CrossLang: the system of cross-lingual plagiarism detection

Oleg Bakhteev
bakhteev@ap-team.ru
Moscow Institute of Physics and Technology,
Antiplagiat Company
Moscow, Russia

Alexandr Ogaltsov
ogaltsov@ap-team.ru
National Research University Higher School of Economics,
Antiplagiat Company
Moscow, Russia

Andrey Khazov
hazov@ap-team.ru
Antiplagiat Company
Moscow, Russia

Kamil Safin
safin@ap-team.ru
Moscow Institute of Physics and Technology,
Antiplagiat Company
Moscow, Russia

Rita Kuznetsova
kuz@zurich.ibm.com
Moscow Institute of Physics and Technology,
IBM Research
Rueschlikon, Switzerland

ABSTRACT

Plagiarism and text reuse become more available with the Internet development. Therefore it is important to check scientific papers for the fact of cheating, especially in Academia. Existing systems of plagiarism detection show the good performance and have a huge source databases. Thus now it is not enough just to copy the text "as is" from the source document to get the "original" work. Therefore, another type of plagiarism become popular — cross-lingual plagiarism. We present a CrossLang system for such kind of plagiarism detection for English-Russian language pair. The key idea for CrossLang system is that we use the monolingual approach. We have a suspicious Russian document and English reference collection. We reduce the task to one language — we translate the suspicious document into English with the help of machine translation system. After this step we perform the subsequent document analysis. There are two main stages at this analysis: source retrieval stage and document comparison stage. Both of these stages are adapted for our task. At source retrieval stage we need to find the most relevant documents from collection for a given translated suspicious document. Therefore the algorithm is based on aggregation of semantically close words into word classes and thus handles the cases of reformulated passages. The following document comparison is based on phrase embeddings that are trained in unsupervised and semi-supervised regimes. We evaluate CrossLang on the existing and generated datasets. We demonstrate the performance of the whole approach. We integrate the CrossLang in Antiplagiat system (most popular and well-known plagiarism detection system in Russia and CIS) and provide technical characteristics. We also provide the analysis of the system performance.

KDD 2019, August 2019, Anchorage, Alaska - USA

CCS CONCEPTS

• **Information systems** → **Information retrieval**; • **Computing methodologies** → *Natural language processing*; Learning settings; • **Applied computing** → Education.

KEYWORDS

Natural Language Processing; Cross-lingual plagiarism detection; Semantic clustering; Source Retrieval; Semi-supervised approach; Text Comparison

ACM Reference Format:
Oleg Bakhteev, Alexandr Ogaltsov, Andrey Khazov, Kamil Safin, and Rita Kuznetsova. 2019. CrossLang: the system of cross-lingual plagiarism detection. In *Proceedings of KDD 2019*. ACM, New York, NY, USA, 8 pages.

1 INTRODUCTION

Plagiarism detection and originality checking has become a major problem in Academia. Unauthorized text reuse often occurs in research papers [41] and even in Ph.D. theses [22]. There are several plagiarism detection systems (Turnitin, Antiplagiat.ru, Plagiarism.org, URKUND) that show good performance on verbatim plagiarism detection task. Possessing huge indexed collections of sources they detect copy-and-paste text reuse with high recall. Because of it another type of plagiarism becomes popular — when reused text was translated from another language [6, 7]. The translation can be both manual or automatic — modern machine translation systems could provide high quality text. Thus it is a very simple way to obtain "original" text without making any effort. There exist some articles described the problem of cross-lingual plagiarism detection for some language pairs [16–18]. Unfortunately none of described approaches are production-ready. On the other side, the existing industrial tools are also unable to detect such kind of plagiarism. Thus there is a need for such tool that allows us to solve this problem at industrial scale with high quality.

In this paper, we focus entirely on the case when unauthorized text reuse comes from English to Russian language. The problem is formulated as follows: given a suspicious Russian document and English reference collection. Suspicious document could possibly contain passages translated from some documents from the collection. The problem is to find all translated passages in the suspicious

document and their corresponding source passages in the documents from the collection. In general case, the language pair could be any. CrossLang is the new extension of the existing system for plagiarism detection — Antiplagiat [1], which is the most known system at Russia and CIS. Most universities in Russia and CIS use the Antiplagiat system for checking the originality of students works as well as publishing houses for scientific articles.

Antiplagiat text reuse detection engine detects text reuse in any language and for any popular file type. It has more than 300 million sources in search databases and 3 million users. Antiplagiat system performs a comparative analysis between a given suspicious text document and a large collection of source documents. At first it searches sources of the reused passages. After that it performs pairwise comparison between the sources and the suspicious document. The result of the analysis is a report — a list of text blocks found both in the document and texts in the corpus.

For reason that Antiplagiat checks a lot of documents in Russian, one of the languages in the pair is Russian. We choose the English as the second language for the following reasons:

(1) English prevails on the Internet and databases such as Web of Science, Scopus. In other words, English is the most reliable and common language for the scientific purposes.
(2) Based on the analysis of doctoral degree theses [22] from the Digital Library of RSL[2], it can be argued that plagiarism in a single language takes place in research papers. The existing tools detect text reuse within the confines of one language with high recall, so we suppose that there are growing cases of cross-lingual text reusing.
(3) High quality translation from English to Russian due to the progress of machine translation systems could also increase the number of such cases.

CrossLang is a service, consists of a set of microservices organized in five main components. Each of these microservices interacts with others via gRPC protocol. The microservise paradigm allows us to build more complex system — we can easily embed other microservices into the current solution if required. To analyze CrossLang performance we decompose the paper by following sections: in section (3) we introduce the design of the whole system and its components; in section (4) we provide two benchmarks to evaluate the CrossLang quality; we analyze the performance of the proposed system not only for the cross-lingual plagiarism detection task but also for the monolingual task when the suspicious document and source documents are written in the same language; in section (5) we analysed the production performance of the system because it has already deployed. Finally, in section (6) we discuss about architecture details. The experiment results show that CrossLang is comparable to other methods for either cross-lingual or monolingual plagiarism detection. Moreover the proposed system is designed for the high-performance document processing and therefore can be used as a production-ready solution.

