Компьютерная лингвистика и интеллектуальные технологии

По материалам ежегодной международной конференции «Диалог» (2017)

Выпуск 16

Том 1 из 2
Компьютерная лингвистика:
практические приложения

Computational Linguistics and Intellectual Technologies

Papers from the Annual International Conference “Dialogue” (2017)

Issue 16

Volume 1 of 2
Computational Linguistics: Practical Applications

УДК 80/81; 004
ББК 81.1
К63

Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной международной конференции «Диалог» (Москва, 31 мая — 3 июня 2017 г.). Вып. 16 (23): В 2 т. Т. 1 — М.: Изд-во РГГУ, 2017.

Сборник включает 71 доклад международной конференции по компьютерной лингвистике и интеллектуальным технологиям «Диалог 2017», представляющих широкий спектр теоретических и прикладных исследований в области описания естественного языка, моделирования языковых процессов, создания практически применимых компьютерных лингвистических технологий.

Для специалистов в области теоретической и прикладной лингвистики и интеллектуальных технологий.

Предисловие

16-й выпуск ежегодника «Компьютерная лингвистика и интеллектуальные технологии» содержит избранные материалы 23-й международной конференции «Диалог». На основании мнений наших рецензентов для публикации в ежегоднике Редсоветом был отобран 71 доклад из числа примерно ста работ, которые были рекомендованы по результатам рецензирования для представления на конференции в 2017 году.

Работы в сборнике отражают все основные направления исследований в области компьютерного моделирования и анализа естественного языка, представленные на конференции:

- Компьютерные лингвистические ресурсы
- Компьютерный анализ документов (классификация, поиск, анализ тональности и т.д.)
- Корпусная лингвистика (создание, разметка, методики применения и оценка корпусов)
- Лингвистические онтологии и автоматическое извлечение знаний
- Лингвистический анализ Social media
- Лингвистический анализ речи
- Машинный перевод текста и речи
- Модели и методы семантического анализа текста
- Модели общения
- Теоретическая и компьютерная лексикография
- Типология и компьютерная лингвистика
- Формальные модели языка и их применение в компьютерной лингвистике

В соответствии с традициями «Диалога», старейшей и крупнейшей конференции по компьютерной лингвистике в России, отбор работ основывается на представлении о важности соединения новых методов и технологий анализа языковых данных с полноценным лингвистическим анализом и моделированием. Одной из важнейших целей конференции была и остается поддержка создания современных компьютерных ресурсов, моделей и технологий для русского языка.

В годовом цикле проведения конференции в рамках программы Dialogue Evaluation проводится тестирования технологий решения отдельных задач компьютерного анализа языка. На конференции подводятся итоги проведенных тестов, а статьи организаторов и наиболее успешных участников представляются в настоящем сборнике.

В этом году было проведено два тестирования:

1. По идентификации внешних заимствований (External Plagiarism Detection)
2. По оценке методов морфологического анализа русского языка, с акцентом на тексты Social Media.

Как обычно, результатом проведенных тестирований стали не только объективные данные о качестве работы различных методов и алгоритмов, но также и открытые для использования эталонные размеченные корпуса, т. н. золотые стандарты, позволяющие любым исследователям проводить сравнительные оценки эффективности своих технологий.

Все направления «Диалога» важны, но каждый год какие-то темы занимают особое место в программе конференции и в составе ежегодника. В этом году можно назвать две таких темы:

1. Применение методов глубинного машинного обучения: прежде всего — нейросетей и таких результатов их применения как word embeddings, как для прикладных задач, так и в лингвистических исследованиях.
2. В программе конференции этого года особенно заметны работы по использованию параллельных корпусов для лингвистических исследований. Такие корпуса уже давно и успешно используются в NLP, например, для обучения статистических моделей машинного перевода, автоматической дизамбигуации, автоматического построения языковых моделей. Но параллельные корпуса оказываются также и важным инструментом контрастивных лингвистических исследований.

Статьи в сборнике публикуются на русском и английском языках. При выборе языка публикации действует следующее правило:

- доклады по компьютерной лингвистике должны подаваться на английском языке. Это расширяет их аудиторию и позволяет привлекать к рецензированию международных экспертов.

- доклады, посвященные лингвистическому анализу русского языка, предполагающие знание этого языка у читателя, подаются на русском языке (с обязательной аннотацией на английском).

Несмотря на традиционную широту тематики представленных на конференции и отобранных в сборник докладов они не могут дать полной картины направлений «Диалога». Ее можно получить с помощью сайта конференции www.dialog-21.ru, на котором представлены обширные электронные архивы «Диалогов» последних лет и все результаты проведенных тестирований Dialogue Evaluation.

Мы обращаем внимание авторов и читателей сборника, что его бумажный вариант, который вы держите в руках, является вторичным по отношению к сборнику, который размещается на сайте конференции и индексируется Scopus. Мы рекомендуем при цитировании использовать именно сетевую версию.

Программный комитет конференции «Диалог»

Редколлегия сборника «Компьютерная лингвистика и интеллектуальные технологии»

Организаторы

Ежегодная конференция «Диалог» проводится под патронажем Российского фонда фундаментальных исследований при организационной поддержке компании ABBYY.

Учредителями конференции являются:

- Институт лингвистики РГГУ
- Институт проблем информатики РАН
- Институт проблем передачи информации РАН
- Компания ABBYY
- Филологический факультет МГУ

Конференция проводится при поддержке Российской ассоциации искусственного интеллекта.

