TÜBİTAK SMT System Submissions for WMT 2016

Emre Bektaş, Ertuğrul Yılmaz, Coşkun Mermer, İlknur Durgar El-Kahlout
TÜBİTAK-BİLGEM
Gebze 41470, Kocaeli, Turkey
{emre.bektas,yilmaz.ertugrul,coskun.mermer,ilknur.durgar}@tubitak.gov.tr

Abstract

We describe the TÜBİTAK Turkish-English machine translation systems submissions in both directions for the WMT 2016: News Translation Task. We experiment with phrase-based and hierarchical phrase-based systems for both directions using word-level and morpheme-level representations for the Turkish side. Finally we perform system combination which results in 0.5 BLEU increase for Turkish-to-English and 0.3 BLEU increase for English-to-Turkish.

1 Introduction

This paper presents TÜBİTAK's submissions for the news translation task of the First Conference on Machine Translation (WMT16) held at ACL 2016. Overview of the systems can be described as follows: We use both word-level and morphological feature-based representation of Turkish for both directions. We experiment with both phrase-based and hierarchical phrase-based systems. A large 5-gram language model is trained with data extracted from the common crawl corpus provided in Turkish and a 4-gram gigaword language model is used for English. Augmenting the training data with its content words (add a new parallel corpora to training consisting of only the content words for both languages) and using reversed training data on the source side in order to achieve better alignments at the root-word level and surface forms, are amongst the methods we employ. Finally system combination of systems with different paradigms is performed.

This paper is organized as follows: Section 2 introduces the challenges of practicing SMT for the Turkish-English language pair and summarizes the previous work. Section 3 provides background on the base SMT approaches we experiment with. Section 4 provides the experimental specifications and reports on the results in both directions. We conclude with section 5.

2 Turkish-English Statistical Machine Translation

Development of statistical machine translation (SMT) systems of typologically different languages have traditionally been quite challenging. The morphological complexity of Turkish compared to English as well as the constituency order difference between these languages makes the SMT practices especially challenging. English language structurally conforms to the Subject-Verb-Object (SVO) constituent order unlike Turkish which has a very flexible constituent order of mostly Subject-Object-Verb (SOV).

Turkish is an agglutinative language wherein words are created by concatenating morphemes (stems and affixes). These combinations are conditioned by certain morphological rules such as vowel harmony and consonant assimilation which are set to preserve the overall gentleness of the language. This means a morpheme can change its form while preserving its meaning in order to suit these rules. After a number of derivations word forms can become quite complex which results in a larger vocabulary. Such complex Turkish words typically align with whole phrases on the English side when sentence pairs are aligned at the word level. Such a morphologically complex language proves to be quite challenging from an SMT point of view.

To reduce the large vocabulary size and to force more one-to-one word alignments, researchers prefer a sub-word representation of the morphologically richer foreign language while translating to/from English.

Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers, pages 246–251,
Berlin, Germany, August 11-12, 2016. ©2016 Association for Computational Linguistics

Mapping the rich morphology of Turkish to the limited morphology of English has been addressed by several researchers. El-Kahlout et al. (2012) and Oflazer (2008) used morphological analysis to separate some Turkish inflectional morphemes that have counterparts on the English side in English-to-Turkish SMT. Along the same direction, Yeniterzi and Oflazer (2010) applied syntactic transformations such as joining function words on the English side to the related content words.

On the other hand Mermer and Akin (2010) used an unsupervised learning algorithm to find the segmentations automatically from parallel data. A series of segmentation schemes has been presented (Ruiz et al., 2012) to explore the optimal segmentation for statistical machine translation of Turkish to English. In addition, an important amount of effort was spent by several research groups on Turkish-to-English SMT in the IWSLT'09 (Paul, 2009) and IWSLT'10 (Paul et al., 2010) BTEC tasks, IWSLT'12 (Federico et al., 2012) and IWSLT'13 (Cettolo et al., 2013) TED tasks.

Several components such as the morphological analyzer and the Turkish word generator that were used in this submission were adopted from the experiments that had been conducted for IWSLT'13 TED tasks by Yilmaz et al. (2013).

