Pristine Sentence Translation: A New Approach to a Timeless Problem

Meenu Ahluwalia, Brian Coari, and Ben Brock

[1]Master of Science in Data Science
Southern Methodist University
Dallas, Texas USA
{mahluwalia, bcoari, bbrock}@smu.edu

Abstract. Translating text from one language to another is a continuous technological challenge. Although many technologies, such as Google Translate, have used machine learning and neural networks to close the translation gap, there are still many translation problems to be solved. Issues such as multiple word meanings, proper sentence structure, slang, colloquialisms, and determining the literal meaning of words vs contextual intent of those words are areas where we sometimes still see Google Translate struggle. In this paper we explore an original strategy that provides a solution to these translation issues, demonstrate a proof-of-concept of the solution, and examine the feasibility of a large-scale solution. For our translation solution we populated a database with translations of entire sentences from one language to another, instead of the words in a sentence. Since a sentence represents an entire thought instead of an assembly of words, the translation did not suffer from the issues that plague Google Translate. We also used Natural Language Processing (NLP) and predictive modeling in order to find sentences close to the sentence requested, which provides the user examples of common grammatically-correct sentences. With these approaches we were able to translate sentences that seemed impossible using traditional translation methods.

1 Introduction

The ability to easily communicate with people in another language is one of the most powerful and satisfying experiences in life. Technology has come a long way from the discovery of the Rosetta Stone in 1799, which allowed us to translate Egyptian hieroglyphics to ancient Greek in a mere 23 years. In the modern day, tools such as Google translate can be used in real time to convert between languages and allow people to connect from different cultures[1]. The latest iterations of Google Translate even use machine learning and neural nets to parse more than just single words, delivering a more satisfying user experience[2].

As far as we have come, however, the areas where we struggle are still painfully obvious. While Google Translate will usually allow you to find a bathroom and order off a menu, the intricacies and complexities of a normal, native

conversation still can cause a non-fluent speaker issues. For example, if an American coworker mentions to a Brazilian coworker about their performance on a project with "You hit one out of the park.", the Brazilian coworker could translate the words, but without familiarity with the context of a baseball game, the Brazilian would be confused and would have to ask for clarification if it was possible. It would be even harder if the Brazilian was reading a book in English with a colloquialism, since there would be no human to ask for help.

If you consider these kinds of issues from a high level they might seem unsolvable. How can you train a translation tool to look at the meaning behind sentences using on the words provided? We think we have a possible answer: an original concept we are calling Pristine Sentence Translations (PSTs). The concept of PST is that instead of translating words or phrases using neural nets and machine learning, we simply store an entire sentence in a database, and we have entire sentences in other languages that represent the meaning of that sentence.

For example, using the example above we would have an entry for the English sentence "You hit one out of the park", and we would have an entry for a Portuguese sentence mapped to that English sentence that says "Você foi ótima" which translates in English to the meaning behind the phrase: "You did great". For another example, in Portuguese there's a sentence "Eu adoro Cafuné" Google Translate does not have a translation for "Cafuné", because it's a complicated word which loosely means "the act of running fingers through hair" . Our program's goal is to return an English translation "I love the feeling of fingers running through my hair" when asked to translate "Eu adoro Cafuné" into English. Using this method there is no sentence or concept we will not be able to translate into another language given enough time and resources.

One main issue with the approach outlined above is that if we do not have an exact match for the sentence, our method return nothing. so if we tried to translate "You really hit one out of the park" from English into Portuguese we would not get any results. We decided to address this concern using Natural Language Processing (NLP) to filter out the noise in a sentence, and then use Predictive Modeling in order to find the sentence "most like" the input sentence. Using this method, "You really hit one out of the park" would ideally map most closely to "You hit one out of the park", and return the same translation: "Você foi ótima". The front-end will indicate that the translation is not for the original input sentence, instead it will indicate that it is "Showing Results for: You hit one out of the park."

Due to the strictly educational and academic nature of the project, we are not attempting to provide a full translation solution. We will limit our translations to English, Portuguese, and Hindi, and we will only provide translations for a few hundred phrases. This will be sufficient to demonstrate the appeal and power of this technique, and we will show how this solution could grow into a complete, living solution using crowdsourcing and time.

2 State of Translations

Meenu or Brian - Meenu to pick either "State of Translations" or "A New Approach to Translations"

2.1 Existing Tools and Methods

2.2 Outstanding Issues

3 A New Approach to Translations

Meenu or Brian - Meenu to pick either "State of Translations" or "A New Approach to Translations"

3.1 Pristine Sentence Translations Theory

3.2 Pristine Sentence Translations In Action

3.3 Database Design

4 Predictive Modeling

Meenu citing https://machinelearningmastery.com/develop-neural-machine-translation-system-keras/

4.1 Data Cleansing

4.2 Building the Neural Translation Model

4.3 Evaluating the Neural Translation Model

5 Full Demo

Brian, but not by Friday. Maybe Sunday.

5.1 Conclusions

6 Ethical Considerations

Meenu or Brian - Meenu to pick either "Ethical Considerations" or "Conclusions and Other Work"

7 Conclusions and Other Work

Meenu or Brian - Meenu to pick either "Ethical Considerations" or "Conclusions and Other Work"

8 References

1. Google's new translation software is powered by brainlike artificial intelligence
(2016, September 27), Retrieved February 04, 2019, from https://www.sciencemag.org/news/2016/09/google-s-new-translation-software-powered-brainlike-artificial-intelligence?r3f_986=https://www.google.com/

2. Found in translation: More accurate, fluent sentences in Google Translate
(2016, November 15), Retrieved February 04, 2019, from https://www.blog.google/products/translate/found-translation-more-accurate-fluent-sentences-google-translate