Russian Computational Linguistics: Topical Structure in 2007-2017 Conference Papers

Amir Bakarov Andrey Kutuzov Irina Nikishina

Dialogue
Conference on Computational Linguistics and Intelligent Technologies

02.06.2018

Aim and Contribution

Part of a larger project to analyze Russian NLP publishing activity (RusNLP).

- Analysis of Russian NLP publications topical structure and topical drift (for papers written in English):
  - between venues;
  - diachronically (2007-2017).
- A large dataset of papers published at Russian NLP venues
  - considerable amount of metadata, available in a machine-readable format

Setup

- 3 primary Russian NLP venues;
  1. AINL;
  2. AIST;
  3. Dialogue.
- More than 1700 papers were manually assessed to parse metadata;
- LDA with varying hyperparameters was trained on this dataset to extract the topics.

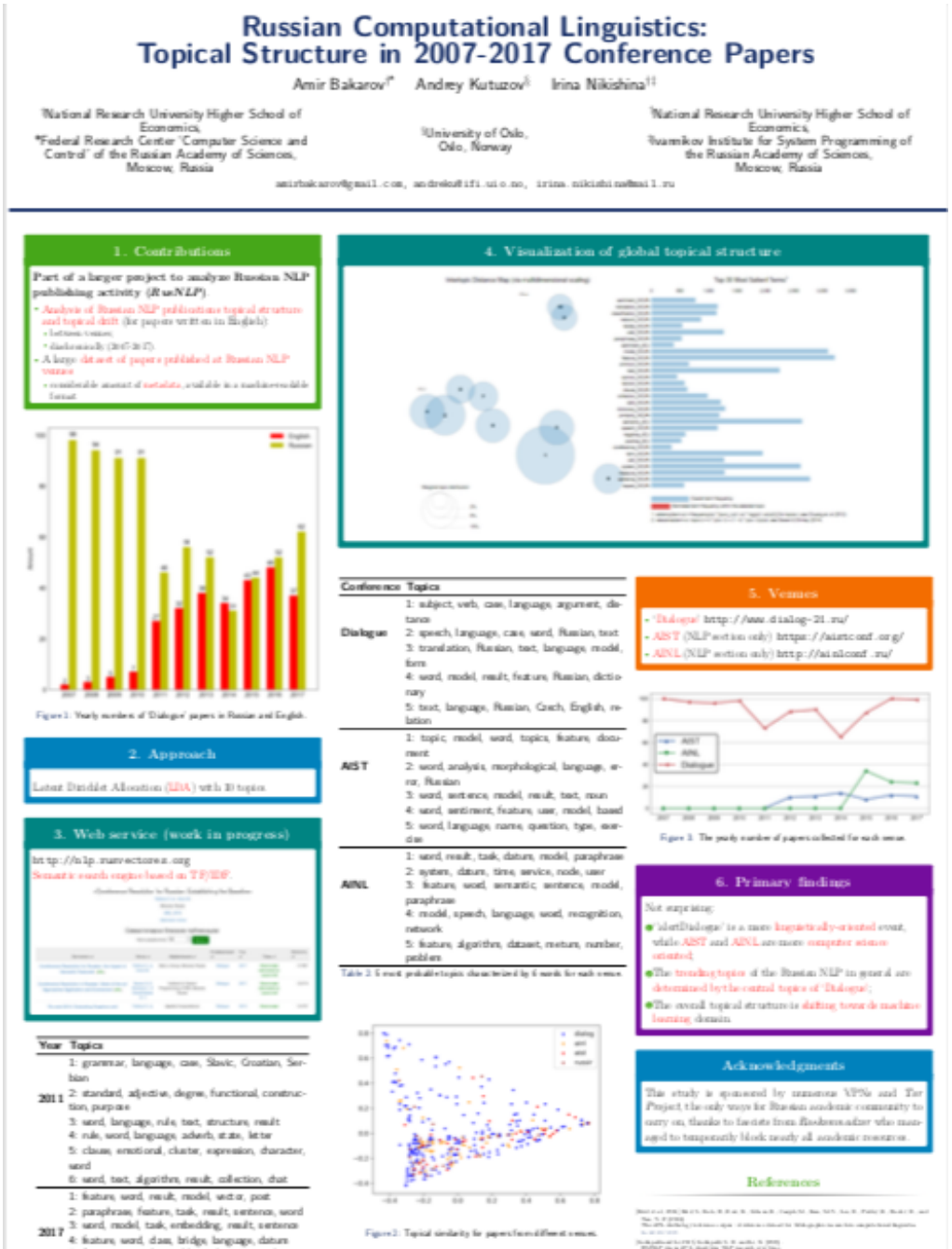
Web service is available at http://nlp.rusvectores.org (work in progress!).

