FRI Natural language processing course 2023
Project report

University *of Ljubljana*
Faculty *of Computer and Information Science*

Project 3: Paraphrasing sentences

Matjaž Zupančič Muc, Marko Novak, and Klemen Laznik

Abstract

The abstract goes here. The abstract goes here.

Keywords
Sentence paraphrasing,

Advisers: Slavko Žitnik

Introduction

Paraphrasing plays an important role in language understanding tasks, such as question answering, machine translation and semantic parsing. Additionally, it serves as a useful method for data augmentation. The goal of paraphrase generation is to produce alternative versions of a given sentence that may feature different phrasing or structure, yet still accurately convey the original meaning. Creating high-quality paraphrases is a challenging problem in natural language processing (NLP), as it requires a deep understanding of the underlying semantics and syntax of the input sentence. In recent years, there has been a growing interest in using machine learning techniques, particularly transformer models, to automatically generate paraphrases.

Authors of [1] compare translation based paraphrase gathering using human (experts and non-experts), automatic and hybrid techniques. The automatic technique is based on back translation using a neural machine translation (NMT) model. They first translate a source sentence from one language to multiple target languages (i.e pivot languages) and than back to the original language. They found that NMT based paraphrases have a higher diversity compared to paraphrases written by human non experts, but NMT based paraphrases don't reach the adequacy or fluency level provided by expert paraphrases. They found that NMTs corrupt inputs such as slang and typing errors leading to bad paraphrases. They also found that NMTs struggle with negation, i.e NMTs tend to loose or add negation to a sentence. Finally they found that paraphrases generated by pivot languages which are closely related have a higher adequacy and fluency level, while paraphrases generated by pivot languages that are not closely related have higher diversity. They conclude that NMT based paraphrase generation is cheap and diverse, although NMTs produce less fluent outputs post editing could be used to improve the quality with little additional expenditure.

[TODO: Add at least 2 other references;]

Placeholder for third reference [2]

Methods

[TODO: Initial Ideas; Matjaz: We have to talk about this !]

Results

Discussion

Acknowledgments

References

[1] Christian Federmann, Oussama Elachqar, and Chris Quirk. Multilingual whispers: Generating paraphrases with translation. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 17–26, Hong Kong, China, November 2019. Association for Computational Linguistics.

[2] Lingfeng Shen, Lemao Liu, Haiyun Jiang, and Shuming Shi. On the evaluation metrics for paraphrase generation, 2022.