Using Kazakh Morphology Information to Improve Word Alignment for SMT

Amandyk Kartbayev

Abstract We propose an word alignment model with two core features: the ability to handle uncertainty in the morpheme matching process and in the selecting correct phrase alignments after its creation. These processes are based on the use of a morphological analysis tool and a large monolingual corpora, which is used for improving the alignment elements correspondence. The consideration of this tool is language-dependent for a special pair of the languages, although an Wikipedia data represents an adequate source of the training text that can be used in many cases, and even allows an unsupervised word segmentation. Based on these features, we propose an approach that captures the morphotactics which is common to the source text. The paper describes experiments in the general domain by using a tagset, and has been compared to a classical word alignment by the help of human judgment.

Keywords Word alignment · Kazakh morphology · Word segmentation · Machine translation · Information retrieval

1 Introduction

The growing demand of machine-translation applications that shows an witness of the creation of complex systems performing similar or even identical functions in real world. These systems, excess in the spare nature of their creation, have a limited functionality cause of mismatches in application purposes. However, integration of these systems is desirable to find information in business information systems. In recent years, a considerable research effort has been directed to evaluate the relationships between word alignment and machine translation performance, which aim at obtaining a certain degree of coordination between various kind of language pairs by automatically detecting correspondences between the elements of these alignments.

A. Kartbayev (✉)
Laboratory of Intelligent Information Systems, Al-Farabi Kazakh National University,
Almaty, Kazakhstan
e-mail: a.kartbayev@gmail.com

A. Abraham et al. (eds.), *Proceedings of the Second International Afro-European Conference for Industrial Advancement AECIA 2015*, Advances in Intelligent Systems and Computing 427, DOI 10.1007/978-3-319-29504-6_34

However, there is no theoretical support from the view of providing a formulation to describe the relationship between word alignments and machine translation performance.

We examine the Kazakh language, which is the majority language in the Republic of Kazakhstan. Kazakh is part of the Kipchak branch of the Turkic language family and part of the majority Ural-Altay family, in comparison with languages like English, is very rich in morphology.

The Kazakh language which words are generated by adding affixes to the root form is called an agglutinative language. We can derive a new word by adding an affix to the root form, then make another word by adding another affix to this new word, and so on. This iteration process may continue several levels. Thus a single word in an agglutinative language may correspond to a phrase made up of several words in a non-agglutinative language [1] (Table 1).

Although the phenomena of word alignment has learned considerably after many challenges by El-Kahlout [2], Bisazza and Federico [3], these contributions are related to the use of morphology, as well as the consideration of probability distribution within the phrase pairs and the resulting alignments. These research issues in word alignment was also handling a precision of the matching process [4].

In order to demonstrate our objective, which can be used to build high quality machine translation systems [5], several applications of word alignment can be found, such as adaptation of the context-semantic disclosure. An word alignment processing is more convenient with respect to obtain m-to-n alignments, where several source words are aligned to several target words than barely segmenting the strings before the matching process. The common approaches of word alignment training are IBM Models [6] and hidden Markov model (HMM)[7], which practically use expectation-maximization (EM) algorithm [8]. The EM algorithm finds the parameters that increases the likelihood of the dependent variables. EM transfers the sentences by overlapping the actual parameters, where some rare words align to many words on the opposite sentence pair.

Since generating segments and modeling the relative features of phrases comprise a similarity measure within the parallel corpora, it makes the system too general to be applied to other kind of language pair, with the different morphotactics. However, our approach can be applied to the potential areas, which include improvements in machine translation, machine learning methods and information retrieval. Anyway,

Table 1 An example of Kazakh agglutination

Word form	English translation
stem[shol'+]	Desert
stem[shol'+]+plural[+der]	Deserts
stem[shol'+]+plural[+der]+1-st pl.[+imiz]	Our deserts
stem[shol'+]+plural[+der]+1-st pl.[+imiz]+locative[+de]	In our deserts
stem[shol'+]+plural[+der]+1-st pl.[+imiz]+locative[+de]+[+gi]	Is in our deserts

many concepts and definitions are pretty vague, which needs to be dealt within the word alignment process.

Using a Morfessor tool [9], we can find grammatical features of the word and can retrieve syntactic structure of input sentence. It clearly demonstrates the benefit of using similarly to the rule-based morphological analyzers [10], which consist in deep language expertise and a exhaustive process in system development. Unsupervised approaches use actually unlimited supplies of text and widely studied for a number of languages [11]. However, for a comprehensive survey of the rule-based morphological analyze we refer a reader to the research by Altenbek [12] and Kairakbay [13].

