Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference “Dialogue 2017”

Moscow, May 31—June 3, 2017

AUTOMATIC GENERATION OF VERBATIM AND PARAPHRASED PLAGIARISM CORPUS

Khazov A. V. (hazov@ap-team.ru)
JSC Anti-Plagiat, Moscow, Russia

Kuznetsova M. V. (kuznetsova@ap-team.ru)
JSC Anti-Plagiat, Moscow, Russia; Moscow Institute of Physics and Technology (MIPT), Moscow, Russia

There are two most common approaches to enhancing text originality: including text blocks from various sources and altering the target text itself. In the first case, the text blocks are copied from external sources to the original text, whereas in the second case the original textual content is changed by means of paraphrasing. Addressing the issue of locating borrowing and detecting plagiarism is one of the topics included into the special track of the 2017 Dialogue Conference and the PlagEvalRus contest. In this paper, we develop a method for automatic generation of corpus for plagiarism detection for the abovementioned contest. We also propose a corpus generation algorithm using verbatim and paraphrased text block insertions from source texts to target documents. The research regards the options for machine generation of texts containing paraphrased blocks and evaluates the paraphrasing quality as well.

Keywords: paraphrasing, natural language processing, plagiarism detection, corpus generation

АВТОМАТИЧЕСКАЯ ГЕНЕРАЦИЯ КОРПУСА ДОСЛОВНЫХ И ПЕРЕФРАЗИРОВАННЫХ ЗАИМСТВОВАНИЙ

Хазов А. В. (hazov@ap-team.ru)
ЗАО «Анти-Плагиат», Москва, Россия

Кузнецова М. В. (kuznetsova@ap-team.ru)
ЗАО «Анти-Плагиат», Москва, Россия; Московский физико-технический институт (государственный университет), Москва, Россия

1. Introduction

Verbatim and paraphrased plagiarism are current issues for research [2, 15] and actual problems in education, SEO. Paraphrasing here stands for change of textual contents with retaining its meaningfulness and semantic similarity with the primary source. The development and widespread use of the Internet and IT made the cases of incorrect information borrowing via automatic paraphrasing tools possible.

Verbatim plagiarism is currently being widely reported [13]. The issue of paraphrasing got the most attention and addressing during the PAN'16 conference on "Author Obfuscation" workshop. Participants of this competition proposed three different approaches to paraphrasing. The first approach consisted in replacing the most frequent words with their synonyms [8]. This approach is benevolent as it allows maintaining the original meaning of the text in the majority of cases. The disadvantage of this method is a measurably small amount of modifications in the original text. The second team proposed using the translation-based approach for the text obfuscation [6]. From the source language (English), the text is being translated into an intermediate language before it gets eventually translated back into English ($English \rightarrow German \rightarrow French \rightarrow English$ chain). The main advantage of this method is a strong modification of the original text. The most significant draw of this approach is a considerably weak semantic cohesion of the derived text and a vast amount of foreign words. The third approach [9] combines substantial context alteration of the target text with preserving the original sense. The authors used splitting and merging sentences, synonymising and removing stop words, spelling corrections and common mistake insertion, working with symbols and numbers and other. This algorithm gave the best result by the sum of metrics used in the contest [7].

In this paper, we offer an algorithm of automatic generation of the corpus of copypasted and paraphrased plagiarism in the Russian language. From the sources [3–5] and the discussed above papers possible paraphrase types were selected. We proposed applying cosine similarity, integrity and coherence as metrics to evaluate the quality of the paraphrased text.

2. Plagiarism generation

There is a set of candidate documents in Russian $T = \{t_i\}_{i=1}^{n}$ and source documents $W = \{w_j\}_{j=1}^{m}$, $W \cap T = \emptyset$ for corpus generation. For each candidate t_i we randomly choose from one to K sources $\{w_k\}_{k=1}^{K} \subset W$. Text of t_i candidate is split by sentences $C_i = \{c_{il}\}_{l=1}^{r}$. Random subset of sentences $\{c_{il}\}_{l=1}^{z} \subset C_i$ is chosen from them. Each of c_{ij} is changed by one or more randomly chosen consecutive source sentences

$$c_{il} \rightarrow \left(s_{w_k}, s_{w_{k+1}}, \dots, s_{w_{k+v}}\right), v \in \{1, \dots, V\}$$

where s_{w_k} — sentence from w_k source.

