# Improving Pivot Translation by Remembering the Pivot

**Akiva Miura, Graham Neubig, Sakriani Sakti, Tomoki Toda, Satoshi Nakamura**
Graduate School of Information Science
Nara Institute of Science and Technology
8916-5 Takayama-cho, Ikoma-shi, Nara, Japan
{miura.akiba.lr9, neubig, ssakti, tomoki, s-nakamura}@is.naist.jp

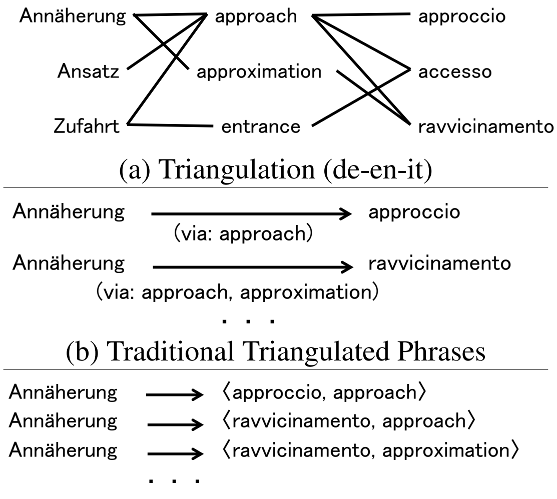
## Abstract

Pivot translation allows for translation of language pairs with little or no parallel data by introducing a third language for which data exists. In particular, the triangulation method, which translates by combining source-pivot and pivot-target translation models into a source-target model, is known for its high translation accuracy. However, in the conventional triangulation method, information of pivot phrases is forgotten and not used in the translation process. In this paper, we propose a novel approach to *remember* the pivot phrases in the triangulation stage, and use a pivot language model as an additional information source at translation time. Experimental results on the Europarl corpus showed gains of 0.4-1.2 BLEU points in all tested combinations of languages[1].

## 1 Introduction

In statistical machine translation (SMT) (Brown et al., 1993), it is known that translation with models trained on larger parallel corpora can achieve greater accuracy (Dyer et al., 2008). Unfortunately, large bilingual corpora are not readily available for many language pairs, particularly those that don't include English. One effective solution to overcome the scarceness of bilingual data is to introduce a pivot language for which parallel data with the source and target languages exists (de Gispert and Mariño, 2006).

Among various methods using pivot languages, the triangulation method (Cohn and Lapata, 2007; Utiyama and Isahara, 2007; Zhu et al., 2014), which translates by combining source-pivot and pivot-target translation models into a source-target model, has been shown to be one of the most effective approaches. However, word sense ambiguity and interlingual differences of word usage cause difficulty in accurately learning correspondences between source and target phrases.


Annäherung
Ansatz
Zufahrt
approach
approximation
entrance
approccio
accesso
ravvicinamento
Annäherung → approccio
(via: approach)
Annäherung → ravvicinamento
(via: approach, approximation)
. . .
Annäherung → ⟨approccio, approach⟩
Annäherung → ⟨ravvicinamento, approach⟩
Annäherung → ⟨ravvicinamento, approximation⟩
. . .

(a) Triangulation (de-en-it)

(b) Traditional Triangulated Phrases

(c) Proposed Triangulated Phrases

Figure 1: An example of (a) triangulation and the resulting phrases in the (b) traditional method of forgetting pivots and (c) our proposed method of remembering pivots.

Figure 1 (a) shows an example of three words in German and Italian that each correspond to the English polysemic word "approach." In such a case, finding associated source-target phrase pairs and estimating translation probabilities properly becomes a complicated problem. Furthermore, in the conventional triangulation method, information about pivot phrases that behave as bridges between source and target phrases is lost after learning phrase pairs, as shown in Figure 1 (b).

To overcome these problems, we propose a novel triangulation method that *remembers* the pivot phrase connecting source and target in the records of phrase/rule table, and estimates a joint translation probability from the source to target

[1] Code to replicate the experiments can be found at https://github.com/akivajp/acl2015

*Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*
*and the 7th International Joint Conference on Natural Language Processing (Short Papers)*, pages 573–577,
Beijing, China, July 26-31, 2015. ©2015 Association for Computational Linguistics

and pivot simultaneously. We show an example in Figure 1 (c). The advantage of this approach is that generally we can obtain rich monolingual resources in pivot languages such as English, and SMT can utilize this additional information to improve the translation quality.

