Evaluation Tracks on Plagiarism Detection Algorithms for the Russian Language

Ivan Smirnov, Institute for Systems Analysis, FRC CSC RAS, Moscow, Russia
Rita Kuznetsova, Antiplagiat JSC, Moscow, Russia
Mikhail Kopotev, University of Helsinki, Finland
Andrey Khazov, Antiplagiat JSC, Moscow, Russia
Olga Lyashevskaya, Higher School of Economics, Moscow, Russia
Lyubov Ivanova, Higher School of Economics, Moscow, Russia
Andrey Kutuzov, University of Oslo, Norway

Motivation

- In recent research, 36 percent of respondents in Russia admitted to regularly copying the texts of others in different forms (Kicherova et al. 2013)
- In 2004, it was estimated that 10 percent of student works in the United States and Australia involved plagiarism
- Academic plagiarism is especially crucial problem (see GuttenPlag, Dissernet communities, etc.)
- There are several services that are able to detect plagiarism in Russian-language texts, but thus far there has been no systematic evaluation of these services

Related events

- RUSSE – the shared task on word-level semantic similarity (Russian)
- ParaPhrase – the shared task on sentence-level paraphrase detection (Russian)
- Semantic Textual Similarity at SemEval with the shared task on semantic equivalence between two snippets of text in English and some other languages (not for Russian)
- ROMIP (2003-2010) - Russian Information Retrieval Evaluation Seminar with Similar Documents Search Track and Adhoc search Track
- PAN workshop at CLEF (2009-2015) with shared tasks on Text Reuse Detection (aka Plagiarism Detection) (not for Russian)

Goals and tasks

- Propose methodology and create dataset for evaluation of plagiarism detection algorithms for the Russian language
- Organize evaluation of plagiarism detectors on PlagEvalRus workshop
- Tracks:
  - Track 1: Plagiarized sources retrieval. For each suspicious text provide a list of sources, sorted according to the number of reused fragments in descending order
  - Track 2: Copy and paste plagiarism detection. For a pair of texts, fragments taken from one text need to be found in a second text.
  - Track 3: Paraphrased plagiarism detection. For a pair of texts, fragments taken from one text need to be found in a second text.

Dataset

- Academic texts in Russian
- Source text – a text, from which fragments are supposedly reused
- Suspicious text – a text that supposedly contains fragments from source texts.
- Suspicious texts contain the following types of plagiarism:
  - Automatically generated copy&paste plagiarism
  - Automatically generated paraphrased plagiarism
  - Manually copy&paste plagiarism
  - Manually paraphrased plagiarism

Collection of sources

- The “potential sources” collection 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. Each text was stored in a separate file with a name containing a unique identifier.
- ~130Gb (~30Gb zipped)

Automatically generated plagiarism

- **Automatically generated copy and paste plagiarism**. Randomly selected sentences from a target text; each of them is replaced by one or more randomly chosen consecutive sentences from the texts, which did not belong to the target collection. The resulting target texts contain from 10 to 80 percent of plagiarized material (calculated in sentences).
- **Automatically generated paraphrased plagiarism**. This collection was created the same way as the copy-and-paste texts, except that sentences of the source texts were automatically paraphrased by using one or more of the following techniques:
  - Replacing words with their synonyms;
  - Adding and removing synonym chains;
  - Abbreviation and amplification;
  - Adding and removing diminutives;
  - Singular/plural replacement.

Manual 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.

