RusNLP: Semantic search engine for Russian NLP conference papers

Irina Nikishina[1][0000−0003−4910−8586], Amir Bakarov[1][0000−0003−4268−5979], and Andrey Kutuzov[2][0000−0003−2540−5912]

[1] National Research university Higher School of Economics, Moscow, Russia
ianikishina_1@edu.hse.ru, amirbakarov@gmail.com
[2] University of Oslo, Oslo, Norway
andreku@ifi.uio.no

Abstract. We present RusNLP, a web service implementing semantic search engine and recommendation system over proceedings of three major Russian NLP conferences (Dialogue, AIST and AINL). The collected corpus spans across 12 years and contains about 400 academic papers in English. The presented web service allows searching for publications semantically similar to arbitrary user queries or to any given paper. Search results can be filtered by authors and their affiliations, conferences or years. They are also interlinked with the NLPub.ru service, making it easier to quickly capture the general focus of each paper. The search engine source code and the publications metadata are freely available for all interested researchers.
In the course of preparing the web service, we evaluated several well-known techniques for representing and comparing documents: TF-IDF, LDA, and Paragraph Vector. On our comparatively small corpus, TF-IDF yielded the best results and thus was chosen as the primary algorithm working under the hood of RusNLP.

Keywords: information retrieval, semantic similarity, scientific literature search, document representations, academic communities

1 Introduction

Russian natural language processing community is thriving and has a strong and interesting scholarly legacy: three leading venues indexed in the Scopus database, 20 years of active publishing, more than 1800 conference papers and dozens of smaller workshops, schools and conferences. Unfortunately, it is sometimes difficult to find NLP papers relevant for processing of Russian (for example, Russian named entity recognition) with mainstream scholarly search engines (like *Google Scholar*, *Semantic Scholar* or *Sci-Hub*). Russian NLP community lacks a unified database that could incorporate all knowledge accumulated by previous researchers and present it in a structured and interlinked way. Such a database (with the corresponding front-end user interface and a recommendation system) can be of great help to those dealing with NLP problems related to Russian or those who study the structure of NLP academic and industrial communities.

In this paper, we describe the preparation and implementation of such a resource available online for everyone. The database and the corresponding semantic search engine are a part of the larger *RusNLP* project aimed at preserving and analyzing the structure and tendencies in the Russian computational linguistics community.

Note that it was not clear from scratch which document analysis technique is most efficient for recommending semantically similar texts on a relatively small collection of domain-restricted documents (in our case, papers from the Russian NLP conferences). Thus, we had to conduct a bunch of experiments with the well-known document representation algorithms and human annotators before choosing the best one for our task. These experiments are described below.

The rest of the paper is organized as follows. In section 2, we briefly present the collected dataset. Section 3 delves into the evaluation experiments comparing different approaches to document representation regarding our task. Section 4 describes the features of the *RusNLP* web service and typical use cases. In section 5, we put our work in the context of the previous research, while in section 6 we conclude and outline future plans for the *RusNLP* project.

2 Dataset

The structure of our dataset (and thus, our publications corpus) is thoroughly described in [1]. That is why here we outline it very briefly.

In general, the *RusNLP* dataset contains texts and metadata for 1,794 academic papers (and 900 authors) related to computational linguistics and natural language processing, similar to the ACL Corpus [3]. Globally, NLP as a scientific field tends to value conferences at least as high as journals; the same is true for the Russian NLP scene. Thus, the papers in our collection come from the main program proceedings of 3 Russian conferences related to computational linguistics: **Dialogue**[3], **AIST**[4] and **AINL**[5]. We do not claim that this covers the whole Russian NLP landscape, but at least the most important venues are included, with the hope that in the future more will come.

The papers themselves were crawled from the **Dialogue** website and from the Springer digital library (for **AIST** and **AINL**). Each paper was parsed to extract metadata (title, abstract, authors' names, affiliations and emails). The author names and affiliations were normalized manually (for example, we merged '*Bauman Moscow State Technical University*' and '*МГТУ имени Н.Э. Баумана*' into one entity, etc.).

The database contains all the papers, but the *RusNLP* web service (see below) for the time being deals only with those written in English: 392 in total. This decision was made to streamline text processing workflow, and considering the current tendency to publish most important research in English (**AINL** and **AIST** proceedings contain only papers in English, so this distinction is relevant

[3] `http://www.dialog-21.ru/en/`
[4] `https://aistconf.org/`
[5] `http://ainlconf.ru/`

only for **Dialogue**). In the future, we plan to integrate papers written in Russian into the search engine as well.

3 Experimental Setup

The purpose of our system is to find and recommend NLP papers (from the dataset described above) most similar by their content to a given paper or a given user query. Thus, some way to represent and compare texts had to be chosen. In this section, we describe the experiments we conducted to this end. Note that first all the 392 texts were pre-processed: that is, cleared from non-alphanumeric characters, and then lemmatized and PoS-tagged with *UDPipe* [15].

We tested three statistical approaches to document representation:

1. *TF-IDF* (term frequency - inverted document frequency), a term weighting scheme ubiquitous in information retrieval [14];
2. *LDA* (Latent Dirichlet Allocation), a widespread distributional topic modeling technique [4]; we tried variants with 10 and 20 topics;
3. *Paragraph Vector* (also known as *doc2vec*), a newer distributional parametric algorithm based on shallow feed-forward neural networks [9]; we tried variants with vector size 40 and 100.

These techniques are conceptually very different, with *TF-IDF* being the simplest and taking into account only word frequencies in the documents, *LDA* trying to model the hidden distribution of topics in the text collections, and *Paragraph Vector* extending the well-known *SGNS* and *CBOW* word embedding algorithms [11] to learn embeddings for sentences, paragraphs, or documents. We aimed to find out which of them would be the best in ranking documents by their similarity in our corpus[6].

Our evaluation workflow was quite simple and measured only the precision of the algorithms. First, 20 papers were randomly sampled from the collection (8 from **Dialogue**, 6 from **AINL**, and 6 from **AIST**). A set of 10 most similar documents (nearest neighbors) was produced for each of these papers, using each of the 5 models described above (*TF-IDF*, two variants of *LDA* and two variants of *Paragraph Vector*). Then, three human assessors independently annotated the sets with the number of non-relevant neighbors t, reflecting the amount of noise in the output of the model. The precision of each set is thus $1 - \frac{t}{10}$, and the precision of the model is the averaged precision for all 20 sets: $\frac{\sum_{t=1}^{20} t_i}{20}$. Note that there is no recall measurement in this setup, as this would require assessors to read through all the 400 papers to find relevant publications not selected by the models. This was not feasible with our time and human resources, and thus these experiments are limited to measuring only the models' precision (their ability to avoid absolutely non-relevant papers in the 10 nearest neighbors). We

[6] The models were trained on our English sub-corpus, using the algorithm implementations in the *Gensim* library [13].

leave finding the recall scores for future work. Additionally, we calculated the inter-rater agreement using Krippendorff's alpha [8].

We also considered using author-provided keywords as a 'gold standard' for evaluation (for example, with Jaccard similarity for keyword lists as the ground truth semantic distance between documents), but we found that in most of the documents the keywords were not actually representative since keyword choice for a paper is a highly subjective process.

Model	Assessor 1	Assessor 2	Assessor 3	Average	Agreement
TF-IDF	0.6	0.65	0.68	**0.64**	0.73
Paragraph Vector - 40	0.28	0.33	0.4	0.33	0.66
Paragraph Vector - 100	0.36	0.36	0.45	0.39	0.5
LDA - 10	0.17	0.2	0.21	0.2	0.66
LDA - 20	0.23	0.19	0.39	0.27	0.52

Table 1: Precision scores and inter-rater agreement for the tested models.

Table 1 reports the results of the comparison. First, for all models, the inter-rater agreement is at quite decent level, meaning the results are trustworthy[7]. The second outcome is somewhat unexpected: the simplistic *TF-IDF* model produced way better results than the sophisticated *LDA* and *Paragraph Vector* distributional algorithms.

We hypothesize that the reason for this is the size of our corpus, which is comparatively small. It seems that 400 documents with total word count of 1,340,957 are not enough to train a distributional model able to outperform the frequency-based approach. Note though that *TF-IDF* is in fact often used in production systems instead of modern algorithms, providing a good trade-off between performance, speed and model size: for example, [5] employ it as the main ranking model for their life science publications search engine. So, following these experiments, we used TF-IDF to implement the search engine and recommendation system described in the next section.

4 RusNLP web service: key features

The *RusNLP* search engine provides web access to the collection of Russian NLP papers described above. Under the hood, it employs an *SQLite* database to store documents metadata and a *TF-IDF* model to find similar papers. For simplicity, and because of comparatively small absolute size of the corpus, at this moment

[7] Initially, the agreement levels for the *TF-IDF* and *LDA-10* models were below 0.5. We performed a reconciliation round with the assessors discussing their choices for these models, which resulted in changing some of the scores for particular documents. This increased the inter-rater agreement, but did not influence the final ranking.

we do not use the existing production-level text search libraries like *Sphinx*, *Solr* or *ElasticSearch*. However, in the future we plan to test whether using these libraries will improve user experience in our case.

Right now, the service already allows its users to:

- Find and rank papers most similar to an arbitrary user query (list of keywords); see Figure 1 in Appendix A for an example.
- Find and rank papers most similar to any given paper; see Figure 2 in Appendix A for an example.
- Filter search results by any combination of user-defined criteria:
  - publication venue,
  - publication year,
  - author names,
  - author affiliations,
  - paper title.

We envisage that *RusNLP* can be used either for the search of relevant previous work ('*I know this paper, what other similar papers are there in Russian NLP?*') or for studies in the history and structure of academic communities ('*What was published in 2008 by NLP scholars from Moscow State University?*' or '*Were there any papers about paraphrases detection at the **AINL** conference in 2015?*'). Figure 3 in Appendix A shows an example of a query for all papers presented at **Dialogue** or **AIST** by any author whose name starts with 'Chernj'.

For many papers, we also provide the so called 'task tags'. They link to the corresponding resource sections of the *NLPub* web service[8] [16]. For example, a hypothetical paper using a thesaurus for emotion detection would be interlinked with the *NLPub* pages listing existing tools for tasks related to thesauri and sentiment analysis. To this end, we manually selected sets of English keywords for each *NLPub* resource section. At query time, our service calculates the *TF-IDF* similarity of all papers found to each of these sets. If the similarity exceeds a predefined threshold (after some experimenting, we set it to 0.03), the paper is assigned the corresponding 'task tag'. This feature allows the users to find out at a glance what the paper is about and to quickly get an idea about the existing tools or resources relevant to this task.

The *RusNLP* web service we describe in this paper is publicly available at `http://nlp.rusvectores.org`.

5 Related Work

The importance of compiling reference in-domain datasets of texts related to NLP was expressed more than 10 years ago in [3]. This paper was the first towards describing and releasing a corpus of computational linguistics academic papers and their metadata: this dataset is now known as the *Association for Computational Linguistics Anthology*[9].

[8] `https://nlpub.ru`
[9] `http://aclanthology.info/`

However, the *ACL Anthology* is not enriched with semantic search or a recommendation system. There exist more functional search engines and databases that are less related to NLP and computational linguistics: see, for instance, *Pulp* that includes semantic search engine, visualization of search topics (with topic modeling), and document-ranking algorithm [10]. Another system, *Bib2Vec* [17], embeds bibliographic information in a vector space and then uses these vectors to show relationships among entities in the ACL Reference Corpus. There are more and more similar service appearing each year, like *Semedico*, a life science semantic search engine [5], or *Citeomatic* which 'recommends papers you might want to cite' [2]. As for the Russian academic community, we are unaware of any domain-specific Russian scholarly search engines: there exist only global bibliographic services for all fields of Russian science like *CyberLeninka* or *Russian Science Citation Index* (RSCI).

Note that our research is highly related to studies in information retrieval (IR) on scientific publications. For instance, proceedings of International Workshop on Mining Scientific Publications (WOSP) provide a substantial amount of papers devoted to 'Automatic categorization and clustering of scholarly data' and 'Academic recommendation systems'. One can mention [6] and [12], that implement rather sophisticated approaches for metadata extraction which we plan to test and possibly use in the *RusNLP* project.

Considering previous research in the field of Russian NLP, the analysis of Russian NLP landscape was arguably started by Vladimir Khoroshevsky [7]. Their research is similar to our project: they described academic communities, conference activity and links between scholar clusters for Russian computational linguistics community. We improve on this work in 3 aspects:

1. [7] is published in 2012, and uses data from 2009; our project mines conference proceedings up to 2017, and includes novel publishing venues (AIST and AINL). We believe that since 2009, the Russian NLP community significantly changed with regards to its primary trends and topics of interest; this is supported by our analysis of diachronic topical drift in Russian NLP academic papers [1].
2. We implement the semantic search engine available online as a web service.
3. In [7], no datasets were released; we publish all our source code[10] and data[11]. Note that our contribution is not only crawling papers from the venues web sites, but also in consistent manual extraction and normalization of papers' metadata).

Thus, we argue that our project is novel, and is a step towards unification and structuring the research within the Russian NLP community.

6 Conclusion

In this paper, we described *RusNLP*, a web service for scholarly search in the collection of natural language processing papers. This is the first publicly available

[10] `https://github.com/bakarov/rusnlp/tree/master/code/web`
[11] `http://nlp.rusvectores.org/about/`

and manually curated database of publications in the most important Russian computational linguistics venues. Of course, contemporary science is global and as a rule one shouldn't limit herself to only papers published in this or that country. However, sometimes (especially in research dealing with particular language) it can be useful to focus more on the local academic landscape. This can be also interesting for those studying academic communities and the trends in NLP publishing activities.

Additionally, we described a set of experiments conducted to find out the most efficient algorithm for finding similar papers in our dataset. The outcome was that the simple *TF-IDF* model outperformed the sophisticated distributional algorithms of *LDA* and *Paragraph Vector* by a large margin, most probably because of comparatively small training corpus size. This model powers the nearest neighbors search under the hood of our web service (`http://nlp.rusvectores.org`).

The efforts described in this paper are part of a larger project aiming to analyze Russian NLP landscape (thus, the dataset is not complete and will be updated yearly). In the future, we plan to make it possible to use *RusNLP* to search for papers written in Russian, not only in English, thus making the resource multilingual. Another important direction in our nearest plans is the construction of the citation graph from all the papers in our database, which will allow to explore connections between scientific communities. Finally, we are considering the possibility of improving our keywords search by expanding user queries with the help of pre-trained word embedding models.

Acknowledgments

We thank numerous VPNs and *Tor Project*. At the time of finalizing this paper, they were the only ways for Russian-based scholars to collaborate with the colleagues abroad, because of Internet censorship carried by the Russian governmental agency called *Roskomnadzor*. It accidentally managed to temporarily block a whole bunch of academic resources, including *Softconf*, *Overleaf*, etc.

References

1. Bakarov, A., Kutuzov, A., Nikishina, I.: Russian computational linguistics: topical structure in 2007-2017 conference papers. In: Proceedings of Dialogue-2018, online papers. ABBYY (2018), `http://www.dialog-21.ru/media/4249/bakarov_kutuzov.pdf`
2. Bhagavatula, C., Feldman, S., Power, R., Ammar, W.: Content-based citation recommendation. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). pp. 238–251. Association for Computational Linguistics (2018), `http://aclweb.org/anthology/N18-1022`
3. Bird, S., Dale, R., Dorr, B., Gibson, B., Joseph, M., Kan, M.Y., Lee, D., Powley, B., Radev, D., Tan, Y.F.: The ACL Anthology reference corpus: A reference dataset for

bibliographic research in computational linguistics. In: LREC 2008 (2008), http://www.aclweb.org/anthology/L08-1005
4. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. Journal of Machine Learning Research **3**(Jan), 993–1022 (2003)
5. Faessler, E., Hahn, U.: Semedico: a comprehensive semantic search engine for the life sciences. Proceedings of ACL 2017, System Demonstrations pp. 91–96 (2017)
6. Kern, R., Jack, K., Hristakeva, M., Granitzer, M.: Teambeam - meta-data extraction from scientific literature. In: Knoth, P., Zdrahal, Z., Juffinger, A. (eds.) Special Issue on Mining Scientific Publications, D-Lib Magazine, vol. 18, number 7/8. Corporation for National Research Initiatives (July 2012)
7. Khoroshevsky, V.: Пространства знаний в сети Интернет и Semantic Web, Часть 3 (Knowledge spaces in the Internet and Semantic Web, part 3); in Russian. Искусственный интеллект и принятие решений (Artificial Intelligence and Decision Making) pp. 3–38 (2012)
8. Krippendorff, K.: Content analysis: An introduction to its methodology. Sage (2012)
9. Le, Q., Mikolov, T.: Distributed representations of sentences and documents. In: International Conference on Machine Learning. pp. 1188–1196 (2014)
10. Medlar, A., Ilves, K., Wang, P., Buntine, W., Glowacka, D.: Pulp: A system for exploratory search of scientific literature. In: Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval. pp. 1133–1136. ACM (2016)
11. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. Advances in Neural Information Processing Systems 26 pp. 3111–3119 (2013)
12. Nanni, F., Dietz, L., Faralli, S., Glavaš, G., Ponzetto, S.P.: Capturing interdisciplinarity in academic abstracts. D-lib magazine **22**(9/10) (2016)
13. Řehůřek, R., Sojka, P.: Software Framework for Topic Modelling with Large Corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. pp. 45–50. Valletta, Malta (May 2010)
14. Sparck Jones, K.: A statistical interpretation of term specificity and its application in retrieval. Journal of documentation **28**(1), 11–21 (1972)
15. Straka, M., Straková, J.: Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In: Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. pp. 88–99 (2017)
16. Ustalov, D.: NLPub: a catalogue and a community for Russian linguistic resources. In: Selected Papers of XVI All-Russian Scientific Conference "Digital libraries: Advanced Methods and Technologies, Digital Collections". vol. 1297, pp. 56–60. RWTH (2014)
17. Yoneda, T., Mori, K., Miwa, M., Sasaki, Y.: Bib2vec: Embedding-based search system for bibliographic information. In: Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics. pp. 112–115. Association for Computational Linguistics (2017), http://aclweb.org/anthology/E17-3028

Appendix A

Fig. 1: Searching scholarly papers in the database by user-provided query words (in this case, 'syntax neural classification').

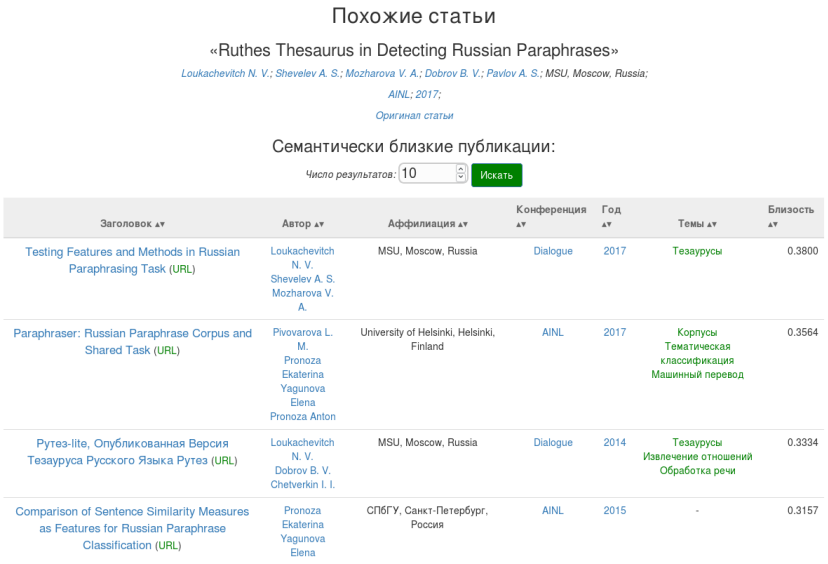

Fig. 2: RusNLP recommending papers similar to the query paper.

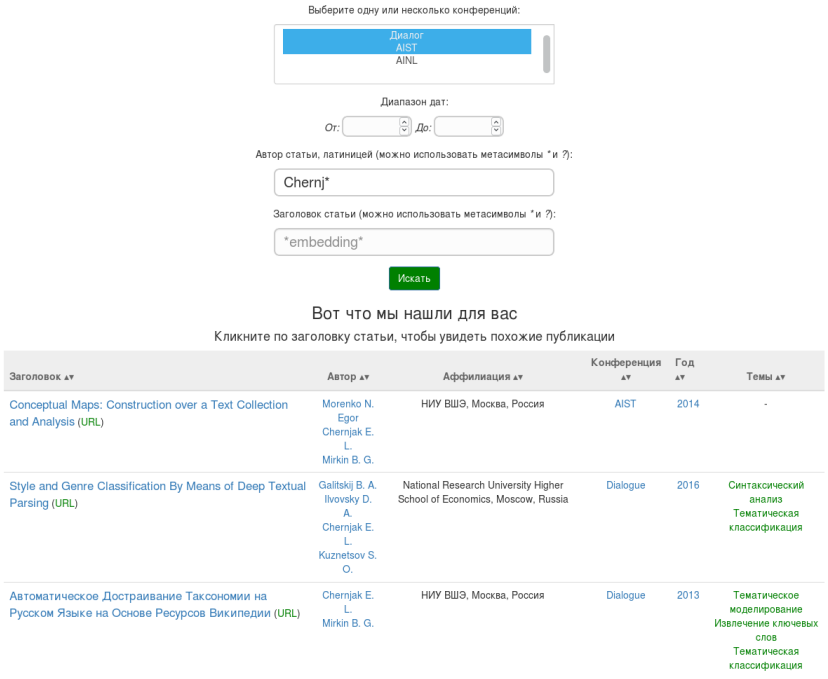

Fig. 3: RusNLP searching for all papers by a certain author in the **Dialogue** and **AIST** conferences.