Multi-Task Word Alignment Triangulation 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

We present a multi-task learning approach that jointly trains three word alignment models over disjoint bitexts of three languages: source, target and pivot. Our approach builds upon model triangulation, following Wang et al., which approximates a source-target model by combining source-pivot and pivot-target models. We develop a MAP-EM algorithm that uses triangulation as a prior, and show how to extend it to a multi-task setting. On a low-resource Czech-English corpus, using French as the pivot, our multi-task learning approach more than doubles the gains in both F- and B scores compared to the interpolation approach of Wang et al. Further experiments reveal that the choice of pivot language does not significantly affect performance.

1 Introduction

Word alignment (Brown et al., 1993; Vogel et al., 1996) is a fundamental task in the machine translation (MT) pipeline. To train good word alignment models, we require access to a large parallel corpus. However, collection of parallel corpora has mostly focused on a small number of widely-spoken languages. As such, resources for almost any other pair are either limited or non-existent.

To improve word alignment and MT in a low-resource setting, we design a multitask learning approach that utilizes parallel data of a third language, called the *pivot* language (§3). Specifically, we derive an efficient and easy-to-implement MAP-EM-like algorithm that jointly trains source-target, source-pivot and pivot-target alignment models, each on its own bitext, such that each model benefits from observations made by the other two.

Our method subsumes the model interpolation approach of Wang et al. (2006), who independently train these three models and then interpolate the source-target model with an approximate source-target model, constructed by combining the source-pivot and pivot-target models.

Pretending that Czech-English is low-resource, we conduct word alignment and MT experiments (§4). With French as the pivot, our approach significantly outperforms the interpolation method of Wang et al. (2006) on both alignment F- and B scores. Somewhat surprisingly, we find that our approach is insensitive to the choice of pivot language.

2 Triangulation and Interpolation

Wang et al. (2006) focus on learning a word alignment model without a source-target corpus. To do so, they assume access to both source-pivot and pivot-target bitexts on which they independently train a source-pivot word alignment model $\Theta_{\mathtt{sp}}$ and a pivot-target model $\Theta_{\mathtt{pt}}$. They then combine the two models by marginalizing over the pivot language, resulting in an approximate source-target model $\Theta_{\widetilde{\mathtt{st}}}$. This combination process is referred to as *triangulation* (see §5).

In particular, they construct the triangulated source-target t-table $\mathtt{t}_{\widetilde{\mathtt{st}}}$ from the source-pivot and pivot-target t-tables $\mathtt{t}_{\mathtt{sp}}, \mathtt{t}_{\mathtt{pt}}$ using the following approximation:

$$\begin{aligned} \mathtt{t}_{\widetilde{\mathtt{st}}}(t \mid s) &= \sum_p \mathtt{t}(t \mid p, s) \cdot \mathtt{t}(p \mid s) \\ &\approx \sum_p \mathtt{t}_{\mathtt{pt}}(t \mid p) \cdot \mathtt{t}_{\mathtt{sp}}(p \mid s) \end{aligned} \quad (1)$$

Subsequently, if a source-target corpus is available, they train a standard source-target model $\Theta_{\mathtt{st}}$, and tune the interpolation

$$\hat{\mathtt{t}}_{\mathtt{st}} = \lambda_{\text{interp}} \mathtt{t}_{\mathtt{st}} + (1 - \lambda_{\text{interp}}) \mathtt{t}_{\widetilde{\mathtt{st}}}$$

with respect to λ_{interp} to reduce alignment error rate (Koehn, 2005) over a hand-aligned development set.

Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL, pages 1221–1226,
Denver, Colorado, May 31 – June 5, 2015. ©2015 Association for Computational Linguistics

Wang et al. (2006) propose triangulation heuristics for other model parameters; however, in this paper, we consider only t-table triangulation.

3 Our Method

We now discuss two approaches that better exploit model triangulation. In the first, we use the triangulated t-table to construct a prior on the source-target t-table. In the second, we place a prior on each of the three models and train them jointly.

