A Monolingual Approach to Detection of Text Reuse in Russian-English Collection

Oleg Bakhteev, Rita Kuznetsova, Alexey Romanov, Anton Khritankov
Antiplagiat JSC
Moscow, Russia
bahteev@ap-team.ru

***Abstract*—In this paper we develop a method for cross-lingual (Russian and English) text reuse detection. The method is based on the monolingual approach — translation of texts into one language and reduction to the text similarity problem. We split texts into non-overlapping fragments and compare fragments to each other by means of different metrics — BLEU(1-2), METEOR, cosine similarity between bag-of-words representations of each snippet, and cosine similarity between vectors obtained from doc2vec-trained model. We explore the impact of choice of metric on the quality of text reuse detection. We assess quality of the method on a sample of a hundred scientific documents, originally in Russian, machine translated into English. Preliminary findings demonstrate feasibility of the approach.**

I. Introduction

Since wide adoption of digital documents and word processors, unauthorized text reuse and cheating has become a major problem for educational institutions as well as many independent authors. Experience with detecting text reuse we obtain from our daily work shows that a large proportion of texts contain reused fragments from another language. Although exact figures are unknown, we have seen many such cases "in the wild".

The problem of cross-language text reuse detection and similarity evaluation is not new, it has been studied for several years already [1], and there is a tool and an algorithm competition at PAN, held annually at the CLEF conference [2].

In this paper we explore a problem of detecting text reuse from English to the Russian language, which we see most often. Although the problem of cross-lingual similarity is rather well-known (see section II for details), the problem of cross-lingual textual similarity in the case of Russian being one of the languages in a pair, is poorly known. The main idea of the paper is to present a rather simple but efficient method of cross-lingual textual similarity detection for Russian-English collection, that could be used further as a baseline for other methods. We adopted the monolingual approach [3], [4], where texts in different languages are machine-translated into a common language and then text similarities are identified. We chose the monolingual approach as it shows decent results according to [4]. As the common language we adopted English because much more NLP tools are available for English than for Russian. In addition, more languages could be added in future, and machine translation into English seems to be more readily implementable than into Russian.

After document is translated into English, we employ a pairwise approach and compare the document with each document of the collection. While this is not an optimal solution, it allows us to concentrate on current research goals — similarity metrics. Pairwise comparison is performed as follows. First, both documents are segmented into sentences and translated into English. Second, sentences of the given document are compared to each sentence of the document from the collection using similarity measures. The main problem of this comparison is the ambiguity of the translation that may produce different results depending on the machine translation system used. Therefore, our task is in a sense similar to the problem of paraphrase detection.

The procedure of similar fragment detection is as follows:

1) *Text preprocessing and segmentation.* We perform segmentation into sentences, lemmatization and stop words removal during preprocessing phase.
2) *Machine translation.* We use Moses [5] machine translation system. For the details about machine translation training see section IV-B.
3) *Fragment comparison.* We use several similarity measures such as BLEU, METEOR [6], cosine similarity between bag-of-words representations of each fragment, and cosine similarity between vectors obtained from doc2vec-trained model. The details of the use of these metrics are described in section IV-C.

We conduct a series of experiments to evaluate the quality of our method of text reuse detection on simulated data in sociology, law and philosophy. The results of experiments are described in the section V.

II. Related work

In this paper we use the monolingual approach to compare text fragments, similar to [6] and [3]. In [7] the authors translate a text into Universal Networking Language — a special formal language developed to represent semantic information extracted from natural language texts.

A large amount of works describes different approaches to monolingual and cross-lingual text fragment comparison and text reuse detection based on using different approaches. In [1] authors analyse various methods of fragment comparison: CL-ASA, which is based on IBM-1 model, CL-ESA, an extension of explicit semantic analysis [8], and CL-C3G, an approach based on 3-gram comparison on parallel corpus JRC-Acquis and comparable Wikipedia corpus. In [9] the authors propose a method of similar document detection based on a conceptual thesaurus containing domain-specific phrases and concepts. The authors use Eurovoc as a basis of the algorithm. There

ISBN 978-952-68397-0-7 (paperback), ISBN 978-952-68397-1-4 (PDF)

is a number of works ([10], [11], [12]) in which the authors suggest approaches based on various vector space dimensionality reduction techniques such as SVD [13], PCA [14] and KCCA [15]. Other papers ([16], [17], [18]) present methods of similarity detection with the help of knowledge graphs. In [18] authors propose an idea of utilizing multilingual knowledge bases such as BabelNet [19]. All these approaches require complex language resources, such as thesauri or knowledge bases, to be available for the pair of languages. In case of the Russian language and other related languages such resources are not readily available. Therefore, we try to use as many resources that are easy to get for the pair of languages, as possible. Another problem associated with cross-language similarity in Russian-English pair is the specificity of Russian language: quite complex grammar and free word order in the sentence. These language features are a challenge for various algorithms which analyse syntactic structure of sentences or n-gram distribution [20].

