The Machine Translation Model

Ziyuan Zhao*

Basis International School Shenzhen, Shenzhen, 518060, China
Ziyuan.Zhao11406-bisz@basischina.com

Abstract. Machine Translation was invented in the late-20th century, when the first IBM model could automatically translate Russian sentences into English. Proceeding to the 21st century, linguists have invented new types of Machine Translation models and improved on them by altering and adding parts. Of all the various types of Machine Translation models, this paper will mainly focus on Statistical Machine Translation models, which put into play the statistical data with their calculation of the probabilities of the words and phrases, and Neural Machine Translation models, which are built by the stacking and connecting of neural layers through encoders and decoders, by comparing the benefits and flaws of the types within the specific Machine Translation models. A specific translation will be mentioned using the Neural Machine Translation models, revealing the flaws within its translation. From the flaws of Neural Machine Translation models, this paper will also examine attempts to improve them by solving existing problems. This paper will build a basic understanding of machine translation models and possibly inspire future experiments and extensions to improve the translation models both mentioned and not mentioned in this paper.

Keywords: Translation model, Source language, Target language, Monolingual, Multilingual

1 Introduction

Though the first Machine Translation was built two decades earlier, there are still advances in the 21st century, namely Statistical Machine Translation and Neural Machine Translation. The former was based on statistical models and probabilities calculated from large amounts of parallel data on language pairs. Some sub-types of this model include the Chunk-based Statistical Translation model, Phrase-based Statistical Translation model, and Syntax-based Statistical Translation model, all of which have non-negligible flaws. The latter model was built off of the stacks of connections between neural layers encoded into word embeddings and decoded into the target language translations. Some sub-types of this model include the Multilingual Neural Translation model and RNN-based Neural Translation model. Considering the flaws in Neural Machine Translation models, this paper mentions some extensions attempting to improve translation quality by addressing specific problems such as connection issues when neural layers become more complex. This paper will discuss Statistical Machine Translation in Section 2, with its subtypes in Sections 2.1, 2.2, and 2.3. Then, it will move

A. Holl et al. (Eds.): ICHESS 2022, ASSEHR 720, pp. 2153–2160, 2022.
https://doi.org/10.2991/978-2-494069-89-3_247

on to Neural Machine Translation in Section 3, with its subtypes in Sections 3.1 and 3.2. The whole of section 4 will be dedicated to presenting extensions on NMT, building on top of the NMT basis from Section 3.

2 Statistical Machine Translation (SMT)

Statistical models laid a foundation for Statistical Machine Translation (SMT) with large amounts of parallel texts to translate one language to another. The quality of translation depends on the availability of parallel data between the source-target language pair [1]. Some common types of SMT include Chunk-based, Phrase-based, and Syntax-based.

2.1 The Chunk-based Statistical Translation (CSTM)

The paper by Watanabe, Sumita and Okuno proposed the Word Alignment Based Translation model (WABTML). In WABTML, a sentence is divided into words that are either individually translated according to the Lexicon Model or deleted and assigned a zero. This blind assignment weakens this model. These translated words are then selected and inserted into their corresponding positions according to the grammatical structure in the target language. The local reordering of words (or local alignment) cannot consider long-distance phrasal constraints, which also weakens this model. Apart from these two weaknesses, WABTML cannot perfectly ideal situations where the exact words of the source language do not match the exact words of the target language, bringing ambiguity in translations that might even make those translations lose meaning. The Chunk-based Statistical Translation model (CSTM) emerged to overcome the two weaknesses mentioned in WABTML [2].

The CSTM divides a sentence into chunks, where each chunk conveys meaning without needing to add or subtract words. After each chunk is translated, the words inside the chunk are reordered. Then the chunk will be assigned according to the grammatical structure, similar to the process in the WABTML [2].

This model, however, also has its flaws. The chunks might be hard to divide, given the potentially vast difference between the source language and targeted language. The grammatical structure of the two languages might differ to the extent that it will make more sense to translate individual words than to translate divided chunks. Additionally, if the associated word is in another chunk, the gender of the word might not be correctly translated into the targeted language. For example, if we are to translate this sentence from English to Spanish: "the girl, to which he blew a kiss, was beautiful," "the girl," "to which he blew," "a kiss," and "was beautiful" could be divided into chunks. But when translating, the adjective "beautiful" needs to be feminine, as it modifies the girl. That is, "beautiful" should be translated to "Preciosa" instead of “precioso.”

2.2 Phrase-based Statistical Translation (PSTM)

Another model is the Phrase-based Statistical Translation Model (PSTM), where a sentence is first divided into phrases of any subsequence of words, translated and reordered according to the probabilities in the targeted language [3].