3.1 Triangulation as a Fixed Prior

We first propose to better utilize the triangulated t-table $\mathtt{t}_{\widetilde{\mathtt{st}}}$ (Eq. 1) by using it to construct an informative prior for the source-target t-table $\mathtt{t}_{\mathtt{st}} \in \Theta_{\mathtt{st}}$.

Specifically, we modify the word alignment generative story by placing Dirichlet priors on each of the multinomial t-table distributions $\mathtt{t}_{\mathtt{st}}(\cdot \mid s)$:

$$\mathtt{t}_{\mathtt{st}}(\cdot \mid s) \sim \text{Dirichlet}(\alpha_s) \qquad \text{for all } s. \tag{2}$$

Here, each $\alpha_s = (\ldots, \alpha_{st}, \ldots)$ denotes a hyperparameter vector which will be defined shortly.

Fixing this prior, we optimize the model posterior likelihood $P(\Theta_{\mathtt{st}} \mid \mathtt{bitext}_{\mathtt{st}})$ to find a maximum-*a-posteriori* (MAP) estimate. This is done according the MAP-EM framework (Dempster et al., 1977), which differs slightly from standard EM. The E-step remains as is: fixing the model $\Theta_{\mathtt{st}}$, we collect expected counts $E[\mathtt{c}(s,t)]$ for each decision in the generative story. The M-step is modified to maximize the *regularized* expected complete-data log-likelihood with respect to the model parameters $\Theta_{\mathtt{st}}$, where the regularizer corresponds to the prior.

Due to the conjugacy of the Dirichlet priors with the multinomial t-table distributions, the sole modification to the regular EM implementation is in the M-step update rule of the t-table parameters:

$$\mathtt{t}_{\mathtt{st}}(t \mid s) = \frac{E[\mathtt{c}(s,t)] + \alpha_{st} - 1}{\sum_t (E[\mathtt{c}(s,t)] + \alpha_{st} - 1)} \tag{3}$$

where $E[\mathtt{c}(s,t)]$ is the expected number of times source word s aligns with target word t in the source-target bitext. Moreover, through Eq. 3, we can view $\alpha_{st} - 1$ as a pseudo-count for such an alignment.

To define the hyperparameter vector α_s we decompose it as follows:

$$\alpha_s = C_s \cdot m_s + 1 \tag{4}$$

where $C_s > 0$ is a scalar parameter, m_s is a probability vector, encoding the mode of the Dirichlet and 1 denotes an all-one vector. Roughly, when C_s is high, samples drawn from the Dirichlet are likely to concentrate near the mode m_s. Using this decomposition, we set for all s:

$$m_s = \mathtt{t}_{\widetilde{\mathtt{st}}}(\cdot \mid s) \tag{5}$$

$$C_s = \lambda \cdot \mathtt{c}(s)^{\gamma} \cdot \frac{\sum_{s'} \mathtt{c}(s')}{\sum_{s'} \mathtt{c}(s')^{\gamma}} \tag{6}$$

where $\mathtt{c}(s)$ is the count of source word s in the source-target bitext, and the scalar hyperparameters $\lambda, \gamma > 0$ are to be tuned (We experimented with completely eliminating the hyperparameters γ, λ by directly learning the parameters C_s. To do so, we implemented the algorithm of Minka (2000) for learning the Dirichlet prior, but only learned the parameters C_s while keeping the means m_s fixed to the triangulation. However, preliminary experiments showed performance degradation compared to simple hyperparameter tuning). Thus, the distribution $\mathtt{t}_{\mathtt{st}}(\cdot \mid s)$ arises from a Dirichlet with mode $\mathtt{t}_{\widetilde{\mathtt{st}}}(\cdot \mid s)$ and will tend to concentrate around this mode as a function of the frequency of s.

The hyperparameter λ linearly controls the strength of all priors. The last term in Eq. 6 keeps the sum of C_s insensitive to γ, such that $\sum_s C_s = \lambda \sum_s \mathtt{c}(s)$. In all our experiments we fixed $\gamma = 0.5$. Setting $\gamma < 1$ down-weights the parameter C_s of frequent words s compared to rare ones. This makes the Dirichlet prior relatively weaker for frequent words, where we can let the data speak for itself, and relatively stronger for rare ones, where a good prior is needed.

