TÜBİTAK TURKISH-ENGLISH SUBMISSIONS for IWSLT 2013

Ertuğrul Yılmaz, İlknur Durgar El-Kahlout, Burak Aydın,
Zişan Sıla Özil, Coşkun Mermer

TÜBİTAK-BİLGEM
Gebze 41470, KOCAELİ, TURKEY
{yilmaz.ertugrul,ilknur.durgar,burak.aydin,zisan.ozil,coskun.mermer}@tubitak.gov.tr

Abstract

This paper describes the TÜBİTAK Turkish-English submissions in both directions for the IWSLT'13 Evaluation Campaign TED Machine Translation (MT) track. We develop both phrase-based and hierarchical phrase-based statistical machine translation (SMT) systems based on Turkish word- and morpheme-level representations. We augment training data with content words extracted from itself and experiment with reverse word order for source languages. For the Turkish-to-English direction, we use Gigaword corpus as an additional language model with the training data. For the English-to-Turkish direction, we implemented a wide coverage Turkish word generator to generate words from the stem and morpheme sequences. Finally, we perform system combination of the different systems produced with different word alignments.

1. Introduction

We participated in the IWSLT Evaluation Campaign for the Turkish-English MT track for both directions. The typological, morphological and word order differences of this language pair make the implementation of SMT systems a challenging task. English and Turkish are typologically rather distant languages. English has a very limited morphology and rather fixed Subject-Verb-Object (SVO) constituent order. However, Turkish is an agglutinative language with very flexible (but Subject-Object-Verb (SOV) dominant) constituent order, and has a very rich and productive derivational and inflectional morphology where word structures might correspond to complete phrases of several words in English when translated.

Overview of the systems can be summarized as follows: (1) We used the feature-based representation of Turkish in order to aggregate the statistics from different morphological forms of the words in addition to the word representation as in [1], (2) We compared phrase-based SMT systems with hierarchical phrase-based systems, (3) We augmented the training data with the content words extracted from itself to bias the stem word alignments as in [2], (4) We used reverse word order of the source language in order to obtain alternative translations similar to [3], (5) We preferred to use WIT corpus for the translation model training, (6) We combined both SETIMES and WIT corpora as one language model for English-to-Turkish systems, (7) We implemented a wide-coverage Turkish morphological word generator to generate Turkish words from stem and morpheme sequences, (8) We added Gigaword corpus as an additional language model for Turkish-to-English systems, (9) We combined systems (2), (3) and (4) to improve the translation quality from different word alignments. As a result, we improved $+1.4$ BLEU points on *test2011* and $+1.5$ BLEU points on *test2012* compared to the best system of IWSLT'12 Turkish-to-English MT track.

This paper is organized as follows: Section 2 introduces the challenges of implementation of SMT systems for the Turkish-English language pair and summarizes the previous work. Section 3 describes the data sets and explains the experimental setups. Section 4 shows the experimental results in both directions and reports the official submission scores. We conclude with Section 5.

2. Turkish-English Statistical Machine Translation

Turkish exhibits interesting properties from an SMT point of view. Its agglutinative structure has very productive inflectional and derivational processes for word formation. Words are created by concatenating morphemes to the stem word or to other morphemes. Generally, word formation is done by suffixation. Except for very few cases, surface realizations of the morphemes are conditioned by various regular morphophonemic processes such as vowel harmony, consonant assimilation, and elisions. The morphotactics of word forms could be quite complex when multiple derivations are involved. The average number of bound morphemes (i.e., excluding the stem) in words is about two. The productive morphology of Turkish potentially implies a very large vocabulary size. In most cases, single Turkish words typically tend to align with whole phrases on the English side when sentence pairs are aligned at the word level.

