НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Факультет гуманитарных наук, Компьютерная лингвистика

Диахронические изменения репрезентаций слов в контекстуализированных дистрибутивно-семантических моделях

Родина Юлия

Москва, 2020

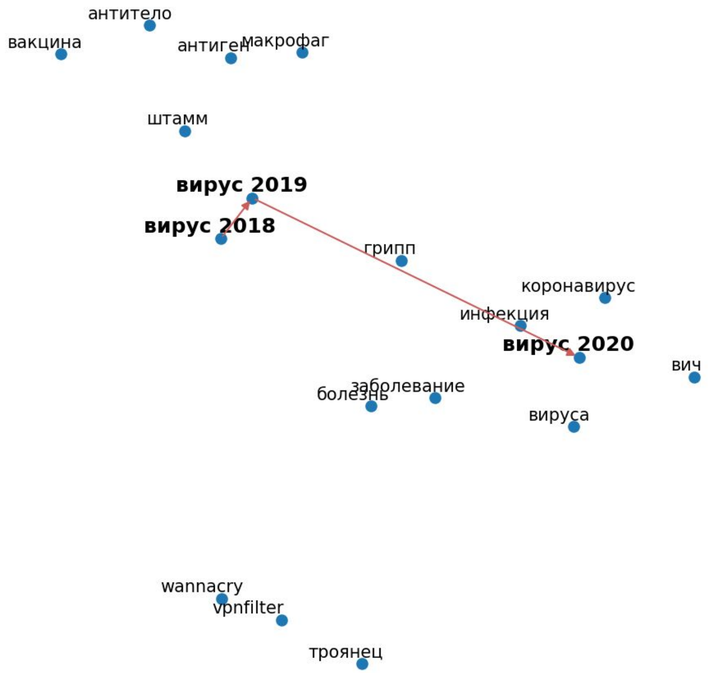
Введение

Выявление семантических изменений слов с помощью дистрибутивных контекстуализированных моделей

~ Дистрибутивная гипотеза [Firth, 1957]:
- значение слова представляется как вектор совместной встречаемости слов
- рассматриваем изменение значения (*семантический сдвиг*), как изменение контекстов

~ “Статические” модели: один вектор для каждого слова
~ *Контекстно-зависимые представления*: вектор для каждого употребления слова, как функция всего предложения

~ Культурные vs. внутрилингвистические изменения



Диахронические ~~эмбеддинги~~ векторные представления

Исследования о выявлении семантических сдвигов во времени

Обзоры

- *Kutuzov*, A. et al.: Diachronic word embeddings ar semantic shifts: A survey. In: Proceedings of the 27th international conference on computational linguistics. pp. 1384–1397 Association for Computational Linguistics, Santa Fe, New Mexico, USA (2018).