Международный программный комитет

Богуславский Игорь Михайлович	Институт проблем передачи информации РАН им. А. А. Харкевича, Россия
Буате Кристиан	Университет Джозефа Фурье — Гренобль 1, Франция
Гельбух Александр Феликсович	Национальный политехнический институт, Мехико
Иомдин Леонид Лейбович	Институт проблем передачи информации РАН им. А. А. Харкевича, Россия
Кобозева Ирина Михайловна	Московский государственный университет им. М. В. Ломоносова, Россия
Козеренко Елена Борисовна	Институт проблем информатики РАН, Россия
Корбетт Гревил	Университет Суррея, Великобритания
Кронгауз Максим Анисимович	НИУ «Высшая школа экономики», Россия
Лукашевич Наталья Валентиновна	НИВЦ МГУ им. М. В. Ломоносова, Россия
Маккарти Диана	Кембриджский университет, Великобритания
Мельчук Игорь Александрович	Монреальский университет, Канада
Нивре Йоаким	Уппсальский университет, Швеция
Ниренбург Сергей	Университет Мэриленда, Балтимор, США
Осипов Геннадий Семёнович	Институт системного анализа РАН, Россия
Раскин Виктор	Университет Пердью, США
Селегей Владимир Павлович	Компания ABBYY, Россия
Хови Эдуард	Университет Карнеги — Меллон, США
Шаров Сергей Александрович	Университет Лидса, Великобритания

Организационный комитет

Селегей Владимир Павлович, *председатель*	Компания ABBYY
Байтин Алексей Владимирович	Компания Yandex
Беликов Владимир Иванович	Институт русского языка им. В. В. Виноградова РАН
Браславский Павел Исаакович	Уральский федеральный университет
Добров Борис Викторович	НИВЦ МГУ им. М. В. Ломоносова
Захаров Леонид Михайлович	Московский государственный университет им. М. В. Ломоносова
Иомдин Леонид Лейбович	Институт проблем передачи информации РАН им. А. А. Харкевича
Кобозева Ирина Михайловна	Московский государственный университет им. М. В. Ломоносова
Козеренко Елена Борисовна	Институт проблем информатики РАН
Лауфер Наталия Исаевна	Компания Yandex
Ляшевская Ольга Николаевна	Институт русского языка им. В. В. Виноградова РАН
Толдова Светлана Юрьевна	НИУ «Высшая школа экономики»
Федорова Ольга Викторовна	Московский государственный университет им. М.В. Ломоносова
Шаров Сергей Александрович	Университет Лидса

Секретариат

Атясова Анастасия Леонидовна, *координатор оргкомитета*	Компания ABBYY
Белкина Александра Андреевна, *секретарь оргкомитета*	Компания ABBYY
Гусева Анна Александровна, *координатор Dialogue Evaluation*	Компания ABBYY
Севергина Екатерина Александровна, *администратор оргкомитета*	Компания ABBYY

Рецензенты

Августинова Таня
Антонова Александра Александровна
Азарова Ирина Владимировна
Андрианов Андрей Иванович
Апресян Валентина Юрьевна
Архангельский Тимофей Александрович
Байтин Алексей Владимирович
Баранов Анатолий Николаевич
Беликов Владимир Иванович
Бенко Владимир
Бердичевский Александр Сергеевич
Богданов Алексей Владимирович
Богданова-Бегларян Наталья Викторовна
Богуславский Игорь Михайлович
Бочаров Виктор Владиславович
Браславский Павел Исаакович
Васильев Виталий Геннадьевич
Галинская Ирина Евгеньевна
Галицкий Борис Александрович
Гельбух Александр Феликсович
Гецевич Юрий Станиславович
Гращенков Павел Валерьевич
Губин Максим Вадимович
Даниэль Михаил Александрович
Диконов Вячеслав Григорьевич
Добров Борис Викторович
Добровольский Дмитрий Олегович
Добрушина Нина Роландовна
Зализняк Анна Андреевна
Захаров Виктор Павлович
Захаров Леонид Михайлович
Ильвовский Дмитрий Алексеевич
Иомдин Борис Леонидович
Иомдин Леонид Лейбович
Катинская Анисья Юрьевна
Клышинский Эдуард Станиславович
Кибрик Андрей Александрович
Князев Сергей Владимирович
Кобозева Ирина Михайловна
Козеренко Елена Борисовна
Копотев Михаил Вячеславович
Коротаев Николай Алексеевич
Котельников Евгений Вячеславович
Котов Артемий Александрович
Кронгауз Максим Анисимович
Левонтина Ирина Борисовна
Лобанов Борис Мефодьевич
Лопухин Константин Александрович
Лукашевич Наталья Валентиновна
Лютикова Екатерина Анатольевна
Мисюрев Алексей Владимирович
Наков Преслав
Недолужко Анна Юрьевна
Падучева Елена Викторовна
Пазельская Анна Германовна
Паперно Денис Аронович
Панченко Александр Иванович
Переверзева Светлана Игоревна
Петрова Мария Андреевна
Пивоварова Лидия Михайловна
Пиперски Александр Чедович
Подлесская Вера Исааковна
Рахилина Екатерина Владимировна
Скулачева Татьяна Владимировна
Смирнов Иван Валентинович
Селегей Владимир Павлович
Слюсарь Наталия Анатольевна
Соколова Елена Григорьевна
Сомин Антон Александрович
Сорокин Алексей Андреевич
Сорокин Виктор Николаевич
Старостин Анатолий Сергеевич
Степанова Мария Евгеньевна
Тихомиров Илья Александрович
Толдова Светлана Юрьевна
Турдаков Денис Юрьевич
Урысон Елена Владимировна
Федорова Ольга Викторовна
Хохлова Мария Владимировна
Циммерлинг Антон Владимирович
Шаров Сергей Александрович
Шелманов Артём Олегович
Янко Татьяна Евгеньевна

Contents[1]

Приглашенные доклады

Компьютерная лингвистика: практические приложения

[*] Доклады упорядочены по фамилии первого автора в соответствии с английским алфавитом. The reports of each section are ordered by the surname of the first author in compliance with the English alphabet.

