DBMS-KU Interpolation for WMT19 News Translation Task

Sari Dewi Budiwati[1,2], **Al Hafiz Akbar Maulana Siagian**[1,3],
Tirana Noor Fatyanosa[1], **Masayoshi Aritsugi**[4],
[1]Computer Science and Electrical Engineering
Graduate School of Science and Technology, Kumamoto University, Japan
[2]School of Applied Science, Telkom University, Indonesia
[3]Indonesian Institute of Sciences, Indonesia
[4]Big Data Science and Technology
Faculty of Advanced Science and Technology, Kumamoto University, Japan
{saridewi, fatyanosa, alha002}@st.cs.kumamoto-u.ac.jp, aritsugi@cs.kumamoto-u.ac.jp

Abstract

This paper presents the participation of DBMS-KU Interpolation system in WMT19 shared task, namely, Kazakh-English language pair. We examine the use of interpolation method using a different language model order. Our Interpolation system combines a direct translation with Russian as a pivot language. We use 3-gram and 5-gram language model orders to perform the language translation in this work. To reduce noise in the pivot translation process, we prune the phrase table of source-pivot and pivot-target. Our experimental results show that our Interpolation system outperforms the Baseline in terms of BLEU-cased score by +0.5 and +0.1 points in Kazakh-English and English-Kazakh, respectively. In particular, using the 5-gram language model order in our system could obtain better BLEU-cased score than utilizing the 3-gram one. Interestingly, we found that by employing the Interpolation system could reduce the perplexity score of English-Kazakh when using 3-gram language model order.

1 Introduction

This paper describes our participation in the WMT19 shared task. We call our system DBMS-KU (Database Management System - Kumamoto University) Interpolation as we use our laboratory and university name, as well as we utilize Interpolation method in our experiments. We choose news translation task and focus on Kazakh-English (and vice versa) language pair.

Kazakh-English is a new shared task for this year, that is, no experience system description from previous WMT. Kazakh-English could be considered as low resource language pair due to the limitation of parallel corpora and morphological tools. Another challenge is the difference in the writing system between Kazakh and English languages. Kazakh uses Cyrillic letters, while English uses the alphabet. Different writing system between language pair needs specific attention in the tokenization step because of its segmentation results that affect the BLEU-cased score. Thus, we are motivated to solve this intriguing and challenging task.

Kazakh to English machine translation has been explored in Statistical Machine Translation (SMT) (Assylbekov and Nurkas, 2014; Kuandykova et al., 2014; Kartbayev, 2015a,b) and Neural Machine Translation (NMT) (Myrzakhmetov and Kozhirbayev, 2018). Assylbekov and Nurkas (2014) have shown an interesting result that different n-gram and neural LSTM-based language models were able to reduce the perplexity score, i.e., giving better translation result. For this reason, we consider investigating different n-gram language model order in this work.

Interpolation has been used in Language Model (LM) (Allauzen and Riley, 2011; Liu et al., 2013; Heafield et al., 2016) and in Translation Model (TM) (Bisazza et al., 2011; Sennrich, 2012; Rosa et al., 2015). Also, the interpolation has been used in pivot language as a strategy to overcome the limitation of parallel corpora (Dabre et al., 2015; Hoang and Bojar, 2016; Kunchukuttan et al., 2017). Pivot strategy arises as a preliminary assumption that there are enough parallel corpora between source-pivot (SRC-PVT) and pivot-target (PVT-TRG) languages. Currently, English as lingua franca has more datasets compared to other languages. Thus, pivot researchers commonly use English as a bridge between source to target (Paul et al., 2013; El Kholy et al., 2013; Ahmadnia et al., 2017; Dabre et al., 2015; Trieu, 2017). However, Paul et al., (2013) and Dabre et al., (2015) have shown that using non-English as pivot language could be a better option to improve the translation results for particular language pair. Since Kazakh-English is categorized as low resource language pair, we adopt the pivot and interpolation strategies in our translation model.

Proceedings of the Fourth Conference on Machine Translation (WMT), Volume 2: Shared Task Papers (Day 1) pages 141–146
Florence, Italy, August 1-2, 2019.