The first subdivision of PSTM, the Standard PSTM, uses the Phrase Translation Table [4], built off of the Word Alignment Table, as shown in Table 1. A sentence is presented in the first column in English, and its Spanish translation is presented in the first row. The boxes are shaded to match the specific English word(s) with its (or their) Spanish translation. For instance, "the" matches with "la" and "girl" matches with "chica," so the box corresponding to "the" and "la" and the one corresponding to "girl" and "chica" is shaded. The Standard PSTM then produces the table in both directions of the translation of the two languages by switching the two axes (that is, putting the Spanish sentence in the first column and its English translation in the first row) [4]. The two tables will then be merged by keeping the boxes shaded on both tables. The resulting table will then be examined to ensure that the alignment points, or the individual translations of words, for all the words are included [4].

Another subdivision is the Hierarchical Phrase-based Statistical Translation, where lexical features are assigned according to a tree-based preordering system. Some propose to reorder the words according to the grammatical structure of the targeted language before translation. However, it may have some drawbacks when some information cannot be included in the translation process [5]. A complete Hierarchical PSTM utilizes a translation model, target language model, and reordering model. This model utilizes the word-based reordering model to guarantee accuracy when reordering the source language. The hierarchical tree sorts the source-side aligned words in the same order as the targeted language. The source-side unaligned words would be moved to their corresponding position after the translation. This model would better utilize the reordering of the source language sentence to bring accuracy to the translation [5].

Unlike CSTM, PSTM does not encounter difficulty when the source and target languages differ vastly in their grammatical structure. Preordering words into phrases and hierarchical structures would make the translation process more apparent and accurate. PSTM also categorizes and reorders the words into phrases, which solves the gender problem mentioned above. That is to say, the words that determine the gender of another word would more likely be organized into the exact phrase, which would assist the model in accurately determining the gender of the word. This model also considers the local context and improves as more data (with longer phrases) decreases the ambiguity in the final translation.

Table 1. Example of a Phrase Translation Table (own-drawn)

	La	chica	ambiciosa	no	tuvo	valor	para	gritar.
The								
ambitious								
girl								
did								

not								
have								
the								
courage								
to								
shout.								

2.3 Syntax-based Statistical Translation (SSTM)

This model applies operations that capture linguistic differences between the source and targeted languages to translate the source language into the target string. Like PSTM, the SSTM requires the preprocessing of nodes by reordering child nodes and inserting and deleting extra words from the nodes. Reordering is to consider language pairs with different grammatical structures (e.g. Subject-Verb-Object and Subject-Object-Verb structures) while inserting and deleting to ensure that the linguistic differences are captured in specific cases of the source and target languages. Unlike PSTM, the output of SSTM is a string rather than a tree [6].

When the nodes' child nodes are reordered, N number of child nodes would have N factorial number of possible orderings, and the probability of each order would be calculated. A word might be inserted either to the left or right of a node, with each insertion's probability calculated. Using statistics and probability, the final child node reordering and insertion accuracy would be better guaranteed. The individual words are then translated without consulting the context. Still, the nodes' probabilities are added for each translation, and the most considerable probability would indicate the best translation [6].

This model enhances the accuracy of the translation by using the concept of probability in each node. However, the problem of the gender of words is still not fully considered, and some correlating words might exist as inaccurately translated exceptions in this model.

3 Neural Machine Translation (NMT)

The Neural Machine Translation (NMT) utilizes a neural network system, or model, to form learning that produces the best translation based on calculations and predictions of occurrence likelihood [7].

NMT first divides the sentence into parts and, using neural layers, converts those parts into word embeddings, which assign words with similar meanings and contexts similar representations, or the encoder representations. The decoder then uses these encoder representations to generate the decoder representations. The details of this process and the specific encoders and decoders will be mentioned later in section 3.2. The next layer would be built based on the biases of the preceding layer, which means that the stacking of multiple layers would result in higher-quality translations [8].

Some common types of NMT include Multilingual Neural Machine Translation (MNMT) and Recurrent Neural Network (RNN)-based model (RNMT). These NMT sub-types use translation processes similar to those of NMT, only modifying parts of the NMT translation process or adding parts to it.

3.1 Multilingual Neural Machine Translation (MNMT)

MNMT uses diverse languages in its translation process, which brings two main benefits: 1) it decreases the ambiguity and therefore increases the translation quality for low-resource language pairs. For instance, a language pair like Spanish and English that has more resources and data can be referred to or used to develop the translation model of the language pair Catalan and English; 2) it is more efficient and compact, as a single model could complete translations of multiple language pair. This efficiency is significant when a large number of languages are required in a translation. Other translation models designed only for one or a few language pairs would need a more extensive deployment of translation models, whereas MNMT would suffice [8].