3 Phrase-Based vs. Hierarchical Phrase-Based Systems

Although phrase-to-phrase translation (Koehn et al., 2003) overcomes many problems of word-to-word translation (Brown et al., 1993) and has been successful for some language pairs during the last decade, the continuity of phrases is its main shortcoming. In general, this is a problem for language pairs with very different word orders such as Chinese-English. For such language pairs, in order to generate the target phrase, we may need sub-phrases from different parts of the source sentence which are distant from each other. To overcome the limitations of the phrase-based model, Chiang (2007) has introduced a hierarchical phrase-based model that uses bilingual phrase pairs to generate hierarchical phrases that allow gaps and enable longer distance reorderings.

Previous work (El-Kahlout et al., 2012; Ruiz et al., 2012) showed that hierarchical phrase-based (HPB) systems outperform phrase-based (PB) systems for Turkish-English.

4 Experiments

4.1 Overview

In the experiments the SETIMES parallel corpora provided were used as training data. The systems were tuned with newsdev2016 consisting of 1000 sentences and tested with the test set newstest2016 of 3000 sentences. GIZA++ toolkit (Och and Ney, 2003) for the word alignment and Moses' base decoders for both HPB and PB systems were utilized. For the PB decoders lexicalized reordering was turned on, the distortion limit was set to 6 (dl = 6) unless no distortion limit (dl = -1) was explicitly indicated. For the HPB decoder cube pruning pop limit was set to 5000.

4.2 Word Representation vs. Full Segmentation

We implemented both the word-level representation and feature based representation of Turkish as baseline systems. As mentioned in Section 2, incorporating morphology when working with morphologically rich(er) languages in SMT is expected to perform better than the word-level approach.

Data Set	Sentences	# of Tokens
Turkish(Word)	208k	3.6M
Turkish(Feature)	208k	7.4M
English	208k	4.4M

Table 1: SETIMES parallel training data statistics.

Table 1 shows the training data statistics before and after morphological analysis. As it was commonplace for sentences to become quite a bit longer due to the morphological segmentation, in our experiments we used a maximum sentence length of 100.

4.3 Pre-processing

We normalized all the data used in the experiments. This includes removing extra spaces, dealing with unicode punctuation, normalizing quotation marks and commas. The word-level representation of Turkish and English were produced using the default Moses tokenizer.

A morphological analyzer (Oflazer, 1994) was used to produce the feature-based representation of the Turkish language. Each word is passed through the analyzer which outputs all the possible interpretations of that word containing the stem and the morphological features. Then morphological disambiguation is performed on the morphological analyses (Sak et al., 2007).

Once the contextually salient morphological interpretation is selected, we removed the redundant morphological features that do not correspond to a surface morpheme such as part-of-speech features (Noun, Verb etc.), 3rd singular agreement feature (A3sg), and positiveness feature (Pos) and so on. There only remained features that correspond to lexical morphemes making up a word such as dative (Dat), accusative (Acc), past participle (PastPart) and so on. We segmented the morphologically-analyzed Turkish sentences at every feature boundary, denoted by the (_) symbol. A typical sentence pair with Turkish word representation and full segmentation is as follows:

- **Word representation:** Kosova'nın özelleştirme süreci büyüteç altında.
- **Feature representation:** Kosova _Gen özel _Become _Caus _Inf2 süreç _P3sg büyüteç alt_P3sg _Loc.
- **Reference:** Kosova's privatisation process is under scrutiny.

4.4 Language Models

The language models were trained using SRILM (Stolcke, 2002) toolkit. For Turkish to English we used a 4-gram Gigaword language model. For English to Turkish experiments we used the monolingual Common Crawl Corpus hosted by Amazon Web Services as a public data set. While being quite large, the crawl data consisted of mostly out of domain grammatically and semantically broken sentences. Even though the provided data was supposed to be de-duplicated we encountered duplicates of sub-sentences embedded within a single sentence which may have been missed by the de-duplication script. We encountered sentences that include only a word, bad UTF-8 characters, sentences containing Turkish characters that were replaced with a UTF-8 place-holder character which were irreversible since all the non-Latin characters were mapped to the same place-holder.

Therefore we processed the monolingual data to train a stronger language model. Firstly we employed the same normalization process as was done on the training, tuning and the test corpus described in Section 4.3. We lowercased the sentences that included fully upper-cased words and phrases. Then we removed the parts in which some characters were irreversibly swapped by UTF-8 place-holder, empty lines, the sentences that consisted of only numbers or characters, URL's and dates.

In addition to the language models trained from the crawl data, two 5-gram language models were trained using the parallel corpora which were then interpolated with the aforementioned language models using SRILM.

