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 developing history, underlying concepts, and current state of the art in Machine Translation(MT). 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)

1.1 History

The origins of machine translation can be traced back to mid-20th century, around 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, which is an example of rule-based machine translation approaches. After that, in 1966 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.

2 PRIOR WORKS

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.

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

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 in general, all machine translations 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

Rule-based machine translation uses linguistic rules and bilingual dictionaries developed by language experts to translate content accurately[1]. 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 final product of the experiment was able to demonstrate the translation of 49 sentences from Russian to English to the public[2]. 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[3]. 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 lexical 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[2].

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

2.3 Neural Machine Translation

2.4 Hybird Machine Translation

3 POSITION THE CURRENT RESEARCH (2024)

REFERENCES

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

[2] W. John Hutchins. Machine translation over fifty years. *Histoire Épistémologie Langage*, 23(1):7–31, 2001.

[3] W. John Hutchins. *The Georgetown-IBM Experiment Demonstrated in January 1954*, page 102–114. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004.