

RusNLP: Semantic search engine for Russian NLP conference papers

Irina Nikishina^{†‡}, Amir Bakarov^{*†}, Andrey Kutuzov[§]

[†] National Research University Higher School of Economics, Moscow, Russia

[‡] Ivannikov Institute for System Programming of the Russian Academy of Sciences, Moscow, Russia

* Federal Research Center 'Computer Science and Control' of the Russian Academy of Sciences, Moscow, Russia

[§] University of Oslo, Oslo, Norway

AIST
5 July 2018

Contents

- 1 What is RusNLP?
- 2 Typical use cases
- 3 What is novel in that?
- 4 Key features of RusNLP
- 5 Conclusion

What is RusNLP ?

- **RusNLP** is a web service :
- search engine and recommendation system...
- ...over proceedings of **three major Russian NLP conferences** :
 - 1 **Dialogue**
 - 2 **AIST**
 - 3 **AINL**

What is RusNLP ?

- **RusNLP** is a web service :
- search engine and recommendation system...
- ...over proceedings of **three major Russian NLP conferences** :
 - 1 **Dialogue**
 - 2 **AIST**
 - 3 **AINL**
- 18 years of publishing ;

What is RusNLP ?

- **RusNLP** is a web service :
- search engine and recommendation system...
- ...over proceedings of **three major Russian NLP conferences** :
 - 1 **Dialogue**
 - 2 **AIST**
 - 3 **AINL**
- 18 years of publishing ;
- 400 academic papers in English ;

What is RusNLP ?

- **RusNLP** is a web service :
- search engine and recommendation system...
- ...over proceedings of **three major Russian NLP conferences** :
 - 1 **Dialogue**
 - 2 **AIST**
 - 3 **AINL**
- 18 years of publishing ;
- 400 academic papers in English ;
- metadata (titles, abstracts authors, affiliations) extracted automatically and **normalized manually** ;
- `http://nlp.rusvectors.org`
- Search for yourself or your institution :-)
- NB : we do not provide full texts of the papers, but there are hyperlinks to source URLs.

What is RusNLP ?



2002-2017

<http://dialog-21.ru>

281 papers in English

Source : conference website



AIST

2014-2017

<https://aistconf.org>

45 papers in English

Source : Springer

ARTIFICIAL INTELLIGENCE
& NATURAL LANGUAGE



2015-2017

<http://ainlconf.ru>

67 papers in English

Source : Springer

The dataset is described in details in [Bakarov et al., 2018].

Contents

- 1 What is RusNLP ?
- 2 Typical use cases**
- 3 What is novel in that ?
- 4 Key features of RusNLP
- 5 Conclusion

Generally, *RusNLP* can be used to :

- **Discover academic knowledge** you were not aware of :
 - may be someone has already done what you are doing?

Generally, *RusNLP* can be used to :

- Discover academic knowledge you were not aware of :
 - maybe someone has already done what you are doing?
- Identify 'gaps' in Russian NLP, where we still lack knowledge ;

Generally, *RusNLP* can be used to :

- Discover **academic knowledge** you were not aware of :
 - may be someone has already done what you are doing?
- Identify **'gaps'** in Russian NLP, where we still lack knowledge ;
- Analyze **academic communities and publishing patterns** :
 - who published what and where ?
 - sort of **'who is who'** for the Russian NLP community.

Questions we can answer

- *'I know this paper, what other similar papers are there in Russian NLP?'*

Questions we can answer

- *'I know this paper, what other similar papers are there in Russian NLP?'*
- *'What was published in 2008 by NLP scholars from Moscow State University?'*

Questions we can answer

- *'I know this paper, what other similar papers are there in Russian NLP ?'*
- *'What was published in 2008 by NLP scholars from Moscow State University ?'*
- *'Were there any papers about paraphrases detection at the **AINL** conference in 2015 ?'*
- etc...

Contents

- 1 What is RusNLP ?
- 2 Typical use cases
- 3 What is novel in that?**
- 4 Key features of RusNLP
- 5 Conclusion

What is novel in that ?

There are many scholar search engines available :

What is novel in that ?