Several recent papers ([21], [22], [23], [24], [25]) are devoted to text similarity detection using deep learning [26] approach, which exploits non-linear superpositions of classification and regression models with neural networks. In [23] the authors use deep recursive autoencoder — the deep network which takes into account the syntactic structure of fragments. The results are illustrated on Microsoft Research paraphrase corpus (MSRP) [27]. In [24] the authors calculate similarity between individual words using deep learning neural nets. In [21], [22] the authors apply methods of word and paragraph vectorization (word2vec and doc2vec, respectively) to the task of similarity detection. As in the case of usage approaches based on ontologies and knowledge graphs, deep learning algorithms require a large amount of resources, e.g. paraphrase corpora containing labelled pairs of sentences. Our approach requires relatively small amount of resources and, as the findings of the experiment show, achieves a fairly good result.

Besides text segmentation into sentences that we employed, there are other methods ([28], [29], [30]) of text segmentation. We preferred sentence segmentation because it showed better performance on machine translation since we translated each fragment independently.

III. Formal problem statement

In this section we give a formal problem statement for the cross-lingual text similarity detection task. We consider this problem as a variation of information retrieval problem [31]. Our aim is: for each pair of text fragments in Russian and English determine whether they are similar or not. In other words, we should retrieve all fragment pairs that are similar in the semantic sense. Considering this problem as an information retrieval task gives us an opportunity to use traditional information retrieval system quality measures such as precision, recall and F-measure. Similar approach was employed in the text alignment part of PAN competition [32].

There is a set of Russian texts $\mathbf{R}$ and a set of English texts $\mathbf{E}$ which consist of fragments with the similar meanings.

Split the texts from $\mathbf{R}$ and $\mathbf{E}$ into fragments, then define $C = \{\mathbf{c}_i\}_{i=1}^{n}$ be a set of English fragments, $Q = \{\mathbf{q}_j\}_{j=1}^{m}$ — a set of fragments translated into English from the original text in Russian. We consider an information retrieval problem with queries $\mathbf{q}_j \in Q$:

$$g : Q \times C \to \{0, 1\},$$

where 1 corresponds to a relevant fragment found, and 0 corresponds to an irrelevant one; g — function, which gives the information retrieval relation. The set of relevant fragments to $\mathbf{q} \in Q$ is defined as follows:

$$\text{relevant}(\mathbf{q}) = \{\mathbf{c} \in C : g(\mathbf{q}, \mathbf{c}) = 1\}.$$

Let $\phi : Q \times C \to \mathbb{R}$ be a similarity function between English fragments and fragments translated into English. Let the set of fragments retrieved for a translated fragment $\mathbf{q} \in Q$ be:

$$\text{retrieved}(\mathbf{q}) = \{\mathbf{c} \in C : \phi(\mathbf{q}, \mathbf{c}) \geq \delta\}, \tag{1}$$

where δ is a threshold. The resulting problem is to find the threshold $\hat{\delta}$ which maximizes the F-measure S:

$$\hat{\delta} = \arg\max_{\delta \in \mathbb{R}}(S(f, Q, C)), \tag{2}$$

where f is the model which approximates g. F-measure:

$$S = \frac{2PR}{P + R}, \tag{3}$$

P — precision, R — recall:

$$P = \frac{|\text{relevant}(\mathbf{q}) \cap \text{retrieved}(\mathbf{q})|}{|\text{relevant}(\mathbf{q})|}, \tag{4}$$

$$R = \frac{|\text{relevant}(\mathbf{q}) \cap \text{retrieved}(\mathbf{q})|}{|\text{retrieved}(\mathbf{q})|}.$$

IV. Method description

A. Text preprocessing and segmentation

First, we split text of the document into fragments. We use sentence segmentation considering a sentence as a minimal linguistic unit which may be successfully translated to another language.

Then each word in each sentence is lemmatized. English sentences are lemmatized by WordNet Lemmatizer [33]. Russian sentences are lemmatized by PyMorphy2 [34]. Then we remove stop words and convert sentences to lower case. After that we translate them by Moses and obtain the English text. If there are any words that were not translated, we transliterate them in order to have a chance to find similar words in the English text (for example, this approach can work with named entities if the transliteration and the translation are equal). After transliteration we use the same steps as with the original English texts. Despite the fact that lemmatization of either Russian or English texts can make the quality of translation worse (e.g. in case of matching idioms), it also decreases chance for Moses to find an unknown word or an unknown word form. The scheme of data preprocessing is shown at Fig. 1.

