Refining Kazakh Word Alignment Using Simulation Modeling Methods for Statistical Machine Translation

Amandyk Kartbayev(✉)

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

Abstract. Word alignment play an important role in the training of statistical machine translation systems. We present a technique to refine word alignments at phrase level after the collection of sentences from the Kazakh-English parallel corpora. The estimation technique extracts the phrase pairs from the word alignment and then incorporates them into the translation system for further steps. Although it is a pretty important step in training procedure, an word alignment process often has practical concerns with agglutinative languages. We consider an approach, which is a step towards an improved statistical translation model that incorporates morphological information and has better translation performance. Our goal is to present a statistical model of the morphology dependent procedure, which was evaluated over the Kazakh-English language pair and has obtained an improved BLEU score over state-of-the-art models.

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

1 Introduction

In this paper, we present the work done for improving a baseline statistical machine translation (SMT) system from an agglutinative Kazakh language to English. In this pair of translation, English word correspond to Kazakh suffixes that can fit more than one of the suffixes to the word. For instance, using the Kazakh lemma el - 'state' we can generate 'eldin' - 'of the state', 'elge' - 'to the state' and so on. The Kazakh language, which is the majority language in the Republic of Kazakhstan, has poor open resources and there are a small available parallel corpora unlike to other languages that more widely used as English or French. The parallel corpora for this work is 70k sentences for Kazakh and English, much bigger one than this corpus we had before.

In previous work[1], we described an approach to word alignment intended to address these problems. A research more relevant to that work was done by Bisazza and Federico[2]. The main goal of this research, different from the previous works, that is to make proposals that increase the expected benefit from

J. Li et al. (Eds.): NLPCC 2015, LNAI 9362, pp. 421–427, 2015.
DOI: 10.1007/978-3-319-25207-0_38

a word alignment estimation, rather than learning a morphology of Kazakh text for a processing activities. The given model is estimated from two principles: a morphological processing technique that gives the correct word segmentation, and a alignment model that determines the correct segment alignment in another case. Simulated results shows that this model has a potential to decrease a sparse words level, and to reach an overall consistency, which was observed during an evaluation.

The idea of using word alignment as a problem of determining correspondence at the word level was introduced by Brown et al.[3] is becoming as one of crucial components of all statistical machine translation approaches. Word alignment refinement can't be seen as a form of the relationship between word alignment quality and translation quality, what is well explained[4]. The common approaches of word alignment training are IBM Models and hidden Markov model (HMM)[5], which practically use expectation-maximization (EM) algorithm[6]. Compared to the word-based alignment models - in which a pair made of two words, each one from a different text that is certainly common to occur, but the words often are not a correct translation of each one - a phrase alignment model focuses on acquiring translations of phrases.

Phrase-based SMT systems usually train a phrase translation table, which may be produced after processing of word alignment and their probabilities for phrases. Phrase-based model has set of advantages over word-based alignment. At first, it naturally integrates context of the phrases and provide possibility to use these contexts in the translation. Eventually, a phrase is a consecutive sequence of words and the model allows the translation of unseen phrases unsupervised way. This makes the model generally applicable to similar language pairs we have learned.

We use Morfessor tool[7] to out grammatical features of word and can find the benefit of using morphological analysis in machine translation. We also explored rule-based morphological analyzer[8], which consist in deep language expertise and a exhaustive process in system development. For a comprehensive survey of the rule-based morphological analyze we refer a reader to the research by Altenbek[9] and Kairakbay[10].

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

2 Description of Our Method

Simulation modeling is too costly and time-consuming in this area of science as it exists. That we describe a new method that has to be more optimal and faster in some applications. In this approach, system starts a simple generic model and then incrementally replace its parts with more special pieces from a systematically organized processing components. Eventually, the system changes its subpart, and then automatically creates a new model or shows the step where further manual changes is necessary. For instance, we used the Helsinki Finite-State Toolkit (HFST)[11] to treat rule-based analyze and could to conduct a

study of its benefits for morpheme based alignment. Also we use the GIZA++[12] tool, which intersects two word alignments and get an union of the alignments, finally it produces a nearly symmetric result. Our study is based on the set of experiments, which have the goal of most properly extraction a phrase table from the word alignment. That actually leads to higher BLEU scores[13] and rises overall translation quality by reduction the level of sparse phrases.

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: $\forall e_i \in E : (e_i, f_j) \in a \Rightarrow f_j \in F$ and $\forall f_j \in F : (e_i, f_j) \in a \Rightarrow e_i \in E$, clearly, there are $\exists e_i \in E, f_j \in F : (e_i, f_j) \in A$.

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 $a \mid e \sim ToUniform\left(a; I+1\right)$, then

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

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(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):

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

We conduct an extensive experiment with using the descriptions mentioned above to construct dynamic models of alignment object in a SMT pipeline. The experiment included developing a comprehensive morpheme analysis that incorporates morphemes and text chunks to the phrase table. We create simulation

models by moving actual linguistic objects onto the bijective space constraints, how it could be used to form a symmetric matrix mxn. A Kazakh-English surjective space consists of two "word" vectors with a certain number of more elementary parts $\{e_i, f_j\}$, and a set of linkages $\{x_{ij}\}$ between these vector elements. We require that all individual elements able to change their links, so every element may be linked to other one or more. Also we use the term "matrix", when we emphasize packing the probabilities $p_{i,j}(x_{ij}|e_i, f_j)$ to the building block, and the term "link constraints" emphasize the components of some grammar, a simulation model.

