Proceedings of the 5th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/13/20230734

The development and advance of machine translation

Yilin Liu

University of Toronto, Toronto, ON M5S, Canada

riddle.liu@mail.utoronto.ca

Abstract. Since the advent of global online information exchanging and communicating, there are growing number of cases showing how effective and useful machine translation can be. This paper offers a brief overview of Machine Translation, including Statistical Machine Translation and Neural Machine Translation. In the Statistical Machine Translation part, different translations using corresponding different bases are introduced. Then some key parts in Neural Machine Translation are discussed: the seq2seq model, LSTM (Long Short-Term Memory) and the encoder and decoder. And finally, the evaluation metrics of machine translation are concluded which contains not only human but automatic evaluation metrics as well.

Keywords: machine translation, Seq2seq model, evaluation metrics, LSTM.

1. Introduction

Machine translation, which can be used on its own or in connect with human proofreaders and post-editors, is useful in many areas. There can be repetitive content which are too costly using manual translation, and high-value dynamic content such as stock market news or real-time communications, where it would not be practical for a human to translate. Therefore, machine translation is needed where there is information exchanging.

A machine translation (MT) system first analyzes the source language input, then an internal representation is built, processed, and translated to a form appropriate for the destination language. Finally, output is generated and printed. On a fundamental level, MT simply replaces words from one natural language with words from another, however, this alone is rarely sufficient to generate a reliable translation of a document because entire phrases and their closest equivalents in the target language often need to be recognized [1].

When given enough information, machine translation programs may often do an adequate job of conveying one language to the people who speak it natively. However, getting enough data to support the method is the always corresponding problem. Grammar-based methods, on the other hand, typically do not require the enormous multilingual corpus of data that is necessary for statistical methods to function.

This article will first introduce the two main stages of machine translation: Statistical Machine Translation (SMT), including the different four types of methods concerning different bases, and Neural Machine Translation (NMT), in which the typical functions like seq2seq model, and the decoder and encoder will be listed. As the process and methods of Machine Translation developed rapidly, evaluation metrics are also needed to be implemented to accord with the new translation form. Therefore, evaluation metrics such as BLEU and WER will be discussed in the last part.

DOI: 10.54254/2755-2721/13/20230734

2. Statistical Machine Translation (SMT)

SMT is a kind of machine translation with better performance in and only in non-limited fields. Prior to the development of neural machine translation, it remained by a significant margin the approach of machine translation that was the most generally known and understood.

The initial concept of statistical machine translation was presented by Warren Weaver in 1949 which applied Claude Shannon's information theory [2]. It wasn't until the latter half of the 1980s and the beginning of the 1990s that IBM scientists at the Thomas J. Watson Research Center brought it back into circulation [3][4][5].

The primary task of statistical machine translation is to construct a reasonable statistical model for language production, and based on this statistical model, define the model parameters to be estimated, and design a parameter estimation algorithm. Therefore, depending on the construction of sentences in languages, using words, phrases, syntax, and hierarchical phrases are the four subcategories that fall under the category of statistical machine translation.

2.1. Word-based Translation

In word-by-word translation, a natural-language word acts as the essential building block of a translation and, in this capacity, serves as the basic unit. This type of translation is referred to as "word-based translation." Due to idioms, morphology, and compound words, translated phrases frequently have a variable word count. The term "fertility" refers to the ratio of the lengths of word sequences in a language that has been translated, and it is used to show the number of foreign words that are produced by each native word. According to the information theory, there is an unavoidable presumption that both theories cover the same general idea.

2.2. Phrase-based Translation

The statistical translation system that uses phrases as input, which has gained prominence recently, applies a discriminative training approach that often calls for reference corpus supervised training. Basically, the goal is to translate entire word sequences, whose durations can vary depending on circumstances to overcome the limitations of word-based translation. Blocks or phrases, the names for word groups, are typically phrasemes discovered using statistical techniques from corpora rather than actual verbal phrases. The selected phrases can be reordered after being further one-to-one mapped using an index of phrase translations. Either by directly learning from a parallel corpus or by basing it on word alignment, one can directly learn this index. Similar to the IBM word-based model, the second model is educated by a process known as expectation maximization [6].