When improving the translation quality for low-resource language pairs, the model loses some extent of translation quality for high-resource language pairs. This loss of translation quality is due to the equal sampling of all data sets with the over-sampling of low-resource language pairs. So, this model aims to maximize the transfer of resources to low-resource language pairs. At the same time, it minimizes the interference in the translation quality of high-resource language pairs [9].

Before the translation process, MNMT divides the sentence from the source language into basic units. These basic units must, to the greatest extent, represent the full meaning of the sentence (later referred to as “coverage”) and lessen their number to minimize computational and spatial costs [9]. Coverage presents a challenge when there is a limited vocabulary, so sub-words and constructing vocabulary are commonly used in preprocessing the sentence [9].

3.2 Recurrent Neural Network-based Neural Machine Translation (RNMT)

In MMT, both the encoder and decoder are Recurrent Neural Networks (RNN). The encoder RNN, which has the word-embedding layer, first converts the source sentence parts into vectors, which are then converted to the target language with attention mechanisms. Before moving on to the next layer, the forward layer and backward layer from the bidirectional RNN encoder are connected to ensure the stability of the translation model [10]. The positional encoding layers add positional encodings or positional vectors to the word-embedding layer to provide information about the word's position. In this case, the relative position of the word is more valuable than the absolute position of the word [11]. For instance, in the sentence “She, when sensing danger, would always call for her mom,” the phrase “when sensing danger” has a relative position behind “she,” so it should be interpreted to be modifying “she” and not “mom.” To incorporate the relative position to the model, the feed-forward layer is replaced by the Simple Recurrent Unites (SRU) [11]. A higher-quality RNN model is built with stacked bidirectional encoders and unidirectional decoders [11].

Self-attention mechanisms work by computing weights of the components with query and keys, each weight representing the amount of attention. An input sentence would be processed into a set in self-attention [10].

Looking at NMT as a whole, the paper by Evgeny Matusov described an experiment where neural machine translations were applied to fictional stories. The experimental results show that there are still existing problems in NMT: 1) ambiguous words/phrases or words with multiple meanings translations were not readily understandable without referring to the source sentences; 2) Out-of-vocabulary (OOV) words, or words not built into the model will be translated separately which might not convert the actual meaning of the word in the source language; 3) translation inconsistencies and mistranslation of pronouns appeared throughout the translation as NMT cannot take into consideration the context information in its translation process; 4) idioms are not correctly translated if translated word-by-word or distorted the meaning of the idiom; 5) repetition of words that have the same meaning (e.g. grassy grass) or insertion of words/phrases into wrong places happen as a result of error in decomposing the source sentence; 6) tone and formality of the source sentences could not be captured through the translation process or presented in the translated sentences. Despite these errors in the translation made by NMT, most translated sentences were still usable. In contrast, others, too, were usable with a bilingual proofreader [12].

4 Extensions to NMT

A zero-shot translation, or the translation between arbitrary translation pairs with no parallel data, is challenging to MNMT. This translation process requires a pivot language, to which both the source and target languages can be translated [9]. This additional translation process brings more significant errors due to the accumulation of errors. A way to reduce error is to synthesize parallel data. Still, the size of the parallel data might grow too large to be easily tractable [9]. By increasing the languages used in the model, regularization could encourage shared representations between languages and therefore increase the quality of zero-shot translations [13].

As the number of neural layers increases and becomes more complex, the connections between them become increasingly fragile. To address this critical problem, Dowmunt (et al.) introduce a method to extract translations from both the source and target languages to achieve a higher word alignment quality [14]. This process is similar to the word-alignment process explained in 2.1 and shown in Table 1. This is also similar to the back-translation process to train monolingual data and to better incorporate the data in the improvement and training of the model [15].

5 Conclusion

This paper examined SMT and NMT according to their different types and weighed their benefits and flaws. Specific types of SMT include Chunk-based SMT, Phrase-based SMT, and Syntax-based SMT. Although some SMTs, such as SSTM, use statistics and calculates probabilities of words and phrases to improve overall translation

quality, all SMTs omit specific characteristics of words, such as gender. This might be due to the decomposition of sentences beforehand, or any preprocesses that do not consider the overall structure or the correlation between words. This is a problematic flaw in SMT. Both the Multilingual NMT and RRN-based NMT have their flaws. Similar to SMT, some NMT processes stack multiple neural layers to ensure translation quality. NMT still fails to consider crucial portions of the source sentences, such as ambiguous words, words outside the model's learned vocabulary set, or idioms. Extension to NMT is also considered to improve the existing model and boost the translation quality.

