Mongolian-English, English-Mongolian independent Neural Machine Translation System

Bat-Erdene Batsukh

School of Information and Communication Management, University of the Humanities, 14200, Ulaanbaatar, Mongolia.
bbat-erdene@humanities.mn

Abstract

For more than a decade, PMT and SMT models have dominated the field of machine translation, and neural machine translation has emerged as a new paradigm for machine translation. The latest neural machine translation not only performs better than systems that consider the structure of ordinary words and sentences, but is also able to find complex relationships between source and target words. Neural machine translation provides a simpler modeling mechanism that makes it easier to use in practice and science. Neural machine translation no longer requires concepts such as word rank, which is a key component of a system that takes into account word and sentence structure. While this simplicity can be seen as an advantage, on the other hand, the lack of careful wording is a loss of control over translation. Systems that take into account word and sentence structure generate translations that consist of word sequences in the training data. On the other hand, neural machine translation is more flexible for translation that does not exactly match the training data. This provides more opportunities for such models, but frees the translation from predefined constraints. Lacking a specific word connection can make it difficult to link the target words you create to the source word. The widespread use of neural machine translation in translation systems has the advantage of allowing users to translate certain terms and translate untrained data to a certain extent, but in some cases often results in distorted sentence structure. This paper aims to address issues such as neural machine translation control, more accurate translation of unrecognized data, correct sentence structure and grammar boundaries, and the creation of independent machine translation system.

Keywords: Mongolian translation, NMT, SMT, Grammar boundary, Hierarchical triple model

1. Introduction

Neural machine translation is a data-driven approach. To translate in this way, a neural network model is used to receive the original sentence as an input and return the target sentence. The first attempts at neural machine translation began in 2013, and by 2015, neural machine translation was recognized as a new paradigm. Compared to the structure that takes into account word and sentence structure, neural machine translation does not require additional intermediate steps, such as word correlation, and produces direct results using a trained model. In addition, neural machine translation performs better than systems that take into account word and sentence structure, especially if the ordered bilingual learning data is sufficient. Although neural network models are

statistical models, neural machine translation is often different from statistical machine translation. Based on this scenario, we have chosen this topic because we have not yet developed a system for Mongolian-English, English-Mongolian independent translation system.

2. Related works

We will consider two different methods of machine translation: first, machine translation that takes into account phrase-based machine translation (PMT) (Koehn et al., 2003) and statical machine translation (SMT) (Brown et al., 1990), and second, neural machine translation (NMT) (Kalchbrenner and Blunsom, 2013), (Tan et al., 2020). Statistical machine translation systems are based on the models proposed in Koehn et al. (2003) and the approach discovered by Vogel et al. (1996). These models vary in context (Vauquois, 1968). Simple models are based solely on the word being translated, but may include more complex concepts for modeling the number of words in one language and the number of words derived from a translation in another language. All of these models are word-based and generate one word per step. Later, a model approach to phrase was proposed by Och and Ney (2000), which laid the foundation for a translation paradigm that takes into account phrase and sentence structure (Brown et al., 1990), (Zens & Ney, 2008). These systems have been widely used as the most advanced machine translation systems for more than a decade, until the introduction of neural machine translation. Models that take into account word and sentence structure differ from word-based models in that they score a whole phrase at each step. For example, "What are you doing right now?" Let's take the sentence via Bayesian decision rules using the minimum error rate training (Och, 2003), each word is described as follows (see Fig. 1).