EVALUATION TRACKS ON PLAGIARISM DETECTION ALGORITHMS FOR THE RUSSIAN LANGUAGE

Smirnov I. (ivs@isa.ru)
Institute for Systems Analysis, FRC CSC RAS, Moscow, Russia; RUDN University, Moscow, Russia

Kuznetsova R. (kuznetsova@ap-team.ru)
Antiplagiat JSC, Moscow, Russia

Kopotev M. (mihail.kopotev@helsinki.fi)
University of Helsinki, Helsinki, Finland

Khazov A. (hazov@ap-team.ru)
Antiplagiat JSC, Moscow, Russia

Lyashevskaya O. (olesar@yandex.ru)
Higher School of Economics, Moscow, Russia; Vinogradov Institute of the Russian Language RAS, Moscow, Russia

Ivanova L. (luben92@gmail.com)
Higher School of Economics, Moscow, Russia

Kutuzov A. (andreku@ifi.uio.no)
University of Oslo, Oslo, Norway

The paper presents a methodology and preliminary results for evaluating plagiarism detection algorithms for the Russian language. We describe the goals and tasks of the PlagEvalRus workshop, dataset creation, evaluation setup, metrics, and results.

Keywords: plagiarism detection, paraphrased plagiarism, source retrieval, text alignment, evaluation workshop

1. Introduction

According to the *MLA Style Manual and Guide to Scholarly Publishing* (Modern Language Association 2008: p. 166),

> [f]orms of plagiarism include the failure to give appropriate acknowledgment when repeating another's wording or particularly apt phrase, paraphrasing another's argument, and presenting another's line of thinking.

Two types of plagiarism are usually distinguished in the scholarly literature: *literal* and *obfuscated* plagiarism (Potthast et al. 2010b: 2) and *disguised* plagiarism (Gipp 2014: 12). Bela Gipp calls these two types of plagiarism *copy & paste* and *shake & paste*. The first type involves taking someone else's text word-for-word without citation, while the second involves minor modifications in another person's words, such as varying the word order, using synonyms or "padding" (Gipp 2014: 11), again without acknowledgment. According to other researchers, the *shake & paste* technique includes insertion of additional paragraphs relevant to the subject as well as mixing paragraphs. This typically leads to a sudden change in style and may remain unnoticed by a reader. When changes in an original, unattributed text are more significant (e.g., a text is paraphrased or translated), plagiarism is described as *obscured*. In paraphrasing, the source texts are reworked with the use of different linguistic tricks such as removal, word replacement, synonym substitution, word order modification, grammatical changes, and patchwriting (i.e., combining fragments from several texts) (Oakes 2014: 60). The nature of these changes depends on whether the paraphrase has occurred through manual text editing or by using automatic methods (Gupta et al. 2011: 1). For example, a manually rewritten text may be better adapted to a plagiarist's personal style than one edited automatically. Still, another case of paraphrasing is *interlingual* plagiarism, when a text is "paraphrased," in a sense, from one source language to another one. The process may include either manual or automatic translation. In the latter, an output of the machine translator usually goes through editing afterwards and obfuscation, which makes comparing the sources with the plagiarized text substantially more difficult while at the same time showing evidence of translation.

In the academic community, the problem is especially crucial in connection with student papers and popular scientific literature. Plagiarism is especially difficult to define in the latter case, since such literature describes facts that are already known and often cannot be reformulated differently. Thus, establishing both the evidence for and the limits of plagiarism becomes more challenging and problematic. In contrast, student plagiarism usually can be detected using basic automated tools. Its widespread occurrence today is primarily the result of the tolerance on the part of educators and the academic community, which makes plagiarism a common practice. In 2004, for instance, it was estimated that 10 percent of student works in the United States and Australia involved plagiarism (Oakes 2014: 60). In more recent research, 36 percent of respondents in Russia admitted to regularly copying the texts of others in different forms (Kicherova et al. 2013: 2). According to a study conducted in 2013 (Maloshonok 2016), as many as 36.7 percent of undergraduate students in eight Russian universities take personal credit for works they have downloaded from the Internet. However, the problem is not limited to students' activity. In 2011 in Germany, two cases of plagiarism were documented in Ph.D. dissertations. Those cases were analyzed in detail by the GuttenPlag community and provided the basis for the monograph *False Feathers: A Perspective on Academic Plagiarism* (Weber-Wulff 2014: 29). In Russia, the same problem has been diagnosed by the Dissernet grassroots movement (www.dissernet.org), whose purpose is to reveal plagiarism in scientific texts (see Golunov 2014; Denisova-Schmidt 2016).

As disguising plagiarism becomes more and more sophisticated, detecting it requires newer and more advanced techniques. At the moment, there are several services that are able to detect plagiarism in Russian-language texts (see Nikitov et al.