An extension regarding the low translation qualities of zero-shot translations encourages the use of similar representations across languages. Another extension to ensure the connection between the neural layers is to translate the words and phrases from the source language to the source language and compare them. Similar to training monolingual data, this process improves the word alignment quality.

NMT and machine translation models fail to consider context information, which might result in awkward translations, various translations of the same reference, or even repetition/neglection of unnecessary phrases. Only under proofreaders' supervision can the translation sentences from the models accurately present the meaning of their sources.

6 Reference

1. G. Haffari, "Active Learning for Statistical Phrase-based Machine Translation", Aclanthology.org. [Online]. Available: https://aclanthology.org/N09-1047.pdf. [Accessed: 27- Aug- 2022].
2. L. Bentivogli, A. Bisazza, M. Cettolo and M. Federico, "Neural versus Phrase-Based Machine Translation Quality: a Case Study", Arxiv.org. [Online]. Available: https://arxiv.org/pdf/1608.04631.pdf. [Accessed: 27- Aug- 2022].
3. T. Watanabe, E. Sumita and H. Okuno, "Chunk-based Statistical Translation", Aclanthology.org. [Online]. Available: https://aclanthology.org/P03-1039.pdf. [Accessed: 27- Aug- 2022].
4. C. Manning, "Natural Language Processing: Phrase-Based Machine Translation", Web.stanford.edu. [Online]. Available: https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1162/handouts/cs224n-lecture4-PhraseBasedMT.pdf. [Accessed: 27- Aug- 2022].
5. K. Hayashi, H. Tsukada, K. Sudoh, K. Duh and S. Yamamoto, "Hierarchical Phrase-based Machine Translation with Word-based Reordering Model", Aclanthology.org. [Online]. Available: https://aclanthology.org/C10-1050.pdf. [Accessed: 27- Aug- 2022].
6. K. Yamada and K. Knight, "A Syntax-based Statistical Translation Model", Aclanthology.org. [Online]. Available: https://aclanthology.org/P01-1067.pdf. [Accessed: 27- Aug- 2022].
7. D. Bahdanau, K. Cho and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate", Arxiv.org, 2016. [Online]. Available: https://arxiv.org/pdf/1409.0473.pdf. [Accessed: 27- Aug- 2022].
8. R. Dabre, C. Chu and A. Kunchukuttan, "A Survey of Multilingual Neural Machine Translation", Dl.acm.org. [Online]. Available: https://dl.acm.org/doi/pdf/10.1145/3406095. [Accessed: 27- Aug- 2022].

9. N. Arivazhagan, A. Bapna and O. Firat, "Massively Multilingual Neural Machine Translation in the Wild: Findings and Challenges", Arxiv.org, 2019. [Online]. Available: https://arxiv.org/pdf/1907.05019.pdf. [Accessed: 27- Aug- 2022].
10. M. Chen, O. Firat and A. Bapna, "The Best of Both Worlds: Combining Recent Advances in Neural Machine Translation", Arxiv.org, 2018. [Online]. Available: https://arxiv.org/pdf/1804.09849.pdf. [Accessed: 27- Aug- 2022].
11. M. Negishi and N. Yoshinaga, "On the Relation between Position Information and Sentence Length in Neural Machine Translation", Aclanthology.org. [Online]. Available: https://aclanthology.org/K19-1031.pdf. [Accessed: 27- Aug- 2022].
12. E. Matusov, "The Challenges of Using Neural Machine Translation for Literature", Aclanthology.org. [Online]. Available: https://aclanthology.org/W19-7302.pdf. [Accessed: 30- Aug- 2022].
13. T. Kudo, "Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates", Arxiv.org, 2018. [Online]. Available: https://arxiv.org/pdf/1804.10959.pdf. [Accessed: 30- Aug- 2022].
14. T. Zenkel, J. Wuebker and J. DeNero, "Adding Interpretable Attention to Neural Translation Models Improves Word Alignment", Arxiv.org, 2019. [Online]. Available: https://arxiv.org/pdf/1901.11359.pdf. [Accessed: 30- Aug- 2022].
15. C. Hoang, G. Haffari, P. Koehn and T. Cohn, "Iterative Back-Translation for Neural Machine Translation", Aclanthology.org. [Online]. Available: https://aclanthology.org/W18-2703.pdf. [Accessed: 30- Aug- 2022].