To utilize information about the pivot language at translation time, we train a Multi-Synchronous Context-free Grammar (MSCFG) (Neubig et al., 2015), a generalized extension of synchronous CFGs (SCFGs) (Chiang, 2007), that can generate strings in multiple languages at the same time. To create the MSCFG, we triangulate source-pivot and pivot-target SCFG rule tables not into a single source-target SCFG, but into a source-target-pivot MSCFG rule table that remembers the pivot. During decoding, we use language models over both the target and the pivot to assess the naturalness of the derivation. We perform experiments on pivot translation of Europarl proceedings, which show that our method indeed provide significant gains in accuracy (of up to 1.2 BLEU points), in all combinations of 4 languages with English as a pivot language.

## 2 Translation Formalisms

### 2.1 Synchronous Context-free Grammars

First, we cover SCFGs, which are widely used in machine translation, particularly hierarchical phrase-based translation (Hiero; Chiang (2007)).

In SCFGs, the elementary structures are rewrite rules with aligned pairs of right-hand sides:

$$X \rightarrow \left\langle \overline{s},\ \overline{t} \right\rangle \tag{1}$$

where $X$ is the head of the rewrite rule, and $\overline{s}$ and $\overline{t}$ are both strings of terminals and non-terminals in the source and target side respectively. Each string in the right side tuple has the same number of indexed non-terminals, and identically indexed non-terminals correspond to each-other. For example, a synchronous rule could take the form of:

$$X \rightarrow \left\langle X_0 \text{ of the } X_1,\ X_1 \text{ 的 } X_0 \right\rangle . \tag{2}$$

In the SCFG training method proposed by Chiang (2007), SCFG rules are extracted based on parallel sentences and automatically obtained word alignments. Each extracted rule is scored with phrase translation probabilities in both directions $\phi(\overline{s}|\overline{t})$ and $\phi(\overline{t}|\overline{s})$, lexical translation probabilities in both directions $\phi_{lex}(\overline{s}|\overline{t})$ and $\phi_{lex}(\overline{t}|\overline{s})$, a word penalty counting the terminals in $\overline{t}$, and a constant phrase penalty of 1.

At translation time, the decoder searches for the target sentence that maximizes the derivation probability, which is defined as the sum of the scores of the rules used in the derivation, and the log of the language model probability over the target strings. When not considering an LM, it is possible to efficiently find the best translation for an input sentence using the CKY+ algorithm (Chappelier et al., 1998). When using an LM, the expanded search space is further reduced based on a limit on expanded edges, or total states per span, through a procedure such as cube pruning (Chiang, 2007).

### 2.2 Multi-Synchronous CFGs

MSCFGs (Neubig et al., 2015) are a generalization of SCFGs that are be able to generate sentences in multiple target languages simultaneously. The single target side string $\overline{t}$ in the SCFG production rule is extended to have strings for $N$ target languages:

$$X \rightarrow \left\langle \overline{s},\ \overline{t_1},\ ...,\ \overline{t_N} \right\rangle . \tag{3}$$

Performing multi-target translation with MSCFGs is quite similar to translating using standard SCFGs, with the exception of the expanded state space caused by having one LM for each target. Neubig et al. (2015) propose a *sequential* search method, that ensures diversity in the primary target search space by first expanding with only primary target LM, then additionally expands the states for other LMs, a strategy we also adopt in this work.

In the standard training method for MSCFGs, the multi-target rewrite rules are extracted from multilingual line-aligned corpora by applying an extended version of the standard SCFG rule extraction method, and scored with features that consider the multiple targets. It should be noted that this training method requires a large amount of line-aligned training data including the source and all target languages. This assumption breaks down when we have little parallel data, and thereby we propose a method to generate MSCFG rules by triangulating 2 SCFG rule tables in the following section.

## 3 Pivot Translation Methods

Several methods have been proposed for SMT using pivot languages. These include *cascade* methods that consecutively translate from source to pivot then pivot to target (de Gispert and Mariño, 2006), *synthetic data* methods that machine-translate the training data to generate a pseudo-parallel corpus (de Gispert and Mariño, 2006), and *triangulation* methods that obtain a source-target phrase/rule table by merging source-pivot and pivot-target table entries with identical pivot language phrases (Cohn and Lapata, 2007). In particular, the triangulation method is notable for producing higher quality translation results than other pivot methods (Utiyama and Isahara, 2007), so we use it as a base for our work.

