Semantic Similarity Detection in a Single Text

Anna Polityuk, Olena Orobinska[0000-0001-8396-4136]
National Technical University "Kharkiv Polytechnic Institute",
Kyrpychova str., 2, 61002, Kharkiv, Ukraine

littlephilologist0@gmail.com, orobinska@khpi.edu.ua

Abstract. To solve many of the problems of automatic natural language processing, it is often necessary to have a dictionary of synonymous terms. To simplify its using is objective of our experiment. We propose the method that realize the lexical approach and provide the detecting all synonyms in a single text and visualize the results directly in the text. The results depend on the completeness of the lexical source. But it is a bottleneck problem of most of thesaurus.

Keywords: Semantic similarity, Synonymy, Semantic Closeness, Proximity Metric, Thesaurus

Semantic similarity is similar to synonymy. In turn, semantic closeness is not limited by the relation of synonymy. A measure of semantic closeness is a quantitative quantity that shows how close the two concepts are to each other. There are many other links between words (other than synonymy), where one can speak of semantic closeness. The reason for most of the differences in the semantic closeness researchers was to determine the degree of "closeness" or "identity" of the meanings of the synonym words.

Semantic similarity is a metric defined on a set of documents or terms, where the idea of the distance between them is based on the similarity of their meaning or semantic content as opposed to the similarity that can be evaluated in terms of their syntactic representation. The semantic closeness, therefore, will depend directly on the number of coexisting nuclei of textual concepts and on the distance between these nuclei in the thesaurus graph.

Computer dictionaries are often shaped by the conversion of plain text dictionaries, but often they require much more complicated and painstaking work. Collections and corpora of texts can be the starting material for obtaining the necessary linguistic information.

We propose the method that allows detecting the series of synonyms in an arbitrary selected text. In this case the statistical approaches don't give a trustworthy result because of single texts usually are not sufficiently long. An acceptable solution would be to use thesaurus dedicated to the information retrieval. But thesaurus can be applied only after additional text processing.

COLINS'2020, Volume II: Workshop. Lviv, Ukraine, April 23-24, 2020, ISSN 2523-4013
http://colins.in.ua, online

Our program is realized on the base of web-interface with processing in Python. The modules in JS were written to provide the data exchange.

There are such steps of processing:

- the content of the file is transformed into a list of tokens;
- each token is labeled with corresponding POS-tag;
- only substantives, verbs, adjectives and pronouns are selected and lemmatized to be sent into thesaurus;

To find the synonyms of any word in the text it is sufficient to click this word on the user interface. All its synonyms presented in the text will be highlighted. It provides the quick visualization of the all synonymic groups in the text.

The results of the testing show that the proposed method can be applied to the solution more complex problems, such as text clustering end classification.

In the course of our work we have considered the problem of semantic closeness estimation. We have analyzed the existing methods and approaches for solving this problem. Existing semantic proximity assessment guidelines and principles have also been considered. To evaluate semantic proximity, a thesaurus can be applied but often it must be completed based on subject terminology based on the extraction of definitions to be considered. The semantic closeness of terms is estimated using two proximity metrics and manual domain expert evaluation. We are looking to continue our experiments with different available thesauruses and to different languages.

References

1. Erbs, Nicolai & Gurevych, Iryna & Zesch, Torsten. (2014). Sense and Similarity: A Study of Sense-level Similarity Measures. 30-39. 10.3115/v1/S14-1004.
2. Suleymanov A.Sh. The semantic analysis and indexing. International Advanced Texnologies, Ankara/ Turkey, 2003.
3. Alexander Budanitsky and Graeme Hirst. 2006. Evaluating WordNet-based Measures of Lexical Semantic Relat.