There are many scholar search engines available :

Google Scholar

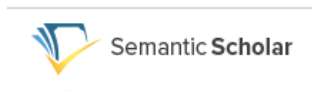
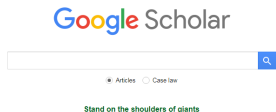
 

Articles Case law

Stand on the shoulders of giants

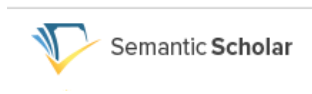
What is novel in that ?

There are many scholar search engines available :



What is novel in that ?

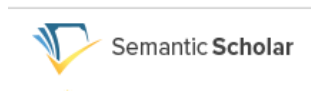
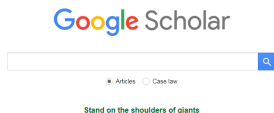
There are many scholar search engines available :



...etc

What is novel in that ?

There are many scholar search engines available :



...etc

What's different about us ?

Feature	Google Scholar scholar.google.com	ArXiv Sanity arxiv-sanity.com	RusNLP nlp.rusvectors.org
International coverage	Green	Green	Red
Recommendation system	Green	Green	Green
Manually parsed metadata	Red	Red	Green
Focus on NLP	Red	Red	Green

What is novel in that ?

Vladimir Khoroshevsky started the analysis of Russian NLP landscape [Khoroshevsky, 2012]. Compared to them, the *RusNLP* project :

- 1 uses much more up-to-date data and includes novel publishing venues (AIST and AINL).

What is novel in that ?

Vladimir Khoroshevsky started the analysis of Russian NLP landscape

[Khoroshevsky, 2012]. Compared to them, the *RusNLP* project :

- 1 uses much more up-to-date data and includes novel publishing venues (AIST and AINL).
- 2 Implements the semantic search engine available online as a web service.

What is novel in that ?

Vladimir Khoroshevsky started the analysis of Russian NLP landscape

[Khoroshevsky, 2012]. Compared to them, the *RusNLP* project :

- 1 uses much more up-to-date data and includes novel publishing venues (AIST and AINL).
- 2 Implements the semantic search engine available online as a web service.
- 3 Publishes the datasets and the source code.

Experiments with similar documents search

Popular approaches for document representations :

- 1 **TF-IDF (term frequency – inverted document frequency)** : a term weighting scheme from information retrieval [Sparck Jones, 1972].
- 2 **LDA (Latent Dirichlet Allocation)** : a widespread distributional topic modeling technique [Blei et al., 2003].
- 3 **Paragraph Vector (also known as *doc2vec*)** : a distributional parametric algorithm based on shallow feed-forward neural networks, extension of *word2vec* [Le and Mikolov, 2014].

Experiments with similar documents search

Popular approaches for document representations :

- 1 **TF-IDF (term frequency – inverted document frequency)** : a term weighting scheme from information retrieval [Sparck Jones, 1972].
- 2 **LDA (Latent Dirichlet Allocation)** : a widespread distributional topic modeling technique [Blei et al., 2003].
- 3 **Paragraph Vector (also known as *doc2vec*)** : a distributional parametric algorithm based on shallow feed-forward neural networks, extension of *word2vec* [Le and Mikolov, 2014].

We tested them all.

TF-IDF wins

Performance of the tested models (3 independent assessors)		
Model	Average precision	Inter-rater agreement
<i>TF-IDF</i>	0.64	0.73
<i>Paragraph Vector, 40 dims</i>	0.33	0.66
<i>Paragraph Vector, 100 dims</i>	0.39	0.50
<i>LDA, 10 topics</i>	0.20	0.66
<i>LDA, 20 topics</i>	0.27	0.52

Precision here is simply the ratio of the results which are at least somewhat relevant in the 10 nearest neighbors of a document. No **recall** evaluation was performed.

