Evaluating Neural Machine Translation With Backtranslation

Group 9

Abstract

This document contains the motivation behind and results from backtranslation analysis of neural machine translation. Backtranslation quality is assessed by measuring edit distance from the original phrase and backtranslated phrase. Three different target languages (Spanish, Chinese, and Afrikaans) are explored and the results are shared. Additionally, we discuss the results and give thoughts on what leads to the observed results. Given time and resource constraints, this report is limited to presenting introductory findings and possible future work is suggested.

1 Introduction

Neural machine translation (NMT) has advanced significantly in the last decade. With the rise of deep learning and the increased availability of large scale data sets, NMT models are continuing to improve. Currently, the most popular way to evaluate NMT quality is via Bilingual Evaluation Understudy (BLEU) score. BLEU "compares the hypothetical translation to one or more reference translations" and gives a better score when the "candidate translation shares many strings with the reference translation"[1]. We propose an alternate metric for evaluation utilizing the distance between the original phrase and the backtranslation from the target language. Our reasoning for using backtranslation is that the translated sentence should contain the same functional meaning, to the point where, when translated back to the source language, the phrase should remain intact. Any variation from the original source phrase represents a mistake in the translation, at least on a syntax level. We will explore the backtranslation quality across different languages and analyze the results based on edit distance from the original phrase. We then discuss our findings and potential extensions of our work.

2 Related Work

Utilizing backtranslation to evaluate statistical language models' ability to translate English phrases coherently is a novel concept (at least in terms of the research we did). Given that our approach to analyzing the success of a translation is novel, we have no relayed work to build upon. Furthermore, there are no established results for such a task that we can reference or use as a baseline.

3 Model

The translation model used for our experiments is contained in *deep-translator*, a Python library that leverages Google Translate. According to *Google's Neural Machine Translation System: Bridging the Gap Between Human and Machine Translation* (Wu et. al), the model "consists of a deep LSTM network with 8 encoding and 8 decoder layers using residual connection as well as attention connections from the decoder network to the encoder"[2].

4 Dataset

The dataset that we use is a list of translations from Wikipedia[3]. Although we selected certain pairs of translations from source to target language, since we are doing both forward and backtranslation, all we need is a standardized set of English sentences (and not actually the target language translations). The dataset of English phrases originally contained 1,714,910 phrases that contained both alphanumeric and special characters. In order to reduce the size of the data as well as clean the

[1] https://mastertcloc.unistra.fr/2019/04/29/methods-of-the-machine-translation-evaluation/

[2] https://arxiv.org/pdf/1609.08144.pdf

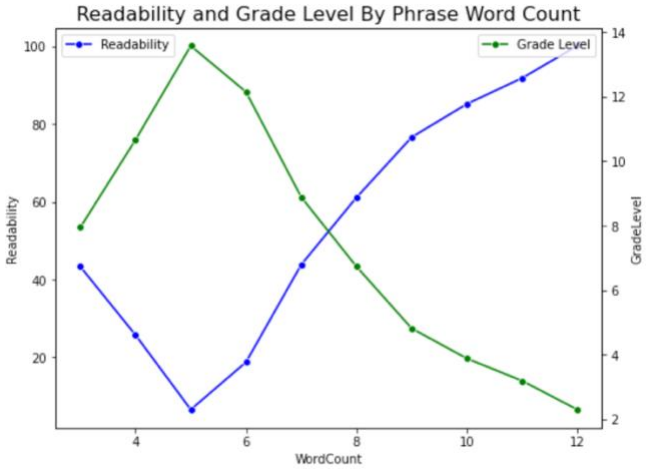
[3] https://opus.nlpl.eu/Wikipedia.php

data, phrases were subset to contain at least 97.5% alphanumeric characters (in order to exclude special characters). Additionally, the dataset was further subset to the first 10,000 phrases due to runtime complexities.

	PhraseLength	WordCount	AlphaNumeric%	Readability	GradeLevel
count	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000
mean	49.691300	8.374000	0.976599	64.735266	6.316520
std	2.559465	1.267077	0.002136	25.880458	3.399286
min	13.000000	3.000000	0.975610	-93.330000	-2.700000
25%	49.000000	8.000000	0.975610	46.440000	3.700000
50%	50.000000	8.000000	0.976190	69.790000	6.000000
75%	51.000000	9.000000	0.976744	85.690000	8.800000
max	52.000000	12.000000	1.000000	119.190000	27.300000

After the dataset was finalized, we assessed the readability and the grade level of the phrase to be translated. The readability of a phrase is defined by the Flesch Reading Ease and is scored up to 121.22 with a higher score meaning the phrase is easier to read (note: there is no lower bound and negative values are valid). The grade level is defined by the Flesch-Kincade Grade and the output is a decimal representing the estimated grade level required for reading of the phrase[4].

