Initial Experiments on Russian to Kazakh SMT

Bagdat Myrzakhmetov and Aibek Makazhanov

National Laboratory Astana
Astana, Kazakhstan
{bagdat.myrzakhmetov, aibek.makazhanov}@nu.edu.kz

Abstract. We present our initial experiments on Russian to Kazakh phrase-based statistical machine translation. Following a common approach to SMT between morphologically rich languages, we employ morphological processing techniques. Namely, for our initial experiments, we perform source-side lemmatization. Given a rather humble-sized parallel corpus at hand, we also put some effort in data cleaning and investigate the impact of data quality vs. quantity trade off on the overall performance. Although our experiments mostly focus on source side pre-processing we achieve a substantial, statistically significant improvement over the baseline that operates on raw, unprocessed data.

1 Introduction

Machine translation from Russian to Kazakh poses certain challenges of both linguistic and technical nature. First of all, both languages belong to a group of so called morphologically rich or complex languages that have sophisticated inflection systems, and notoriously raise data sparseness and long range dependencies issues. This fact practically demands the data sparseness problem to be addressed. At the same time, however, both, source and target, languages being morphologically complex hinders the use of a common remedy – morphological segmentation.

When dealing with morphologically rich languages, including Kazakh [1, 2], most of the previous work performed morphological segmentation of words into sub-lexical units, such as isolated morphemes [1, 3], groups of morphemes [4] (both free and bound), or morpheme-like units obtained by automatic segmentation [5]. The motivation behind this approach is to reduce data sparseness by splitting off and/or removing some of the units, and hopefully improve alignment quality, as certain units in one language may correspond to words in another, e.g. case affixes often correspond to adpositions [1, 4]. The choice of a particular segmentation scheme, i.e. the decision of which sub-lexical units to isolate, remove, or group, is usually guided by the nature of translation languages in question, and based on the results of empirical comparison[1]. When only one of the translation languages is morphologically complex and the other is not (e.g. mostly-analytic language such as Mandarin Chinese or English) candidate

[1] A notable exception is the work by Mermer [5] who proposed a language independent method that does not require any linguistic knowledge, and is capable of automatically choosing an optimal segmentation scheme by training a generative segmentation-translation model, which maximizes posterior probability of the training corpus.

Table 1. Realization of grammatical categories for Kazakh, Russian, and English: M - Morphological; A - Analytic; L - Lexical

#	Grammatical category	Kazakh	Russian	English
1	Number	M	M/L	M/L
2	Possession	M	A	A
3	Case	M/A	M/A	A/L
4	Gender	–	M/L	L
5	Agreement	M	M	M/L
6	Voice	M	M/A	A
7	Tense	M/A	M/A	M/A/L
8	Mood	M/A	M/A	A
9	Aspect	A+L	L	M/A/L
10	Negation	M/A	A	A

segmentation schemes can be devised intuitively (but not easily), provided sufficient familiarity with the grammar. However, when both languages are morphologically rich, and on top of it, typologically distant, devising such schemes may prove to be more complicated. This is exactly the problem with Russian and Kazakh.

While both, Russian and Kazakh, are morphologically complex, they employ different inflection strategies. Russian, a Slavic language, is usually classified as fusional, i.e. a language that "packs" several grammatical categories into a single affix. In contrast Kazakh, as a Turkic language, employs agglutination and "stacks" grammatical categories one per affix[2]. While corresponding many-to-one mapping of affixes through common grammatical categories is possible, it has to account for idiosyncrasies of grammar of both languages, e.g. Russian numerals agree with nouns in case, and adjectives – in case, number and gender, while neither type of agreement is found in Kazakh[3]. Even if such a sensitive segmentation scheme could be designed, the issue does not end there: richness of morphology does not, of course, imply a language's complete reliance on it. Many things in both languages can be also (or exclusively) expressed analytically and/or lexically. A very general, and by no means authoritative, summary of surface realization of major grammatical categories for both languages (and English as an example of morphologically "poor" language) is given in Table 1. Notice that only verbal agreement is realized by the same means, namely morphologically, and for other nine categories means of grammeme encoding allow certain variation. Thus, from a linguistic perspective, for the given language pair we are faced with sparseness on both sides and difficulty in designing morphological segmentation schemes.