Primary Findings

Not surprising:

1. Dialogue is a more linguistically-oriented event, while AIST and AINL are more computer science oriented;
2. The trending topics of the Russian NLP in general are determined by the central topics of 'Dialogue';
3. The overall topical structure is shifting towards machine learning domain.

Come to Our Poster to Learn More!



Russian Computational Linguistics: Topical Structure in 2007-2017 Conference Papers

Amir Bakarov[1,2] Andrey Kutuzov[3] Irina Nikishina[1,4]

[1]National Research University Higher School of Economics, [2]Federal Research Center 'Computer Science and Control' of the Russian Academy of Sciences, Moscow, Russia

[3]University of Oslo, Oslo, Norway

[4]National Research University Higher School of Economics, [4]Ivannikov Institute for System Programming of the Russian Academy of Sciences, Moscow, Russia

amirbakarov@gmail.com, andreku@ifi.uio.no, irina.nikishina@mail.ru

1. Contributions

Part of a larger project to analyze Russian NLP publishing activity (*RuNLP*)

- Analysis of Russian NLP publications' topical structure and topical drift (for papers written in English)
  - between venues;
  - diachronically (2007-2017).
- A large dataset of papers published at Russian NLP venues
  - considerable amount of metadata, available in a machine-readable format.

Figure 1: Yearly numbers of 'Dialogue' papers in Russian and English.

2. Approach

Latent Dirichlet Allocation (LDA) with 10 topics

3. Web service (work in progress)

http://nlp.rusvectores.org

Semantic search engine based on TF/IDF.

Year	Topics
2011	1: grammar, language, case, Slavic, Croatian, Serbian 2: standard, adjective, degree, functional, construction, purpose 3: word, language, rule, text, structure, result 4: rule, word, language, adverb, state, letter 5: clause, emotional, cluster, expression, character, word 6: word, text, algorithm, result, collection, chat
2017	1: feature, word, result, model, vector, pos 2: paraphrase, feature, task, result, sentence, word 3: word, model, task, embedding, result, sentence 4: feature, word, class, bridge, language, dataset

4. Visualization of global topical structure

Conference	Topics
Dialogue	1: subject, verb, case, language, argument, distance 2: speech, language, case, word, Russian, text 3: translation, Russian, text, language, model, form 4: word, model, result, feature, Russian, dictionary 5: text, language, Russian, Czech, English, relation
AIST	1: topic, model, word, topics, feature, document 2: word, analysis, morphological, language, error, Russian 3: word, sentence, model, result, text, noun 4: word, sentiment, feature, user, model, based 5: word, language, name, question, type, exercise
AINL	1: word, result, task, dataset, model, paraphrase 2: system, dataset, time, service, node, user 3: feature, word, semantic, sentence, model, paraphrase 4: model, speech, language, word, recognition, network 5: feature, algorithm, dataset, measure, number, problem

Table 2: 5 most probable topics characterized by 6 words for each venue.

Figure 2: Topical similarity for papers from different venues.

5. Venues

- 'Dialogue' http://www.dialog-21.ru/
- AIST (NLP section only) https://aistconf.org/
- AINL (NLP section only) http://ainlconf.ru/

Figure 3: The yearly number of papers collected for each venue.

6. Primary findings

Not surprising:

- 'dialogue' is a more linguistically-oriented event, while AIST and AINL are more computer science oriented;
- The trending topics of the Russian NLP in general are determined by the central topics of 'Dialogue';
- The overall topical structure is shifting towards machine learning domain.

Acknowledgments

This study is sponsored by numerous VPNs and Tor Project, the only ways for Russian academic community to carry on, thanks to fascists from Roskomnadzor who managed to temporarily block nearly all academic resources.

References

[1] Bakarov, A. (2018). A Survey of Word Embeddings Evaluation Methods. CoRR, abs/1801.09536.

[2] Bakarov, A., Kutuzov, A., and Nikishina, I. (2018). Predicting Topical Structure: Russian Natural Language Processing 2007-2017. In Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference 'Dialogue 2018', pages 79–91, Moscow, Russia. RSUH.

Thank you for your attention!

Amir Bakarov, Andrey Kutuzov, Irina Nikishina
amirbakarov@gmail.com, andreku@ifi.uio.no, irina.nikishina@mail.ru
http://bakarov.github.io