Semantic Similarity Identification for Short Text Fragments

Viktoriia Chuiko[0000-0002-4393-3260] and Nina Khairova[0000-0002-9826-0286]

National Technical University «Kharkiv Polytechnic Institute»,
Kyrpychova str., 2, 61002, Kharkiv, Ukraine

viktoriia.chuiko@gmail.com, nina_khajrova@yahoo.com

Abstract. The paper contains review of the existing methods for semantic similarity identification, such as methods based on the distance between concepts and methods based on lexical intersection. We proposed a method for measuring the semantic similarity of short text fragment, i.e. two sentences. Also, we created corpus of mass-media text. It contains articles of Kharkiv news, that were sorted by their source and date. Then we annotated texts. We defined semantic similarity of sentences manually. In this way, we created learning corpus for our future system.

Keywords: semantic similarity, short text fragments, corpus of mass-media text, automatic identification.

The goal of the research is to develop a method for measuring the semantic similarity of short text fragment and to create a program of automatic semantic similarity identification. Existing methods of evaluating similarity have focused mainly on either large documents or individual words [1]. In this work, we focus on computing the similarity between two sentences using corpora of mass-media texts.

The semantic similarity is a quantitative measure that shows how two concepts are close (that is, related or similar) to each other. There are many other connections between words (other than synonymy), in the presence of which one can speak of semantic closeness.

Most of nowadays studies are based on the fact that two sentences that have most of the same words are likely to paraphrase each other. Thereby, we can say that they have semantic similarity [2]. The problem lies in the fact that there are many sentences that convey the same information, but have little resemblance to the surface.

The task of quantitative evaluation of semantic similarity is deeply examined, and now there are many solutions based on different algorithms. Systems, such as Texterra, Semanticus, S-Space, Semantic Vectors and their counterparts, use a semantic distance algorithm. The methods of this group are based on finding the distance between two concepts in a semantic network (for example, WordNet or EuroWordNet). So, between two concepts lies the shortest path and, on its basis, determines the semantic closeness between the words. One of the first such measures

was offered by Reznik [3]. The obvious drawback of it was that for some concepts the nesting of the classes to which they belong is greater than for others. To solve this problem, Leacock and Chodorow [4] proposed a method that normalizes the length of the path, considering the depth of the general hierarchy.

Another group is the methods based on lexical intersection. The first algorithm of this type was developed by Lesk [5]. He constructed an algorithm, that basically has the assumption, that related concepts are defined or explained by the same words. Lesk used this approach to solve the problem of finding the correct meaning of a word in some context. The disadvantage of this approach is that articles in common vocabularies are rather short, and may therefore badly reflect the semantic similarity of some words.

There are also systems that offer the use of semantic or parser analyzers to construct the corresponding trees of two comparable sentences, with further analysis and comparison of these trees. An example of such a system can be the MaltParser utility.

All analyzed methods evaluate semantic similarity only for words, but not for larger part of sentences. So, a task of semantic similarity identification for short text fragment (i.e., sentences) is still relevant.

First of all, to make such evaluation we need to have a learning corpus. There are a lot of different corpuses for English language, for example Microsoft Paraphrase Corpus [6]. But only few of them are for Ukrainian and Russian languages.

In this way, the first step of the research was to create corpora of mass-media texts. For this task we choose some sites of Kharkiv news. Then, we automatically extracted news articles content and sorted them according to source.

The next point was to annotate texts. We defined semantic similarity of sentences manually. In this way, we created learning corpus for our future system.

The third stage of the research involves development of method for measuring the semantic similarity of short text fragments using corpora created on previous step. After deep analyze of the existed methods with all their advantages and disadvantages, we propose the following algorithm:

1. From all texts select sentences, which have from 1 to 3 common words.
2. Select two of them and evaluate.
3. Analyze if selected sentences have synonyms.
4. Review the word order of each sentence, using information received on previous stages.
5. Determine a semantic similarity of these two sentences.
6. Repeat steps 2-5 until all texts will be analyzed.

The last stage of the research is to identify the semantic similarity between sentences in the texts of mass-media using developed method.

References

1. Anisimov, A., Glybovets,M., Marchenki,O., Kysenko, V.: The method for calculating of semantic closeness of natural language words meanings (2011)
2. Islam, A., Inkpen, D.: Semantic Similarity of Short Texts. University of information technology & sciences (2009)
3. Resnik, P.: Using information content to evaluate semantic similarity in a taxonomy. In:

International Joint Conference for Artifcial Intelligence (IJCAI-95),pp. 448-453(1995)
4. Leacock, C., Chodorow, M., Miller, G. A.: Using corpus statistics and wordnet relations for sense identifcation. Computational Linguistics, 24(1),pp. 147-165(1998)
5. Lesk, M.: Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In:SIGDOC'86: Proceedings of the 5th annual international conference on Systems documentation, pp. 24-26. New York, NY, USA. ACM(1986)
6. Microsoft Research Paraphrase Corpus Homepage,https://www.microsoft.com/en-us/download/details.aspx?id=52398, last accessed 2019/03/01.