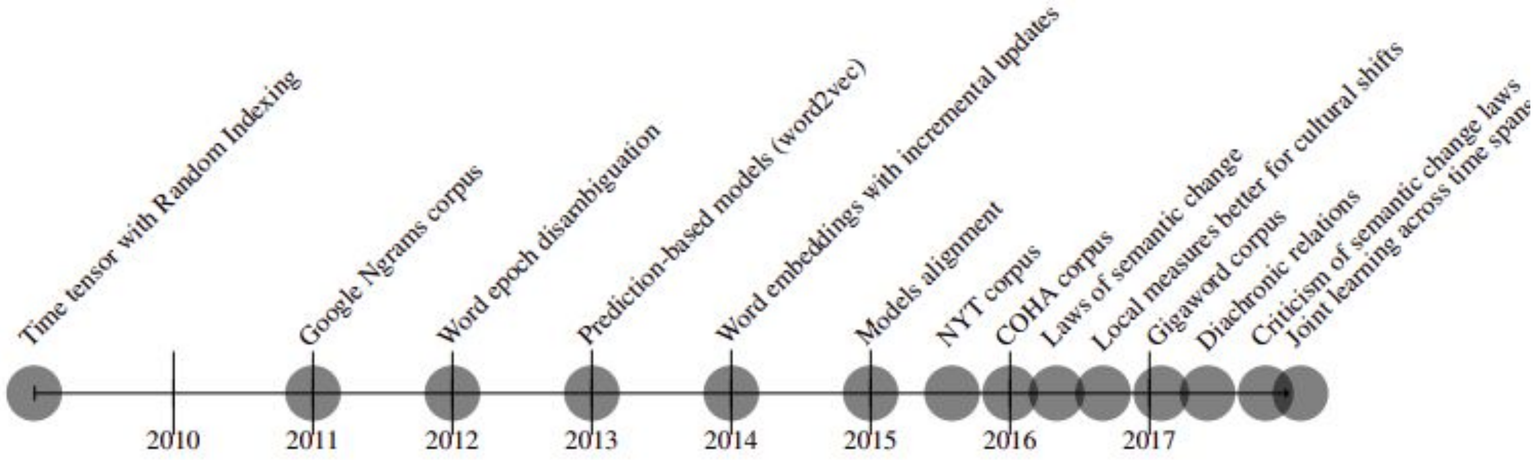


Figure 1: Distributional models in the task of tracing diachronic semantic shifts: research timeline

- *Tang*, X.: A state-of-the-art of semantic change computation. Natural Language Engineering. 24, 5, 649–676 (2018).

Диахронические векторные представления

Исследования о выявлении семантических сдвигов во времени

- 1st International Workshop on Computational Approaches to Historical Language Change, ACL-2019
- SemEval 2020 Task 1: Unsupervised Lexical Semantic Change Detection

Для русского языка

- Книга “*Два века в двадцати словах*”, 2016. Очень подробный “ручной” анализ 20 слов
- Two centuries in two thousand words: Neural embedding models in detecting diachronic lexical changes *[Kutuzov & Kuzmenko, 2018]*. Первая работа такого рода для русского языка
- Tracing cultural diachronic semantic shifts in Russian using word embeddings: test sets and baselines *[Fomin et al., 2019]*. Исследование на более гранулярных корпусах (периодом в 1 год, новостные тексты)
- Measuring diachronic evolution of evaluative adjectives with word embeddings: the case for English, Norwegian, and Russian *[Rodina et al., 2019]*. Проверка гипотезы о том, что оценочные прилагательные изменяются быстрее (нет)
- ShiftRy: Web Service for Diachronic Analysis of Russian News [*Kutuzov et al., 2020*]. Веб-сервис для работы с семантическими сдвигами в русском языке, новостные корпуса с 2010 по 2020 год

Диахронические векторные представления

Исследования о выявлении семантических сдвигов во времени

Контекстуализированные модели

Paper	Model	Data
Hu, Li, and Liang, 2019	BERT	COHA Oxford dictionary
Giulianelli, 2019; Giulianelli, Del Tredici, and Fernández, 2020	BERT	COHA GEMS dataset
Martinc, Novak,and Pollak, 2020	BERT	LiverpoolFC, Brexit news, Immigration news
Martinc et al., 2020	BERT	COHA
Kutuzov and Giulianelli, 2020	BERT ELMo	Wikipedia corpus (English, German, Latin, Swedish) SemEval-2020 test sets
Ours	***ELMo***	***RNC*** ***new datasets***

- BERT - архитектура трансформер с механизмом внимания.
- ELMo - двунаправленная двуслойная LSTM сеть поверх сверточной сети для символьных репрезентаций.

- ELMo быстрее обучается при том, что не намного уступает модели BERT во многих задачах.

