DOI: 10.15514/ISPRAS-2019-31(5)-9

Cross-lingual similar document retrieval methods

D.V. Zubarev, ORCID: 0000-0*002-9687-6650 <zubarev@isa.ru>*
I.V. Sochenkov, ORCID: 0000-0003-3113-3765 <sochenkov@isa.ru>
Federal Research Center «Computer Science and Control» of Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russia

Abstract. In this paper, we compare different methods for cross-lingual similar document retrieval. We focus on Russian-English language pair. We compare well-known methods like Cross Lingual Explicit Semantic Analysis (CL-ESA) with methods based on cross-lingual embeddings. We use approximate nearest neighbor (ANN) search to retrieve documents based entirely on distances between learned document embeddings. Also we employ a more traditional approach with usage of inverted index, with extra step of mapping top keywords from one language to other with the help of cross-lingual word embeddings. We use Russian-English aligned Wikipedia articles to evaluate all approaches. Conducted experiments show that an approach with inverted index achieves better performance in terms of recall and MAP than other methods.

Keywords: cross-lingual document retrieval; cross-lingual plagiarism detection; cross-lingual word embeddings.

For citation: Zubarev D.V., Sochenkov I.V. Cross-lingual similar document retrieval methods. Trudy ISP RAN/Proc. ISP RAS, vol. 31, issue 5, 2019, pp. 127-136. DOI: 10.15514/ISPRAS-2019-31(5)-9

Acknowledgements. This study was funded by RFBR according to the research project № 18-37-20017. The reported research is also partially funded by the project “Text mining tools for big data” as a part of the program supporting Technical Leadership Centers of the National Technological Initiative “Center for Big Data Storage and Processing” at the Moscow State University (Agreement with Fund supporting the NTI-projects No. 13/1251/2018 11.12.2018).

Методы кросс-языкового поиска похожих документов

Д.В. Зубарев, ORCID: 0000-0002-9687-6650 <zubarev@isa.ru>
И.В. Соченков, ORCID: 0000-0003-3113-3765 <sochenkov@isa.ru>
Федеральный исследовательский центр «Информатика и управление» РАН, 119333, Россия, г. Москва, ул. Вавилова, д. 44, к. 2

Аннотация. В этой статье сравниваются различные методы кросс-языкового поиска похожих документов. Для сравнения используется русско-английская языковая пара. Сравниваются известные методы, такие как CL-ESA, с методами, основанными на кросс-языковых эмбеддингах. Для поиска документов используется приближенный поиск ближайшего соседа (ANN), использующий расстояния между векторами, представляющими документы. Также применяется более традиционный подход с использованием инвертированного индекса, с дополнительным шагом: отображение ключевых слов с одного языка на другой с помощью кросс-языковых эмбеддингов. Для экспериментальной оценки всех методов используются русские статьи из Википедии, которые имеют аналоги в англоязычной версии. Проведенные эксперименты показывают, что подход с инвертированным индексом показывает лучшие результаты по двум метрикам: полнота и средняя точность (MAP).

Ключевые слова: кросс-языковой поиск похожих документов; кросс-языковой поиск заимствований; кросс-языковые эмбеддинги

Для цитирования: Зубарев Д.В., Соченков И.В. Методы кросс-языкового поиска похожих документов. Труды ИСП РАН, том 31, вып. 5, 2019 г., стр. 127-136 (на английском языке). DOI: 10.15514/ISPRAS-2019-31(5)-9

Благодарности. Работа выполнена при поддержке гранта РФФИ № 18-37-20017. Исследование выполнено также при частичной финансовой поддержке проекта "Средства интеллектуального анализа больших массивов текстов" в рамках программы Центров компетенций Национальной технологической инициативы на базе Московского государственного университета им. М.В. Ломоносова (соглашение о финансовой поддержке проектов НТИ № 13/1251/2018 от 11.12.2018)

1. Introduction

Document retrieval from a large collection of texts is important information retrieval problem. This problem is extensively studied for short queries, such as user queries to search engines.