2012), but thus far there has been no systematic evaluation of these services. This paper and the PlagEvalRus workshop it stems from are the first attempts to define the problem of how to evaluate plagiarism and outline ways of handling it.

There are several related workshops and events on similarity detection on both word and sentence levels. The Russian language is a primary target for two of them: 1) RUSSE (Panchenko et al. 2015), the shared task on word-level semantic similarity; and 2) ParaPhrase (Pivovarova et al. 2016), the shared task on sentence-level paraphrase detection, i.e. identification of sentences that have similar meaning but not necessarily similar in structure. The series of related workshops, SemEval, includes a task on Semantic Textual Similarity (Agirre et al. 2016), which is aimed to measure degree of semantic similarity between two text snippets, written in English and some other languages (but not in Russian). However, in plagiarism detection tasks, snippets of reused texts are not given, but supposed to be retrieved from source texts, thus this task is significantly more complicated to accomplish. The most closely related to the PlagEvalRus seminar are PAN workshops (e.g. Potthast et al. 2010a) that have several tasks on plagiarism detection.

2. Goals and tasks

In this article the methodology we propose for detecting plagiarism in the Russian language is based on years of experience of the PAN network (a series of events on digital text forensics [e.g., Potthast et al. 2010a, Potthast et al. 2010b, 2014]; see more on http://pan.webis.de). We have focused on evaluating algorithms oriented toward monolingual Russian plagiarism with an emphasis on scientific texts (academic plagiarism). In our workshop, called PlagEvalRus and held during 2016-2017, we offered the following tracks after holding preliminary discussion:

- Track 1: Plagiarized sources retrieval
- Track 2: Copy and paste plagiarism detection
- Track 3: Paraphrased plagiarism detection.

Track 1 corresponds to the Source Retrieval (SR) task evaluated at the PAN competitions. The participants received a dataset, which includes potential sources and suspicious texts; the latter contained both literal and paraphrased plagiarism. The participants are required to provide a list of sources for each suspicious text (more details below), sorted according to the number of reused fragments in descending order; unlike the PAN Source Retrieval task does not require any sorting of detected text pieces. Track 1 was thus quite similar to the search tracks on the Russian Information Retrieval Evaluation Seminar (see http://romip.ru/en), the difference being that the search queries in our case were much longer textual excerpts.

Tracks 2 and 3 entirely correspond to the Text Alignment (TA) task evaluated at the PAN competitions; i.e., in a pair of texts given to participants, fragments taken from one text need to be found in a second text. A **fragment** is a sequence of at least five tokens excluding stop words. **Literal reusing** means a full correspondence of character strings ignoring blank and hidden characters. **Paraphrased reusing** is rewriting the original text preserving the idea of a reused fragment. Thus, Track 2 was intended to detect literal plagiarism, while Track 3 involved detecting illicit paraphrasing.

3. Dataset

For each track, the organizers provided two datasets, training and testing, along with a text collection that contains, among other things, potential sources. Participants were supposed to train their algorithms on the training dataset, which was provided to all participants and could be read on the Workshop's site, www.dialog-21.ru/en/evaluation/2017/plageval, well in advance. The participants received clear instructions on how to handle the data. All scripts, datasets and instructions are freely accessible at https://plagevalrus.github.io.

3.1. Collection of sources

The "potential sources" dataset contains about 5.7 million Russian texts, compiled from the following resources:

- Russian Wikipedia: about 1.3 million texts;
- Student essays from open online collections: about 3.3 million texts;
- Open-sourced book-sized academic texts: about 12,000 texts;
- Academic papers from the open access resource Cyberleninka.ru: 1 million texts.

All texts were converted to the plain-text format in UTF-8. Evident duplicates were preliminarily removed, and the remaining files were then mixed. Each text was stored in a separate file with a name containing a unique identifier.

3.2. Suspicious Texts

The test dataset was created under the same conditions as the training dataset. In line with the PAN workshops (Potthast et al. 2010a), the following types of texts were specified:

1) **Automatically generated copy and paste plagiarism**. To do this, we randomly selected sentences from a target text and changed each of them by one or more randomly chosen consecutive sentences from source texts, which did not belong to the target collection. Each fragment was identified by its beginning and its length in characters.

   The resulting target texts contain from 10 to 80 percent of plagiarized material (calculated in sentences).

2) **Automatically generated paraphrased plagiarism**. The collection containing this type of reused texts was created in the same way as the copy and paste texts, except that the sentences of the source texts were automatically paraphrased by using one or more of the following methods:
   - Replacing words with their synonyms;
   - Adding and removing synonym chains;
   - Abbreviation and amplification;
   - Adding and removing diminutives;
   - Singular/plural replacement.

   For a detailed description of the procedure, see (Khazov and Kuznetsova 2017).

3) **Manually generated copy and paste plagiarism**. This dataset was compiled from academic texts, the sources of which are known and available on the Internet. The texts with the manually created word-for-word fragments were used only for Track 1.

