A Machine Learning Approach to Evaluating Translation Quality

Brenda Reyes Ayala
University of North Texas
Department of Information Science
Denton, Texas, USA
Brenda.Reyes@unt.edu

Jiangping Chen
University of North Texas
Department of Information Science
Denton, Texas, USA
Jiangping.Chen@unt.edu

ABSTRACT

We explored supervised machine learning (ML) techniques to understand and predict the adequacy and fluency of English-Spanish machine translation. Five experiments were conducted using three classifiers in Weka, an open-source ML tool. We found that the highest performance was achieved by applying a dimensionality reduction approach to the classification task, which included collapsing a numeric scale of quality to two categories: high quality and low quality. Our results showed that the Support Vector Machine classifier performed the best at predicting the adequacy (65.65%) and fluency (65.77%) of the translations. More research is needed to explore the methodologies of applying ML to translation evaluation.

Keywords

Machine translation evaluation, machine learning, Weka

1. INTRODUCTION

The evaluation of machine translation (MT) has played a crucial role in advancing the research, development, and application of MT. MT evaluation can be conducted manually and automatically, but both approaches have their challenges. Especially, when no reference translations are available, it is impossible to conduct automatic evaluation. Is it possible to predict the performance of an MT system without reference translations? Generating referent translation is usually a big undertaking for MT researchers and digital library developers.

In this study, we explored the use of supervised machine learning techniques to predict the adequacy and fluency of machine translations. Using Machine Learning methods has an advantage over human evaluation. As has been noted by others [2], hiring humans to evaluate the adequacy and fluency of documents is difficult, expensive, and time-consuming, therefore, human evaluation is difficult to scale to larger data sets. The purpose of this study was to explore the possibility of applying ML for MT evaluation. The question we wanted to answer was: How well do classifiers trained to distinguish between high-quality and low-quality translations predict the adequacy and fluency of metadata records translated from English to Spanish?

JCDL 2017, June 19-12, 2017, Toronto, Ontario, Canada.

DOI: http://dx.doi.org/10.1145/12345.67890

2. RELATED STUDIES

When manually evaluating the performance of an MT system's output, two measures are typically used: *adequacy* and *fluency*. The Linguistic Data Consortium [3] defined fluency as "the degree to which the target is well formed according to the rules of Standard Written English", while adequacy was "the degree to which information present in the original is also communicated in the translation." In the literature, a few studies have been conducted to explore machine learning approaches for evaluating machine translation. Corston-Oliver, Gamon, and Brockett[2] explored how Machine Learning classifiers could be used to distinguish translations created by humans from those created by an MT translation system. Their classifier focused on evaluating the well-formedness of the output sentences (similar to our concept of fluency) according to 46 features picked and extracted by the authors. Finch, and Sumita [4] used as inputs MT outputs whose translation quality had been previously assessed by human evaluators. Rather than having classifiers evaluate the adequacy and fluency of translations on a scale of 1 to 5, they employed a *reduction of classification ambiguity* method, turning a multi-class classification problem into a set of binary classification problems. Their results showed their approached achieved a higher correlation with human judgments (from 0.63 to 0.77) at the sentence level compared to standard automatic evaluation measures.

3. METHODOLOGY

Our data came from a previous project evaluating machine translation for digital metadata records[1]. The data set contained 2,000 English metadata records, and their Spanish translations by three online MT services, Google Translate, Bing Translator and Yahoo! Babel Fish (now inactive), and human evaluations of the performance of the translations using adequacy and fluency. Each metadata record had three Spanish translations, performed by three online MT systems. Each translation had two sets of accuracy and fluency scores produced by human evaluators. In total we had 21,734 score pairs.

We used Weka, an open-source machine learning tool to predict fluency and adequacy. Weka does not directly perform text classification with textual contents, so we used Weka's built-in *StringToWordVector* filter for text classification. Three classifiers were applied to the data sets: Decision Tree, Naïve Bayes, and Support Vector Machine (SVM). For all three classifiers, we applied the default parameters in order to establish a baseline and evaluated their performance using 10-fold cross-validation. Five experiments or runs were conducted. For the first run, the three classifiers were applied separately for each element (title, description, coverage, subject) of each system (Bing, Google, Yahoo!). The second run merged the four metadata elements into a single record and applied the three classifiers separately for each of the systems (Bing, Google, and Yahoo!). The other runs were:

Run 3: Same as Run 2, but this time combine all the elements of the three systems together (Bing, Google, and Yahoo!) together, resulting in a single dataset. Each document is classified as having an adequacy or fluency between 1 and 5.

Run 4: From Run 3's dataset, we created a new one, this time using three classes instead of five: "low quality", "adequate quality", and "high quality".

Run 5: From Run 3's dataset, we created a new one, this time, using only two classes: "low quality" and "high quality"

Due to space constraints, in the next section we report the results of the last three runs. The first two runs did not produce better results than the third run.