The document retrieval with texts as queries impose some difficulties, among them inability to capture the main ideas and topics from the long text. The problem becomes even harder when we enter the field of cross-lingual document retrieval. Some tasks require to use a text (possibly long) as query to retrieve documents that are somehow similar to it. One of these tasks is plagiarism detection that is divided into two stages: source retrieval and text alignment.

- On the source retrieval stage for a given suspicious document, we need to find all sources of probable text reuse in a large collection of texts. For this task, a source is a whole text, without details of what parts of this document were plagiarized. Typically, we get a large set of documents (around 500 or more) as a result of this stage.
- On the text alignment stage: we compare the suspicious document to each candidate to detect all reused fragments, and identify its boundaries [1-4].

In this work, we study only the first task. The same stages are valid for cross-lingual plagiarism detection. Given a query document in one language the goal is to find the most similar documents from the collection in another language.

2. Related work

Some works were recently devoted to the monolingual document retrieval for long texts. In [5], the authors introduce a siamese multi-depth attention-based hierarchical recurrent neural network that learns the long text semantics. They conducted multiple experiments including retrieval of similar Wikipedia articles. In [6], the authors try to employ standard approximate nearest neighbor (ANN) search instead of the usual discrete inverted index, for retrieving documents. They learned similarity function and showed that it can improve performance on two similar-question retrieval tasks. However, using the custom similarity functions makes impossible to employ existing frameworks for ANN, consequently they used exact search in experiments.

In [7], a framework is introduced for monolingual and cross-lingual information retrieval based on cross-lingual word embeddings. They represent user queries and documents as averaged embeddings of words and employ exact search to find similar documents for a given query. The overview of different approaches for cross-lingual source retrieval is presented in [8] and [9]. Also, there made an evaluation and a detailed comparison of some featured methods.

In [10], NMT (neural machine translation) is used to translate a query document to other language. They solve source retrieval task by employing shingles (overlapping word N-grams) method. They use word-class shingles, instead of word shingles, where each word is substituted by the label of the class it belongs to. To obtain word classes they apply agglomerative clustering on word embeddings learned from English Wikipedia.

The work [11] describes a training of word embeddings on comparable monolingual corpora and learning the optimal linear transformation of vectors from one language to another (there were

used Russian and Ukrainian academic texts). Also there were discussed usage of those embeddings in source retrieval and text alignment subtasks. This work focuses on comparison of retrieval-based approaches with ANN approach for distant language pair.

3. Document retrieval methods

In this section, we describe various methods that we used for document retrieval.

3.1 Preprocessing

On a preprocessing stage, we split each sentence into tokens, lemmatize tokens and parse texts. We use AOT for the Russian language and Udpipe[1] [12] for English language. Besides, we removed words with non-important part of speech: conjunction, pronoun, preposition, etc., and common stop-words (be, являться, etc.).

3.2 Cross-lingual embeddings

We train cross-lingual word embeddings for a Russian-English pair on parallel sentences available on the Opus site [13] namely:

- News Commentary;
- TED Talks 2013;
- MultiUN (first 2 million sentences);
- Wiki;
- JW300;
- QED;
- Tatoeba.

We extend this corpus with sentences from the Yandex Parallel corpus[2] [14].

All parallel sentences are preprocessed. After that, all pairs that have a difference in the size of more than 10 words are filtered out. We use syntactic phrases up to 4 words in length to enrich the vocabulary. We take only those phrases (noun phrases and some prepositional phrases for the English language) that are common for the corpus (>10 occurrences). We duplicate one sentence multiple times if there are some overlapping phrases. For example, from the sentence with the phrase «Russian presidential election ...» will be generated three variations with different phrases:

- «Russian_presidential_election ... »;
- «Russian_election presidential_election ... »;
- «Russian presidential election ... ».

Finally, we assembled a corpus of more than 5.1 million sentences (more than 10 million sentences with phrases variations). The dictionary size was around 680 000 words/phrases.

We apply two different methods for learning cross-lingual embeddings [15].