Данные

Для обучения моделей и получения репрезентаций слов

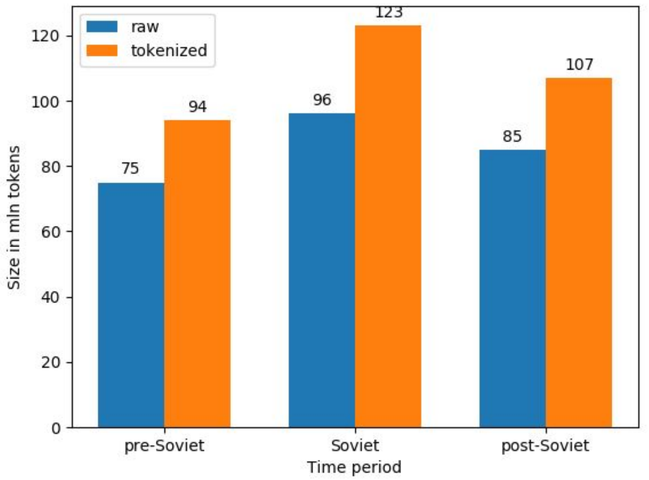
Национальный корпус русского языка

Содержит тексты на русском языке с конца 18 века по настоящее время (в работе - до 2017 года).
Корпус был разделен на 3 подкорпуса *[Kutuzov & Kuzmenko, 2018]*, периоды по несколько десятков лет:

- тексты **досоветского** периода (до 1917 года);
- тексты **советского** периода (1918-1990);
- тексты **постсоветского** периода (1991-2017).

Предобработка (*UDPipe*):

- сегментация
- токенизация (с сохранением пунктуации)
- лемматизация



Объемы корпусов в миллионах токенов до и после предобработки

Дистрибутивные контекстуализированные модели

Обучение

6 моделей в 3-х режимах обучения[1]

- модель, обученная *на текстах всего корпуса*, т.е. разделение на временные периоды происходит во время получения векторных представлений (отдельно для употреблений каждого подкорпуса);
- 3 модели, обученные *отдельно на каждом подкорпусе* (pre-Soviet model, Soviet model, post-Soviet model);
- 2 модели, обученные *инкрементально*, т.е. поверх модели, обученной на предыдущем периоде (Soviet incremental model, post-Soviet incremental model).

Сравниваем 2 пары периодов: изменение слов *с досоветского по советский* период и *советский/постсоветский*.

Для каждого анализируемого слова получаем 6 матриц размера (*n*, 512), где *n* - количество употреблений слова в корпусе.

1. Для обучения и инференса использовались модификации оригинальной имплементации из библиотеки AllenNLP: https://github.com/ltgoslo/simple_elmo_training, https://github.com/ltgoslo/simple_elmo
   3 эпохи, размер батча 192, размерность lstm уменьшена до 2048, объем словаря топ-100.000 наиболее частотных слов корпуса/подкорпуса

Датасеты

Предварительный выбор слов

досоветский/советский

- создан в работе *[Kutuzov & Kuzmenko, 2018]*
- слова отобраны вручную из “Два века в двадцати словах” и “К вопросу об изменения русского языка в советскую эпоху” Ожегова (1953)
- 43 слова: 38 существительных и 4 прилагательных

советский/постсоветский

- полностью новый датасет
- слова отобраны вручную преимущественно из словаря “Новые слова и значения” ИЛИ РАН (2009)
- 42 слова: 35 существительных и 7 прилагательных

Для каждого из этих слов генерировалось случайное слово той же части речи с похожей частотой в качестве филлера. Всего для разметки было предоставлено 142 слова: 72 и 70 слов соответственно.

Датасеты

Разметка по аналогии с фреймворком Diachronic Usage Relatedness (**DURel**) *[Schlechtweg, Walde, and Eckmann, 2018]*

- Оценка степени изменения употреблений слова по 4-балльной шкале (+ 0):

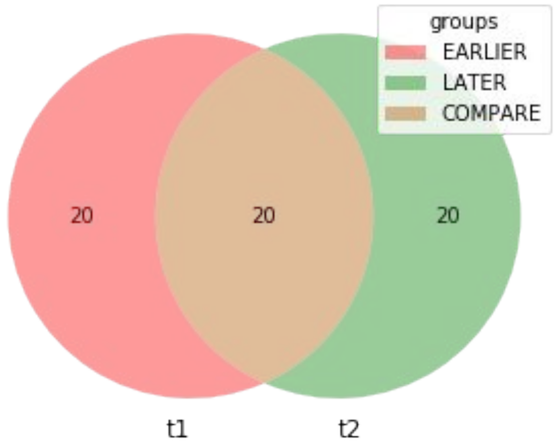
value	original meaning	translated meaning
0	Cannot decide	Не знаю
1	Unrelated	Значения разные
2	Distantly Related	Значения отдаленно похожи
3	Closely Related	Очень похожие значения
4	Identical	Употреблено в одном значении