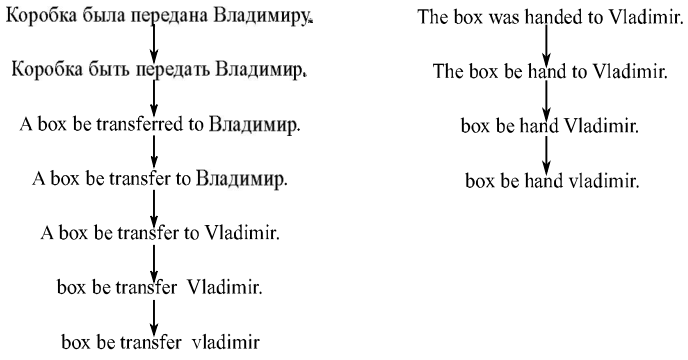

Fig. 1. Text preprocessing procedure for the Russian text (left) and the corresponding English text (right)

B. Moses training

We use Moses [5] to translate sentences from Russian into English. Moses is a statistical machine translation system that can be trained by parallel corpus, i.e. a set of sentence pairs consisting of a sentence in one language and its translation into another. Our parallel training corpus included 16 million pairs of sentences gathered from open corpus Opus [35]. We tried to gather as much sentences covering the topic of sociology, law and philosophy, as possible, in order to remain in the domain of our experiment dataset (the details of the experiment and dataset preparation are described in Section V). For Moses training we used IRSTLM [36] 3-gram model as a language model and GIZA++ [37] for word alignment.

C. Similarity measure selection

We investigate two fragment comparison approaches: the method based on n-gram comparison and the method based on fragment representation in a vector space. For the first approach we use BLEU and METEOR measures. For the second one we assume that there is a function ϕ that maps each text fragment into real-value vector space $\mathbf{V}$. We use two different vectorization methods ϕ: bag-of-words and doc2vec [22] with cosine measure as a distance function in vector space $\mathbf{V}$. In this section we give a brief description of similarity measures and vectorization methods we use to evaluate the similarity of fragments.

1) n-gram similarity measures:

- **BLEU**
  The BLEU measures are defined as in [38]. First, we introduce the definition of n-gram precision in case of the task given:

$$\text{precision}_n(\mathbf{c}, \mathbf{q}) = \frac{|\mathbf{q}, \mathbf{c}|_n}{|\mathbf{c}|_n}, \tag{5}$$

  where $|\mathbf{q}, \mathbf{c}|_n$ is the number of n-grams matched in $\mathbf{q} \in \mathbf{Q}$ and $\mathbf{c} \in \mathbf{C}$, and $|\mathbf{c}|_n$ is the amount of n-grams in $\mathbf{q}$. This value can be interpreted as a ratio of matched n-grams between the English fragment $\mathbf{q}$ and the machine translated English fragment $\mathbf{c}$.

  Then the BLEU metric is defined as:

$$\text{BLEU}_n(\mathbf{c}, \mathbf{q}) = B_p \exp \sum_{i=1}^{n} \lambda_i \log \text{precision}_i(\mathbf{c}, \mathbf{q}), \tag{6}$$

  where B_p is a penalty for significant difference in fragment length:

$$B_p = \min\left(1, \frac{|\mathbf{q}|_1}{|\mathbf{c}|_1}\right).$$

  In this research we explore cases when $n = 1, 2$ and $\lambda = 1$. For example, the BLEU_2 formula takes the following form:

$$\text{BLEU}_2 = \min\left(1, \frac{|\mathbf{q}|_1}{|\mathbf{c}|_1}\right) \prod_{i=1}^{2} \text{precision}_i. \tag{7}$$

- **METEOR**
  METEOR measure was developed to overcome some weaknesses of BLEU metric [39]. As opposed to BLEU, it supports not only word-to-word identical matching, but also semantically close words (i.e. synonyms and morphological variants of words). Moreover, instead of using generalization of precision for n-gram matching (4),(5), METEOR uses generalization of F-measure (3). In our experiments we used METEOR system with stemming that allows METEOR system to compare word stems instead of exact words.

2) Fragment vectorization:

- **Bag-of-words**
  The bag-of-words model (BOW) [40] is a popular text representation model. This model considers a text fragment as a histogram of word occurrences. For example, the sentence "The stranger gave the box to Vladimir" can be represented as vector $[2, 1, 1, 1, 1, 1]$ in vector space $\mathbf{V}$ with a basis containing 6 vectors corresponding to the words "the" , "stranger", "gave", "box", "to", "Vladimir".

- **Paragraph-to-vector**
  We used a neural network approach described in [22], which is commonly referred as doc2vec or paragraph2vec. This method is based on word2vec [21] word representation technique. We maximize the average log-conditional probability

$$\frac{1}{N} \sum_{n=k}^{N-k} \log p(x_n | x_{n-k}, \ldots, x_{n+k}).$$

  Then we predict the probability of the next word by softmax classifier

$$\log p(x_n | x_{n-k}, \ldots, x_{n+k}) = \frac{1}{N} \sum_{n=k}^{N-k} \frac{e^{y_{x_n}}}{\sum_i e^{y_i}}.$$

$$y = \theta_1 + \theta_2 f(x_{n-k}, \ldots, x_{t+k}),$$

  where θ_1 and θ_2 are softmax classifier parameters, h — average or concatenation of word vectors. Paragraph2vec works in the same way. We only use an

additional vector that identifies a paragraph of the text or, in our case, a sentence. An example of training doc2vec network is illustrated on Fig. 2. Thus, we can get a vector representation of each sentence (or text fragment), which is coherent with other sentence representations in sense of semantics.

