SMT: A Case Study of Kazakh-English Word Alignment

Amandyk Kartbayev(✉)

Laboratory of Intelligent Information Systems,
Al-Farabi Kazakh National University, Almaty, Kazakhstan
a.kartbayev@gmail.com

Abstract. In this paper, we present results from a set of experiments to determine the effect on translation quality, it depends on the particular kind of morphological preprocessing that can be represented by finite-state transducers. A high agglutinative nature of the Kazakh language under the condition of poor language resources makes an issue in the processing of derivational morphology. Our methods are focused on useful phrase pairs in word alignment and provide a most language independent approach, which may improve a translation into other morphological complex languages. We processed our algorithms over the Kazakh Wikipedia base of about 1.5 million unique lexeme and 230 million words overall. Our best translation system increases 3 BLEU points over the Kazakh-English baseline on a blind test set.

Keywords: Word alignment · Kazakh morphology · Word segmentation · Machine translation

1 Introduction

In this work we focus on the word alignment process, which is the most important part of information recovery from a source with a lot of inflection. Particularly, we are interested in the sources where the given sentence pairs contain more new words with a less prior information about their nature. This is a challenging problem in machine translation and it is a hard to learn from the lexicon and usually repeats the similar errors again and again.

Morphological segmentation process intended to break words into morphemes, which are the basic semantic units and a key component for natural language processing systems. This is our current subtask in the machine translation project and we also desired to show that a simple segmentation scheme can perform pretty well as the most sophisticated one.

Most papers in statistical machine translation (SMT) oriented morphology analysis presents experiments that they consist of numerous experimentation to choose the best among a set of segmentation schemes. These morphological preprocessing schemes focused on various level of decomposition and compare the resulting translation performances, but usually use a subset of morphology and apply only a few simple rules in a segmentation process.

F. Daniel and O. Diaz (Eds.): ICWE 2015 Workshops, LNCS 9396, pp. 40–49, 2015.
DOI: 10.1007/978-3-319-24800-4_4

In the paper, well known to the intended audience, El-Kahlout and Oflazer[1] explored this task for English to Turkish translation, which is an agglutinative language as Kazakh. Their methods used in the survey were a morphological analyzer and token disambiguation, though translation models trained throw morphemes obviously degrades the translation quality. But they outperformed the baseline results after some morpheme grouping techniques. A research more relevant to this work was done by Bisazza and Federico[2].

Our segmentation model incorporates simple ideas inspired by finite state features such as morphemes and their contexts in the range of situations, where the lexeme is likely a morpheme, as any other cases, it is a word boundary. We develop a segmentation scheme using syntactic and morphological rules are implemented as finite-state transducers. We focus on derivational morphology and tested our approach on Kazakh wiki and news datasets, which was crawled from Web. The affix system, which will be the focus in this paper, is described in more detail in Table 1.

Table 1. An example of Kazakh agglutination

Stem	Plural affixes	Possesive affixes	Case affixes
stem[kol'+]	plural[+der]	1-st pl.[+imiz]	locative[+de]
stem[kol'+]	-	1-st s.[+im]	locative[+de]
stem[kol'+]	-	-	locative[+de]

Our system, using monolingual features only, is one of the most realistic application for Kazakh and compared to Morfessor tool[3], so it can be readily applied to supervised and semi-supervised learning of morphological inflations of the language even on speech processing. Also morphological adjustment gives a improved statistical machine translation performance over the pair of the morphological rich and poor languages. A substantial improvement in translation performance is achieved, when we used word alignments learned from the output of the processing technique, but we found that some of the segmentation errors are caused by morphological analyzer. These kind of errors could be avoided using data selection, which demonstrates the ability of the method fix it successfully. Using morphological analysis we out grammatical features of word and can find syntactic structure of input sentence, which further demonstrates the benefit of using this method in machine translation.

In this paper, we present a systematic comparison of preprocessing techniques for a Kazakh-English pair. Previous researches that we explored on our approaches are rule-based morphological analyzers[4], which consist in deep language expertise and a exhaustive process in system development. Unsupervised approaches use actually unlimited supplies of text to cover very few labeled resource and it has been widely studied for a number of languages[5]. However, existing systems are too complicated to extend them with random overlapping dependencies that are crucial to segmentation.

