RUSNLP MAP

Кураторы:
Амир Бакаров,
Андрей Кутузов,
Ирина Никишина

Участники:
Анна Сафарян,
Пётр Фильченков,
Вэйцзя Янь

О чём этот проект

RusNLP_MAP - это поисковик по статьям, опубликованным на российских NLP-конференциях: Диалог, AIST, AINL

Вопросы, на которые может ответить **RusNLP**:

- *Какие есть публикации в российском NLP, похожие на уже известную мне статью?*
- *Что публиковали в 2008 году компьютерные лингвисты из МГУ?*
- *Представляли ли в 2015 году на конференции AINL какие-либо исследования, связанные с перифразированием?*

Похожие проекты

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

Что уже есть

Сайт https://nlp.rusvectores.org/ru/

2 публикации по проекту на Диалоге и Аисте:

- *Bakarov A*, *Kutuzov A.*, *Nikishina I.* Russian computational linguistics: topical structure in 2007-2017 conference papers // Dialogue-2018
- *Nikishina I.*, *Bakarov A.*, *Kutuzov A.* RusNLP: Semantic search engine for Russian NLP conference papers // AIST-2018

Датасет, содержащий:

- около 2 000 статей
- 500 ст. на английском, остальные – на русском
- во всех статьях размечены метаданные: авторы, аффилиации, аннотации и т.д.

Векторизация через **TF-IDF**

RusNLP Тематическая структура О проекте EN/RU

Ключевые слова через пробел:

syntax Russian

Показать дополнительные фильтры >>>

Конференции:

Диалог
AIST
AINL

Годы (от-до): 2002 2018

Автор: Начните набирать фам

Аффилиация: Начните набирать афф

Заголовок: *embedding*

Искать

«Application of a Hybrid Bi-lstm-crf Model to the Task of Russian Named Entity Recognition»

Burtsev M. S.; Arkhipov M. Y.; Anh L. T.; Moscow Institute of Physics and Technology, Moscow, Russia;

AINL, 2017;

Похожие публикации:

Число результатов: 10 Искать

Заголовок	Автор	Аффилиация	Конференция	Год	Задачи	Близость к запросу
Named Entity Recognition in Russian with Word Representation Learned by a Bidirectional Language Model (URL)	Filchenkov Andrey Konoplich G. Putin E. Rybka R. B.	ITMO University, Saint-Petersburg, Russia	AINL	2018	Извлечение именованных сущностей Дистрибутивная семантика	0.1981
Deep recurrent neural networks for multiple language aspect-based sentiment analysis of user reviews (URL)	Tarasov D. S.		Dialogue	2015	Извлечение отношений Анализ тональности	0.1603
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	0.1459
Comparison of Neural Network Architectures for Sentiment Analysis of	Skorniakov K. Turdakov D.	Lomonosov Moscow State University, Moscow, Russia	Dialogue	2016	Анализ тональности	0.1275

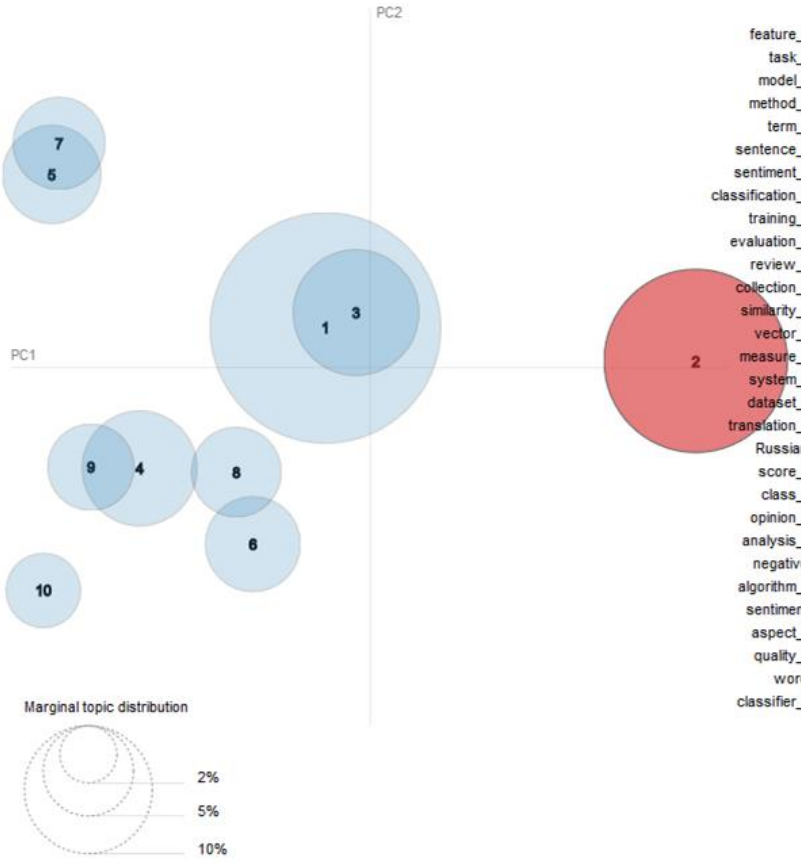
Selected Topic: 2 | Previous Topic | Next Topic | Clear Topic

Slide to adjust relevance metric:(2)