Finally, note that this EM procedure reduces to an interpolation method similar to that of Wang et al. by applying Eq. 3 only at the very last M-step, with α_s, m_s as above and $C_s = \lambda \sum_t E[\mathtt{c}(s,t)]$.

3.2 Joint Training

Next, we further exploit the triangulation idea in designing a multi-task learning approach that jointly trains the three word alignment models $\Theta_{\mathtt{st}}$, $\Theta_{\mathtt{sp}}$, and $\Theta_{\mathtt{pt}}$.

To do so, we view each model's t-table as originating from Dirichlet distributions defined by the triangulation of the other two t-tables. We then train

Algorithm 1 Joint training of $\Theta_{\mathtt{st}}, \Theta_{\mathtt{sp}}, \Theta_{\mathtt{pt}}$

Parameters: $\lambda, \gamma > 0$

- Initialize $\left\{\Theta_{\mathtt{st}}^{(0)}, \Theta_{\mathtt{sp}}^{(0)}, \Theta_{\mathtt{pt}}^{(0)}\right\}$
- Initialize $\{C_{\mathtt{s}}\}, \{C_{\mathtt{p}}\}, \{C_{\mathtt{t}}\}$ as in Eq. 6
- For each EM iteration i:

  Estimate hyperparameters α:

  1. Compute $\mathtt{t}_{\widetilde{\mathtt{st}}}^{(i)}$ from $\mathtt{t}_{\mathtt{sp}}^{(i-1)}$ and $\mathtt{t}_{\mathtt{pt}}^{(i-1)}$ (Eq. 1)
  2. Set $\alpha_{st}^{(i)} := C_s \cdot \mathtt{t}_{\widetilde{\mathtt{st}}}^{(i)}(t \mid s) + 1$

  E: collect expected counts $E[\mathtt{c}(\cdot)]^{(i)}$ from $\Theta_{\mathtt{st}}^{(i-1)}$

  M: Update $\Theta_{\mathtt{st}}^{(i)}$ using $E[\mathtt{c}(\cdot)]^{(i)}$ and $\alpha_{\mathtt{st}}^{(i)}$ (Eq. 3)

  Repeat for $\Theta_{\mathtt{sp}}^{(i)}, \Theta_{\mathtt{pt}}^{(i)}$ using Eq. 7 as required

the models in a MAP-EM like manner, updating both the model parameters and their prior hyperparameters at each iteration. Roughly, this approach aims at maximizing the posterior likelihood of the three models with respect to both model parameters and their hyperparameters (see Appendix).

Procedurally, the idea is simple: In the E-step, expected counts $E[\mathtt{c}(\cdot)]$ are collected from each model as usual. In the M-step, each t-table is updated according to Eq. 3 using the current expected counts $E[\mathtt{c}(\cdot)]$ and an estimate of α from the triangulation of the *most recent* version of the other two models. See Algorithm 1.

Note, however, that we cannot obtain the triangulated t-tables $\mathtt{t}_{\widetilde{\mathtt{sp}}}, \mathtt{t}_{\widetilde{\mathtt{pt}}}$ by simply applying the triangulation equation (Eq. 1). For example, to construct $\mathtt{t}_{\widetilde{\mathtt{sp}}}$ we need both source-to-target and target-to-pivot distributions. While we have the former in $\mathtt{t}_{\mathtt{st}}$, we do not have $\mathtt{t}_{\mathtt{tp}}$. To resolve this issue, we simply approximate $\mathtt{t}_{\mathtt{tp}}$ from the reverse t-table $\mathtt{t}_{\mathtt{pt}} \in \Theta_{\mathtt{pt}}$ as follows:

$$\mathtt{t}_{\mathtt{tp}}(p \mid t) := \frac{\mathtt{c}(p)\mathtt{t}_{\mathtt{pt}}(t \mid p)}{\sum_p \mathtt{c}(p)\mathtt{t}_{\mathtt{pt}}(t \mid p)} \tag{7}$$

where $\mathtt{c}(p)$ denotes the unigram frequency of the word $\mathtt{p}$. A similar transformation is done on $\mathtt{t}_{\mathtt{sp}}$ to obtain $\mathtt{t}_{\mathtt{ps}}$, which is then used in computing $\mathtt{t}_{\widetilde{\mathtt{pt}}}$.