Paraphrasing corpus generation uses the algorithm as described above with one main difference. One of paraphrase types $F = \{f_q\}$ is applied for each source sentence s_{w_k}: $s'_{w_k} = f_q(s_{w_k})$.

3. Paraphrasing types

We allocated and released the following paraphrase types:

- synonym replacement;
- adding and removing synonym chains;
- abbreviation and disabbreviation;
- adding and removing diminutives;
- singular/plural replacement.

3.1. Preprocessing

As candidate texts and source texts we regarded documents containing not less than 30 sentences. We used the NLTK [11] toolbox especially the *nltk.tokenzie* package to extract separate sentences from the original text. Sentences shorter than 10 symbols or containing less than two words were excluded from the analysis.

3.2. Synonym replacement

We used Yet Another RussNet (YARN) [16] as the thesaurus. From the sentences, we allocated random subsamples of word sequences. For each subsample, we checked if it existed in the thesaurus. If it did, we changed it for random synonym sequence. It is possible to replace both single words and word subsequences. Example:

(1) ***Скорее всего**, это пациент клиники для душевнобольных.*
***Должно быть**, это пациент клиники для душевнобольных.*

3.3. Adding and removing synonym chains

To accomplish paraphrasing of this type we also used the YARN thesaurus.

In case of adding synonyms chains, we obtained the list of synonyms for each random word, including the original word itself. From the list d words were selected ($d \in \{1, ..., D\}, 2 \leq D \leq 4$) and the result was written in random order, separated by commas (separated by conjunction "и/или" (and/or) between the penultimate and the last word) instead of the original word.

In case of removing synonyms chains, the pairs of words located on either side of the conjunction "и/или" were looked up in sentence. Grammeme set (POS, case, number, tense, person) were defined with morph analyser [10] for each word in each pair. If the grammemes sets were equal, one of the two words was removed.

Example:

(2) *Малому следовало **отдать** должное.*
*Малому следовало **вернуть, возвратить или отдать** должное.*

3.4. Abbreviation and Disabbreviation

We used *The newest abbreviations dictionary of the Russian language* [12] for generating this type of paraphrasing. The list of abbreviations is a dictionary, where the key is the abbreviation, and its value is a list of its possible decoding word combinations.

Algorithm of the abbreviation process:

- extracting all first letters from the sentence words;
- getting all possible n-grams ($n \in \{2,3,4,5\}$) from consecutive letters;
- removing n-grams not present in the abbreviation dictionary;
- checking if normalized words related to their n-grams are in the list of abbreviation
- decoding;
- if the word combination is present in the list, it is replaced with its abbreviation.

The disabbreviation process is significantly simpler. For disabbreviation, we checked upon the existence of each word in the abbreviation dictionary keys. If word had a match in the dictionary keys, it was replaced with its random decoding.

Example:

(3) *Положение о взыскании налогов и неналоговых платежей, утвержденных постановлением* ***ЦИК*** *и СНК СССР от 17 сентября 1932 года.*
Положение о взыскании налогов и неналоговых платежей, утвержденных постановлением ***Центральная избирательная комиссия*** *и СНК СССР от 17 сентября 1932 года.*

3.5. Adding and removing diminutives

For execution of this paraphrase type we used the Russian nouns diminutive suffix list [14]: *-оньк-, -еньк-, -иньк-, -инк-, -ичк-, -очк-,* etc.

During the process of adding, random diminutive suffix was being inserted into the original word one or two characters prior to its ending. Through morph analyser the following conditions were checked:

- both target and result words are lexicalized;
- the grammeme sets of the two words are identical.

If both conditions were met, the original word was replaced with the result word.

Removing the diminutive suffix was the same up to the direction: if suffix was present in the word, it was removed using a regular expression. The same two conditions were being checked for both the original and the result words using pymorphy2.

Example:

(4) *Вчера утром Карл-Хайнц притащил три* ***бутылки*** *вина и галеты.*
Вчера утром Карл-Хайнц притащил три ***бутыли*** *вина и галеты.*

3.6. Singular/Plural form replacement

From each sentence we choose a random word and determine its part of speech. If it was a noun, an adjective, a verb, a participle, a pronoun, or an ordinal numeral, its form was possible to determine. If the word had no grammatical number, it was

skipped. If it was singular, via the morph analyser it is changed to plural one and vice versa. One by one, all the words to the left and to the right from the original word were checked this way. If a word's number was the same as the original word number, we changed it. If the word we were checking had no number, we skipped it. If the current word had the opposite number, we stopped processing in the current direction (left or right from the start word).

