Unsupervised Turkish Morphological Segmentation for Statistical Machine Translation

Coşkun Mermer[1,2]
[1]TÜBİTAK-UEKAE
Gebze, Kocaeli, Turkey
coskun@uekae.tubitak.gov.tr

Murat Saraçlar[2]
[2]Boğaziçi University
Bebek, Istanbul, Turkey
murat.saraclar@boun.edu.tr

We present our investigations on unsupervised Turkish morphological segmentation (TMS) for statistical machine translation (SMT), which has not been addressed in previous work (Table 1). We perform *in vivo* testing of the TMS performance in a phrase-based SMT system developed for the Turkish-English task at IWSLT[1]. We compare unsupervised segmentation against two baselines: (1) no segmentation, i.e., word-based translation, (2) a state-of-the-art supervised segmentation comprising morphological analysis + disambiguation + manually-crafted rules (Mermer et al., 2009) that performed very well in IWSLT 2010.

We set out with an existing unsupervised segmentation tool, Morfessor (Creutz and Lagus, 2007). The original search algorithm of Morfessor aims to satisfy the MAP objective by greedily searching for the segmentation that results in the highest posterior probability according to the generative model. However, the greedy search in the high-dimensional combinatorial search space often gets stuck in local optima. We instead propose to approximate the posterior distribution of segmentations via Gibbs sampling. We decide the segmentation location for a word by drawing a sample from the distribution proportional to the posterior probability of the model given the existing state of segmentation for the rest of the words. We ran the Gibbs sampler for 2000 iterations over the (dynamic) sub-word vocabulary. Table 2 shows that Gibbs sampling is able to find better segmentations in terms of model scores (previously unattainable in greedy search). However, this search improvement does not translate over to the translation performance (Table 3). This suggests a model mismatch, which can be expected in this case since the segmentation model uses only monolingual observations.

Hence we extend the generative model to incorporate both sides of the parallel corpus via translation from hidden segmentations: $P(e, f, M_f) = \sum_{f_{seg}} P(M_f)P(f_{seg}|M_f)P(f|f_{seg})P(e|f_{seg})$. Here, e and f are the two sides of the parallel corpus and M_f is the segmentation model for f that results in the segmentation f_{seg}. Note that $P(f|f_{seg})$ is either 1 or 0 indicating legal segmentations of f. In searching for the MAP segmentation model M_f^*, we approximate the summation with the $\max$ operation.

We model the first two components as in the monolingual case while for the translation component $P(e|f)$ we use IBM Model 1. To cope with the increased computational load, we propose a search algorithm that enables parallel computation instead of the original sequential search. We also devise a method of computing approximate IBM Model-1 translation likelihood incrementally from an adjacent segmentation state to speed-up the computation.

Preliminary results show that the BLEU scores

Previous work	A	B	C	D
(Nguyen et al., 2010; Xu et al., 2008)		√	√	√
(Poon et al., 2009)		√	√	
(Luong et al., 2010)	F	√	√	
(Durgar El-Kahlout and Oflazer, 2010; Bisazza and Federico, 2009)	T	√		
(Creutz and Lagus, 2007; Arısoy, 2009)	T,F		√	
This work	**A**	**B**	**C**	**D**
(Mermer et al., 2010)	T	√	√	
(Mermer and Akın, 2010)	T	√	√	√

Table 1: Morphological segmentation literature relevant to this work. Features: A: Tested on agglutinative languages (F: Finnish, T: Turkish), B: Tested on SMT, C: Unsupervised, D: SMT-guided segmentation learning.

[1]International Workshop on Spoken Language Translation. http://iwslt2010.fbk.eu

Search	Model score
Original	1559831
	1559315
	1559527
Gibbs	1554433

Table 2: Segmentation model scores (in negative log probability) reached after segmentation training: (Top) Original search with three different random vocabulary scan orders. (Bottom) 2000 iterations over the vocabulary via Gibbs sampling.