In this work, we consider examining two systems, namely, Baseline and Interpolation. The Baseline system is a direct translation between each language pair, while Interpolation one is a combination of pivot and direct translation models. We use Russian as our pivot language with 3-gram and 5-gram language model orders in each system. Our experimental results are encouraging and indicate that using Interpolation system could obtain better BLEU-cased score than employing Baseline one when translating both Kazakh to English (KK-EN) and English to Kazakh (EN-KK).

This paper is organized as follows. Section 2 explains the data preprocessing and experiment setup for each system. Section 3 shows and discusses the obtained results. Section 4 provides the conclusion and future direction of this work.

2 Case Study and Experiment Setup

In this section, we describe the case study, dataset, and experiment of this study.

2.1 Kazakh to English Machine Translation

Kazakh language is an agglutinative and highly inflected language that belongs to the Turkic group (Makhambetov et al., 2013). This rich morphology leads to a different length of phrases when translating from English to Kazakh (Assylbekov and Nurkas, 2014). Therefore, the translation of KK-EN and vice versa is a challenging task. Moreover, the KK-EN is considered as low resource language pair due to the limitation of parallel corpora and morphological tools.

2.2 Data and preprocessing

We used a dataset provided by WMT19 organizer. Thus, our system was considered as a constrained system. To prepare parallel datasets, we cleaned the dataset by using our script because the original dataset had blank lines and unsynchronized sentences between source and target parallel corpora. In the Interpolation system, we used Russian-English dataset from WMT18. The dataset statistics of training *(train)* and development (*dev*) for Baseline and Interpolation systems are given in Table 1.

After cleaning the dataset, we followed dataset preprocessing as in (Myrzakhmetov and Kozhirbayev, 2018), namely, tokenizing, normalizing punctuation, recasing, and filtering the sentences. Tokenizing was used to separate the token and punctuation by inserting spaces. Our tokenization results were based on words. Thus, the obtained sentences of the tokenization results were longer than the original sentences. Since long sentences could cause problems in the training process, we removed the sentences with a length of more than 80 words. This process was called filtering the sentences. Normalizing punctuation was to convert the punctuation for being recognized by the decoder system. Recasing was to change the initial words into their most probable casing in order to reduce the data sparsity. All preprocessing steps were done by using scripts from Moses (Koehn et al., 2007).

2.3 Experiment setup

We used open source Moses decoder (Koehn et al., 2007) and Giza++ for word alignment, Ken-LM (Heafield, 2011) for language model, and MERT (Och, 2003) for tuning the weight. The translation results were measured by five automatic evaluations provided by the organizer, namely BLEU, BLEU-cased, TER, BEER 2.0, and CharacTER. However, in this paper, we used the BLEU-cased because it is the main comparison metric in the evaluation system[1].

We built two systems, namely, Baseline and Interpolation. The Baseline system is a direct translation between KK-EN and vice versa. Meanwhile, the Interpolation system is the combination of direct translation with pivot phrase table. Pivot phrase table was produced by merging the source to pivot (SRC-PVT) and pivot to target (PVT-TRG) by using Triangulation method (Hoang and Bojar, 2015). We built the Interpolation phrase table as follows:

- Constructing a phrase table from SRC-PVT and PVT-TRG systems and pruning the phrase table with *filter-pt* (Johnson et al., 2007). The pruning activity was intended to minimize the noise of SRC-PVT and PVT-TRG phrase tables.
- Merging two pruned phrase tables by using the Triangulation method (Hoang and Bojar, 2015). The result was called `TmTriangulate` phrase table.
- Combining `TmTriangulate` and direct translation model with *dev* phrase table as