### 3.1 Traditional Triangulation Method

In the triangulation method by Cohn and Lapata (2007), we first train source-pivot and pivot-target rule tables, then create rules:

$$X \rightarrow \langle \bar{s},\ \bar{t} \rangle \tag{4}$$

if there exists a pivot phrase $\bar{p}$ such that the pair $\langle \bar{s}, \bar{p} \rangle$ is in source-pivot table $T_{SP}$ and the pair $\langle \bar{p}, \bar{t} \rangle$ is in pivot-target table $T_{PT}$. Source-target table $T_{ST}$ is created by calculation of the translation probabilities using phrase translation probabilities $\phi(\cdot)$ and lexical translation probabilities $\phi_{lex}(\cdot)$ for all connected phrases according to the following equations (Cohn and Lapata, 2007):

$$\begin{aligned}
\phi\left(\bar{t}|\bar{s}\right) &= \sum_{\bar{p} \in T_{SP} \cap T_{PT}} \phi\left(\bar{t}|\bar{p}\right) \phi\left(\bar{p}|\bar{s}\right), & (5)\\
\phi\left(\bar{s}|\bar{t}\right) &= \sum_{\bar{p} \in T_{SP} \cap T_{PT}} \phi\left(\bar{s}|\bar{p}\right) \phi\left(\bar{p}|\bar{t}\right), & (6)\\
\phi_{lex}\left(\bar{t}|\bar{s}\right) &= \sum_{\bar{p} \in T_{SP} \cap T_{PT}} \phi_{lex}\left(\bar{t}|\bar{p}\right) \phi_{lex}\left(\bar{p}|\bar{s}\right), & (7)\\
\phi_{lex}\left(\bar{s}|\bar{t}\right) &= \sum_{\bar{p} \in T_{SP} \cap T_{PT}} \phi_{lex}\left(\bar{s}|\bar{p}\right) \phi_{lex}\left(\bar{p}|\bar{t}\right). & (8)
\end{aligned}$$

The equations (5)-(8) are based on the memoryless channel model, which assumes $\phi\left(\bar{t}|\bar{p}, \bar{s}\right) = \phi\left(\bar{t}|\bar{p}\right)$ and $\phi\left(\bar{s}|\bar{p}, \bar{t}\right) = \phi\left(\bar{s}|\bar{p}\right)$. Unfortunately, these equations are not accurate due to polysemy and disconnects in the grammar of the languages. As a result, pivot translation is significantly more ambiguous than standard translation.

### 3.2 Proposed Triangulation Method

To help reduce this ambiguity, our proposed triangulation method remembers the corresponding pivot phrase as additional information to be utilized for disambiguation. Specifically, instead of marginalizing over the pivot phrase $\bar{p}$, we create an MSCFG rule for the tuple of the connected source-target-pivot phrases such as:

$$X \rightarrow \langle \bar{s},\ \bar{t},\ \bar{p} \rangle. \tag{9}$$

The advantage of translation with these rules is that they allow for incorporation of additional features over the pivot sentence such as a strong pivot LM.

In addition to the equations (5)-(8), we also estimate translation probabilities $\phi(\bar{t}, \bar{p}|\bar{s})$, $\phi(\bar{s}|\bar{p}, \bar{t})$ that consider both target and pivot phrases at the same time according to:

$$\begin{aligned}
\phi\left(\bar{t}, \bar{p}|\bar{s}\right) &= \phi\left(\bar{t}|\bar{p}\right) \phi\left(\bar{p}|\bar{s}\right), & (10)\\
\phi\left(\bar{s}|\bar{p}, \bar{t}\right) &= \phi\left(\bar{s}|\bar{p}\right). & (11)
\end{aligned}$$

Translation probabilities between source and pivot phrases $\phi(\bar{p}|\bar{s})$, $\phi(\bar{s}|\bar{p})$, $\phi_{lex}(\bar{p}|\bar{s})$, $\phi_{lex}(\bar{s}|\bar{p})$ can also be used directly from the source-pivot rule table. This results in 13 features for each MSCFG rule: 10 translation probabilities, 2 word penalties counting the terminals in $\bar{t}$ and $\bar{p}$, and a constant phrase penalty of 1.