Data Set	**Lines**	**Total Words**
TR-CC-lm	28M	796M

Table 2: Filtered crawl-data language model statistics.

4.5 Methods

In our experiments we used both HPB and PB decoders for both directions. For Turkish-to-English we observed that the HPB systems outperformed the PB systems and for English-to-Turkish PB systems outperformed the HPB ones. For both directions we augmented the training data with its content words in order to increase the alignments at root-word level. For the English side this was achieved by using TreeTagger (Schmid, 1994) to tag the sentences and remove all the non-content words (the remaining part-of-speech tags and conjunctions etc.) (Yilmaz et al., 2013). For the Turkish side the morphological analyzer we described in section 4.3 was used to strip the corpus of any non-content words, in this case part-of-speech features. Finally the Turkish and English corpora that consisted of only the content words were then added to the original parallel corpora effectively doubling its size and enlarging their vocabularies.

For another experiment, reversed corpora for the source side for each direction were used in hopes of achieving a more accurate word alignment.

Table 3 shows the experimental results of the official test set for the Turkish-to-English direction. We observed that removing the distortion limit (dl = -1) on re-ordering improves the performance of the PB system. Later the strength of the diverse systems were combined using the open-source system combination tool MEMT (Heafield

Experiment	newstest2016
1. HPB Word Rep.	12.78
2. HPB Feature rep.	14.68
3. 2 + GW_LM	15.46
4. 3 + Content Corpus	14.94
5. 3 + Reverse Corpus	13.42
6. 1 with PB	11.20
7. 2 with PB	13.36
8. 7 without dl	15.06

Table 3: BLEU scores of individual systems for Turkish-to-English.

and Lavie, 2010).

Experiment	newstest2016
1. PB Word Rep.	8.34
2. PB Feature Rep.	8.25
3. 1 + CC_LM	8.59
4. 3 + Content Corpus	8.00
5. 3 + Reverse Corpus	7.96
6. 1 with HPB	7.65
7. 2 with HPB	7.57

Table 4: BLEU scores of individual systems for English-to-Turkish.

Table 4 shows the experimental results of the official test set for the English-to-Turkish direction. We observe that the PB system with a word-level representation gives us the best result.

4.6 Post-processing

4.6.1 Turkish Word Generation

When using a feature-based translation model, a word generation step is required to generate the correct Turkish word from the outputs of systems which contain words represented with stems and sequence of morphemes. We used an in-house morphological generation tool that, given a text with words in a format where each morpheme is concatenated to the previous morpheme or stem, transforms these representations to the correct single-word form. This generation tool has been trained on a large Turkish corpus and works by simply creating a reverse-map through morphological segmentation of the corpus. This map contains stem+morpheme sequences as keys and their corresponding single-word forms as values. While creating this map, the disambiguation step of morphological segmentation is omitted to increase the coverage, as keeping multiple resolutions for a single-word form increases the number of keys for the reverse-map. We augmented the map to further increase the coverage.

The following are the working steps of the generation tool:

1. The system outputs and the combined map of ”stem+morphemes to surface form” is taken as input.
2. Iterating through tokens, if an encountered token is:
   (a) a stem; simply output the token.
   (b) a ”stem+morphemes” that is in the map; output its value.
   (c) otherwise; drop the trailing morpheme, and go to 2a.

An example of word generation is as follows:

- **Stem + Morpheme:** git_Aor_A1sg
- **Output Surface Form:** giderim
- **English:** I go

4.6.2 System Combination

System combination attempts to improve the quality of machine translation output by combining the outputs of different translation systems which usually are based on different paradigms such as phrase-based, hierarchical, etc. aiming to exploit and combine strengths of each system. The outputs of some of our translation systems, which are based on different methods as explained in the previous sections, were put into a combination task. We combined the outputs of some of the best performing (best tuning run in terms of BLEU score) hierarchical phrase-based systems using the open-source system combination tool, MEMT. We trained the system combination decoder over different development sets and selected the best ones as our primary submissions to the WMT 2016.

5 Conclusions

This paper described TÜBİTAK's submissions to the WMT'16 news translation task for the Turkish-English language pair. We used Moses in our submissions as well as other open source tools such as MEMT and TreeTagger. For the English-Turkish direction the crawl-data provided was processed and used to generate a 5-gram language model.

