The JAIST Machine Translation Systems for WMT 17

Hai-Long Trieu and **Trung-Tin Pham** and **Le-Minh Nguyen**
School of Information Science
Japan Advanced Institute of Science and Technology
{trieulh, tinpt, nguyenml}@jaist.ac.jp

Abstract

We describe the JAIST phrase-based machine translation systems that participated in the news translation shared task of the WMT17. In this work, we participated in the Turkish-English translation, in which only a small amount of bilingual training data is available, so that it is an example of the low-resource setting in machine translation. In order to solve the problem, we focus on two strategies: building a bilingual corpus from comparable data and exploiting existing parallel data based on phrase pivot translation. In order to utilize the strategies to enhance machine translation on the low-resource setting most effectively, we introduce a system combining the extracted corpus, the pivot translation, and the direct training data. Experimental results showed that our combined systems significantly improved the baseline models, which were trained on the small bilingual data.

1 Introduction

We participated in the WMT 17 news translation shared task for the Turkish-English language pair. The amount of bilingual training data for this language pair is small, which means that this machine translation task poses the problem of a low-resource setting. The problem causes a bottleneck for current data-driven machine translation methods including phrase-based and neural-based machine translation because there are few large bilingual corpora for most language pairs in the world (Irvine, 2013; Wang et al., 2016).

In our systems, we focus on two strategies to enhance machine translation for the low-resource setting: building a bilingual corpus from comparable data, and exploiting existing parallel corpora based on the phrase pivot translation (Wu and Wang, 2007; Cohn and Lapata, 2007; Utiyama and Isahara, 2007). First, we built a bilingual corpus for Turkish-English based on parallel titles of Wikipedia articles. The parallel titles were extracted from Wikipedia articles' titles and interlanguage link records. Bilingual articles were collected based on the title pairs. Then, bilingual sentences were extracted from the article pairs using the Microsoft sentence aligner (Moore, 2002). Second, we exploited the phrase pivot translation method using six pivot languages to bridge the translation between Turkish and English. Finally, the two resources of the extracted corpus and the pivot translation were utilized with the direct bilingual training data in a combined system. Our combined systems achieved a significant improvement compared with the baseline model, which was trained on the direct bilingual data. The code and datasets used in our systems can be found at the repository.[1]

2 Methods

We describe approaches used in our systems. The Turkish-English bilingual data in this shared task embodies only 207k parallel sentences, which is an instance of machine translation task in low-resource setting. Our goal is to enhance the phrase-based machine translation on the low-resource setting by using two approaches: building a Turkish-English bilingual corpus from comparable data, and exploiting existing parallel corpora based on the phrase pivot translation method. The two approaches were then combined to enhance machine translation on the low-resource setting most effectively.

[1] https://github.com/nguyenlab/WMT17-JAIST

2.1 Building A Turkish-English Bilingual Corpus from Comparable Data

We built a bilingual corpus for Turkish-English from comparable data to improve machine translation on the low-resource setting. We used Wikipedia, a free accessible resource containing articles in the same domain and topics in different languages, to build the corpus. In order to build a bilingual corpus from Wikipedia, we based on parallel titles of Wikipedia articles. Then, pairs of articles were crawled based on the parallel titles. Finally, sentences in the article pairs were aligned to extract parallel sentences. We describe these steps in more detail in this section.

Extracting Parallel Titles The content of Wikipedia can be obtained from their database dumps.[2] In order to extract parallel titles of Wikipedia articles, we used two resources for each language from the Wikipedia database dumps: the articles' titles and IDs in a particular language (ending with *-page.sql.gz*) and the interlanguage link records (file ends with *-langlinks.sql.gz*).

Collecting Parallel Articles After parallel titles of Wikipedia articles were extracted, we collected the article pairs using the parallel titles. We implemented a Java crawler for collecting the articles. The collected data was then preprocessed including sentence split and word tokenization using the Moses scripts.[3]

Sentence Alignment For each article pair, bilingual sentences were aligned using the Microsoft bilingual sentence aligner (Moore, 2002), one of the most powerful sentence alignment algorithms as shown in (Singh and Husain, 2005). After the sentence alignment step, we obtained a Turkish-English bilingual corpus with 48k parallel sentences, which is presented in Table 1.

	Turkish	English
Input articles	188,235	192,512
Input sentences	2,030,931	3,023,324
Bilingual articles	184,154	184,154
Aligned articles	22,100	22,100
Aligned sentences	48,554	48,554

Table 1: Building a bilingual corpus of Turkish-English from Wikipedia.