[1] http://matrix.statmt.org/

Dataset	Sentences	Average Sentence Length	Vocab
Baseline system			
Train			
news-commentary-v14.en-kk.kk	9,619	18.0857	29,142
news-commentary-v14.en-kk.en	9,619	22.1487	16,742
Dev			
newsdev2019-enkk.kk	2,068	18.0164	11,389
newsdev2019-enkk.en	2,068	22.2316	7,726
Language Model			
news-commentary-v14.kk	12,707	17.2109	-
news-commentary-v14.en	532,560	21.5762	-
Interpolation system			
Train			
news-commentary-v14.kk-ru.ru	7,230	23.6836	27,819
news-commentary-v14.kk-ru.kk	7,230	20.1187	24,627
news-commentary-v14.en-ru.en	97,652	23.0416	51,566
news-commentary-v14.en-ru.ru	97,652	21.3508	126,476
Dev			
news-commentary-v14.kk-ru.ru	2,000	20.8755	11,841
news-commentary-v14.kk-ru.kk	2,000	18.048	10,561
newstest2018-ruen.dev.en	3,000	20.975	10,108
newstest2018-ruen.dev.ru	3,000	17.3293	17,091
Language Model			
news-commentary-v14.kk	12,707	17.2109	
news-commentary-v14.en-ru.ru	114,375	21.2678	
news-commentary-v14.en-ru.en	114,375	22.9811	

Table 1: Dataset statistic for Baseline and Interpolation systems

Language Pair	3-gram LM	5-gram LM
KK-EN		
1. Baseline system	2.6	2.9
2. Interpolation system	2.7	3.4
EN-KK		
1. Baseline system	0.8	0.8
2. Interpolation system	0.9	0.9

Table 2: BLEU-cased score results

references. We used linear interpolation with backoff mode and exploited *combine-ptables* tools (Bisazza et al., 2011). The result was called `Interpolation` phrase table.

3 Results and Discussions

In this section, we show the obtained automatic evaluation results using BLEU-cased score. We also discuss the effect of the different language model order with the BLEU-cased score. Furthermore, we analyze the perplexity score on Interpolation system.

3.1 Language model effects on BLEU-cased score

In this paper, we conducted experiments for two language model orders, i.e., 3-gram and 5-gram, and two systems, viz., Baseline, and Interpolation. As shown in Table 2, the 5-gram language model order had more significant influence than the 3-gram one on the BLEU-cased score for KK-EN translation in both Baseline and Interpolation systems. The improvement in KK-EN was obtained by +0.3 and +0.7 points for Baseline and Interpolation systems, respectively. However, the BLEU-cased score for EN-KK could not be improved in terms of the language model order. These results might indicate that the language model order influenced the BLEU-cased score.

In terms of the translation system, the Interpolation system obtained higher BLEU-cased score than the Baseline one for all language model and translation directions. The improvement of BLEU-cased score from Baseline to Interpolation system for KK-EN using 3-gram and 5-gram was +0.1 and +0.5 points, respectively. Meanwhile, the improvement from Baseline to Interpolation System for EN-KK was +0.1 for both 3-gram and 5-gram orders. These results indicated that the use of pivot language in the Interpolation system combined with longer language model also had a significant influence on the BLEU-cased score.

Also, we found that the KK-EN obtained higher BLEU-cased score than the EN-KK in terms of the translation direction. This result might be influenced by the number of target LM datasets in each translation direction. As shown in Table1, KK-EN had 532,560 sentences, while EN-

KK had 12,707 sentences. The translation direction of KK-EN, that is, having almost 42 times larger number of sentences than EN-KK, could obtain a higher BLEU-cased score than that of EN-KK. This result indicated that the number of the target LM dataset in the experiments might be able to improve the BLEU-cased score.

Although our obtained BLEU-cased score was relatively low, we showed that by combining Baseline and pivot parallel corpora with different LM order was a valuable effort compared with using direct parallel corpora only. Moreover, the improvement of BLEU-cased score could be influenced by the language model order, the translation system, and the target monolingual LM dataset.

3.2 Perplexity effects on Interpolation system

Language model (LM) is one of the SMT components to ensure how good is the model by using perplexity as measurement. Lower perplexity score indicates better language models, while high perplexity score represents that the language model has poor quality. We show the perplexity score of the target language test dataset according to each n-gram language model trained on the respective training dataset in Table 3.

As shown in Table 3, the lowest perplexity score for KK-EN was obtained by the 5-gram Baseline system, i.e., 45.51. Thus, the best model for KK-EN was 5-gram Baseline system. However, we found that the difference of perplexity score for 5-gram model between Baseline and Interpolation systems was not quite significant, i.e., 5.42. Specifically, the perplexity of 5-gram of Baseline was 45.51, while the perplexity of 5-gram of Interpolation was 50.93. This finding might indicate that pivot language with interpolation system could be a beneficial approach in the translation process.