From technical perspective, there is another challenge that concerns mostly Kazakh in its lack of resources for our particular purposes. By and large the language is being actively studied, and there exist monolingual corpora [6, 7], and ongoing research on

[2] For example, consider plural genitive of "language": (*rus*) *jazyk-ov* – language-PL.GEN; (*kaz*) *til-der-ding* – language-PL-GEN. Notice how number and case are encoded in a single suffix in Russian (*rus*), and in two successive suffixes in Kazakh (*kaz*).

[3] Thus if we were to segment Russian numerals and adjectives and split off corresponding suffixes their alignment to anything except null would be incorrect.

morphological processing [8–13] and syntactic parsing [14–16]. However, except for a rather small and noisy OPUS corpus [17] there are no Russian-Kazakh parallel corpora[4] and the only tool for automatic morphological disambiguation of Kazakh available to us[5] was reported to have accuracy of 86%, which we considered to be low enough to question the results of experiments with segmentation: would possible misalignments be shortcomings of a chosen segmentation scheme or results of incorrect morphological analysis and disambiguation. Thus, due to the lack of resources to process the target side and no solid approach to the design of suitable segmentation scheme for the particular pair of languages, in our first attempt to tackle the problem we settle for discussing potential challenges and conducting straightforward experiments.

To our knowledge, the present work is the first to address the problem of Russian to Kazakh SMT. Our contribution is a rather modest one: for the initial experiments, we are testing waters by performing a source-side lemmatization and dictionary-based data cleaning. The aim is to see what improvement one can get from a basic, light-weight morphological pre-processing, such as lemmatization, and to quantify data quality vs. quantity trade-off. The intuition behind lemmatization was to reduce sparseness by performing a "poor man's segmentation", where all the inflections of the source side morphologically complex language get trimmed off. The data cleaning step was necessitated by the fact that bulk of the data we used was misaligned and noisy. The results we obtained show that just source-side pre-processing gives a net improvement of 0.98 BLEU points (6.3% relative), and, when coupled with a data cleaning procedure, the approach results in 1.49 net and 9.5% relative improvement over the baseline.

The rest of the paper is organized as follows. In Section 2 we discuss the related work. In Section 3 we proceed to the description of the parallel corpus that we used for training, testing and tuning our models and the process of its collection and alignment. In Section 4 we describe our experiments and report results. Lastly, we provide concluding remarks and discuss future works in Section 5.

2 Related Work

There is a number of works on rule-based machine translation involving Kazakh language with the main focus on building bilingual dictionaries and structural transfer rules, and developing word sense disambiguation techniques for the open source Apertium system [19] to translate to/from Russian [20] and English [21, 22].

As for the statistical machine translation research involving Kazakh, two studies concerned with Kazakh to English SMT [1, 2] perform morphological segmentation on the source side, using respectively Morphessor [23] and the HFST-based [24] Kazakh FST [11]. Both studies report relative improvement over the word-based baselines. Assylbekov and Nurkas [1] devise several segmentation schemes, and evaluate each of them. Their results suggest that removing 3rd person possession and genitive, accusative case markers from both nominals and non finite verbs, and doing the same for

[4] At the time of experiments a Russian-Kazakh parallel corpus of approximately 342K sentence pairs described by Assylbekov et al. [18] was not available to us.

[5] At the time of experiments a disambiguation tool with reported accuracy of almost 91% developed by Assylbekov et al. [8] was not available.

agreement markers, while splitting off much of the rest of the inflections, yields the best results. Interestingly enough, when faced with the problem of the absence of an accurate morphological disambiguation tool for Kazakh, the authors chose to reduce (not to resolve completely) ambiguity by using a constraint grammar-based tool, and to replace remaining ambiguous analyses by the first analysis returned by this tool. Unfortunately, the authors did not measure or in any way analyze the effect of incorrect disambiguation on the quality of produced alignments and translation.

Assylbekov et al. [18] report on building a Russian-Kazakh parallel corpus of around 342K sentence pairs, which, unfortunately, was not available at the time when we conducted our experiments. The authors describe various data pre- and post-processing techniques that improve the quality of sentence alignment. Among other methods the authors describe a dictionary-based re-alignment of lemmatized sentence pairs, a technique that we use in our data cleaning step. According to the authors this technique increases the portion of correctly aligned sentences (as measured by an automatic learner-based estimator) in a given bitext by 2% compared to the standard length-based alignment procedure.

