MUSS: Multilingual Unsupervised Sentence Simplification by Mining Paraphrases

Presentation

Dumitrascu Claudiu Cristian - 507

What is text simplification

Text simplification is an operation used in natural language processing to modify, enhance, classify or otherwise process an existing corpus of human-readable text in such a way that the grammar and structure of the prose is greatly simplified, while the underlying meaning and information remains the same.

Examples

Original: The mouse was eaten by the cat.
Simplified: The cat ate the mouse.

Original: Saint Petersburg, formerly known as Petrograd and later Leningrad, is the second-largest city in Russia.
Simplified: Saint Petersburg is the second largest city in Russia. It used to be called Leningrad.

What is MUSS

Multilingual Unsupervised Sentence Simplification system that does not require labeled simplification data. MUSS uses a novel approach to sentence simplification that trains strong models using sentence-level paraphrase data instead of proper simplification data.

Which provides us with the ability to create a text simplification model on any language which contains an machine translation model to use as an pivot for the dataset.

Why MUSS

- Unsupervised methods removes the need to complex labeled data

- Achieves best results for text simplification on ASSET and TurkCorpus datasets

Text simplification metrics: SARI

Is a lexical simplicity metric that measures “how good” are the words added, deleted and kept by a simplification model. The metric compares the model’s output to multiple simplification references and the original sentence. SARI has shown high correlation with human judgements of simplicity gain

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.

How MUSE works

- Extract sentences from Common Crawl(open source dataset with snapshots of websites)
- Extract sentence embedding using LASER, a joint multilingual sentence embedding in 93 languages
- Extract top 8 nearest neighbours and filtering using Levenshtein distance
- Use ACCESS tokens to control pretrained BART/MBART model to target sequence
- Select control values at inference to obtain desired simplification type.
- Finetune BART/MBART on the new corpus

- LASER (BiLSTM encoder with a shared BPE vocabulary for all languages, coupled to auxiliary decoder trained using parallel corpus)
- ACCESS (Seq2Seq with parametrization training such as for target sentence to control char length ratio, levenshtein distance, word lexical complexity and maximum dependency, to approximate syntactic complexity)
- BART (denoising autoencoder for pretraining sequence-to-sequence models used for machine translation task on our case)

How to apply for romanian

- Extract romanian sentences corpus
- Apply the same steps only by using the MBART model or any english to romanian translation model
- Finetune with the desired ACCESS tokens

Thank you