2 RELATED WORK

Based on the fact that we did not find the tools for cross-lingual plagiarism detection task (none of plagiarism detection systems announced about that), we provide the research papers that are dedicated to this problem (or considering the related topics).

Similar methods proposed in [6, 29], where the suspicious document is translated into the language of collection using machine translation systems. In [7] IBM-1 model is used to obtain information about text similarity. The authors in [3, 14] propose methods based on the use of n-gram and term statistics. These approaches do not use machine translation directly, but try to obtain a translation of specific phrases and words using external resources. In contrast to them we use a collection of parallel corpora, which is available for many language pairs. Since growing number of papers devoted to unsupervised machine translation [4, 11, 24, 25] the proposed method can potentially work with any language pair. The use of additional resources, such as thesauruses and ontologies is a common practice for the cross-lingual plagiarism problem. In [17, 18, 20] the authors propose to use BabelNet [30] and WordNet [28] to obtain the information about texts similarity. Current state-of-the-art [16, 17], propose to construct semantic graph for each document. Text similarity evaluation is based on the similarity of the structures of these graphs. The main drawback of this approach is the resources requirement: the approach requires using multilingual ontologies, such as BabelNet [30], which cannot be used for commercial products.

Another class of papers similar to our approach is devoted to the document retrieval. In [31, 47] various methods of the document retrieval are compared. A number of works [12, 27] proposes to use paragraph or document vectors for this problem. One of the challenges of such methods is its computational expensiveness. In [9] authors propose to use approximate nearest neighbors method for fast document retrieval, which allows to retrieve documents faster at the cost of significant memory usage. A number of works [37–39] use methods for determining the text similarity like to latent semantic indexing [26], using decomposition of word-document matrix. This approach focuses on a significant text reuse, while our main task is to develop a system which can work with small-size text reuse cases.

As we use the monolingual approach the problem is very close to paraphrase detection task. Many approaches [23, 45, 53] have been developed for the paraphrase detection with neural sequence embedding. In [42, 45] authors propose to use recursive neural networks with dependency or constituency grammars. In [23, 51] authors propose to use long short-term memory (LSTM) and gated recurrent unit (GRU). For the cross-lingual paraphrase detection one can employ deep learning methods based on bilingual autoencoders [10, 54] or on siamese neural networks [53]. Opposing to works [21, 23, 43] we consider neural network outputs as embeddings in vector space for further approximate nearest neighbor search [49].

Our work deals with the task of cross lingual plagiarism for Russian-English language pair. The study of this pair is not common, to the best of our knowledge, the only paper devoted to this pair is our previous paper [6], where methods based on machine translation metrics were analyzed. In this paper we also present a dataset for the cross-lingual plagiarism detection. We believe that our impact will help to investigate new cross-lingual plagiarism detection methods for this language pair.

[1] https://www.antiplagiat.ru
[2] http://olden.rsl.ru/en

3 CROSSLANG DESIGN

In this section we want to describe the service architecture and its components. We also would like to point on the main differences of separate service components from the existing approaches, proposed in some papers.

The key idea for CrossLang system is that we use the monolingual approach. We have suspicious Russian document and English reference collection. We reduce the task to the one language — we translate the suspicious document into English, because the reference collection is in English. After this step we perform the subsequent document analysis. Due to this fact the main challenge with the CrossLang design is that the algorithms for the plagiarism detection task should be stable to the translation ambiguity. We have to deal with this problem twice. First, suspicious document in Russian *already* has the translated passages from English. Second, when we use the monolingual approach, we translate this document *again*. Since the translation ambiguity, here we have the situation which reminds so named "noisy channel model" [40] — the original English text passed through the noisy channel №1 and transformed into translated Russian text, which than passed through the noisy channel №2 and transformed into English text unlike the original. This "double noise" creates additional difficulties in our work (Figure 1).

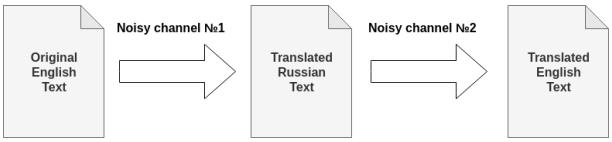

Figure 1: Double noisy channel model.

The main stages of CrossLang service is depicted in Figure 2. CrossLang receives the suspicious document from Antiplagiat system, when user send it for originality checking. Then it goes to *Entry point* — main service, that routes tha data between following stages:

(1) *Machine Translation system* — microservice, that translates suspicious document into English. For these purposes we use Transformer [48], open-source neural machine translation framework. For the details see section (3.1).
(2) *Source retrieval* — this stage unites two microservices: *Shingle index* and *Document storage*. Entry point receives the translated suspicious document's shingles (n -grams) and Shingle index returns to it the documents ids from the reference English collection. To deal with the translation ambiguity we use modified shingle-based approach. Document storage returns the Source texts from the collection by these ids. For the details see section (3.2).
(3) *Document comparison* — this microservice performs the comparison between translated suspicious document and source documents. We compare not the texts themselves, but the vectors corresponding to the phrases of these texts. Thus we deal with the translation ambiguity problem. For the details see section (3.3). After this stage the plagiarism report is formed and sent to user. Figure 3 is an example of a report. The marked passages correspond to reused text in the suspicious document. The elements on the right side of the screenshot gives brief information about the percentage of the reused text and the source documents.

