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

Moscow, May 31—June 3, 2017

THE PARAPLAG: RUSSIAN DATASET FOR PARAPHRASED PLAGIARISM DETECTION

Sochenkov I. V. (sochenkov_iv@rudn.university)[1,2],
Zubarev D. V. (zubarev@isa.ru)[1,2], **Smirnov I. V.** (ivs@isa.ru)[3,1]

[1]RUDN University, Moscow, Russia; [2]Federal Research Center “Computer Science and Control” of Russian Academy of Sciences, Moscow, Russia; [3]Institute for Systems Analysis, Federal Research Center “Computer Science and Control” of Russian Academy of Sciences, Moscow, Russia

The paper presents the ParaPlag: a large text dataset in Russian to evaluate and compare quality metrics of different plagiarism detection approaches that deal with big data. The competition PlagEvalRus-2017 aimed to evaluate plagiarism detection methods uses the ParaPlag as a main dataset for source retrieval and text alignment tasks. The ParaPlag is open and available on the Web. We propose a guide for writers who want to contribute to the ParaPlag and extend it. The analysis of text rewrite techniques used by unscrupulous authors is also presented in our research.

Keywords: paraphrased plagiarism detection, text reuse detection, dataset for plagiarism detection evaluation.

PARAPLAG: КОРПУС ДЛЯ ВЫЯВЛЕНИЯ ПЕРЕФРАЗИРОВАННЫХ ТЕКСТОВЫХ ЗАИМСТВОВАНИЙ НА РУССКОМ ЯЗЫКЕ

Соченков И. В. (sochenkov_iv@rudn.university)[1,2],
Зубарев Д. В. (zubarev@isa.ru)[1,2],
Смирнов И. В. (ivs@isa.ru)[3,1]

[1]Российский университет дружбы народов, Москва, Россия; [2]Федеральный исследовательский центр «Информатика и управление» Российской академии наук, Москва, Россия; [3]Институт системного анализа, Федеральный исследовательский центр «Информатика и управление» Российской академии наук, Москва, Россия

Ключевые слова: выявление перефразированных заимствований, текстовые заимствования, оценка качества методов выявления текстовых заимствований

1. Introduction

Modern information technologies have given to unscrupulous authors a simple and effective tool to usurp someone else's results with minimal effort. Modern plagiarism detection systems (PDS), such as TurnitIn[1], Antiplagiat[2] and others detect "copy and paste" plagiarism in research or student papers with high recall and precision. Therefore unscrupulous authors use various obfuscation techniques (noise bringing) to their text documents without substantial modification of its meaning and content. This obfuscation can be done automatically by disturbing plagiarized text fragments using hidden text, pictures, non-standard symbols, formulas etc. to prevent correct text extraction from documents in common formats (PDF, Microsoft Word). PDS developers considered this problem, so most of such obfuscation techniques can be successfully detected. The other ways to hide the plagiarism from PDS is to use a little more time consuming techniques: paraphrasing the original text and/or translation from another language. In such case, it expands from "text stealing" (improper citation) to wrongful appropriation of thoughts and ideas. Therefore, the task is beyond the detection of improper citations and goes to judgment of novelty and originality of information presented in scientific or student papers. This judgment is a vital part of scientific expertise or student works checking. Obviously, such expertise is impossible without PDS. However, the detection of rewritten paraphrased and/or translated texts is challenging for modern PDS. Therefore, the research of new methods for detection of paraphrased and translated plagiarism on big data sources is an important scientific task related to information retrieval.

The best and complete solution could not be found without well-grounded and representative evaluation of different approaches for the addressed problems. The evaluation of information retrieval methods for modern PDS requires a large dataset containing original and plagiarized texts for training and testing.

This research addresses the task of creating a dataset for plagiarism detection. It focuses on paraphrased text plagiarism that comes from the real world practice. The main research goal is to create the ParaPlag—a large text dataset in Russian to evaluate and compare quality metrics of different plagiarism detection approaches that deal with a big data. The analysis of text rewrite techniques used by unscrupulous authors is also in focus in our research. We also propose a guide for writers who want to contribute to the ParaPlag and extend it. The competition PlagEvalRus-2017 aimed to evaluate quality of PDS uses the ParaPlag as a main dataset for source retrieval and text alignment tasks.