During the development of SMT systems, morphological preprocessing is useful and sometimes crucial when at least one of the languages is morphologically complex. Turkish is one of the languages that need special attention as

several derivational and inflectional processes can produce very complex Turkish words. Mapping the rich morphology of Turkish to the limited morphology of English has been addressed by several researchers. To reduce the large vocabulary size and to force more one-to-one word alignments, researchers prefer a sub-word representation of the foreign language while translating to/from English. The research showed that replacing the Turkish word representation with a sub-word representation performs better in the translation process in both directions. [1, 2, 4] 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, [5] applied syntactic transformations such as joining function words on the English side to the related content words. [6] used an unsupervised learning algorithm to find the segmentations automatically from parallel data. [7] presented a series of segmentation schemes to explore the optimal segmentation for statistical machine translation of Turkish. In addition, an important amount of effort was spent by several research groups on Turkish-to-English SMT in the IWSLT'09[1] BTEC task, IWSLT'10[2] and IWSLT'12[3] TED tasks.

3. Experiments

In the experiments, we used all supplied monolingual and parallel texts for the system development. We tuned systems with *dev2010* and used *test2010* as the internal test set. In terms of official evaluation of the translation systems, we submitted the last two years' test sets *test2011*, *test2012* and a new test set *test2013*. As we noticed that some portions of the Turkish texts in WIT corpus [8] are asciified, we employed a deasciifier tool[4] to clean these portions of the data.

Tables 1 and 2 show the Turkish (with word and morpheme-based representations) and English statistics after the pre-cleaning step. One can notice that with full segmentation, the number of unique words of Turkish and English are closer than the word representation. As the vocabulary sizes of the languages get closer, we expect better word alignments.

Table 1: WIT training data statistics

	Sentences	Unique words	Total words
Turkish (Word)	130K	158K	1.8M
Turkish (Feature)		35K	2.9M
English		45K	2.5M

[1] www2.nict.go.jp/univ-com/multi_trans/WS/IWSLT2009
[2] iwslt2010.fbk.eu
[3] hltc.cs.ust.hk/iwslt
[4] turkce-karakter.appspot.com

Table 2: SETIMES training data statistics

	Sentences	Unique words	Total words
Turkish (Word)	173K	143K	3.9M
Turkish (Feature)		43K	5.5M
English		60K	4.6M

3.1. Phrase-based vs. Hierarchical Phrase-based Systems

Although phrase-to-phrase translation [9] overcomes many problems of word-to-word translation [10] and has been successful for some language pairs during the last decade, the continuity of phrases is its main shortcoming. Clearly, this is a problem for language pairs with very different word orders such as Chinese-English. For that kind of language pairs, 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 [11] 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 [1, 7] showed that using hierarchical phrase-based (HPB) decoder outperforms the phrase-based (PB) systems for Turkish-English.

For this reason, we performed experiments mainly with HPB decoders but also implemented systems with PB decoders in order to use the output of the PB systems in the internal system combination.

3.2. Sub-word Representation

We implemented the baseline experiments with the word-level representation of Turkish. As mentioned in Section 2, incorporating morphology when working with morphologically rich(er) languages in SMT performs better than the word-level. For this reason, in the further experiments, we preferred using a feature-based representation of Turkish in both directions as this representation dramatically reduces the vocabulary size on the Turkish side as shown in Tables 1 and 2. To produce the feature-based word representation, we first pass each word through a morphological analyzer [12]. The output of the analyzer contains the morphological features encoded for all possible analyses and interpretations of the word. Then we perform morphological disambiguation on the morphological analyses [13]. Once the contextually-salient morphological interpretation is selected, we remove 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 remain 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:** Organize edeceğim , yöneteceğim ve onu dünyaya sunacağım .
- **Full segmentation:** Organize et _Fut _A1sg , yönet _Fut _A1sg ve o _Acc dünya _Dat sun _Fut _A1sg .
- **Reference:** I'm going to organize it and direct it and get it going in the world .

3.3. Content Words

From the morphologically segmented corpora, we also extract for each sentence in the training corpus, the sequence of stems for open-class content words (Nouns, Adjectives, Adverbs, and Verbs). For Turkish, this corresponds to removing *all* morphemes and any stems for closed classes.