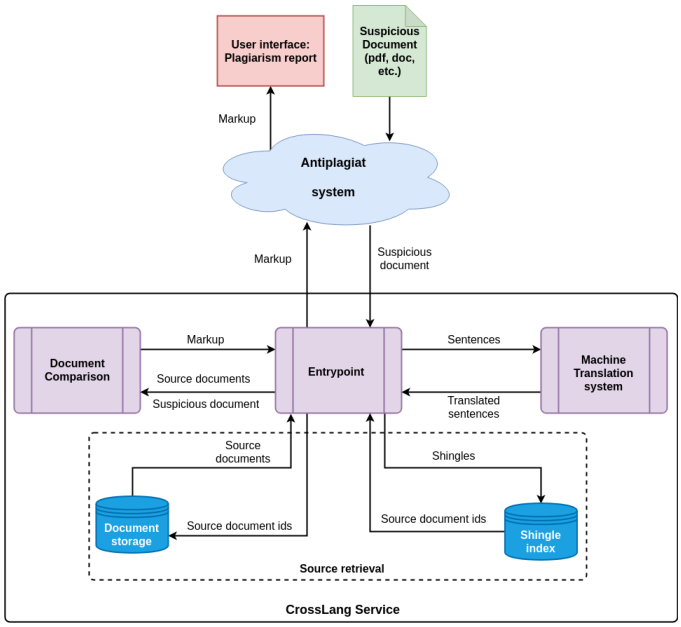

Figure 2: CrossLang service design.

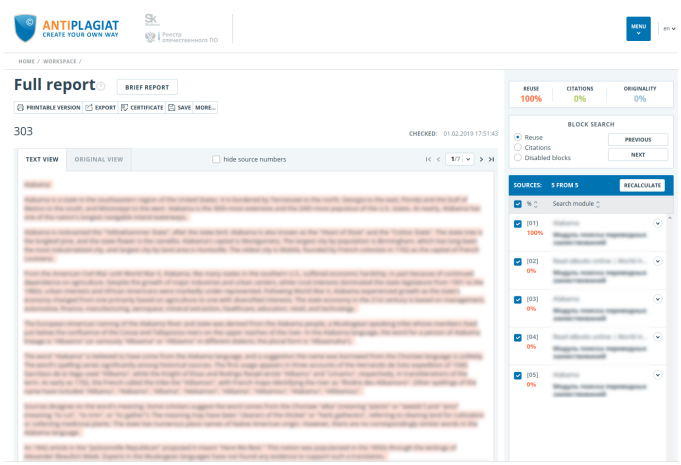

Figure 3: Plagiarism report.

Also in this section we would like to highlight the main differences of our work:

- The best of our knowledge it is the first system for cross-lingual plagiarism detection for English-Russian language pair. It is deployed on production and we could analyze the results. We could not find another examples of such system (even for other language pairs).

- The Source retrieval stage is often employed using rather simple heuristical algorithms such as shingle-based search [33, 47] or keyword extraction [13, 31] because of simplicity of such methods and their computational efficiency. However, these methods can significantly suffer from word replacements and usually detect only near-duplicate paraphrase. We employ modified shingle-based method for this stage. In order to handle translation ambiguity we clusterize the words using word embedding model. We use semantic classes instead of words during this stage.
- Many articles on the cross-lingual plagiarism detection topic investigate the solutions based on bilingual or monolingual word embeddings [15, 17] for documents comparison, but almost none of them uses the phrase embeddings for this problem solution.

In the next sections we introduce how the main stages work.

3.1 Machine Translation system

We create machine translation system using state-of-the-art Transformer algorithm [48]. We utilize Tensorflow realization [3] of it. Training dataset consists of approximately 30M parallel sentences. They were obtained from open-source parallel OPUS [46] corpora, but also we mine parallel sentences from Common Crawl.[4] Algorithm was trained for 5 epochs with batch size equals to 128 on Amazon p2.xlarge instance[5] We evaluate BLEU score [34] for *Russian* → *English* translation on news test 2018 dataset [6] and compare it with Google translator via API [7]. Results are in Table 1.

Table 1: BLEU of different systems

System	BLEU
Google	31.34
CrossLang Transformer	28.18

The CrossLang BLEU score lower than Google's BLEU score — this was to be expected. But it is very important to notice that we are not interested in ideal translation. Our main goal is to translate with sufficient quality for the next stages: Source retrieval and Document comparison.

3.2 Source retrieval

The method of source retrieval in the case of verbatim plagiarism is inverted index construction,where a document from the reference collection is represented as a set of its shingles, i.e. overlapping word n -grams, and a suspicious document's shingles are checked for matches with the indexed documents. The collection documents is subsequently ranked according to a selected pairwise document similarity measure that correlates with the number of shingles they have in common with the suspicious document. There is one major problem with using the standard shingles — in our case the machine translation stage generates texts that differ too much from the sources of plagiarism. We argue that the source retrieval task can be solved with the help of a similar method that performs better than the method mentioned above; this improvement is achieved by moving from word shingles to word-class shingles, where each word is substituted by the label of the class it belongs to:

$$\{\text{word}_1, \ldots, \text{word}_n\} \to \{\text{class}(\text{word}_1), \ldots, \text{class}(\text{word}_n)\}.$$

Those classes may be obtained in several ways, with the core idea being their semantic unity. If the classes follow this assumption, two semantically close phrases that do not share a words may map into a single shingle as long as two sequences of the word-class labels match. We also remove stop words and sort the words in each n-gram to take into account possible phrases differences. In our experiments, we examine word embeddings as source of word classes. Distributional semantic models are known to provide word vector representations that can be used to estimate pairwise semantic similarity of words at cosine similarity of their corresponding vectors. Clustering the vectors is thus a convenient and relatively fast way of obtaining semantic word classes. The examples of resulting classes are provided below:

- *[beer, beers, brewing, ale, brew, brewery, pint, stout, guinness, ipa, brewed, lager, ales, brews, pints, cask]*
- *[survey, assessment, evaluation, evaluate, examine, assess, surveys, analyze, evaluating, assessments, examining, analyzing, assessing, questionnaire, evaluations, analyse, questionnaires, analysing]*
- *[brilliant, excellent, exceptional, finest, outstanding, super, terrific]*

For the word embedding model we used `fastText` [8] trained on English Wikipedia. The dimension for word embedding model was set to 100. For the semantic word classes construction we applied agglomerative clustering on word embeddings with the cosine similarity measure to group words into word classes. We got 777K words clustered into 30K classes.