On our general task we refer to the methodology exposed by Oflazer and El-Kahlout on the Turkish-English task. Because, Turkish is also morphological rich language like Kazakh and not all affix combinations looks grammatical. This means the linguistic knowledge is the key to finding significant segmentation schemes among many possible combinations of the rules. Only rule-based approaches are provided and have done detailed analyses of the Kazakh morphological parsing task. For a comprehensive survey of the rule-based morphological analyze we refer a reader to the research by Altenbek[6] and Kairakbay[7].

The paper is structured as follows: Section 2 discusses the key challenges of translating Kazakh to English. In Section 3 we described the different segmentation techniques we study. And Section 4 presents our evaluation results.

2 Translation Task

In our work, we experiment with a range of segmentation technique totally giving a five best distinct schemes. Our results show that the proper selection of the segmentation scheme has a significant impact on the performance of a phrase-based system in a large corpora. The translation experiments described in this paper are carried out with a standard phrase-based Moses[8] system (not with Experiment Management System) and the target-side language models were trained on the MultiUN[9] corpora.

Generally, breaking up a process of generating the data into smaller steps, modeling the smaller steps with probability distributions, and combining the steps into a coherent story is called generative modeling. The phrase-based models are generative models that translate sequences of words in f_j into sequences of words in $e_j(1)$, in difference from the word-based models that translate single words in isolation.

$$P\left(e_j \mid f_j\right) = \sum_{j=1}^{J} P\left(e_j, a_j \mid f_j\right) \tag{1}$$

Improving translation performance directly would require training the system and decoding each segmentation hypothesis, which is computationally impracticable. That we made various kind of conditional assumptions using a generative model and decomposed the posterior probability(2). In this notation e_j and f_i point out the two parts of a parallel corpus and a_j marked as the alignment hypothesized for f_i.

$$P\left(e_j^J, a_j^J \mid f_i^I\right) = \frac{f_i}{(I+1)^J} \prod_{j=1}^{J} p\left(e_j \mid f_{a_j}\right) \tag{2}$$

The use of phrases as translation units is motivated by the observation that sometimes one word in a source language translates into multiple words.

Because, a Kazakh word can correspond to a single English word, up to phrases of various lengths, or even to a whole sentence as shown in Table 2.

Our objective is to produce alignments, which can be used to build high quality machine translation systems[10]. These are pretty close to human annotated alignments that often contain m-to-n alignments, where several source words are aligned to several target words and the resulting unit can not be further decomposed. Using segmentation, we describe a new generative model which directly models m-to-n non-consecutive word alignments. There is a very small improvement in alignment if a source word occurs only once in the parallel text, the probability assigned to it, generates each of the words to the each sentence, will be too high. This problem is solved by smoothing the word-to-word translation probabilities with a coincident distribution.

3 Improving Word Alignment

In order to look through this task, we did a series of experiments and found morpheme alignment can be employed to increase the similarity between languages, therefore enhancing the quality of machine translation for Kazakh-English language pair. Our experiments consist of two parts: one is on Kazakh-English morphological segmentation; the other is a case study of the benefits of morpheme based alignment.

We use following heuristic methods that improve the generative models for phrase alignment. At first, the tags were assigned to the obtained phrase pair pieces, then we make classification and clustering the phrases according their contexts, also we extract phrase pairs that are not linked within the word alignments, like the phrases containing multiword entities that can not be correctly aligned. We obtained word alignments in both translation directions by the GIZA++ toolkit[11], which is based on the IBM models[12]. We prefer a grow-diag-final symmetrization method to others for both alignment directions.

As the first part of our experiments we morphologically segmented Kazakh input sentences to compute morpheme alignment. For these purposes we used Morfessor, an unsupervised analyzer and Helsinki Finite-State Toolkit (HFST)[13]. Helsinki Finite State Toolkit is an open-source implementation of the Xerox finite-state toolkit, that implements the lexc, twol and xfst formalisms for modeling morphlogical rules. After these Kazakh stems and suffixes is converted into labeled morphemes, as well as particular English verbs. We append a plus sign to the end of each open tag to know boundaries of internal morphemes from final ones, e.g., [stem+] and [stem] are assumed as different tokens.

3.1 Morphological Segmentation

Our preprocessing job starts from morphological segmentation, which includes running Morfessor tool and HFST to each entry of the corpus dictionary. The first step of word segmentation aims to get suffixes and roots from a vocabulary consisting of 1500k unique word forms taken from Kazakh Wikipedia dump[14].

Accordingly, we take surface forms of the words and generate their all possible lexical forms.