2. Related work

It is very important to have a standardized dataset to evaluate new plagiarism detection methods. Having such dataset, it is possible to get comparable results of the evaluation. The dataset should be large and represent different text reuse techniques. Therefore, one can test a plagiarism detection method in conditions, which are close to the terms of a real world.

---

[1] http://turnitin.com

[2] http://antiplagiat.ru

Actually there are a few open datasets for such evaluation and mostly used is PAN-PC-11 corpus (Potthast et al., 2010). This corpus was used in PAN competition that held yearly since 2009 until 2015 year. The corpus consists of texts in English that were created by borrowing text of books from Gutenberg collection. Reused text was modified automatically and manually. Since text is borrowed randomly from any book, the suspicious documents do not belong to the same topic as sources. This is the main concern related with this corpus and it makes it suitable only for evaluation of the text alignment task.

The paper (Gollub, T. et al., 2012) introduces the TIRA platform as a standard framework for PDS evaluation. It is a playground for text alignment and source retrieval tracks. TIRA contains Webis-TRC-12 (Potthast et al., 2013) that aims to address the aforementioned issue. Each suspicious document from this corpus was created manually and writers should have tried to hide the plagiarism. Web pages from ClueWeb dataset[3] are the sources for manually written essays (Potthast et al., 2013). The writers found the materials for their plagiarized essays using Chat Noir (Potthast et al., 2012) and Indri search engines. Therefore, the source documents are safely hidden among tens of millions of web pages. Thus, this corpus is suitable for source retrieval task on PAN competition. Under the TIRA this task can be solved using the aforementioned search engines as an entry point to the ClueWeb dataset. In common, suspicious passages are queries and search results are candidates for deep analysis. However the real world PDS require their own indexes and special data structures to deal with plagiarism with high efficiency. Therefore, one needs to index about 504 million web pages (the size of the part of ClueWeb in English) to deal with source retrieval and text alignment in real applications.

Other related research includes the Semeval workshop, which has a corpus for methods that estimate Semantic Textual Similarity (Agirre, E., et al., 2015). This task is close to the text alignment for plagiarism detection, but it focuses on more precise alignments between chunks. The corpus contains sentences in English and Spanish (news headlines, image descriptions, answer pairs from a tutorial dialogue system, etc.) but it is relatively small and therefore it could not be used as a dataset for complex plagiarism detection.

The competition on paraphrase detection in Russian texts uses the specially created and extensible open corpus containing sentence pairs (Pronoza, E., et al. 2016). The task follows the standard procedure: the participating method takes a pair of sentences and returns the similarity class as a response. There are three cases: precise paraphrase, near paraphrase and non-paraphrase.

Other research (Burrows, S. et al., 2013) studies some paraphrase techniques (including translation) and discusses the construction of a paraphrase corpus via crowdsourcing. It also gives a brief review for some other datasets mostly containing paraphrases at sentence-level developed in English: (Dolan, W. B., Brockett, C., 2005), (Clough, P. et al., 2002). The research by Madnani and Dorr (Madnani, N., Dorr, B. J., 2010) discusses the automatic generating of phrasal and sentential paraphrases, and gives a review of paraphrasing techniques.

---

[3] http://www.lemurproject.org/clueweb09/index.php

As we have seen, none of the discussed researches presents a dataset for paraphrased plagiarism detection. The standard solution (TIRA) for PDS evaluation does not have a dataset for Russian. There is no study for Russian covering techniques that unscrupulous authors use (can use) during the writing of plagiarized texts to hide the fact of plagiarism. The current paper will study all these aspects.

3. Creating the dataset

3.1. Common considerations

The creation of the ParaPlag was inspired by our own need to evaluate quality of our PDS which implements some original plagiarism detection methods for English and Russian. The PAN CLEF provides a great opportunity to test them but has some significant limitations. Participants need to use two standard search engines, and tasks do not contain texts in Russian. An alternative approach to evaluate the quality of plagiarism detection for texts in Russian is to use the available results of plagiarism investigations done by experts based on Russian Ph.D. theses repository—Dissernet[4]. However, in most cases each document from Dissernet contains text reuse from a few sources as a pure "copy-paste" with minimal changes. So it does not contain any significant paraphrase (or it was not be marked by experts as text reuse).

At the early stage of our research, we have considered approaches for automated generation of paraphrased plagiarized data. However, the automatic paraphrase (even synonymization) of the given text is a quite challenging task if we want to keep the original sense. The synonymization tools are widespread but their automatic usage makes text meaningless and ugly. Therefore, we asked some students to be writers of plagiarized texts. They were motivated to produce non-original texts and hide plagiarism whenever possible. However, our writers are not professional plagiarists in general. They are not also experts in linguistics or information retrieval. Therefore, we provide a guide describing the writing process and set up general requirements.