For English, we used the TreeTagger [14] to tag the sentences and removed all words tagged as closed class words along with the tags such as *+VVG* that signal a morpheme on an open-class content word. We use this data to augment the training corpus and bias content word alignments, with the hope that such stems may get a better chance to align without any additional "noise" from morphemes and other function words. An example of a content word (bold) sentence pair of is as follows:

- **Turkish content words: Organize et** _Fut _A1sg , **yönet** _Fut _A1sg ve o _Acc **dünya** _Dat **sun** _Fut _A1sg .
- **English content words:** I +PP am +VB **go** +VVG to +TO **organize** +VV it +PP and +CC **direct** +VV it +PP and +CC **get** +VV it +PP **go** +VVG in +IN the +DT **world** +NN . +SENT

Table 3 shows the Turkish and English content word corpus statistics after the pre-cleaning step.

Table 3: WIT content word statistics

	Sentences	Unique words	Total words
Turkish	128K	45K	1.1M
English	128K	39K	1M

3.4. Reverse Translation

Word order differences affect many steps of the translation process such as word alignment, phrase extraction, and thus the translation quality. It has been observed that one gets better alignments and hence better translation results when the word orders of the source and target languages are more or less the same. When word orders are different, it can be useful to systematically reorder the tokens of source sentences to an order matching or very close to the target language word order so that alignments could be very close to a monotonic one. Thus instead of forcing the decoders to employ reordering schemes, the source sentences are similarly reordered and then decoded with the decoder employing more simple reordering models. As the word orders of Turkish (SOV) and English (SVO) differ, reordering of the source sentence may allow to produce an alternative translation table thus alternative translation performance. In order to make the word orders especially *Verbs* a bit closer, one approach can be to use the reverse order of the source side of the language pair. In these experiments, we reversed the order of the source language similar to [3] before the word alignment step as generally reordering target language is not preferred because of the need of an additional post-processing. Reverse sentence examples of the source language for two translation directions are as follows:

- **Turkish reverse sentence:** . _A1sg _Fut sun _Dat dünya _Acc o ve _A1sg _Fut yönet , _A1sg _Fut et Organize
- **English reverse sentence:** . world the in going it get and it direct and it organize to going I'm

4. Results

For the IWSLT'13 Evaluation Campaign, we performed several SMT experiments for Turkish-English with different settings. All available data was tokenized with an in-house Turkish tokenizer and then truecased. We generated word alignments using MGIZA [15] with default settings. We implemented both the phrase-based and the hierarchical phrase-based systems with Moses Open Source Toolkit [16]. The system parameters were optimized with the minimum error rate training (MERT) algorithm [17] on the tune set *dev2010*, evaluated on the test set *test2010*, and scores are reported in terms of BLEU [18]. We trained conventional 5-gram language models (LMs) from the parallel corpus for both directions but also performed tests with 4-gram Gigaword language model for the Turkish-to-English systems. All language models were trained with the SRILM toolkit [19] using modified Kneser-Ney smoothing [20] and then binarized using Kenlm [21] .

For phrase-based systems, we allowed unlimited jumps for word reordering (*distortion-limit* $= -1$). At each step, systems were tuned with five different seeds with lattice-samples and minimum Bayes risk decoding; *-mbr* [22] is employed during the decoding.

For hierarchical phrase-based systems, we relaxed the rule table extraction by allowing sub-phrases of any size to be replaced by a non-terminal(*-MinHoleSource* $= -1$), and set *-cube-pruning-pop-limit* to 5000 to increase the number of hypotheses created in each span.

4.1. Turkish-to-English

The baseline experiment was conducted with the hierarchical phrase-based system and Turkish word representation (Exp. #1), then we employed the morpheme-based representation as explained in Section 3.2 which results in an improvement of $+2.5$ BLEU points (Exp. #2). We experimented to remove the out-of-domain data *SETIMES* corpus from the training that gave us a $+1.1$ BLEU point increase (Exp. #3). Further, including the 4-gram Gigaword corpus as an additional language model improved the performance of the system by 1.1 BLEU points (Exp. #4). We performed two more experiments with augmenting the corpus with content words (Exp. #5) and using the reverse word order on the source side (Exp. #6) which resulted in a -0.4 and -1.0 BLEU points decrease, respectively. We also repeated the experiments 4, 5, and 6 with the phrase-based framework.

