Asian Journal of Social Science Studies; Vol. 7, No. 3; 2022
ISSN 2424-8517 E-ISSN 2424-9041
Published by July Press

English-Mongolian, Mongolian-English Neural Machine Translation

Bat-Erdene Batsukh[1], Chuluundorj Begz[2] & Baigaltugs Sanjaa[3]

[1] School of Information and Communication Management, University of the Humanities, Ulaanbaatar, Mongolia

[2] University of the Humanities, Ulaanbaatar, Mongolia

[3] Mongolian University of Science and Technology, Mongolia

Correspondence: Bat-Erdene Batsukh, School of Information and Communication Management, University of the Humanities, Sukhbaatar square 20, Ulaanbaatar, Mongolia. Tel: 976-8805-9362. E-mail: bbat-erdene@humanities.mn

Received: February 5, 2022 Accepted: March 1, 2022 Online Published: March 28, 2022

doi:10.20849/ajsss.v7i3.999 URL: https://doi.org/10.20849/ajsss.v7i3.999

Abstract

The latest neural machine translation not only performs better than systems that consider simple words and sentence structures, but also finds a delicate connection between source and target words. Neural machine translation provides a simple modeling mechanism that is easy to use in practice and science. Thus, it does not require concepts such as word ranking, a key component of the system that takes into account the structure of words and sentences. While this simplicity may be seen as an advantage, on the other hand, the lack of careful spelling is to lose control of the translation. Systems that take into account the structure of words and sentences create translations consisting of word sequences in the curriculum data. On the other hand, the neural machine translation is more flexible in terms of translations that don't exactly match the training data. This provides more opportunities for such models, but exempts translation from pre-determined restrictions. Failure to connect specific words can make it difficult to connect the target words you create to the original word. The widespread use of neural networks in the translation system has the advantage of allowing users to translate certain terms and translate uneducated data to a certain extent. In some cases, however, the structure of a sentence is often distorted. The paper is intended to address issues such as the control of neural machine translation, more accurate translation of unidentified data, the accuracy of sentence structure and grammar boundaries.

Keywords: Mongolian cyrillic translation, nmt, smt, grammar boundary, hierarchical model

1. Introduction

In today's globalized world, translation plays a vital role in removing barriers to communication. The need for translation arises from the understanding, study, and expression of some kind of content prepared in a language other than one's native language, no matter where one is in the world. Thanks to social media platforms, users are more likely to view content written in other users' foreign languages. The need for translation is growing. Because professional translation is labor-intensive. Automatic translation, also known as machine translation, has played an important role in helping millions of users understand content written in a foreign language. Machine translation can be used not only by ordinary users for independent translation, but also to help professional translators translate faster. The new machine translation is a data-driven approach. To translate in this way, the neural network model is used to accept the original sentence as an entry and to reverse 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 structures that take into account the structure of words and sentences, the translation of a neural machine does not require additional intermediate steps, such as word dependency, and produces direct results using an accustomed model. In addition, neural machine translation performs better than systems that take into account the structure of words and sentences. If there is plenty of data to learn, especially in the two languages ordered. Although neural network-based models are statistically based models, neural machine translations are often different from statistical machine translations. Based on this example, we chose this topic because an independent English-Mongolian, Mongolian-English translation system has not yet been established.

2. Literature Review

Machine translation is the process of automatically translating text written in one native language into another. We can identify three different approaches to machine translation (Vauquois, 1968). First, the tendency to translate directly from the text into the target language. This approach focuses on translating one text to another, regardless of sentence meaning, syntax, or semantics. The second method is the "transfer method", which is a step-by-step translation between the text and the abstract representation of the target text. This abstract representation is obtained by analyzing the text. Text representation creates the final target text through a transfer step to create an abstract target representation. The translation of the text in this way takes place in the order of the analysis and translation of the text and the creation of the target text. The third method of machine translation is to translate the text into a non-linguistic representation between languages, and the target text is extracted from the abstract representation of all these languages. It can be divided into rule-based and data-based machine translation. Grammar-based methods focus on manually defined translation rules for a given bilingual. This method requires human knowledge and is usually expensive to obtain. On the other hand, data-based methods, such as statistical machine translation, do not require such human knowledge, but are based on data examples when modeling translations.