For describing how we organize vector of words, it is useful to see some detailed schemes of morphological segmentation. Our morphological segmentation has run Morfessor tool and HFST to each entry of the corpus. Accordingly, we take surface forms of the words and generate their all possible lexical forms. The schemes presented below are different combinations of outputs determining the removal of affixes from the analyzed words. We mainly focused on detection of a few techniques for the segmentation of 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.

Table 1. 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

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. Because Kazakh derivational suffixes cannot occur freely, only in conjunction with a word stem, so each input word was reduced to its lemma and POS tagged word parts.

Since we have applied morphological segmentation, we will use word translation probabilities as random variables. This enables IBM Model 1 integration with an associated prior distribution. Dirichlet prior Θ is placed on word translation probabilities are basically a parameters t_{ij} of a linear relation function. In the mathematical notation, integration is represented as a rectangle, a stylized covariant matrix. A relation is represented as a symmetric row, like a pipe. A policy controlling a process is represented as the sub-policies are represented as labels connected by curved arrows. Every relation comes from one element and goes into another. An advantage of a matrix is that it simplifies the creation of new models by allowing one to build up a more specialized submatrixes by replacing elements in a smaller one.

Many of the constraints we have added through our previous experiments with this method are the complex side of its benefits. For instance, the fact that relations accumulate a context that in the some stages of using our process has an access to a hierarchy of context free grammar.

In some modeling cases, the real benefit of modeling comes from the numerical result of a simulation after deeper analyses of the finite state loops, which generate patterns of rules. Usually, analysis proceeds as slowly as creating the model. If there were no technology to allow finite state analysis, that an advantage of the modeling would be lost. Although we have not enough combined this feature with the other components of the system, our further approaches will provide a scalable flexibility in model analysis as well. The study of the effect of the model is pretty difficult because a morpheme ambiguity influences to the overall result so much.

3 Evaluation

We evaluate the SMT with the phrase-based Moses[14] 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, and the target-side language models were trained on the MultiUN[15] 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.

The model is implemented like a middle tier component, that 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.

Therefore, we expect the accuracy of the alignment will be measured using precision, recall, and F-measure, we present equations given in the below; here, A represents the reference alignment; T, the output alignment; A and T intersection, the correct alignments.

$$pr = \frac{|A \cap T|}{|T|}, re = \frac{|A \cap T|}{|A|}, F - measure = \frac{2 \times pr \times re}{pr + re} \tag{5}$$

The system parameters were optimized with the minimum error rate training (MERT) algorithm [16], and we trained 5-gram language models with the IRSTLM toolkit[17] and then were converted to binary form using KenLM for a faster execution[18].

Table 2 shows metric scores, which were computed using the MultEval[19]: BLEU, TER[20] and METEOR[21]; the survey shows that translation quality measured by BLEU metrics is not strictly related with lower AER. The final values show that the model can work consistently to give a greater improvement, despite the independent assumptions.

Table 2. Metric scores for all systems

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

4 Conclusions

In this work, we address a morpheme alignment problems concerned the Kazakh language. We compared our approach against a baseline of the Moses translation pipeline and have found it is able to obtain translation quality better than the baseline method by substantial level.

The system results can be transferred to other fields of application, where exists an alignment problem in natural language processing and the incorporation of word segments is useful. Subjects of future research include improvements in the phrase selection method and a context disambiguation. A special experiment with different learning methods may change the interpretation of the results. The improved model works at the same speed as the previous one, and gives an increase of about 3 BLEU in translation quality. This is a modest improvement, but we feel the potential of simulation modeling for this application, and we plan to conduct more sophisticated approaches in the future.

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, Astana, pp. 1–5 (2014)
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. Brown, P.F., DellaPietra, V.J., DellaPietra, S.A., Mercer, R.L.: The mathematics of statistical machine translation: Parameter estimation. Computational Linguistics **19**, 263–311 (1993). MIT Press Cambridge, MA
4. Moore, R.: Improving IBM word alignment model 1. In: 42nd Annual Meeting on Association for Computational Linguistics, Barcelona, pp. 518–525 (2004)
5. Vogel, S., Ney, H., Tillmann, C.: HMM-based word alignment in statistical translation. In: 16th International Conference on Computational Linguistics, Copenhagen, pp. 836–841 (1996)
6. Dempster, A., Laird, N., Rubin, D.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B **39**, 1–38 (1977). Wiley-Blackwell, UK
7. 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)

8. Beesley, K.R., Karttunen, L.: Finite State Morphology. CSLI Publications, Palo Alto (2003)
9. 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)
10. 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)
11. Lindén, K., Axelson, E., Hardwick, S., Pirinen, T.A., Silfverberg, M.: HFST–framework for compiling and applying morphologies. In: Mahlow, Cerstin, Piotrowski, Michael (eds.) SFCM 2011. CCIS, vol. 100, pp. 67–85. Springer, Heidelberg (2011)
12. Och, F.J., Ney, H.: A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics **29**, 19–51 (2003). MIT Press, Cambridge, MA
13. 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)
14. 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. 177–180 (2007)
15. Tapias, D., Rosner, M., Piperidis, S., Odjik, J., Mariani, J., Maegaard, B., Choukri, K.h., 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)
16. 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)
17. 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)
18. Heafield, K.: Kenlm: faster and smaller language model queries. In: Sixth Workshop on Statistical Machine Translation, Edinburgh, pp. 187–197 (2011)
19. 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)
20. 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)
21. 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)