The results of writings are "essays" on different topics chosen on authors wish and interest. We tried to avoid the duplication of topics so we maintain a registry of topics for essays. By doing this we address the sources duplication problem, which will be discussed later.

Essays were written in Russian using special format (Microsoft Excel sheets), so we can extract a markup and transform it into different tasks related to a PDS evaluation. The file with an essay contains the following fields: number of fragment, filename of source document (empty for original fragments), rewritten fragment (the text of an essay), source fragments (taken from source document), and applied rewrite techniques.

Writers are free to find sources for their topics on the Web and use documents in common formats (plain text, HTML, PDF, Microsoft Word).

According to the guide, our writers should work at the sentence level, so atomic text fragment, which could be reused, is one sentence. The motivation is that each sentence expresses a statement, question, exclamation, request, command or suggestion,

[4] https://www.dissernet.org/

which could be taken from source text and paraphrased. In general, modern PDS perform well in case when unscrupulous authors change the sentence order in non-original text. Therefore, we did not introduce special requirements on sentence ordering. Writers can mix sentences from different sources and sometimes insert original sentences between plagiarized sentences.

To summarize, essays contain original and paraphrased fragments, which are produced by writers with the following rewrite techniques.

3.2. Text rewrite techniques

We consider the following most common techniques, which are often used by authors to modify the original sentences and hide reused text from PDS:

1. DEL—Delete some words (about 20%) of the original sentence;
2. ADD—Add some words (about 20%) into the original sentence;
3. LPR (Light Paraphrase)—for **Essays-1:** Replace some words or phrases of the original sentence with synonyms, reorder clauses, add new words. For **Essays-2:** change word forms (number, case, form and verb tense, etc.) for some words (about 30%) in the original sentence;
4. SHF (shuffling)—Change the order of words or clauses in the original sentence;
5. CCT (concatenation)—Concatenate two or more original sentences into one sentence;
6. SEP or SSP (sentence splitting/separation)—Split the original sentence into two or more sentences (possibly with a change in the order they appear in the text).
7. SYN (synonymizing)—Replace some words or phrases of the original sentence with synonyms (e.g. "sodium chloride"—"salt"), replace abbreviations to their full transcripts, and vice versa, replace the person's name with the name initial, etc.
8. HPR (Heavy Paraphrase)—Complex rewrite of the original sentence, which combines 3–5 or even more aforementioned techniques. This type involves significant changes of the source text by paraphrase using idioms, synonyms for complex structures, a permutation of words or parts of a complex sentence, etc. Usage of this technique produces strongly paraphrased texts. So in some cases even the experts hardly to consider the rewritten text as plagiarized.
9. CPY—Copy the sentence from source and paste it into essay almost with no changes.

3.3. Writing the essays

We have prepared two tasks for our writers. The rules for the first task (**Essays-1**) were the following:

a) Each essay must contain at least *150* sentences (sentences shorter than 3 words are not taken into account);
b) Plagiarized sentences should be taken from at least 5 different source documents;
c) Each sentence must be ether rewritten from source using one of aforementioned techniques (*1–3, 5–6, 8–9*) or must be original. An author should specify an applied technique for each sentence;

d) The ratio of sentences with techniques used to rewrite them and amount of original fragments was limited. The soft limits were set as follows: *original sentences ~10–40%., CPY ~5–30%, DEL ~20–30%, ADD ~15–25%, ~LPR~10–30%, CCT ~5–15%, SEP ~5–15%, HPR ~5–20%.* However, these limits can vary from writer to writer;
e) Techniques 5–6 allow light modification of sentence (addition /deletion of 10–15% of words).

After collecting some data for **Essays-1** task, we have changed the rules and formed the second task (**Essays-2**):

a) Each essay shall contain at least *100* sentences (sentences shorter than 3 words are not taken into account), and *at least 150 sentences from sources should be used*;
b) Plagiarized sentences should be taken from at least 5 different source documents;
c) Each sentence either must be rewritten from source using *several* (more than one!) aforementioned techniques (*1–8*) or must be original. *"Copy-and paste" text reuse is not allowed*. An author should specify all applied techniques for each sentence;
d) The ratio of sentences with techniques used to rewrite them and amount of original fragments was limited as follows: *original sentences ~5–10%., CPY ~0%, HPR ~20–40%. Other technique at least 10% for each type.* However, these limits can vary from writer to writer. *If some fragment was strongly changed, so one cannot clearly define the applied techniques, it is possible to mark this fragment with the HPR type. In other cases, all techniques must mark the considered fragment.*

There is the additional limitation for the both tasks: each writer shall prepare no more than 10 essays.