Experiment	**newstest2016**
1. HPB Feature Rep.	15.46
2. 1 + Content Corpus	14.94
3. PB Word Rep.	11.20
4. PB Feature Rep. dl -1	15.06
5. sys-comb	16.01

Table 5: BLEU scores of system combinations for Turkish-to-English.

Experiment	**newstest2016**
1. PB Feature Rep.	8.25
2. PB Word Rep.	8.59
3. HPB Word Rep.	7.65
4. 2 + Content Corpus	8.00
5. sys-comb	8.90

Table 6: BLEU scores of system combinations for English-to-Turkish.

A 4-gram gigaword language model for English was used. Due to the morphological discrepancy between the two languages, a morphological analyzer was used to apply full segmentation to the Turkish side. A word-generation tool was used to generate back the word forms of the Turkish sentences from its morphologically analysed counter-parts for English-to-Turkish. We observed that morphological-analysis performed quite well for the Turkish-to-English direction. We experimented with training data with its source side in reverse order and with its content words added to it. Employing system combination of different SMT paradigms resulted in 0.5 BLEU increase for Turkish-to-English and 0.3 BLEU increase for English-to-Turkish.

References

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

Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, and Marcello Federico. 2013. Report on the 10th IWSLT evaluation campaign. *The 10th International Workshop on Spoken Language Translation, Heidelberg, Germany*.

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

Ilknur Durgar El-Kahlout, Coskun Mermer, and Mehmet U. Dogan. 2012. *Recent Improvements In Statistical Machine Translation Between Turkish and English*. Cambridge Scholars Publishing.

Marcello Federico, Mauro Cettolo, Luisa Bentivogli, Michael Paul, and Sebastian Stüker. 2012. Overview of the IWSLT 2012 evaluation campaign. *Proceedings of the International Workshop on Spoken Language Translation, Hong Kong, HK*.

Kenneth Heafield and Alon Lavie. 2010. Combining machine translation output with open source: The Carnegie Mellon multi-engine machine translation scheme. *The Prague Bulletin of Mathematical Linguistics*, 93:27–36.

Phillip Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical phrasebased translation. *Proceedings of the Human Language Technology Conference and the North American Chapter of the Association for Computational Linguistics (HLT/NAACL)*, pages 127–133.

Coskun Mermer and Ahmet A. Akin. 2010. Unsupervised search for the optimal segmentation for statistical machine translation. *Proceedings of the ACL 2010 Student Research Workshop*, pages 31–36.

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

Kemal Oflazer. 1994. Two-level description of Turkish morphology. *Literary and Linguistic Computing*, 9:137–148.

Kemal Oflazer. 2008. Statistical machine translation into a morphologically complex language. *Computational Linguistics and Intelligent Text Processing, 9th International Conference, CICLing*, 4919:376–387.

Michael Paul, Mauro Federico, and Sebastian Stüker. 2010. Overview of the IWSLT 2010 evaluation campaign. *International Workshop on Spoken Language Translation (IWSLT 2010), Paris - France*.

Michael Paul. 2009. Overview of the IWSLT 2009 evaluation campaign. *Proceedings of IWSLT 2009, Tokyo - Japan*.

Nicholas Ruiz, Arianna Bisazza, Roldano Cattoni, and Marcello Federico. 2012. FBK's machine translation systems for IWSLT 2012's TED lectures. *Proceedings of the seventh International Workshop on Spoken Language Translation (IWSLT)*, pages 61–68.

Hasim Sak, Tunga Gungor, and Murat Saraclar. 2007. Morphological disambiguation of Turkish text with perception algorithm. *Proceedings of CICLING*, pages 107–118.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. *Proceedings of the International Conference on New Methods in Language Modeling*.

Andreas Stolcke. 2002. SRILM an extensible language modeling toolkit. *Proceedings of ICSLP*.

Reyyan Yeniterzi and Kemal Oflazer. 2010. Statistical phrasebased translation. *Syntax-to-morphology mapping in factored phrase-based statistical machine translation from English to Turkish," in Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL '10. Stroudsburg, PA, USA*, pages 454–464.

Ertugrul Yilmaz, Ilknur Durgar El-Kahlout, Zisan S. Ozil, and Coskun Mermer. 2013. Tubitak Turkish-English submissions for IWSLT 2013. *Proceedings of the 10th International Workshop on Spoken Language Translation*, pages 152–159.