We inspect the dataset in terms of the number of words in the English phrases to be translated. Interestingly, the phrases with a higher word count were simpler to read (see plot below). Given that the readability and grade level of phrases are very similar metrics we see an inverse relationship in the plot above: as the readability level increases, the grade level required to understand the phrase decreases.



5 Experiments

The experiments that we conducted focused on measuring the difference between an English phrase and it's backtranslation - translating first to a target language and then translating that phrase back into English. The evaluation metric we used is edit distance, specifically Levenshtein distance. Levenshtein distance counts the number of single character changes (insertions, deletions, substitutions) between two phrases[5]. To account for the length of each phrase, the edit distance was then divided by the number of characters in the original phrase, leaving us with, what we termed, *edit distance %*. We chose to inspect three different target languages; Spanish, Chinese (simplified), and Afrikaans. The motivation behind choosing these languages was so that we would be exposed to a high-resource language with a similar character alphabet (Spanish), a character-based language (Chinese), and a low-resource language (Afrikaans).

6 Results

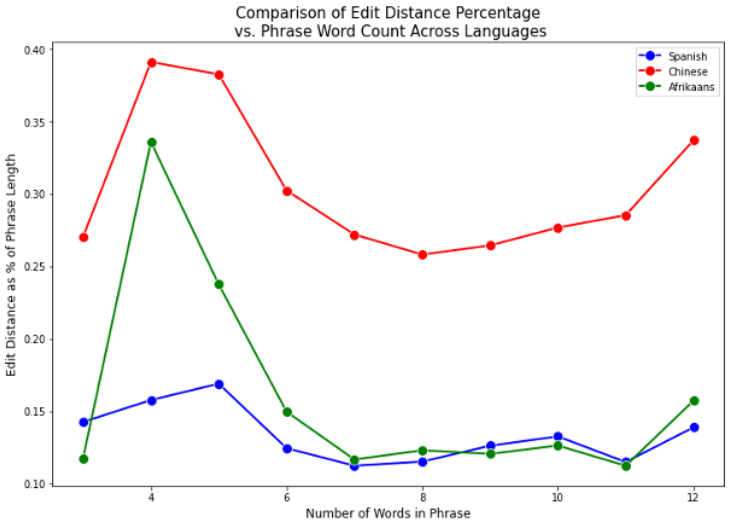
We ran the above experiments for the same 10,000 English phrases, translating first to the target language and then translating the output translation back to English. On average, across all 10,00 phrases, the English-Spanish-English (ESE) backtranslation performed the best, with a mean edit distance % of 12.18%. English-Afrikaans-English (EAE), was second best with a mean edit distance % of 12.46%, surprisingly, only about 2% worse than English-Spanish-English. Lastly, the English-Chinese (simplified)-English (ECE) backtranslation scored much worse, with a mean edit distance % of 26.97%. Following the above results, we see the same order in terms of standard deviation of edit distance %. Both metrics can be seen in the below table.

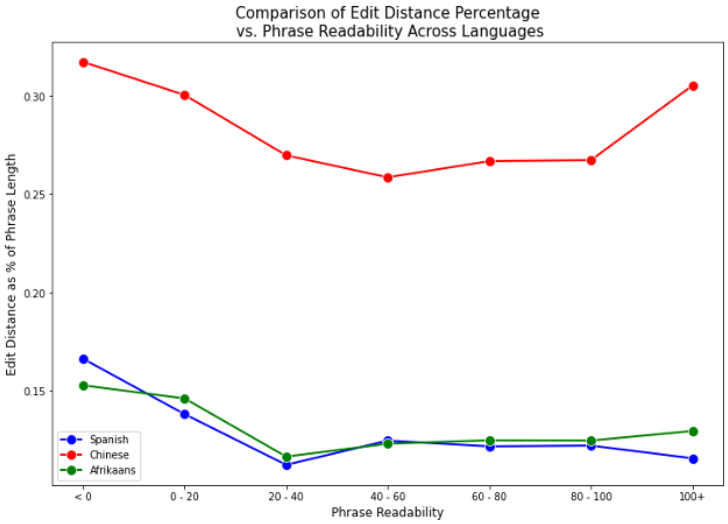
Furthermore, we were interested in evaluating each set of translations by the length and complexity of the phrase (using readability). The results of grouping the data by phrase length and complexity followed closely to the results of the overall performance. Aside from short phrases, on which ESE performed the best, backtranslations with ESE and EAE performed very similarly, while ECE once again lagged behind.