3.3 Document Comparison

For the comparison between retrieved documents and translated suspicious documents we introduce the phrase embedding model. Since in the final plagiarism report we must highlight phrases, we need to compare separate text fragments. We split documents (retrieved and suspicious) into phrases s and compare its vectors. Our goal is to learn representations for variable-sized phrases. For this purpose we learn a mapping: $s \to \mathbf{s}$, where $s = (\text{word}_1, \ldots, \text{word}_n)$. We learn this mapping both in unsupervised and semi-supervised training regimes. For mapping the word sequence into low dimensional space we use the encoder-decoder scheme. An encoder learns a vector representation of the input phrase and the decoder uses this representation to reconstruct the phrase in reverse order. During the training error between input phrase and reconstructed output phrase is minimized.

$$E_{rec} = \| \mathbf{s} - \hat{\mathbf{s}} \|^2 . \quad (1)$$

Encoder-decoder model is completely unsupervised and does not use any information whether the phrase pair is paraphrased or not. We train Seq2Seq model with attention [5]. As initial word vector representations for word_i we used word vectors from `fastText`

[3] https://tensorflow.github.io/tensor2tensor/
[4] http://commoncrawl.org/
[5] https://aws.amazon.com/ec2/instance-types/p2/
[6] http://www.statmt.org/wmt18/translation-task.html
[7] https://cloud.google.com/translate/docs/

model. For reconstruction error minimization E_{rec} (1) 10M sentences from Wikipedia articles were used.

In order to use information about phrase similarity we extend the objective function. We employ the margin-base loss from [51] with the limited number of similar phrase pairs $\mathcal{S} = \{(s_i, s_j)\}$:

$$E_{me} = \frac{1}{|\mathcal{S}|}\left(\sum_{(s_i,s_j)\in\mathcal{S}} \max(0, \delta - c_-) + \max(0, \delta - c_+)\right), \quad (2)$$

where $c_- = \cos(\mathbf{s}_i, \mathbf{s}_j) - \cos(\mathbf{s}_i, \mathbf{s}_{i'})$, $c_+ = \cos(\mathbf{s}_i, \mathbf{s}_j) + \cos(\mathbf{s}_j, \mathbf{s}_{j'})$, δ is the margin,
$\mathbf{s}_{i'} = \arg\max_{\mathbf{s}_{i'}\in\mathcal{S}_b\setminus(\mathbf{s}_i,\mathbf{s}_j)} \cos(\mathbf{s}_i, \mathbf{s}_{i'})$, $\mathcal{S}_b \in \mathcal{S}$ — current mini-batch.

The sampling of so named "false neighbour" $\mathbf{s}_{i'}$ during training helps to improve the final quality without strict limitations on what phrases we should use at dissimilar.

This part of objective requires a dataset of similar sentences $\mathcal{S} = \{(s_i, s_j)\}$. We used double translation method [52] as a method of similar sentences generation comparable to paraphrase. Consider a parallel corpus with pairs of Russian and English sentences. We translate Russian sentences back to English. This method of generation allows us to obtain pairs of sentences we process in CrossLang: both during training and during system usage we process a pairs of English sentences sentences translated from Russian into English by our translation system. We believe that this method gives us the opportunity to make phrase embedding model robust to our translation system errors since the machine translation errors can significantly influence the total performance of our framework. We used 100K pairs of sentences from OpenSubtitles [46] corpus. The examples of resulting pairs (original sentence and doubly translated sentence) are listed below:

- *You know, I remember you pitched me the idea for this thing five years ago.*
- *I remember you pitched me the idea for this to the cause of 5 years ago.*

The final objective function is:

$$\alpha E_{rec} + (1 - \alpha)E_{me}, \quad (3)$$

where α is a tunable hyperparameter that weights both of errors. The dimensions for word embedding model and phrase embedding model[8] were set to 100.

For each phrase embedding from the suspicious document find M nearest vectors by cosine similarity from source documents using *Annoy*[9] library. The main idea of this function is to reduce the number of fragments pairs with a simple decision rule: for phrase embeddings pairs $(\mathbf{s}_i, \mathbf{s}_j)$ we consider that it is the plagiarism case if $\cos(\mathbf{s}_i, \mathbf{s}_j) > t_1$, where t_1 is a cosine measure threshold[10].

4 EXPERIMENTS

In this section we provide the experiment settings and results. As there are no results for cross-lingual plagiarism detection task for language pair English-Russian, we perform the two following benchmarks:

(1) We propose a dataset for plagiarism detection tool evaluation in the case of Russian-English language pair. We describe the generation method and provide the link on the dataset. We evaluate the general framework performance on it.
(2) The crosslingual plagiarism detection task in our case is very close to paraphrase detection task because we use the machine translation methods. So we can use the datasets for evaluation plagiarism detection algorithms in the confines of one language. These datasets consist different type of *obfuscation*, i.e. paraphrasing. We evaluate our system without machine translation stage at this experiment (for this we specifically evaluated the quality of machine translation system separately at (3.1)), because we work with English data. We argue that this settings are very close to our situation, because we also work with *modified* text. It is important to notice than modification after translation and after paraphrasing are similar.