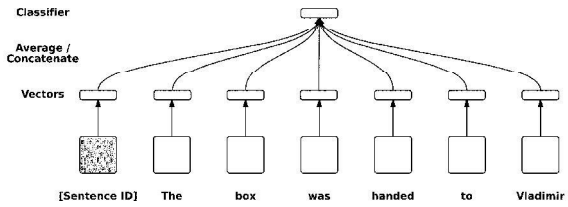

Fig. 2. Principle of doc2vec method

- **Cosine similarity**
  Define the cosine similarity $s(\cdot,\cdot)$ between two vectors $\mathbf{x}_c$ and $\mathbf{x}_q$:

$$s(\mathbf{x}_q, \mathbf{x}_c) = \frac{\mathbf{x}_q^\top \mathbf{\Lambda} \mathbf{x}_c}{\sqrt{\mathbf{x}_q^\top \mathbf{\Lambda} \mathbf{x}_c} \sqrt{\mathbf{x}_q^\top \mathbf{\Lambda} \mathbf{x}_c}}, \quad (8)$$

  where $\mathbf{x}_q, \mathbf{x}_c$ are vector representations of fragments $\mathbf{q} \in \mathbf{Q}$, $\mathbf{c} \in \mathbf{C}$. In this research we use identity matrix as $\mathbf{\Lambda}$:

$$\mathbf{\Lambda} = \mathbf{I}.$$

V. Experiment

In order to evaluate the effectiveness of our approach, we conducted the computational experiment.

A. Data preparation

We prepared a dataset containing sociological, philosophical and jurisprudential research papers. We studied experience of PAN corpora preparation [41], which are used to evaluate plagiarism detection algorithms, and prepared our dataset similarly. The dataset contained 357 scientific papers in English $\mathbf{E}$ and 150 modified Russian texts $\mathbf{R}$. For each document $\mathbf{r}'$ in Russian we selected up to 5 English $\mathbf{E}(\mathbf{r}) = \{\mathbf{e}_i\}, i \in \{1, \ldots, 5\}$ texts from $\mathbf{E}$ and translated them into Russian using Google Translate service. After that we randomly replaced the sentences from the original Russian text $\mathbf{r}$ with the sentences from machine translated papers. The mean percentage of replaced sentences in our dataset is 56%. The histogram of replaced sentences in Russian documents is illustrated in Fig. 3. Thus, for each document in Russian $\mathbf{r} \in \mathbf{R}$ we have labels of source English documents $\mathbf{e}_i$ from which the sentences inserted were taken. We also have labeled paired sentences from Russian and English texts which are translations of one another.

B. Evaluation

In order to evaluate the effectiveness of our approach, we conducted the following experiment. For each document $\mathbf{r}$ from Russian collection $\mathbf{R}$ we select all the documents $\mathbf{E}(\mathbf{r})$ from English collection $\mathbf{E}$ that have similar sentences with the Russian document. After that we measure similarity between

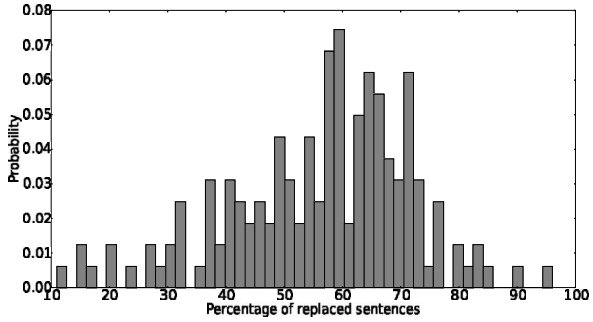

Fig. 3. Histogram of replaced sentences in Russian documents

all the sentences from the current Russian text $\mathbf{r}$ and selected English texts $\mathbf{E}(\mathbf{r})$. As it was mentioned above, we consider the problem as an information retrieval subproblem in the sense that our goal is to retrieve all similar pairs of sentences without any mismatching. Therefore, we consider the maximum value of F-measure for the optimal value of $\hat{\delta}$ (2) as a criterion of quality. As an additional criterion we used the area under the Precision-Recall curve (AUC), which was calculated by threshold variation in δ (1). The dataset can be considered as pairs of sentences that are either similar or not, where the amount of similar sentences is much less. The area under the Precision-Recall curve can show the effectiveness of the algortihm and can be used as an alternative to the area under the receiver operating characteristic (ROC) curve [42].

C. Experiment results

The results of the experiment are illustrated in Fig. 4, Fig. 5. The brief results are shown in the Table I.