Statistical machine translation is a data-based approach developed in the late 1980s. Its main purpose is to develop a translation template that can be taught using a collection of texts and target texts. Statistical-based templates are used to translate text into the target language without the need for manually generated translation rules. Statistical machine translation often returns the most probable results based on trained words and phrases. Previous systems of machine translation based on statistics were word-based, and each translation step consisted of generating a single word. In the early 2000s, a system that considered word and sentence structure was proposed (Zens & Ney, 2008). These systems have been widely used for more than a decade as the latest machine translation systems. Later, neural network-based machine translation became the leading trend in machine translation. 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 & 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. 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 (Och & Ney, 2000), which laid the foundation for a translation paradigm that takes into account phrase and sentence structure (Brown et al., 1990). 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, "Where are you going right now?" Let's take the sentence Using Bayesian decision rules using the minimum error rate training (Och, 2003), each word is described as follows (see Figure 1).

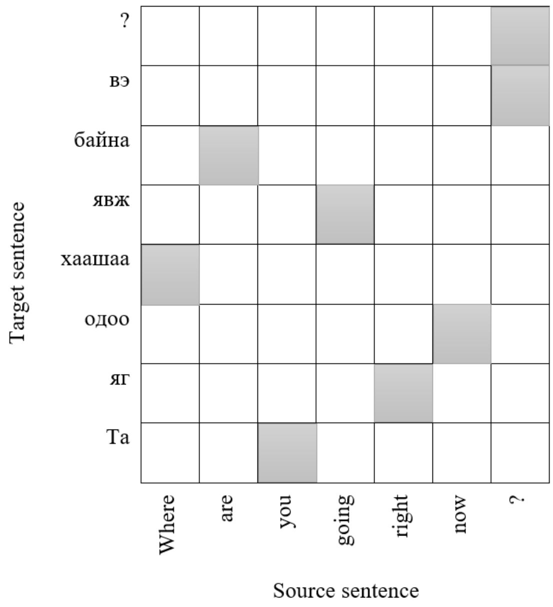

Figure 1. Word alignment

Word correlation is the word-level relationship between words in the original and target order. Usually, parallel syllables are not marked at the word level. Therefore, the word correlation is calculated automatically. The basic idea is that the correspondence of the words $T \subseteq \{1,2, ..., K\} \times \{1,2, ..., L\}$ is the correlation of the text indices $k \in \{1,2, ..., K\}$. and the indexes of the target sentence are $l \in \{1,2, ..., L\}$. An example of word matching is shown in Figure 14. Word correlations can be introduced as a sequence of hidden variables (Brown et al., 1993) (Vogel et al., 1996). Using this method, it is possible to define the word and sentence structure in more detail and incorporate it into the translation template. Sentence endings do not need to be taken into account when determining sentence structure and grammar boundary. We define this range using an algorithm developed by Stanford University (H. Wang & Huang, 2003). 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 Figure 2).

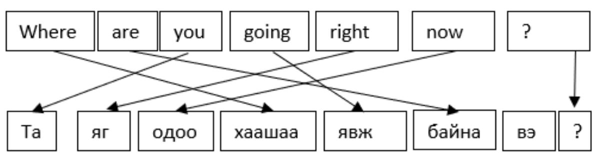

Figure 2. Word based Model

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 Figure 3).

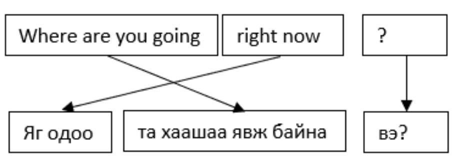

Figure 3. Phrase based model

The widespread use of neural machine translation has the advantage of allowing users to translate terms and untrained data to a certain extent, but in some cases the results often deviate from the sentence structure. For this reason, research into the design and improvement of neural machine translation models has been extensively conducted in the field of applied and computational linguistics in the form of combinations and hierarchies based on basic statistical machine translation models.