2.3. Syntax-based Translation

In contrast to phrase-based machine translation, which translates individual words or strings of words, a translation approach known as syntax-based translation prioritizes the translation of syntactic units over the translation of individual words or word combinations [7]. It has been there for a very long time in MT, but its statistical counterpart did not start gaining traction in the industry until the 1990s, when strong stochastic parsers emerged on the scene. However, one of the difficulties that comes with using a syntax-based approach is how quickly translations can be produced.

2.4. Hierarchical Phrase-based Translation

The first example of hierarchical phrase-based translation can be found in Chiang's Hiero system [8]. In this system, there is a statistical model for machine translation that includes Hierarchical phrases that are composed of subexpressions. The model is technically a grammar that is context-free, synchronous, and is picked up from a parallel text, despite the fact that it does not contain any syntactic information annotations. As a result, it can be thought of as fusing the core concepts of a translation that incorporates both phrase- and syntax-based translation. However, due to the large amount of data needed to be recorded, corpus creation can be costly. And because this way of translation usually does not find specific rules or formulae on translation, certain errors are difficult to predict and fix.

DOI: 10.54254/2755-2721/13/20230734

3. Neural Machine Translation and its Functions

There are mainly two reasons to migrate from SMT to NMT: higher quality translation output and translation speed, which can sometimes improve several ten times of efficiency.

A new encoder-decoder structure for machine translation was proposed by Nal Kalchbrenner and Phil Blunsom in 2013 [9]. In order to encode a piece of source text into a continuous vector, the model can make use of a convolutional neural network (CNN). With a view to translating the state vector into the desired language, a recurrent neural network, also known as an RNN, is utilized as a decoder. Their research results can be said to be the birth of Neural Machine Translation (NMT); a method of using deep learning neural networks to obtain mapping relationships between natural languages.

3.1. Encoder and Decoder

From a probabilistic point of view, finding a target sequence y that maximizes the likelihood of obtaining y given a source sentence x is essentially what the challenge of machine translation is similar to. In the NMT task, parallel training corpora (i.e.: x, y are two languages with the same content) to fit the parameterized model is used so that the model can learn to be able to optimize the probability distribution under the various conditions of the target sentence. Once NMT learns this conditional probability distribution, given a source sentence, it can find the sentence with the highest conditional probability as the corresponding translation by searching. The encoder goes through the steps of processing each individual item in the input sequence in order to extract the contextual information included in the input sequence. Following the completion of the processing of the complete input sequence, the contextual information is transferred from the encoder to the decoder, which then begins the process of producing the output sequence item by item.

The source embeddings are translated into concealed continuous representations by the encoder network. The encoder must be able to simulate the complex dependencies and ordering that existed in the source language in order to learn expressive representations. Variable-length sequences can be modelled using recurrent neural networks (RNNs).

In NMT architectures, the encoder and decoder are essential elements. There are numerous ways to create effective encoders and decoders, which can be broadly categorized into three groups: methods that are based on recurrent neural networks (RNN), methods that are based on convolution neural networks (CNN), as well as methods that are based on self-attention networks (SAN). When creating an encoder and decoder, various factors such as receptive field, computational complexity, sequential operations, and position awareness must be considered.

3.2. Sequence to Sequence Model

In 2014, Sutskever et al. and Cho et al. devised a method called sequence-to-sequence (seq2seq) learning that makes use of RNNs for both the encoder and the decoder [10][11]. They also presented Long Short Term Memory (LSTM, a form of RNN) for NMT. With the help of a gate mechanism (allowing to delete and update explicit memories in LSTMs), the "explosion/vanishing gradient" problem is controlled, allowing the model to capture the "long range" in sentences far better dependency".

