FRI Natural language processing course 2021
Project report

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

Paraphrasing sentences

Uroš Škrjanc, Jaka Mesarec, and Grega Rovšček

Abstract

The abstract goes here. The abstract goes here.

Keywords
Keyword1, Keyword2, Keyword3 ...

Advisors: Slavko Žitnik

Introduction

Paraphrasing sentences is an important aspect of language processing, especially in machine translation, text summarisation, extracting information from text, answering questions and in syntactic and semantic analysis of sentences and parsing. Although various methods have been used in the past for paraphrasing sentences, the use of artificial neural networks has recently been on the rise, as in many other areas of machine learning and artificial intelligence. We intend to approach the task ourselves by means of translation, where, by translating a sentence into an intermediate language and back into the original language, we obtain different sentences which, if translated correctly, should retain the same meaning. Such machine paraphrasing is usually less accurate in terms of similarity of meaning, when compared with human paraphrasing, but it is much faster, cheaper, and usually results in a more varied vocabulary, especially if the intermediate language is from another language group and there is a sufficiently large pool of texts.

There exist multiple solutions to the problem of paraphrasing. An article on paraphrase generation with translation by Federmann et al. [1] reports impressive results. It reports better effectiveness of strong neural machine translation at paraphrase generation compared to humans. This is because it is cheap, adequate and diverse.

Zhou and Bhat [2] reported on great advances in the state-of-the-art results, but labeled them as still not satisfactory enough. In their article, they also discuss potential directions of research that the authors believe are worth studying and could help bring forward better results.

A paper on the topic of evaluation metrics in paraphrase generation by Shen et al.[3] considers different metrics of paraphrase evaluation and investigates their correlation with human judgment. It poses questions about quality and limitations of existing metrics and presents a detailed analysis. It also proposes a new metric, called ParaScore, which, by their findings, significantly outperforms existing metrics.

Prakash et al. [4] conducted a research where they explored the usage of neural machine translation to generate paraphrases via pivoting. Their evaluation of the results demonstrate that their model outperforms a few known methods such as sequence to sequence, bi-directional LSTM models on BLEU, METEOR, and TER.

Existing solutions

Existing applications

Writing assistants: https://instatext.io/

Paraphrasing sentences is used for:

- plagiarism checking
- to avoid plagiarism by changing sentences that sound too similar
- text summarization
- simplifying the text - when the text is too complex or technically demanding, part of the text can be changed to make it more understandable
- machine translation

- to optimise the content of a website (SEO - Search Engine Optimisation) so that it is easier for search engines to find it or rank it higher in search results

Some web applications:

- Prepostseo
  https://www.prepostseo.com/paraphrasing-tool Has the feature to manually add synonyms.
- QuillBot
  https://quillbot.com/
- SpinBot
  https://www.spinbot.com/ Actually redirects to QuillBot
- AI Article Spinner
  https://aiarticlespinner.co/
- Small SEO Tools
  https://smallseotools.com/paraphrasing-tool/ (not free)

Short comparison

Original text: (https://www.bbc.com/news/world-europe-64935449): *Analysts say Bakhmut has little strategic value, but has become a focal point for Russian commanders who have struggled to deliver any positive news to the Kremlin.*

- **Prepostseo - standard:** *Analysts say Bakhmut has little strategic value but has become a flashpoint for Russian commanders struggling to get positive messages across to the Kremlin.*
- **Prepostseo - fluency:** *Analysts say Bakhmut has little strategic value but has become a focal point for Russian commanders who have struggled to relay positive messages to the Kremlin.*
- **QuillBot:** *Despite having minimal strategic worth, according to analysts, Bakhmut has become a focus for Russian commanders who have found it difficult to bring any good news to the Kremlin.*
- **AI Article Spinner standard:** *Analysts say Bakhmut has little strategic value, but he has become a focal point for Russian commanders who have struggled to deliver positive news to the Kremlin.*
- **AI Article Spinner standard:** *Analysts say Bakhmut has little strategic value, but has become a focal point for Russian commanders struggling to convey good news to the Kremlin.*

Methods

First of all, we would take a suitable amount of sentences for paraphrasing from the ccGigafida corpus. For simplicity, we would focus on sentences containing between 5 and 20 words, clean them up, separate them and turn them into lower case. The sentences would then be translated using the Clarin translator, some of them used for learning and some for testing the model.

Once we have a suitable base, we would try to learn one of the autoregressive models for paraphrasing Slovene sentences. In our opinion, an autoregressive model would be better for such a task, as it also takes into account the context of the sentence and does not only translate single words.

Both in the translation of the sentences and in the evaluation of the resulting model, we would try to assess:

1. The distance or similarity between the original and the translated sentence
2. How many words are different between the translated and the original sentence
3. Syntactic correctness of the translated sentence
4. We would also try to manually assess how the translated sentence sounds on a given sample.

When evaluating the resulting model, we would try to add the following metrics:

1. BLEU
2. METEOR
3. ParaScore

BLEU

Basic metric, made for machine translation, measures the similarity between machine and human translated sentences. One of the standard measures. [5]

METEOR

Addresses some of the shortcomings of BLEU. [6]

ParaScore

The assessment takes into account both the semantic similarity between the words and the lexical variety (as similar a meaning as possible and as different words as possible). [3]

CLARIN

Common Language Resources and Technology Infrastructure, Slovenia CLARIN.SI is the Slovenian national consortium of the European research infrastructure network CLARIN. It provides language resources and technologies. We will use it for translation between English and Slovene.
https://www.clarin.si/repository/xmlui/handle/11356/1736
https://github.com/clarinsi/Slovene_NMT

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] Jianing Zhou and Suma Bhat. Paraphrase generation: A survey of the state of the art. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5075–5086, 2021.

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

[4] Aaditya Prakash, Sadid A Hasan, Kathy Lee, Vivek Datla, Ashequl Qadir, Joey Liu, and Oladimeji Farri. Neural paraphrase generation with stacked residual lstm networks. *arXiv preprint arXiv:1610.03098*, 2016.

[5] Wikimedia Commons. Bleu.

[6] Wikimedia Commons. Meteor.