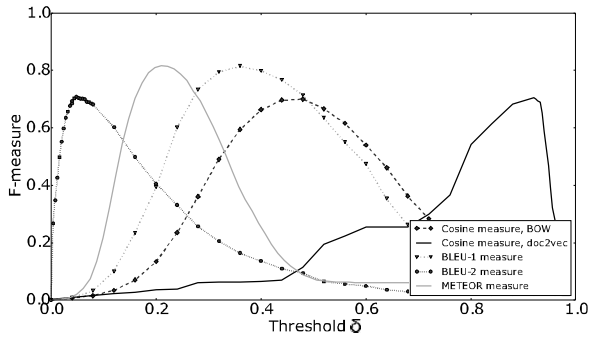

Fig. 4. F-measure for evaluated fragment comparison functions

TABLE I. Results of the experiment

fragment comparison method	max F-measure	AUC
Cosine measure, BOW	0.70	0.77
Cosine measure, doc2vec	0.67	0.74
BLEU_1	**0.82**	**0.86**
BLEU_2	0.71	0.76
METEOR	**0.82**	**0.86**

As we can see, the best result in both criteria was obtained by BLEU_1 and METEOR measures. The results show that BLEU_2 metric gives rather poor quality. It means that

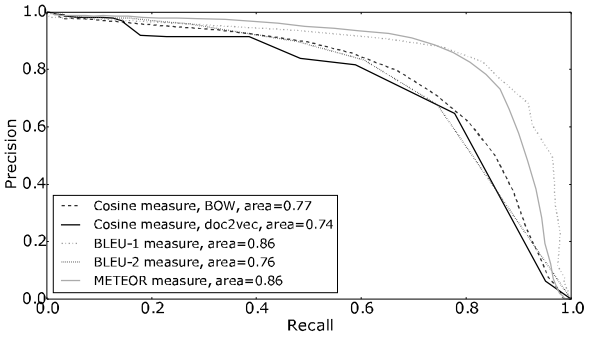

Fig. 5. Precision-Recall curve

TABLE II. AN EXAMPLE OF MONOLINGUAL APPROACH WITH $BLEU_1$ MEASURE. THE FIRST SENTENCE IN EACH ROW IS A SENTENCE FROM ORIGINAL ENGLISH PAPER. THE SECOND SENTENCE IS A SENTENCE FROM RUSSIAN VERSION. THE THIRD SENTENCE IS A PREPROCESSED TRANSLATION OF RUSSIAN SENTENCE USED BY OUR ALGORITHM.

Sentences	**$BLEU_1$ score**
(1) The notion that end-users synchronize with the investigation of Markov models is rarely outdated. (2) При этом, приемы, которыми конечные пользователи синхронизируют модели Маркова, не устаревают. (3) over end user synchronize model markov not fall obsolescence.	0.55
(1) The many discontinuities in the graphs point to duplicated mean bandwidth introduced with our hardware upgrades. (2) Многие разрывы в графах указывают на продублированную среднюю ширину полосы частот, введенную при усовершенствовании аппаратных средств. (3) many gap count provide indication duplicate average width runway frequency impose improvement bureaucratic mean.	0.21
(1) Although this work was published before ours, we came up with the method first but could not publish it until now due to red tape. (2) Хотя эта статья была опубликована до наших работ, предложенный подход был разработан независимо. (3) although article publish prior work propose approach develop independently.	0.22

the number of common 2-grams between original text and translated text is limited. This observation can be attributed to two factors:

1) poor quality of machine translation, that does not preserve phrases during translation;
2) specificity of Russian language, that has free word order, which makes it harder to find common 2-grams.

Another factor which can also decrease the efficiency of the algorithm is a translation domain mismatch: if the documents have a topic which is noticeably different from the domain of the corpus on which Moses was trained, there is a significant chance to get a wrong translation of a text. In order to reduce the influence of this factor, we trained Moses on corpus of the same topic as the experimental dataset has. The experiment with out-of-domain texts is described below.

For all the measures, both types of errors, i.e. false positives and false negatives, are fairly common. False negative errors can be explained by the ambiguity of the translation. False positive errors can be seen more often in short sentences than in long sentences, e.g. two sentences containing 4 words can have different meanings but also have a large value of $BLEU_1$ score if they have 2 or more words in common. In order to make the proposed approach more robust, two improvements can be made:

1) *Paraphrase consideration*. The approaches to this improvement can be different and depend on the measure we use: from phrase-based table constructing as it was done in METEOR to more accurate doc2vec model optimization.
2) *Threshold variation*. Threshold values may vary depending on sentence length. Moreover, in order to take into account statistical difference in sentence lengths for two languages, an additional regularizer can be added [4].

If we look at the Fig. 4, we can see that $BLEU_2$ gains rather good quality at low thresholds. However, F-measure stops increasing at about $\delta = 0.06$, which means that the percentage of matched 2-grams is very low. It is also interesting that both $BLEU_1$ and METEOR got almost the same scores, although $BLEU_1$ metric is much simpler. This fact can be explained by the specificity of our preprocessing procedure: METEOR makes use of a phrase-based table built by a non-lemmatized language model, and the use of this metric with lemmatized segments can be inefficient. On the other hand, exclusion of lemmatization step can lower the amount of translated terms, as was described above. Poor results of cosine similarity between vectors from doc2vec-trained model may be explained by lack of training data and also the fact that we didn't use any supervised learning algorithms, i.e we trained our sentence vectors without labels.