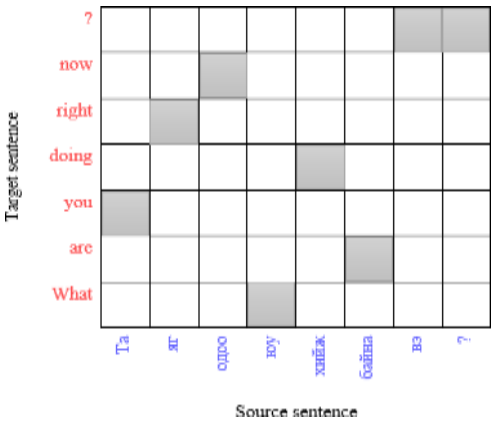

Figure 1: Word alignment

Sentence endings do not need to be taken into account when determining sentence structure and scope. We define this range using an algorithm developed by Wang and Huang (2003) at Stanford University. Word-based models must model a long context to generate such a sentence, and the search must be flexible enough not to stop the partial assumptions that lead to such a translation (see Fig. 2).

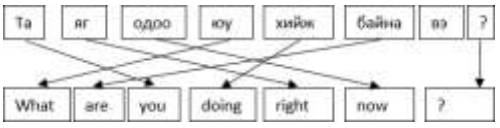

Figure 2: Word based model

However, phrase-based systems that take into account word and sentence structure are sufficient to store such entries in the sentence table. During the search, all expressions can be assumed to be a single atomic unit (see Fig. 3).

Figure 3: Phrase based model

3. Mongolian-English, English-Mongolian independent neural machine translation system

Auli (2013) suggests keeping hidden recurrent states in the search state, and suggests a way to reconcile the states and decide whether they are correct when comparing search states. Although state reassembly is not abstract, the method of repeating the model is used to approximate the iteration of the node, as it only retains the latent state corresponding to the best path when the node is reassembled. Schwenk (2012), on the other hand, uses transfer models to calculate additional language scores, while Le et al. (2012) use short lists to evaluate translation models using class-based output layers and transmission networks. Kalchbrenner & Blunsom (2013) have used recurrent neural networks to describe the original sentence obtained by using sequential alignments in the source sentence. Textual descriptions fall into the hidden layer of repetition on target words. The best translation is created by segmenting all possible translations and their key phrases. In practice, this type of search does not have an exact tag, and a similar search procedure is used to find it. For example, if the source sequence of sentences in a text of length K is $M = m_1^K = m_1 m_2 \dots m_K$, then the corresponding MOSE format, or the sequence of sentences in the target language corresponding to the same length L, must be $E = e_1^L = e_1 e_2 \dots e_L$. In our case, we want to translate from Mongolian to English, we get a (M, E) ranked pair. Based on this, $t_1^L = t_1 t_2 \dots t_L$ is the alignment path of the position of each word in the target language to the position of the words in the target language, the position of each word in the target language to the position of the words in the target language $s_1^K = s_1 s_2 \dots s_K$, (Wang et al., 2017) and let $g_1^K = g_1 g_2 \dots g_K$ be the grammar and sentence boundary. Let A be the probability of the translation pattern, B the probability of the model of expression used in language modeling and BPE, and C the probability of the pattern of words, sentence structure, and sentence scope. Since we are looking for the best English sentence for a given Mongolian sentence, we need to find the best option for both A, B, and C. Existing neural network-based machine translation models have solved the problem of machine translation as a combination of these three models. In other words, it seeks to create a complex model that is interdependent. On the one hand, this makes it possible for every researcher to do and test machine translation, but it also requires a very high capacity for training machines. For us, however, we prefer a more modular device that requires less capacity. This is due to the lack of Mongolian translation in the field of machine translation, the lack of Mongolian vocabulary and sentence structure in the international UD, the lack of experiments with BPE, and the lack of high-capacity experimental equipment. By definition of probability, $P(B/A) = P_A(B)$ is the probability of event B under condition A. The model we are currently developing is a hierarchical version of the three models mentioned above, and the final translation is based on each of the independent models. In the future, each time a different condition is added to these models, it will be necessary to find the conditional probability of each. In this case, we can increase the condition to n by an increasing number as the hierarchical model, such as $A = A_1, B = A_2, C = A_3$, increases (Equation 3).

$$P(A) \cdot P\left(B/A\right) = P(B) \cdot P\left(A/B\right) \quad (1)$$

Since the above formula is valid, consider it for any $A_1, A_2, \dots, A_n$.

