INITIAL EXPLORATIONS IN KAZAKH TO ENGLISH STATISTICAL MACHINE TRANSLATION

Zh. Assylbekov*, A. Nurkas

School of Science and Technology, Nazarbayev University, Astana, Kazakhstan; *zhassylbekov@nu.edu.kz

Introduction. The availability of considerable amounts of parallel texts in Kazakh and English has motivated us to apply statistical machine translation (SMT) paradigm for building a Kazakh-to-English machine translation system using publicly available data and open-source tools.

Corpus preparation. We mined around 20,000 Kazakh-English parallel sentences from the official website of the President of the Republic of Kazakhstan located at http://akorda.kz.

Morphological segmentation schemes. We performed morphological analysis for Kazakh side of our corpora using an open-source finite-state morphological transducer *apertium-kaz* [1]. After that simple regular expressions were used to describe different segmentation rules which are combinations of splitting and removal of tags from the analyzed lexical forms. Overall we developed 7 schemes: MS2, MS6, MS7, MS11, MS11a, MS12, MS13 [2]. The benefit of segmentation for word alignment in Kazakh-to-English direction is shown in Figure 1.

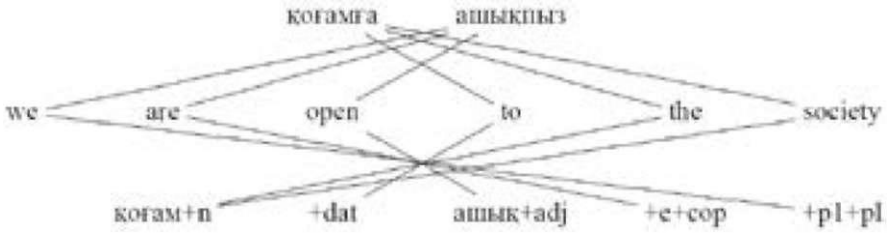

Figure 1. Word alignment before (up) and after (down) morphological segmentation

Experiments. The open-source SMT toolkit Moses [3] was used to build the baseline system. Since the number of words in each sentence has grown on average after segmentation, we allowed the distortion (DL) to be unlimited.

Results and discussion. Table 1 shows how morphological preprocessing and unlimited distortion affects translation performance. The experiments have shown that a selective morphological segmentation improves the performance of an SMT system for morphologically rich languages such as Kazakh.

Scheme	DL=6	DL=∞
Baseline	22.75	23.70
MS2	23.77	25.23
MS6	23.77	25.06
MS7	23.90	25.41
MS11	23.62	25.21
MS11a	23.95	25.30
MS12	23.82	25.18
MS13	24.05	25.46

Table 1: BLEU scores

Acknowledgments. We would like to thank Dr. Francis Morton Tyers and Jonathan North Washington for their constant attention to this work. This research was financially supported by the grant of the Corporate Fund "Fund of Social Development".

References.

1. J. N. Washington, et al. Finite-state morphological transducers for three Kypchak languages. *Proceedings of LREC 2014,* 3378-3385.
2. Zh. Assylbekov and A. Nurkas. Initial Explorations in Kazakh to English Statistical Machine Translation. *Proceedings of CLiC-it 2014,* accepted.
3. Ph. Koehn, et al. Moses: Open source toolkit for statistical machine translation. *Proceedings of the ACL 2007. Demo and poster sessions,* Prague, Czech Repuplic, 177-180.