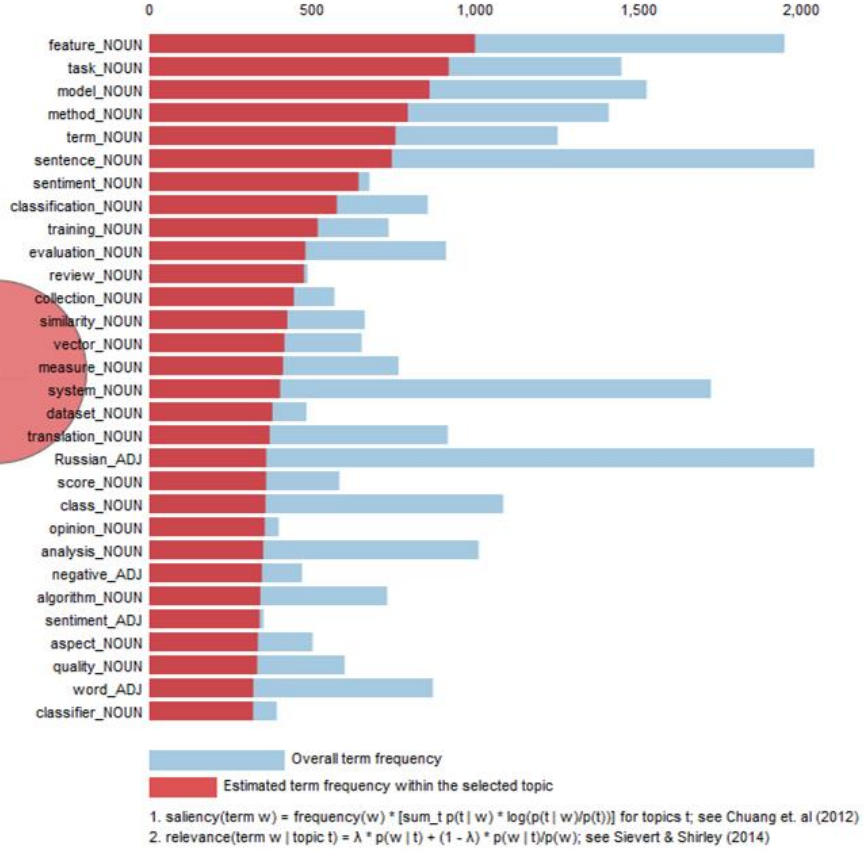
λ = 1

0.0 0.2 0.4 0.6 0.8 1.0

Intertopic Distance Map (via multidimensional scaling)



Top-30 Most Relevant Terms for Topic 2 (20.5% of tokens)



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Что еще может быть

Данные

- извлечение информации из новых статей
- структурирование базы данных
- автоматизированное извлечение данных из xml, pdf

Веб-интерфейс

RusNLP Map

Мультиязычность

- кросс-языковой поиск по статьям
- предложение похожих статей на другом языке

Академическая полиция

- выявление двойных публикаций и плагиата на основании косинусной близости

Сети

- составление графа цитирований
- выявление academic communities

Что мы будем делать

- Одновременный поиск по статьям на английском и русском
- Предложение статей на другом языке в числе похожих
- Учёт статей на обоих языках при построении тематической структуры
- Статья по результатам работы

Ближайшие Задачи

1. Разметить свежие статьи
2. Разобраться в существующих способах векторного представления текстов
3. Изучить способы проецирования кросс-языковых эмбеддингов в одно пространство
4. Оценить эти методы на тестовом заднии (рус. и англ. википедия)
5. Выбрать лучший подход и применить предобученную модель к нашим данным
6. Реализовать поиск семантически близких статей на разных языках

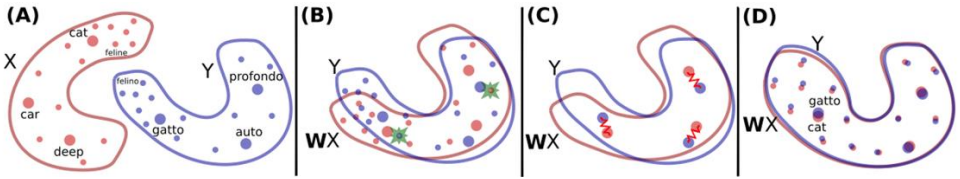
На что мы ориентируемся

Обзорная статья: Ruder S., Vulić I., Søgaard A. A Survey Of Cross-lingual Word Embedding Models. Journal of Artificial Intelligence Research 65 (2019) 569-631 (https://arxiv.org/abs/1706.04902)

	Parallel	Comparable
Word —Mapping	Mikolov et al. (2013b) Faruqui and Dyer (2014)	Bergsma and Van Durme (2011) Kiela et al. (2015)
Word —Joint	Klementiev et al. (2012) Kočiský et al. (2014)	

Approach	Monolingual	Regularizer	Joint?	Description
Klementiev et al. (2012)	$\mathcal{L}_{\text{MLE}}$	Ω_{MSE}	✓	Joint
Mikolov et al. (2013b)	$\mathcal{L}_{\text{SGNS}}$	Ω_{MSE}		Projection-based
Zou et al. (2013)	$\mathcal{L}_{\text{MMHL}}$	Ω_{MSE}		Matrix factorization
Hermann and Blunsom (2013)	$\mathcal{L}_{\text{MMHL}}$	Ω^{*}_{MSE}	✓	Sentence-level, joint

Реализация Facebook Research: MUSE: Multilingual Unsupervised and Supervised Embeddings (https://github.com/facebookresearch/MUSE)



Трудности и риски

- Сложно разобраться в теме практически с нуля
- Необходимо устанавливать соотвествия между англоязычными и русскоязычными именами и из вариантами (Анастасия Бонч-Осмоловская = Бонч-Осмоловская А. А. = Bonch-Osmolovskaya A. A. = Anastasia Bonch-Osmolovskaya)

Финальная статья

Публикация на одной из обработанных конференций:

- Обзор подходов к кросс-языковым эмбеддингам
- Архитектура кросс-языкового поиска
- Тематическая структура текстов русскоязычного NLP-сообщества с учетом добавленных статей
- Планы на будущее

Спасибо за внимание!