3.3 Adjustment of the t-table

Note that a t-table resulting from the triangulation equation (Eq. 1) is both noisy and dense. To see why, consider that $\mathtt{t}_{\widetilde{\mathtt{st}}}(t \mid s)$ is non-zero whenever there is a pivot word p that co-occurs with both s and t. This is very likely to occur, for example, if p is a function word.

To adjust for both density and noise, we propose a simple product-of-experts re-estimation that relies on the available source-target parallel data. The two experts are the triangulated t-table as defined by Eq. 1 and the exponentiated pointwise mutual information (PMI), derived from simple token co-occurrence statistics of the source-target bitext. That is, we adjust:

$$\mathtt{t}_{\widetilde{\mathtt{st}}}(t \mid s) \quad := \quad \mathtt{t}_{\widetilde{\mathtt{st}}}(t \mid s) \cdot \frac{p(s,t)}{p(s)p(t)}$$

and normalize the result to form valid conditional distributions.

Note that the sparsity pattern of the adjusted t-table matches that of a co-occurrence t-table. We applied this adjustment in all of our experiments.

4 Experimental Results

Pretending that Czech-English is a low-resource pair, we conduct two experiments. In the first, we set French as the pivot language and compare our fixed-prior (Sec. §3.1) and joint training (Sec. §3.2) approaches against the interpolation method of Wang et al. and a baseline HMM word alignment model (Vogel et al., 1996).

In the second, we examine the effect of the pivot language identity on our joint training approach, varying the pivot language over French, German, Greek, Hungarian, Lithuanian and Slovak.

4.1 Data

For word alignment, we use the Czech-English News Commentary corpus, along with a development set of 460 hand aligned sentence pairs. For the MT experiments, we use the WMT10 tuning set (2051 parallel sentences), and both WMT09/10 shared task test sets. See Table 1.

For each of the 6 pivot languages, we created Czech-pivot and pivot-English bitexts of roughly the same size (ranging from 196k sentences for English-Greek to 223k sentences for Czech-Lithuanian). Each bitext was created by forming a Czech-pivot-English tritext, consisting of about 500k sentences

from the Europarl corpus (Koehn, 2005) which was then split into two disjoint Czech-pivot and pivot-English bitexts of equal size. Sentences of length greater than 40 were filtered out from all training corpora.

4.2 Experiment 1: Method Comparison

We trained word alignment models in both source-to-target and target-to-source directions. We used 5 iterations of IBM Model 1 followed by 5 iterations of HMM. We tuned hyperparameters to maximize alignment F-score of the hand-aligned development set. Both interpolation parameters λ_{interp} and λ were tuned over the range $[0, 1]$. For our methods, we fixed $\gamma = 0.5$, which we found effective during preliminary experiments. Alignment F-scores using grow-diag-final-and (gdfa) symmetrization (Koehn, 2010) are reported in Table 2, column 2.

We conducted MT experiments using the Moses translation system (Koehn, 2005). We used a 5-gram LM trained on the Xinhua portion of English Gigaword (LDC2007T07). To tune the decoder, we used the WMT10 tune set. MT B scores are reported in Table 2, columns 3–4.

Both our methods outperform the baseline and the interpolation approach. In particular, the joint training approach more than doubles the gains obtained by the interpolation approach, on both F- and B .

We also evaluated the Czech-French and French-English alignments produced as a by-product of our joint method. While our French-to-English MT experiments showed no improvement in B , we saw a +0.6 (25.6 to 26.2) gain in B on the Czech-to-French translation task. This shows that joint training may lead to some improvements even on high-resource bitexts.