Using the two tasks for our writers, we have collected the two testing subsets: **Essays-1** and **Essays-2**, which have different characteristics. **Essays-1** contains mostly essays with large amount of "atomic" usage of the paraphrasing techniques. **Essays-2** is a little bit more complex test set with large amount of heavily paraphrased fragments.

We have had very responsible writers but always remember the principle "errare humanum est". Therefore, we developed validating tools to ensure writers understand and fulfill our requirements. Tools automate detection of common errors, so supervision process gets simple. Tools control some characteristics of written essays such as percentage of techniques used, misspells in names of techniques. Tools check that sentences rewritten with DEL have less word, and sentences rewritten with ADD have more words than original. For LPR-sentences, tools control the grammatical form changing. Tools also ensure that source sentences can be found in corresponding source documents, and original sentences are not taken from this sources. Thus, all essays written under the first and second tasks are validated with a help of these tools, and writers usually correct found errors.

Both subsets will be yet another dataset for text alignment with paraphrased plagiarism. To set up a complex task for source retrieval we must hide source documents in large dataset and deal with some issues with it.

3.4. Building the background dataset for source retrieval

Building the background set of documents comprises the two preliminary steps: web crawling and plain text extraction. Both steps were done using Exactus Expert crawling subsystem (Osipov, G., et al., 2016). Documents were crawled from the Web sources: Russian Wikipedia[5], Cyberleninka[6] and Student Essays[7]. We have added sources from written essays to it. After that, a plain text was extracted from all documents and a unique numeric id was assigned to each document. We also provide a mapping from essays to sources into the dataset for training PDS on source retrieval task.

In fact a simple combining a large dataset of documents from the Web and sources from essays can give biased results in source retrieval competition, since there could be (and actually there are) a lot of near-duplicates. Near-duplicates share almost identical content, so if there are near duplicates for some sources of essays, they likely will be found by competitors. However, these findings will be treated as false positives, since they are not in original mapping that comes with essays. PAN source retrieval track deals with this problem using near-duplicate detection. The same problem appears even if source and some other document are not near-duplicates but share some text fragments. Obviously, this could affect the results of PDS evaluation.

We decided to address this problem on the stage of building of our dataset. We have indexed all crawled documents using TextApp: the search and analytical engine—the successor of Exactus Expert (Osipov, G., et al., 2016). After that, we filtered out all near-duplicates to sources, which came from essays. We use the function of TextApp[8] that searches for topically similar documents for a given query document (Suvorov, R. E., Sochenkov, I. V., 2015). It is rather similar to the inverted index based approaches (Ilyinski, S., et al., 2002), (Ageev, M. S., Dobrov, B. V., 2011), but uses not only single words but also noun phrases as features to represent documents. Thus, we are ready to present the first version of ParaPlag: the Russian dataset for paraphrased plagiarism detection.

3.5. The dataset statistics

As of writing, our volunteers continue to work on additional essays of type 2 (**Essays-2**), which will be suitable for future PDS training and testing. However, we are ready to present the current statistics on our dataset (table 1).

The subset **Essays-1** contains 118 documents, whilst the subset **Essays-2** contains 34 documents currently.

Table 2 presents the statistics on distribution of text rewrite techniques used by writers in **Essays-1**. As we have said before, in this subset each fragment is marked with the technique used to rewrite it.