Table 4: Turkish-to-English BLEU scores

System	dev2010	test2010
1. HPB - Word Rep.	11.31	12.47
2. HPB - Feature Rep.	13.54	14.96
3. 2 + WIT only	14.00	16.10
4. 3 + Gigaword	15.33	17.14
5. 4 + Content Corpus	14.80	16.68
6. HPB Reverse Corpus	14.18	16.18
7. 4 with PB	13.22	15.69
8. 5 with PB	13.53	15.95
9. 6 with PB	13.00	14.77

Table 4 shows the experimental results on the development and test sets. All of the experiments run with five tuning seeds and the one with the maximum score is selected after each step. We observed that the reported improvements are consistent in all tuning runs[5]. Although not reported here, using Turkish-specific tokenizer improved the performance by $+0.3$ BLEU points. As expected, the HPB systems outperform the PB systems by approximately 1.5 BLEU points. Adding content word corpus degraded the performance in the HPB framework but induced a slight improvement (0.3 BLEU points) in the PB systems. Experiments showed that using out-of-domain data without performing any domain adaptation method hurts the performance of the systems. Reverse word order in the source language is slightly worse than the exact word order individually but this system can be used as a candidate in the system combination which will be explained later. We also performed experiments with combined language model where SETIMES and WIT corpora are concatenated and trained together but observed a decrease of 0.2 BLEU points.

[5]Variation between tunes are approximately 0.4 BLEU points

4.2. English-to-Turkish

In this case, the target language is the morphologically-complex Turkish. This presents a challenge in predicting the correct word forms (or their morphological composition) using a sparser target language model data. In the morpheme-based system, there is a need for a word-generation tool that generates Turkish words from stem and morpheme sequences. The performance of this tool will directly affect the translation quality of the morpheme-based system. The challenge in generating Turkish word forms is that Turkish word features can be mapped to several suffixes and each combination leads to a different Turkish word. Moreover, during the generation process the vowel harmony should be taken into consideration.

Most of the experiments of Section 4.1 were repeated for the English-to-Turkish direction. Similar to the Turkish-to-English experiments, the baseline experiment was conducted with the hierarchical phrase-based system using Turkish word representation (Exp. #1), then we experimented with Turkish morpheme-based representation which results in an improvement of $+0.6$ BLEU points (Exp. #2). We also removed the out-of-domain data *SETIMES* corpus from the training, which resulted in an increase of $+0.1$ BLEU points (Exp. #3). We performed experiments with the combined language model which induced a $+0.1$ BLEU improvement (Exp. #4). Above that, we performed experiments by augmenting the corpus with content words (Exp. #5) and using the reverse word order on the source side (Exp. #6) which resulted in a -0.3 and -0.4 BLEU points decrease, respectively. Again, we also repeated the experiments 4, 5, and 6 with the phrase-based framework.

Table 5: English-to-Turkish BLEU scores

System	dev2010	test2010
1. HPB - Word Rep.	6.11	7.70
2. HPB - Feature Rep.	7.14	8.31
3. 2 + WIT only	6.34	8.41
4. 3 + Combined LM	6.07	8.52
5. 3 + Content Corpus	6.54	8.24
6. HPB Reverse Corpus	5.99	8.13
7. 4 with PB	4.91	7.40
8. 5 with PB	4.91	7.23
9. 6 with PB	4.32	6.83

Table 5 shows the experimental results on the development and test set. Similar to Turkish-to-English direction, the HPB systems outperform the PB systems by approximately 1.1 BLEU points. Adding content word corpus and reverse word order hurts the performance in both HPB and PB systems but they were kept for the system combination. Employing combined language model increased the system performance in the test set contrary to the Turkish-to-English experiments.

