Machine Translation Subfield Survey

Shio Huang
xiaoxiaohuang@berkeley.edu

Abstract

[1] This paper is a subfield survey concentrating on the course of Machine Translation (MT) of Natural Language Process(NLP). It discusses the historical development, major schools, and important events and their impacts in this area, aiming to provide a thorough outline for newcomers who seek to investigate it in a full picture. To make this survey legitimate and inclusive, this paper draws references from 25 sources that are a combination of the most important papers published by ACL and other major conferences.

1 Introduction

Machine translation (MT) is a subfield of natural language processing that focuses on developing algorithms and models that enable automatic translation from one human language to another. The earliest effort of MT could be traced back to the 1940s when Rule-based Machine Translation (RBMT) was the mainstream. After that, huge developments were made over the years and resulted in multiple schools that focused on different approaches to address efficiency, low resources, and other problems in this field. In the 1960s, the idea of Statistical Machine Translation (SMT) started to become popular and largely influenced further research by introducing statistical methods into the algorithms. In the 1990s, Example-based Machine Translation (EMBT) showed up as a means to optimize the model with limited resources. When it comes to the 21st century, researchers of the new age seek to combine the findings of previous schools and have achieved extraordinary progress with a hybrid style. Nowadays, Neural Machine Translation (NMT) becomes another great breakthrough. Compared to the former strategies, its core is even more complex. This paper will introduce each of the strategies and their historical development in chronological order. The encoder-decoder and the later transformer-based models with their progress in recent decades, will be another topic to focus on in the later parts.

2 Rule-based Machine Translation

In the early stage of Machine Translation, researchers developed a rule-based system to explore this brand-new field. They created hand-written grammatical rules and dictionaries as tools for translation. According to such pre-defined rules, the system was able to generate translation in a restricted manner. One of the most famous attempts of this age is the Georgetown-IBM experiment of 1954, a collaboration of Georgetown University and the business company IBM, with the purpose to translate sixty Russian sentences into English(Hutchins, 2004). This translation was facilitated using a set of hand-written rules by linguists and programmers and breaking down the Russian sentences into component parts. The researchers then assigned a grammatical role to each part and used the rule to translate based on which. Historically, this experiment has demonstrated the potential of translating languages by computer and spurred strong motivation for further research. After that, efforts to improve the rules are done in many languages by countless researchers. In the Internation Conference on Computational Linguistics, for example, Hiroyuki Kaji proposed a new method based on two-level grammar and a table-driven parsing technique and demonstrated the effectiveness of such methods on English-Japanese translations(Kaji, 1988). However, despite the continuous contributions made in this strategy, more efficient means of MT quickly emerged.

3 Statistical Machine Translation

The experiments for Statistical Machine Translation could be dated back to the 1960s when researchers used statistical models to learn the patterns of language based on large amounts of bilin-

[1] Wordcount: 2119

gual text. Nevertheless, this approach didn't gain enough exposure until the 1990s. In 1966, an infamous report by Automatic Language Processing Advisory Committee (ALPAC), announced the limitations of RMBT's incapability of producing high-quality translations. Additionally, it also acknowledged the dilemma of SMT due to the unsatisfying performance of contemporary computers and the lack of bilingual data sources(ALPAC, 1966). This report concluded that machine translation was still in its early stages and much more progress in related fields needs to be made before achieving usefulness in this technique.

In the 1990s, however, breakthroughs in computing technologies fostered the developments in machine translation. Many initiatives started, for instance, the EUROTRA project that uses a transfer model derived from RMBT to promote internal language translation within the European communities(Maegaard, 1989). Likewise, SMT also experienced a boom in recognition. In the paper *A statistical approach to machine translation*(Brown et al., 1990) and *The Mathematics of Statistical Machine Translation: Parameter Estimation* (Brown et al., 1993), the authors reintroduced the idea of combining statistical methods with machine translation. The first paper has proven the value of statistical methods in speech recognition, lexicography, and NLP, while the second substantiated such ideas by applying statistical models for predicting the probability of word-by-word alignment.

This strategy quickly developed from word-based to phrase-based. Compared to the former approach, phrase-based SMT allows for greater flexibility and accuracy and is able to translate more complex sentence structures. Not only offering a better way to analyze sentences, but the phrase-based SMT also indicates the model's potential for deeper learning. In the famous paper *Statistical Phrase-based Translation*, the authors proposed a new phrase-based translation model with decoding algorithms. This paper signified the correctness of the phrase-based analysis, paving the way for the later development of more advanced SMT models and even Neural models(Koehn et al., 2003).