- Сравнивались значения слова в двух разных предложениях

Датасеты

Разметка

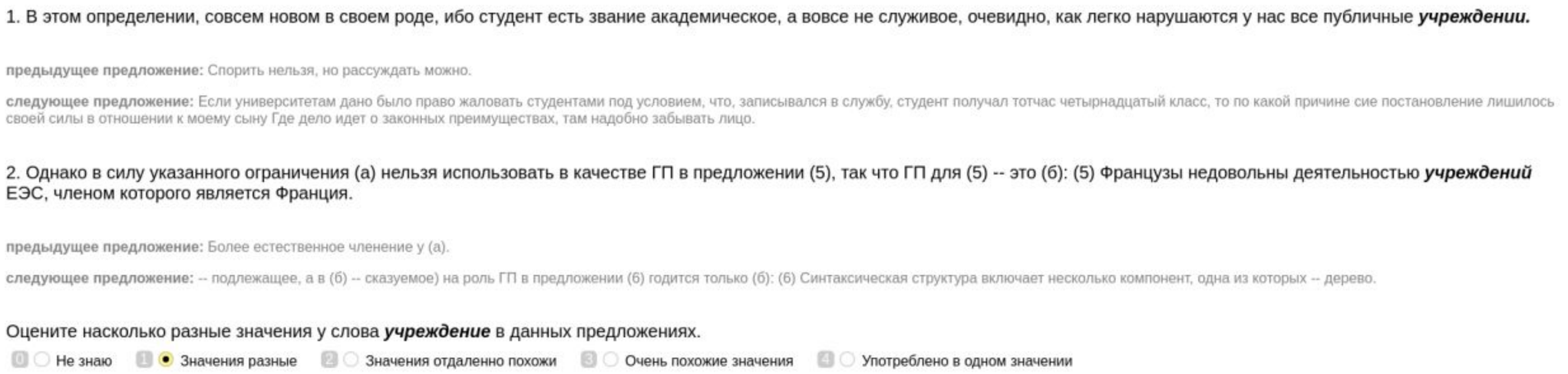
- Идея: оценивать степень изменения значения слова как разницу средних оценок пар предложений в разных периодах
- Случайным образом выбирается 20 ПАР уникальных предложений, в которых используется данное слово, из более раннего периода (EARLIER) и 20 пар из более позднего (LATER).
- Метрика изменения значения слова *w* $\Delta\text{LATER}(w) = \text{Mean}_{\text{later}}(w) - \text{Mean}_{\text{earlier}}(w)$. Разница средних оценок всех пар предложений в группах LATER и EARLIER.
- + группа COMPARE: 20 пар предложений, в которых 1 предложение из раннего периода и 1 из позднего
- Метрика $\text{COMPARE}(w) = \text{Mean}_{\text{compare}}(w)$. Среднее по всем парам предложений в этой группе



Датасеты

Разметка

- Разметка проводилась на краудсорсинговой платформе *Яндекс.Толока*
- Всего было размечено 7846 пар контекстов для 142 слов
- Каждую пару предложений размечало 5 аннотаторов
- Интерфейс задания:



Датасеты

Качество разметки

- Фильтры:
  - по возрасту (старше 30 лет)
  - высшее образование
  - 10% лучших исполнителей
  - блокировка по скорости выполнения заданий
  - блокировка по качеству контрольных заданий (в них попадали однозначные случаи, которые можно проверить в словаре, например, *Государственная Дума* и *думать думу* - значения разные)