Manually paraphrased plagiarism

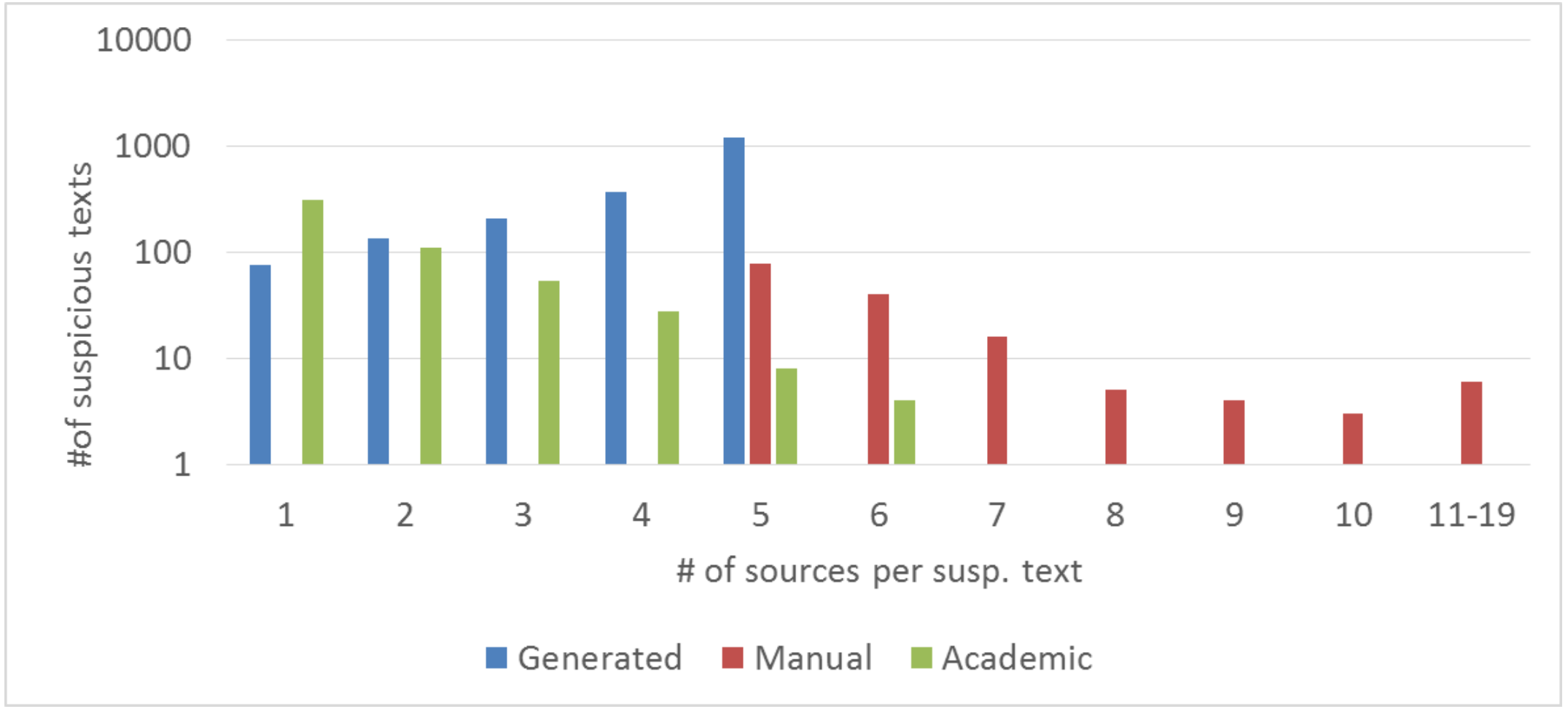
- This collection consists of the essays on a given unique topic
- Instruction for students: select texts from the source collection, mark the fragments (at least one sentence) and paraphrase them
- Rewriting techniques:
  - Deleting some words (about 20%) of the original sentence;
  - Adding some words (about 20%) into the original sentence;
  - Replacing some words or phrases of the original sentence with synonyms, changing word forms (number, case, form and verb tense, etc.) for some words (about 30%) in the original sentence;
  - Changing the order of words or clauses in the original sentence;
  - Concatenating two or more original sentences into one sentence;
  - Splitting the original sentence into two or more sentences (possibly with a change in the order they appear in the text);
  - Complex rewriting of the original sentence, which combines 3-5 or even more aforementioned techniques.

