10.12871/CLICIT201413

Initial Explorations in Kazakh to English Statistical Machine Translation

Zhenisbek Assylbekov, Assulan Nurkas
School of Science and Technology
Nazarbayev University
53 Kabanbay batyr ave., Astana, Kazakhstan
{zhassylbekov, anurkas}@nu.edu.kz

Abstract

English. This paper presents preliminary results of developing a statistical machine translation system from Kazakh to English. Starting with a baseline model trained on 1.3K and then on 20K aligned sentences, we tried to cope with the complex morphology of Kazakh by applying different schemes of morphological word segmentation to the training and test data. Morphological segmentation appears to benefit our system: our best segmentation scheme achieved a 28% reduction of out-of-vocabulary rate and 2.7 point BLEU improvement above the baseline.

Italiano. *Questo articolo presenta dei risultati preliminari relativi allo sviluppo di un sistema di traduzione automatica statistica dal Kazaco all'Inglese. Partendo da un modello di base, addestrato su 1.3K e 20K coppie di frasi, proviamo a gestire la complessa morfologia del Kazaco utilizzando diversi schemi di segmentazione morfologica delle parole sui dati di addestramento e di valutazione. La segmentazione morfologica sembra apportare benefici al nostro sistema: il nostro migliore schema di segmentazione ottiene una riduzione del 28% del "Out-of-Vocabulary Rate" ed un miglioramento di 2.7 punti della misura "BLEU" rispetto al sistema di base.*

1 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. The main ideas of SMT were introduced by researchers at IBM's Thomas J. Watson Research Center (Brown et al., 1993). This paradigm implies that translations are generated on the basis of statistical models whose parameters are derived from the analysis of bilingual text corpora. We show how one can compile a Kazakh-English parallel corpus from publicly available resources in Section 2.

It is well known that challenges arise in statistical machine translation when we deal with languages with complex morphology, e.g. Kazakh. However recently there were attempts to tackle such challenges for similar languages by morphological pre-processing of the source text (Bisazza and Federico, 2009; Habash and Sadat, 2006; Mermer, 2010). We apply morphological preprocessing techniques to Kazakh side of our corpus and show how they improve translation performance in Sections 5 and 6.

2 Corpus preparation

In order to build an SMT system for any languages one needs to obtain a substantial amount of parallel texts in those languages.

2.1 Small corpus

First we decided to mine a parallel corpus from e-mail messages circulated within one of Kazakhstani organizations with a considerable amount of international staff. In that organization e-mail messages that are addressed to all employees are usually written in three languages: Kazakh, English and Russian. But sometimes they are written solely in English. To identify among all messages only those that contained at least Kazakh and English parts we examined several such e-mails, and we found out that most of them had 'Dear', 'Құрметті' and 'Уважаемые' as beginnings of English, Kazakh and Russian parts respectively as in the example below:

Dear Library Patrons, Please see the …
Құрметті оқырмандар, Қосымшадан …
Уважаемые читатели, Пожалуйста, …

Statistical analysis showed that at 0.9 confidence level a simple heuristic method that classified an e-mail message as trilingual if it contained the words 'Dear', 'Құрметті' and 'Уважаемые' would get not less than 77% of such e-mails.

Out of 1,609 e-mails addressed to all employees that were dumped in April 2014 from one of the company workers' mailbox, we could get 636 trilingual messages. In order to extract Kazakh and English parts from each text chunk we assumed that the Kazakh part began with 'Құрметті', the English part began with 'Dear' and the Russian part began with 'Уважаемые' as in the example above. There are better approaches to detect languages in a multilingual document, e.g. Compact Language Detector (`https://code.google.com/p/cld2/`) or `langid.py` (Lui and Baldwin, 2012), and we are going to use them in our future work.

We trained the *Punkt* sentence splitter from NLTK (Loper and Bird, 2002) on Kazakh side of the corpus and used it along with the pre-trained model for English to perform sentence segmentation for each e-mail message. Then sentence alignment for each pair of e-mails was performed using *hunalign* (Varga et al., 2005). After removing all repeating sentences we obtained 1,303 parallel sentences. We sampled 100 sentence pairs for tuning and 100 sentence pairs for testing purposes.