Dataset	Original search	Gibbs sampling
Tuning (dev1)	0.5941	0.5909
Test (dev2)	0.5442	**0.5455**
IWSLT09 test	**0.5215**	0.5190
IWSLT10 test	**0.4983**	0.4860

Table 3: BLEU scores in IWSLT 2010 with different segmentation search strategies.

of the bilingually-informed segmentation (currently implemented with the original greedy search) are similar to monolingual segmentation (Table 4). However, the correlation between the BLEU scores and the segmentation model scores is higher in the bilingual case than in the monolingual case (Mermer and Akın, 2010). Therefore, we are hopeful that better search, e.g., via Gibbs sampling, this time improves the translation performance now that we expect the model to be more suited towards translation. This line of research as well as improving the model (e.g., incorporating the HMM morpheme generation model of Morfessor Categories-MAP (Creutz and Lagus, 2007)) and testing the segmentations on more data sets and other morphologically-rich languages constitute our next steps.

Overall, experimental results show that while unsupervised segmentation improves translation BLEU scores over the word-based baseline for this task, it does not (yet) reach the performance of task-optimized supervised segmentation (Table 5). Even though up to now we have tested our results on Turkish, the applied methods are entirely language-independent (save affixation) and we expect them to be applicable particularly to other agglutinative languages as well.

Segmentation	BLEU
None	0.5204
Monolingual	**0.5273**
Bilingual	0.5271

Table 4: BLEU scores on 1512-sentence BTEC test set averaged over multiple searches with different random vocabulary scans ("Monolingual" also utilizes parallel search, since it gives higher BLEU scores).

Method	Tune	Test	2009	2010
A	0.5665	0.5140	0.4948	0.4749
B	0.5941	0.5442	0.5215	0.4983
C	0.6269	0.5478	0.5303	0.5091
D	0.6462	**0.5946**	**0.5640**	**0.5332**

Table 5: BLEU scores of different segmentation methods in IWSLT 2010 (Mermer et al., 2010). A: Word-based, B: Morfessor, C: Morfessor Categories-MAP, D: Morphological analyzer (Oflazer, 1994) + postprocessing.

References

E. Arısoy. 2009. *Statistical and Discriminative Language Modeling for Turkish Large Vocabulary Continuous Speech Recognition*. Ph.D. thesis, Boğaziçi University.

A. Bisazza and M. Federico. 2009. Morphological Pre-Processing for Turkish to English Statistical Machine Translation. In *IWSLT*, pages 129–135, Tokyo, Japan.

M. Creutz and K. Lagus. 2007. Unsupervised models for morpheme segmentation and morphology learning. *ACM Trans. Sp. and Lang. Proc.*, 4(1):1–34.

I. Durgar El-Kahlout and K. Oflazer. 2010. Exploiting morphology and local word reordering in english-to-turkish phrase-based statistical machine translation. *IEEE Trans. Aud., Sp. and Lang. Proc.*, 18(6):1313–1322, August.

M.-T. Luong, P. Nakov, and M.-Y. Kan. 2010. A hybrid morpheme-word representation for machine translation of morphologically rich languages. In *EMNLP*, pages 148–157, Cambridge, MA, October.

C. Mermer and A.A. Akın. 2010. Unsupervised search for the optimal segmentation for statistical machine translation. In *ACL 2010 Student Research Workshop*, pages 31–36, Uppsala, Sweden, July.

C. Mermer, H. Kaya, and M.U. Doğan. 2009. The TUBITAK-UEKAE statistical machine translation system for IWSLT 2009. In *IWSLT*, Tokyo, Japan.

C. Mermer, H. Kaya, and M.U. Doğan. 2010. The TUBITAK-UEKAE statistical machine translation system for IWSLT 2010. In *IWSLT (to appear)*.

T.L. Nguyen, S. Vogel, and N.A. Smith. 2010. Nonparametric word segmentation for machine translation. In *COLING*, Beijing, China, August.

K. Oflazer. 1994. Two-level description of Turkish morphology. *Literary and Linguistic Computing*, 9(2).

H. Poon, C. Cherry, and K. Toutanova. 2009. Unsupervised morphological segmentation with log-linear models. In *HLT-NAACL*, pages 209–217, Boulder, Colorado.

J. Xu, J. Gao, K. Toutanova, and H. Ney. 2008. Bayesian semi-supervised chinese word segmentation for statistical machine translation. In *COLING*, Manchester, UK.