First, we learn monolingual embeddings for each language. We use word2vec skip-gram model [16] with the following parameters: dimensionality of embeddings was 300, a window size of 10 words, the minimal corpus frequency of 10, negative sampling with 10 samples, no down-sampling, 20 iterations over the corpus. Then we use vecmap [17, 18] framework to learn a transformation matrix that maps representations in one language to the representations of the

[1] english-ewt-ud-2.4-190531 model

[2] Англо-русский параллельный корпус: https://translate.yandex.ru/corpus?lang=en

other language. We use 20 000 random word pairs from the bilingual dictionary of MUSE project[3] [19] as the training data.

Second, we apply the method proposed in [20], designed for learning bilingual word embeddings from a non-parallel document-aligned corpus, but it can be used for learning on parallel sentences too. We assume that the structures of the two sentences are similar. Words are inserted into the pseudo-bilingual sentence relying on the order in which they appear in their monolingual sentences and based on length ratio of two sentences. For example, if we were given two sentences: «Мама мыла раму» and «Mother washed beautiful frame», the result of their merging is «мама mother мыла washed раму beautiful frame». Since we removed auxiliary words from sentences, we assume that corresponding Russian and English words are in the same context window. It would not be the case if there were a different word order, so we experimented with different window sizes and chose size == 10. After that, the word2vec skip-gram model is used on the resulting bilingual corpus. We use gensim word2vec implementation with those parameters: dimensionality of embeddings was 300, a window size of 10 words, the minimal corpus frequency of 10, negative sampling with 10 samples, no down-sampling, 20 iterations over the corpus.

3.3 Retrieval-based approach

We use a custom implementation of inverted index [21], which maps each word to a list of documents in which it appears along with weight (e.g. TF) that represents the strength of association of this word with a document. Along with words, we index syntactic phrases up to 4 words, which occur in a document more than once.

At query time, we extract the top words/phrases from the query document according to some weighting scheme. Then we map each keyword to N other language keywords with cross-lingual embeddings. We precompute the most similar words for each word in our vocabulary to speed up this operation. We preserve the weights of keywords from the original top. The searcher iterates over the top keywords, retrieves corresponding documents from the inverted index, and merges them into weighted vectors of keywords that represent the other documents. Then we compare the query vector with all other vectors. It should be noted that comparison is asymmetrical since vectors of other documents consist only of words from the query vector. Although it is not the most accurate representation of these documents, the comparison is very efficient, and retrieval performance (recall) is not affected much by that. To compute the similarity score between vectors we employ some similarity measure (e.g. cosine similarity).

3.4 Approximate nearest neighbor search (ANN)

In this approach, we represent each document as a dense vector. It is done by averaging vectors of the top K keywords of the document. After that, we index all vectors with ANN index. At query time, the given document is transformed into the vector representation, and the approximate nearest neighbor search is employed to retrieve the most similar documents.

3.5 Explicit semantic analysis (ESA)

We implemented CL-ESA method described in [9] and firstly introduced for solving monolingual semantic relatedness task in [22]. This method represents the document as a weighted vector of concepts. Concepts are defined by Wikipedia articles. In the original work, the authors used all English Wikipedia articles as concepts. We selected around 800 000 English articles that are aligned with Russian Wikipedia articles (articles that identified as comparable across languages by the Wikipedia community). For a given document *D* the weight of a concept

[3] A library for Multilingual Unsupervised or Supervised word Embeddings, https://github.com/facebookresearch/MUSE

C is defined as cosine similarity between top M keywords of D and matched keywords of an article that is linked with the concept C:

$$\frac{\sum_{w_i \in D} v_i \cdot c_i}{\sqrt{\sum_{w_i \in D} v_i^2} \sqrt{\sum_{w_i \in D} c_i^2}}$$

where v_i is the weight of a word w_i for D (e.g. TF-IDF), c_i is the weight of a word w_i for a Wikipedia article linked with the concept (e.g. TF-IDF).

We precomputed vector of concepts for each document in text collection and stored them with the same inverted index implementation that was used for the retrieval-based approach.