$$P(A_1 A_2 \dots A_n) = P\left({}^{A_n}/_{A_1 A_2 \dots A_{n-1}}\right) \cdot P(A_1 A_2 \dots A_{n-1}) \quad (2)$$

If this is repeated until $P(A_1)$, the probability of our model is as follows.

$$P(A_1 A_2 \dots A_n) = P\left({}^{A_n}/_{A_1 A_2 \dots A_{n-1}}\right) \cdot P\left({}^{A_{n-1}}/_{A_1 A_2 \dots A_{n-2}}\right) \cdot \dots \cdot P\left({}^{A_2}/_{A_1}\right) \cdot P(A_1) \quad (3)$$

Our system training based on simplified version of alignment based neural machine translation by Alkhouli et al. (2016). Main difference is in the search procedure we applied grammar and sentence boundary detection (Equation 4).

$$m_1^K \rightarrow \hat{e}_1^{\hat{L}}(m_1^K) = \underset{L, e_1^L}{arg\max} \max_{t_1^L} \left\{ \frac{1}{L} \left(\sum_{l=1}^{L} \lambda \log p\left(e_l \middle| e_1^{l-1}, t_1^l, g_1^K, m_1^K\right) + (1-\lambda) \log p\left(\Delta_l \middle| e_1^{l-1}, t_1^{l-1}, g_1^K, m_1^K\right) \right) \right\} \quad (4)$$

When modeling grammar and sentence boundary, the general relationship of sentences in Mongolian is first plotted. " Барак Обама Хавайд төрсөн." Given the sentence, the graph looks like this (see Fig. 4).

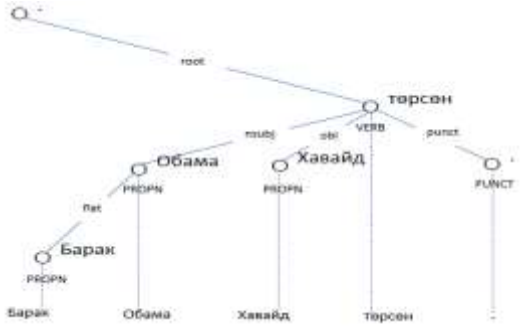

Figure 4: Dependency tree

For us, the UD, which combines Mongolian grammar and sentence boundaries, is inspired by Stanford's method (Dozat et al., 2017), which studies neural network-based words and sentence structures and relationships. For example, "Барак Обама Хавайд төрсөн." The Stanford dependency of the Mongolian

language is as follows (see Fig. 5).

Figure 5: Stanford dependency

When learning grammar and sentence boundary in a total of 1000 steps, sentence recognition loss was reduced to 0.02 (see Fig. 6).

Figure 6: Training loss reduction

During the training, development scores were automatically evaluated for every 200 steps, and the final development score reached 98,653 (see Fig. 7).

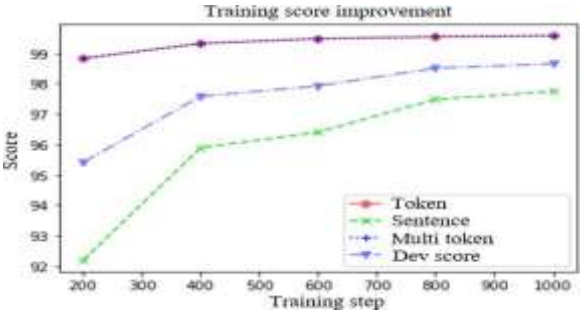

Figure 7: Training score improvement

By including this dependency in the search for neural translation model, we have become a gateway to better understanding of grammar and sentence boundary.