In EN-KK, the lowest perplexity score was obtained by 5-gram Baseline system, i.e., 77.18. Thus, the best model for EN-KK was 5-gram Baseline system. However, we found that the difference of perplexity score between 5-gram Baseline and 3-gram Interpolation systems was not quite significant, i.e., 2.16. Specifically, the perplexity of 5-gram of Baseline was 77.18, while the perplexity of 3-gram of Interpolation was 79.34. This finding might indicate that using the interpolation system with 3-gram model only could reduce the perplexity score of EN-KK that using the longer n-gram language model, i.e., 5-gram. Nevertheless, it would be better to study further the cause of this finding in the future.

4 Conclusion and future work

We examined the effect of different LM order with linear interpolation method for participating in WMT19 shared task, namely, Kazakh-English language pair. Our Interpolation system utilized the combination of direct translation, i.e., Baseline, with Russian as our pivot language. We used 3-gram and 5-gram language model orders in our Baseline and Interpolation systems. The BLEU-cased score of using Interpolation system could outperform that of utilizing Baseline one. This good performance of Interpolation system was obtained by using 3-gram and 5-gram language model orders for both Kazakh to English (KK-EN) and English to Kazakh (EN-KK) translations. We found that the Interpolation system indicated a different effect on each of KK-EN and EN-KK in terms of the perplexity score. In KK-EN, the pivot language with interpolation system could be an option in the translation process because the difference of perplexity score between Baseline and Interpolation was not quite significant. Interestingly, we found that the Interpolation system using 3-gram language model order could reduce the perplexity score compared with utilizing longer n-gram one in EN-KK.

In this shared task, we used standardized tokenizer from Moses. In the future, it must be worthwhile to use specific Kazakh and Russian tokenizers as their results will affect the BLEU-case scored. Another pivot language that has the same language family or has the same word order with the Kazakh language could also be a valuable effort. In addition, the use of different n-gram can also be taken into account for the next future research. Furthermore, the utilization of morph-based language modeling can also be applied to the system. Finally, the different interpolation scheme in another MT model, i.e., NMT, with out-domain dataset should be investigated to overcome the sparse of Kazakh resources.

Acknowledgments

A. H. A. M. Siagian would like to express his gratitude for the scholarship support from RISET-Pro (Research and Innovation in Science and Technology Project) KEMENRISTEKDIKTI (Ministry of

Language pair	3-gram LM	5-gram LM
KK-EN		
1. Baseline system	- Incl OOVs: 829.59 - Excl OOVs: 77.79	- Incl OOVs: 617.36 - Excl OOVs: 45.51
2. Interpolation system	- Incl OOVs: 1034.50 - Excl OOVs: 94.72	- Incl OOVs: 762.79 - Excl OOVs: 50.93
EN-KK		
1. Baseline system	- Incl OOVs: 328.940 - Excl OOVs: 103.27	- Incl OOVs: 256.138 - Excl OOVs: 77.185
2. Interpolation system	- Incl OOVs: 256.13 - Excl OOVs: 79.34	- Incl OOVs: 276.85 - Excl OOVs: 85.40

Table 3: Perplexity results

Research, Technology and Higher Education of the Republic of Indonesia).

References

Benyamin Ahmadnia, Javier Serrano, and Gholamreza Haffari. 2017. Persian-spanish low-resource statistical machine translation through english as pivot language. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017, Varna, Bulgaria, September 2 - 8, 2017*, pages 24–30.

Cyril Allauzen and Michael Riley. 2011. Bayesian language model interpolation for mobile speech input. In *INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27-31, 2011*, pages 1429–1432.

Zhenisbek Assylbekov and Assulan Nurkas. 2014. Initial explorations in kazakh to english statistical machine translation. In *Proceedings of the The First Italian Conference on Computational Linguistics CLiC-it 2014*, page 12.

Arianna Bisazza, Nick Ruiz, and Marcello Federico. 2011. Fill-up versus interpolation methods for phrase-based SMT adaptation. In *2011 International Workshop on Spoken Language Translation, IWSLT 2011, San Francisco, CA, USA, December 8-9, 2011*, pages 136–143.