Together with newly emerged SMT models, the need for a reliable evaluation metric for the models is more and more urgent. In 2003, Och and Ney published a paper for a comparison of various statistical alignment models(Och and Ney, 2003). In that paper, the authors compared the performance of models using alignment error rate (AER) as a key indicator. Despite its usefulness for analyzing word alignment, it is not as widely adopted as the metric that came up earlier in 2002, the BLEU score. In the paper *Bleu: a Method for Automatic Evaluation of Machine Translation*, the new metric called Bilingual Evaluation Understudy was proposed based on comparing machine-generated translations to a reference translation using a modified n-gram overlap score (Papineni et al., 2002). This new metric had a significant impact on the development of MT systems later and became a generally accepted standard for evaluating MT output.

4 Example-based Machine Translation and Hybrid Machine Translation

Example-based Machine Translation (EMBT) is an alternative to RBMT and SMT which utilizes previously translated examples as resources for translation. Instead of rules or statistical methods, EMBT focused on the existing materials and could achieve high accuracy because of the exponential growth of corpora after the 1990s. In What is Example-Based Machine Translation, the author Turcato identified its advantages to handle certain complex translation problems for both RBMT and SMT, such as idiomatic expressions, phrasal verbs, and complex sentence structures (Turcato and Popowich, 2003). It also requires less human intervention compared to the other two as it relies on existing translations and possesses the potential to improve itself as the examples accumulate over time. On the other hand, EMBT has limitations in that it cannot interpret texts whose examples were not provided as training material. Moreover, EMBT could suffer from overfitting and fail to generalize its prediction to new examples that have different traits to its training set. The explosion of the internet in the 21st century provides the best database for EMBT. As early as 1999, the idea of exploiting the World Wide Web as a resource for EMBT translation tasks was mentioned in the paper by Grefenstette (Grefenstette, 1999). While the other strategy pursued advancement in algorithms, EMBT seeks to exploit currently available resources for improvement.

As more and more approaches appear in this field, researchers began the trial to take advantage of each through combination. The result of which was what's called hybrid machine translation. Among different means to combine the strategies, Lagarda *et al* tried to apply a phrase-based SMT model to the RBMT system for translation to generate suggestions for editing the original output (Lagarda et al., 2009). The author reported a significant improvement in the quality of the translation after editing. In *A Machine-Learning Framework for Hybrid Machine Translation*, Federmann also proposed a framework for combining RMBT and SMT by uniting the output of the two systems and assigning weights to each (Federmann, 2012). Those two and many other publications have demonstrated the possibility of utilizing hybrid approaches to achieve better quality.

5 Neural Machine Translation

5.1 Encoder-decoder Architecture with Attention

As time proceeds to the recent decades, the research of machine translation is dominated by a new paradigm, Neural Machine Translation (NMT). This new technique relies on the powerful deep-learning capability of large neural networks and gets rid of the need for manmade rules or statistical models. While there are many new models viewed as stretches of NMT, this approaches in reality has experienced revolutionary changes in barely ten years.

One of the most influential models in the 2010s, regarded as the basis for the majority of further research, is the Encoder-decoder model proposed by Sutskever *et al* in their paper *Sequence to Sequence Learning with Neural Networks* published in 2014. In this paper, the authors introduced the sequence-to-sequence (seq2seq) architecture which consists of an encoder that processes the input sequence and generates a fixed-length representation, and a decoder that generates the output sequence based on such a representation (Sutskever et al., 2014). This model outperformed traditional phrase-based SMT and was quickly adopted as the presiding approach in the field of NMT. It also triggered hundreds of new studies as extensions of this model. Bahdanau *et al*, for example, built upon the encoder-decoder structure by introducing an attention mechanism that allows the model to selectively focus on different parts of a sentence (Bahdanau et al., 2016). Combining the model with the attention mechanism quickly becomes mainstream in this field. In 2015, Thang Luong and his colleagues analyzed the different types of attention in this architecture and found that a global attention system performs best (Luong et al., 2015). There is also a study that tried to use this architecture together with SMT phrase-based ideologies and showed that it has achieved state-of-the-art that outperforms existing models (Cho et al., 2014) Studies like those further substantiated the effectiveness of the encoder-decoder model combined with the attention mechanism, lighting up a promising direction for researchers to concentrate on.

Besides academia, this model was also adopted by business companies. In 2016, Google built its own NMT system based on the encoder-decoder architecture with attention. Compared to previous algorithms, this model performs well in processing input sentences of arbitrary length, again demonstrating its capability to produce a fluent translation with semantic and syntactical accuracy (Wu et al., 2016). In a study in 2017, the team proposed a new model based on this architecture that focused on the same problem, the efficiency in processing sequences of variable length, by using a convolutional neural network as encoder and decoder, and proved its productivity through their findings (Gehring et al., 2017). As a result, the encoder-decoder architecture with attention becomes a universal standard in the second decade of the 21st century.