4.3 Other Pivot Languages

We examined how the choice of pivot language affects the joint training approach by varying it over 6 languages (French, German, Greek, Hungarian, Lithuanian and Slovak), while keeping the size of the pivot language resources roughly the same.

	train	dev	WMT09	WMT10
sentences	85k	460	2525	2489
cz tokens	1.63M	9.7k	55k	53k
en tokens	1.78M	10k	66k	62k

Table 1: Czech-English sentence and token statistics.

	F	B	
method/dataset	**dev**	**WMT09**	**WMT10**
baseline	63.8	16.2	16.6
interpolation (Wang)	66.2	16.6	17.1
fixed-prior (§3.1)	67.3	16.9	17.3
joint (§3.2)	**70.1**	**17.2**	**17.7**

Table 2: F- and B scores for Czech-English via French. The joint training method outperforms all other methods tested.

	fr	fr, sk	fr, el	fr, sk, el	all 6
Tune	16.1	16.4	16.4	16.4	16.4
WMT09	17.2	17.2	17.2	17.3	17.4
WMT10	17.7	17.8	17.8	17.8	17.8

Table 3: Czech-English B scores over pivot language combinations. Key: fr=French, sk=Slovak, el=Greek.

Somewhat surprisingly, all models achieved an F-score of about 70%, which resulted in B scores comparable to those reported with French (Table 2). Subsequently, we combined all pivot languages by simply concatenating the aligned parallel texts across pairs, triples and all pivot languages. Combining all pivots yielded modest B score improvements of +0.2 and +0.1 on the test datasets (Table 3).

Considering the low variance in F- and B scores across pivot languages, we computed the pairwise F-scores between the predicted alignments: All scores ranged around 97–98%, indicating that the choice of pivot language had little effect on the joint training procedure.

To further verify, we repeated this experiment over Greek-English and Lithuanian-English as the source-target task (85k parallel sentences), using the same pivot languages as above, and with comparable amounts of parallel data (~200k sentences). We obtained similar results: In all cases, pairwise F-scores were above 97%.

5 Related Work

The term "triangulation" comes from the *phrase-table* triangulation literature (Cohn and Lapata, 2007; Razmara and Sarkar, 2013; Dholakia and

Sarkar, 2014), in which source-pivot and pivot-target phrase tables are triangulated according to Eq. 1 (with words replaced by phrases). The resulting triangulated phrase table can then be combined with an existing source-target phrase table, and is especially useful in increasing the source language vocabulary coverage, reducing OOVs. In our case, since word alignment is a closed vocabulary task, OOVs are never an issue.

In word alignment, Kumar et al. (2007) uses *multilingual* parallel data to compute better source-target alignment posteriors. Filali and Bilmes (2005) tag each source token and target token with their most likely translation in a pivot language, and then proceed to align (source word, source tag) tuple sequences to (target word, target tag) tuple sequences. In contrast, our word alignment method can be applied without multilingual parallel data, and does not commit to hard decisions.

6 Conclusion and Future Work

We presented a simple multi-task learning algorithm that jointly trains three word alignment models over disjoint bitexts. Our approach is a natural extension of a mathematically sound MAP-EM algorithm we originally developed to better utilize the model triangulation idea. Both algorithms are easy to implement (with closed-form solutions for each step) and require minimal effort to integrate into an EM-based word alignment system.

We evaluated our methods on a low-resource Czech-English word alignment task using additional Czech-French and French-English corpora. Our multi-task learning approach significantly improves F- and B scores compared to both baseline and the interpolation method of Wang et al. (2006). Further experiments showed our approach is insensitive to the choice of pivot language, producing roughly the same alignments over six different pivot language choices.

For future work, we plan to improve word alignment and translation quality in a more data restricted case where there are very weak source-pivot resources: for example, word alignment of Malagasy-English via French, using only a Malagasy-French dictionary, or Pashto-English via Persian.

Acknowledgements

The authors would like to thank Kevin Knight, Daniel Marcu and Ashish Vaswani for their comments and insights as well as the anonymous reviewers for their valuable feedback. This work was partially supported by DARPA grants DOI/NBC D12AP00225 and HR0011-12-C-0014 and a Google Faculty Research Award to Chiang.