In the Kazakh language, as in other agglutinative languages,the morphemes are affixed to the root due to the morphotactic rules of the language. These morphotactic rules define the states and the suffixes that can be added to the stem, then change the state of the affixed word. These rules often represented by the certain finite state transducers. Where the transitions are marked as the derivational morphemes, that come in same order as the affixation of the word. Also we use the lexicon to label the initial states as the root words by parts of speech such as noun, verb, etc. The final states represent a lexeme created by affixing morphemes in each further states.

The schemes presented below are different combinations of outputs determining the removal of affixes from the analyzed words. The baseline approach is not perfect since a scheme includes several suffixes incorrectly segmented. In this case, we mainly focused on detection a few techniques for the segmentation of such word forms. In order to find an effective rule set we tested several segmentation schemes named S[1..5], some of which have described in the following Table 2.

Table 2. The segmentation schemes

Id	Schema	Examples	Translation
S1	stem	el	state
S2	stem+case	el + ge	state + dative
S3	stem+num+case	el + der + den	state + num + ablativ
S4	stem+poss+	el + in	state + poss2sing
S5	stem+poss+case	el + i +ne	state + poss3sing + dative

Nominal cases that are expected to have an English counterpart are split off from words: these are namely dative, ablative, locative and instrumental, often aligning with the English prepositions to, from, in and with/by. The remaining case affixes nominative, accusative and genitive are not have English counterparts. After treating case affixes we split of possessive suffixes from nouns of all persons except the 1st singular, which doesnt need removed.

There are large amount of verbs presenting ambiguity during segmentation, as suppositional verbs ‘eken’ - ‘to seem’ and ‘goi’. Which do not take personal endings, but follow conjugated main verbs. The verb ‘to become’ has the forms ‘bolu’ - ‘to become’, ‘bolar’ - ‘will become’, and ‘bolmau’ - ‘to not become’. There are also the verbs ‘bar’ - ‘to exist/have’ and ‘jok’ - ‘to not exist/not have’. These are special verbs because they do not take personal endings. Also a verbs generally refer to group action, e.g. ‘oinasu’ - ‘to play together’, ‘soilesu’ - ‘to converse’ produce an ambiguity, e.g. a stem ‘soile’ - ‘say’ and a suffix ‘su’ - ‘water’. During the process, we hardly determined the border between stems and inflectional affixes, especially when the word and the suffix matches entire word in the language. For instance, a progressive auxiliary word ‘jat’ - ‘alien’ and the negation morphemes like ‘ba’, ‘ma’, etc, though an irregular form of several verbs. Under many situations, the type of words, which we described, made an inaccurate stemming. In fact, there are lack of syntactic information we cannot easily distinguish among similar cases.

While GIZA++ tool produces a competitive alignment between words, the Kazakh sentences must be segmented as we already have in the first step. Therefore our method looks like an word sequence labeling problem, the contexts can be presented as POS tags for the word pairs.

Table 3. Part of Speech tag patterns

Tag	Sample	Tag	Sample
NN (Noun)	“el”-“state”	JJS (Adjective, super.)	“tym”-“most”
NNP (Proper noun)	“biz”-“we”	VB (Verb, base form)	“bar”-“go”
JJ (Adjective)	“jasyl”-“green”	VBD (Verb, past tense)	“bardy”-“went”
JJR (Adj, comp.)	“ulkenirek”-“bigger”	VBG (Verb, gerund)	“baru”-“to go”
RB (Adverb)	“jildam”-“speedy”	CC (Conjunction)	“jane”-“and”

3.2 Alignment Model

We extend the alignment modeling process of Brown et al. at the following way. We assume the alignment of the target sentence e to the source sentence f is a. Let c be the tag(from Penn Treebank) of f for segmented morphemes. This tag is an information about the word and represents lexeme after a segmentation process. This assumption is used to link the multiple tag sequences as hidden processes, that a tagger generates a context sequence c_j for a word sequence f_j(3).

$$P\left(e_1^I, a_1^I \mid f_1^J\right) = P\left(e_1^I, a_1^I \mid c_1^J, f_1^J\right) \tag{3}$$

Then we can show Model 1 as(4):

$$P\left(e_i^I, a_i^I \mid f_j^J, c_j^J\right) = \frac{1}{(J+1)^I} \prod_{i=1}^{I} p\left(e_i \mid f_{a_i}, c_{a_i}\right) \tag{4}$$

The training is carried out in the tagged Kazakh side and the untagged English side of the parallel text. If we estimate translation probabilities for every possible context of a source word, it will lead to problems with data sparsity and rapid growth of the translation table. We applied expectation maximization(EM) algorithm to cluster a context of the source sentence using similar probability distributions, avoiding problems with data sparsity and a size of the translation table another case.

