Using Django and MongoDB to Develop a Translation Quality Checking Tool

Subhashree Chowdhury, Ryan Egbert, Poorna K Narasimhan, Suyash Sukthankar, Matthew A. Lanham
Purdue University, Krannert School of Management
krish212@purdue.edu; regbert@purdue.edu; chowdh30@purdue.edu; ssuktha@purdue.edu; lanhamm@purdue.edu

ABSTRACT

As literacy continues to spread throughout the world, more documents are being translated into languages that did not have prior access to such information. To ensure accurate and complete translations, this project provides a system through which the quality of a translation can be determined. A visual interface was built using the Django web framework to allow users the ability to check the quality of translations. The metrics used to measure the quality of the translations are semantic similarity, comprehensibility, and readability.

INTRODUCTION

There are 7117 known languages in the world and 4065 languages currently have a written form. Due to the vast number of languages, and an increase of worldwide literacy, translations are being written for texts ranging from medical information to poetry.

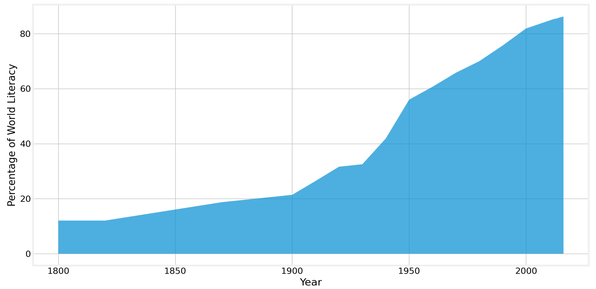

Figure 1. World literacy growth over time

Because much of this information could be vitally important, ensuring accurate translations is key. There are three aspects of translations that need to be considered when looking at translations: semantic similarity, comprehensibility, and text readability.

These three individual metric scores can be combined to determine the overall quality of a specific translation. This overall “score” can be used to determine which areas of a translation must be improved or, more generally, if the quality of the overall translated text must be improved.

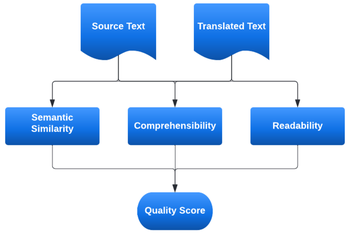

Figure 2. Information flow

QUALITY METRICS

Semantic Similarity – Semantically speaking, how accurate is a specific translation?

Comprehensibility – Is important information captured in a piece of translated text?

Readability – How complex is a passage of text?

METHODOLOGY

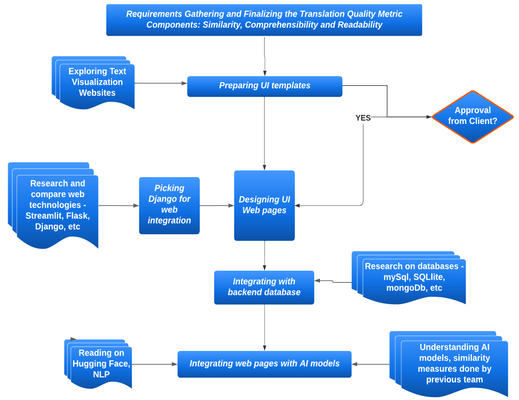

Figure 3. Methodology workflow

TECH STACK

One of the benefits of a software development project is the chance to learn how to utilize new software and technology. Each of these technologies introduced new concepts which contributed to becoming better programmers and critical thinkers.

Figure 4. Technology utilized in the project

FUNCTIONALITY

The application has many different features, all contributing to its overall functionality.

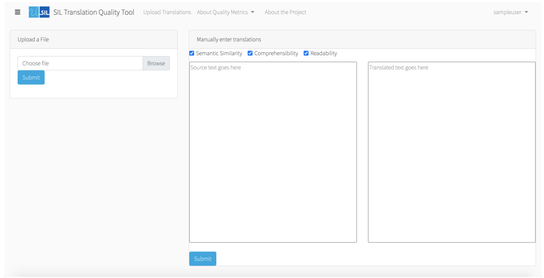

Figure 5. Text upload and entry page

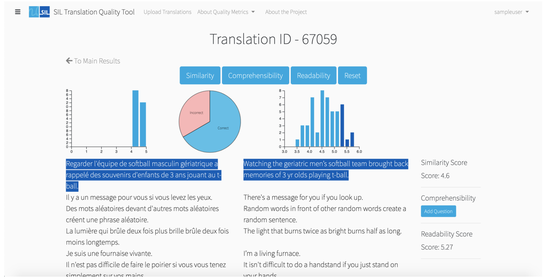

Figure 6. Results page

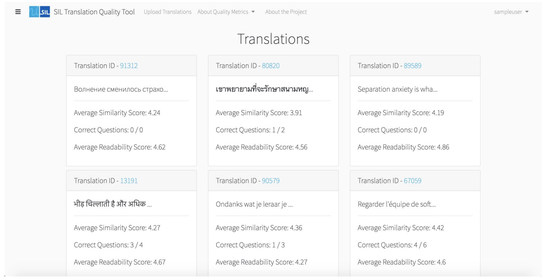

Figure 7. User translation page

EXPECTED IMPACT

A translation quality checking tool does not exist in this current context. This project provides a system in which users of many different types can quickly determine how their work could be improved.

Not only could this service be externally monetizable, but internally, it could increase productivity and output, resulting in more information being widely available to larger groups of people.

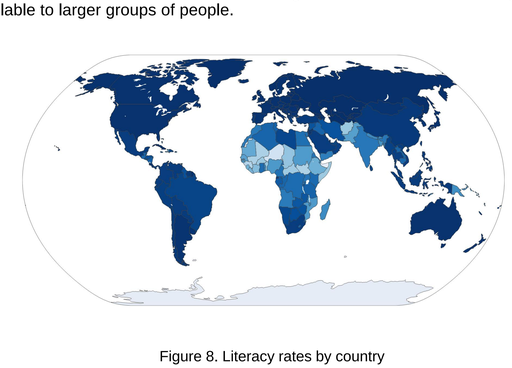

Figure 8. Literacy rates by country

CONCLUSIONS

The Django web framework provided a solid foundation for rapid development of a translation quality checking tool. MongoDB databases consist of a simple, yet effective, structure that could be utilized to store, manage, and process large amounts of unstructured data.

Pre-trained models that were used to check semantic similarity and comprehensibility between two separate texts were very effective at measuring quality scores. There are many different readability algorithms, each of which provides a different understanding of text complexity.

All types of users were provided with ease of access and many different visual metrics of the efficacy and quality measure of their translations.

ACKNOWLEDGEMENTS

We would like to thank Professor Matthew Lanham and our client correspondent at SIL International, Joshua Nemecek.

PURDUE UNIVERSITY
Krannert School of Management

SIL