4.1 Metrics

We use plagiarism detection metrics proposed in [35]. Detailed definition of the metrics is follows. Let S be a set of plagiarism cases and R be a set of cases that were detected by algorithm. Let's also define s, r such that $s \in S$ and $r \in R$. Therefore:

$$Prec(S, R) = \frac{1}{|R|}\sum_{r\in R}\frac{\left|\bigcup_{s\in S}(s \cap r)\right|}{|r|}, \quad (4)$$

$$Rec(S, R) = \frac{1}{|S|}\sum_{s\in S}\frac{\left|\bigcup_{r\in R}(s \cap r)\right|}{|s|}, \quad (5)$$

$$F(S, R) = \frac{2 \times Prec(S, R) \times Rec(S, R)}{Prec(S, R) + Rec(S, R)}, \quad (6)$$

Following [35] we define a granularity of R for given S by average size of existing covers:

$$Gran(S, R) = \frac{1}{|S_R|}\sum_{s\in S}|C_s|, \quad (7)$$

where $S_R = \{s | s \in S \wedge \exists r \in R : s \cap r \neq \emptyset\}$ and $C_s = \{r | r \in R \wedge s \cap r \neq \emptyset\}$. Overall metric is called *Plagdet* and defined by combination of above:

$$Plagdet(S, R) = \frac{F(S, R)}{log_2(1 + Gran(S, R))}. \quad (8)$$

4.2 Experiment on synthesized collection

For this experiment with synthesized collection we used Russian and English Wikipedia. Before the dataset generation we preliminary analyzed the behavior of our system for the monolingual verbatim plagiarism detection. The most part of the documents with detected text reuse used up to 10 collection documents as text reuse sources. We used parameters similar to parameters of PAN dataset generation: we syntesized documents with percentage of plagiarism between 20 to 80% and with number of souce documents from 1 to 10. We believe these parameters are rather natural and close to the real-world text reuse cases. Since the most part of the

[8] For the phrase embedding model we used AdaDelta with parameters: $\epsilon = 10^{-6}$, $\mu = 0.95$. We used L2-regularization with parameter $\lambda_2 = 10^{-6}$. The objective (3) had the following value: $\alpha = 0.1$.. In the objective (2) $\delta = 0.3$
[9] https://github.com/spotify/annoy
[10] We set $t_1 = 0.6$

documents processed by our system are student works we focus on the documents with quite small size. For the monolingual plagiarism detection it's quite often case when the checked document is a practically copy-paste of another work except title, author name and introduction part. Therefore we synthesize documents with plagiarism percentage up to 80%, which is rather high. On the other hand we also would like to detect cases when the text reuse not so high, therefore we also synthesize documents with rather low plagiarism percentage.

As a reference English collection D we used 100K articles from English Wikipedia. As a collection of suspicious documents D_{susp} we used a random sample of documents from Russian Wikipedia. For each document $d^i_{susp} \in D_{\text{susp}}$ we did the following:

(1) Select source documents $\{d^i\}$ from the collection. In order to have a similar topic for the suspicious document and source documents we use the following scheme. We translate suspicious document into English and find a subsample of 500 most relevant collection documents by *tf· idf*. After that we randomly choose from 1 to 10 documents from this subsample.
(2) Pick sentences randomly from source documents $\{d^i\}$ and translate them into Russian.
(3) Replace random sentences from document d^i_{susp} by the translated sentences from source documents. For each document from D_{susp} we replaced from 20 to 80% sentences.

The whole dataset with PAN-format markup can be found at[11].

We conducted the experiment the whole framework, which allows to assess the general performance. For the whole framework we got Precision = 0.83, Recall = 0.79 and $F1 = 0.80$.

4.3 Monolingual plagiarism detection

Since our system translates the suspicious document into the language of the document collection it's quite natural to analyze the performance of our system not only for cross-lingual plagiarism detection problem but also for monolingual problem. For such experiment we do not use the machine translation service. In order to check performance of monolingual plagiarism detection we exploit PAN'11 [36]. This corpora is suitable for us as a test because of various plagiarism cases with different obfuscation level. Obfuscation distribution is in Table 2. For more information of approaches to plagiarism obfuscation see [35]. Dataset consists of collection with

Table 2: Obfuscation distribution

Type	% of plagiarism cases
No obfuscation	2
Low	50
High	48

suspicious documents (approx. 11000) that we need to check on plagiarism and collection with reference documents (also approx. 11000) from those one can potentially plagiarize.

Since we use PAN'11 corpora, it is naturally to compare algorithm performance with PAN'11 participants and other works that were tested on this corpora. Results of CrossLang and top-5 known previous methods are in Table 3.

[11] http://tiny.cc/cl_ru_en

Table 3: PAN'11 performance comparison

Model	P	R	F	Plagdet
CrossLang	**0.94**	**0.76**	**0.84**	**0.83**
PDLK [2]	0.90	0.70	0.79	0.79
Sys-1 [50]	0.86	0.69	0.76	0.75
Sys-2 [19]	0.75	0.66	0.7	0.69
Sys-3 [44]	0.89	0.55	0.68	0.68
Sys-4 [32]	0.87	0.56	0.68	0.67

5 PRODUCTION PERFORMANCE

Our service was successfully deployed and connected to Antiplagiat system. We analyzed the performance of the service from May to July 2018. During this period students in Russia take exams and the average load on the system increases. For the production version of our service we indexed 30M documents from the Internet in addition to Wikipedia and arxiv we used earlier.