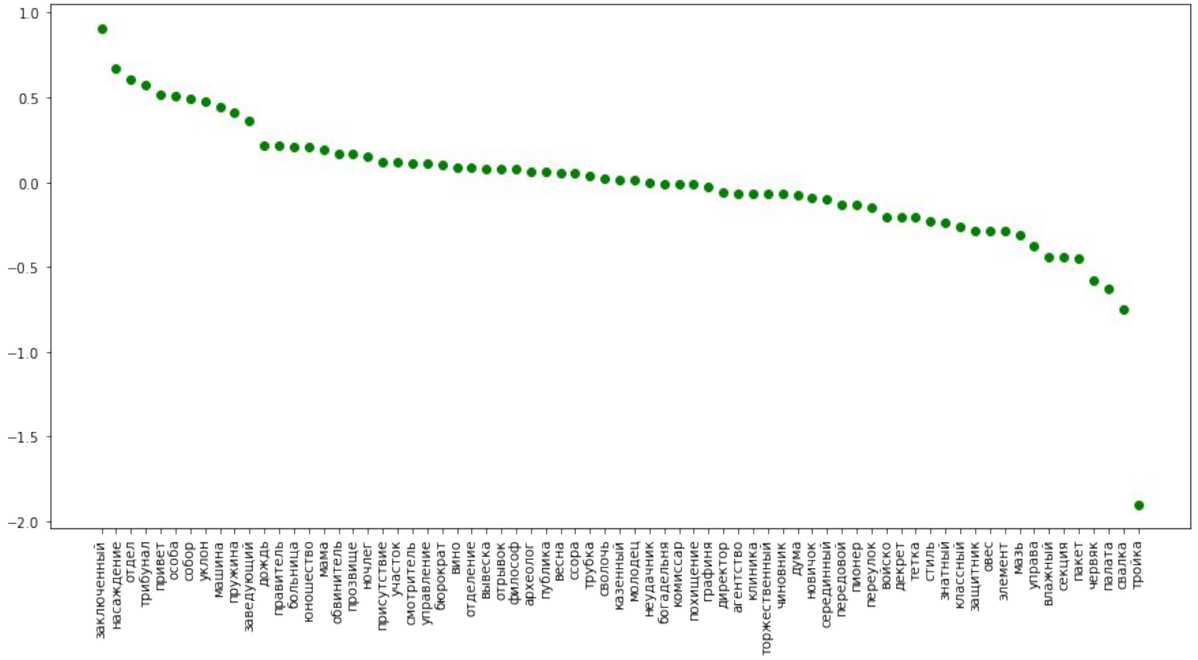
- Согласованность разметчиков в среднем по двум датасетам = 0..52 (альфа Криппендорффа)
- Из датасетов были исключены слова с согласованностью < 0.2 (fair agreement): 24 для первого датасета и 19 для второго
- Таким образом для оценки осталось 48 и 51 слово соответственно

