ISSN 2616-5783 Vol.6, Issue 21: 6-12, DOI: 10.25236/AJHSS.2023.062102

An Analysis of the Factors of Mistranslation in Statistical Machine Translation—Taking Prose Text Translation as Example

Yon Jee Kwun (Yang Yikun)[1,a], Yon Jee Han (Yang Yihan)[2,b]

[1]*Foreign Language School, Gannan Normal University, Ganzhou, China*
[2]*School of Event and Communication, Suibe, Shanghai, China*
[a]*1026915492@qq.com,* [b]*19914719632@163.com*

Abstract: *This article analyzes the factors of mistranslation in Statistical Machine Translation (STM) by taking prose translation as an example. First, it reviews the basics of STM and prose translation, and then discusses the common mistranslation types in STM, including lexical, syntactic, semantic, and pragmatic mistakes. Next, it identifies possible factors contributing to these mistranslations, including the quality of training data, the selected translation model, and the limitations of machine learning algorithms. Finally, it proposes some possible solutions to reduce mistranslation in STM, such as improving the quality of training data, selecting more appropriate translation models, and exploring new mistranslation detection techniques. The analysis presented in this article provides valuable insights into the challenges and opportunities in STM, and helps improve the accuracy and quality of machine translation.*

Keywords: *Statistical Machine Translation; Prose Text Translation; Mistranslation*

1. Introduction

1.1 Background

Machine translation has become an increasingly popular tool for translating text across different languages. Statistical machine translation (SMT) is one of the most widely used approaches, which relies on algorithms that learn statistical patterns from large bilingual corpora to generate translations. However, despite its widespread use, SMT systems still suffer from mistranslations, which can lead to significant errors and misunderstandings.

One major factor contributing to mistranslation in SMT is the complexity of natural language, particularly in the context of prose text. Prose text often contains idiomatic expressions, cultural references, and other nuances that are difficult for SMT systems to accurately translate. Additionally, SMT systems are prone to errors when translating between languages with different syntax or grammatical structures.

To address these issues, researchers have proposed various approaches to improve the accuracy of SMT systems for translating prose text. These include improving the quality of training data, selecting more appropriate translation models, and exploring new mistranslation detection techniques.

1.2 Objective

The objective of this research is to analyze the factors contributing to mistranslation in statistical machine translation (SMT) systems, which aims to identify the specific challenges and limitations of SMT when it comes to accurately translating prose text, and propose strategies for improving translation accuracy, with a focus on translating prose text.

1.3 Methodology

This research will employ a mixed-methods approach to analyze the factors contributing to mistranslation in statistical machine translation (SMT) systems for translating prose text. The study will begin with a comprehensive review of existing literature on SMT, prose text and the types of

ISSN 2616-5783 Vol.6, Issue 21: 6-12, DOI: 10.25236/AJHSS.2023.062102

mistranslation. This will involve a systematic review of relevant academic journals, conference proceedings, and other scholarly publications.

Moreover, according to the literature review, the study will propose strategies and approaches to mitigate mistranslation in SMT systems for prose text translation, drawing on the findings from the literature review and case studies. These strategies will be evaluated using a combination of expert feedback and user testing to determine their effectiveness in improving translation accuracy. The research methodology employed in this study will provide a comprehensive analysis of the factors contributing to mistranslation in SMT systems for translating prose text and propose practical solutions for enhancing translation accuracy.

1.4 Significance

The significance of this research lies in its potential to improve the accuracy and reliability of statistical machine translation (SMT) systems for translating complex prose text. Mistranslation in SMT systems can have significant consequences, particularly in fields such as international affairs, culture heritage, education and academia, where accurate translation is essential for effective communication and understanding.

By analyzing the factors contributing to mistranslation in SMT systems for translating prose text, this research will provide valuable insights into the limitations and challenges of current SMT approaches. The proposed strategies and approaches for mitigating mistranslation will be of significant practical value to developers and users of SMT systems, particularly in fields where accurate translation is critical.

2. Literature review

2.1 Statistical Machine Translation

Statistical machine translation (SMT) is a widely used approach to machine translation that relies on algorithms that learn statistical patterns from large bilingual corpora to generate translations. The use of SMT has grown rapidly in recent years, driven by advances in computing power and the availability of large bilingual corpora.