It should be noted that remembering the pivot results in significantly larger rule tables. To save computational resources, several pruning methods are conceivable. Neubig et al. (2015) show that an effective pruning method in the case of a main target $T_1$ with the help of target $T_2$ is the $T_1$-pruning method, namely, using $L$ candidates of $\overline{t_1}$ with the highest translation probability $\phi(\overline{t_1}|\bar{s})$ and selecting $\overline{t_2}$ with highest $\phi(\overline{t_1}, \overline{t_2}|\bar{s})$ for each $\overline{t_1}$. We follow this approach, using the $L$ best $\bar{t}$, and the corresponding 1 best $\bar{p}$ .

## 4 Experiments

### 4.1 Experimental Setup

We evaluate the proposed triangulation method through pivot translation experiments on the Europarl corpus, which is a multilingual corpus including 21 European languages (Koehn, 2005) widely used in pivot translation work. In our work, we perform translation among German (de), Spanish (es), French (fr) and Italian (it), with English (en) as the pivot language. To prepare the data for these 5 languages, we first use the Gale-Church alignment algorithm (Gale and Church, 1993) to retrieve a multilingual line-aligned corpus of about 900k sentences, then hold out 1,500 sentences each for tuning and test. In our basic

| Source | Target | BLEU Score [%] | | | | | |
|---|---|---|---|---|---|---|---|
| | | Direct | Cascade | Tri. SCFG (baseline) | Tri. MSCFG -PivotLM | Tri. MSCFG +PivotLM 100k | Tri. MSCFG +PivotLM 2M |
| de | es | 27.10 | 25.05 | 25.31 | 25.38 | 25.52 | † **25.75** |
| | fr | 25.65 | 23.86 | 24.12 | 24.16 | 24.25 | † **24.58** |
| | it | 23.04 | 20.76 | 21.27 | 21.42 | † 21.65 | ‡ **22.29** |
| es | de | 20.11 | 18.52 | 18.77 | 18.97 | 19.08 | † **19.40** |
| | fr | 33.48 | 27.00 | 29.54 | † 29.87 | † 29.91 | † **29.95** |
| | it | 27.82 | 22.57 | 25.11 | 25.01 | 25.18 | ‡ **25.64** |
| fr | de | 19.69 | 18.01 | 18.73 | 18.77 | 18.87 | † **19.19** |
| | es | 34.36 | 27.26 | 30.31 | 30.53 | † 30.73 | ‡ **31.00** |
| | it | 28.48 | 22.73 | 25.31 | 25.50 | † 25.72 | ‡ **26.22** |
| it | de | 19.09 | 14.03 | 17.35 | † 17.99 | ‡ 18.17 | ‡ **18.52** |
| | es | 31.99 | 25.64 | 28.85 | 28.83 | 29.01 | † **29.31** |
| | fr | 31.39 | 25.87 | 28.48 | 28.40 | 28.63 | † **29.02** |

Table 1: Results for each method. Bold indicates the highest BLEU score in pivot translation, and daggers indicate statistically significant gains over Tri. SCFG ($\dagger : p < 0.05, \ddagger : p < 0.01$)

training setup, we use 100k sentences for training both the TMs and the target LMs. We assume that in many situations, a large amount of English monolingual data is readily available and therefore, we train pivot LMs with different data sizes up to 2M sentences.

As a decoder, we use Travatar (Neubig, 2013), and train SCFG TMs with its Hiero extraction code. Translation results are evaluated by BLEU (Papineni et al., 2002) and we tuned to maximize BLEU scores using MERT (Och, 2003). For trained and triangulated TMs, we use $T_1$ rule pruning with a limit of 20 rules per source rule. For decoding using MSCFG, we adopt the sequential search method.

We evaluate 6 translation methods:

**Direct:** Translating with a direct SCFG trained on the source-target parallel corpus (not using a pivot language) for comparison.

**Cascade:** Cascading source-pivot and pivot-target translation systems.

**Tri. SCFG:** Triangulating source-pivot and pivot-target SCFG TMs into a source-target SCFG TM using the traditional method.