Raj Dabre, Fabien Cromieres, Sadao Kurohashi, and Pushpak Bhattacharyya. 2015. Leveraging small multilingual corpora for smt using many pivot languages. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1192–1202. Association for Computational Linguistics.

Ahmed El Kholy, Nizar Habash, Gregor Leusch, Evgeny Matusov, and Hassan Sawaf. 2013. Language independent connectivity strength features for phrase pivot statistical machine translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 412–418. Association for Computational Linguistics.

Kenneth Heafield. 2011. KenLM: faster and smaller language model queries. In *Proceedings of the EMNLP 2011 Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland, United Kingdom.

Kenneth Heafield, Chase Geigle, Sean Massung, and Lane Schwartz. 2016. Normalized log-linear interpolation of backoff language models is efficient. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*.

Duc Tam Hoang and Ondrej Bojar. 2015. Tmtriangulate: A tool for phrase table triangulation. *Prague Bull. Math. Linguistics*, 104:75–86.

Duc Tam Hoang and Ondrej Bojar. 2016. Pivoting methods and data for czech-vietnamese translation via english. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation, EAMT 2017, Riga, Latvia, May 30 - June 1, 2016*, pages 190–202.

Howard Johnson, Joel Martin, George Foster, and Roland Kuhn. 2007. Improving translation quality by discarding most of the phrasetable. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*.

Amandyk Kartbayev. 2015a. Learning word alignment models for kazakh-english machine translation. In *Integrated Uncertainty in Knowledge Modelling and Decision Making - 4th International Symposium, IUKM 2015, Nha Trang, Vietnam, October 15-17, 2015, Proceedings*, pages 326–335.

Amandyk Kartbayev. 2015b. SMT: A case study of kazakh-english word alignment. In *Current Trends in Web Engineering - 15th International Conference, ICWE 2015 Workshops, NLPIT, PEWET, SoWEMine, Rotterdam, The Netherlands, June 23-26, 2015. Revised Selected Papers*, pages 40–49.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran,

Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180. Association for Computational Linguistics.

Ayana Kuandykova, Amandyk Kartbayev, and Tannur Kaldybekov. 2014. English -kazakh parallel corpus for statistical machine translation. In *International Journal on Natural Language Computing (IJNLC)*, page 65.

Anoop Kunchukuttan, Maulik Shah, Pradyot Prakash, and Pushpak Bhattacharyya. 2017. Utilizing lexical similarity between related, low-resource languages for pivot-based SMT. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 283–289, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Xunying Liu, Mark John Francis Gales, and Philip C. Woodland. 2013. Use of contexts in language model interpolation and adaptation. *Computer Speech & Language*, 27(1):301–321.

Olzhas Makhambetov, Aibek Makazhanov, Zhandos Yessenbayev, Bakhyt Matkarimov, Islam Sabyrgaliyev, and Anuar Sharafudinov. 2013. Assembling the Kazakh language corpus. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1022–1031, Seattle, Washington, USA. Association for Computational Linguistics.

Bagdat Myrzakhmetov and Zhanibek Kozhirbayev. 2018. Extended language modeling experiments for kazakh. In *Proceedings of 2018 International Workshop on Computational Models in Language and Speech, CMLS 2018*. CEUR-WS.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of Association for Computational Linguistics - Volume 1*, ACL '03, pages 160–167, Stroudsburg, PA, USA. Association for Computational Linguistics.

Michael Paul, Andrew Finch, and Eiichiro Sumita. 2013. How to choose the best pivot language for automatic translation of low-resource languages. *ACM Trans. Asian Lang. Inf. Process.*, 12(4):14:1–14:17.

Rudolf Rosa, Ondrej Dusek, Michal Novak, and Martil Popel. 2015. Translation model interpolation for domain adaptation in tectomt. In *Proceedings of the 1st Deep Machine Translation Workshop (DMTW 2015)*, volume 27, pages 89–96.

Rico Sennrich. 2012. Perplexity minimization for translation model domain adaptation in statistical machine translation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 539–549, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hai-Long Trieu. 2017. *A Study on Machine Translation for Low-Resource Languages*. Ph.D. thesis, Japan Advanced Institute of Science and Technology.