One of the key strengths of SMT is its ability to handle large volumes of text, making it well-suited for use in industries such as e-commerce international business, literature translation and cultural communication. SMT systems typically consist of three components: a language model that generates probable translations, a translation model that maps source language to target language, and a decoding algorithm that selects the most likely translation. [1]

While SMT has shown promise in many applications, it is not without limitations. One major challenge is the difficulty of accurately translating idiomatic expressions, cultural references, and other linguistic nuances. Additionally, SMT systems are prone to errors when translating between languages with different syntax or grammatical structures.[2]

To address these limitations, researchers have proposed various approaches to improve the accuracy of SMT systems. One such approach is the use of neural machine translation (NMT) models, which rely on deep learning algorithms to generate translations. NMT has shown promise in improving the accuracy of SMT systems, particularly in handling complex linguistic structures.[3] Another approach to improving SMT accuracy is the incorporation of more linguistic knowledge into the translation process. This can include using syntactic and semantic information to guide translation, such as leveraging human feedback to improve translations.[4]

Despite these advances, SMT still faces significant challenges in accurately translating complex prose text. For example, SMT systems struggle with correctly translating idiomatic expressions and other linguistic nuances that are common in prose text. Additionally, SMT systems may produce mistranslation when translating between languages with different syntax or grammatical structures.

2.2 Prose Text

Prose text is a form of written language that is typically used in literature, journalism, and other forms of non-poetic writing. Prose text is characterized by its use of sentences and paragraphs, and its focus on narrative or informational content. In the context of machine translation, translating prose text is

ISSN 2616-5783 Vol.6, Issue 21: 6-12, DOI: 10.25236/AJHSS.2023.062102

particularly challenging due to the presence of idiomatic expressions, cultural references, and other linguistic nuances.[5]

One of the key challenges in translating prose text is accurately capturing the meaning of idiomatic expressions. Idiomatic expressions are phrases that have a meaning that is different from the literal meaning of the words used. For example, the expression "kick the bucket" means to die, but has no literal connection to kicking a bucket. Translating idioms accurately requires an understanding of the cultural context in which they are used, as well as an understanding of the specific nuances of the language being translated. Another challenge in translating prose text is accurately conveying cultural references. Cultural references are elements of a text that are specific to a particular culture or society, such as historical events, literary allusions, or pop culture references. Translating cultural references accurately requires an understanding of the cultural context in which they are used, as well as an understanding of the specific nuances of the language being translated.

In addition to these challenges, translating prose text also requires an understanding of the specific linguistic nuances of the language being translated. For example, languages may have different grammatical structures or word orders that can affect the meaning of a sentence. Translating accurately requires an understanding of these nuances and the ability to convey them effectively in the target language.

Due to the long history of Chinese prose and the division between classical and vernacular Chinese, translating Chinese prose into English presents special challenges that are different from those of other languages. It is not easy to define what prose is in just one sentence because the concept of Chinese prose has been dynamically changing over time with its application. In ancient China, prose was not a literary concept but a linguistic one that was opposed to rhymed verse. Rhymed verse was a language that emphasized sound and rhythm, while prose was a language that did not have such requirements. In modern China, the concept of prose, especially modern vernacular prose, is not the same as ancient Chinese prose as a linguistic material, but rather comes from the Western essay, which combines narrative and argumentative functions.[6] Therefore, translating Chinese prose is not an easy task.

2.3 Mistranslation

Mistranslation is a common problem in machine translation, particularly in the context of statistical machine translation (SMT). In the process of translating prose text, mistranslation can occur for a variety of reasons, including the presence of idiomatic expressions, cultural references, and other linguistic nuances that are difficult for SMT systems to accurately translate.

Generally speaking, all these reasons can be concluded as lexical, syntactic, semantic and pragmatic mistakes. To mitigate mistranslation when translating prose text with SMT, there are possible solutions such as improving the quality of training data, selecting more appropriate translation models, and exploring new mistranslation detection techniques.

3. Common Mistranslation Types in Statistical Machine Translation

3.1 Lexical Mistranslation

SMT systems are useful for automatic language translation and can be trained using large amounts of bilingual or multilingual text. The quality of these systems can be evaluated by human judges or through automated metrics. Many SMT systems today do not rely on linguistic knowledge and instead use a direct translation approach at the lexical level.[7]

One of the primary causes of lexical mistranslations in SMT is the over-reliance on statistical probabilities to determine the most likely translation of a word or phrase. This approach can be effective for translating simple, straightforward sentences, but it often fails to capture the nuances and complexities of more complex prose text. As a result, SMT systems may incorrectly translate idioms or expressions, or may miss important contextual cues that affect the meaning of a particular word or phrase.

Another factor that contributes to lexical mistranslations in SMT is the lack of context awareness. SMT systems typically analyze text in isolation, without considering the broader context in which it is used. This can lead to errors such as ambiguous translations of homonyms or incorrectly translated named entities, which may have multiple possible translations depending on the context.