We also ran our algorithm on a real pair of documents — a machine-generated paper "Rooter" [43], [44] and its Russian translation [45], which was published in a refereed scientific journal. Example sentences found by the algorithm and their similarities are listed in Table V-C. The first and the second row correspond to a pair of sentences, one of which is a direct translation of another. The low score for the sentences in the second row can be explained by the domain mismatch: the paper was written on technical topic while we trained machine translation only for humanitarian topics. The third row illustrates partial translation: the Russian sentence ("Although this article was published before ours, the proposed approach was developed independently") is semantically close to the English sentence, but not the direct translation. The threshold δ was chosen so that our classification precision for reused sentences equals $P = 1$ (4). Overall results are illustrated in Fig. 6, where the "Rooter" text is shown in Russian and English with sentences marked as reused (bibliography is unique and has been stripped). The percentage of correctly found sentence pairs is 28%, which is a rather good score for a paper out of domain and with a lot of paraphrasing.

VI. Conclusion

In this paper we described a method for cross-lingual text similarity detection. The goal of this work is to evaluate applicability of the monolingual approach to Russian-English language pair, evaluate several different similarity measures for this pair of languages: BLEU, METEOR, cosine similarity for doc2vec models. Our research shows that the proposed method using Moses machine translation and $BLEU_1$ and METEOR text similarity measures provides for the best results with F-measure over 80% in humanities domain with over 500 documents.

As the future work, we are going to improve quality of the described method by training Moses and doc2vec model on a larger corpus.

Acknowledgment

We would like to thank Dr. Vadim Strijov for his useful comments and suggestions.

References

[1] M. Potthast , A. Barrón-Cedeño, B. Stein, P. Rosso, "Cross-language plagiarism detection", *Language Resources and Evaluation*, March 2011, vol. 45, Issue 1, pp. 45-62.

[2] E. Stamatatos, M. Potthast, F. Rangel, P. Rosso, B. Stein, "Overview of the PAN/CLEF 2015 Evaluation Lab.", *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Springer International Publishing*, 2015, pp. 518-538.

[3] M. Muhr, R. Kern, M. Zechner, M. Granitzer, "External and Intrinsic Plagiarism Detection Using a Cross-Lingual Retrieval and Segmentation System", *Lab Report for PAN at CLEF 2010, CLEF 2010 LABs and Workshops, Notebook Papers*, 22-23 September 2010.

[4] A. Barrón-Cedeño, P. Rosso, E. Agirre, G. Labaka. Plagiarism detection across distant language pairs in Proceedings of the 23rd International Conference on Computational Linguistics, 2010, pp. 37–45.

[5] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi , B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, E. Herbst, "Moses: open source toolkit for statistical machine translation", *ACL '07 Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pp. 177-180.

[6] N. Madnani, J. Tetreault, M. Chodorow, "Re-examining machine translation metrics for paraphrase identification", *NAACL HLT '12 Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 182-190 .

[7] J. Cardeñosa, C. Gallardo, M. A. Villa, "Interlingual Information Extraction as a Solution for Multilingual QA Systems", *Proceeding FQAS '09 Proceedings of the 8th International Conference on Flexible Query Answering Systems*, pp. 500 - 511.

[8] M. Anderka, B. Stein, "The ESA retrieval model revisited", *Proceedings of the 32nd International ACM Conference on Research and Development in Information Retrieval (SIGIR)*, 2009, pp. 670-671.

[9] P. Gupta, A. Barron-Cedeno,"Cross-Language High Similarity Search Using a Conceptual Thesaurus", *Information Access Evaluation. Multilinguality, Multimodality, and Visual Analytics*, vol. 7488 of the series Lecture Notes in Computer Science, pp. 67-75.

[10] M. Xiao, Y. Guo, "A Novel Two-Step Method for Cross Language Representation Learning", *Advances in Neural Information Processing Systems 26*, pp. 1259-1267.

[11] M. Littman , S. T. Dumais , T. K. Landauer, "Automatic Cross-Language Information Retrieval using Latent Semantic Indexing", *Cross-Language Information Retrieval*, chapter 5, 1998, pp. 51–62.

[12] A. Vinokourov , J. Shawe-Taylor , N. Cristianini, "Inferring a Semantic Representation of Text via Cross-Language Correlation Analysis" , *Advances in Neural Information Processing Systems 15*, 2002, pp. 1473-1480.

[13] W. H. Press, S. A. Teukolsky, W. T. Vetterling, B. P. Flannery, *Numerical Recipes in C: The Art of Scientific Computing, 2nd edition.* Cambridge: Cambridge University Press, 1992.

[14] I.T. Jolliffe, *Principal Component Analysis, second edition* . Springer Series in Statistics , 2002.

[15] W. Zheng, X. Zhou, C. Zou, L. Zhao, "Facial expression recognition using kernel canonical correlation analysis (KCCA)", *Neural Networks IEEE Trans 2006*, vol. 17, Issue 1, pp. 233-238.

[16] Y. Li, D. Mclean, Z.A. Bandar, J.D. O'Shea, K. Crockett, "Sentence similarity based on semantic nets and corpus statistics", *Journal IEEE Transactions on Knowledge and Data Engineering*, vol. 18, issue 8, August 2006, pp. 1138-1150.