---

[5] https://ru.wikipedia.org

[6] http://cyberleninka.ru

[7] http://studopedia.ru, http://www.bestreferat.ru, http://allbest.ru, http://do.gendocs.ru,

[8] http://demo.textapp.ru/

Table 1. ParaPlag documents statistics[9]

Source	Documents count	Comments
Cyberleninka	1,037,540	Crawled on August, 2016
Russian Wikipedia	1,330,783	Used the official dump on August, 20169
Student Referats	3,325,255	Crawled on November, 2016
Academic texts	12,183	
Sources from essays	2,037	
TOTAL:	**5,707,798**	

Table 2. Distribution of text rewrite techniques in **Essays-1**

Technique	Fragments count
CPY:	1,596
LPR:	2,870
HPR:	1,839
ORIG:	1,956

Technique	Fragments count
DEL:	3,970
ADD:	2,930
CCT:	1,198
SSP:	1,627
TOTAL:	**17,986**

Table 3. Distribution of text rewrite techniques in **Essays-2**

Technique	Fragments count
LPR:	993
HPR:	938
ORIG:	274
DEL:	1,450
ADD:	1,231

Technique	Fragments count
CCT:	490
SSP:	29
SHF:	750
SEP:	366
SYN:	1,508
TOTAL:	**8,029**

Table 4. Most popular combinations of text rewrite techniques in **Essays-2** and their distribution[10]

Techniques	Fragments count
DEL, SYN:	709
ADD, SYN:	669
DEL, ADD:	625
LPR, SYN:	518
LPR, DEL:	487
SHF, SYN:	409

Techniques	Fragments count
LPR, ADD:	400
DEL, SHF:	359
DEL, ADD, SYN:	327
ADD, SHF:	315
HPR, SYN:	303
HPR, ADD:	296

Tables 3 and 4 show the distribution of text rewrite techniques and their combinations used by writers in **Essays-2**. Each fragment of essay in this subset is marked by at least two techniques used to paraphrase it.

[9] https://dumps.wikimedia.org/

[10] The top 12 frequent combinations, which appear at least 99 times

We performed comparison of plagiarized sentences with sentences from sources. Like in (Potthast et al., 2010) we used N-gram vector space model (VSM) where N ranges from 1 to 8 words. We performed following preparations: words were normalized (via pymorphy2 (Korobov, 2015)), stop words were removed (prepositions, conjunctions, participles), N-grams were TF-weighted. The cosine measure was employed to compute similarity between sentences. Figures 1 and 3 show the obtained similarities for **Essays-1** and **Essays-2** respectively. The box plots show the middle 50% of the respective similarity distributions as well as median similarities. The high value of similarity under 1-gram VSM indicates that essays and sources are about the same topic, since they share considerable amount of their vocabulary. The varying decrease of similarity under N-gram VSM (N>2) pinpoints the difference between two collections.

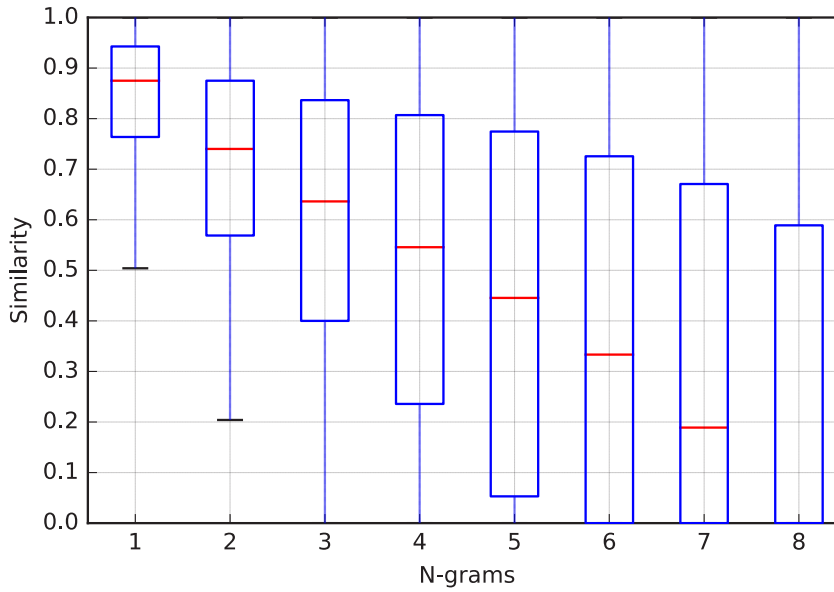

Fig. 1. Distribution of measured similarities for **Essays-1**

Each box plot shows the middle range of the distribution of measured similarities. The top of each box is the 75th percentile, the bottom is the 25th percentile, and the line in a box is a median of distribution. The upper and lower caps show 95th and 5th percentiles respectively.