Датасеты

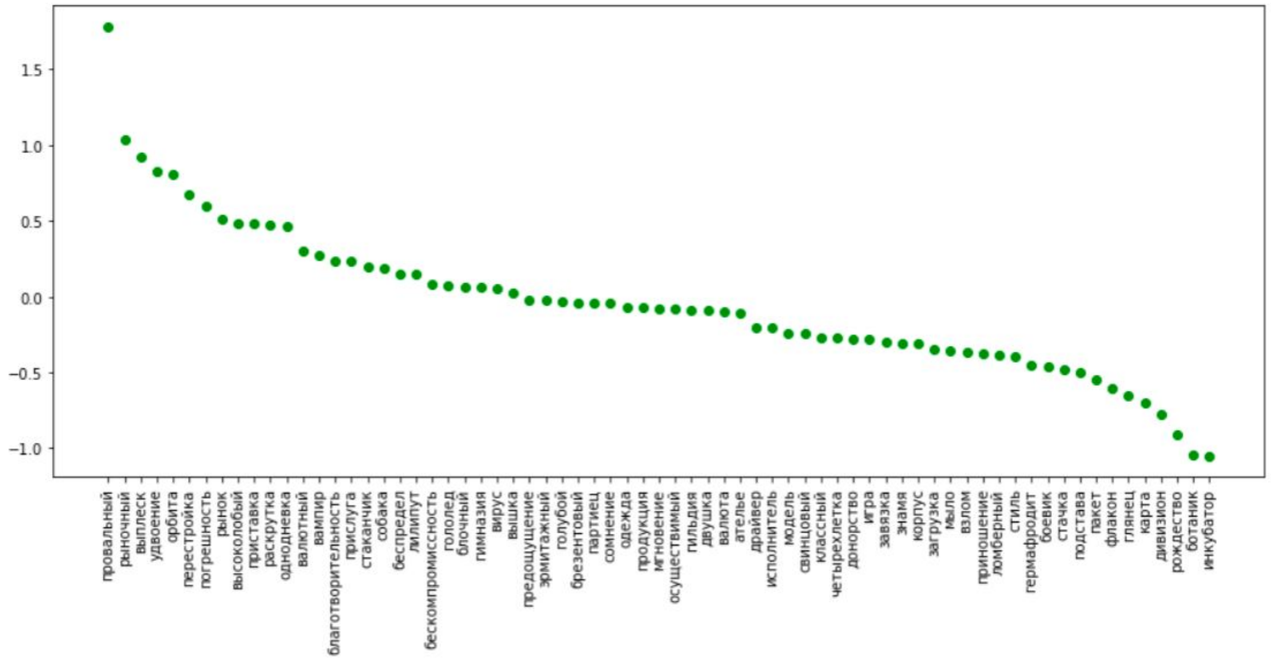
Результаты разметки

	word	COMPARE	EARLIER	LATER	delta_later	frequency_sov/postsov
0	ателье	3.29	3.30	3.19	-0.11	288/326
1	блочный	2.15	2.57	2.63	0.06	67/157
2	боевик	2.25	3.28	2.82	-0.46	231/2918
3	ботаник	2.52	3.20	2.16	-1.04	410/219
4	взлом	3.26	3.40	3.03	-0.37	99/99