At query time, the query document is converted to a vector of weighted concepts, i.e., identificators of Wikipedia articles. Then those identificators are mapped to articles in other language, and similar documents are retrieved via search in the inverted index.

4. Dataset

We use Russian-English aligned Wikipedia articles as a dataset for evaluation of retrieval methods (Wikipedia dump of June 2019). We exclude all articles, which title starts with words "List of", which size in symbols is less than 800, and which number of sentences is less than 10.

Then we divide all remaining pairs of articles into two groups and each group into five bins by the size of a Russian article in sentences:

- comparable by size: those articles that satisfy the following requirement:

$$|len(a_{\text{ru}}) - len(a_{\text{en}})| < min(len(a_{\text{ru}}), len(a_{\text{en}}))/4$$

- non-comparable by size: Those articles that satisfy the following requirement:

$$|len(a_{\text{ru}}) - len(a_{\text{en}})| > min(len(a_{\text{ru}}), len(a_{\text{en}}))$$

Table 1: Statistic of comparable by size articles

Size in ru sents	count	Mean sife of ru texts	Mean size of en texts
(9, 50]	62291	2560.34	2626.16
(50, 100]	20012	5878.33	5989.66
(100, 200]	9163	11100.2	11301.2
(200, 400]	3526	21275	21693.2
(400, 1000]	1628	43478.6	44023.1

Table 2: Statistic of non-comparable by size articles

Size in ru sents	count	Mean sife of ru texts	Mean size of en texts
(9, 50]	170902	2336.02	11471.6
(50, 100]	42958	6076.27	17771.8
(100, 200]	22491	11519.4	22309.7
(200, 400]	9318	21425.2	26847.4
(400, 1000]	3517	42744.7	26895

Then we sampled 100 documents from each group. That gives us a dataset that contains 1000 document pairs[4].

[4] http://nlp.isa.ru/ru-en-src-retr-dataset/

4.1 Indexing of Wikipedia

We indexed all articles from English (5.8M) and Russian (1.5M) Wikipedia dumps (June 2019).

4.1.1 Retrieval-based approach

We use TF-IDF weighting scheme with $\log_{len(D)+1}(Cnt(w_i) + 1)$ as TF weight for word w_i from a document D, and $max(0, \log_{10}(N - w_{\text{cnt}} + 0.5)/(w_{\text{cnt}} + 0.5))$ as IDF, where N is total amount of documents in a collection.

4.1.2 Approximate nearest neighbor search

We take top keywords with weight > 0.05 and average embeddings of those words. We use Faiss IVFFlatIndex [23] for indexing document embeddings with the following parameters: number of centroids - $4 * \sqrt{|V|}$, where V set of all vectors that we need to index, training set size – $5 \times min\_points\_per\_centroid \times num\_centroids$, where $min\_points\_per\_centroid$ is equal 39 by default, nprobe – 16, compression – SQfp16. Our experiments showed that these parameters result in efficient search time and search precision greater than 90%.

4.1.3 ESA

When precomputing concept vectors for ESA method, we used 200 top keywords (with weight > 0.05) of a document to compute weights of concepts. We kept the maximum 1200 concepts with the largest weight per document. Since we build vectors of Wikipedia articles using Wikipedia articles as concepts, we excluded a concept that represents the same article from the vectors.

5. Evaluation Results

We used grid search for parameters tuning on 400 documents that were sampled independently of the testing data. We performed a search on all 1000 documents using various methods, retrieved the most similar 600 documents and measured standard metrics: Recall, MAP. We use the following abbreviation in the table 3 and below:

- RBA – Retrieval-based approach;
- EMB – Embeddings that were used: BIL - embeddings built on bilingual corpus, MAP - embeddings mapped via Vecmap framework;
- MP – Maximum phrase size (1-4), if 1 the keywords may only contain single words;
- N – Number of similar words in other language that were taken for each word when mapping keywords (1 if not specified explicitly);
- MTS – Number of keywords in other language (mapped top size) (100 if not specified explicitly);
- SK – Similarity score: cosine (cos) or hamming (ham) similarity measures (cos if not specified explicitly);
- ANN – Approximate nearest neighbor search;
- DIM – Dimensionality of embeddings (300 if not specified explicitly);
- K – Document is an average of vectors of the top K keywords;
- ESA – Explicit semantic analysis;
- CTOP – Number of concepts to use for retrieval.