Example:

(5) ***В конце приведено графическое приложение*** *в формате А3 "Геоизотермы западной части Ново-Грозненского месторождения".*
В концах приведены графические приложения *в формате А3 "Геоизотермы западной части Ново-Грозненского месторождения".*

4. The corpus generation

4.1. Matching content generation

Both verbatim and paraphrased extracts were generated using the data provided by the contest [1] organizers. The source collection includes 5,707,798 plain text files in the UTF-8 encoding. It was gathered using the following sources: Russian Wikipedia.org, students' dissertations and reports, Internet report collections and scientific articles from CyberLeninka. For each target text one to $K = 5$ source texts were selected. We took not more than Z sentences for the replacement, $Z \leq 0.3 * total$, where $total$ — the total sentence amount in the target text. Thus, the Z coefficient was selected individually for each target document. Each target sentence was replaced with not more than $T = 10$ consecutive sentences from the source.

4.2. Paraphrasing quality metrics

The adapted PAN assessed paraphrasing quality [7] task metrics:

- cosine similarity between the original sentence and the paraphrased sentence;
- integrity — preservation of the original sentence's sense in the paraphrased, peer reviewed using a three-point scale, where 0 is the lowest value of metrics, 2 is the highest;
- coherence — correct morphological structure of the paraphrased sentence, measured using a three-point scale of expert assessment as well.

Table 1. Integrity measurement examples

0	Sense lost	*В основе такой деятельности лежит экспертный* ***процесс*** *реализации специальных знаний.* *В основе такой деятельности лежит экспертный* ***рассказ*** *реализации специальных знаний.*
1	Sense lost partially	*Поэтому, по* ***мере*** *становления рыночных отношений в системе здравоохранения, ...* *Поэтому, по* ***мерочке*** *становления рыночных отношений в системе здравоохранения, ...*

2	Sense preserved	*Сокращение **посевных площадей** было обусловлено рядом причин.* *Сокращение **посевной площади** было обусловлено рядом причин.*

Table 2. Coherence measurement examples

0	More than one morphological mistake	*С **момента прихода** туда у меня была только одна мысль.* *С **моментов приходов** туда у меня была только одна мысль.*
1	One morphological mistake	*Однако чаще всего наблюдалось повышение **ЧСС**, ...* *Однако чаще всего наблюдалось повышение **частота сердечных сокращений**, ...*
2	No morphological mistakes	*В **структуре системы** выделяются ядро системы и модули.* *В **структурах систем** выделяются ядро системы и модули.*

Morphological mistake here is mistake in the usage of the noun's case, number, tense or person in the paraphrased sentence.

We regarded 20 random sentences for each paraphrased type and calculated the quality metrics described above. Assessors evaluated integrity and coherence and results were averaged.

Table 3. Evaluation results

Paraphrase type	Cosine similarity	Integrity	Coherence
Synonym replacement	0.916	1.17	1.83
Adding and removing synonym chains	0.929	0.85	1.55
Abbreviation and disabbreviation	0.912	1.32	1.43
Adding and removing diminutive forms	0.942	1.07	1.87
Singular/plural form replacement	0.734	1.78	1.47

4.3. Paraphrasing mistakes analysis

Most of the integrity errors were related to the polysemy in Russian language, when one word or word combination can be replaced with another, which is not appropriately synonymous in the current context:

(6) *Почему **социальный** заказ на философию экономики сформировался в конце XX века.*
*Почему **общительный** заказ на философию экономики сформировался в конце XX века.*

Various transcripts of the same abbreviation also caused errors:

(7) *В контексте политических и экономических противоречий, существующих между странами* ***СНГ****.*
В контексте политических и экономических противоречий, существующих между странами ***Сургутнефтегаз****.*

Incorrect diminutive replacement:

(8) *Так, в деятельности Государственной Думы Федерального Собрания Российской Федерации сосредотачивается и используется огромное количество информационных* ***потоков****.*
Так, в деятельности Государственной Думы Федерального Собрания Российской Федерации сосредотачивается и используется огромное количество информационных ***потов****.*

Most of the coherence errors appeared in the abbrevation replacement:

(9) *Азбукин был утвержден* ***ВАК*** *в ученой степени доктора медицинских наук и звании профессора.*
Азбукин был утвержден ***Высшая аттестационная комиссия*** *в ученой степени доктора медицинских наук и звании профессора.*