4) **Manually paraphrased plagiarism**. Compiling such a collection was motivated by the activity of those "authors" who reuse fragments from various sources trying to obfuscate their borrowings by paraphrasing. This collection is built of essays reflected different topics; creators were instructed to select a text from the source collection, to mark and paraphrase fragments, and then to insert them into a Microsoft Excel table. The procedure like this makes it possible to extract the markups and transform it into different tasks related to a plagiarism detection evaluation. A fragment that has been restructured should contain at least one sentence. The creators were allowed mixing sentences from different sources and inserting original sentences between plagiarized ones. Therefore, the resulted essays contain both original and paraphrased fragments, which are produced by creators under the following condictions:
   - deleting words (about 20%) from an original sentence;
   - adding words (about 20%) into an original sentence;
   - replacing words or phrases with synonyms, reordering clauses, adding new words, changing word forms (number, case, form and verb tense, etc.); about 30% in an original sentence;
   - changing the order of words or clauses in an original sentence;
   - concatenating two or more original sentences into one;
   - splitting an original sentence into two or more (with a possible changing in order of how they appear in a text);
   - replacing words or phrases of an original sentence with synonyms (e.g. "sodium chloride" → "salt"), replacing abbreviations to their full transcripts and vice versa, replacing personal names with their initials, etc.;
   - complex rewriting of an original sentence, which combines 3-5 or more aforementioned techniques. This type involves significant changes in a source text by paraphrasing idioms, synonymic modification of structures, permutation of words or parts of a complex sentence, etc. Using this technique effectively produces paraphrased texts: in some cases even experts could hardly consider the rewritten text as plagiarized;
   - coping a sentence from a source and pasting it into an essay with no significant changes.

Therefore, each essay contains no fewer than 100 paraphrased sentences, 90 percent of the texts being taken from at least five sources. For a detailed description of the procedure, see (Sochenkov et al. 2017). Table 1 shows the number of texts and pairs <suspicious text, source text> in both training and the testing data.

Table 1. The Training and the Test Data sets: size in the number of texts and pairs

	Training set		Test set		
	Texts for SR and TA	**Pairs**	**Texts for SR**	**Texts for TA**	**Pairs**
Automatically generated copy&paste plagiarism	1,000	4,257	5,000	100	268
Automatically generated paraphrased plagiarism	2,000	4,251	5,000	100	297
Manually copy&paste plagiarism	519	—	519	—	—
Manually paraphrased plagiarism	152	913	38	39	234
Total	3,671	9,421	10,557	239	799

Figure 1 shows the texts we suspected of plagiarizing (from 1 to 19 sources).

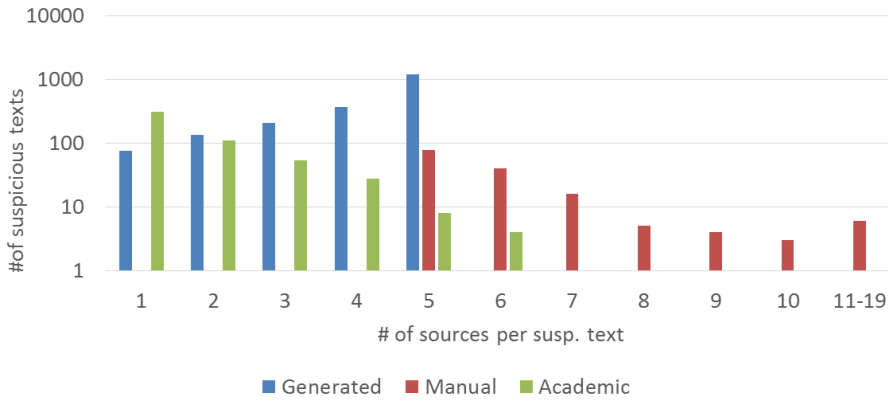

Figure 1. Texts suspected of plagiarizing N sources (where N ranges from 1 to 19)

4. Evaluation

4.1. Evaluation Setup

The evaluation of the results on Track 1, Source Retrieval, differs significantly from that on Tracks 2 and 3, Text Alignment. On Track 1, the participants downloaded the collection of sources and searched for copied fragments using a system of the participant's own devising. The result is supposed to be a list containing sources for each suspicious text, ranked (in descending order) according to the number of fragments detected. Those following this track were asked to deliver results for a maximum of 5 runs. In evaluating the runs, the participants' responses were automatically

evaluated against the benchmark created by the PlagEvalRus Workshop's organizers. A baseline was not offered for the source retrieval track due to both the complexity of the task and lacking time needed for it development.

For Tracks 2 and 3, plagiarism is considered successfully detected if a fragment found by a system is located or completely within a text marked as such in the test collection. Coincidences in texts were not taken into account. Therefore, any fragment detected, but not marked in the test collection was not registered for the evaluation. The PAN baseline method was used in comparing results. This brute method is based on a simple shingles approach with chunks of 50 symbols length.

To evaluate the systems on Tracks 2 and 3, we used the TIRA platform (http://www.tira.io),1 which ensured reproducibility and neutrality in evaluating the algorithms. Each participant in Track 2 or 3 was provided with a virtual machine on the TIRA server in order to run his/her system on a given test set. The evaluation was performed automatically on the server and the results were available to the participant. The overall results are available only to the administrator of the TIRA service.

4.2. Evaluation Metrics

4.2.1. Source Retrieval

Let T_{src} denote a set of source texts for suspicious text t_{plg}, and let T_{ret} denote the set of texts that is retrieved by a source retrieval algorithm when given t_{plg}. Then precision (P) is defined as

$$P = \frac{|T_{ret} \cap T_{src}|}{|T_{ret}|}$$

and recall (R) as

$$R = \frac{|T_{ret} \cap T_{src}|}{|T_{src}|}$$

The PAN metrics (Potthast et al. 2014) measures the effect of near-duplicate web documents, but we do not take into account similar texts from T_{ret}. Furthermore, full duplicates were preliminarily removed from the collection of sources.

