Paraphrase Detection Using Machine Translation and Textual Similarity Algorithms

Dmitry Kravchenko(✉)

Department of Computer Science, Ben-Gurion University of the Negev,
Beer Sheva, Israel
to.dmitry.kravchenko@gmail.com

Abstract. I present experiments on the task of paraphrase detection for Russian text using Machine Translation (MT) into English and applying existing sentence similarity algorithms in English on the translated sentences. But since I use translation engines - my method to detect paraphrases can be applied to any other languages, which translation into English is available on translation engines. Specifically, I consider two tasks: given pair of sentences in Russian – classify them into two (non-paraphrases, paraphrases) or three (non-paraphrases, near-paraphrases, precise-paraphrases) classes. I compare five different well-established sentence similarity methods developed in English and three different Machine Translation engines (Google, Microsoft and Yandex). I perform detailed ablation tests to identify the contribution of each component of the five methods, and identify the best combination of Machine Translation and sentence similarity method, including ensembles, on the Russian Paraphrase data set. My best results on the Russian data set are an Accuracy of 81.4% and F1 score of 78.5% for an ensemble method with the translation using three MT engines (Google, Microsoft and Yandex). This compares favorably with state of the art methods in English on data sets of a similar size which are in the range of Accuracy 80.41% and F1-score of 85.96%. This demonstrates that, with the current level of performance of public MT engines, the simple approach of translating/classifying in English has become a feasible strategy to address the task. I perform detailed error analysis to indicate potential for further improvements.

Keywords: Paraphrase detection · Semantic similarity algorithms
Machine translation · Supervised classification

1 Introduction

1.1 Motivation

Paraphrase identification is useful in many natural language applications such as search engines (to calculate relevance of one sentence to the other), in plagiarism detection systems, authorship identification, patents and copyright detection systems, question-answering bots (to compute the semantic similarity between a

A. Filchenkov et al. (Eds.): AINL 2017, CCIS 789, pp. 277–292, 2018.
https://doi.org/10.1007/978-3-319-71746-3_22