difference ($p < 0.01$).

Method	β	tune	test
triangulation	–	12.2	11.1
triangulation+our scores	0.6	12.4	**11.6**

Table 5: Malagasy-French B . Supervision with a dictionary significantly improves upon simple triangulation by +0.5 B .

5 Conclusion

In this paper, we argued that constructing a triangulated phrase table independently from even very limited source-target data (a small dictionary or parallel corpus) underutilizes that parallel data.

Following this argument, we designed a supervised learning algorithm that relies on word translation distributions derived from the parallel data as well as a distributed representation of words (embeddings). The latter enables our algorithm to assign translation probabilities to word pairs that do not appear in the source-target bilingual data.

We then used our model to generate new lexical weights for phrase pairs appearing in a triangulated or interpolated phrase table and demonstrated improvements in MT quality on two tasks. This is despite the fact that the distributions (w^{sup}) we fit our model to were estimated automatically, or even naïvely as uniform distributions.

Acknowledgements

The authors would like to thank Daniel Marcu and Kevin Knight for initial discussions and a supportive research environment at ISI, as well as the anonymous reviewers for their helpful comments. This research was supported in part by a Google Faculty Research Award to Chiang.

References

Trevor Cohn and Mirella Lapata. 2007. Machine translation by triangulation: Making effective use of multi-parallel corpora. In *Proc. ACL*, pages 728–735.

John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *J. Machine Learning Research*, 12:2121–2159, July.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proc. NAACL HLT*, 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, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proc. ACL, Interactive Poster and Demonstration Sessions*, pages 177–180.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proc. EMNLP*, pages 388–395.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proc. MT Summit*, pages 79–86.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proc. ICLR, Workshop Track*.

Masao Utiyama and Hitoshi Isahara. 2007. A comparison of pivot methods for phrase-based statistical machine translation. In *Proc. HLT-NAACL*, pages 484–491.

Hua Wu and Haifeng Wang. 2007. Pivot language approach for phrase-based statistical machine translation. In *Proc. ACL*, pages 856–863.