We estimate the phrase pairs that are consistent with the word alignments, and then assign probabilities to the obtained phrase pairs. Context information is incorporated by the use of part-of-speech tags in both languages of the parallel text, and the EM algorithm is used to improve estimation of word-to-word translation probabilities. The probability p_k of the word w to the corresponding context k is:

$$p_k\left(w\right) = \frac{p_k f_k\left(w \mid \phi_k\right)}{\sum p_i f_i\left(w \mid \phi_i\right)} \tag{5}$$

Where, ϕ is the covariance matrix, and f are certain component density functions, which evaluated at each cluster. After we use association measures to filter infrequently occurring phrase pairs by log likelihood ratio r estimation[15]. For n pairs of the phrases, we can obtain the phrase pairs whose comparative values are larger than a threshold value as follows(6):

$$R\left(f, e\right) = \frac{r\left(f, e\right)}{Max_e r\left(f, e\right)} \tag{6}$$

Our algorithm, like a middle tier component, processes the input alignment files in a single pass. Current implementation reuses the code from https://github.com/akartbayev/clir that conducts the extraction of phrase pairs and filters out low frequency items. After the processing all valid phrases will be stored in the phrase table and be passed further. This algorithm proposes refinement by adding morphological constraints between the direct and the reverse directions of the alignment, which may improve the final word alignments.

4 Evaluation

Though our final objective is an improvement of the translation quality of SMT systems, we evaluate the alignment relies with the phrase-based system on the Kazakh-English parallel corpus of approximately 60K sentences, which have a maximum of 100 morphemes. Our corpora consists of the legal documents from http://adilet.zan.kz, a content of http://akorda.kz, and Multilingual Bible texts. We conduct all experiments on a single PC, which runs the 64-bit version of Ubuntu 14.10 server edition on a 4Core Intel i7 processor with 32 GB of RAM in total. All experiment files were processed on a locally mounted hard disk. Also we expect the more significant benefits from a larger training corpora, therefore we are in the process of its construction.

We did not have a gold standard for phrase alignments, so we had to refine the obtained phrase alignments to word alignments in order to compare them with our word alignment techniques.

Table 4 shows the change in alignment error rate (AER) of the alignments, that the improved model produce a decrease in AER and leads to a better translation quality, measured by BLEU score[16]. A high recall apparently improves translation quality, but low precision may decrease it and a relation between recall and precision is substantial. A high recall and low precision in alignment

Table 4. Alignment quality results

System	Precision	Recall	F-score	AER
Baseline	57.18	28.35	38.32	36.22
Morfessor	71.12	28.31	42.49	20.19
Rule-based	89.62	29.64	45.58	09.17

Table 5. Metric scores for all systems

System	BLEU	METEOR	TER
Baseline	30.47	47.01	49.88
Morfessor	31.90	47.34	49.37
Rule-based	33.89	49.22	48.04

pretty significant for the amount of generated phrases. The best situation takes place on well maintained recall and precision, which is a result of our study.

We employed an approach of the morpheme-based representation as explained in Section 3 about the morphological analysis, which impacts an improvement of +2 BLEU points. The system parameters were optimized with the minimum error rate training (MERT) algorithm [17], and evaluated on the out-of and in-domain test sets. Monolingual corpora from News Commentary was partially used, when we trained 5-gram language models. All language models were trained with the IRSTLM toolkit[18] and then were converted to binary form using KenLM for a faster execution[19].

Table 5 visualizes the best BLEU scores, which were computed using the MultEval[20]: BLEU, TER[21] and METEOR[22]; and we ran Moses three times per experiment setting, and report the highest BLEU scores obtained. Our survey shows that translation quality measured by BLEU metrics is not strictly related with lower AER.

5 Conclusions

In this work, we address a morpheme alignment problems concerned highly inflected languages. We compared our approach against a baseline of the Moses translation pipeline and another common approach to inflected languages segmentation. By using our method for phrase selection we were able to obtain translation quality better than the baseline method produce, while the phrase table size and the noise phrase pairs have been reduced by substantial level. Although memory requirements of the processing environment are increased, but they are still within manageable limits.

Our method is comparable to other language-specific works, and there are many possible directions for future research. As our approach may produce improvements in alignment quality, any downstream changes of the translation model also possible. We learned that processing the features are integrated into the standard phrase table is an area for improvement. That was our initial investigation into alignment models and further translation experiments will be carried out.