[2]https://dumps.wikimedia.org/backup-index.html

[3]https://github.com/moses-smt/mosesdecoder/tree/master/scripts/tokenizer

From the results, for 184k input bilingual articles, a small ratio of 22k articles were aligned. One of the main reasons is the characteristic of Wikipedia bilingual articles, in which each article in a language of a bilingual article pair is created separately by different authors with different styles of writing, background knowledge, etc. This leads to various challenges for aligning parallel sentences such as: the small portion of overlap in the article pair's content, the unbalance of sentence length, the unbalance of numbers of sentences in the articles. Further investigations on the Wikipedia data as well as different aligners and methods are needed to improve the performance on this task.

2.2 Phrase Pivot Translation

In order to enhance machine translation for the low-resource setting, we exploited existing bilingual corpora using the phrase pivot translation method (Cohn and Lapata, 2007; Utiyama and Isahara, 2007; Wu and Wang, 2007). In the phrase pivot translation method, source-pivot and pivot-target bilingual corpora are used to train phrase tables. Then, the source and target phrases are connected via common pivot phrases.

Given a source phrase s and a target phrase t of the source-pivot phrase table T_{SP} and the pivot-target phrase table T_{PT}, the phrase translation probability is estimated via common pivot phrases p based on the following feature function.

$$\phi(t|s) = \sum_{p \in (T_{SP}) \cap (T_{PT})} \phi(p|s)\phi(t|p) \quad (1)$$

Previous research showed the effectiveness of this method when source-target bilingual corpora are unavailable or in a limited amount.

In our systems, we used bilingual data sets of the SETIMES2 corpus (Tiedemann, 2009)[4], the same resource of the Turkish-English training data in this shared task, for training phrase pivot translation. We used six pivot languages to bridge the translation between Turkish and English: Bulgarian, Bosnian, Greek, Macedonian, Romanian, and Albanian. The bilingual corpora for the pivot translation are presented in Table 2.

For phrase pivot translation, we implemented the triangulation method of (Wu and Wang, 2007) using Java. One of the issues of the triangulation

[4]http://opus.lingfil.uu.se/SETIMES2.php

No.	Pivot	tr-pvt	pvt-en	tr-en	en-tr
1	bg	206k	213k	393k	490k
2	bs	133k	138k	321k	374k
3	el	206k	226k	390k	472k
4	mk	202k	207k	387k	469k
5	ro	205k	212k	382k	457k
6	sq	206k	227k	379k	446k

Table 2: Bilingual corpora for Turkish-English pivot translation (the number of parallel sentences) and the number of pivoted phrase pairs in Turkish-English (**tr-en**) and English-Turkish (**en-tr**); **Pivot** languages: **bg** (Bulgarian), **bs** (Bosnian), **el** (Greek), **mk** (Macedonian), **ro** (Romanian), **sq** (Albanian); **tr-pvt (pvt-en)**: the bilingual corpus of Turkish and the pivot language (pivot-English)

method is that the number of pivoted phrase pairs is exploded (El Kholy et al., 2013). Therefore, we filtered the pivoted phrase tables by using a *n-best* technique in which for a set of *n* best target phrases was extracted for each source phrase (*n was set to 10 in our experiments*).

2.3 Combining Additional Resources

We exploited two resources to enhance machine translation for the low-resource setting: a bilingual corpus extracted from Wikipedia, and bilingual corpora of Turkish and English paired with the six pivot languages. Our goal now is to utilize the resource most effectively. We introduce a system incorporating the following components. First, we trained a phrase table based on the Wikipedia bilingual corpus, called *align* component. Second, using the phrase pivot translation, we obtained pivoted phrase table, called the *pivot* components. Additionally, we trained a phrase table using the Turkish-English training data, called *baseline* component. The components were combined to generate a phrase table for decoding. We adapted the linear interpolation (Sennrich, 2012) for combining phrase tables. Equation 2 describes the combination of the components.

$$\begin{aligned} p(t|s) =& \lambda_d p_d(t|s) + \lambda_a p_a(t|s) \\ &+ \lambda_{p_1} p_1(t|s) + \lambda_{p_2} p_2(t|s) + \lambda_{p_3} p_3(t|s) \\ &+ \lambda_{p_4} p_4(t|s) + \lambda_{p_5} p_5(t|s) + \lambda_{p_6} p_6(t|s) \end{aligned} \quad (2)$$

Where $p_d(t|s)$, $p_a(t|s)$ stand for the translation probability of the *baseline* and the *align* components, respectively. $p_i(t|s), i = 1..6$ stand for the translation probability of the six pivoted phrase tables.