Essays-1 collection contains copy-paste fragments, therefore its average similarity is relatively high even for large N (such as 6, 7). It means that essays from this collection can be found quite easily with the common methods (so-called shingles) and do not pose serious difficulties for participants of the source retrieval task. However, the collection contains disguised plagiarized text—64% of all sentences, among them: 38% with the light obfuscation techniques (ADD, DEL) and 26% with moderate or heavy obfuscations (LPR, HPR). It makes this collection appropriate for text alignment task. The measured similarity for each obfuscation type is presented in the figure 2.

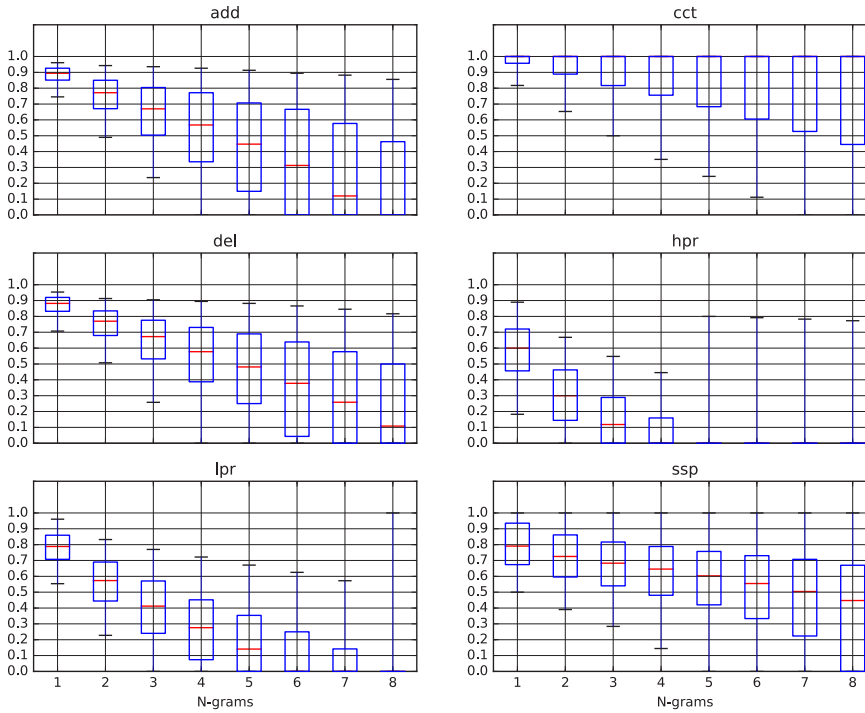

Fig. 2. Distribution of measured similarities per obfuscation type for **Essays-1**

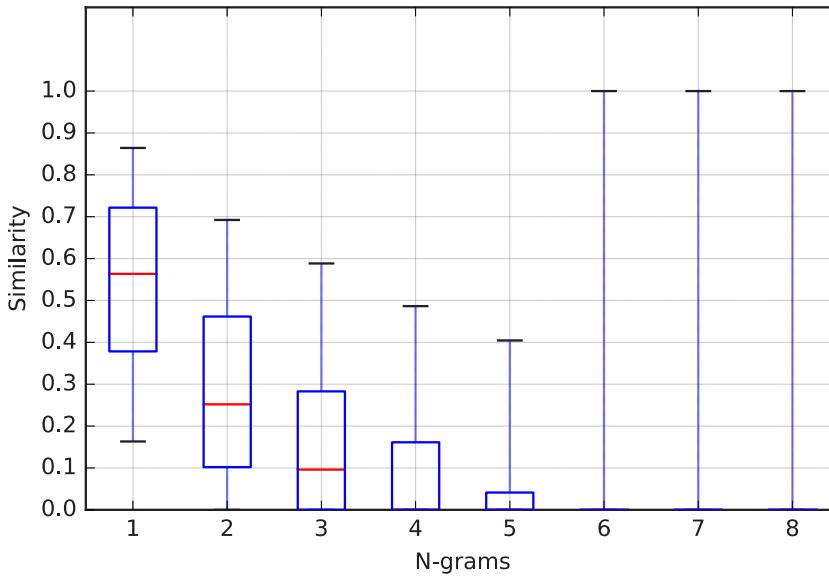

Fig. 3. Distribution of measured similarities for **Essays-2**

This figure emphasizes the aforementioned characteristics of obfuscation techniques: as DEL, ADD being easy for detection, LPR, HPR being moderately or heavily paraphrased text and CCT, SSP being almost verbatim compilation/decompilation of a source text. CCT and SSP were meant to introduce obfuscation via destruction of the structure of reused sentences and there were no special requirements for disguising of a text.