Table 6: The word statistics of morphological generation for outputs of Exp. #4. (#stems: words with no morphemes, hence no word generation is required, #sequences: words of the form stem+morphemes, found: sequence words for which an exact single-word-form is found; sub-found: sequence words resolved after elimination of some trailing morphemes)

	total	**#stems**	**#sequences**	**found (%)**	**sub-found (%)**	**missed (%)**
test2010	23056	13604	9452	9065 (95.9%)	167 (1.8%)	220 (2.3%)
test2011	19447	11312	8135	7793 (95.8%)	124 (1.5%)	218 (2.7%)
test2012	22021	12609	9352	8878 (94.9%)	174 (1.9%)	300 (3.2%)
test2013	16410	9414	6996	6643 (95.9%)	132 (1.9%)	221 (3.2%)

4.2.1. Turkish Word Generation

In morpheme-based translation, a word generation tool 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 by 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. An additional map is generated through morphological segmentation of WIT and SETIMES corpora to further increase coverage. These two maps are combined giving the preference to the latter map in case of disagreements.

The following are the working steps of the generation tool:

1. The system outputs and the combined map of 'stem+morphemes to single-word form' is taken.
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.

Examples of word generation are as follows:

- **Stem+Morpheme Sequence:** et_Aor_A1sg
  Surface Form: ederim[6]
- **Stem+Morpheme Sequence:** duy_PastPart_P3sg
  Surface Form: duydu[7]

[6] I do it
[7] he/she/it heard

Step 2c in this procedure can help in cases where an extraneous morpheme is found at the end of a word, which in turn would increase the coverage of the generator. Table 6 shows the coverage of the word generator for outputs of (Exp. #4) for all the test data. For about 95% of the tokens of the form stem+morpheme sequences, the procedure finds an exact single-word form match. An additional 1-2% match is achieved by following the process of dropping the trailing morpheme and re-checking the map for the resulting sequence. For 2% to 3% of the words of the form stem+morphemes, all morphemes are eliminated and only the stem is represented in the output (missed).

4.3. 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 [23]. We also experimented with adding phrase-based systems to the combination task but did not observe any improvements, hence we provide results for combination of different hierachical phrase-based systems.

MEMT should ideally be tuned by a separate held-out data that is different from system training and tuning data. As we did not have additional tuning data for system combination tuning, we primarily used *dev2010* data to tune the system combination decoder. To see how having separate tuning data for system combination would have effected the quality of the combined system outputs, we trained the system combination decoder with *test2010* data evaluating the performance on *test2011*, *test2012*, and *test2013* data (not tested for *test2010* as it would not be valid). Tuning the system combination decoder with *test2010* data yielded comparable results with the system tuned by *dev2010* data. Also, tuning with the combination of *dev2010* and *test2010* data yielded similar results. The results we provide in this paper are for system combination tasks that employed either only

dev2010, or both *dev2010* and *test2010* data as tuning data.

The language models used for system combination training and decoding were i) a 4-gram language model constructed from the Gigaword database for Turkish-English translations, and ii) a 5-gram language model constructed from the combination of WIT and SETIMES corpora for English-Turkish translations.

Table 7: BLEU scores of individual systems and their system combinations for English-to-Turkish. (*Exp. #3 with different tuning seed)

Experiment	test2010	test2011	test2012	test2013
3*	8.84	8.85	8.81	8.50
4	8.52	8.86	9.20	8.65
5	8.24	8.74	8.70	8.08
Sys. Combo.	8.82	**9.16**	**9.29**	**8.97**
6	8.13	7.99	8.57	8.05
Sys. Combo.	8.77	**9.34**	**9.48**	**8.86**