The interpolation parameters λ_d, λ_a, and $\lambda_{p_i}(i = 1..6)$ in which $\lambda_d + \lambda_a + \lambda_{p_i} = 1$ were tuned based on the interpolation method (Sennrich, 2012) using the development set (*newsdev2016*) provided by the shared task.

3 Experiments

We describe the data sets, settings, and results of our systems in this section. We discuss the experimental results on three settings: building a bilingual corpus, using phrase pivot translation, and using the system combining the two components.

3.1 Training Data

We used the training, development, and test sets provided by the WMT 17 shared task. The Turkish-English training data contain 207k parallel sentences. For the development set, we used the *dev2016*. We evaluated our systems on the *tst2016*, and submitted the translation output for the *tst2017* test set.

For monolingual datasets to train language models, we used the monolingual datasets provided by the shared task: 40M sentences of Turkish and 40M sentences of English.

3.2 Baseline Systems

We conducted baseline experiments for phrase-based machine translation using the Moses toolkit (Koehn et al., 2007). The word alignment was trained using GIZA++ (Och and Ney, 2003) with the configuration *grow-diag-final-and*. 5-gram language models of Turkish and English were trained using KenLM (Heafield, 2011). For tuning, we used the batch MIRA (Cherry and Foster, 2012). The system's outputs were evaluated using the NIST-BLEU on the online system.[5]

3.3 Experimental Results

The results of the JAIST systems are presented in Table 3 and Table 4. We discuss the results for the three different settings.

3.3.1 Building A Bilingual Corpus

Although the aligned Wikipedia corpus contains a small number of parallel sentences (48k) compared with the direct training data (207k), the phrase-based models trained on the Wikipedia

[5] http://matrix.statmt.org/

Model	newsdev2016	newstest2016	newstest2017
baseline	12.28	12.3	12.0
align	7.67	8.1	7.9
pivot (bs)	7.47	11.0	7.6
baseline-align	13.35	12.9 (+0.6)	12.7 (+0.7)
baseline-pivot(bs)	12.39	13.1 (+0.8)	12.4 (+0.4)
baseline-pivot(bs)-align	13.02	13.0 (+0.7)	12.7 (+0.4)
baseline-pivot(6)-align	14.04	**13.7 (+1.4)**	**13.1 (+1.1)**

Table 3: Experimental results on the Turkish-English (BLEU); **baseline (align)**: the system trained on the baseline (the aligned Wikipedia) bilingual corpus; **pivot (bs), pivot (6)**: the phrase pivot translation system using one pivot language (bs: Bosnian) or using all of the 6 pivot languages; **baseline-pivot(6)-align**: the combined system of the baseline, align, and 6 pivot components.

Model	newsdev2016	newstest2016	newstest2017
baseline	8.66	9.3	9.9
align	5.96	6.3	6.6
pivot (bs)	6.01	8.2	6.3
baseline-align	8.87	9.3	10.0 (+0.1)
baseline-pivot	9.01	9.6 (+0.3)	9.7
baseline-pivot(bs)-align	8.98	9.6 (+0.3)	9.9
baseline-pivot(6)-align	10.11	**9.7 (+0.4)**	**10.4 (+0.5)**

Table 4: Experimental results on the English-Turkish translation (BLEU).

corpus showed a quite promising result: 7.9 BLEU point on the Turkish-English and 6.6 BLEU point on the English-Turkish. When the baseline model was combined with the align model, we achieved a significant improvement: +0.6 and +0.7 BLEU points on the Turkish-English of the *newstest2016* and *newstest2017*, respectively. The results showed the effectiveness of the extracted corpus to enhance machine translation on the low-resource setting. Nevertheless, the task becomes more challenging on the English-Turkish. Although the Wikipedia corpus showed the contribution on the Turkish-English translation, there was no improvement on the English-Turkish translation when we achieved only +0.1 BLEU point on the *newstest2017*.

3.3.2 Phrase Pivot Translation

For the phrase pivot translation models, using one pivot language (bs: Bosnian) showed the competitive performance on the newstest2016 of the Turkish-English: 11.0 BLEU point vs. 12.3 BLEU point (baseline), or 8.2 BLEU point vs. 9.3 BLEU point (baseline) on the English-Turkish.