Various segmentation schemes are also considered in works dealing with translation from and to Russian. Lo et al. [25] employ source-side lemmatization for the Russian-to-English translation task. However, unlike what we do in the present work, the authors use lemmatization only at the word alignment step and restore original surface forms before estimating the remaining parameters of the model. The intuition is to obtain accurate word alignments through lemmatization, while being able to use valuable information, such as case and agreement markers, encoded in inflected forms. For the same translation task Borisov and Galinskaya [26] propose a sophisticated segmentation scheme, consisting of a number of rules, that cover nouns, adjectives, various verb forms, and a catch-all rule that covers all possible remaining cases, except ambiguous analyses, which are simply skipped. Apart from "usual suspects" such as case and agreement markers, the authors consider isolation of comparison degree affixes from adjectives (for some reason living out this category for adverbs) to account for irregular forms in English. Remaining inflections for this parts of speech, i.e. gender, number, and case, are removed as they are not defined for English adjectives. The authors report a small improvement over an already strong baseline, and respective reduction of 35% and 29% (best case) in word types and OOV rate for Russian.

3 Data Set

Our data set consists of a portion of the OPUS corpus [17], on-line news (`https://ortcom.kz`), a collection of historical essays (`https://e-history.kz`), and legal texts (`https://akorda.kz`). When obtained the Russian-Kazakh portion of the OPUS corpus contained 92035 parallel sentences that came mostly from software documentation and movie sub-titles. However, upon visual inspection we have noticed some repetitions, corrupted text bits, and clearly misaligned pairs. After manually removing such instances, we ended up with more than 56K sentences. News, essay, and legal data were aligned using Hunalign [27]. The aligned data were manually checked and around 2.5K sentence pairs were held out for testing and tuning, while remaining

Table 2. Quantitative description of the data set

	# sentences	**# tokens**	**# unique tokens**
Kazakh			
Training	69889	447499	68547
Testing	1004	15563	5653
Tuning	1510	21691	7065
Russian			
Training	69889	475388	68884
Testing	1004	16509	5958
Tuning	1510	22936	7251

Table 3. Domain distribution on the data set

	Training		**Testing**		**Tuning**	
Domain	# sen-s (%)	# tok-s (%)	# sen-s (%)	# tok-s (%)	# sen-s (%)	# tok-s (%)
OPUS	56695 (81.5)	425431 (46.1)	-	-	-	-
News	10027 (14.3)	383079 (41.5)	491 (48.9)	14766 (46.0)	1260 (83.3)	37606 (84.3)
Essay	1402 (2.0)	49963 (5.4)	72 (7.2)	2241 (6.9)	-	-
Law	1765 (2.2)	64414 (7.0)	441 (43.9)	15065 (47.1)	252 (16.7)	7071 (15.7)

13K+ pairs together with the cleaned OPUS data comprised the training set. Table 2 shows the counts of sentence pairs, running and unique tokens contained in the data set. The distribution of the data across domains is shown in Table 3.

4 Experiments and Evaluation

In our experiments we build and compare two models: (i) the baseline model that assumes no pre-processing of the input, and (ii) the source lemmatized (SL) model which is run on the data whose source side (Russian) has been lemmatized. For lemmatization we use Mystem [28], a freely available stemmer for Russian. We proceed to evaluate both models in two settings: on complete and cleaned training sets respectively. For cleaning we again use hunalign [27], but this time we also use a Russian-Kazakh dictionary[6] with about 116K entries. The tool scores existing alignments, and we remove those that ranked among the bottom 10%. Because dictionary entries are lemmatized, during cleaning we perform lemmatization on both source and target sides of the training set, and later restore the target side of the cleaned data. For target side lemmatization we use a data-driven morphological disambiguator for Kazakh [10].

We implement the models using the Moses toolkit [29], setting the distortion limit parameter to -1 (infinity) to account for long range dependencies and free word order of the languages. Remaining parameters are estimated on the tuning set with the help of the MERT [30] procedure. We train a 3-gram language model smoothed with the modified

[6] `http://mtdi.kz/til-bilimi/sozdikter/oryssha-kazaksha`

Table 4. Main results

	Training set	
Model	Complete	Cleaned
Baseline	15.62±0.08	16.02±0.11
Source lemmatized	16.60±0.08	**17.11±0.09**
Statistical significance, %	98.6	100.0

Kneser-Ney [31] algorithm on the target side of the training data and a portion of the Kazakh Language Corpus [6] that contains slightly more than 1.3M running tokens. The models are evaluated in terms of BLEU [32] metric. The results are averaged over three independent tuning runs and reported in Table 4 together with the standard deviations and results of the statistical significance tests. Statistical significance is calculated using bootsrap resampling technique [33] for 1000 samples under the null-hypothesis that the SL model outperforms the baseline.