Target Language	Mean Edit Distance %	Edit Distance % STD
Spanish	12.2%	15.4%
Afrikaans	12.5%	15.6%
Chinese (simplified)	27.0%	24.0%

[4] https://readable.com/blog/the-flesch-reading-ease-and-flesch-kincaid-grade-level/

[5] http://people.cs.pitt.edu/~kirk/cs1501/Pruhs/Spring2006/assignments/editdistance/Levenshtein%20Distance.htm





In addition to comparing each backtranslation target language based on average distance from source phrase, we also looked at how many (and which) phrases were backtranslated with no errors. The ESE backtranslation perfectly matched the original phrase on 36.5% of the phrases, the EAE on 36.6% of the phrases, and the ECE on only 12.4% of phrases.

Of the perfectly translated phrases, 2156 of the phrases were common between ESE and EAE, 768 common between EAE and ECE, 857 common between ESE and ECE, and 597 common between all three. The table to the right details summary statistics of subsets of perfectly backtranslated phrases, all non-perfectly backtranslated phrases, and the worst 20% of backtranslated phrases.

	Perfectly Backtranslated Phrases			
Target Language	**Count**	**Avg Word Count**	**Readability**	**Grade Level**
Spanish	3855	8.23	63.16	6.50
Afrikaans	3659	8.30	63.75	6.43
Chinese (simplified)	1236	8.23	62.91	6.54
	Imperfectly Backtranslated Phrases			
Target Language	**Count**	**Avg Word Count**	**Readability**	**Grade Level**
Spanish	6145	8.47	65.72	6.20
Afrikaans	6341	8.42	65.30	6.25
Chinese (simplified)	8764	8.39	64.99	6.29
	Worst 20% of Backtranslated Phrases			
Target Language	**Count**	**Avg Word Count**	**Readability**	**Grade Level**
Spanish	2005	8.45	64.25	6.04
Afrikaans	2018	8.30	64.06	6.39
Chinese (simplified)	1985	8.27	63.37	6.48

Although edit distance was our primary method of evaluation, we also took a brief look at the percentage of words from the original phrase that were present in the backtranslated phrase, regardless of the position. Similar to previous results, ESE and EAE performed very similar, with the each retaining 87.7% and 86.4% of words, respectively. The ECE backtranslation only retained 72.8% of words from original phrase in the backtranslated phrase.

7 Analysis of Results

Counterintuitively, the results presented in the previous section suggest that word count, readability, and grade level have no significant effect on the accuracy of the backtranslation. What appears to have the more significant effect is the type of symbol set used. For both ESE and EAE, the two better performing backtranslations, the set of alphabetic characters in source and target language are extremely similar, if not identical. Chinese, however, is a character based language, and thus, shares no common symbols with English. Our theory is that this lack of similarity between source and target language led to the large gap in performance between the backtranslations.

The comparison of edit distance and retained words shows that, although the backtranslations preserve a high rate of the words in the original phrase, the order of the specific characters, and therefore words, is changing. Without manually inspecting data samples it is difficult to determine whether or not meaning is retained when words appear in different order.

Unfortunately, given time and resource limitations, we were unable to extensively inspect the individual subsets of phrases (perfectly backtranslated, imperfectly backtranslated, and worst 20% of backtranslations). Minor inspection of the ESE backtranslation showed that a common error was changing genders between original and backtranslated phrase. This led us to believe that, while the model is doing well to backtranslate the semantic meaning of

the phrases, less vital details, such as gender, are resulting in mistakes.

8 Conclusion

In our eyes, evaluating backtranslation by edit distance % is a useful and effective way to quantify the quality of translations done by NMT. Initial research shows that phrase complexity (in terms of length, readability, and grade level) have little to no impact on the performance of the backtranslations. In comparison, shared attributes in terms of language representation symbols appear to be the leading factor in backtranslation success.

Future research can look more extensively into the specific phrases that comprise each of the above defined subsets. Moreover, additional language combinations should be explored, both extending upon our choices of language characteristics (high/low resource, character based, similar alphabet/characters) as well as extending into other language type combinations such as comparing two character based language, two low resource language, or any other myriad possible combinations.

Additionally, analysis should be done regarding the accuracy of the forward and reverse translations. Since a backtranslation is affected by both the conversion from source to target and from target back to source, it will be affected differently in each step. Analyzing individual directions will lead to more answers regarding where in the process errors are occurring most often.