The article is structured as follows. Section 2 discusses the proposed model and describes the different segmentation techniques we study. And Sect. 3 presents our evaluation results.

2 Description of Our Method

Hybrid methods comprise two major groups of approaches: those that use morpheme analysis, and those that rely on probability distribution combined with techniques from machine learning in order to compare the similarity of the stems and their synonymy and ambiguity problems. An alignment process was understood as the process of establishing relations between the elements of a parallel language pair, which results in an alignment between equivalent phrases. Different alignment techniques, which enhance the quality of machine translation for Kazakh-English lasnguage pair, have been introduced in the past years in order to resolve different types of morphological segmentation of Kazakh words, relying on methods coming from areas of machine learning and linguistics. For these purposes we used Morfessor, an unsupervised analyzer and Helsinki Finite-State Toolkit (HFST) [14] for the rule-based analyze; finally we use the GIZA++ [15] tool to produce IBM Model 4 word alignment. Our morpheme analysis approach is concerned with word segmentation and as a result comparing groups of morphemes to another and detects the relations that exists between them.

Our studies try to investigate the impact of pruning technique to the overall translation quality by reduction the level of sparse phrases, which leads to higher BLEU scores [16]. We don't use a manually annotated gold standard word alignment set that the similarity measured on new sets of alignments reflects the personal opinion about a translation similarity between the instances.

2.1 Word Alignment

We suppose a phrase pair is denoted by (F, E) and with an alignment A, if any words f_j in F have a correspondence in a, with the words e_i in E. Formal definition can be described as follows: $\forall e_i \in E\ :\ \left(e_i, f_j\right) \in a \Rightarrow f_j \in F$ and $\forall f_j \in F\ :\ \left(e_i, f_j\right) \in a \Rightarrow e_i \in E$, clearly, there are $\exists e_i \in E, f_j \in F\ :\ \left(e_i, f_j\right) \in A$.

Generally, the phrase-based models are generative models that translate sequences of words in f_j into sequences of words in e_j, 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. 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. If $a \mid e \sim ToUniform\,(a; I+1)$, then

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

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 of f for segmented morphemes. This tag is an information about the word and represents lexeme after the 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):

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

We applied EM algorithm to estimate the phrase pairs that are consistent with the word alignments, and then assign probabilities to the obtained phrase pairs. The probability p_k of the word w to the corresponding context k is:

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

where, ϕ is the covariance matrix, and f are certain component density functions, which evaluated at each sequence. Consecutive word subsequences in the sentence pair are not longer than w words. After we use association measures to filter infrequently occurring phrase pairs by log likelihood ratio estimation [17].

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 alignment processing all valid phrases have to be stored in the phrase table and should be passed further.

2.2 Morphological Segmentation

Kazakh is a morphologically complex language with many differences from English. We describe here the main grammar features of Kazakh that are relevant to its English translation and mostly are associated separately in English by a different order. Case suffixes are attached to the noun in Kazakh often represent the preposition in English, also the word order is pretty challenging in the context of translation to English. For Kazakh noun phrases, which correspondence to English phrases may lead to the long phrases problem that exceed the size of phrases in a phrase table.

Our job usually starts from word segmentation, which includes running morphological tools to each entry of the phrase pair. At the first step, an word segmentation process aims to get suffixes and roots from the word. Therefore, we take surface forms of the words and generate their all possible lexical forms. Also we use the vocabulary 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[1–8], some of which have described in the following Table 2.

Table 2 The segmentation schemes

Id	Schema	Examples	Translation
S1	stem	el	A nation
S2	stem+case	el + ge	To the nation
S3	stem+num+case	el + der + den	From the nation
S4	stem+poss+	el + in	His/her nation
S5	stem+poss+case	el + i +ne	To his/her nation
S6	stem+num+poss+case+case	el+der+in+de+gi	It is within their nation
S7	stem+tense	oina+dy	Played
S8	stem+suffixN+suffixN+case	oina+gan+dyk+tan	Because he/she has played

There are large amount of verbs presenting ambiguity during segmentation, which do not take personal endings, but follow conjugated main verbs. 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. In fact, there are lack of syntactic information we cannot easily distinguish among similar cases.