2.2 Larger corpus

A larger corpus was mined from the official site of the President of the Republic of Kazakhstan located at `http://akorda.kz`. Text extraction from HTML was performed through a Perl-script that used HTML::TreeBuilder module from CPAN. After sentence splitting and sentence alignment we obtained 22,180 parallel sentences. Unfortunately, there were misalignments and sometimes Russian sentences found their way into Kazakh side of the corpus. This happened because the President of Kazakhstan sometimes gave bilingual speeches in Kazakh and Russian and the Russian parts were not translated. We sampled 2,200 sentence pairs from the larger corpus, and 242 of them turned out to be misaligned. So, it seems that approximately $242/2200 = 11\%$ of all sentence pairs are "bad" and the data is subject to further cleaning. We used the "good" 1,958 sentence pairs out of 2,200 for tuning and testing purposes.

3 Kazakh morphology and MT

Kazakh is an agglutinative language, which means that words are formed by joining suffixes to the stem. A Kazakh word can thus correspond to English phrases of various length as shown in Table 1.

дос	friend
дос**тар**	friend**s**
достар**ым**	**my** friends
достар**ымыз**	**our** friends
достарымыз**да**	**at** our friends
достарымызда**мыз**	**we are** at our friends

Table 1: Example of Kazakh suffixation

The effect of rich morphology can be observed in our corpora. Table 2 provides the vocabulary sizes, type-token ratios (TTR) and out-of-vocabulary (OOV) rates of Kazakh and English sides of larger corpus.

	English	Kazakh
Vocabulary size	18,170	35,984
Type-token ratio	3.8%	9.8%
OOV rate	1.9%	5.0%

Table 2: Vocabulary sizes, TTR and test set OOV rates

It is easy to see that rich morphology leads to sparse data problems for SMT that make translation of rare or unseen word forms difficult. That is why we need to use morphological segmentation to reduce data sparseness.

4 Related work

Few small-sized (0.2K–1.3K sentences) and one medium-sized (69.8K sentences) parallel corpora for Kazakh-English pair are available within the OPUS project (Tiedemann, 2012). We were not aware of these resources at the beginning of our research, and therefore we decided to compile our own corpora.

Rule-based approach and preliminary ideas on statistical approach for Kazakh-to-English machine translation were discussed by Tukeyev et al. (2011). Sundetova et al. (2013) presented

structural transfer rules for English-to-Kazakh machine translation system based on Apertium platform (Forcada et al., 2011).

To our knowledge, this is the first paper on the application of SMT methods and morphological segmentation to Kazakh language. However preprocessing of morphologically-rich languages was considered previously in several works: for the Arabic-to-English task Habash and Sadat (2006) presented morphological preprocessing schemes; for the Turkish-to-English direction Bisazza and Federico (2009) developed morphological segmentation schemes and Mermer (2010) presented unsupervised search for the optimal segmentation. In our work we implemented four schemes suggested by Bisazza and Federico (2009), and developed three new schemes for verbs and gerunds.

5 Morphological segmentation schemes

5.1 Preprocessing technique

We performed morphological analysis for our corpora using an open-source finite-state morphological transducer *apertium-kaz* (Washington et al., 2014). It is based on Helsinki Finite-State Toolkit and is available within the Apertium project (Forcada et al., 2011). The analysis was carried out by calling `lt-proc` command of the Lttoolbox (Ortis-Rojas et al., 2005). Since more than one analysis was possible, disambiguation was performed through a Constrained Grammar rules (Karlsson et al., 1995) by calling the `cg-proc` command, which decreased ambiguity from 2.4 to 1.4 analyses per form (see an example of disambiguation in Table 3). In cases when ambiguity still remained we used the first analysis from the output of `cg-proc`.

‘in 2009 , we started the construction works .’		
2009 *жылы* *біз құрылысты бастадық* .		
	жылы⟨adj⟩	‘warm’
	жылы⟨adj⟩⟨advl⟩	‘warmly’
→	**жыл⟨n⟩⟨px3sp⟩⟨nom⟩**	**‘year’**
	жылы⟨adj⟩⟨subst⟩⟨nom⟩	‘warmth’

Table 3: Morphological disambiguation of a Kazakh word in context.

Consequently, each surface form is changed to one of its lexical forms. Now simple regular expressions can be used to describe different segmentation rules on lexical forms.

5.2 Segmentation schemes

Below we present segmentation schemes which are combinations of splitting and removal of tags from the analyzed lexical forms. Segmentation rules MS2–MS11 were suggested by Bisazza and Federico (2009).