**Tri. MSCFG:** Triangulating source-pivot and pivot-target SCFG TMs into a source-target-pivot MSCFG TM in our approach. -PivotLM indicates translating without a pivot LM and +PivotLM 100k/2M indicates a pivot LM trained using 100k/2M sentences respectively.

### 4.2 Experimental Results

The result of experiments using all combinations of pivot translation tasks for 4 languages via English is shown in Table 1. From the results, we can see that the proposed triangulation method considering pivot LMs outperforms the traditional triangulation method for all language pairs, and translation with larger pivot LMs improves the BLEU scores. For all languages, the pivot-remembering triangulation method with the pivot LM trained with 2M sentences achieves the highest score of the pivot translation methods, with gains of 0.4-1.2 BLEU points from the baseline method. This shows that remembering the pivot and using it to disambiguate results is consistently effective in improving translation accuracy.

We can also see that the MSCFG triangulated model without using the pivot LM slightly outperforms the standard SCFG triangulation method for the majority of language pairs. It is conceivable that the additional scores of translation probabilities with pivot phrases are effective features that allow for more accurate rule selection.

Finally, we show an example of a translated sentence for which pivot-side ambiguity is resolved in the proposed triangulation method:

**Input (German):** ich bedaure , daß es keine gemeinsame annäherung gegeben hat .

**Reference (Italian):** sono spiacente del mancato approccio comune .

**Tri. SCFG:** mi rammarico per il fatto che non si ravvicinamento comune . (BLEU+1: 13.84)

**Tri. MSCFG+PivotLM 2M:**

mi dispiace che non esiste un approccio comune . (BLEU+1: 25.10)
i regret that there is no common approach . (Generated English Sentence)

The derivation uses an MSCFG rule connecting "approccio" to "approach" in the pivot, and we can consider that appropriate selection of English words according to the context contributes to selecting relevant vocabulary in Italian.

## 5 Conclusion

In this paper, we have proposed a method for pivot translation using triangulation of SCFG rule tables into an MSCFG rule table that remembers the pivot, and performing translation with pivot LMs. In experiments, we found that these models are effective in the case when a strong pivot LM exists. In the future, we plan to explore more refined methods to devising effective intermediate expressions, and improve estimation of probabilities for triangulated rules.

## Acknowledgements

The authors thank anonymous reviewers for helpful suggestions. This work was supported in part by the Microsoft CORE project.

## 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*, 19:263–312.

Jean-Cédric Chappelier, Martin Rajman, et al. 1998. A Generalized CYK Algorithm for Parsing Stochastic CFG. *TAPD*, 98(133-137):5.

David Chiang. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201–228.

Trevor Cohn and Mirella Lapata. 2007. Machine Translation by Triangulation: Making Effective Use of Multi-Parallel Corpora. In *Proc. ACL*, pages 728–735, June.

Adrià de Gispert and José B. Mariño. 2006. Catalan-English Statistical Machine Translation without Parallel Corpus: Bridging through Spanish. In *Proc. of LREC 5th Workshop on Strategies for developing machine translation for minority languages*.

Christopher Dyer, Aaron Cordova, Alex Mont, and Jimmy Lin. 2008. Fast, easy, and cheap: construction of statistical machine translation models with MapReduce. In *Proc. WMT*, pages 199–207.

William A Gale and Kenneth W Church. 1993. A program for aligning sentences in bilingual corpora. *Computational linguistics*, 19(1):75–102.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.

Graham Neubig, Philip Arthur, and Kevin Duh. 2015. Multi-Target Machine Translation with Multi-Synchronous Context-free Grammars. In *Proc. NAACL*.

Graham Neubig. 2013. Travatar: A Forest-to-String Machine Translation Engine based on Tree Transducers. In *Proc. ACL Demo Track*, pages 91–96.

Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proc. ACL*, pages 160–167.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proc. ACL*, pages 311–318.

Masao Utiyama and Hitoshi Isahara. 2007. A Comparison of Pivot Methods for Phrase-Based Statistical Machine Translation. In *Proc. NAACL*, pages 484–491.

Xiaoning Zhu, Zhongjun He, Hua Wu, Conghui Zhu, Haifeng Wang, and Tiejun Zhao. 2014. Improving Pivot-Based Statistical Machine Translation by Pivoting the Co-occurrence Count of Phrase Pairs. In *Proc. EMNLP*.