Текст эссе	Текст источника
В психиатрии под настроением понимается эмоциональное состояние, которое характеризуется сменой радости и печали в зависимости от различных обстоятельств.	Настроение – это эмоциональное состояние, характеризующееся сменой радости и печали в зависимости от обстоятельств.
Расстройства настроения проявляются в особой группе аффективных расстройств.	
Это клинические состояния, проявляющиеся в нарушении настроения, ощущении тяжелых эмоциональных страданий и неспособности управлять своими аффектами.	Аффективные расстройства представляют собой группу клинических состояний, характеризующихся нарушением настроения, потерей способности контролировать свои аффекты и субъективным ощущением тяжелых страданий.
Одним из расстройств настроения является биполярное аффективное расстройство.	К основным расстройствам настроения относятся депрессивные расстройства и биполярное аффективное расстройство.
С биполярным аффективным расстройством прежде всего ассоциируются резкие переходы от эмоционального подъема к эмоциональному упадку. Такое состояние ведет к серьезной, а иногда даже опасной, нестабильности настроения.	Перепады от эмоционального подъёма к эмоциональному упадку – эти крайние противоположности ассоциируются с биполярным расстройством, душевным заболеванием, которое характеризуется серьёзной, а порой даже опасной нестабильностью настроения.

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	1000	4257	5000	100	268
Automatically generated paraphrased plagiarism	2000	4251	5000	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

Texts suspected in plagiarizing N sources (Training set)



Evaluation Setup

- On Track 1, the participants downloaded the dataset and retrieved sources for suspicious texts using a system of the participant's own devising
- Tracks 2 and 3 were evaluated on TIRA – 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.
- PAN Baseline for Text Alignment is based on simple shingles approach with chunks of 50 symbols length

Evaluation Metrics for 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}

$$P = \frac{|T_{ret} \cap T_{src}|}{|T_{ret}|} \qquad R = \frac{|T_{ret} \cap T_{src}|}{|T_{src}|} \qquad F = \frac{2*R*P}{R+P}$$

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

$$AP(t) = \frac{1}{|K|} \sum_{k \in K} P@k \qquad MAP = \frac{1}{|T_{plg}|} \sum_{t_{plg} \in T_{plg}} AP(t_{plg})$$

Evaluation Metrics for Text Alignment

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.

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

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

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

$$granularity(S,R) = \frac{1}{|S_R|}\sum_{s\in S_R}|R_s| \qquad plagdet(S,R) = \frac{F_1}{\log_2(1+gran(S,R))}$$

PlagEvalRus-2017 Workshop

- 22 participants registered
- Dates
  - 16/12/2016 – Training set
  - 21/02/2017 – Evaluation set for Track 1
  - 27/02/2017 – Evaluation set for Track 2 and Track 3 in TIRA
  - 31/03/2017 – Deadline for submitting results
- Each participant was supposed to submit results for a maximum of 5 runs
- Only 1 participant submitted his runs

Evaluation results for Track 1: Plagiarized source detection

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

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

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

Evaluation results for Track 2: Copy and paste plagiarism detection

		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

Evaluation results for Track 3: Paraphrased plagiarism detection

Automatically-generated paraphrased plagiarism detection

		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**

Manually paraphrased plagiarism detection

		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**

Evaluation results for overall text alignment tasks

		Macro			Micro		
team.run	Granularity	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**

Results

- The methodology for evaluation of plagiarism detection algorithms in monolingual Russian texts is prepared and available in ready-to-use format.
- Datasets of different types of plagiarism is created
- Preliminary experimental results are received
- Text Alignment task is continuously available for evaluation on the TIRA site http://www.tira.io/tasks/pan/#text-alignment; the dataset “pan17-text-alignment-test-dataset-dialogue17-russian-2017-02-22”

Problems

- Preparation of manually paraphrased texts was the most time-consuming phase of the Workshop:
  - preparing one essay takes in average from 4 to 10 hours
  - students often make trash
  - essays should be automatically verified
- The decision to use TIRA was maybe incorrect, as participants had to invest time to study this evaluation framework
- Computational complexity and lack of both high-performance computing facilities and large-scale storage systems

Further advances

We plan:

- to enlarge collection of sources and increase the size of training datasets
- to announce joint plagiarism detection track, in which Source Retrieval and Text Alignment are not separated
- to announce cross-language (translated) plagiarism detection track
- to discuss refusing to TIRA as an evaluation platform

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)

Thanks!
Questions?

See https://plagevalrus.github.io/ for details and instructions on how to use the data