The Table 3 displays the evaluation results obtained on the wiki dataset.

Table 3. Evaluation results

Method	Rec@1	Rec@10	Rec@20	Rec	MAP
RBA (EMB=BIL,MP=1)	0.415	0.622	0.66	0.831	0.48
RBA (EMB=BIL,MP=2)	**0.418**	**0.632**	**0.67**	0.843	**0.49**
RBA (EMB=BIL,MP=4)	0.415	**0.635**	**0.67**	0.845	**0.49**
RBA (EMB=BIL,DIM=600,MP=4)	**0.428**	0.629	**0.671**	**0.856**	**0.5**
RBA (EMB=BIL,MP=4,SK=ham)	0.387	0.611	0.661	**0.849**	0.467
RBA (EMB=MAP,MP=4)	0.263	0.478	0.533	0.767	0.336
ANN (EMB=BIL,MP=2,K=25)	0.313	0.508	0.548	0.715	0.379
ANN (EMB=BIL,MP=2,K=50)	0.337	0.508	0.548	0.728	0.398
ANN (EMB=BIL,MP=2,K=100)	0.266	0.433	0.475	0.689	0.323
ANN (EMB=BIL,DIM=600,MP=2,K=50)	0.374	0.527	0.577	0.724	0.433
ANN (EMB=MAP,MP=2,K=50)	0.197	0.36	0.426	0.665	0.254
ESA (CTOP=200,SK=cos)	0.254	0.453	0.501	0.833	0.318

The results show that the retrieval-based approach is better in terms of Recall and Map than other methods. The embeddings, built on the bilingual corpus (EMB=BIL), give better results for this task than embeddings obtained via mapping (EMB=MAP). The results of experiments show that syntactic phrases give no significant boost in performance for RBA and ANN approaches. Doubling the number of components of embeddings from 300 to 600 results in better ranking for ANN aproach, but almost has no effect for RBA. ESA shows good recall, but ranking of found documents is worse than for the RBA and ANN methods.

It should be pointed out that the performance of the methods differs significantly depending on the size of the documents (table 4).

Table 4: RBA (EMB=BIL,MP=4, MTS=100/50) Metrics per each size group

No	size in ru sents	comparable by size?	MAP (MTS=100)	Rec (MTS=100)	MAP (MTS=50)	Rec (MTS=50)
1	(9, 50]	False	0.346	0.82	0.346	0.82
2	(9, 50]	True	0.338	0.65	0.338	0.65
3	(50, 100]	False	0.419	0.79	0.419	0.8
4	(50, 100]	True	0.44	0.88	0.445	0.88
5	(100, 200]	False	0.461	0.81	0.453	0.82
6	(100, 200]	True	0.542	0.88	0.535	0.9
7	(200, 400]	False	0.451	0.79	0.473	0.8
8	(200, 400]	True	0.730	0.98	0.742	0.97
9	(400, 1000]	False	0.306	0.85	0.341	0.81
10	(400, 1000]	True	0.87	1	0.871	1

RBA method works better with long texts that are comparable by size. Short texts (groups 1,2) are likely to have some specific out-of-vocabulary lexis, and remaining words do not help to retrieve the article in other language and rank it highly. The lowest MAP is for the combination 9, i.e. the non-comparable by size long texts, whereas the best MAP is for the longest texts also, but this time for comparable ones (group 10). It can be seen from table 2 that Russian texts from this group are longer than English texts the factor of two. These articles may be devoted to Russian concepts that have short descriptions in English. In this case, similar articles with longer

texts have more chances to share lexis than shorter articles. Therefore, shorter texts are lower in the rank list.

Similar behavior is observed for other methods too. For example, table 5 presents the comparison of results for different K (50, 25) for ANN method.