As it can be seen on the complete data set the SL model scores at 16.6 points against 15.6 points of the baseline, achieving 6.3% relative improvement. Data cleaning boosts the performance of both models on approximately half a point. On the cleaned data the SL model outperforms the baseline again, and achieves relative improvement of 6.8%. Thus, after cleaning and lemmatization, we improve 9.5% over the baseline that operates on the raw, unprocessed data.

Lastly, in terms of combating data sparseness, let us note that after lemmatization, the number of unique tokens on the source side has been reduced to 62.5% (from 68884 to 25824), and the size of the phrase table has been reduced to 12.3% (from 1070916 to 939087 entries).

5 Conclusion and Future Work

We have conducted initial experiments with Russian to Kazakh SMT. Our findings suggest that even light weight morphological processing, such as lemmatization on the source side, provides substantial improvement over the word-based baseline that assumes no pre-processing of the input. We have also showed that noise reduction in the training set can be beneficial as well, although the improvement is a less drastic one.

For the future work we plan to closely investigate various strategies of morphological segmentation for these languages. We also plan to enrich our data with grammatical annotation in order to experiment with factored models [34].

Acknowledgments

This work has been funded by the Committee of Science of the Ministry of Education and Science of the Republic of Kazakhstan under the targeted program O.0743 (0115PK02473) and by the Nazarbayev University under the research grant 064-2016/ 013-2016.

References

1. Assylbekov, Z., Nurkas, A.: Initial explorations in Kazakh to english statistical machine translation. In: The First Italian Conference on Computational Linguistics, CLiC-it. (2014)
2. Kartbayev, A.: Learning word alignment models for kazakh-english machine translation. In: Integrated Uncertainty in Knowledge Modelling and Decision Making - 4th International Symposium, IUKM 2015. (2015) 326–335
3. Bisazza, A., Federico, M.: Morphological pre-processing for turkish to english statistical machine translation. In: IWSLT 2009. (2009)
4. Oflazer, K., El-Kahlout, I.D.: Exploring different representational units in english-to-turkish statistical machine translation. In: Proceedings of the Second Workshop on Statistical Machine Translation, Association for Computational Linguistics (2007) 25–32
5. Mermer, C.: Unsupervised search for the optimal segmentation for statistical machine translation. In: Proceedings of the ACL 2010 Student Research Workshop, Association for Computational Linguistics (2010) 31–36
6. Makhambetov, O., Makazhanov, A., Yessenbayev, Z., Matkarimov, B., Sabyrgaliyev, I., Sharafudinov, A.: Assembling the kazakh language corpus. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, Washington, USA, Association for Computational Linguistics (October 2013) 1022–1031
7. Altenbek, G., Xiao-long, W.: Kazakh segmentation system of inflectional affixes. In: Joint Conference on Chinese Language Processing, CIPS-SIGHAN (2010) 183–190
8. Assylbekov, Z., Washington, J., Tyers, F., Nurkas, A., Sundetova, A., Karibayeva, A., Abduali, B., Amirova, D.: A free/open-source hybrid morphological disambiguation tool for Kazakh. In: TurCLing 2016. (2016) 18–26
9. Kessikbayeva, G., Cicekli, I.: A rule based morphological analyzer and a morphological disambiguator for kazakh language. Linguistics and Literature Studies **4**(1) (2016) 96–104
10. Makhambetov, O., Makazhanov, A., Sabyrgaliyev, I., Yessenbayev, Z.: Data-driven morphological analysis and disambiguation for kazakh. In: Proceedings of the 2015 Computational Linguistics and Intelligent Text Processing, Part I, Cairo, Egypt, Springer International Publishing (2015) 151–163
11. Washington, J., Salimzyanov, I., Tyers, F.: Finite-state morphological transducers for three kypchak languages. In Chair), N.C.C., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S., eds.: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland, European Language Resources Association (ELRA) (may 2014)
12. Makazhanov, A., Makhambetov, O., Sabyrgaliyev, I., Yessenbayev, Z.: Spelling correction for kazakh. In: Proceedings of the 2014 Computational Linguistics and Intelligent Text Processing, Kathmandu, Nepal, Springer Berlin Heidelberg (2014) 533–541
13. Makazhanov, A., Yessenbayev, Z., Sabyrgaliyev, I., Sharafudinov, A., Makhambetov, O.: On certain aspects of kazakh part-of-speech tagging. In: Application of Information and Communication Technologies (AICT), 2014 IEEE 8th International Conference on. (Oct 2014) 1–4
14. Altenbek, G., Wang, X., Haisha, G.: Identification of basic phrases for kazakh language using maximum entropy model. In: COLING. (2014) 1007–1014
15. Tyers, F.M., Washington, J.: Towards a free/open-source universal-dependency treebank for kazakh. In: 3rd International Conference on Turkic Languages Processing (TurkLang 2015), Kazan, Tatarstan (2015) 276–290
16. Makazhanov, A., Sultangazina, A., Makhambetov, O., Yessenbayev, Z.: Syntactic annotation of kazakh: Following the universal dependencies guidelines. a report. In: 3rd International Conference on Turkic Languages Processing (TurkLang 2015), Kazan, Tatarstan (2015) 338–350