4. Methods and results

An attempt was made to integrate neural network results with a model that takes into account word and sentence structure, and for the first time proposed a model of re-alignment by changing the position of words (Wang et al., 2017). In practice, this integrated model of neural machine translation uses phrases to train neural networks. The difference between our experiments is that in this study, we selected three hierarchical models which including language model, translation model and grammar boundary model, the basic model of which was obtained using OpenNMT. During the development phase, each system component can be trained on a separate training corpus, but setting up the system on that data is too costly in terms of computation. Therefore, a separate development package (consisting of hundreds to thousands of original sentences and relevant reference translations) is used to optimize the log-linear design combination for optimal translation performance to avoid overloading. In our training, we created a local English-Mongolian mixed bilingual corpus by translating the following corpora (see Table. 1).

Table 1. Mongolian – English, English-Mongolian mixed bilingual corpus via back translation (Graça et al., 2019), (Cotterell and Kreutzer, 2018), (Edunov et al., 2018).

Corpus	File size	Translated sentences
United Nations Parallel Corpus (Ziemski et al., 2016)	3.44 gb	25,173,399
Wikimatrix (Schwenk et al., 2019)	227 mb	1,661,908
OpenSubtitles (Lison and Tiedemann, 2016)	832 mb	25,910,106

In order to present the results of the study more clearly and in more detail, we have considered some statistical indicators. The probability of translation was calculated by randomly sampling sentences from a set of 300,000 sentences not included in the training package to check how the quality of the translation depends on the coherence of the training data and the hierarchy model (see Table. 2).

Table 2. Sets of mixed bilingual corpus selected data

Corpus data		Mongolian	English
train	Sentence	2,402,138 line	
	Word	39,298,174	43,170,480
dev	Sentence	300,000 line	
	Word	4,895,610	5,378,414
test	Sentence	300,000 line	
	Word	4,893,721	5,382,746

The average number of words in the original sentences was 15.919942984124976, the average number of characters was 112.4927664438929, the smallest line consisted of 2 characters with 1 word, and the line with the most words consisted of 2149 words with 2,039,369 indexes. In order to present the results of the study more clearly and in more detail, we have considered some statistical indicators. A translation test using a hierarchical triple model-based system resulted in a 95% confidence interval of 0.9514 mean, a standard deviation of 0.0233, and a standard error of 0.0007.

The above experiments showed that a neural network-based triple hierarchy model translation quality was highly effective on top of bootstrap result (0.9465159565110001, 0.9560245841607502) (see Fig. 8).

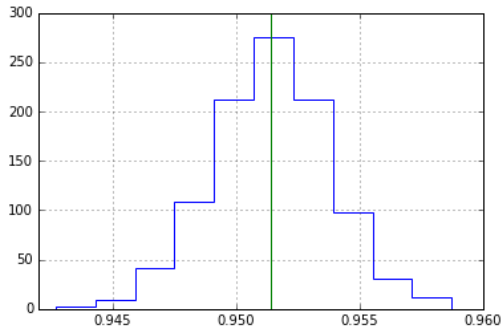

Figure 8: Bootstrap result

Based on this result, we compared our system to the best Mongolian-English, English-Mongolian neural machine translator currently in use (see Table. 1).

Table 3. Proposed system and Google translator comparison

Features	Our NMT	GNMT
Text length	unlimited	5000 chars
API	free	cloud service
File length	unlimited	2,400 pages
OpenOffice	full support	partial
Interactivity	typing	typed
Boundary	full support	partial
Connectivity	on/offline	online
Speed	mid-high	high

5. Conclusion

In recent times, the neural machine translation has become a new paradigm that will dominate the machine translation research and manufacturing market. In this sense, this type of translation model and systematic research have entered the field of applied and computational linguistics. The usage of neural machine translations individually or in two stages reduces system output controls on systems that take into account word, sentence structure and grammar boundaries, so we have developed triple model of order and neural machine translation systems to improve the boundaries of sentence grammar. The results of the neural network were then staged in a three-step model that worked by correctly defining the sentence grammar boundary by linking it to a pattern that took into account word and sentence structure. In addition, a neural machine translation or can generate direct output without waiting for a complete input sentence, allowing the user to translate directly or in real time.

