Evaluating Google Translate using chrF++ metric for select language pairs

Yifu Mu
Georgetown University
ym431@georgetown.edu

Abstract

This paper presents the results of our evaluation of Google Translate for certain language pairs using the chrF++ metric, a recent, nonetheless state-of-the-art Machine Translation evaluation metric that has been frequently discussed in WMT.

1 Introduction

Recent advances in Machine Translation (MT) evaluation systems attempts to pick out nuances in translation errors. The widely-used BLEU (Papineni et. al,.) metric has its empirical advantages, but researchers are looking at ways to zoom in on translation errors beneath the word level. In this paper, I will present an application of one such evaluation metric, chrF++ (Popovic, 2017), on the freely-available Google Translate API.

1.1 Background

chrF++ is an extension of chrF (Popovic, 2015), which exclusively uses character *n*-gram F-scores to determine translation accuracy. Usage of character *n*-gram F-score was chosen because it had been shown in recent literature that it has a high degree of correlation with human judgment. chrF++ builds on top of chrF by incorporating word-level *n*-gram F-scores where n is small, typically 1 or 2. Experiments have been done to show that the addition of word unigrams and bigrams improve the correlation with target translations.

1.2 Objective

Each language pair represents a translation system. Through attaining the chrF++ scores for sentence pairs in a language pair, we can make conclusions regarding the effectiveness of that translation system. When we have attained the scores for different translation systems, it is then possible to determine whether one translation system is significantly better than another. We construct translation systems by building various language pairs, each having a different source language, but all having English as the target language. Upon knowing which system performs better, the next step would be to reverse-engineer and investigate into which variables led to the outcome. However, this paper does not address those further endeavors.

1.3 Statistical testing

Although MT research has made breakthroughs in recent years, few focus on statistical testing of MT evaluation metrics. We believe that it is valuable to present statistical methods employed in the methodology thoroughly in order to show the meaning behind numbers. Obviously, merely obtaining a BLEU score is not enough, and we hope to see more works in the future that ground evaluation in statistical theory. This paper draws inspiration from Koehn (2004), who outlines approaches that researchers can take to more comprehensively explain MT evaluation results. However, due to certain limitations, we were not able to perform as many tests as desired.

2 Data

To conduct experiments that fits our objective, it is important to find a dataset that satisfies the requirements. The kind of corpus that best serves our purpose is one that provides parallel sentences in two languages. Luckily, the Europarl dataset, one of the premier corpuses in MT research, is publicly available, and so we decided to utilize it.

2.1 Dataset

Europarl is a parallel corpus of 11 European languages. It contains transcriptions of the European Parliament proceeding from year 1996-2011 (most recent version - v7). The data is collected from the European Parliament website through crawling, and is aligned sentence by sentence. The English data files contain generally one sentence per line. The corresponding sentence in another language takes that same line number in another file. The dataset is organized such that each language pair (some language – English) with the parallel sentences is put into a folder specific to that language pair. The full size of each file is generally somewhere around 100,000.

2.2 Preprocessing

Due to limited time and computational power, the dataset needed to be severely trimmed down in size. We recognize that tasks need to be run in a reasonable amount of time for grading.

We read all the lines in the provided files, and took the first 300 sentences, and put them in another file. We then made API requests to PyPI's version of Google Translate, and translated each sentence in the condensed dataset.

3 Methodology

For preprocessing techniques, see **section 2.2.** After the initial data preprocessing step, we calculate the chrF++ score[1] for each translation, and then conduct statistical significance tests to determine whether certain translation systems perform better.

Due to a limited-sized dataset and a sub-par performing API, we could not employ interesting statistical methods popular in the field of MT such as bootstrap resampling. There is simply not enough data for us to break up the dataset in chunks and bring in additional data. Therefore, we perform the Student's *t*-test and the paired *t*-test for statistical significance.

3.1 Student's *t*-test

The t-test is one of the most popular parametric tests for statistical significance. It makes several assumptions about the datasets and determines if the means of the sample data are significantly different. The most integral assumption that the *t*-distribution makes is that the data points have a Gaussian distribution. This implies that the individual data points are independent from each other.

For our purposes, the t-test works because we have a collection of chrF++ scores – one for each translation. And from there we can attain the mean from the set of numbers. The independence assumption is easy to make, but could have been more thoroughly addressed if the size of the dataset allowed us to perform broad sampling (Koehn, 2004), where context clues and similar discussion topics in the parliament proceedings could be effectively neutralized by sampling as randomly as possible on a given dataset. Nevertheless, we can rather safely say that all the requirements are satisfied for the t-test.

4 Results

Source language 2	Source language 1				
	it	de	el	fi	nl
it		0.327	**0.043**	0.867	0.754
de			0.339	0.420	0.188
el				0.068	**0.015**
fi					0.630
nl					

Table 1: *p*-value when running Student's *t*-test on the chrF++ scores for all the possible combinations of source language. Significant results are bolded. These results reflect the default chrF++ score implementation, with character *n*-gram set to 6, and word *n*-gram set to 2.

We performed the *t*-test on the chrF++ score sample on every possible combination of source language, the result of which can be seen in Table 1. Assuming a 95% confidence interval, with alpha = 0.05, the evaluation yielded some significant results. We can see that there are two *p*-value scores under 0.05, and they both involve Greek as the one of the two source languages.

Another set of tests were conducted, but with modified parameters – character *n*-gram = 8, and word *n*-gram = 1. These set of parameters give even more predictive power to elements on the sub-word level, and scales back the predictive power of word-level *n*-grams. The results were quite similar, and no changes in significant results.

If we examine the average chrF++ score for Greek, we would find that it is higher than all other

[1] Code available on Github: https://github.com/m-popovic/chrF

source languages. This is somewhat surprising, considering that Greek uses a different writing script than the other four languages (Latin alphabet). This interesting fact deserves some further investigations in the future.

5 Discussion

Obviously, the results of this paper is rather preliminary, due to said limitations mentioned in previous sections. However, there are some surprising findings here. It is within expectation that Greek would perform differently than some of the other source languages, but it is not expected that it would do better, since it writes in a completely different script, and is not grammatically more similar to English than some of the other source languages (Germanic). It would be interesting to investigate why Greek translation into English performs at this level. Is it because of its rich grammatical content?

The results of this paper would be much more convincing if more thorough statistical methods were employed. We plan on carrying this project into the future, go more in depth into the numbers of chrF++. If given enough flexibility with our data, proven statistical practices such as bootstrap resampling could be utilized in order to explain the results more comprehensively.

References

Philipp Koehn. 2004. *Statistical Significance Tests for Machine Translation Evaluation: A Parallel Corpus for Statistical Machine Translation*. Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, MA

Philipp Koehn. 2005. *Europarl: A Parallel Corpus for Statistical Machine Translation*. School of Informatics, University of Edinburgh, Scotland.

Maja Popović. 2015. *chrF: character n-gram F-score for automatic MT evaluation*. Humboldt University of Berlin, Germany

Maja Popović. 2017. *chrF++: Words Helping Character n-grams*. Humboldt University of Berlin, Germany