3. Method

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, if you want to translate from Mongolian to English, you get a ranked pair (M, E), and if you want to translate from English to Mongolian, you get a ranked pair (E, M). 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 $s_1^K = s_1 s_2 \dots s_K$ (W. 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 boundary. 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 (Equation 1).

$$m_1^K \to \hat{e}_1^{\hat{L}}(m_1^K) = \underset{L,e_1^L}{argmax}\{P_r(e_1^L|m_1^K)\}. \quad (1)$$

Existing neural 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 4).

$$P(A) \cdot P\left({}^{B}/_{A}\right) = P(B) \cdot P\left({}^{A}/_{B}\right) \quad (2)$$

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

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

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

$$P(A_1 A_2 \dots A_n) = P\left({}^{A_n}/_{A_1 A_2 \dots A_{n-1}}\right) \cdot P\left({}^{A_{n-1}}/_{A_1 A_2 \dots A_{n-2}}\right) \cdot \dots \cdot P\left({}^{A_2}/_{A_1}\right) \cdot P(A_1) \quad (4)$$

When modeling grammar and sentence boundary, the general relationship of sentences in Mongolian is first plotted. "Тэрээр 2008 онд ерөнхийлөгчөөр сонгогдсон." Given the sentence, the graph looks like this (see Figure 4).

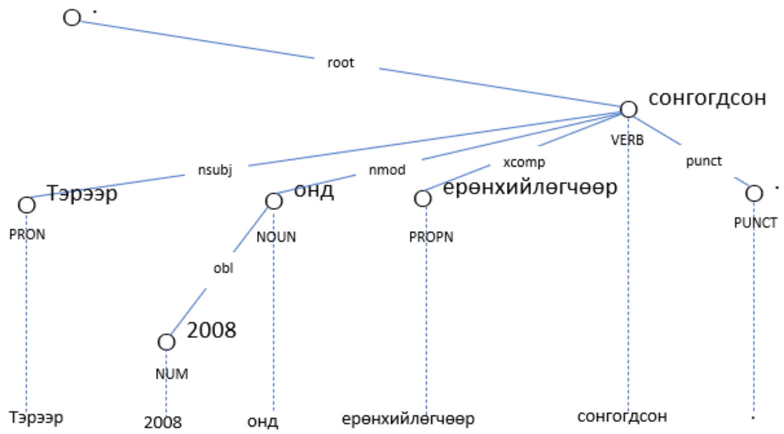

Figure 4. Dependency tree

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, "Тэрээр 2008 онд ерөнхийлөгчөөр сонгогдсон." The Stanford dependency of the Mongolian language is as follows (see Figure 5).

```
1   Тэрээр  тэр PRON    PRP Case=Nom|Gender=Masc|Number=Sing|Person=3|PronType=Prs  5   nsubj:pass  _   _
2   2008    2008    NUM CD  NumType=Card    3   obl _   SpaceAfter=No
3   онд он  NOUN    NN  Number=Sing 5   nmod    _   _
4   ерөнхийлөгчөөр  ерөнхийлөгч PROPN   NNP Number=Sing 5   xcomp   _   _
5   сонгогдсон  сонгох  VERB    VBN Tense=Past|VerbForm=Part|Voice=Pass 0   root    _   _
6   .   .   PUNCT   .   _   5   punct   _   _
```

Figure 5. Stanford dependency

When learning grammar and sentence boundary in a total of 500 steps, sentence recognition loss was reduced to 0.002 (see Figure 6).

