Machine Translating From English to Chinese for E-Commerce Product Categorization

Kitty Duong
University of Windsor
Windsor, Canada
duongy@uwindsor.ca

Miaomiao Zhang
University of Windsor
Windsor, Canada
zhang3s2@uwindsor.ca

ABSTRACT

This paper provides an overview of the history, underlying concepts, and current state of the art in Machine Translation(MT) models. We will also explain how we will utilize and build up the current Machine Translation algorithms to improve the translation of E-Commerce Product Categorization from English to Chinese.

KEYWORDS

E-Commerce Product Categorization, English-Chinese translation, Machine Translation, Multilingual NLP

1 MACHINE TRANSLATION (MT)

Machine translation is an automatic translating system from one language to another while utilizing Artificial Intelligence (AI) without human involvement. The current model of MT is not only a simple word-to-word translation, but it also analyzes the input texts to understand and recognize how words are being influenced by one another. As a result, produces the most accurate translation from one language to another as possible[1].

1.1 History

The origins of machine translation can be traced back to the mid-20th century, around the 1940s to 1950s. Scientists and researchers envisioned automated systems that could assist in translating documents for military and diplomatic purposes. In 1954, researchers collaborated to develop the "Georgetown-IBM Experiment Model 1." This system translated Russian sentences into English, focusing on scientific and technical texts, an example of rule-based machine translation approaches. After that, in 1966, the U.S. government commissioned "The Automatic Language Processing Advisory Committee" (ALPAC) report, which was critical of the limited success achieved at that time and led to a reduction in funding for machine translation projects. In the 1990s, there was a shift towards statistical approaches to machine translation, instead of relying on explicit linguistic rules. The 2010s witnessed a significant breakthrough with the introduction of neural machine translation (NMT). NMT relies on artificial neural networks, particularly recurrent neural networks (RNNs) and later, transformer models. From 2017 till now, Transformer models, such as OpenAI's GPT and Google's BERT, have further advanced the capabilities of machine translation. These models leverage attention mechanisms and pre-training on large datasets to achieve state-of-the-art performance in various natural language processing tasks, including translation.

Conference'17, July 2017, Washington, DC, USA
2024. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 MACHINE TRANSLATION METHODS

There are multiple different approaches when using Machine Translation software/algorithms, such as rule-based, statistical, neural, and hybrid machine translation. Each approach has its pros and cons, but all machine translations generally follow a basic two-step process. First, they decode the source language for the meaning of the original text, and then they encode that meaning to the target language[1].

2.1 Rule-based Machine Translation (RBMT)

The methods of Rule-based Machine Translation are mainly rely on linguistic rules and dictionaries and based on explicit rules crafted by linguists. It improved over early handcrafted approaches by incorporating more sophisticated rules and increased rule coverage for better translation accuracy. The dataset was based on custom-built rule databases and bilingual dictionaries.

One of the earliest experiments utilizing Machine Translation happened in early 1954, known as the Georgetown-IBM Experiment, which was developed by utilizing the rule-based approach to machine translation. This experiment was a collaboration between IBM and Georgetown University led by Léon Dostert and Cuthbert Hurd. The experiment's final product demonstrated the translation of 49 sentences from Russian to English to the public[4]. The goal of this experiment was to figure out any grammatical and morphological problems with the algorithm and predict what is doable with the algorithm going forward. The experiment was planned to be conducted using a small number of sentences from organic chemistry and other general topics, with only 250 lexical items and six syntax rules for the computer to follow[5]. Another rule-based machine translation research was conducted at the University of Washington, led by Erwin Reifler. This research utilized the construction of multiple bilingual dictionaries, where the lexicographic information was used to select the equivalents lexically and solve grammatical problems without analyzing the syntax. From 1959, the results of this research were used by IBM to develop a Russian-English system used by the US Air Force for translation purposes for many years. However, Systran later replaced this system in 1970[4].

Given that rule-based machine translation works by implementing different dictionaries, it can be customized to use for many different purposes, topics, and industries. However, due to the reliance on dictionaries and rules developed by the language experts, if the source texts include any misspelled words, or if the words do not exist in the dictionaries, the final translation will be incorrect. The only way to improve the accuracy of this approach requires the dictionaries to be updated constantly[1].

2.2 Statistical Machine Translation (SMT)

In terms of method, this category of Machine Translation utilizes statistical models trained on large bilingual corpora. Translation is generated based on learned probabilities of word and phrase occurrences. For the improvements, translation quality was enhanced by capturing statistical patterns. The handling of context and phrase-based translation was improved as well. Parallel corpora are used as the dataset for training, containing aligned sentences in source and target languages.