MS2. Dative, ablative, locative and instrumental cases are split off from words, since they often align with the English prepositions ‘to’, ‘from’, ‘in’ and ‘with/by’, respectively. The remaining case tags – nominative, accusative and genitive – are removed from the words because they are not expected to have English counterparts.

MS6. After treating case tags we split off from nouns the possessive tags of all persons except the 3rd singular ⟨*px3sp*⟩, which is removed.

MS7. This rule splits off copula from words, in addition to MS6’s rules.

MS11. This rule splits off person suffixes from finite verb forms and copula, in addition to MS7’s rules.

MS11a. This rule removes person suffixes from finite verb forms, in addition to MS7’s rules.

MS12. In addition to MS11a’s rules this rule splits off dative, ablative, locative and instrumental cases from gerunds that are derived from verbs in active form. The remaining case tags – nominative, accusative and genitive – are removed.

MS13. In addition to MS12’s rules this rule splits off from gerunds the possessive tags of all persons except the 3rd singular ⟨*px3sp*⟩, which is removed.

The Kazakh side of our corpora was preprocessed by the aforementioned segmentation schemes. After that angle brackets ‘⟨⟩’ around tags were replaced by plus sign ‘+’ at the beginnings of tags for compatibility with SMT toolkit Moses (Koehn et al., 2007). The benefit of segmentation for word alignment in Kazakh-to-English direction is shown in Figure 1.

6 Experiments

6.1 Baseline

The open-source SMT toolkit Moses (Koehn et al., 2007) was used to build the baseline system. Phrase pairs were extracted from symmetrized word alignments generated by GIZA++ (Och and Ney, 2003). The decoder features a statistical log-linear model including a phrase-based translation model, a 5-gram language model, a lexicalized dis-

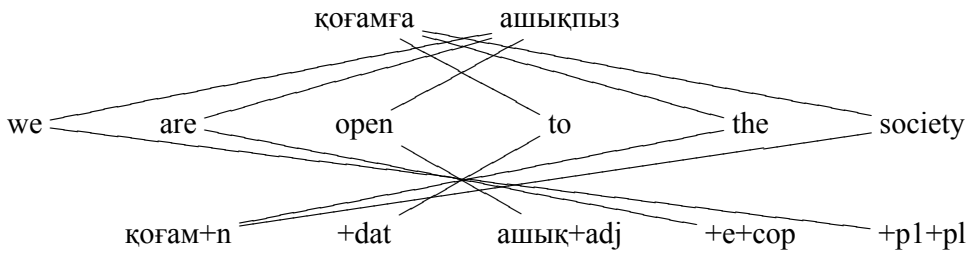

Figure 1: Word alignment before (up) and after (down) morphological segmentation MS11.

tortion model and word and phrase penalties. Distortion limit is 6 by default.

The weights of the log-linear combination were optimized by means of a minimum error rate training procedure (Och, 2003) run on tuning sets mentioned in section 2. Evaluation was performed on test sets.

6.2 Morphological segmentation

The impact of morphological segmentation on training corpus dictionary size and the test set OOV rate is shown in Table 4. One can see that better segmentation schemes lower the vocabulary size and OOV rate.

Scheme	Small corpus		Larger corpus	
	Vocab.	OOV	Vocab.	OOV
baseline	6,143	19.4	35,984	5.0
MS2	5,754	16.0	31,532	4.1
MS6	5,404	15.0	29,430	3.9
MS7	5,393	15.0	29,270	3.9
MS11	5,368	14.1	28,928	3.7
MS11a	5,362	14.8	28,923	3.8
MS12	5,283	14.5	28,079	3.7
MS13	5,241	14.3	27,792	3.6

Table 4: Effect of preprocessing on Kazakh side's training corpus vocabulary size and test set OOV rate.

6.3 Distortion limit

Since the number of words in each sentence has grown on average after segmentation, it seems reasonable to increase the distortion limit (DL) consequently. Thus, we allowed the distortion to be unlimited.

Table 5 shows how morphological preprocessing and unlimited distortion affects translation performance. In each system the same preprocessing was applied to the training, tuning and test data. Each system was run with limited and unlimited distortion but the set of weights for both cases was optimized with the default DL equal to 6.