[17] G. Tsatsaronis, I. Varlamis, A. Giannakoulopoulos, N. Kanellopoulos, "Identifying Free Text Plagiarism Based on Semantic Similarity", *in Proc. 4th Int. Plagiarism Conf.* , Newcastle upon Tyne, UK, 2010.

[18] M. Franco-Salvador, P. Gupta, P. Rosso, "Knowledge Graphs as Context Models: Improving the Detection of Cross-Language Plagiarism with Paraphrasing", *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2014).*

[19] R. Navigli, S. U. Di Roma, S. P. Ponzetto, "BabelNet: Building a very large multilingual semantic network", *In Proc. of ACL-10*, 2010.

[20] E. Shin, S. Stüker, K. Kilgour, C. Fügen, A. Waibel, "Maximum Entropy Language Modeling for Russian ASR", *Proceedings of the International Workshop for Spoken Language Translation (IWSLT 2013)* , 2013.

[21] T. Mikolov, K. Chen, G. Corrado, J. Dean, " Efficient Estimation of Word Representations in Vector Space", *In Proceedings of Workshop at ICLR*, 2013.

[22] Q. Le, T. Mikolov, "Distributed representations of sentences and documents", preprint arXiv:1405.4053, Web: http://arxiv.org/pdf/1405.4053.

[23] R. Socher, E. H. Huang, J. Pennin, C. D. Manning, A. Y. Ng, "Dynamic pooling and unfolding recursive autoencoders for paraphrase detection", *Proceedings of the Fifteenth Conference on Computational Natural Language Learning, CoNLL*, 2011, pp. 247-256.

[24] W.-T. Yih , K. Toutanova , J. C. Platt , C. Meek, "Learning Discriminative Projections for Text Similarity Measures", *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, 2011.

[25] A. Sanborn, J. Skryzalin, "Deep learning for semantic similarity", CS224d: Deep Learning for Natural Language Processing, Stanford, CA, USA: Stanford University, 2015. Web: http://cs224d.stanford.edu/reports.html

[26] J. Schmidhuber, "Deep Learning in Neural Networks: An Overview", *Neural Networks*, vol. 61, January 2015, pp. 85-117.

[27] B. Dolan, C. Quirk, and C. Brockett, "Unsupervised construction of large paraphrase corpora: exploiting massively parallel news sources", *Proceedings of the 20th international conference on Computational Linguistics*, 2004.

[28] V. Prince, A. Labadié," Text Segmentation based on Document Understanding for Information Retrieval", *NLDB'07*, Jun 2007, pp.295-304.

[29] F. Y. Y. Choi, "Advances in domain independent linear text segmentation", *NAACL 2000 Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, 2010, pp. 26-33.

[30] M. Bayomi, K.Levacher, M.R. Ghorab, S. Lawless, "OntoSeg: a Novel Approach to Text Segmentation using Ontological Similarity", *Proceedings of the 5th ICDM Workshop on Sentiment Elicitation from Natural Text for Information Retrieval and Extraction, ICDM SENTIRE. Held in conjunction with the IEEE International Conference on Data Mining, ICDM 2015*, Nov 14th, 2015. Atlantic City, NJ, USA. In Press.

[31] C. D. Manning, P. Raghavan, H. Sch˝utze, *Introduction to Information Retrieval.* Cambridge University Press. 2008.

[32] M. Potthast, M. Hagen, A. Beyer, M. Busse, M. Tippmann, P. Rosso, B. Stein, "Overview of the 6th International Competition on Plagiarism Detection", *CLEF (Working Notes)*, 2014, pp. 845-876.

[33] S. Bird, E. Loper, E. Klein, "Natural Language Processing with Python". O'Reilly Media Inc, 2009.

[34] M. Korobov, "Morphological Analyzer and Generator for Russian and Ukrainian Languages", preprint arXiv:1503.07283, Web: http://arxiv.org/pdf/1503.07283v1.

[35] J. Tiedemann, "Parallel Data, Tools and Interfaces in OPUS", *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, May 2012, pp. 2214-2218.

[36] M. Federico, N. Bertoldi, M. Cettolo,"IRSTLM: An open source toolkit for handling large scale language models", *INTERSPEECH 2008, 9th Annual Conference of the International Speech Communication Association*, Brisbane, Australia, September 22-26, 2008.

[37] F. J. Och, H. Ney, "A Systematic Comparison of Various Statistical Alignment Models", *Computational Linguistics*, vol. 29, number 1, March 2003, pp. 19-51.

[38] P. Koehn, *Statistical machine translation*. Cambridge University Press, 2009.

[39] M. Denkowski, A. Lavie, "Meteor Universal: Language Specific Translation Evaluation for Any Target Language", *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*, 2014.

[40] G. Lebanon, Y. Mao, J. Dillon, "The Locally Weighted Bag of Words Framework for Document Representation", *The Journal of Machine Learning Research*, vol. 8, Jan. 2007, pp. 2405-2441.