5.2 Transformer Architecture

The transformer architecture was another great progress built upon the encoder-decoder system. In the influential paper *Attention Is All You Need*, the authors introduced the transformer architecture, a neural network that does not rely on recurrent or convolutional layers, but solely on attention mechanisms. The transformer model has shown extraordinary efficiency with less computational cost and has since become a cornerstone of a wide range of applications (Vaswani et al., 2017). This transformer-based model inspired many further explorations.

In 2019, a team published their pre-trained model based on the transformer architecture and achieved excellent accuracy using a bidirectional mechanism. In specific, this model, Bidirectional Encoder Representations from Transformers or BERT, was trained on a large corpus of unannotated text and fine-tuned with a bidirectional transformer architecture that could model the dependency of words from both directions (Devlin et al., 2019). BERT has achieved state-of-the-art results on a wide range of NLP tasks and is then used by most industry practitioners.

There are many other new algorithms and developments built upon the transformer architecture. For instance, XLNet, a model proposed by Yang and his team was aiming to address some limitations of the previous transformer-based models like BERT, by introducing an autoregressive approach that models the joint distribution of all possible permutations of a sequence(Yang et al., 2020). Another study tried to facilitate translation among a large number of languages, potentially thousands of languages, using a multilingual transformer-based architecture(Aharoni et al., 2019). In the same year, another research explored the text-to-text algorithm and trained their model called T5 (Raffel et al., 2020). From the other perspective, Beltagy *et al* proposed the "Longformer", a transformer-based model focused on processing long documents which is applicable for documents up to 4 times longer than previous state-of-the-art models (Beltagy et al., 2020).

6 Conclusion

In conclusion, starting from the second half of the 20th century, Machine Translation has developed from a potential direction toward an actually applicable technique and an important component of the current Machine Learning industry. From RMBT to SMT, and then to NMT, despite the limitations of the corpus and computer capabilities, researchers have made continuous progress through algorithms and accumulated more and more resources for newcomers. Nowadays, State-of-the-Art is no longer a supreme standard but something that is always challenged by new models and techniques. This field burst into unprecedented vigors while new directions emerge one after another. Along with the boom of Artificial Intelligence in the new decade, machine translation is going to experience its golden age and hence embrace a new life from its tortuous history.

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

ALPAC. 1966. *Languages and machines: computers in translation and linguistics*. National Academy of Sciences, National Research Council, Washington, D.C. A report by the Automatic Language Processing Advisory Committee, Division of Behavioral Sciences, National Academy of Sciences, National Research Council. (Publication 1416.) 124pp.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2016. Neural machine translation by jointly learning to align and translate.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer.

Peter F Brown, John Cocke, Stephen A Della Pietra, Vincent J Della Pietra, Fredrick Jelinek, John D Lafferty, and Robert L Mercer. 1990. A statistical approach to machine translation. *Computational linguistics*, 16(2):79–85.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Christian Federmann. 2012. A machine-learning framework for hybrid machine translation. In *KI 2012: Advances in Artificial Intelligence*, volume 7526 of *Lecture Notes in Computer Science*, pages 79–90, Berlin, Heidelberg. Springer, Springer.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1243–1252. JMLR. org.

Gregory Grefenstette. 1999. The world wide web as a resource for example-based machine translation tasks.

WJ Hutchins. 2004. The georgetown-ibm experiment demonstrated in january 1954. *Proceedings of the 6th Conference of the Association for Machine Translation in the Americas (AMTA-2004)*, pages 1–13.

Hiroyuki Kaji. 1988. An efficient execution method for rule-based machine translation. In *Coling Budapest 1988 Volume 2: International Conference on Computational Linguistics*.

Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 127–133.

A-L Lagarda, V Alabau, F Casacuberta, R Silva, and E Díaz-de Liaño. 2009. Statistical post-editing of a rule-based machine translation system. In *Proceedings of NAACL HLT 2009: Short Papers*, pages 217–220. Association for Computational Linguistics.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

Bente Maegaard. 1989. Eurotra: The machine translation project of the european communities. *Machine Translation*, 6(1-2):3–16.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks.

Davide Turcato and Fred Popowich. 2003. What is example-based machine translation? In Michael Carl and Andy Way, editors, *Recent Advances in Example-Based Machine Translation*, volume 21 of *Text, Speech and Language Technology*, pages 1–18. Springer.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2020. Xlnet: Generalized autoregressive pretraining for language understanding.