We define F-measure (F) as

$$F = \frac{2 * R * P}{R + P}$$

The results of Track 1 were supposed to be ranked in descending order according to number of reused fragments detected, so that we could assess the quality of ranking. The text t_{ret} is relevant to t_{plg} if $t_{plg} \in T_{ret} \cap T_{src}$. Precision at k ($P@k$) is a measure of ranking performance for t_{plg} and is defined as the number of relevant texts among the first k retrieved results, divided by k.

The average precision (AP) for t_{plg} is the average of $P@k$ for all relevant texts:

---

1 TIRA is currently one of the few platforms (if not the exclusive one) that support software submissions with a little extra effort; it has been utilized for several similar shared tasks within PAN@CLEF, CoNLL, etc.

$$AP(t) = \frac{1}{|K|} \sum_{k \in K} P@k$$

where K stands for a set of positions of all relevant texts. The mean average precision (*MAP*) is the mean of the average precision for each text from a set of suspicious texts denoted T_{plg}.

$$MAP = \frac{1}{|T_{plg}|} \sum_{t_{plg} \in T_{plg}} AP(t_{plg})$$

4.2.2. Text Alignment

Following (Potthast et al. 2010b), let S denote the set of plagiarism cases in the corpus, and let R denote the set of detections reported by a plagiarism detector for the suspicious documents. A plagiarism case $s = \langle s_{plg}, d_{plg}, s_{src}, d_{src} \rangle$, $s \in S$, is represented as a set s of references to the characters of t_{plg} and t_{src}, specifying the passages s_{plg} and s_{src}. Likewise, a plagiarism detection $r \in R$ is represented as r. Based on this notation, both macro- and micro-averaged precision and recall of R under S can be measured as follows:

$$precision_{micro}(S, R) = \frac{\left| \cup_{(s,r) \in (S \times R)} (s \sqcap r) \right|}{|\cup_{r \in R^r}|}$$

$$precision_{macro}(S, R) = \frac{1}{|R|} \sum_{r \in R} \frac{|\cup_{s \in S} (s \sqcap r)|}{|r|}$$

$$recall_{micro}(S, R) = \frac{\left| \cup_{(s,r) \in (S \times R)} (s \sqcap r) \right|}{|\cup_{s \in S^s}|}$$

$$recall_{macro}(S, R) = \frac{1}{|S|} \sum_{s \in S} \frac{|\cup_{r \in R} (s \sqcap r)|}{|s|}$$

where

$$s \sqcap r = \begin{cases} s \cap r \text{ if } r \text{ detects } s, \\ \varnothing \text{ otherwise.} \end{cases}$$

The macro-averaged variants are allotted equal weight in each plagiarized case, regardless of length. Conversely, the micro-averaged variants favor detecting long plagiarized fragments, which are generally easier to identify.

To address the fact that plagiarism detectors sometimes reported overlapping or multiple detections for a single plagiarism case, let a detector's granularity be defined as:

$$granularity(S, R) = \frac{1}{|S_R|} \sum_{s \in S_R} |R_S|$$

where $S_R \subseteq S$ are cases detected by detections in R, and $R_S \subseteq R$ are detections of s; i.e., S_R $\{s | s \in S \wedge \exists r \in R$:$r$ detects $s\}$ and R_S $\{r | r \in R \wedge r \in R$:$r$ detects $s\}$. The three above-mentioned measures taken individually do not allow single ranking based on these approaches. To make a uniform ranking, the measures are combined into a single overall score, called the Plagdet score and calculated as follows:

$$plagdet(S,R) = \frac{F_1}{\log_2(1 + gran(S,R))}$$

where F_1 is the equally weighted harmonic mean of precision and recall.

4.3. Evaluation Results

Only one team participated in all offered Tracks (Hereafter referred to as **zubarev**; see Zubarev and Sochenkov 2017). The results of all runs are shown in Tables 2–8.

4.3.1. Track 1: Plagiarized source detection

The evaluation results for the track are presented in Tables 2–4.

Table 2. Evaluation results for the automatically-generated copy and paste and paraphrased plagiarism retrieval tasks

team	Run	generated copy&paste plagiarism				generated paraphrased plagiarism			
		MAP	P	R	F1	MAP	P	R	F1
zubarev	zubarev.1	0.603	0.222	0.779	0.346	0.593	0.234	0.745	0.357
	zubarev.2	0.151	0.005	0.785	0.011	0.202	0.005	0.750	0.011

Table 3. Evaluation results for manual copy and paste and paraphrased plagiarism retrieval task

team	run	manual copy&paste plagiarism				manually paraphrased plagiarism			
		MAP	P	R	F1	MAP	P	R	F1
zubarev	zubarev.1	0.851	0.106	0.974	0.191	0.608	0.441	0.830	0.576
	zubarev.2	0.610	0.003	0.978	0.006	0.390	0.009	0.989	0.019

Table 4. Evaluation results for overall source retrieval tasks

team	runs	Total			
		MAP	P	R	F1
zubarev	zubarev.1	0.664	0.251	0.832	0.368
	zubarev.2	0.338	0.005	0.876	0.012

The participant has submitted 36 sources in average for each suspicious text in zubarev.1 run and 579 sources in average for each suspicious text in zubarev.2 run, so the second run is obviously optimized for higher recall. As one can see, the best F1 and MAP was gained on manual plagiarism detection. We suppose the reason behind that is a topical heterogeneity of automatically generated texts that might affect participant's algorithms. The results in general correspond to average results of PAN participants, who showed the highest F1 equaled 0.47 at PAN2015.