17. Tiedemann, J.: Parallel data, tools and interfaces in opus. In: LREC. (2012)
18. Assylbekov, Z., Myrzakhmetov, B., Makazhanov, A.: Experiments with Russian to Kazakh Sentence Alignment. Izvestija KGTU im.I.Razzakova **38**(2) (2016) 18–23
19. Forcada, M.L., Ginestí-Rosell, M., Nordfalk, J., O'Regan, J., Ortiz-Rojas, S., Pérez-Ortiz, J.A., Sánchez-Martínez, F., Ramírez-Sánchez, G., Tyers, F.M.: Apertium: A free/open-source platform for rule-based machine translation. Machine Translation **25**(2) (June 2011) 127–144
20. Rakhimova, D.: Research of problem of the semantic analysis and sythesis of pretext in the russian-kazakh machine translation. In: 3rd International Conference on Turkic Languages Processing (TurkLang 2015), Kazan, Tatarstan (2015) 59–67
21. Assem Shormakova, Aida Sundetova, A.S.: Features of machine translation of different systemic languages using an apertium platform (with an example of english and kazakh languages). JSCSE (2013) 255–259
22. Sundetova, A., Forcada, M., Tyers, F.: A free/open-source machine translation system for english to kazakh. In: 3rd International Conference on Turkic Languages Processing (TurkLang 2015), Kazan, Tatarstan (2015) 78–91
23. Creutz, M., Lagus, K.: Unsupervised models for morpheme segmentation and morphology learning. ACM Transactions on Speech and Language Processing (TSLP) **4**(1) (2007) 3
24. Linden, K., Silfverberg, M., Axelson, E., Hardwick, S., Pirinen, T. In: HFST-Framework for Compiling and Applying Morphologies. Volume Vol. 100 of Communications in Computer and Information Science. (2011) 67–85
25. Lo, C.k., Cherry, C., Foster, G., Stewart, D., Islam, R., Kazantseva, A., Kuhn, R.: Nrc russian-english machine translation system for wmt 2016. In: Proceedings of the First Conference on Machine Translation, Association for Computational Linguistics (2016) 326–332
26. Borisov, A., Galinskaya, I.: Yandex school of data analysis russian-english machine translation system for wmt14. In: Proceedings of the Ninth Workshop on Statistical Machine Translation, Association for Computational Linguistics (2014) 66–70
27. Varga, D., Halácsy, P., Kornai, A., Nagy, V., Németh, L., Trón, V.: Parallel corpora for medium density languages. AMSTERDAM STUDIES IN THE THEORY AND HISTORY OF LINGUISTIC SCIENCE SERIES 4 **292** (2007) 247
28. Segalovich, I.: A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine. In: MLMTA. (2003)
29. Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al.: Moses: Open source toolkit for statistical machine translation. In: Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions, Association for Computational Linguistics (2007) 177–180
30. Och, F.J.: Minimum error rate training in statistical machine translation. In: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1, Association for Computational Linguistics (2003) 160–167
31. Kneser, R., Ney, H.: Improved backing-off for m-gram language modeling. In: Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on. Volume 1., IEEE (1995) 181–184
32. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting on association for computational linguistics, Association for Computational Linguistics (2002) 311–318
33. Koehn, P.: Statistical significance tests for machine translation evaluation. In: Proceedings of EMNLP 2004, Association for Computational Linguistics (2004) 388–395
34. Koehn, P., Hoang, H.: Factored translation models. In: EMNLP-CoNLL. (2007) 868–876