Furthermore, the quality of the training data used to train SMT systems can also impact the occurrence

ISSN 2616-5783 Vol.6, Issue 21: 6-12, DOI: 10.25236/AJHSS.2023.062102

of lexical mistranslations. If the training data is not representative of the types of texts that will be translated, or if it contains errors or inconsistencies, this can lead to errors in the resulting translations.

3.2 Syntactic Mistranslation

Syntactic mistranslations are a common challenge in statistical machine translation (SMT), particularly when translating complex prose text. A well-executed translation should ideally appear as though it was originally composed in the target language. To achieve this, a human translator must adhere to a complex set of grammatical and semantic guidelines, among other considerations, when determining the word order of the translated sentence. [8]

One of the primary causes of syntactic mistranslations in SMT is the lack of understanding of the underlying grammatical structures of the source and target languages. SMT systems rely on statistical models to identify the most likely translations of words and phrases, but they often fail to capture the complex syntactic structures that are common in natural language. This can lead to errors such as incorrect word order, missing or extraneous words, or incorrect use of tense or mood.

Another factor that contributes to syntactic mistranslations in SMT is the lack of semantic understanding. SMT systems often rely on surface-level patterns in the training data to identify translations, without considering the underlying meaning of the text. This can lead to errors such as incorrect translations of idioms or expressions, or incorrect use of synonyms or antonyms.

Furthermore, the quality and quantity of training data can also impact the occurrence of syntactic mistranslations. When the training data is not a representative type of texts translated, or when it is erroneous or inconsistent, the resulting translations can also be erroneous or inconsistent.

3.3 Semantic Mistranslation

Semantic mistranslations pose a significant challenge within the realm of statistical machine translation (SMT), particularly when dealing with the translation of intricate prose text. SMT systems heavily rely on statistical models to determine the most probable translations of words and phrases. However, these models often fall short in capturing the underlying meaning and contextual nuances of the text. Consequently, this can result in various errors, such as inaccurately translating idioms or expressions, or misusing synonyms and antonyms. For SMT system to learn new words and phrases, it is necessary to have word alignments that make sense in the context of phrases. [9]

One of the primary factors contributing to semantic mistranslations in SMT is the limited awareness of context. Typically, SMT systems analyze text in isolation, neglecting the broader context in which it is situated. Consequently, this can lead to errors like translating homonyms ambiguously or incorrectly rendering named entities, which may have multiple possible translations contingent upon the context.

Another factor that contributes to semantic mistranslations in SMT is the deficiency in semantic comprehension. SMT systems often rely on superficial patterns present in the training data to identify translations, without fully considering the underlying meaning of the text. Consequently, this can lead to errors such as misinterpreting idioms or expressions, as well as misusing synonyms and antonyms.

Additionally, the quality and quantity of training data have a significant impact on the occurrence of semantic mistranslations. If the training data fails to represent the text types that will be translated accurately or if it contains errors or inconsistencies, this can result in erroneous translations.

3.4 Pragmatic Mistranslation

When translating prose text using statistical machine translation (SMT), pragmatic mistranslations may occur due to several reasons.

Firstly, the data sources. The data sources used by SMT systems may not cover all possible language expressions, resulting in missing or incorrect translations for some expressions, especially those culture-loaded expressions.

Moreover, the language structure. The structure of different languages, as it is always a research subject of pragmatics, may be different, resulting in different word orders or sentence structures in the original text and the translated text.

Furthermore, the vocabulary and the context. The vocabulary used in different languages may be

ISSN 2616-5783 Vol.6, Issue 21: 6-12, DOI: 10.25236/AJHSS.2023.062102

different, resulting in different translations for the same word in different contexts, and the context of the original text may affect the translation result, but SMT systems may not fully consider these issues during translation.

Nevertheless, the statistical model. The statistical model used by SMT systems may not fully capture the semantics of the original text, for which resulting in pragmatic mistranslations.

4. Possible Factors

4.1 Quality of Training Data

Statistical machine translation (SMT) is a type of machine translation that uses statistical models to automatically translate text from one language to another. One of the most important factors that affects the performance of an SMT module is the quality and quantity of the training data used to build the statistical models. As the training data has been mentioned above for times, it is of importance.

Training data is a collection of parallel texts in the source and target languages that are used to train the statistical models. The more diverse and relevant the training data is, the better the SMT module will be at accurately translating between languages. For example, if an SMT module is being trained to translate from English to Spanish, the training data should include a wide range of English texts, such as news articles, books, and websites, as well as their corresponding Spanish translations.