4.3.2. Track 2: Copy and paste plagiarism detection

The evaluation results for the automatically-generated copy and paste plagiarism retrieval are shown in Table 5.

Table 5. Evaluation results for automatically-generated copy and paste plagiarism detection. Macro- and Micro-average

		Macro			Micro		
team.run	Granularity	Precision	Recall	Plagdet	Precision	Recall	Plagdet
PAN Baseline	1.0046	0.7240	0.9101	0.8038	0.9615	0.9943	0.9744
zubarev17.1	1.5084	0.9496	0.6427	0.5778	0.9828	0.8217	0.6746
zubarev17.2	1.4660	0.9320	0.7013	0.6146	0.9776	0.8588	0.7022

In this track, the PAN baseline outperforms Zubarev's detector by all measures except precision. In general, the task of copy and paste plagiarism detection has been solved well enough.

4.3.3. Track 3: Paraphrased plagiarism detection

The evaluation results for paraphrased plagiarism retrieval are shown in Tables 6–7.

Table 6. Evaluation results for automatically-generated paraphrased plagiarism detection. Macro- and Micro-average

		Macro			Micro		
team.run	Granularity	Precision	Recall	Plagdet	Precision	Recall	Plagdet
PAN Baseline	3.4639	0.9051	0.6895	0.3626	0.9710	0.8334	0.4156
zubarev17.1	1.5404	0.9604	0.6730	0.5884	0.9875	0.8219	0.6670
zubarev17.2	1.4834	0.9473	0.7340	0.6303	0.9812	0.8650	0.7006

Table 7. Evaluation results for manually paraphrased plagiarism detection. Macro- and Macro-average

		Macro			Micro		
team.run	Granularity	Precision	Recall	Plagdet	Precision	Recall	Plagdet
PAN Baseline	1.1414	0.8332	0.0554	0.0946	0.8960	0.0761	0.1277
zubarev17.1	1.0015	0.8068	0.3409	0.4788	0.8845	0.3815	0.5325
zubarev17.2	1.0016	0.6250	0.4715	0.5369	0.8208	0.5312	0.6443

In this track, Zubarev's detector outperforms PAN baseline by all measures. The results of generated paraphrased plagiarism detection are better than results for manually paraphrased texts, though granularity is better for the last. The reason of a granularity gap is probably connected with the difference in length of fragments in the

tasks: in manually paraphrased texts, the reused fragments equal to sentence, while in automatically generated paraphrased texts, the reused fragments equal to a paragraph (up to 10 sentences).

We can see that the measures on copy and paste plagiarized texts are expectedly higher than measures on paraphrased texts almost in all cases. Nevertheless, the most complicated task of paraphrased plagiarism detection is solved by Zubarev detector quiet well while PAN baseline dropped down Recall and Plagdet in this task.

4.3.4. Plagiarism detection for both types

Evaluation results for automatically-generated copy and paste, automatically-generated and manually paraphrased plagiarism detection are shown in Table 8.

Table 8. Evaluation results for overall text alignment tasks. Macro- and Micro-average

team.run	Granularity	Macro			Micro		
		Precision	Recall	Plagdet	Precision	Recall	Plagdet
PAN Baseline	1.9953	0.8525	0.3366	0.3049	0.9637	0.6893	0.5078
zubarev17.1	1.3028	**0.9129**	0.4605	0.5087	**0.9693**	0.7043	0.6780
zubarev17.2	**1.2417**	0.8158	**0.5644**	**0.5729**	0.9460	**0.7737**	**0.7309**

In the overall text alignment task, the Zubarev's detector (which is based on sentence similarity) performed by the Plagdet better than the PAN baseline (which is based on character shingles. The Zubarev's detector also performed better in all types of plagiarism except an automatically-generated copy and paste variation. In the PlagEvalRus test dataset, the PAN baseline demonstrated results comparable to those on the PAN test dataset in English (Potthast et al. 2014). Finally, we should notice that micro-measures are always higher than macro.

5. Conclusions and further advances

In this article, we have presented the methodology and the datasets for plagiarism detection evaluation algorithms in monolingual Russian texts. Owing to circumstances beyond our control, only one of all the teams which signed up for the PlagEvalRus Workshop submitted its results. Participants' feedback showed that computational complexity and lack of both high-performance computing facilities and large-scale storage systems caused no-bid decisions. Our decision to lay upon TIRA technical solutions should obviously be reconsidered in our further workshops, because the participants have had to invest much time in studying this evaluation framework. Nevertheless, the TIRA framework allows and we agreed to make the text alignment task continuously available for evaluation on the TIRA site (http://www.tira.io/tasks/pan/#text-alignment; see the dataset "pan17-text-alignment-test-dataset-dialogue17-russian-2017-02-22"), so that anyone who submits his/her software can obtain the results for comparison.

Preparation of manually paraphrased texts was the most laborious phase in any workshop like ours. According to our estimations, preparing one essay takes in average from 4 to 10 hours; the properly formed essays are not always resulted on the first try, a (semi)automated verification is always required for this time-consuming preparatory work. Taking both our experience and participants' needs into consideration, we intend to hold PlagEvalRus workshop for the second time next year. We plan to enlarge collection of sources and increase the size of training datasets. We will discuss offering a joint plagiarism detection track, where both source retrieval and text alignment are not separated. We also plan to announce a cross-language (translated) plagiarism detection track expecting more participants at our Workshop.