As well as in the singular/plural replacement:

(10) *Соблюдение прав и свобод других лиц входит в* ***общую систему*** *оснований, ограничивающих злоупотребление правами и свободами.*
Соблюдение прав и свобод других лиц входит в ***общих системы*** *оснований, ограничивающих злоупотребление правами и свободами.*

5. Conclusion

In this paper, we proposed an algorithm of generation of a corpus for verbatim and borrowing plagiarism. We conducted an article review on the possible types of paraphrasing. A list of the types of paraphrasing that are suitable for automatic generation in the Russian language was prepared and tested. We also implemented the paraphrase generation algorithm and developed a quality metrics system. We generated the Russian corpus of borrowing options for the competition at the Dialogue 2017 conference. In the future, we plan to make improvements in the proposed paraphrasing algorithm to increase the integrity and coherence of the generated texts.

References

1. 23-rd International Conference on Computer Linguistics and Intellectual Technologies Dialogue 2017, available at: http://www.dialog-21.ru/.
2. *Ion Androutsopoulos, Prodromos Malakasiotis* (2010), A Survey of Paraphrasing and Textual Entailment Methods, Journal of Artificial Intelligence Research, Vol.38, pp.135–187, available at: https://www.jair.org/media/2985/live-2985-5001-jair.pdf.

3. *Rahul Bhagat, Eduard Hovy: What Is a Paraphrase?* (2013), Computational Linguistics, Vol. 39, pp. 463–472, available at: http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00166.
4. *I. Bolshakov* (2004), Two methods of synonymic paraphrasing in linguistic stenography [Dva metoda sinonimicheskogo perefrazirovanija v lingvisticheskoj steganografii], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2004" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2004"], available at: http://www.dialog-21.ru/media/2496/bolshakov.pdf.
5. *Diana Hasibullina* (2016), Lexical and grammar transformation of phraseological units based on the materials of foreign press [Leksiko-grammaticheskie transformacii frazeologicheskih edinic na materiale zarubezhnoj pressy], Philological Sciences. Issues of Theory and Practice [Filologicheskie nauki. Voprosy teorii i praktiki], Vol. 55, pp. 190–192.
6. *Yashwant Keswani, Harsh Trivedi, Parth Mehta, Prasenjit Majumder* (2016), Author Masking through Translation, CLEF2016 Working Notes, available at: http://ceur-ws.org/Vol-1609/16090890.pdf.
7. *Matthias Liebeck, Pashutan Modaresi, and Stefan Conrad* (2016), Evaluating Safety, Soundness and Sensibleness of Obfuscation Systems, CLEF2016 Working Notes, available at: http://ceur-ws.org/Vol-1609/16090920.pdf.
8. *Muharram Mansoorizadeh, Taher Rahgooy, Mohammad Aminiyan, Mahdy Eskandari* (2016), Author obfuscation using WordNet and language models, CLEF2016 Working Notes, available at: http://ceur-ws.org/Vol-1609/16090939.pdf.
9. *Tsvetomila Mihaylova, Georgi Karadjov, Yasen Kiprov, Georgi Georgiev, Ivan Koychev, Preslav Nakov* (2016), SU@PAN'2016: Author Obfuscation, CLEF2016 Working Notes, available at: http://ceur-ws.org/Vol-1609/16090956.pdf.
10. Morph Analyser pymorphy2, available at: http://pymorphy2.readthedocs.io/en/latest/.
11. Natural Language Toolkit, available at: http://www.nltk.org/index.html.
12. Newest Russian Abbreviation Dictionary [Novejshij slovar' abbreviatur russkogo jazyka], available at: http://netler.ru/slovari/abbreviature.htm.
13. *A. V. Nikitov, O. A. Orchakov, Yu. V. Chehovich* (2012), Plagiarism in works of undergraduate and graduate students: problem and methods of counteraction, University Management: Practice and Analysis, Vol. 5, pp. 61–68.
14. Noun Suffixes Database [Suffiksy imen sushhestvitel'nyh], available at: http://tutrus.com/morfemika/suffiksy-imen-sushhestvitelnyx.
15. *Martin Potthast, Matthias Hagen, and Benno Stein* (2016), Author Obfuscation: Attacking the State of the Art in Authorship Verification, CLEF2016 Working Notes, available at: http://ceur-ws.org/Vol-1609/16090716.pdf.
16. Yet Another RussNet, available at: http://russianword.net/.