There were about 1.5M text reuse check in this period. We analyzed the statistics of document checks and found that 467K documents were detected as documents containing text reuse, which is about one of a third of all checked documents. However only a small part of documents contained significant reuse: we had about 70K document checks that contained more than 5% of cross-lingual text reuse. The median of text reuse level is 8.94 for such documents. The distribution of plagiarism in these checks is illustrated in Figure 4. The distribution of sources retrieved for the suspicious documents is illustrated in Figure 5.

A brief analysis of the production performance showed the following:

(1) The developed system successfully copes with a large load, which indicates the effectiveness of the proposed method.
(2) A significant part of the works contains cross-lingual text reuse from the English language. Despite some false positive cases, we found that the system works quite correctly. Unfortunately, there are some cases when the system detects fairly general phrases (for example, introductory phrases like "This work represents").
(3) About 5% of documents contain a significant amount of cross-lingual text reuse. This number is preliminary and requires further analysis: for many sources of text reuse there can be found versions already translated into Russian, therefore often the student uses already translated. Nevertheless, we believe that this number is quite close to real: with the development of systems that allow to find monolingual text reuse, the share of cross-language text reuse should increase.

The findings suggest the applicability of the developed system and determine the possible directions for further development of the system: filtering common phrases and determining of a language of text reuse source.

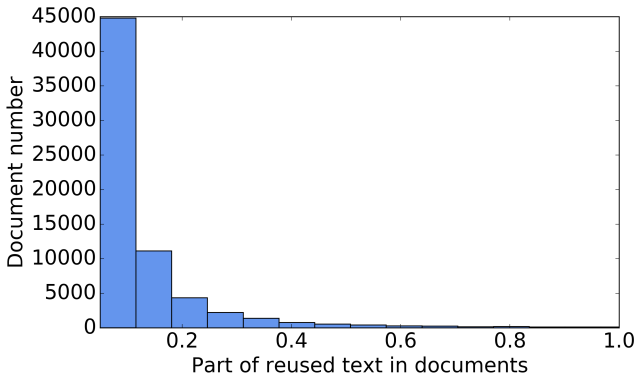

Figure 4: Histogram of percentage text reuse in the real documents. We analyzed only the documents contained more than 5% percent of text reuse.

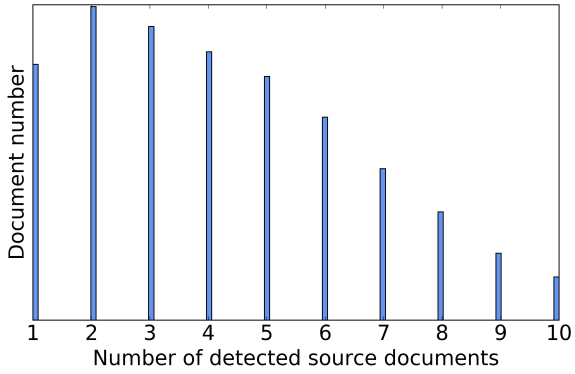

Figure 5: Histogram of source number retrieved for the documents contained more than 5% percent of text reuse.

6 ARCHITECTURE

In this section we briefly describe how the microservices are deployed in our system. Our main technical requirement for the system is the document check speed and an ability to scale with the number of simultaneous document checks. Our microservices are stateless, i.e. they treat all the operations as independent. This allows us to easily replace microservice backends and make the architecture more flexible. Currently we use RocksDB[12] for the Shingle index and Document storage services and Tensorflow [1] for the Machine translation and Document comparison services.

Our service is deployable on an 8-GPU cluster with Tesla-K100 GPUs, 128GB RAM and 64 CPU Cores. Depending on the requirements, the service is able to scale horizontally. For the fast rescaling we use Docker containerization and Consul and Consul-template for the service discovery and automatic load balancing.

The stress testing of our system showed that the system is able to check up to 100 documents in a minute. Despite the fact the average loading on our service is much lower, this characteristic of our service is important for withstanding peak loads.

[12]https://github.com/facebook/rocksdb

7 CONCLUSION

We introduced CrossLang — a framework for cross-lingual plagiarism detection for English Russian language pair. We decomposed the problem of cross-lingual plagiarism detection into several stages and provide a service, consists of a set of microservices. The CrossLang use a monolingual approach — reducing the problem to the one language. For this purpose we trained the neural machine translation system. Another two main algoithmic components are Source Retrieval and Document Comparison stages. For the Source Retrieval problem we used a modification of shingling method that allow us to deal with ambiguity after translation. For the Document Comparison stage we used phrase embeddings that were trained with slight supervision. We proposed method to make documents comparison efficient using approximate nearest neighbors method. We evaluated the effectiveness of our approach on several datasets. We also provided our own dataset. We integrated CrossLang in Antiplagiat system — the most popular and well-known system for plagiarism detection in Russia and CIS and analyzed the real system performance.

In future, we are going to develop the approach in several directions — use the documents vectors instead of shingles in source retrieval stage and modify our phrase embedding model. Also we will monitor system performance and analyze real users documents. We would like to conduct more experiments on the samples of real-world cases as long as corresponding data is available. Since our approach is rather general and does not use any language-specific features we believe that it can be applied to other language pairs. Therefore one of our plans is to implement our approach for language pairs other than English-Russian.

ACKNOWLEDGMENTS

The authors would like to thank Yury Chehovich for his advices and contribution to the development of the CrossLang system. This work was supported by RFBR project No.18-07-01441 and FASIE project No.44116.

REFERENCES

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. https://www.tensorflow.org/ Software available from tensorflow.org.

[2] Asad Abdi, Norisma Idris, Rasim Alguliyev, and Ramiz Aliguliyev. 2015. PDLK: Plagiarism Detection using Linguistic Knowledge. *Expert Systems with Applications* (07 2015). https://doi.org/10.1016/j.eswa.2015.07.048

[3] Zaid Alaa, Sabrina Tiun, and Mohammedhasan Abdulameer. 2016. Cross-language plagiarism of Arabic-English documents using linear logistic regression. *Journal of Theoretical and Applied Information Technology* 83, 1 (2016), 20.