One sequence is converted into another sequence via Seq2seq (sequence transformation). To get over the issue of vanishing gradient, it uses a recurrent neural network (RNN). The output from the preceding stage serves as the context for each item. The primary components consist of a single encoding network and a single decoding network each. The encoder is responsible for creating a concealed vector out of each item that corresponds to it and contains both the object and its context. Using the input context from the prior output, the procedure is performed in reverse by the decoder, which results in the vector being produced as an output item. Optimizations include Attention, which enables the decoder to examine the input sequence in a selective manner, and Beam Search, which stores multiple highly probable choices rather than selecting a single output (word) as the output. Both of these optimizations are referred to as attention and beam search, respectively [12].

DOI: 10.54254/2755-2721/13/20230734

3.3. LSTM (Long Short-Term Memory)

The phrase "long short term memory" (LSTM) refers to an artificial recurrent neural network (RNN) architecture that is utilized in the discipline of deep learning. In contrast to the more common feedforward neural networks, LSTM uses connections that reflect previous activity. In addition to analyzing individual data points (such as photographs), it is able to evaluate entire data sequences (such as speech or video). For illustration, LSTM is useful for a variety of applications, including speech recognition, uncut and linked handwriting identification, and the detection of anomalies in network traffic or in IDSs (intrusion detection systems).

Cells, input gates, output gates, and forget gates are the component parts that make up a typical LSTM unit. The cell is able to remember values for arbitrary periods of time, and the information that enters and leaves the cell is regulated by its three gates, which are responsible for the cell's overall access control [13].

LSTM networks function admirably when it comes to classifying, processing, and making predictions based on time series data. This is because there is a possibility that there will be gaps in time between significant events that occur in a time series that have a duration that cannot be precisely determined. In order to circumvent the issue of disappearing gradients, which can arise during the training of conventional RNNs, LSTMs were developed.

An LSTM-based RNN can be trained under supervision on a series of training drills by computing these gradients required for the optimization procedure and varying every weight of the LSTM network in proportion to the consequence of the inaccuracy (at the output layer of the LSTM network) with respect to proportional weight. This optimization algorithm is called gradient descent combined with backpropagation through time. Even if it is backpropagated from the output layer, the error value is still there in the LSTM unit's cells. This is because the error value was first input into the cells. Until the gates of the LSTM unit learn to shut off the value, this so-called "error carousel" feeds error back to each one of the gates continuously.

4. Evaluation of Machine Translation

4.1. Human Evaluation

A correlation with human judgment ought to be established as the standard of evaluation for metrics. This is typically done on two different levels: at the sentence level, where scores from the metric are calculated for a collection of translated sentences, and then these aggregate scores are associated with human decision-making for the same sentences; and at the corpus level, where scores from the sentences are collected for both human and metric evaluations, and then these cumulative scores are connected.

4.1.1. Automatic Language Processing Advisory Committee (ALPAC). The findings of a study conducted in 1966 and contained in the ALPAC report compared the output of various levels of human translation with that of machine translation. The study recruited human volunteers to serve as judges. In preparation for this, the human judges went through extensive specialized training. In the assessment research, a machine translation (MT) system that can translate from Russian into English was compared to human translators on two different factors.

"Intelligibility" and "fidelity" were the variables under investigation. Intelligibility was used to gauge how "understandable" the statement was, and fidelity was used to gauge how much of the original meaning was carried over into the translated version. A textual description was linked to each point on the scale. In contrast to integrity, intelligence was assessed without reference to the source material. After reading the translated sentence and understanding its meaning, the original sentence was next provided. The first sentence's informativeness was rated by the judges. Therefore, the translation's quality decreases the more informative the original text was.

4.1.2. Advanced Research Projects Agency (ARPA). The Advanced Research Projects Agency (ARPA) developed a technique to evaluate machine translation systems as a part of the Human Language

DOI: 10.54254/2755-2721/13/20230734

Technologies Program, and it continues to carry out evaluations based on this methodology. These evaluations are being carried out at this time.