Contents

- 1 What is RusNLP ?
- 2 Typical use cases
- 3 What is novel in that ?
- 4 Key features of RusNLP**
- 5 Conclusion

Key features of RusNLP

RusNLP

Topical structure

About

RU/EN

Space-separated keywords:

▼ Show advanced filters >>>

Conferences:

Dialogue
AIST
AINL

Years (since-till):

Paper author:

Affiliation:

Paper title:

Search

Generic search with filters

Key features of RusNLP

RusNLP [Тематическая структура](#) [О проекте](#) [ENRU](#)

Ключевые слова через пробел:

► Показать дополнительные фильтры >>>

Искать

Вот что мы нашли для вас

Кликните по заголовку статьи для поиска похожих публикаций

Заголовок - +	Автор - +	Аффилиация - +	Конференция - +	Год - +	Задача - +
Evaluation Tracks on Plagiarism Detection Algorithms for the Russian Language (URL)	Kutuzov A. B. Ivanova L. Ljshchekaja O. N. Smirnov I. V. Khazov A. V. Kuznetsova Rita Kopotev M. V.	Institute for Systems Analysis, FRC CSC RAS, Moscow, Russia; RUDN University, Moscow, Russia НИУ ВШЭ, Москва, Россия Хантымансийский университет; Февальдия Antiplagiat JSC, Moscow, Russia Mail.ru Group, Moscow, Russia	Dialogue	2017	Обнаружение дубликатов
Size Vs. Structure in Training Corpora for Word Embedding Models: Araneum Russicum Maximum and Russian National Corpus (URL)	Kutuzov A. B. Kunilovskaja M. A.	University of Tyumen, Tyumen, Russia University of Oslo, Oslo, Norway	AIST	2017	-
The Impact of Morphology Processing Quality on Automated Anaphora Resolution for Russian (URL)	Ionov M. Kutuzov A. B.	Mail.ru Group, Moscow, Russia	Dialogue	2014	Машинный перевод Обработка реп
Корпус Неосвоенных Переводов как Инструмент для Переводоведческих Исследований (URL)	Oshchepkov A. Ju. Kutuzov A. B. Cherpuikova A. Ju. Kunilovskaja M. A.	University of Tyumen, Tyumen, Russia	Dialogue	2012	Машинный перевод Корпусы

Interlinked authors

RusNLP web service : key features

Ключевые слова через пробел:

► Показать дополнительные фильтры >>>

Искать

Вот что мы нашли для вас

Кликните по заголовку статьи для поиска похожих публикаций

Заголовок + *	Автор + *	Аффилиация + *	Конференция + *	Год + *	Задачи + *
Improving Distributional Semantic Models Using Anaphora Resolution During Linguistic Preprocessing (URL)	Koslova O.	НИУ ВШЭ, Москва, Россия	Dialogue	2010	Морфологический анализ Дистрибутивная семантика Токенизация
Building Dependency Parsing Model for Russian with Maltparser and Mystem Tagset (URL)	Droganova K. A.	НИУ ВШЭ, Москва, Россия	AINL	2015	Синтаксический анализ
Detection of Domain-specific Trends in Text Collections (URL)	Gadelshin Ilnur Antonova Anna Ilvovsky D. A.	НИУ ВШЭ, Москва, Россия	AIST	2014	Извлечение ключевых слов
Совместная Встречаемость Слов: Опыт Классификации (URL)		НИУ ВШЭ, Москва, Россия	Dialogue	2013	Тематическая классификация Извлечение отношений Корпусы
Comparison of Neural Network Architectures for Sentiment Analysis of Russian Tweets (URL)	Skorniakov K. Turdakov D. Gomzin A. Trofimovich J. Ankhipenko K. Kozlov I.	Moscow Institute of Physics and Technology НИУ ВШЭ, Москва, Россия Institute for System Programming of RAS, Moscow, Russia MSU, Moscow, Russia	Dialogue	2016	Анализ тональности

Interlinked affiliations

Key features of RusNLP

RusNLP	Тематическая структура	О проекте	EN/RU					
Entity Based Sentiment Analysis Using Syntax Patterns and Convolutional Neural Network (URL)			Kazorin V. I. Nemov N. R. Kozhevnikov M. V. Karpov I. A.		Dialogue	2016	Анализ тональности	0.031
На Входе Тексты, на Выходе Смысл: Нейронные Языковые Модели для Задач Семантической Близости (На Материале Русского Языка) (URL)			Kutuzov A. B. Andreev I.	Mail.ru Group, Moscow, Russia	Dialogue	2015	Дистрибутивная семантика	0.015
Identifying Disease-related Expressions in Reviews Using Conditional Random Fields (URL)			Miftahutdinov Z. Sh. Tropsha A. E. Tutubalina E. V.		Dialogue	2017	Извлечение именованных сущностей	0.018
Combined Feature Representation for Emotion Classification from Russian Speech (URL)			Verkholyak Oxana Karpov Alexey	ITMO university	AINL	2017	Извлечение именованных сущностей Обработка речи Тематическая классификация Извлечение отношений	0.054