Appendix: Joint Training Generative Story

We argue that our joint training procedure can be seen as optimizing the posterior likelihood of the three models. Specifically, suppose we place Dirichlet priors on each of the t-tables $\mathtt{t_{st}}, \mathtt{t_{sp}}, \mathtt{t_{pt}}$ as before, but define the prior parameterization using a single hyperparameter $\alpha = \{\alpha_{spt}\}$ and its marginals such that:

$$\begin{aligned}
\mathtt{t_{st}}(\cdot \mid s) &\sim D(\ldots, \alpha_{s\cdot t}, \ldots) & \alpha_{s\cdot t} &= \textstyle\sum_p \alpha_{spt} \\
\mathtt{t_{sp}}(\cdot \mid s) &\sim D(\ldots, \alpha_{sp\cdot}, \ldots) & \alpha_{sp\cdot} &= \textstyle\sum_t \alpha_{spt} \\
\mathtt{t_{pt}}(\cdot \mid p) &\sim D(\ldots, \alpha_{\cdot pt}, \ldots) & \alpha_{\cdot pt} &= \textstyle\sum_s \alpha_{spt}
\end{aligned}$$

Intuitively, α_{spt} represents the number of times a source-pivot-target triplet (s, p, t) was observed.

With this prior, we can maximize the posterior likelihood of the three models given the three bitexts (denoted $\mathtt{data} = \{\mathtt{bitext_{st}}, \mathtt{bitext_{sp}}, \mathtt{bitext_{pt}}\}$) with respect to all parameters and hyperparameters:

$$\begin{aligned}
&\arg\max_{\Theta,\alpha} P(\Theta \mid \alpha, \mathtt{data}) = \\
&\quad \arg\max_{\Theta,\alpha} \textstyle\prod_{d\in\{\mathtt{st},\mathtt{sp},\mathtt{pt}\}} P(\mathtt{bitext}_d \mid \Theta_d) P(\Theta_d \mid \alpha)
\end{aligned}$$

Under the generative story, we need only observe the marginals $\alpha_{s\cdot t}, \alpha_{sp\cdot}, \alpha_{\cdot pt}$ of α. Therefore, instead of explicitly optimizing over α, we can optimize over the marginals while keeping them consistent (via constraints such as $\sum_t \alpha_{s\cdot t} = \sum_p \alpha_{sp\cdot}$ for all s).

In our joint training algorithm (Algorithm 1) we abandon these consistency constraints in favor of closed form estimates of the marginals $\alpha_{s\cdot t}, \alpha_{sp\cdot}, \alpha_{\cdot pt}$.

References

Peter F. Brown, Vincent J.Della Pietra, Stephen A. Della Pietra, and Robert. L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19:263–311.

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

A. P. Dempster, N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38.

Rohit Dholakia and Anoop Sarkar. 2014. Pivot-based triangulation for low-resource languages. In *Proc. AMTA*.

Karim Filali and Jeff Bilmes. 2005. Leveraging multiple languages to improve statistical MT word alignments. In *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*.

P. Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proc. Machine Translation Summit X*, pages 79–86.

Philipp Koehn. 2010. *Statistical Machine Translation*. Cambridge University Press, New York, NY, USA, 1st edition.

Shankar Kumar, Franz Och, and Wolfgang Macherey. 2007. Improving word alignment with bridge languages. In *Proc. EMNLP-CoNLL*.

Thomas P. Minka. 2000. Estimating a Dirichlet distribution. Technical report, MIT.

Majid Razmara and Anoop Sarkar. 2013. Ensemble triangulation for statistical machine translation. In *Proc. IJCNLP*, pages 252–260.

Stephan Vogel, Hermann Ney, and Christoph Tillmann. 1996. HMM-based word alignment in statistical translation. In *Proc. COLING*, pages 836–841.

Haifeng Wang, Hua Wu, and Zhanyi Liu. 2006. Word alignment for languages with scarce resources using bilingual corpora of other language pairs. In *Proc. COLING/ACL*, pages 874–881.