Method for ParaphraseExtractionfrom the News Text Corpus

Illia Manuilov, Svitlana Petrasova[0000-0001-6011-135X]

National Technical University "Kharkiv Polytechnic Institute",
Pushkinska str., 79/2, Kharkiv, Ukraine

banger@ukr.net, svetapetrasova@gmail.com

Abstract. The paper discusses the process of automatic extraction of paraphrases used in rewriting. The researchers propose the method for extracting paraphrases from English news text corpora. The method is based on both the developed syntactic rules to define phrases and synsets to identify synonymous words in the designed text corpus of BBC news. In order to implement the method, Natural Language Toolkit, Universal Dependencies parser and WordNet are used.

Keywords: paraphrase extraction, news text corpus, syntactic rules, synsets, Universal Dependencies,WordNet.

In modern computational linguistics, technologies for identifyingsemantic similarity between linguistic units are widely used. Formally, such a mechanism means the synonymous replacement of elements, extension of the structure (the addition of elements) or shortening of the structure (the omission of elements).

For converting complex text into simpler one andwriting unique texts, the following methods for paraphrasing are used.

1. Transformation of the direct speech into indirect one. This technique allows saving the necessary sense in the text, but at the same time makes information unique for search engines.
2. Reducing the size of the text to simplify it and better understand the content.
3. Text structure processing: moving paragraphs of the text, changing grammatical constructions of sentences which adds not only uniqueness to a new text but a new style of writing without changing its meaning [1].

According toa language level for paraphrasing, the vocabulary, syntactic structure, morphological characteristics of words, their number and order are being changed. In this case, one word can be replaced saving the entire structure or we can change the entire structure retaining lexical units.

There are several ways to paraphrase syntactical units of texts:

- changing the grammatical structure of the sentence, for example, replacing the subject and object;
- replacing words of one part of speechby another, for example, a verb by a noun or adjective;

– extending the structure (addition of elements) and vice versa;
– splitting long sentences into several smaller ones and vice versa;
– replacing synonymous words orphrases (collocations) [2].

The paper proposesthe method for paraphrase extraction from the news text corpus based on the developed syntactic rules [3] to define phrases (collocations) and the use of WordNet [4] to identify synonymous words in the text corpus.

The developed corpus consists ofBBCnews articles, the sport section [5].

For preprocessing(POS-tagging), the NLTK's Python language library tools are suggested to use.

Figure 1 shows the synonymous pairs obtained in WordNet.

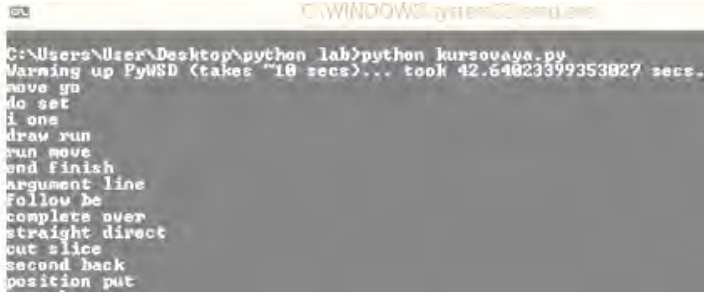

Fig. 1. Synonymous Pairs Extracted from WordNet

For extracting paraphrases, we check the correspondence of the grammatical characteristics of collocates (synonymous words of phrases identified at the previous stage) with the syntactic rules.

Thus, phrases whose grammatical characteristics correspond to the rules are considered to be synonymous.As a result, the proposed method for paraphrase extraction from the news text corpus allows identifying a common information space for topical news.

References

1. Koloiev, A.S.: Rewrite as a new phenomenon in modern journalism. In: SPU Bulletin. Philology, vol. 1, 221-226 (2012)
2. Bolshakov, I.A.: Two methods of synonymous paraphrasing in linguistic steganography. In:Proceedings of the International ConferenceDialogue-2004,http://www.dialog-21.ru/media/2496/bolshakov.pdf, last accessed 2019/02/10.
3. Petrasova, S., Khairova, N., Lewoniewski, W.: Building the semantic similarity model for social network data streams. In:Data Stream Mining & Processing, Proceedings of the 2018 IEEE Second International Conference (DSMP), 21-24 (2018)
4. WordNet: https://wordnet.princeton.edu, last accessed 2019/02/10.
5. BBC, https://www.bbc.com/news,last accessed 2019/02/10.