According to Church et al., the goal of the understanding evaluation was to quantitatively compare various systems using the outcomes of multiple-choice comprehension examinations (1993). The resources that were selected consisted of a compilation of articles written in English on recent developments in the financial sector. These articles were first translated by professionals into a variety of language pairs, and then those language pairs were translated back into English using software designed specifically for that purpose. Due to problems with the translation of meaning from English, it was deemed that this was insufficient as a stand-alone technique of comparing systems and was abandoned as a result.

4.2. Automatic Evaluation

It goes without saying that there is no objective or quantifiable "excellent" when it comes to translation quality. Any measure must therefore offer quality scores for them to correspond with the human evaluation of quality. In other words, a statistic must match the appropriate human rates. Considering that human beings are the ultimate consumers of any translation result, human judgement should be used as the standard for evaluating automatic metrics. Even if a correlation between a measure and human judgment is demonstrated in a study conducted on one corpus, it is still possible that this correlation will not transfer to another corpus. Because it is undesirable to develop a new measure for each assessment or domain, it is valuable to have a metric that is only applicable to text in a specific domain; however, such a metric is less useful than one that is applicable to text in a variety of domains. This is because a metric that is only applicable to text in a particular domain is valuable.

4.2.1. BLEU. BLEU, which stands for "bilingual evaluation understudy," is a piece of software that examines the content's quality after automatic translation between two natural languages. BLEU was one of the first measures to claim a strong connection with human evaluations of quality. Today, it is still one of the most common automated metrics and it is also one of the least expensive metrics [14].

Programmatically, the main task of a BLEU implementor is to examine the n-grams of the candidate translation to those of the reference translation and tally the instances when the two sets of n-grams are identical. These competitions don't consider position. The quality of the candidate translation improves with the number of matches. However, they are unable to quantify the effectiveness of NMT systems in addressing the problems [8].

The first step in computing the improved n-gram precisions' geometric average, p_n, is to use n-grams with lengths up to N and positive weights w_n that add up to one. Next, we'll take the candidate translation's length, denoted by c, and compare it to the length of the reliable reference corpus, denoted by r.

The brevity penalty BP,

$$BP = \begin{cases} 1 & if\ c > r \\ e^{(1-r/c)} & if\ c \leq r \end{cases}$$

Then,

$$\text{BLEU}=\text{BP}\cdot\exp\left(\sum_{n=1}^{N} w_n log p_n\right)$$

The output of BLEU is consistently an integer between 0 and 1. With values closer to 1, texts that are more comparable to the reference texts are considered to be more similar to the candidate text. Rarely would a human translation obtain a score of 1, as this would indicate that the potential translations are identical. Therefore, obtaining a score of 1 is not necessary. The BLEU score will rise when more reference translations are added because there are more chances for matches.

BLEU continues to serve as a baseline for the evaluation of any new evaluation metric because it has consistently been reported to correlate favorably with human judgement. However, there have been several criticisms. It has been highlighted that while BLEU is theoretically capable of assessing

DOI: 10.54254/2755-2721/13/20230734

translations from any language, it is currently unable to handle languages that lack word boundaries. It has been stated that although though BLEU offers a number of benefits, there is no assurance that rising BLEU scores are a sign of higher translation quality.

In addition to simple inertia, one of the reasons BLEU is still in use may be due to its applicability and simplicity in a variety of real-world situations. NIST, which stands for the National Institute of Standards and Technology, is a procedure for measuring the quality of text that has been translated using machine translation. This method is based on the BLEU metric, but it has been modified in a few key ways. NIST additionally determines how informative a certain n-gram is, in contrast to BLEU, which merely calculates n-gram precision by giving each n-gram an equal weight. When a proper n-gram is discovered, an n-gram will be assigned more significance if it is less prevalent. For instance, as it is less likely to happen, the bigram "on the" will be given less weight if it is correctly matched than the bigram "interesting calculations". Small differences in translation length have less of an impact on the final score in NIST's calculation of the brevity penalty than they do in BLEU.s

4.2.2. Word Error Rate. Word error rate, often known as WER, is a statistic that is commonly used to evaluate the performance of speech recognition or machine translation systems. The fact that the length of the recognized word sequence can vary in comparison to the length of the reference word sequence makes it more difficult to gauge performance in general (supposedly the correct one). The WER is a helpful tool that may be used to compare and contrast different systems as well as evaluate the effects of changes made to a single system [15].