When the pivot model (using one pivot language of Bosnian) was combine with the baseline model, we achieved the improvement on both translation directions: +0.8 BLEU point on the Turkish-English, and +0.3 BLEU point on the English-Turkish of the *newstest2016*. For the newstest2017, we achived the improvement only on the Turkish-English (+0.4 BLEU point).

The results confirmed the contribution of the phrase pivot translation. Nevertheless, there was no improvement on some cases. Therefore, we seek to the combination of all components: the baseline, align, and pivot components (from one pivot language to six pivot languages).

3.3.3 Combined Systems

We would like to exploit the components most effectively to improve machine translation on the low-resource setting. The baseline, align, and pivot components were combined in a model. When using one pivot language (Bosnian), we achieved the improvement in most cases: +0.7 and +0.4 BLEU points on the *newstest2016* and *newstest2017* of the Turkish-English. For the English-Turkish, we achieved the improvement of +0.3 BLEU point on the *newstest2016*; however, there was no improvement on the *newstest2017*, in which the pivot model did not showed the contribution.

Interestingly, using six pivot languages showed the significant improvement in all settings. For the Turkish-English, we achieved +1.4 and +1.1 BLEU points on the *newstest2016* and *newstest2017*, respectively. For the English-Turkish, the combined system showed +0.4 BLEU point (newstest2016) and +0.5 BLEU point (newstest2017).

We submitted our systems using the settings

that combine the baseline, align, and six pivot languages in the phrase pivot translation.

4 Conclusion

We describe our phrase-based machine translation systems for Turkish-English participated in the WMT 17 news translation shared task. In this work, our goal is to enhance machine translation for the low-resource setting for Turkish-English, in which a only small training bilingual data is available. Two approaches were exploited in our systems: building a bilingual corpus from Wikipedia, and utilizing existing bilingual corpora using the phrase pivot translation method. In order to exploit the extracted data most effectively, we introduce a combined system of the aligned corpus, the pivot data, and the direct training data. We achieved a significant improvement on the *newstest2016* and *newstest2017*. The results showed the effectiveness of the extracted corpus and the pivot translation in improving machine translation on the low-resource setting. We released the Wikipedia corpus, which can be used to improve machine translation on Turkish-English in future work.

Acknowledgement

This work was supported by JSPS KAKENHI Grant number JP15K16048 and the VNU project "Exploiting Very Large Monolingual Corpora for Statistical Machine Translation" (code QG.12.49).

References

Colin Cherry and George Foster. 2012. Batch tuning strategies for statistical machine translation. In *Proceedings of HLT/NAACL*. Association for Computational Linguistics, pages 427–436.

Trevor Cohn and Mirella Lapata. 2007. Machine translation by triangulation: making effective use of multi-parallel corpora. In *Proceedings of ACL*. Association for Computational Linguistics, pages 728–735.

Ahmed El Kholy, Nizar Habash, Gregor Leusch, Evgeny Matusov, and Hassan Sawaf. 2013. Language independent connectivity strength features for phrase pivot statistical machine translation. In *Proceedings of ACL*. Association for Computational Linguistics, pages 412–418.

Kenneth Heafield. 2011. Kenlm: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, pages 187–197.

Ann Irvine. 2013. Statistical machine translation in low resource settings. In *Proceedings of HLT/NAACL*. Association for Computational Linguistics, pages 54–61.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of ACL*. Association for Computational Linguistics, pages 177–180.

Robert C Moore. 2002. *Fast and accurate sentence alignment of bilingual corpora*. Springer.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics* 29(1):19–51.

Rico Sennrich. 2012. Perplexity minimization for translation model domain adaptation in statistical machine translation. In *Proceedings of EAMT*. pages 539–549.

Anil Kumar Singh and Samar Husain. 2005. Comparison, selection and use of sentence alignment algorithms for new language pairs. In *Proceedings of the ACL Workshop on Building and using Parallel texts*. Association for Computational Linguistics, pages 99–106.

Jörg Tiedemann. 2009. News from OPUS - A collection of multilingual parallel corpora with tools and interfaces. In N. Nicolov, K. Bontcheva, G. Angelova, and R. Mitkov, editors, *Recent Advances in Natural Language Processing*, John Benjamins, Amsterdam/Philadelphia, Borovets, Bulgaria, volume V, pages 237–248.

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

Pidong Wang, Preslav Nakov, and Hwee Tou Ng. 2016. Source language adaptation approaches for resource-poor machine translation. *Computational Linguistics* .

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