In order to solve the problems represented above, we have to split up Kazakh words into the morphemes and some tags which represent the morphological information expressed on the suffixation. Splitting Kazakh words in this way, we expect to reduce the sparseness produced by the agglutination being of Kazakh and the drought of training data. Anyway, the segmentation model takes into account the several segmentation options of both sides of the parallel corpus while looking for the optimal segmentation. As we discovered, words with same Part-Of-Speech (POS) tags often correspond to each other in the word alignment and may help to efficiently handle out-of-vocabulary (OOV) words by incorporating linguistic information, but it can also make the training data more sparse [18]. We also suppose that the discovery of word context relations could lead to better word alignment scores and we apply this idea using a heuristic algorithm for every single training scenarios.

To define the most convenient segmentation for our Kazakh-English system, we checked most of the segmentation options and have measured their impact on the translation quality. This application of morphological processing aims to find several best splitting options that the each Kazakh phrase ideally corresponds to one English phrase, so the deep analysis is more desirable.

3 Evaluation

For evaluation the system, three samples of text data were processed with 50k sentences each one, which were used in raw form and with special segmentation. The expert decisions about a segmentation quality were defined by our university undergraduate students. The data samples were stored randomly into a training set and a test set had one sample for each of the phrase-based Moses [19] system run. After the most of the samples were found processed correctly, which means the same interpretation of data was selected as acceptable by the experts, we decided the system was trained well, and that is a good result.

Our corpora consists of the legal documents from http://adilet.zan.kz, a content of http://akorda.kz, and Multilingual Bible texts, and the target-side language models were trained on the MultiUN [20] corpora. 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. We measure the accuracy of the alignment using

Table 3 The performance of word alignment on 50 K

Alignment	Precision	Recall	F-score
Intersection	83.0	40.8	59.0
Union	43.1	61.0	55.0
Grow-diag	51.0	57.2	54.0
Grow-diag-final	42.8	70.0	51.5

Table 4 Best performance scores

System	Precision	Recall	F-score	AER	BLEU	METEOR	TER
Baseline	57.18	28.35	38.32	36.22	30.47	47.01	49.88
Morfessor	71.12	28.31	42.49	20.19	31.90	47.34	49.37
Rule-based	89.62	29.64	45.58	09.17	33.89	49.22	48.04

precision, recall, and F-measure, as given in the equations below; here, A represents the reference alignment; T, the output alignment; A and T intersection, the correct alignments (Table 3).

$$pr = \frac{|A \cap T|}{|T|}, re = \frac{|A \cap T|}{|A|}, F - measure = \frac{2 \times pr \times re}{pr + re} \tag{6}$$

The alignment error rate (AER) values for the trained system show distinct tendencies which were consistent through the iteration of different training parameters. The values show completely the higher rates for raw lexeme than for segmented one, which seems suitable for an alignment task. Another tendency is that the differences of context receive smaller impact than the precision of segmentation. This was not clear since removing or normalization causes a change in word structure. A problem in interpreting these training results depend on the scaling of the morpheme probability, which can be of different variation, and the scale needs to be appropriate to the text domain and segmentation schemes. We assume that phrase alignment connects word classes rather than words. Consequently, the phrase translation table has to be learned directly from phrase alignment models, and an estimation of phrase distribution probability is internally part of the process (Table 4).

The system parameters were optimized with the minimum error rate training (MERT) algorithm [21], and evaluated on the out-of and in-domain test sets. All 5-gram language models were trained with the IRSTLM toolkit [22] and then were converted to binary form using KenLM for a faster execution [23]. The translation performance scores were computed using the MultEval [24]: BLEU, TER [25] and METEOR [26]; and we ran Moses several times per experiment setting, and report the best BLEU/AER combinations obtained. Our survey shows that translation quality measured by BLEU metrics is not strictly related with lower AER.

4 Conclusion and Future Work

In this paper, we learned the effect of morphological processing on SMT by making the source and target languages more similar than they usually are. The methods we use to solve most common problems are implemented as a pre-processing steps script and a middle-tier component for word alignment processing. As far as we know, dealing with nominal agglutination only does not considerable change the BLEU score of the baseline translation. However, we expected the combination of morphological analysis and phrase table refining have a positive effect on translation quality. As a result, our experiments produced not only more perfectly matching phrases, but also obtained new alignments that did not produce from the training data. Taking a closer look, we found that morphological features extracted from the source language are a valuable resource for alignment prediction. Our evaluation shows that morphological processing leads to better translations where the quality can not be measured by BLEU score. The improved model performs at slightly the same speed as the previous one, and gives an increase of about 3 BLEU over baseline translation. I think that it is a demonstration of the potential of word alignments for SMT quality, and we plan to investigate more complicated methods in the future researches, possibly adding the new alignment features to the model.