There are two main SMT methods, word-based and phrase-based. The idea of SMT was proposed in 1990 by Brown et al. In 1999, research was performed at Johns Hopkins, introducing an SMT toolkit called Egypt as the result. Two word-based SMT toolkits, GIZA and GIZA++, were also released shortly after. In 2003, the phrase-based SMT was introduced, which promised translation quality improvement. Based on the phrase-based method, the open-source MT system "Pharaoh" was released, and was later upgraded to "Moses". These toolkits and systems greatly improved the SMT adopted rate by the public. As a result, phrase-based SMT was used by Google to develop and launch its translation system in 2006, followed by other companies in the next few years[6].

Given the success, many researchers have begun to propose new models to improve the performance of SMT, such as factored SMT, hierarchical SMT, and syntax-based SMT. However, SMT also introduced reordering issues when translating distant language pairs. This is due to SMT models utilizing log-linear models to implement multiple designed components, such as translation model, language model, etc[6].

2.3 Neural Machine Translation (NMT)

Neural Machine Translation (NMT) represents a paradigm shift in machine translation, moving away from traditional rule-based and statistical methods. The core method involves the use of artificial neural networks, particularly sequence-to-sequence models, which take a sequence of words (source language) as input and generate another sequence of words (target language) as output. Neural Machine Translation also utilizes recurrent neural networks (RNNs) and transformers, and end-to-end learning with direct mapping from source to target language[3]. Large parallel corpora were used as the dataset for training neural networks, which are collections of sentences in the source language paired with their translations in the target language.

In 2014, the idea of NMT was proposed by Bahdanau et al. and Sutskever et al., which was used to map the source language to a dense semantic representation and used an attention mechanism to get the final translation. At the same time, a new multilingual translation framework was introduced by Dong et al. utilizing NMT[6]. Compared with previous approaches, translation quality was significantly improved. NMT allows for end-to-end learning, meaning that the model directly learns the mapping from source to target language without relying on explicit linguistic rules or intermediate representations. Besides, NMT models excel at capturing contextual information and dependencies between words, leading to more fluent and context-aware translations. Given the new idea and framework, many companies started to implement NMT in their software and platforms. For example, in 2015, the first large-scale NMT system was deployed by Baidu. In 2016, Google also introduced its new NMT system, which also led other companies to introduce their version of the NMT system as well[6].

2.4 Hybrid Machine Translation (HMT)

Each Hybrid Machine Translation software utilizes multiple MT models, which greatly improves the effectiveness of only using a single translation model. HMT is a combination of RBMT, SMT, and NMT. Given the HMT model nature, there have been multiple system combinations purposed, developed, and implemented. Table 1 below presents the three main MT combinations: sentence, subsequential, and search graph; and some references for them[2].

Table 1: System combination summary[2]

Granularity	References
Sentence-level	Callison-Burch and Flournoy (2001), Nomoto (2004), Akiba et al. (2002), Costa-jussà et al. (2007), Formiga et al. (2013)
Subsentencial-level	Jayaraman and Lavie (2005), Matusov et al. (2006), Sim et al. (2007), Rosti et al. (2008), He et al. (2008), Mellebeek and van Genabith (2006)
Graph-level	Li et al. (2009), DeNero et al. (2010), Duan et al. (2011), Okita and van Genabith (2012)

2.5 Comparison and Conclusion

Table 2 below summarizes the definitions and differences between the 4 Machine Translation approaches. For this project, we have decided to utilize NMT to develop, train, and implement our algorithm to improve upon the current machine translation system. Our main objective for this project will be analyzing datasets that include a list of all product categories from Amazon US, the most popular E-Commerce platform. Then by utilizing Machine Translation, translate these product categories from English to Chinese and vice versa. The result of this research will provide great support in developing the translation software we describe in our proposal.

Table 2: Comparison of Machine Translation Approaches

Approach	Improvements	Dataset
RBMT	Enhanced rule coverage	Custom-built rule databases
SMT	Improved translation quality	Parallel corpora
NMT	Superior context handling	Large parallel corpora
Hybrid	Enhanced adaptability	Diverse datasets

REFERENCES

[1] Amazon. What is machine translation? - neural machine translation explained - aws, Feb 2024.

[2] Marta R. Costa-jussà, Reinhard Rapp, Patrik Lambert, Kurt Eberle, Rafael E. Banchs, and Bogdan Babych. *Hybrid Approaches to Machine Translation.* Springer, Jul 2016.
[3] Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. A survey of multilingual neural machine translation. *ACM Computing Surveys*, 53(5):1–38, Oct 2020.
[4] W. John Hutchins. Machine translation over fifty years. *Histoire Épistémologie Langage*, 23(1):7–31, 2001.
[5] W. John Hutchins. *The Georgetown-IBM Experiment Demonstrated in January 1954*, page 102–114. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004.
[6] Haifeng Wang, Hua Wu, Zhongjun He, Liang Huang, and W Church, Kenneth. Progress in machine translation. Mar 2021.