Table 7 shows the BLEU scores of some individual systems as well as the the BLEU score of their combined outputs for English-to-Turkish translations. Combining the outputs of experiments 3, 4, and 5 yields about the same BLEU score for *test2010* and better BLEU scores for test sets 2011, 2012, and 2013 compared to the best individual system. Combination of the outputs of those three systems achieves a relative BLEU improvement of about 3.5%, 0.98%, and 3.7% over the best performing individual systems for test sets 2011, 2012, and 2013, respectively. Integration of a fourth system, experiment 6, to the combination task provides better improvements for *test2011*(5.5%) and *test2012*(3.0%) data, but yields a lower score for *test2010* and *test2013* data compared to the combination of 3, 4, and 5. The official results we submitted to IWSLT'13 are the combined outputs of systems 3, 4, and 5. For the submitted combined outputs, the improvements over the best performing individual systems for *test2011* and *test2013* were computed to be statistically significant ($p<0.05$).

Table 8 shows the BLEU scores of individual systems and combined systems for the opposite translation direction, Turkish to English. Using only *dev2010* data for system combination decoder tuning (Sys. Combo. (dev only)), the combined system outputs in this direction provided about 1.17% improvements for both *test2010* and *test2012* over the best performing individual systems, and no improvements for *test2011* and *test2013* data. Adding *test2010* data into the tuning of the system combination decoder (Sys. Combo. (dev+test)) provided some improvement for *test2011* and *test2013* over (Sys. Combo (dev only)). The combined systems -compared to the best individual system- provided statistically significant improvements for this translation direction for *test2010* and *test2012* data ($p<0.05$), and performed worse or about the same for *test2011* and *test2013* data.

Our official submissions for English-to-Turkish and Turkish-to-English are the fourth rows of Tables 7 and 8.

5. Conclusions

This paper presented the experiments and results of the TÜBİTAK Turkish-English submissions in both directions for the IWSLT'13 Evaluation Campaign TED Machine Translation (MT) track. Due to the rich morphological and syntactic properties of Turkish, statistical machine translation involving Turkish implies processes that are more complex than standard statistical translation models.

In our implemented systems, we improved from 12.47 to 17.34 BLEU points in Turkish-to-English SMT systems and 7.70 to 8.82 in English-to-Turkish SMT systems on the *test2010* set. For Turkish-to-English, we improved $+1.4$ BLEU points on *test2011* and $+1.5$ BLEU points on *test2012* compared to the best system of IWSLT'12 Turkish-to-English MT track. Major results of our work can be summarized as follows:

- We compared the feature-based and word representation of Turkish,
- We compared phrase-based SMT systems with hierarchical phrase-based systems,
- We augmented the training data with the content words extracted from itself,
- We used reverse word order of the source language in order to obtain alternative translations,
- We preferred to use WIT corpus for the training,
- We added Gigaword corpus as an additional language model for Turkish-to-English systems,
- We combined both SETIMES and WIT corpora as one language model for English-to-Turkish systems,
- We implemented a wide-coverage Turkish morphological word generator to generate Turkish words from stem and morpheme sequences,
- We applied system combination to hierarchical phrase-based systems that are trained on different representations of the training corpora.

6. References

[1] I. D. El-Kahlout, C. Mermer, and M. U. Doğan, "Recent Improvements In Statistical Machine Translation Between Turkish and English," in *Multilingual Processing in Eastern and Southern EU Languages.- Low-resourced Technologies and Translation*. Cambridge: Cambridge Scholars Publishing, 2012. ??-??. Print.

[2] I. Durgar El-Kahlout and K. Oflazer, "Exploiting morphology and local word reordering in English-to-Turkish phrase-based statistical machine translation,"

Table 8: BLEU scores of individual systems and their system combinations for Turkish-to-English. (*Exp. #5 with different tuning seed)

Experiment	test2010	test2011	test2012	test2013
4	17.14	18.77	18.62	18.88
5*	16.59	18.29	18.53	18.40
6	16.18	17.76	17.61	17.59
Sys. Combo. (dev only)	**17.34**	18.63	**18.93**	18.67
Sys. Combo. (dev+test)	N/A	**18.83**	**18.84**	18.70

IEEE Transactions on Audio, Speech, and Language Processing, vol. 18, no. 6, pp. 1313–1322, 2010.