Interlinking with <http://NLPub.ru> (typical NLP tasks)

Key features of RusNLP

«Morphological Analysis for Russian: Integration and Comparison of Taggers»

Kuzmenko E. A.; НИУ ВШЭ, Москва, Россия;

AIST, 2016;

Similar publications:

Number of results:

Paper title ▲ ▼	Paper author ▲ ▼	Affiliation ▲ ▼	Conference ▲ ▼	Year ▲ ▼	Tasks ▲ ▼	Similarity ▲ ▼
A Close Look at Russian Morphological Parsers: which One is the Best? (URL)	Fishcheva Irina Koteinikov E. V. Razova E. V.		AINL	2017	Морфологический анализ	0.3375
The Beginning of a Beautiful Friendship: Rule-based and Statistical Analysis of Middle Russian (URL)	Gavrilova T. Berdichevskij A.		Dialogue	2016	Морфологический анализ	0.2115
Morphological Analyzer and Generator for Russian and Ukrainian Languages (URL)	Korobov Mikhail		AIST	2015	Морфологический анализ	0.2108
Автоматическое Извлечение Правил для Снятия Морфологической Неоднозначности (URL)	Bocharov V. V. Protoporova	СПбГУ, Санкт-Петербург, Россия	Dialogue	2013	Корпусы	0.2051

Nearest neighbors search

Contents

- 1 What is RusNLP ?
- 2 Typical use cases
- 3 What is novel in that ?
- 4 Key features of RusNLP
- 5 Conclusion**

Conclusion

Done

- **Publicly available search engine** over publications from Russian computational linguistics venues ;

Conclusion

Done

- **Publicly available search engine** over publications from Russian computational linguistics venues ;
- search by keywords, authors, affiliations, years, conferences...

Conclusion

Done

- Publicly available search engine over publications from Russian computational linguistics venues ;
- search by keywords, authors, affiliations, years, conferences...
- manually curated metadata dataset available for downloading (*SQLite* database) ;

Conclusion

Done

- Publicly available search engine over publications from Russian computational linguistics venues ;
- search by keywords, authors, affiliations, years, conferences...
- manually curated metadata dataset available for downloading (*SQLite* database) ;
- (simplistic) evaluation of algorithms for similar documents search.

Conclusion

In the future

- we will **maintain and yearly update the database** ;
- analysis of **citation network** (coming soon !);
- search for papers in Russian (from the **Dialogue** proceedings);
- **temporal topic models** ;
- use **multiword expressions** (word n-grams) ;
- expand user queries with pre-trained word embedding models.

Conclusion

In the future

- we will **maintain and yearly update the database** ;
- analysis of **citation network** (coming soon !);
- search for papers in Russian (from the **Dialogue** proceedings);
- **temporal topic models** ;
- use **multiword expressions** (word n-grams);
- expand user queries with pre-trained word embedding models.

Thank you for your attention !

We welcome any questions, comments or suggestions
you may have :)

<http://nlp.rusvectors.org>

References I

-  Bakarov, A., Kutuzov, A., and Nikishina, I. (2018).
Russian computational linguistics : topical structure in 2007-2017 conference papers.
In Proceedings of Dialogue-2018, online papers. ABBYY.
-  Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003).
Latent Dirichlet allocation.
Journal of Machine Learning Research, 3(Jan) :993–1022.
-  Khoroshevsky, V. (2012).
Knowledge spaces in the Internet and Semantic Web, part 3 ; in Russian.
Artificial Intelligence and Decision Making, pages 3–38.
-  Le, Q. and Mikolov, T. (2014).
Distributed representations of sentences and documents.
In International Conference on Machine Learning, pages 1188–1196.
-  Sparck Jones, K. (1972).
A statistical interpretation of term specificity and its application in retrieval.
Journal of documentation, 28(1) :11–21.