MUSS: Multilingual Unsupervised Sentence Simplification by Mining Paraphrases
Machine Translation Course

1st Semester of 2021-2022

Dumitrascu Claudiu Cristian
claudiu.dumitrascu@s.unibuc.ro

Abstract

Any text can be written in a manner which would look like very fancy and hard to grasp meaning of any word. Let's take as an example a full course of Machine Translation course which would be readed by an toddler, a lot of definitions and lots of complicated words would make his task to understand very hard and nearly impossible. A great help would be an helper to do a text simplification on a fancy text to be transformed in a very easy to grasp text which anyone would be able to do. Using MUSS, we can work in any domain to do paraphrase mining and do unsupervised pretraining at sentence level to achieve text simplification, removing the need for full and complex datasets.

1 Introduction

Sentence simplification is the task of making a sentence easier to read and understand by reducing its lexical and syntactic complexity, while retaining most of its original meaning. The appliance of text simplification can greatly help reduce the difficulty of reading and understanding, having a lot of societal applications for people with cognitive disabilities.

2 Method

Sequence Extraction

is made by doing sentence tokenization with a maximum length of 300 from CCNet, an extraction of Common Crawl (an open source snapshot of the web) that has been split into different languages using fasttext language identification

Creating a Sequence Index Using Embeddings

To automatically mine paraphrase corpora, we first compute n-dimensional embeddings for each extracted sequence using LASER, which provides joint multilingual sentence embeddings in 93 languages that have been successfully applied to the task of bilingual bitext mining, also this can be used for monolingual paraphrase datasets.

Mining Paraphrases

Computed for each language and indexed for nearest neighbor search, in which is applied a query to extract the top-8 nearest neighboor using the upper bound L2 distance, to filter low similarity and then remove identical sequences by using Levenshtein Distance.

Simplyfing with ACCESS

ACCESS is a method to make any seq2seq model controllable by conditioning on simplification-specific control tokens. Where is applied with an seq2seq pretrained transformer models based on the BART. At training time, the model is provided with control tokens that give oracle information on the target sequence, such as the amount of compression of the target sequence relative to the source sequence (length control). For example, when the target sequence is 80the length of the original sequence, we provide the NumChars 80 percent control token. At inference time is controled by selecting a given target control value. Using the original Levenshtein similarity control to only consider replace operations but otherwise use the same.The controls used are therefore character length ratio, replace-only Levenshtein similarity, aggregated word frequency ratio, and dependency tree depth ratio. controls

Selecting Control Values at Inference

Once the model has been trained with oracle controls, we can adjust the control tokens to obtain the desired type of simplifications. Indeed, sentence simplification often depends on the context and target audience

Leveraging Unsupervised Pretraining

Combining the controllable models with unsupervised pretraining to further extend the approach to

text simplification. For English, is finetuning the pretrained generative model BART on the newly created training corpora. BART is a pretrained sequence-to-sequence model that can be seen as a generalization of other recent pretrained models such as BERT. For non-English, the model used is MBART for generalization, which was pretrained on 25 languages.

Pivot Datasets

For datasets in which the datasets for simplification is not available, a pivot model is employed by translating from non-english to english, simplified then translated back to non-english.

Metric

In a normal machine translation task, the BLEU score is employed to score the peformance of an translation. For text simplification the SARI is a metric used for evaluating automatic text simplification systems. The metric compares the predicted simplified sentences against the reference and the source sentences. It explicitly measures the goodness of words that are added, deleted and kept by the system. $Sari = (F1_add + F1_keep + P_del)/3$ where $F1_add$: n-gram F1 score for add operation $FF1_keep$F: n-gram F1 score for keep operation FP_delF: n-gram precision score for delete operation n = 4. Examples
INPUT: About 95 species are currently accepted .
REF-1: About 95 species are currently known .
REF-2: About 95 species are now accepted .
REF-3: 95 species are now accepted .
OUTPUT-1: About 95 you now get in .
OUTPUT-2: About 95 species are now agreed .
OUTPUT-3: About 95 species are currently agreed.

The corresponding SARI scores of these three toy outputs are 0.2683, 0.7594, 0.5890, which match with intuitions about their quality. To put it in perspective, the BLEU scores are 0.1562, 0.6435, 0.6435 respectively. BLEU fails to distinguish between OUTPUT-2 and OUTPUT-3 because matching any one of references is credited the same.

2.1 References

This paper was written based on https://arxiv.org/abs/2005.00352, all the work belongs to the authors of the paper.

Conclusion

In this paper the MUSE model is presented and opens a new way for simplified text based on unsupervised methods which also opens a gate for text simplification for other languages that do no have a corpus using a pivot dataset.