Acknowledgments

We would like to thank the following people and institutions for various kinds of assistance in organizing this Workshop:

- For both technical support and inspiration: Martin Potthast (PAN founder, Digital Bauhaus Lab.);
- For the data provided: Cyberleninka.ru and other institutions;
- For the preparation of datasets: students of RUDN University, students of the Higher School of Economics in Nizhny Novgorod (A. Safaryan, O. Andriyanova, N. Babkin, A. Bazyleva, A. Beloborodova, Ju. Frolova, M. Kurilina, M. Petrova, V. Rybakov, T. Semenova, A. Sorokina, T. Sharipova, A. Tryaskova, V. Vdovina) and Moscow (S. Malinovskaya, Z. Evdaeva, A. Stepanova, D. Suslova).

References

1. *Agirre E. et al.* (2016) Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation //Proceedings of SemEval. — 2016. — pp. 497–511.
2. *Gipp, B.* (2014), Citation-based Plagiarism Detection. Detecting Disguised and Cross-language Plagiarism Using Citation Pattern Analysis. Springer Fachmedien Wiesbaden. http://link.springer.com/book/10.1007%2F978-3-658-06394-8.
3. *Golunov, S.* (2014), The Elephant in the Room: Corruption and Cheating in Russian Universities. Columbia University Press.
4. *Gupta, P., Singhal, K., Majumder, P., Rosso, P.* (2011), Detection of Paraphrastic Cases of Mono-lingual and Cross-lingual Plagiarism. In Proceedings of ICON-2011, Macmillan Publishers, India. http://users.dsic.upv.es/~prosso/resources/GuptaEtAl_ICON11.pdf
5. *Denisova-Schmidt, E.* (2016), Corruption in Russian Higher Education. Russian Analytical Digest, 191: 5–9.
6. *Khazov, A., Kuznetsova, M.,* (2017) Automatic Generation of Verbatim and Paraphrased Plagiarism Corpus. In press.
7. *Kicherova, M., Kyrov, D., Smykova, P., et al.* (2013), Plagiarism in Students' Papers: Toward the Roots of the Problem [Plagiat v studencheskikh rabotakh: analiz sushhnosti problemy]. Online journal Naukovedenie (IGUPIT), 4. http://naukovedenie.ru/PDF/83pvn413.pdf

8. *Modern Language Association* (2008), MLA Style Manual and Guide to Scholarly Publishing (3rd ed.). New York: Modern Language Association of America.
9. *Maloshonok, N.* (2016), How Perception of Academic Honesty at the University Is Linked with Student Engagement: Conceptualization and Empirical Research Opportunities [Kak vospriyatie akademicheskoj chestnosti sredy universiteta vzaimosvyazano so studencheskoj vovlechennost'yu: vozmozhnosti kontseptualizatsii i ehmpiricheskogo izucheniya]. Voprosy obrazovanija, 1: 35–60.
10. *Nikitov, A., et al.* (2012), Plagiarism in Under- and Postgraduate Students' Papers: Problem and Actions Against [Plagiat v rabotakh studentov i aspirantov: problema i metody protivodejstviya]. Universitetskoe upravlenie: praktika i analiz, 5: 61–68.
11. *Oakes, M.* (2014), Literary Detective Work on the Computer. Amsterdam: John Benjamins Publishing Company.
12. *Panchenko, A., Loukachevitch, N. V., Ustalov, D., Paperno, D., Meyer, C. M. and Konstantinova, N.,* (2015) RUSSE: The First Workshop on Russian Semantic Similarity, Computational Linguistics and Intellectual Technologies: papers from the Annual conference "Dialogue", vol. 2, pp. 89–105.
13. *Pivovarova L., Pronoza E., Yagunova E.* (2016) Shared Task on Sentence Paraphrase Detection for the Russian Language // http://www.paraphraser.ru/download/get?file_id=2
14. *Potthast, M., Barrón-Cedeño, A., Eiselt, A., Stein, B., Rosso, P.* (2010a), Overview of the 2nd International Competition on Plagiarism Detection. Martin Braschler and Donna Harman (Eds.): Notebook Papers of CLEF 2010 LABs and Workshops, 22–23 September, Padua, Italy.
15. *Potthast, M., Stein, B., Barrón-Cedeño, A., Rosso, P.* (2010b), An Evaluation Framework for Plagiarism Detection. In Proceedings of the 23rd international conference on computational linguistics: Posters: 997–1005. http://www.uni-weimar.de/medien/webis/publications/papers/stein_2010p.pdf
16. *Potthast, M., Hagen, M., Beyer, A., Busse, M., Tippmann, M., Rosso, P., Stein, B.* (2014), Overview of the 6th International Competition on Plagiarism Detection. In CEUR Workshop Proceedings, 1180: 845–876. https://www.uni-weimar.de/medien/webis/publications/papers/stein_2014k.pdf
17. *Sochenkov, I., Zubarev, D., Smirnov, I.* (2017) PARAPLAG: Russian Dataset for Paraphrased Plagiarism Detection. In press.
18. *Weber-Wulff, D.* (2014), False Feathers. A Perspective on Academic Plagiarism. http://link.springer.com/book/10.1007%2F978-3-642-39961-9.
19. *Zubarev, D., Sochenkov, I.* (2017) Paraphrased plagiarism detection using sentence similarity. In press.