STAP: Simultaneous Translation And Paraphrase

Dmitry Grigorev, DS2
Vakhitova Alsu, DS1
Ilalova Alina, DS2
Kozlov Pavel, RO

Project Description

Machine translation systems typically produce a single output, but in certain cases, it is desirable to have many possible translations of a given input text. This situation is common with language-learning platforms, where some learning happens via translation-based exercises, and grading is done by comparing learners' responses against a large set of human-curated acceptable translations. We believe the processes of grading and/or manual curation could be vastly improved with richer, multi-output translation and paraphrase systems.

A portion of learning on language-learning platforms happens through translation-based exercises. In this task, we focus on the challenges where users are given a prompt in the language they are learning (English), and type a response in their native language (Russian). The developed system will provide the opportunity for students to learn the language in a more flexible way.

The most efficient translation systems nowadays use transformer-based language models as a core (Google translate, Yandex etc.). However, “pure” transformer models have seq2seq architecture, so they are able to produce only one variation of the correct translation. To overcome it some additional features are used. Research of such features will be a part of our work.

Problem Description

This problem represents supervised learning type, as the dataset contains both input and output data. The learning will take place fully offline and no additional training will be initialized during the work with end-users.

One of comparable problems is to translate to another language (learning another language). We may reuse our experience in terms of applying good translation and paraphrasing in English, German and other languages. The application of our experience depends only on dataset which will be used. Another place to restate this tool is to adapt it to real-time translation to help with communication during travelling abroad.

The best human expertise in this case would be linguists or professional translators, but people who speak both Russian and English are also able to validate that translations are correct and relevant.

The problem can be solved manually just by using internet and trying to translate the text on several different translator websites, but in case if this activity has to be performed regularly it can be time and energy consuming. In addition to that, some different websites can have the same translation for some sentences, so the problem becomes even more difficult to solve.

Performance Measurement

The main scoring metric can be F1 score, with respect to the accepted translations. Therefore, the system should be scored based on how well it can return all human-curated acceptable translations.

The performance measure is aligned with the business objective. However, there might be some limitations or minimum applicable score to achieve the business objective.

The minimum performance is translating and paraphrasing such that it hard to determine was it machine or human translation.

Assumptions

1. For this particular project we are tackling only English-to-Russian translation. However, the techniques might be applicable to other domain.
2. Our aim is not beating current SOTA. We want to create a model that produce just adequate results.

Verification of assumptions:

1. We made that assumption because, these are the only languages all members of the team can speak fluently.
2. We are aware of our time and computational power limits.