Table 5. ANN (EMB=BIL, MP=2, K=50/25) Metrics per each size group

No	Size in ru sents	comparable by size?	MAP (K=50)	Rec (K=50)	MAP (K=25)	Rec (K=25)
1	(9, 50]	False	0.228	0.59	0.192	0.55
2	(9, 50]	True	0.226	0.61	0.351	0.69
3	(50, 100]	False	0.237	0.53	0.214	0.52
4	(50, 100]	True	0.482	0.72	0.433	0.69
5	(100, 200]	False	0.334	0.72	0.352	0.74
6	(100, 200]	True	0.562	0.81	0.472	0.79
7	(200, 400]	False	0.328	0.73	0.357	0.71
8	(200, 400]	True	0.56	0.87	0.507	0.81
9	(400, 1000]	False	0.272	0.74	0.229	0.74
10	(400, 1000]	True	0.754	0.96	0.688	0.91

There are some groups where performance is better with a lesser amount of keywords, e.g., 2, 5 (recall and MAP), and 7 (MAP). This result suggests that our strategy of selecting keywords (select up to N words with weight > X) does not work well for all cases.

6. Conclusion

In this article, we compared various methods for cross-lingual retrieval of similar documents. We employed classical inverted indexes combined with cross-lingual embeddings and pure continuous retrieval using ANN. For this task and our dataset, the best result was shown by retrieval-based approach. It achieves the best recall and MAP scores with long comparable by size texts. As future work, we need to focus on improving performance for non-comparable by size texts, since now its ranking is far from the good. Dealing with OOV is another important issue. One way to solve it is to employ subword vector representation to encode the OOV-words [24]. Another way is to extending vocabulary from different comparable corpora: scientific papers, patents, etc. containing a lot of special lexis and terms. One of the possible solutions is to use the system for translated plagiarism detection to extract parallel sentences from comparable corpora [25].

References / Список литературы

[1]. Romanov A., Kuznetsova R., Bakhteev O., Khritankov A. Machine-Translated Text Detection in a Collection of Russian Scientific Papers. In Proc. of the Annual International Conference “Dialogue”, 2016.

[2]. Zubarev D.V., Sochenkov I.V. Cross-language text alignment for plagiarism detection based on contextual and context-free models. In Proc. of the Annual International Conference “Dialogue” 2019, v. 1, pp. 799-810.

[3]. Ferrero J., Agnes F., Besacier L., Schwab D. Using Word Embedding for Cross-language Plagiarism Detection. arXiv:1702.03082, 2017.

[4]. Franco-Salvador M., Gupta P., Rosso P., Banchs R.E. Cross-language plagiarism detection over continuous space and knowledge graph-based representations of language. Knowledge-based systems, vol. 111, 2016, pp. 87-99.

[5]. Jiang J.Y., Zhang M., Li C., Bendersky M., Golbandi N., Najork M. Semantic Text Matching for Long-Form Documents. In Proc. of the World Wide Web Conference, 2019, pp. 795-806.