References

1. Bekbulatov, E., Kartbayev, A.: A study of certain morphological structures of Kazakh and their impact on the machine translation quality. In: IEEE 8th International Conference on Application of Information and Communication Technologies, pp. 1–5. Astana (2014)
2. Oflazer, K., El-Kahlout, D.: Exploring different representational units in English-to-Turkish statistical machine translation. In: 2nd Workshop on Statistical Machine Translation, pp. 25–32. Prague (2007)
3. Bisazza, A., Federico, M.: Morphological pre-processing for Turkish to English statistical machine translation. In: International Workshop on Spoken Language Translation 2009, pp. 129–135. Tokyo (2009)
4. Kartbayev, A.: SMT: A case study of Kazakh-English word alignment. In: Current Trends in Web Engineering, pp. 40–49. Springer, Heidelberg (2015)
5. Moore, R.: Improving IBM word alignment model 1. In: 42nd Annual Meeting on Association for Computational Linguistics, pp. 518–525. Barcelona (2004)
6. Brown, P.F., DellaPietra, V.J., DellaPietra, S.A., Mercer, R.L.: The mathematics of statistical machine translation: parameter estimation. In: Computational Linguistics, vol. 19, pp. 263–311. MIT Press Cambridge, MA (1993)
7. Vogel, S., Ney, H., Tillmann, C.: HMM-based word alignment in statistical translation. In: 16th International Conference on Computational Linguistics, pp. 836–841. Copenhagen (1996)
8. Dempster, A., Laird, N., Rubin, D.: Maximum likelihood from incomplete data via the em algorithm. J. Roy. Stat. Soc. B **39**, 1–38. Wiley-Blackwell, UK (1977)
9. Creutz, M., Lagus, K.: Unsupervised models for morpheme segmentation and morphology learning. In: ACM Transactions on Speech and Language Processing, vol. 4, article 3. Association for Computing Machinery, New York (2007)
10. Beesley, K.R., Karttunen, L.: Finite State Morphology. CSLI Publications, Palo Alto (2003)

11. Goldsmith, J.: Unsupervised learning of the morphology of a natural language. In: Computational Linguistics, vol. 27, pp. 153–98. MIT Press, Cambridge (2001)
12. Altenbek, G., Xiao-long, W.: Kazakh segmentation system of inflectional affixes. In: CIPS-SIGHAN Joint Conference on Chinese Language Processing, pp. 183–190. Beijing (2010)
13. Kairakbay, B.: A nominal paradigm of the Kazakh language. In: 11th International Conference on Finite State Methods and Natural Language Processing, pp. 108–112. St. Andrews (2013)
14. Linden, K., Silfverberg, M., Axelson, E., Hardwick, S., Pirinen, T.A.: HFST—Framework for compiling and applying morphology. In: Systems and Frameworks for Computational Morphology, pp. 67–85. Springer, Heidelberg (2011)
15. Och, F.J., Ney, H.: A systematic comparison of various statistical alignment models. In: Computational Linguistics, vol. 29, pp. 19–51. MIT Press, Cambridge (2003)
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, pp. 311–318. Philadephia (2002)
17. Dunning, T.: Accurate methods for the statistics of surprise and coincidence. In: Computational Linguistics, vol. 19, pp. 61–64. MIT Press, Cambridge (1993)
18. Lee, J.-H., Lee, S.-W., Hong, G., Hwang, Y.-S., Kim, S.-B., Rim, H.-C.: A post-processing approach to statistical word alignment reflecting alignment tendency between part-of-speeches. In: 23rd International Conference on Computational Linguistics, pp. 623–629. Beijing (2010)
19. 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, pp. 177–180. Prague (2007)
20. Tapias, D., Rosner, M., Piperidis, S., Odjik, J., Mariani, J., Maegaard, B., Choukri, Kh., Calzolari, N.: MultiUN: a multilingual corpus from united nation documents. In: Seventh conference on International Language Resources and Evaluation, pp. 868–872. La Valletta (2010)
21. Och, F.J.: Minimum error rate training in statistical machine translation. In: 41st Annual Meeting of the Association for Computational Linguistics, pp. 160–167. Sapporo (2003)
22. Federico, M., Bertoldi, N., Cettolo, M.: IRSTLM: An open source toolkit for handling large scale language models. In: Interspeech 2008, pp. 1618–1621. Brisbane (2008)
23. Heafield, K.: Kenlm: faster and smaller language model queries. In: Sixth Workshop on Statistical Machine Translation, pp. 187–197. Edinburgh (2011)
24. 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, pp. 176–181. Portland (2011)
25. 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, pp. 223–231. Cambridge (2006)
26. 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, pp. 85–91. Edinburgh (2011)