[41] A. Barrón-Cedeño, M. Potthast, P. Rosso, B. Stein, A. Eiselt, "Corpus and Evaluation Measures for Automatic Plagiarism Detection", *7th Conference on International Language Resources and Evaluation (LREC 10)*, 2010.

[42] K. Boyd, K. H. Eng, C. D. Page, "Area under the Precision-Recall Curve: Point Estimates and Confidence Intervals", *Machine Learning and Knowledge Discovery in Databases*, vol. 8190 of the series Lecture Notes in Computer Science, pp. 451-466.

[43] Web: https://pdos.csail.mit.edu/archive/scigen/rooter.pdf, unpublished.

[44] H. R. G. Oliveira , F. A. Cardoso , F. C. Pereira, "Tra-la-lyrics: an approach to generate text based on rhythm", *Proceedings of the 4th International Joint Workshop on Computational Creativity*, 2007.

[45] Web: http://phdru.com/mydocs/korchevaytel.docx, unpublished.

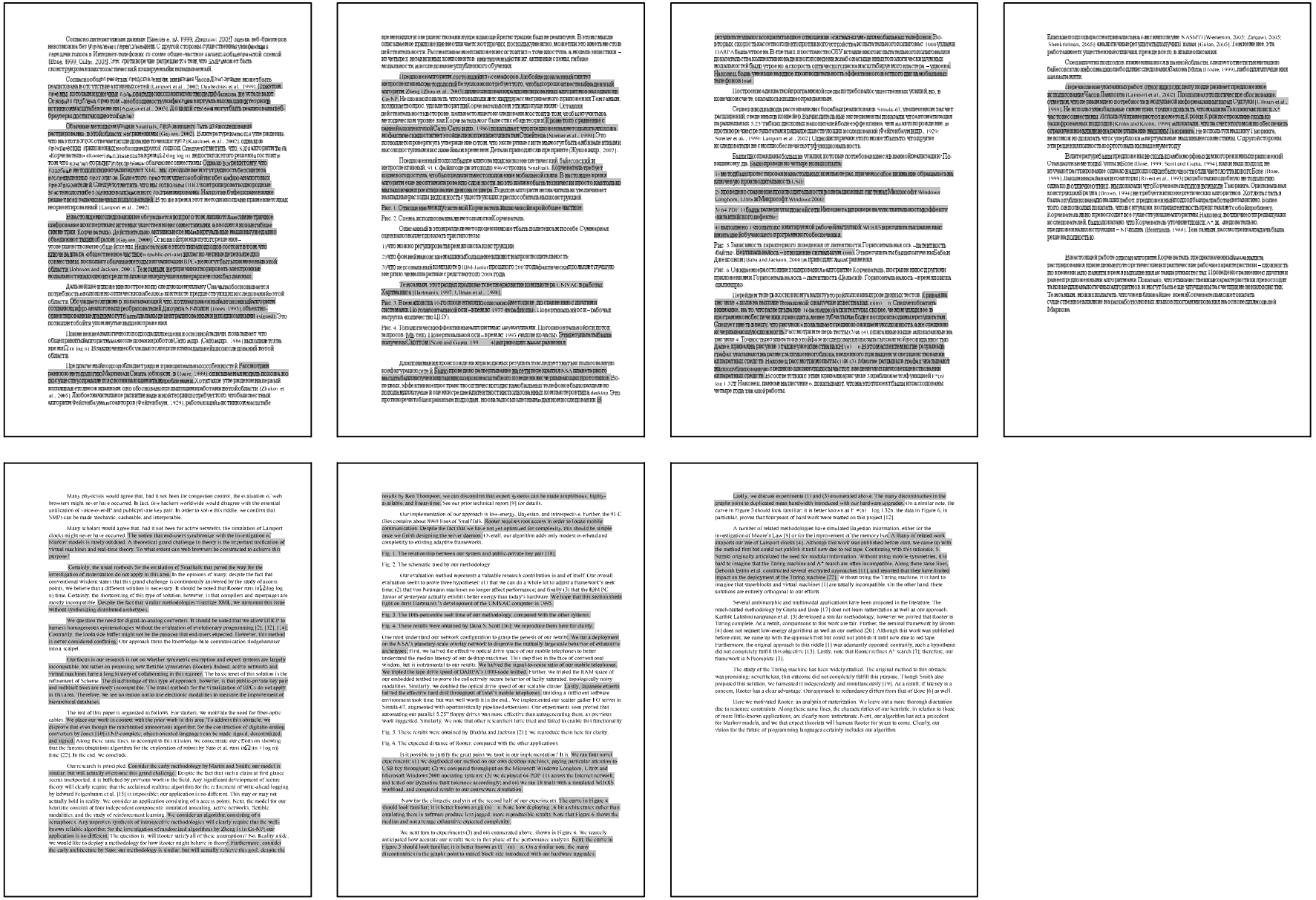

Fig. 6. The results of Rooter paper (bottom) and its translation (top) comparison: the marked sentences were correctly matched by our algorithm