However, this form of evaluation offers little insight into the specifics of the faults that occur in translation. Therefore, additional study is required in order to identify the primary source (or sources) of mistake and to target any research efforts that may be undertaken. This issue can be remedied by first employing dynamic string alignment in order to line up the word sequence that was heard with the word sequence that was referenced (spoken). The power law hypothesis proposes that there is a correlation between the degree of bewilderment and the amount of words that are misspelled. This hypothesis is used to investigate the topic at hand.

Word error rate can then be computed as:

$$\mathrm{WER} = \frac{S+D+I}{N} = \frac{S+D+I}{S+D+C}$$

Where S is the number of word changes, D signifies the number of word deletions, I signifies the number of word additions, C signifies the number of word corrections, and N signifies the total amount of words in the reference (N = S + D + C).

The thought process that lies behind the terms "deletion" and "insertion" focuses on how to proceed from the reference to the hypothesis. Word accuracy, also known as WAcc, is occasionally used instead of reporting a voice recognition system’s performance.

$$\mathrm{WAcc} = 1 - \mathrm{WER} = \frac{N-S-D-I}{N} = \frac{C-I}{N}$$

Due to the fact that N signifies the total amount of words contained in the source, the word mistake rate has the potential to be higher than 1.0, and as a consequence, the word accuracy has the potential to be lower than 0.0.

4.2.3. METEOR. METEOR is an acronym that stands for "Metric for Evaluation of Translation with Explicit Ordering," and it is a statistic used to assess the effectiveness of machine translation. In this metric, which is derived from the harmonic average of the of the unigram precision and recall, recall is given a greater degree of importance than accuracy. Along with the typical exact word matching, it also offers several additional elements that are not present in other measures, like stemming and synonymy matching. The metric was created to address some of the issues that were present in the more widely used BLEU metric and to yield results that were well correlated utilizing human judgment at the segment or phrase level. The BLEU metric is different from this in that it looks for correlation at the corpus level.

DOI: 10.54254/2755-2721/13/20230734

There have been reports of outcomes at the corpus level that yield a correlation of up to 0.964 with human assessment. This is in comparison to the achievement of BLEU, which was only 0.817 on the same data set. At the level of the phrase, the highest correlation with human assessment that could be attained was 0.403 [16].

Since the sentence is the primary evaluative unit in BLEU, the procedure begins by establishing an alignment between two phrases: the reference translation string and the candidate translation string. This alignment is done because the reference translation string and the candidate translation string are both strings that represent translations. A collection of mappings between unigrams makes up the alignment. One way to conceptualize a mapping is as a line connecting a unigram in one string to a unigram in another. Each unigram in the candidate translation must correspond to either 0 or 1 in the source text. To create the above-described alignment, mappings are chosen. The alignment with the fewest crosses, or the one where there are fewer intersections of two mappings, is chosen when there are two alignments that have the same amount of mappings in common. The score is calculated after the final alignment has been determined as follows:

Unigram precision P is calculated as $\frac{\mathrm{m}}{w_t}$, where m represents the amount of unigrams contained inside the candidate translation that are also present in the reference translation, and w_t represents the total amount of unigrams contained within the candidate translation.

Unigram recall R is calculated as $\frac{\mathrm{m}}{w_r}$, where m is determined in the same manner as described, and w_r is the total amount of unigrams contained inside the reference translation. Following is an example of how the harmonic mean is used to combine precision with memory, with recall being weighted nine times more than precision:

$$F_{mean} = \frac{10PR}{R+9P}$$

The current methods only consider congruity between single words, not between longer segments that are present in both the reference and candidate sentences. Longer n-gram matches are used in the computation of a penalty p for the alignment. This is done so that these factors can be taken into account. The penalty will be increased by the number of mappings that are included in both reference sentences and candidate sentences that are not contiguous to one another.