[4] Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised neural machine translation. In *Proceedings of the Sixth International Conference on Learning Representations*.

[5] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).

[6] Oleg Bakhteev, Rita Kuznetsova, Alexey Romanov, and Anton Khritankov. 2015. A monolingual approach to detection of text reuse in Russian-English collection. In *Artificial Intelligence and Natural Language and Information Extraction, Social Media and Web Search FRUCT Conference (AINL-ISMW FRUCT), 2015.* IEEE, 3–10.

[7] Alberto Barrón-Cedeno, Paolo Rosso, Eneko Agirre, and Gorka Labaka. 2010. Plagiarism detection across distant language pairs. In *Proceedings of the 23rd International Conference on Computational Linguistics*. Association for Computational Linguistics, 37–45.

[8] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching Word Vectors with Subword Information. *arXiv preprint arXiv:1607.04606* (2016).

[9] Leonid Boytsov, David Novak, Yury Malkov, and Eric Nyberg. 2016. Off the Beaten Path: Let's Replace Term-Based Retrieval with k-NN Search. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. ACM, 1099–1108.

[10] Sarath Chandar, Stanislas Lauly, Hugo Larochelle, Mitesh Khapra, Balaraman Ravindran, Vikas C Raykar, and Amrita Saha. 2014. An autoencoder approach to learning bilingual word representations. In *Advances in Neural Information Processing Systems*. 1853–1861.

[11] Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word Translation Without Parallel Data. In *International Conference on Learning Representations (ICLR)*.

[12] Andrew M Dai, Christopher Olah, and Quoc V Le. 2015. Document embedding with paragraph vectors. *arXiv preprint arXiv:1507.07998* (2015).

[13] Sandipan Dutta and Debotosh Bhattacharjee. 2014. Plagiarism Detection by Identifying the Keywords. In *Computational Intelligence and Communication Networks (CICN), 2014 International Conference on*. IEEE, 703–707.

[14] Nava Ehsan, Frank Wm. Tompa, and Azadeh Shakery. 2016. Using a Dictionary and N-gram Alignment to Improve Fine-grained Cross-Language Plagiarism Detection. In *Proceedings of the 2016 ACM Symposium on Document Engineering (DocEng '16)*. ACM, New York, NY, USA, 59–68. https://doi.org/10.1145/2960811.2960817

[15] Jérémy Ferrero, Frédéric Agnès, Laurent Besacier, and Didier Schwab. 2017. Using Word Embedding for Cross-Language Plagiarism Detection. In *EACL 2017*, Vol. 2. 415–421.

[16] Marc Franco-Salvador, Parth Gupta, and Paolo Rosso. 2013. Cross-language plagiarism detection using a multilingual semantic network. In *European Conference on Information Retrieval*. Springer, 710–713.

[17] Marc Franco-Salvador, Parth Gupta, Paolo Rosso, and Rafael E Banchs. 2016. Cross-language plagiarism detection over continuous-space-and knowledge graph-based representations of language. *Knowledge-Based Systems* 111 (2016), 87–99.

[18] Marc Franco-Salvador, Paolo Rosso, and Manuel Montes-y Gómez. 2016. A systematic study of knowledge graph analysis for cross-language plagiarism detection. *Information Processing & Management* 52, 4 (2016), 550–570.

[19] Cristian Grozea, Christian Gehl, and Marius Popescu. 2009. ENCOPLOT: Pairwise sequence matching in linear time applied to plagiarism detection. *3rd PAN Workshop. Uncovering Plagiarism, Authorship and Social Software Misuse* 502 (01 2009), 10.

[20] Ezzikouri Hanane, Mohammed Erritali, and Mohamed Oukessou. 2016. Semantic Similarity/Relatedness for Cross language plagiarism detection. In *Computer Graphics, Imaging and Visualization (CGiV), 2016 13th International Conference on*. IEEE, 372–374.

[21] Hua He, Kevin Gimpel, and Jimmy J. Lin. 2015. Multi-Perspective Sentence Similarity Modeling with Convolutional Neural Networks.. In *EMNLP*, LluÃŋs MÃărquez, Chris Callison-Burch, Jian Su, Daniele Pighin, and Yuval Marton (Eds.). The Association for Computational Linguistics, 1576–1586.

[22] Anton S Khritankov, Pavel V Botov, Nikolay S Surovenko, Sergey V Tsarkov, Dmitriy V Viuchnov, and Yuri V Chekhovich. 2015. Discovering text reuse in large collections of documents: A study of theses in history sciences. In *Artificial Intelligence and Natural Language and Information Extraction, Social Media and Web Search FRUCT Conference (AINL-ISMW FRUCT), 2015*. IEEE, 26–32.

[23] Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in neural information processing systems*. 3294–3302.

[24] Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations (ICLR)*.

[25] Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, et al. 2018. Phrase-Based & Neural Unsupervised Machine Translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 5039–5049.

[26] Thomas K Landauer and Susan Dumais. 2008. Latent semantic analysis. *Scholarpedia* 3, 11 (2008), 4356.

[27] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*. 1188–1196.

[28] George A Miller. 1995. WordNet: a lexical database for English. *Commun. ACM* 38, 11 (1995), 39–41.

[29] Markus Muhr, Roman Kern, Mario Zechner, and Michael Granitzer. 2010. External and intrinsic plagiarism detection using a cross-lingual retrieval and segmentation system. In *Notebook Papers of CLEF 2010 LABs and Workshops*.

[30] Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th annual meeting of the association for computational linguistics*. Association for Computational Linguistics, 216–225.