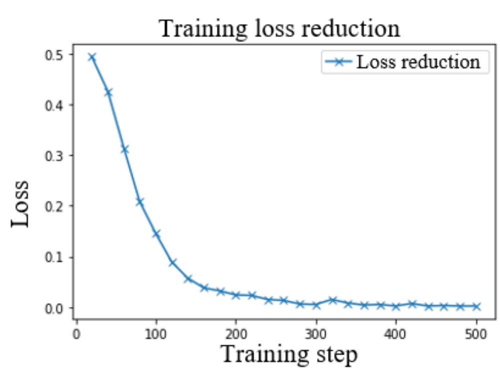

Figure 6. Loss reduction

By including this dependency in the search for neural translation model, we have become a gateway to better understanding of grammar and sentence boundary. Our system training based on hierarchical version of alignment based neural machine translation (Alkhouli et al., 2016). Only key difference is in the search procedure we applied grammar and sentence boundary detection (Equation 5).

$$m_1^K \rightarrow \hat{e}_1^{\hat{L}}(m_1^K) = \ argmax_{L,e_1^L} \max_{t_1^L} \left\{ \frac{1}{L} \left(\sum_{l=1}^{L} \lambda \, log \, p\left(e_l \middle| e_1^{l-1}, t_1^l, g_1^K, m_1^K\right) + (1-\lambda) \log p\left(\Delta_l \middle| e_1^{l-1}, t_1^{l-1}, g_1^K, m_1^K\right) \right) \right\} \quad (5)$$

4. Results and Discussion

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 (W. 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, the basic model of which was obtained using OpenNMT. During the development phase, each system component can be trained on a separate training corps, 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 United Nations Parallel Corpus (Ziemski et al., 2016), Wikimatrix (Schwenk et al., 2019), and OpenSubtitles (Lison & Tiedemann, 2016).

Table 1. English-Mongolian mixed bilingual corpus

Corpus data		Mongolian	English
train	Sentence	2,402,138 line	
	Word	39,298,174	39,298,174
dev	Sentence	2,402,138 line	
	Word	39,298,174	39,298,174
test	Sentence	2,402,138 line	
	Word	39,298,174	39,298,174

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 2,402,138 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.

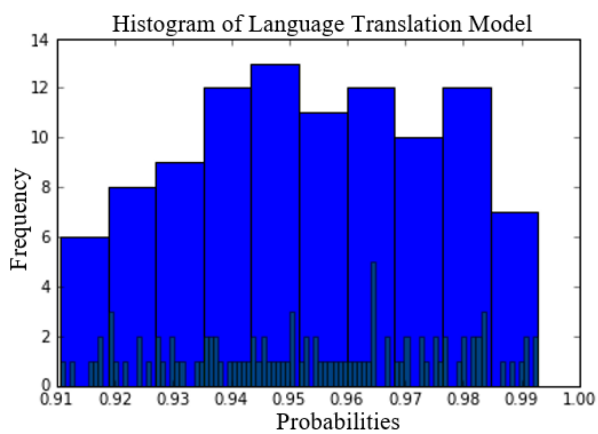

Figure 7. Bootstrap results

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 Figure 7).

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 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 rules. The results of the neural network were then staged in a three-step model that worked by correctly defining the sentence boundary by linking it to a pattern that took into account word and sentence structure. Although the model we have developed has been successful in practical experiments, improvements need to be made to bring it into line with the standard system of neural machine translation. In addition, a neural machine translator can generate direct output without waiting for a complete input sentence, allowing the user to translate directly or in real time.

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*, pp. 54-65. https://doi.org/10.18653/v1/W16-2206

Brown, P. F., Cocke, J., della Pietra, S. A., della Pietra, V. J., Jelinek, F., Lafferty, J. D., … 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*, pp. 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*, pp. 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*, pp. 127-133. https://aclanthology.org/N03-1017

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)*. Retrieved from 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*, pp. 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*, pp. 440-447. https://doi.org/10.3115/1075218.1075274

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*. Retrieved from 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*, pp. 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)*, pp. 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*, pp. 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)*, pp. 3530-3534. Retrieved from https://aclanthology.org/L16-1561