4. RESULTS

We discuss the accuracy statistics for each run, as well as the weighted average measures of precision, recall, and f-measure statistics. During the third run, the SVN classifier achieved higher accuracy for both adequacy and fluency (63.18% and 61.60%, respectively), though the Decision Tree had somewhat higher precision and F-measure scores (for adequacy, precision and F-measure were both 0.58, while for fluency, precision and the F-measure were both 0.57).

We concluded that the low performance of our classifiers in Runs 1-3 was due to the lack of a balanced training set. In order to remedy this, we created a new training set, this time using three classes instead of five: "low quality", "adequate quality", and "high quality". Each class was comprised of 326 documents, for a total of 970 documents. The class "low quality" was created by compiling those documents with adequacy and fluency scores of "1" and "2" from the third experiment dataset. The "adequate quality" class was composed of documents randomly sampled from the third experiment dataset that had adequacy and fluency scores of "3", while the "high quality" class was composed of similar documents with adequacy and fluency scores of "5". Documents with adequacy and fluency scores of "4" were removed from the data altogether.

Table 2. Adequacy and Fluency Results for Run 5

	Adequacy			
Classifier	**Accuracy**	**Precision**	**Recall**	**F-measure**
Decision Tree	58.74	0.60	0.59	0.58
Naive Bayes	60.20	0.61	0.60	0.59
SVN	**65.64**	**0.66**	**0.66**	**0.66**
	Fluency			
Classifier	**Accuracy**	**Precision**	**Recall**	**F-measure**
Decision Tree	62.57	0.65	0.63	0.61
Naive Bayes	62.50	0.63	0.63	0.62
SVN	**65.77**	**0.66**	**0.66**	**0.66**

The results of Run 4 showed little improvement over the results of Run 3. For adequacy, the Decision Tree classifier outperformed the other two (60.63% accuracy), but the SVN classifier achieved higher precision, recall, and F-measure results (0.56 for precision and recall, and 0.55 for the F-measure). For fluency, the SVN outperformed classifier resulted in the best performance across all measures, with 61.81% accuracy and 0.62 for precision and recall, and 0.61 for the F-measure).

Not satisfied with the results of Run 4, we utilized the reduction of classification ambiguity approach in order to reduce a multiple classification problem to a binary classification problem, as in [4]. Instead of employing five or three different classes, we simply collapsed them into "low quality" and "high quality" translations, where low quality were those documents classified as having an adequacy or fluency score from 1-3, and high quality documents were those classified as having an adequacy or fluency score of 4-5. Each class "low quality" and "high quality" was comprised of 652 documents, for a total of 1,304 documents. Both classes were created by creating a random sample of low quality and high quality documents from previous experiments.

This approach yielded our most promising results, as can be seen in Table 2. For both adequacy and fluency, the SVN classifier provided the best scores across all metrics. SVN was able to successfully predict the adequacy of 65.64% of documents and the fluency of 65.77% of them. Precision, recall, and F-measure scores were similarly high. Since we are comparing the performance of classifiers with the judgment of human evaluators, we can also treat the accuracy score as a measure of correlation between the Machine Learning classifiers and human evaluators. Thus, concluding that the SVN classifier has achieved a 65% correlation with human judgments of adequacy and fluency.

5. DISCUSSION AND CONCLUSION

The research presented here provides insight into how to evaluate the translation quality of large, heterogeneous collections of short texts, such as metadata records in libraries. The highest performance was achieved by applying a dimensionality reduction approach to the classification task, which included collapsing our numeric scale of quality to two categories: high quality and low quality. Our results showed that the Support Vector Machine classifier performed the best at predicting the adequacy (65.65%) and fluency (65.77%) of translations.

6. ACKNOWLEDGMENTS

We would like to thank Lokesh Viswavarapu for his help on conducting some of the experiments with Weka for this project.

7. REFERENCES

[1] Chen, J. 2016. *Multilingual access and services for digital collections*. Libraries Unlimited: Santa Barbara, CA.

[2] Corston-Oliver, S., Gamon, M., and Brockett, C. 2001. A machine learning approach to the automatic evaluation of machine translation. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics* (pp. 148-155). Stroudsburg, PA: Association for Computational Linguistics. doi: 10.3115/1073012.1073032

[3] Linguistic Data Consortium. (2005). Linguistic data annotation specification:Assessment of fluency and adequacy in translations revision 1.5. Retrieved from http://wayback.archive.org/web/20100622130328/http://projects.ldc.upenn.edu/TIDES/Translation/TransAssess04.pdf

[4] Paul, M., Finch, A., and Sumita, E. 2012. Predicting human assessment of machine translation quality by combining automatic evaluation metrics using binary classifiers. *International Journal of Computer Applications, 59*(10) doi:http://dx.doi.org/10.5120/9581-4062