A collection of unigrams that are close to one another in both the reference and the hypothesis constitute what is known as a chunk, and unigrams are clustered into as few chunks as feasible to calculate this penalty. The longer the adjacent mappings are that are placed between the candidate and the reference, the fewer the chunks that are present. One chunk will be provided if the translation is exact to the reference. The penalty p is computed by the following formula:

$$p = \frac{1}{2}\left(\frac{c}{u_{\mathrm{m}}}\right)^3$$

Where c represents total amount of chunks and u_m refers to the amount of unigrams that have already been mapped. The complete score for a segment can be determined by using the formula below, denoted by M. In the event that there are no longer or bigram matches, the penalty can have the effect of lowering the F_{mean} by as much as fifty percent.

$$M = F_{mean}(1-p)$$

When calculating a score for an entire corpus, also known as a series of segments, the aggregate values for P, R, and p are obtained and merged employing the same formula. This allows the score to be calculated for the collection as a whole. Comparing a candidate translation against multiple reference translations can also be accomplished with the help of the algorithm. In this particular scenario, the algorithm evaluates the candidate in light of each of the references and picks the one with the best score.

DOI: 10.54254/2755-2721/13/20230734

5. Conclusion

This paper generally talks about the two main methods of machine translation: SMT (Statistical Machine Translation) and NMT (Neural Machine Translation). In the first section of SMT, there are four different based of statistical machine translation. In the second part of NMT, more contents are concerned, including the process and models of machine translation: decoder and encoder, seq2seq Model, and LSTM Model. Then in the last part, evaluation metrics are introduced, with brief comparison between Human and Automatic evaluation, the latter includes BLEU, World Error Rate, and METEOR.

References

[1] M. D. Okpor, (2014). Machine Translation Approaches: Issues and Challenges
[2] W. Weaver (1955). “Machine Translation of Languages” MIT Press, Cambridge, MA.
[3] P. Brown; John Cocke; S. Della Pietra; V. Della Pietra; Frederick Jelinek; Robert L. Mercer; P. Roossin (1988). "A statistical approach to language translation".
[4] P. Brown; John Cocke; S. Della Pietra; V. Della Pietra; Frederick Jelinek; John D. Lafferty; Robert L. Mercer; P. Roossin (1990). "A statistical approach to machine translation".
[5] P. Brown; S. Della Pietra; V. Della Pietra; R. Mercer (1993). "The mathematics of statistical machine translation: parameter estimation".
[6] Koehn, Philipp (2010). “Statistical Machine Translation” Cambridge University Press. ISBN 978-0-521-87415-1
[7] Steve DeNeefe, Kevin Knight, Wei Wang, Daniel Marcu (2007) “What Can Syntax-based MT Learn from Phrase-based MT?”
[8] David Chiang, Adam Lopez, Nitin Madnani, Christof Monz, PHillp Resnik, Michael Subotinf (2015). “The HIero Machine Translation System: Extensions, Evaluation, and Analysis”
[9] Kalchbrenner, N., & Blunsom, P. (2013). “Recurrent Continuous Translation Models. In EMNLP” (Vol. 3, No. 39, p. 413).
[10] Sutskever, I., Vinyals, O., & Le, Q. V. (2014). “Sequence to sequence learning with neural networks.”
[11] Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). “Learning phrase representations using RNN encoder-decoder for statistical machine translation.”
[12] Wadhwa, Mani (2018). "seq2seq model in Machine Learning"
[13] Sepp Hochreiter, Jurgen Schmidhuber (1996). “Long Short Term Memory”
[14] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu (2005). “ BLEU: a Method for Automatic Evaluation of Machine Translation”
[15] Klakow, Dietrich; Jochen Peters (2002) "Testing the correlation of word error rate and perplexity".
[16] Satanjeev Banerjee, Alon Lavie (2005) “METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments”