RUSSIAN COMPUTATIONAL LINGUISTICS: TOPICAL STRUCTURE IN 2007-2017 CONFERENCE PAPERS

Bakarov A. (amirbakarov@gmail.com)
National Research University Higher School of Economics, Russia,
Federal Research Center “Computer Science and Control” of Russian Academy of Sciences, Russia
Kutuzov A. (andreku@ifi.uio.no)
University of Oslo, Norway
Nikishina I. (irina.nikishina@mail.ru)
National Research University Higher School of Economics, Russia

Russian NLP community exists for at least several decades. However, academic works analyzing it are scarce. The present paper fills in this gap by topical modeling of the proceedings of three major Russian NLP conferences (Dialogue, AIST and AINL) for the years from 2007 to 2017. The resulting corpus consists of about 500 academic papers.
We focus on the analysis of developing research trends manifested in topical drift over time. As a result, we show statistically how Russian NLP community interests are moving towards machine learning and how the Dialogue (as the largest venue) influences the whole computational linguistics landscape.

Keywords: document clustering, topical drift, distributional semantics, document representations, academic communities

АНАЛИЗ ТЕМАТИЧЕСКОЙ СТРУКТУРЫ ПУБЛИКАЦИЙ НА РОССИЙСКИХ КОМПЬЮТЕРНО-ЛИНГВИСТИЧЕСКИХ КОНФЕРЕНЦИЯХ ЗА 2007-2017 ГОДЫ

Бакаров А. (amirbakarov@gmail.com)
Национальный исследовательский университет Высшая Школа Экономики, Россия
Кутузов А. (andreku@ifi.uio.no)
Университет Осло, Норвегия
Никишина И. (irina.nikishina@mail.ru)
Национальный исследовательский университет Высшая Школа Экономики, Россия

Ключевые слова: кластеризация документов, сдвиг тематики, дистрибутивная семантика, репрезентации документов, академические сообщества

1. Introduction

The theory of language universals states that there are many patterns that occur systematically across most of natural languages. However, natural language processing (NLP) methods for different languages could also be very different. When talking about an NLP method we usually mean not a method for processing natural language in general, but one for a particular human language. Thus, many NLP research communities are focused on processing their national languages, like French or Spanish (most members of such communities are native speakers of the language). Research questions pursued by such communities can range from the creation of language-specific datasets to testing the applicability of methods proposed for other languages, etc.

Russian language is not an exception, enjoying rich and diverse computational linguistics and NLP community[1]. Its roots go back to the middle of the XX century [Lyashevskaya et al., 2018], when computational linguistics gained global popularity following the Georgetown experiment in machine translation. In those years, NLP was closely related to theoretical linguistics, and it is difficult to track which venues were exclusively related to NLP; arguably, the oldest Russian major computational linguistics venue is the *Models of Communication* workshop which has been held since the 1970s until 1995.

In 2018, there are three major conferences related to Russian NLP: *Dialogue*[2] (started in 1995), *Analysis of Images, Social networks and Texts Conference*[3] (AIST; started in 2012) and *Artificial Intelligence and Natural Language Conference*[4] (AINL; started in 2015). There is also a considerable amount of smaller venues like NLP-related sections and workshops at conferences in various subjects: mostly data science (for example, *Data Analytics and Management in Data Intensive Domains*[5]*)*, and computer science (for example, *Ivannikov ISPRAS Open Conference*[6]). There are also NLP-related summer schools: *Russian Summer School in Information Retrieval (RuSSIR)*[7]*, MIEM Computational Linguistics School*[8], and others.

The proceedings of the major venues constitute a valuable resource promoting NLP research. However, we argue that they can also be an object of study themselves, as a large academic texts collection. One of possible (and, probably, the most interesting) directions for analysis of this corpus is the *detection of temporal topical drifts*. By this, we mean detection of topic change in a collection of documents annotated with timestamps. Solving this problem can provide many interesting insights. For example, we already mentioned that in the past

[1] Further on, we will use the terms ‘computational linguistics’ and ‘natural language processing’ (NLP) interchangeably.

[2] http://www.dialog-21.ru/en

[3] http://aistconf.ru

[4] http://ainlconf.ru

[5] http://damdid2017.frccsc.ru/en/conference_short.html

[6] http://www.isprasopen.ru/en/

[7] http://romip.ru/russir2018/

[8] https://miem.hse.ru/clschool/

Russian NLP community was closely related to general linguistics, but does it still hold up to now? Is Russian NLP more about data science than about linguistics in 2018? Are the topics that were trendy a decade ago still popular now?

To make this research possible, we collected a dataset of academic papers presented at the three aforementioned major conferences (*Russian NLP Dataset* or *RusNLP)*. Certainly, the potential uses of this dataset are not limited by the present paper. *Russian NLP Dataset* can be employed to implement semantically-aware information retrieval applications or to study citation networks. We make this dataset available in the form of SQLite database[9] (it does not contain actual texts of papers, as some of them are under prohibitive licenses, but there is always a URL for each paper) and a simple web service to search for papers in the dataset[10].

In contrast to studies that propose new models and approaches for topical drift detection, our work is focused on describing the topical structure of Russian NLP conference papers during the 2007-2017 time span, highlighting differences between major academic venues in terms of research topics and comparing these topics. As a supplementary research question, we also analyze topical difference in three major Russian NLP conferences. So, we represent texts in our dataset with the help of topic modeling methods, as probability distributions on topics for two cases: a) topics for particular venues and b) topics for particular years. Our findings can be helpful for those who are interested in the current trends in Russian computational linguistics.

To sum up, *the primary contributions of our work* are the following:

1. Analysis of Russian NLP topical structure and topical drift was conducted: both between venues and diachronically over time. Diachronic part is limited to the years 2011-2017, due to small amount of English papers before 2011; in other parts of our research, we use papers dating back to 2007.
2. A large dataset of conference papers was collected. We extracted a considerable amount of metadata from this dataset and made it available in a machine-readable format.

The rest of the paper is organized as follows. Section 2 describes our dataset and provides some insights about the documents metadata. Section 3 introduces the experimental setup. Section 4 discusses the results of the experiments. Section 5 briefly highlights the related work, and Section 6 concludes the article.

2. Russian NLP Dataset

As it was mentioned before, Russian computational linguistics and natural language processing field is represented with 3 leading venues: *Dialogue*, *AIST* and *AINL*. *Russian NLP Dataset* was created from the proceedings of these venues: lifetime proceedings of *AIST* and *AINL*, and 2000-2017 proceedings of the *Dialogue*; each paper in the dataset is accompanied with metadata. Of course, these 3 conferences are quite different in their aims

[9] http://nlp.rusvectores.org/data/rus_nlp.db.gz/
[10] http://nlp.rusvectores.org/

and focuses; however, we believe that they represent the major Russian venues to publish NLP or computational linguistics research and papers published at them can be looked at as one domain-specific corpus.

All proceedings of the *Dialogue* starting from 2000 are available online (although the amount of papers in English remains negligible until at least 2007), and it was straightforward to crawl them. Proceedings of *AIST* and *AINL* were manually downloaded from the Springer digital library. It is important that *AIST* and *AINL* also feature papers on other topics: computer vision, social network analysis, etc. We manually filtered out *AIST* papers which appeared in the non-NLP sections of the conference (note that we relied on how the conference organizers had set up their sections).

The Figure 1 presents the number of papers in each of the considered venues starting from 2007. It is easy to see that the *Dialogue* is still the largest computational linguistics event, but *AIST* and *AINL* have also firmly established themselves.

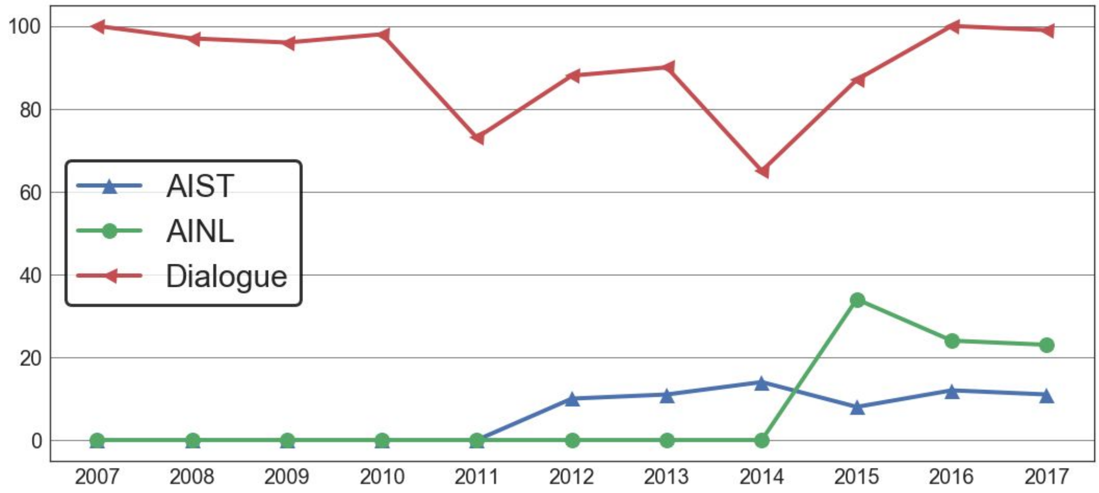

Figure 1. The number of papers collected from each year, depending on the venue

We extracted metadata for each paper (title, abstract, keywords, references and authors' names, affiliations and emails), which sometimes was non-trivial, as each conference has its own conventions on the papers layout. The parsed data was normalized (e.g. we merged 'Higher School of Economics' and 'National Research University Higher School of Economics' into one entity).

English is the primary language of *AIST* and *AINL*, and their proceedings feature English texts only. The problem with the *Dialogue* conference is that it accepts papers both in Russian and in English. Words manifesting topics would obviously differ for papers in different languages, so the experiments had to be performed for both languages separately. Thus, we limited ourselves to the English papers for the time being, leaving cross-linguistic analysis to the future work. The language of the papers was detected via character 3-grams, using the *langid* tool [Lui and Baldwin, 2012]. The Figure 2 shows the amounts of Russian and English papers in different years of the *Dialogue* (including student sections and papers published only in the online version).

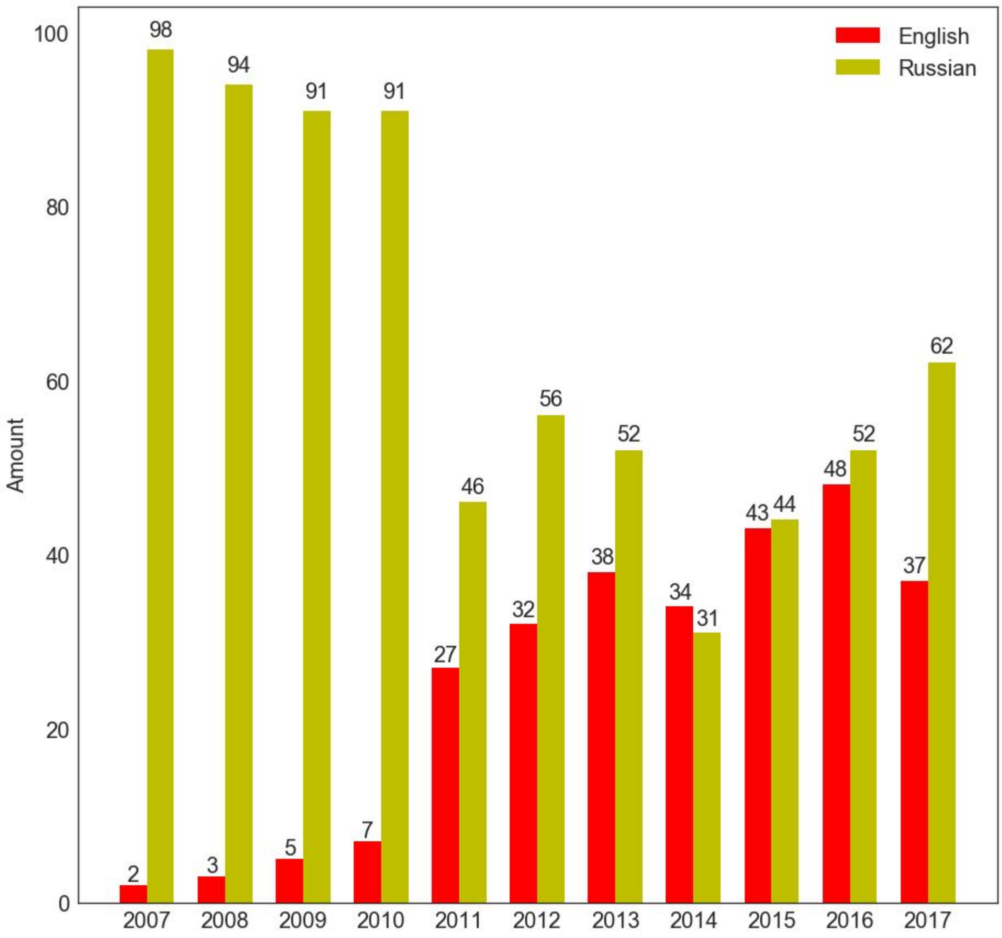

Figure 2. Amounts of papers in Russian and English for different years of the *Dialogue*

The total amount of papers in the dataset is 1866; the corpus containing only English papers (on which our experiments were conducted) consists of 497 papers. The texts were cleared from non-alphanumeric symbols and then lemmatized and POS-tagged with *UDPipe* [Straka et al, 2016]. We also tried to use the versions of texts that were lemmatized with *Stanford CoreNLP* [Manning et al., 2014], stemmed using *Porter Stemme*r, and versions without any morphological normalization; all of them showed lower results in our evaluation experiments. Surely, this pre-processing is very basic: for example, it would most probably be beneficial to glue together multiword expressions. However, for now we leave this for future work.

3. Experimental setting

To detect the latent topics behind the texts in our collection, we used the method of *Latent Dirichlet Allocation (LDA)* [Blei et al., 2003], as implemented in the *Gensim* library [Rehurek & Sojka, 2010]. This model is able to produce the distribution of most probable topics for each document and to give the insights about topical drift over venues and topical drift over years.

We have compared results of modeling with different number of topics up to 10 (we tried higher values, but then the topics became uninterpretable), evaluating the model perplexity on all texts in the dataset. The results of this comparison are presented in Table 1. As one can see, the perplexity gets higher (and thus the models get worse) as the number of topics increases. At the same time, the models with 2 or 3 topics are not informative as well, and perplexity for the models with 4, 5, 7 or 10 topics is almost equal. Thus, we use 10 topics in the further experiments, as this provides most human-interpretable results.

Number of topics	2	3	4	5	7	10
Per-word likelihood bound	-10.73	-10.83	-10.9	-10.95	-11.01	-11.08
Perplexity	1698	1820	1911	1978	2062	2165

Table 1. Perplexity on different amount of topics for LDA on all texts of the dataset.

Figure 3 presents a visualization of our corpus topical structure in the LDA model trained on the whole collection (including papers from the *RuSSIR* summer school). Each document was represented as a 10-D vector of its topic probabilities, which was then projected into 2 dimensions using the Principal Components Analysis (PCA). One can also explore these topics closer at http://nlp.rusvectores.org/topical/ .

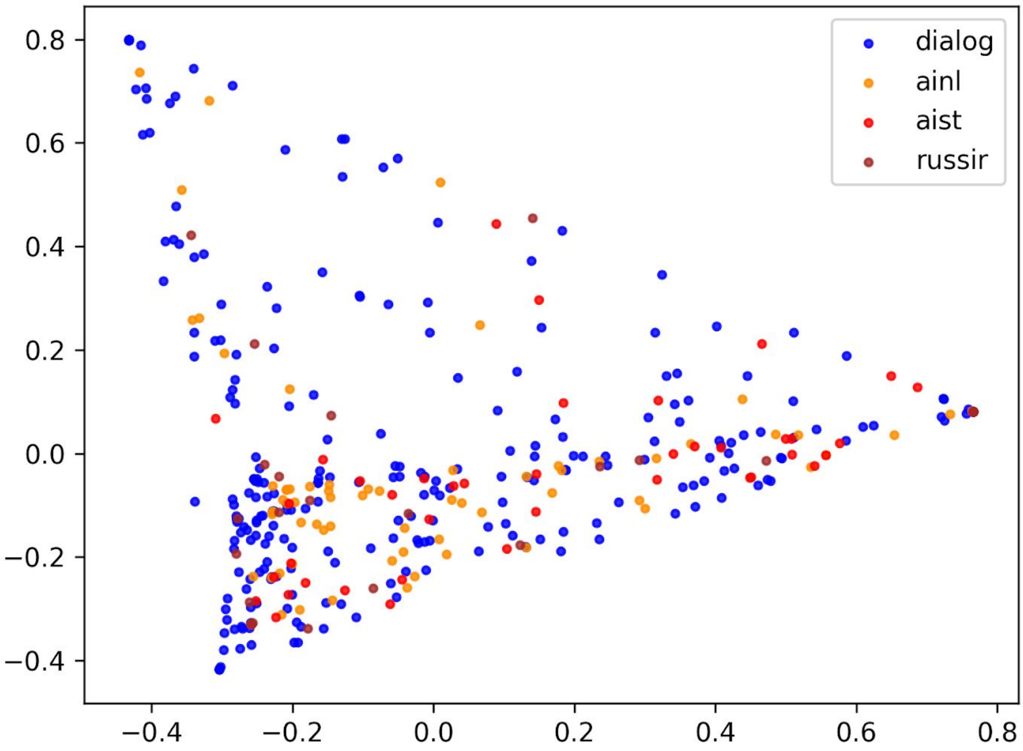

Figure 3. Topical similarity for papers from different Russian NLP venues

Further on, we trained separate LDA models for the papers belonging to each conference and to each year. The tables 2 and 3 report topical structure for these subcorpora.

Dialogue	1: subject, verb, case, language, argument, distance, structure 2: speech, language, case, word, Russian, text, user 3: translation, Russian, text, language, model, form, English 4: word, model, result, feature, Russian, dictionary, corpus 5: text, language, Russian, Czech, English, relation, pronoun 6: semantic, language, word, text, example, concept, structure 7: sentence, feature, word, text, system, result, Russian 8: collocation, word, semantic, list, Russian, datum, result 9: sentiment, review, result, task, classification, analysis, opinion 10: aspect, term, clitic, element, fifth, language, Celsius
AIST	1: topic, model, word, topics, feature, document, detection 2: word, analysis, morphological, language, error, Russian, dictionary 3: word, sentence, model, result, text, noun, context 4: word, sentiment, feature, user, model, based, document 5: word, language, name, question, type, exercise, entity 6: model, word, name, pair, window, female, gender 7: term, word, text, dataset, tree, pattern, method 8: word, problem, datum, relation, information, model, result 9: association, model, gender, specialization, experiment, reaction,

	associative 10: topic, matrix, algorithm, batch, model, thread, BigARTM
AINL	1: word, result, task, datum, model, paraphrase, feature 2: system, datum, time, service, node, user, information 3: feature, word, semantic, sentence, model, paraphrase, result 4: model, speech, language, word, recognition, network, layer 5: feature, algorithm, dataset, metum, number, problem, cluster 6: user, expression, Twitter, relation, discussion, influencer, motion 7: operation, interaction, application, subscription, Smart, combination, space 8: similarity, Jaccard, equality, difference, absolute, subject, Description 9: text, reuse, translation, thes, feature, number, system 10: word, language, result, model, Russian, text, English

Table 2. 10 most probable topics characterized by 7 words for each conference venue

2011	1: grammar, language, case, Slavic, Croatian, Serbian, linguistic 2: standard, adjective, degree, functional, construction, purpose, English 3: word, language, rule, text, structure, result, object 4: rule, word, language, adverb, state, letter, method 5: clause, emotional, cluster, expression, character, word, synonym 6: word, text, algorithm, result, collection, chat, number 7: form, fifth, dictionary, nominalization, label, word, grammatical 8: language, word, text, structure, clitic, sentence, method 9: classi, review, word, class, feature, cation, noun 10: word, language, object, information, similarity, text, name
2012	1: emotional, language, text, corpus, information, form, semantic 2: verb, sound, syntactic, semantic, word, object, lexical 3: emphatic, remark, datum, dialogue, example, topic, emphasis 4: word, Russian, sentence, system, case, domain, result 5: cluster, word, element, translation, clitic, Russian, relation 6: speech, term, word, sentence, feature, synthesis, text 7: sentiment, review, task, class, result, camera, book 8: word, feature, region, model, classification, information, website 9: question, sentence, word, information, verb, Russian, sentiment 10: system, Russian, text, language, information, term, gram
2013	1: verb, subject, semantic, human, impersonal, Russian, argument 2: word, phrase, result, sentence, break, system, text 3: word, speech, extraction, system, based, collocation, relation 4: text, coreference, relation, sentence, idiom, phrase, generalization 5: speech, sentence, word, Celsius, method, review, feature 6: word, feature, semantic, language, translation, opinion, context 7: Celsius, sentiment, word, system, name, case, rule 8: language, sentiment, feature, classification, task, object, result 9: collocation, word, pair, corpus, distance, frequency, association 10: discourse, taxonomy, type, parameter, style, functional, mode
2014	1: category, question, user, subcategory, classification, team, example 2: accuracy, prediction, datum, price, stock, market, Twitter

	3: semantic, feature, text, Russian, type, parser, role 4: subject, name, problem, Russian, verb, sentence, model 5: word, system, result, relation, text, corpus, resolution 6: semantic, role, rule, object, system, information, tree 7: type, semantic, entity, reference, expression, example, language 8: concept, term, detection, domain, ontology, frame, algorithm 9: model, term, word, document, query, translation, topic 10: word, text, language, feature, result, sentence, standard
2015	1: verb, role, semantic, construction, class, pattern, FrameBank 2: word, term, model, feature, result, task, aspect 3: text, language, fifth, feature, result, disease, network 4: system, semantic, datum, case, patient, text, information 5: word, text, relation, language, sense, news, model 6: feature, node, tree, algorithm, text, system, number 7: word, language, text, Russian, dictionary, corpus, number 8: form, case, model, preposition, number, action, argument 9: word, syntactic, example, datum, system, relation, algorithm 10: word, language, text, analysis, pymorphy, Russian, system
2016	1: entity, name, fact, text, Russian, system, word 2: language, text, role, number, corpus, Russian, semantic 3: time, algorithm, path, topic, matrix, sery, batch 4: relation, dataset, text, Russian, possessive, expression, Czech 5: word, user, model, result, based, term, vector 6: word, model, language, feature, result, system, case 7: similarity, Jaccard, equality, difference, absolute, subject, calculate 8: word, datum, dictionary, algorithm, result, number, text 9: word, model, language, text, number, result, Russian 10: word, sentiment, lexicon, Russian, sense, tagger, dictionary
2017	1: feature, word, result, model, vector, post, network 2: paraphrase, feature, task, result, sentence, word, pair 3: word, model, task, embedding, result, sentence, vector 4: feature, word, class, bridge, language, datum, distance 5: feature, network, model, result, mention, layer, coreference 6: question, answer, model, feature, predicate, word, task 7: text, relation, discourse, feature, translation, Russian, news 8: word, model, Russian, result, language, number, table 9: text, sentence, essay, source, word, plagiarism, document 10: search, word, question, topic, classification, result, document

Table 3. 10 most probable topics characterized by 7 words for each year

4. Results and discussion

Table 2 shows comparison between the 3 conferences under analysis. It is clear that despite all the topics being related to the domain of computation and language, each conference is different. The *Dialogue* topics contain linguistically-oriented words that do not appear in other conferences keyphrases lists: for example, *relation*, *pronoun*, *subject*, *case*. On the other hand, *AIST* and *AINL* feature many terms from computer science, like *algorithm*, *batch*, *dataset*. Certainly, *AIST* and *AINL* topics do contain some linguistic terms (e.g. *morphological*, *sentiment*), and some computer science terms can be found in the *Dialogue* topics (*classification*, *analysis*), but their amount is negligible in each case.

Thus, *AINL* and *AIST* are more computationally oriented venues, while the *Dialogue* still retains its linguistic orientation. This is not surprising, since the *Dialogue* historically had a strong background in humanities, while *AINL* and *AIST* were established mostly by the researchers with computer science background.

Less obvious results could be observed in Table 3 which reports most probable topics for each of the analyzed years. It is reasonable to hypothesize that starting from approximately 2014, there should be an increasing amount of topics related to neural networks and word embeddings[11]. However, in fact the topics including words like *network*, *feature* and *layer* are becoming more or less visible only in 2017. The topics from 2011-2016 are not that computer science oriented. Of course, they do contain certain words related to machine learning methods (e.g. *feature, classification*), but their amount is considerably lower than the amount of general linguistic terms.

That lack of machine learning related topics can stem from the fact that the number of the *Dialogue* papers is several times higher than the number of those from *AIST* and *AINL*. Thus, the topics are heavily influenced by the linguistic orientation of this conference. In other words, most of the topics in each year reflect central trends in this year's *Dialogue*. For example, in 2012, the *Dialogue* for the first time featured a track for sentiment analysis, and the global topics for this years contain terms like *emotional* and *sentiment.* In 2015, the *Dialogue* was focused on computational models of semantics: and in the global topics we observe words like *semantics*, *sense*, *word, model,* etc. In 2016, the *Dialogue* featured a shared task on named entity recognition, causing the terms like *entity* and *fact* to appear in the leading topics.

Then, in 2017, the Dialogue announced deep machine learning techniques to be one of its central topics, and the global 2017 topical structure rapidly shifted. Since in 2018 the *Dialogue* continues with the strong focus on machine learning, one can expect further increase of ML-related topics this year.

5. Related work

The task of topical drift tracking was originally proposed in [Allan et al., 1998] (formulated as '*Topic Detection and Tracking'*), and extensively described in [Allan, 2002]. Since then, the task was investigated by many researchers proposing different approaches. They can be roughly divided into *static topic models* (including ubiquitous techniques like the *LDA* we used or less common ones like *Replicated Softmax* [Hinton and Salakhutdinov, 2009]) and *dynamic topic models* which are static models generalized to dynamic setting (like *RNN-RSM)* [Gupta et al., 2017].

Investigations of topical drifts in scientific literature were already proposed for certain English conferences and resources, for example, *NIPS*[12] [Wang and McCallum, 2006] and

[11] Word embeddings frenzy started after the release of *word2vec* [Mikolov et al, 2013] in 2013.

[12] https://nips.cc/

CiteSeer[13] [He et al., 2009]. The most relevant work to our study is [Gollapalli and Li, 2015], who analyzed the topics in two top-tier NLP conferences: *ACL* and *EMNLP*. The authors compared research topics in these venues, and concluded that the topics currently addressed in NLP texts are significantly different from those addressed a decade back, and that the topics within the ACL conference are very close to each other, unlike EMNLP.

Talking about other types of data mining applications to NLP scholarly paper archives, there is a reference dataset for bibliographic research based on the *ACL Anthology* [Bird et al, 2008]. One can also mention *bib2vec*: a tool for processing bibliographic data [Yoneda et al., 2017]. However, we are not aware of any similar work dealing with Russian NLP papers.

Russian NLP community was studied by [Khoroshevsky, 2012] (and some earlier papers by the same author), but the analysis in these series of papers was focused more on research clusters and citation networks than on historical topic development. Also, our dataset, unlike the one used in [Khoroshevsky, 2012], is freely available and contains conference proceedings up to 2018. Another related work is [Sokolova and Kononenko 2012], which presented a bilingual computational linguistics thesaurus. However, no analysis of actual Russian NLP papers was made in this work, and thus their results are not comparable to ours.

6. Conclusions and future work

In this paper, we briefly analyzed the topical structure and drift in the 10 years of Russian computational linguistics. To this end, we compiled a corpus of 497 papers written in English from the proceedings of 3 major NLP conferences held in Russia. The topics extracted from this corpus were compared for different venues and for different years. The obtained results provide insights on the development of Russian NLP. The primary findings are:

1. the *Dialogue* is more a linguistically-oriented event, while *AIST* and *AINL* are more computationally-oriented,
2. topics of Russian NLP in general each year usually reflect the topics that the organizers of the *Dialogue* picked as central for this year's conference
3. there is a clear shift in the topical structure towards machine learning domain.

To the best of our knowledge, we are the first to present the study of evolution and distribution of topics in the field of Russian computational linguistics. However, the analysis of topical structure presented in this work is only a part of a bigger *RusNLP* project aimed at describing the Russian NLP community. Our nearest plans include implementing a Web service allowing to search for semantically related papers; to this end, we plan to compare the performance of topic modelling approaches (like *LDA*) and prediction-based document embedding models (like *Paragraph Vector* from [Le and Mikolov, 2014]).

[13] http://citeseerx.ist.psu.edu/

As a next step, we are going to analyze the citation network in the Russian NLP community, based on the metadata extracted from our dataset. This will allow to trace more granular research communities and informal scientific schools.

Finally, we plan to investigate cross-linguistic topic modeling methods in order to include in our analysis the papers written in Russian as well. With cross-linguistic NLP techniques, we will also be able to compare Russian NLP conferences against other similar communities in other parts of the world: for example, Arabic computational linguistics[14] or Italian computational linguistics[15]. For sure, we will also continue to maintain the *Russian NLP Dataset*, updating it with new proceedings from major Russian NLP conferences.

References

1. Allan, J., Carbonell, J. G., Doddington, G., Yamron, J., & Yang, Y. (1998), Topic detection and tracking pilot study final report.
2. Allan, J. (2002), Introduction to topic detection and tracking. In Topic detection and tracking, Springer, Boston, MA, pp. 1-16.
3. Bird, S., Dale, R., Dorr, B. J., Gibson, B., Joseph, M. T., Kan, M. Y., ... & Tan, Y. F. (2008), The ACL anthology reference corpus: A reference dataset for bibliographic research in computational linguistics.
4. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003), Latent dirichlet allocation. Journal of machine Learning research, Vol. 3, pp. 993-1022.
5. Gupta, P., Rajaram, S., Schütze, H., & Andrassy, B. (2017), Deep Temporal-Recurrent-Replicated-Softmax for Topical Trends over Time. arXiv preprint arXiv:1711.05626.
6. Gollapalli, S. D., & Li, X. (2015), EMNLP versus ACL: Analyzing NLP research over time. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 2002-2006.
7. He, Q., Chen, B., Pei, J., Qiu, B., Mitra, P., & Giles, L. (2009), Detecting topic evolution in scientific literature: how can citations help?. In Proceedings of the 18th ACM conference on Information and knowledge management, pp. 957-966.
8. Hinton, G. E., & Salakhutdinov, R. R. (2009). Replicated softmax: an undirected topic model. In Advances in neural information processing systems (pp. 1607-1614).
9. Khoroshevsky, V. (2012), Пространства знаний в сети Интернет и Semantic Web, Часть 3 (Knowledge spaces in the Internet and Semantic Web, part 3); Искусственный интеллект и принятие решений (Artificial Intelligence and Decision Making) (1), 3–38
10. Le, Q., & Mikolov, T. (2014), Distributed representations of sentences and documents. In International Conference on Machine Learning, pp. 1188-1196.
11. Lui, M., & Baldwin, T. (2012), langid. py: An off-the-shelf language identification tool. In Proceedings of the ACL 2012 system demonstrations, Association for Computational Linguistics, pp. 25-30.
12. Lyashevskaya, O., Kopotev, M., & Mustajoki, A. (2018). Russian challenges for quantitative research. In Quantitative approaches to the Russian language, Routledge, Taylor & Francis Group, pp. 3-29.
13. Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., & McClosky, D. (2014), The Stanford CoreNLP natural language processing toolkit. In Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations, pp. 55-60.
14. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013), Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems, pp. 3111-3119.

[14] https://www.acling.org
[15] http://www.ai-lc.it/en

15. Rehurek, R., & Sojka, P. (2010), Software framework for topic modelling with large corpora. In In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks.
16. Sokolova, E. G., & Kononenko, I. S. (2012), Russian-English thesaurus on computational linguistics. In *Компьютерная лингвистика и интеллектуальные технологии* (pp. 598-606).
17. Straka, M., Hajic, J., & Straková, J. (2016), UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. In LREC.
18. Wang, X., & McCallum, A. (2006), Topics over time: a non-Markov continuous-time model of topical trends. In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 424-433.
19. Yoneda, T., Mori, K., Miwa, M., & Sasaki, Y. (2017), Bib2vec: An Embedding-based Search System for Bibliographic Information. arXiv preprint arXiv:1706.05122.