References

1. Oflazer, K., El-Kahlout, D.: Exploring different representational units in English-to-Turkish statistical machine translation. In: 2nd Workshop on Statistical Machine Translation, Prague, pp. 25–32 (2007)
2. Bisazza, A., Federico, M.: Morphological pre-processing for Turkish to English statistical machine translation. In: International Workshop on Spoken Language Translation 2009, Tokyo, pp. 129–135 (2009)
3. Creutz, M., Lagus, K.: Unsupervised models for morpheme segmentation and morphology learning. ACM Transactions on Speech and Language Processing **4**, article 3. Association for Computing Machinery, New York (2007)
4. Beesley, K.R., Karttunen, L.: Finite State Morphology. CSLI Publications, Palo Alto (2003)
5. Goldsmith, J.: Unsupervised learning of the morphology of a natural language. Computational Linguistics **27**, 153–198 (2001)
6. Altenbek, G., Xiao-Long, W.: Kazakh segmentation system of inflectional affixes. In: CIPS-SIGHAN Joint Conference on Chinese Language Processing, Beijing, pp. 183–190 (2010)
7. Kairakbay, B.: A nominal paradigm of the Kazakh language. In: 11th International Conference on Finite State Methods and Natural Language Processing, St. Andrews, pp. 108–112 (2013)
8. Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E.: Moses: open source toolkit for statistical machine translation. In: 45th Annual Meeting of the Association for Computational Linguistics, Prague, pp. 177–18 (2007)
9. Tapias, D., Rosner, M., Piperidis, S., Odjik, J., Mariani, J., Maegaard, B., Choukri, K., Calzolari, N.: MultiUN: a multilingual corpus from united nation documents. In: Seventh conference on International Language Resources and Evaluation, La Valletta, pp. 868–872 (2010)
10. Moore, R.: Improving IBM word alignment model 1. In: 42nd Annual Meeting on Association for Computational Linguistics, Barcelona, pp. 518–525 (2004)
11. Och, F.J., Ney, H.: A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics **29**, 19–51 (2003)
12. Brown, P.F., Della-Pietra, V., Del-Pietra, S., Mercer, R.L.: The mathematics of statistical machine translation: Parameter estimation. Computational Linguistics **19**, 263–311 (1993)
13. Lindén, K., Axelson, E., Hardwick, S., Pirinen, T.A., Silfverberg, M.: HFST—framework for compiling and applying morphologies. In: Mahlow, C., Piotrowski, M. (eds.) SFCM 2011. CCIS, vol. 100, pp. 67–85. Springer, Heidelberg (2011)
14. Gabrilovich, E., Markovitch, S.: Computing semantic relatedness using wikipedia-based explicit semantic analysis. In: 20th International Joint Conference on Artificial Intelligence, Hyderabad, pp. 1606–1611 (2007)

15. Dunning, T.: Accurate methods for the statistics of surprise and coincidence. Computational Linguistics **19**, 61–64 (1993)
16. Papineni, K., Roukos, S., Ward, T., Zhu, W.: BLEU: a method for automatic evaluation of machine translation. In: 40th Annual Meeting of the Association for Computational Linguistics, Philadephia, pp. 311–318 (2002)
17. Och, F.J.: Minimum error rate training in statistical machine translation. In: 41st Annual Meeting of the Association for Computational Linguistics, Sapporo, pp. 160–167 (2003)
18. Federico, M., Bertoldi, N., Cettolo, M.: IRSTLM: an open source toolkit for handling large scale language models. In: Interspeech 2008, Brisbane, pp. 1618–1621 (2008)
19. Heafield, K.: Kenlm: faster and smaller language model queries. In: Sixth Workshop on Statistical Machine Translation, Edinburgh, pp. 187–197 (2011)
20. Clark, J.H., Dyer, C., Lavie, A., Smith, N.A.: Better hypothesis testing for statistical machine translation: controlling for optimizer instability. In: 49th Annual Meeting of the Association for Computational Linguistics, Portland, pp. 176–181 (2011)
21. Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J.: A study of translation edit rate with targeted human annotation. In: Association for Machine Translation in the Americas, Cambridge, pp. 223–231 (2006)
22. Denkowski, M., Lavie, A.: Meteor 1.3: automatic metric for reliable optimization and evaluation of machine translation systems. In: Workshop on Statistical Machine Translation EMNLP 2011, Edinburgh, pp. 85–91 (2011)