Scheme	small corpus		larger corpus	
	DL=6	DL=∞	DL=6	DL=∞
baseline	17.69	17.32	22.75	23.70
MS2	18.50	**18.54**	23.77	25.23
MS6	17.29	17.32	23.77	25.06
MS7	17.63	17.43	23.90	25.41
MS11	14.95	15.13	23.62	25.21
MS11a	18.03	17.97	23.95	25.30
MS12	17.80	17.84	23.82	25.18
MS13	**18.74**	18.49	**24.05**	**25.46**

Table 5: BLEU scores.

7 Discussion and Future Work

The experiments have shown that a selective morphological segmentation improves the performance of an SMT system. One can see that in contrast to Bisazza and Federico's results (2009), in our case MS11 downgrades the translation performance. One of the reasons for this might be that Bisazza and Federico considered translation of spoken language in which sentences were shorter on average than in our corpora.

In this work we mainly focused on nominal suffixation. In our future work we are planning to: increase the dictionary of morphological transducer – currently it covers 93.3% of our larger corpus; improve morphological disambiguation using e.g. perceptron algorithm (Sak et al., 2007); develop more segmentation rules for verbs and other parts of speech; mine more mono- and bilingual data using official websites of Kazakhstan's public authorities.

Acknowledgments

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

References

Arianna Bisazza and Marcello Federico. 2009. Morphological Pre-Processing for Turkish to English Statistical Machine Translation. *Proceedings of IWSLT 2009*, 129–135.

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics,* 19(2):263–311.

Mikel L. Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M. Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation,* 25(2):127–144.

Nizar Habash and Fatiha Sadat. 2006. Arabic preprocessing schemes for statistical machine translation. *Proceedings of the Human Language Technology Conference of the NAACL 2006*, 49–-52.

Fred Karlsson, Atro Voutilainen, Juha Heikkilae, and Arto Anttila, eds. 1995. *Constraint Grammar: A language independent system for parsing unrestricted text*. Mouton de Gruyter.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan et al. 2007. Moses: Open source toolkit for statistical machine translation, *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics. Demo and Poster Sessions,* Prague, Czech Republic, 177–180.

Steven Bird. 2006. Nltk: the natural language toolkit. *Proceedings of the COLING/ACL on Interactive presentation sessions,* 69–72. Association for Computational Linguistics.

Marco Lui and Timothy Baldwin. 2012. langid.py: An Off-the-shelf Language Identification Tool. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)*, Demo Session, Jeju, Republic of Korea.

Coşkun Mermer and Ahmet Afşin Akin. 2010. Unsupervised search for the optimal segmentation for statistical machine translation. *Proceedings of the ACL 2010 Student Research Workshop*, 31–36. Association for Computational Linguistics.

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

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics,* 160–-167.

Sergio Ortiz Rojas, Mikel L. Forcada, and Gema Ramírez Sánchez. 2005. Construccion y minimización eficiente de transductores de letras a partir de diccionarios con paradigmas. *Procesamiento del Lenguaje Natural*, 35:51-–57.

Haşim Sak, Tunga Güngör, and Murat Saraçlar. 2007. Morphological disambiguation of Turkish text with perceptron algorithm. *Computational Linguistics and Intelligent Text Processing,* 107–118. Springer Berlin Heidelberg.

A. Sundetova, M. L. Forcada, A. Shormakova, A. Aitkulova. 2013. Structural transfer rules for English-to-Kazakh machine translation in the free/open-source platform Apertium. *Компьютерная обработка тюркских языков. Первая международная конференция: Труды.* Астана: ЕНУ им. Л. Н. Гумилева, 2013. – с. 317–326.

Jörg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'2012)*, 2214–2218

U. A. Tukeyev, A. K. Melby, Zh. M. Zhumanov. 2011. Models and algorithms of translation of the Kazakh language sentences into English language with use of link grammar and the statistical approach. *Proceedings of the IV Congress of the Turkic World Mathematical Society*, 1(3):474.

D. Varga, L. Németh, P. Halácsy, A. Kornai, V. Trón, V. Nagy. 2005. Parallel corpora for medium density languages. *Proceedings of the RANLP 2005*, 590–596.

Jonathan N. Washington, Ilnar Salimzyanov and Francis Tyers. 2014. Finite-state morphological transducers for three Kypchak languages. *Proceedings of the 9th edition of the Language Resources and Evaluation Conference*, 3378–3385.