EVALUATING THE ROBUSTNESS OF QUESTION-ANSWERING MODELS TO PARAPHRASED QUESTIONS

Paulo Alting von Geusau (p.geusau@gmail.com)
Peter Bloem (vu@peterbloem.nl)

Understanding questions expressed in natural language is a fundamental challenge studied under different applications such as question answering (QA). We explore whether recent state-of-the-art models are capable of recognising two paraphrased questions using unsupervised learning.

The contributions of this paper are threefold:

1. We test the QA models' performance on an existing paraphrased testset (Dev-Para) to evaluate their robustness to question paraphrases.
2. We create a new paraphrased evaluation set (Para-SQuAD) that consists of question pairs from the original SQuAD development set, the question pairs are semantically similar but have a different syntactic structure.
3. We create and visualize useful sentence embeddings of Para-SQuAD by the language models.

GPT-2

Legend:
- Rhine
- Economic_inequality
- Civil_disobedience
- Immune_system
- Packet_switching
- Ctenophora
- Amazon_rainforest
- European_Union_law
- Oxygen
- Nikola_Tesla
- Super_Bowl_50
- Warsaw

BERT

XLNet