The decrease of similarity for **Essays-2** is quite steep. The main difference from **Essays-1** is the lack of any copy-paste text from the sources. There are 30–60% of 3-gram in common only for 25% of all sentence pairs. It means that source retrieval and text alignment performed on this collection can be a challenging task. Figure 4 shows the distribution of similarity for each obfuscation type of **Essays-2** collection.

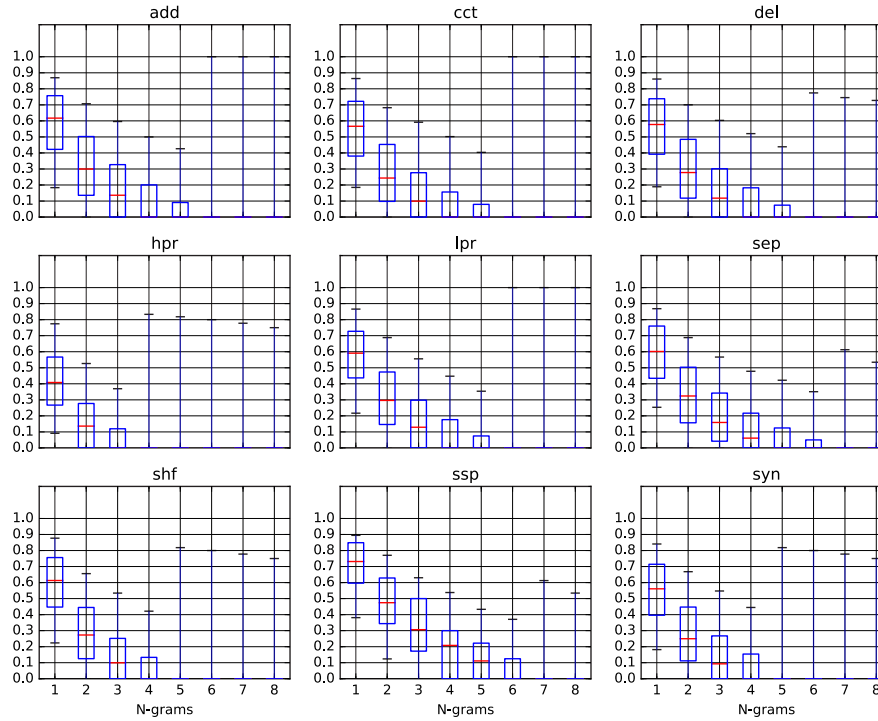

Fig. 4. Distribution of measured similarities per obfuscation type for **Essays-2**

There is a similar pattern for all distributions, which reflects the distribution of similarity for all pairs. It can be explained that these techniques were used usually together, not separately as was the case in the collection **Essays-1**. Another difference from **Essays-1** is the usage of CCT technique. It commonly indicates that a passage of a text (3–6 sentences) were used to produce short summary.

4. ParaPlag as a training dataset on PlagEvalRus-2017

The Russian Plagiarism Evaluation Seminar uses the ParaPlag as a primary dataset. The organizing committee decided to use **Essays-1** and **Essays-2** subsets as a training data for source retrieval and text alignment tracks. They also have automatically generated copy-paste and paraphrased essays to evaluate quality metrics on a big test set. The independent essay writers were encouraged to prepare additional test set similar to **Essays-2**. For this test set writers use TextApp as a search engine with indexed ParaPlag to find sources for their topics. Therefore, they do not extend the sources set. Three testing subsets (manually written essays, generated copy-paste and generated paraphrased essays) were merged and offered to competitors as a tasks in the form of plain text. Competitors do not know the mappings to sources and alignment for this training data. Thus, they should send the results of their PDS on these tasks. Finally, the organizers will calculate quality metrics according the source retrieval and text alignment tracks.

5. Conclusion and future work

We presented the ParaPlag: the Russian dataset for evaluating methods for paraphrased plagiarism detection. The ParaPlag is open and available on the Web[11]. It is used as one of the main datasets on PlagEvalRus-2017 competition. We plan to analyze the participants feedback and provide the updated version of this dataset. Hope it helps to advance the quality of modern PDS. We will continue our work on the typology of techniques used to paraphrase text and hide the plagiarism.

We plan to develop an integrated plagiarism detection task that encourages competitors to solve both source retrieval and text alignment in one track using their own plagiarism retrieval engines. The idea is that competitors need to find sources first and then to align plagiarized fragments, so these two stages could not be optimized separately.