[3] M. Huck, S. Peitz, M. Freitag, M. Nuhn, and H. Ney, "The RWTH Aachen machine translation system for WMT 2012," in *Proceedings of the Seventh Workshop on Statistical Machine Translation (WMT)*, 2012, pp. 304–311.

[4] K. Oflazer, "Statistical machine translation into a morphologically complex language," in *Computational Linguistics and Intelligent Text Processing, 9th International Conference, CICLing 2008, Haifa, Israel*, ser. Lecture Notes in Computer Science, vol. 4919, 2008, pp. 376–387.

[5] R. Yeniterzi and K. Oflazer, "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*, ser. ACL '10. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 454–464. [Online]. Available: http://dl.acm.org/citation.cfm?id=1858681.1858728

[6] C. Mermer and A. A. Akin, "Unsupervised search for the optimal segmentation for statistical machine translation," in *Proceedings of the ACL 2010 Student Research Workshop*, ser. ACLstudent '10. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 31–36. [Online]. Available: http://dl.acm.org/citation.cfm?id=1858913.1858919

[7] N. Ruiz, A. Bisazza, R. Cattoni, and M. Federico, "FBK's machine translation systems for IWSLT 2012's TED lectures," in *Proceedings of the seventh International Workshop on Spoken Language Translation (IWSLT)*, 2012, pp. 61–68.

[8] M. Cettolo, C. Girardi, and M. Federico, "WIT3: Web inventory of transcribed and translated talks," in *Proceedings of EAMT*, 2012, pp. 261–268.

[9] P. Koehn, F. J. Och, and D. Marcu, "Statistical phrase-based translation," in *Proceedings of the Human Language Technology Conference and the North American Chapter of the Association for Computational Linguistics (HLT/NAACL)*, 2003, pp. 127–133.

[10] P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer, "The mathematics of statistical machine translation: Parameter estimation," *Computational Linguistics*, vol. 19, pp. 263–311, 1993.

[11] D. Chiang, "Hierarchical phrase-based translation," *Computational Linguistics*, vol. 33, no. 2, pp. 201–228, 2007.

[12] K. Oflazer, "Two-level description of Turkish morphology," *Literary and Linguistic Computing*, vol. 9, pp. 137–148, 1994.

[13] H. Sak, T. Güngör, and M. Saraçlar, "Morphological disambiguation of Turkish text with perception algorithm," in *Proceeding of CICLING, LNCS 4394*, 2007, pp. 107–118.

[14] H. Schmid, "Probabilistic part-of-speech tagging using decision trees," in *Proceedings of the International Conference on New Methods in Language Processing*, 1994.

[15] Q. Gao and S. Vogel, "Parallel implementations of word alignment tool," in *Proceedings of ACL WSETQANLP*, 2008, pp. 49–57.

[16] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, "Moses: Open source toolkit for statistical machine translation," in *Proceedings of ACL Demo and Poster Session*, 2007, pp. 177–180.

[17] F. J. Och, "Minimum error rate training in statistical machine translation," in *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, 2003, pp. 160–167.

[18] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, 2002, pp. 311–318.

[19] A. Stolcke, "SRILM - an extensible language modeling toolkit," in *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, 2002, pp. 901–904.

[20] R. Kneser and H. Ney, "Improved backing-off for m-gram language modeling," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, 1995, pp. 181–184.

[21] K. Heafield, "Kenlm: Faster and smaller language model queries," in *Proceedings of the Sixth Workshop on Statistical Machine Translation*, 2011, pp. 187–197.

[22] S. Kumar and W. Byrne, "Minimum Bayes-risk decoding for statistical machine translation," in *Proceedings of HLT-NAACL*, 2004, pp. 169–176.

[23] K. Heafield and A. Lavie, "Combining machine translation output with open source: The Carnegie Mellon multi-engine machine translation scheme," *The Prague Bulletin of Mathematical Linguistics*, vol. 93, pp. 27–36, 2010.