The quality of the training data can also impact the accuracy of the translations produced by the SMT module. If the training data contains errors or inconsistencies, these may be reflected in the translations produced by the module. For example, if the training data includes poorly translated texts or texts with grammatical errors, the SMT module may learn to produce translations with similar errors.

Another important factor to consider when selecting training data is domain relevance. Different domains have their own unique vocabulary and syntax, and an SMT module that is trained on general texts may not perform as well when translating domain-specific texts. For example, an SMT module that is trained on general English texts may not perform as well when translating medical texts, as medical texts have their own unique vocabulary and syntax.

In addition to the quantity and quality of the training data, the choice of statistical models used in the SMT module also plays a role in its performance. There are several different types of statistical models used in SMT, including phrase-based models, hierarchical phrase-based models, and neural machine translation models. Each model has its own strengths and weaknesses, and the choice of model depends on factors such as the size of the training data and the desired level of accuracy.

4.2 Elected Translation Model

When translating prose text with statistical machine translation (SMT), the choice of translation module can have a significant impact on the quality of the translation output. There are several different types of translation modules that can be used in SMT, including phrase-based models, hierarchical phrase-based models, and neural machine translation models.

Phrase-based models are one of the oldest and most widely used types of SMT models. These models break down the source text into small units called phrases, which are then translated into the target language. The translation decision is based on a statistical model that estimates the probability of a given target phrase given a source phrase. Phrase-based models are relatively simple and easy to implement, but they can struggle with long-distance dependencies and may produce translations that are overly literal.

Hierarchical phrase-based models (HPBMT) are an extension of the phrase-based model that attempt to address some of its limitations. HPBMT models use a hierarchical structure to capture longer-range dependencies between phrases, allowing them to produce more fluent translations. However, these models are more complex and computationally expensive than phrase-based models.

Neural machine translation (NMT) models are a more recent development in SMT and have quickly become the dominant approach in the field. NMT models use deep neural networks to directly model the mapping between source and target languages. These models can capture complex linguistic structures and produce fluent translations that are often more accurate than those produced by phrase-based or HPBMT models. However, NMT models require large amounts of training data and can be computationally expensive to train and run.

ISSN 2616-5783 Vol.6, Issue 21: 6-12, DOI: 10.25236/AJHSS.2023.062102

The choice of translation module depends on several factors, including the size and quality of the training data, the desired level of accuracy, and the computational resources available. In general, NMT models are considered to be the state-of-the-art in SMT and tend to produce the most accurate translations. However, they require large amounts of training data and computational resources, which may not be feasible for all applications.

4.3 Limitations of Machine Learning Algorithms

Machine learning algorithms are widely used in statistical machine translation (SMT) to automatically learn the mapping between source and target languages. While these algorithms have led to significant improvements in translation quality, they still have several limitations when it comes to translating prose text.

One of the main limitations of machine learning algorithms is their reliance on training data. Machine learning algorithms require large amounts of high-quality training data to learn the statistical patterns in the data and make accurate predictions. However, it can be difficult to obtain large amounts of high-quality training data for all language pairs and domains. This can lead to poor translation quality, especially for less common language pairs or specialized domains.

Another limitation of machine learning algorithms is their inability to capture context and meaning beyond the local context. Machine learning algorithms typically operate on a sentence-by-sentence basis, and are not able to capture the broader context or meaning of a text. This can lead to errors in translation, especially for texts that rely heavily on context or idiomatic expressions.

Machine learning algorithms also struggle with rare or out-of-vocabulary words. These are words that are not present in the training data, and therefore the algorithm has not learned how to translate them. While some SMT models attempt to address this issue by using techniques such as phrase-based translation or subword units, rare words can still pose a challenge for machine learning algorithms.

Another limitation of machine learning algorithms is their lack of interpretability. Machine learning models can be very complex, with millions of parameters and hidden layers. It can be difficult to understand how the model is making its predictions or to identify errors in the model. This can make it challenging to improve the performance of the model or to diagnose errors in the translation output.

Finally, machine learning algorithms are limited by their inability to handle multiple languages simultaneously. Most machine learning algorithms are trained on a single language pair, and are not able to handle translations between multiple languages. This can be a significant limitation for applications that require translation between multiple languages, such as multilingual websites or international organizations.

5. Possible Solutions

5.1 Improving the Quality of Training Data

To improve the quality of training data when translating prose text using statistical machine translation (SMT), human feedback can be leveraged to identify errors or inconsistencies in the translations produced by the SMT module. This feedback can then be used to correct the errors and improve the quality of the training data. Additionally, neural machine translation (NMT) models can be used to further improve the quality of the training data. NMT models are able to capture more complex linguistic structures and produce more fluent translations than traditional SMT models. By using NMT models to generate additional training data, the quality and diversity of the training data can be improved, leading to better performance of the SMT module.