[6]. Gillick D., Presta A., Tomar G.S. End-to-End Retrieval in Continuous Space. arXiv:1811.08008, 2018.
[7]. Vulić I., Moens M.F. Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In Proc. of the 38th international ACM SIGIR conference on research and development in information retrieval, 2015, pp. 363-372.
[8]. Barrón-Cedeño A., Gupta P., Rosso P. Methods for cross-language plagiarism detection. Knowledge-Based Systems, vol. 50, 2013, pp. 211-217.
[9]. Potthast M., Barrón-Cedeño A., Stein B., Rosso P. Cross-language plagiarism detection. Language Resources and Evaluation, vol. 45, issue 1, 2011, pp. 45-62.
[10]. Bakhteev O., Ogaltsov A., Khazov A., Safin K., Kuznetsova R. CrossLang: the system of cross-lingual plagiarism detection. In Proc. of the KDD Workshop on Deep Learning for Education, 2019. Available at: https://truth-discovery-kdd2019.github.io/papers/crosslang.pdf, accessed 15.11.2019
[11]. Kutuzov A., Kopotev M., Sviridenko T., Ivanova L. Clustering Comparable Corpora of Russian and Ukrainian Academic Texts: Word Embeddings and Semantic Fingerprints. In Proc. of the Ninth Workshop on Building and Using Comparable Corpora, 2016, pp. 3-10
[12]. Straka M., Hajic J., Straková J. UDPipe: trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, pos tagging and parsing. In Proc. of the tenth international conference on language resources and evaluation, 2016, pp. 4290-4297.
[13]. Tiedemann J. Parallel Data, Tools and Interfaces in OPUS. In Proc. of the language resources and evaluation (LREC), 2012, pp. 2214-2218.
[14]. Antonova A., Misyurev A. Building a web-based parallel corpus and filtering out machine-translated text. In Proc. of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web, 2011, pp. 136-144.
[15]. Ruder S., Vulić I., Søgaard A. A survey of cross-lingual word embedding models. Journal of Artificial Intelligence Research, vol. 65, issue 1, 2019, pp. 569-631.
[16]. Mikolov T., Sutskever I., Chen K., Corrado G.S., Dean J. Distributed representations of words and phrases and their compositionality. In Proc. of the 26th International Conference on Neural Information Processing Systems, 2013, vol. 2, pp. 3111-3119.
[17]. Artetxe M., Labaka G., Agirre E. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In Proc. of the Thirty-Second AAAI Conference on Artificial Intelligence, 2018, pp. 5012-5019.
[18]. Glavas G., Litschko R., Ruder S., Vulic I. How to (Properly) Evaluate Cross-Lingual Word Embeddings: On Strong Baselines, Comparative Analyses, and Some Misconceptions. arXiv:1902.00508, 2019.
[19]. Conneau A., Lample G., Ranzato M. A., Denoyer L., Jégou H. Word translation without parallel data. arXiv:1710.04087, 2017.
[20]. Vulić I., Moens M.F. Bilingual word embeddings from non-parallel document-aligned data applied to bilingual lexicon induction. In Proc. of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, 2015, vol. 2, pp. 719-725.
[21]. Sochenkov I.V., Zubarev D.V., Tikhomirov I.A. Exploratory patent search. Informatics and its Applications, vol. 12, issue 1, 2018, pp. 89-94 (in Russian). / Соченков И.В., Зубарев Д.В., Тихомиров И.А. Эксплоративный патентный поиск. Информатика и ее применения, том 12, вып. 1, 2018 г., стр. 89-94..
[22]. Gabrilovich E., Markovitch S. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In Proc. of the 20th international joint conference on Artifical intelligence, 2007, pp. 1606-1611.
[23]. Johnson J., Douze M., Jégou H. Billion-scale similarity search with GPUs. arXiv:1702.08734, 2017.
[24]. Bojanowski P., Grave E., Joulin A., Mikolov T. Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 2017, vol. 5, pp. 135-146.
[25]. Zweigenbaum P., Sharoff S., Rapp R. Overview of the third BUCC shared task: Spotting parallel sentences in comparable corpora. In Proc. of 11th Workshop on Building and Using Comparable Corpora, 2018, pp. 39-42.

Информация об авторах / Information about authors

Денис Владимирович ЗУБАРЕВ – инженер-исследователь ФИЦ ИУ РАН. Сфера научных интересов: информационный поиск, компьютерная лингвистика, поиск текстовых заимствований, анализ больших массивов данных.

Denis Vladimirovich ZUBAREV – Engineer of FRC CSC RAS. Research interests: information retrieval, natural language processing, plagiarism detection, big data.

Илья Владимирович СОЧЕНКОВ – кандидат физико-математических наук, заведующий отделом «Интеллектуальные технологии и системы» ФИЦ ИУ РАН. Область научных интересов: интеллектуальные методы поиска и анализа информации, компьютерная лингвистика, компьютерный анализ естественного языка, машинное обучение, анализ больших массивов данных.

Ilya Vladimirovich SOCHENKOV – PhD, Head of the Department of Intelligent Technologies and Systems of FRC CSC RAS. Research interests: information retrieval, natural language processing, machine learning, big data.