In addition, the ParaPlag can be developed in other directions. As we have previously said, unscrupulous authors can use tools and methods to prevent correct text extraction from plagiarized documents in common formats. It is possible to investigate the ways that such tools “bring noise” to the documents to disturb text extraction procedures. Developing the test dataset of such obfuscated documents in different formats can boost methods that can detect and withstand “noise bringing” tools. Another challenging task for the future is to create a parallel (i.e. English–Russian) dataset for translated plagiarism detection.

Acknowledgments

We are grateful to students of Peoples’ Friendship University of Russia (RUDN University), and Higher School of Economics (HSE) for writing essays. Our special thanks to Ph.D. student Margarita Kamenskaya for management of that process. The research is supported by RFBR, project № 16-37-60048 (mol_ad_dk).

[11] https://plagevalrus.github.io/content/corpora/paraplag.html

References

1. *Ageev, M. S., Dobrov, B. V.* (2011), An efficient nearest neighbours search algorithm for full-text documents. Vestnik S.-Petersburg Univ. Ser. 10. Prikl. Mat. Inform. Prots. Upr., (3), pp. 72–84.
2. *Agirre, E., Baneab, C., Cardiec, C., Cerd, D., Diabe, M., Gonzalez-Agirrea, A., & Mihalceab, R.* (2015), Semeval-2015 task 2: Semantic textual similarity, english, spanish and pilot on interpretability. In Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015), pp. 252–263.
3. *Burrows, S., Potthast, M., & Stein, B.* (2013), Paraphrase acquisition via crowdsourcing and machine learning. ACM Transactions on Intelligent Systems and Technology (TIST), 4(3), p. 43.
4. *Clough, P., Gaizauskas, R., Piao, S. S., & Wilks, Y.* (2002), METER: Measuring text reuse. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pp. 152–159.
5. *Dolan, W. B., Brockett, C.* (2005), Automatically constructing a corpus of sentential paraphrases. In of The Third International Workshop on Paraphrasing. M. Dras and K. Yamamoto, Eds. Kazuhide Yamamoto, Jeju, South Korea, pp. 1–8.
6. *Gollub, T., Stein, B., Burrows, S. and Hoppe, D.,* (2012), TIRA: Configuring, executing, and disseminating information retrieval experiments. In Database and expert systems applications (DEXA), 2012, pp. 151–155.
7. *Korobov M.* (2015), Morphological Analyzer and Generator for Russian and Ukrainian Languages, Analysis of Images, Social Networks and Texts, pp 320–332.
8. *Madnani, N., Dorr, B. J.* (2010), Generating phrasal and sentential paraphrases: A survey of data-driven methods. Computational Linguistics, 36(3), pp. 341–387.
9. *Osipov, G., Smirnov, I., Tikhomirov, I., Sochenkov, I., Shelmanov, A.* (2016), Exactus Expert—Search and Analytical Engine for Research and Development Support. In Novel Applications of Intelligent Systems. Springer International Publishing, pp. 269–285.
10. *Ilyinski, S., Kuzmin, M., Melkov, A. & Segalovich, I.* (2002), An efficient method to detect duplicates of web documents with the use of inverted index, in "Proceedings of 11th International Conference on World Wide Web", Honolulu, Hawaii,
11. *Potthast M., Stein B., Barrón-Cedeño A., Rosso P.* (2010), An evaluation framework for plagiarism detection, Proceedings of the 23rd international conference on computational linguistics: Posters, Beijing, pp. 997–1005.
12. *Potthast, M., Hagen, M., Stein, B., Graßegger, J., Michel, M., Tippmann, M., & Welsch, C.* (2012), ChatNoir: a search engine for the ClueWeb09 corpus. In Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, pp. 1004–1004.
13. *Potthast M., Hagen M., Völske M., Stein B.* (2013), Crowdsourcing Interaction Logs to Understand Text Reuse from the Web, Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 13) pp. 1212–1221
14. *Pronoza, E., Yagunova, E., & Pronoza, A.* (2016). Construction of a Russian paraphrase corpus: unsupervised paraphrase extraction. In Information Retrieval. Springer International Publishing, pp. 146–157.
15. *Suvorov, R. E., Sochenkov, I. V.* (2015), Establishing the similarity of scientific and technical documents based on thematic significance. Scientific and Technical Information Processing, 42(5), pp. 321–327.