5.2 Selecting More Appropriate Translation Models

When selecting appropriate translation modules for translating prose text using statistical machine translation (SMT), several factors should be considered.

First, the size and quality of the training data should be taken into account, as different translation modules may perform better or worse depending on the amount and quality of the training data available. Second, the desired level of accuracy and fluency should be considered, as different translation modules may produce translations that are more or less accurate or fluent. Finally, the computational resources available should also be taken into account, as some translation modules may require more computational

ISSN 2616-5783 Vol.6, Issue 21: 6-12, DOI: 10.25236/AJHSS.2023.062102

power than others.

In general, neural machine translation (NMT) models are considered to be the state-of-the-art in SMT and tend to produce more accurate and fluent translations than traditional SMT models. However, NMT models also require more computational resources and larger amounts of training data than traditional SMT models.

5.3 Exploring New Mistranslation Detection Techniques

To explore new mistranslation detection techniques for prose text translation using statistical machine translation (SMT), a combination of approaches can be employed. This includes analyzing the output of the SMT module to identify common types of mistranslations, developing machine learning models to automatically detect and correct errors, leveraging human feedback to identify inconsistencies, and exploring existing technologies like natural language processing for error detection. By iteratively refining and evaluating these techniques, it is possible to develop novel methods for effectively detecting and addressing mistranslations in the prose text translation process.

6. Conclusion

An analysis of the factors contributing to mistranslation in Statistical Machine Translation (SMT) has been presented above, using prose text translation as an example. An overview of SMT and its application in prose text translation is provided, followed by a discussion of the potential sources of mistranslation in SMT, including data sources, language structures, vocabulary, context, and statistical models.

It is found that data sources are one of the key factors affecting the quality of SMT translations. If the data sources used do not cover all possible language expressions, SMT systems may produce missing or incorrect translations. In addition, language structures, vocabulary, and context also play an important role in mistranslation. For example, differences in word orders or sentence structures in different languages may lead to misunderstandings in translations. Similarly, differences in vocabulary usage in different contexts may result in incorrect translations.

Statistical models are another important factor in mistranslation. If the statistical model used does not fully capture the semantics of the original text, SMT systems may produce semantic mistranslations. To address these issues, researchers are working on improving SMT systems through better data collection and processing, improved language models, and more advanced translation algorithms.

In conclusion, mistranslation in SMT is a complex issue that requires attention from both researchers and practitioners. By understanding the potential sources of mistranslation in SMT, we can better evaluate and improve the quality of SMT translations. It is believed that through continued research and development, SMT systems will continue to improve and provide more accurate and natural translations for users.

References

[1] Cohen, K. Bretonnel, and Andrew Dolbey. "Foundations of Statistical Natural Language Processing (Review)." Language, vol. 78, no. 3, Jan. 2002, pp. 599–599, doi:https://doi.org/10.1353/lan.2002.0150.
[2] Artetxe, Mikel, et al. Unsupervised Statistical Machine Translation. Sept. 2018, doi:https://doi.org/10.18653/v1/d18-1399.
[3] Niehues, Jan, and Eunah Cho. Exploiting Linguistic Resources for Neural Machine Translation Using Multi-Task Learning. Aug. 2017, doi:https://doi.org/10.18653/v1/w17-4708.
[4] Peris, lvaro, et al. "Interactive Neural Machine Translation." Computer Speech & Language, vol. 45, Sept. 2017, pp. 201–20, doi:https://doi.org/10.1016/j.csl.2016.12.003.
[5] Toral, Antonio, and Andy Way. "Machine-Assisted Translation of Literary Text." Translation Spaces, vol. 4, no. 2, Jan. 2015, pp. 240–67, doi:https://doi.org/10.1075/ts.4.2.04tor.
[6] Yang Haoou. "24 styles of Chinese prose". Sichuan people's publishing house, April. 2023, pp. 003-005.
[7] Karageorgakis, P., et al. Towards Incorporating Language Morphology into Statistical Machine Translation Systems. Jan. 2005, doi:https://doi.org/10.1109/asru.2005.1566533.
[8] Khalilov, Maxim, and José A. R. Fonollosa. "Syntax-Based Reordering for Statistical Machine Translation." Computer Speech & Language, vol. 25, no. 4, Oct. 2011, pp. 761–88, doi:https://doi.org/10.1016/j.csl.2011.01.001.
[9] Wong, Yuk Shan, and Raymond J. Mooney. Learning for Semantic Parsing with Statistical Machine Translation. June 2006, doi:https://doi.org/10.3115/1220835.1220891.