Although the model we have developed has been successful in practical experiments, improvements need to be made to bring it into line with the speed of standard neural machine translation system.

References

Alkhouli, T., Bretschner, G., Peter, J.-T., Hethnawi, M., Guta, A., & Ney, H. (2016). Alignment-Based Neural Machine Translation. *Proceedings of the First Conference on Machine Translation: Volume 1, Research Papers*, 54–65. https://doi.org/10.18653/v1/W16-2206

Auli, M., Galley, M., Quirk, C., & Zweig, G. (2013). Joint Language and Translation Modeling with Recurrent Neural Networks. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1044–1054. https://aclanthology.org/D13-1106

Brown, P. F., Cocke, J., della Pietra, S. A., della Pietra, V. J., Jelinek, F., Lafferty, J. D., Mercer, R. L., & Roossin, P. S. (1990). A Statistical Approach To Machine Translation. *Computational Linguistics*, 79–85.

Dozat, T., Qi, P., & Manning, C. D. (2017).

Stanford's Graph-based Neural Dependency Parser at the CoNLL 2017 Shared Task. *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 20–30. https://doi.org/10.18653/v1/K17-3002

Kalchbrenner, N., & Blunsom, P. (2013). Recurrent Continuous Translation Models. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1700–1709. https://aclanthology.org/D13-1176

Koehn, P., Och, F. J., & Marcu, D. (2003). Statistical Phrase-Based Translation. *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, 127–133. https://aclanthology.org/N03-1017

Le, H. S., Allauzen, A., & Yvon, F. (2012). Continuous Space Translation Models with Neural Networks. *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 39–48. https://aclanthology.org/N12-1005

Lison, P., & Tiedemann, J. (2016). OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*. http://www.opensubtitles.org.

Och, F. J. (2003). Minimum Error Rate Training in Statistical Machine Translation. *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, 160–167. https://doi.org/10.3115/1075096.1075117

Och, F. J., & Ney, H. (2000). Improved Statistical Alignment Models. *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, 440–447. https://doi.org/10.3115/1075218.1075274

Schwenk, H. (2012). Continuous Space Translation Models for Phrase-Based Statistical Machine Translation. *Proceedings of COLING 2012*: Posters, 1071–1080. https://aclanthology.org/C12-2104

Schwenk, H., Chaudhary, V., Sun, S., Gong, H., & Guzmán, F. (2019). WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia. *CoRR, abs/1907.05791*. http://arxiv.org/abs/1907.05791

Tan, Z., Wang, S., Yang, Z., Chen, G., Huang, X., Sun, M., & Liu, Y. (2020). Neural machine translation: A review of methods, resources, and tools. *AI Open*, *1*, 5–21. https://doi.org/10.1016/j.aiopen.2020.11.001

Vauquois, B. (1968). A survey of formal grammars and algorithms for recognition and transformation in mechanical translation. *IFIP Congress*.

Vogel, S., Ney, H., & Tillmann, C. (1996). HMM-Based Word Alignment in Statistical Translation. *International Conference on Computational Linguistics*, 836–841.

Wang, H., & Huang, Y. (2003). *Bondec-A Sentence Boundary Detector*.

Wang, W., Alkhouli, T., Zhu, D., & Ney, H. (2017). Hybrid Neural Network Alignment and Lexicon Model in Direct HMM for Statistical Machine Translation. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 125–131. https://doi.org/10.18653/v1/P17-2020

Zens, R., & Ney, H. (2008). Improvements in Dynamic Programming Beam Search for Phrase-based Statistical Machine Translation. *International Workshop on Spoken Language Translation*, 195–205.

Ziemski, M., Junczys-Dowmunt, M., & Pouliquen, B. (2016). The United Nations Parallel Corpus v1.0. *Proceedings of the Tenth International Conference on*

Language Resources and Evaluation (LREC'16), 3530–3534. https://aclanthology.org/L16-1561