[31] Hui Ning, Leilei Kong, Mingxing Wang, Cuixia Du, and Haoliang Qi. 2015. Comparisons of keyphrase extraction methods in source retrieval of plagiarism detection. In *Computer Science and Network Technology (ICCSNT), 2015 4th International Conference on*, Vol. 1. IEEE, 661–664.

[32] Gabriel Oberreuter, SebastiÃąn A. RÃąos, and Juan D. VelÃąsquez. [n. d.]. FASTDOCODE: Finding Approximated Segments of N-Grams for Document Copy Detection Lab Report for PAN at CLEF 2010.

[33] Ahmed Hamza Osman, Naomie Salim, Yogan Jaya Kumar, and Albaraa Abuobieda. 2012. Fuzzy Semantic Plagiarism Detection. In *International Conference on Advanced Machine Learning Technologies and Applications*. Springer, 543–553.

[34] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 311–318.

[35] Martin Potthast, Alberto Barrón-Cedeño, Andreas Eiselt, Benno Stein, and Paolo Rosso. 2009. Overview of the 1st international competition on plagiarism detection. In *CLEF*.

[36] Martin Potthast, Andreas Eiselt, Alberto BarrÃşn-cedeÃśo, Benno Stein, and Paolo Rosso. [n. d.]. Overview of the 3rd international competition on plagiarism detection. In *In Working Notes Papers of the CLEF 2011 Evaluation*.

[37] Martin Potthast, Benno Stein, and Maik Anderka. 2008. A Wikipedia-based multilingual retrieval model. *Advances in Information Retrieval* (2008), 522–530.

[38] Anak Agung Putri Ratna, F. Astha Ekadiyanto, Mardiyah, Prima Dewi Purnamasari, and Muhammad Salman. 2016. Analysis on the Effect of Term-Document's Matrix to the Accuracy of Latent-Semantic-Analysis-Based Cross-Language Plagiarism Detection. In *Proceedings of the Fifth International Conference on Network, Communication and Computing (ICNCC '16)*. ACM, New York, NY, USA, 78–82. https://doi.org/10.1145/3033288.3033300

[39] Anak Agung Putri Ratna, Prima Dewi Purnamasari, Boma Anantasatya Adhi, F Astha Ekadiyanto, Muhammad Salman, Mardiyah Mardiyah, and Darien Jonathan Winata. 2017. Cross-Language Plagiarism Detection System Using Latent Semantic Analysis and Learning Vector Quantization. *Algorithms* 10, 2 (2017), 69.

[40] Claude Elwood Shannon. 1948. A mathematical theory of communication. *Bell system technical journal* 27, 3 (1948), 379–423.

[41] I. Smirnov, R. Kuznetsova, M. Kopotev, A. Khazov, O. Lyashevskaya, L. Ivanova, and A. Kutuzov. 2017. Evaluation Tracks on Plagiarism Detection Algorithms for the Russian Language. In *Proceedings on the âĂIJComputational Linguistics and Intellectual TechnologiesâĂİ*.

[42] Richard Socher, Andrej Karpathy, Quoc V. Le, Christopher D. Manning, and Andrew Y. Ng. 2014. Grounded Compositional Semantics for Finding and Describing Images with Sentences. *TACL* 2 (2014), 207–218. https://tacl2013.cs.columbia.edu/ojs/index.php/tacl/article/view/325

[43] Richard Socher, Jeffrey Pennington, Eric H. Huang, Andrew Y. Ng, and Christopher D. Manning. 2011. Semi-supervised Recursive Autoencoders for Predicting Sentiment Distributions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '11)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 151–161. http://dl.acm.org/citation.cfm?id=2145432.2145450

[44] Åăimon Suchomel, Jan Kasprzak, and Michal Brandejs. 2012. Three Way Search Engine Queries with Multi-feature Document Comparison for Plagiarism Detection.

[45] Kai Sheng Tai, Richard Socher, and Christopher D. Manning. 2015. Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks. *CoRR* abs/1503.00075 (2015). http://arxiv.org/abs/1503.00075

[46] JÃűrg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)* (23-25), Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association (ELRA), Istanbul, Turkey.

[47] Serhii Vashchilin and Halyna Kushnir. 2017. Comparison plagiarism search algorithms implementations. In *Advanced Information and Communication Technologies (AICT), 2017 2nd International Conference on*. IEEE, 97–100.

[48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 5998–6008. http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf

[49] Jingdong Wang, Heng Tao Shen, Jingkuan Song, and Jianqiu Ji. 2014. Hashing for similarity search: A survey. *arXiv preprint arXiv:1408.2927* (2014).

[50] Shuai Wang, Haoliang Qi, Leilei Kong, and Cuixia Nu. 2013. Combination of VSM and Jaccard coefficient for external plagiarism detection. In *2013 International Conference on Machine Learning and Cybernetics*, Vol. 04. 1880–1885. https://doi.org/10.1109/ICMLC.2013.6890902

[51] John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. Towards Universal Paraphrastic Sentence Embeddings. *CoRR* abs/1511.08198 (2015). http://arxiv.org/abs/1511.08198

[52] John Wieting, Jonathan Mallinson, and Kevin Gimpel. 2017. Learning paraphrastic sentence embeddings from back-translated bitext. *arXiv preprint arXiv:1706.01847* (2017).

[53] Wen-tau Yih, Kristina Toutanova, John C. Platt, and Christopher Meek. 2011. Learning Discriminative Projections for Text Similarity Measures. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning (CoNLL '11)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 247–256. http://dl.acm.org/citation.cfm?id=2018936.2018965

[54] Biao Zhang, Deyi Xiong, and Jinsong Su. 2017. BattRAE: Bidimensional Attention-Based Recursive Autoencoders for Learning Bilingual Phrase Embeddings. In *Proc. of AAAI*.