- Распределение слов по значению ΔLATER



датасет 1

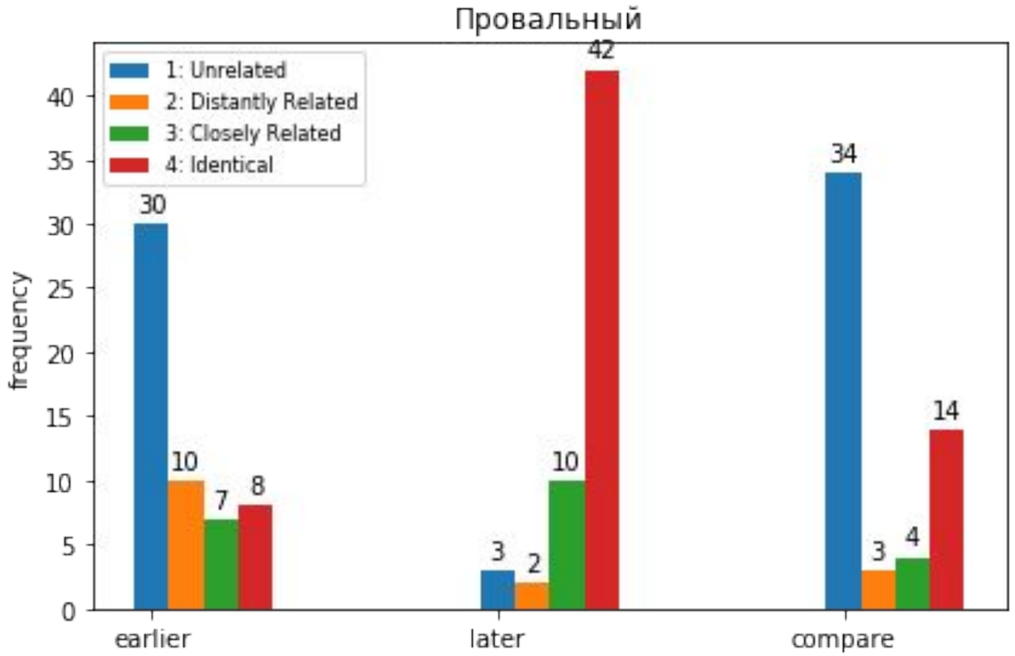


датасет 2

Датасеты

Результаты разметки

- Пример распределения оценок в разных группах контекстов одного слова



group	sentence 1	sentence 2
EARLIER	<...> выпили "микстуры"и спят где-нибудь сном провальным, пьяным <...> *'<...> they drank a "potion" and sleeping somewhere in **deep** drunken **sleep** <...>'* [Viktor Astaf'ev. Pechal'nyj detektiv (1982-1985)]	<...> Кщаре доходит до семидесяти пяти метров и что вообще здесь много провальных озер <...> *'<...> Kshchare reaches seventy-five meters and in general there are many **deep lakes** <...>'* [V. A. Solouhin. Vladimirskie prosyolki (1956-1957)]
LATER	<...> самые "бюджетные"программы оказываются самыми провальными в исполнении <...> *'<...> the cheapest programs turn out to be the most **disastrous**<...>'* [Ivan Golikov. Dohodnoe mesto // «Vsluh o...», 2003.05.19]	<...> Подготовка к референдуму в Чечне до прошлой недели носила провальный характер <...> *'<...>Preparation for referendum in Chechnya until last week was **disastrous**<...>'* [Aleksej Makarkin. Krizisnoe upravlenie «chechenizaciej» // «POLITKOM.RU», 2003.03.03]
COMPARE	<...> Наденька на минутку забылась провальным сном и когда открыла глаза <...> *'<..> Nadenka **fainted** for a minute and when she opened her eyes <..>'* [B. S. ZHitkov. Viktor Vavich. Kniga vtoraya (1941)]	Провальное выступление сборной России на Олимпиаде <...> *'The **failed** performance of the Russian team at the Olympics <...>'* [Sergej Podushkin. Opasnye igry na al'pijskom vozduhe. Gornolyzhniki nachinayut sezon // «Izvestiya», 2002.10.25]

Методы оценки

- Решаем задачу ***ранжирования*** по степени изменения смысла слова (по двум золотым метрикам - ∆LATER и COMPARE)

- **Косинусное расстояние** между усредненными эмбеддингами двух матриц
- **Дивергенция Дженсена-Шеннона** (JSD):
  - кластеризуем вектора обеих матриц алгоритмом Affinity Propagation
  - считаем количество векторов из каждого временного периода в кластерах
  - нормализованные распределения используются для подсчета JSD метрики:

$$JSD = \sqrt{\frac{D(p \parallel m) + D(q \parallel m)}{2}}$$

где D - дивергенция Кульбака-Лейблера, p и q - распределения репрезентаций двух периодов, m - их поэлементное среднее

Результаты

Корреляция Спирмена с золотым стандартом

time periods	frequency difference ΔLATER	frequency difference COMPARE
pre-Sov./Sov.	-0.235	0.003
Sov./post-Sov.	-0.002	0.001

Простая разница относительных частот слова в сравниваемых корпусах

model type	time periods	Cosine distance ΔLATER	Cosine distance COMPARE	JSD ΔLATER	JSD COMPARE
single	pre-Sov./Sov.	**0.351**	-0.246	0.158	**-0.385**
	Sov./post-Sov.	**0.3**	**-0.541**	0.147	**-0.364**
incremental	pre-Sov./Sov.	-0.068	-0.055	0.117	0.074
	Sov./post-Sov.	-0.028	-0.186	-0.028	-0.186
separate	pre-Sov./Sov.	-	-	-0.201	0.283
	Sov./post-Sov.	-	-	0.064	-0.100

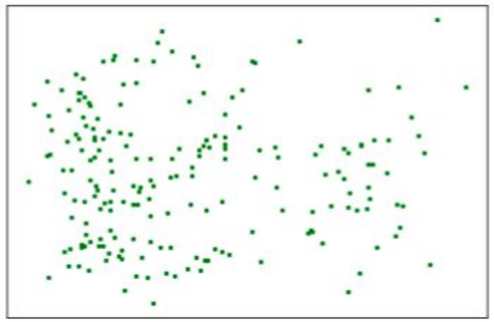
Выделены значения коэффициента для статистически значимых корреляций (p-value < 0.05)

Результаты

PCA-визуализации

2D-проекции ELMo-эмбеддингов употреблений слова пионер
в досоветском и советском периодах

‘пионер’ in pre-Soviet period; number of points: 197



‘пионер’ in Soviet period; number of points: 2241

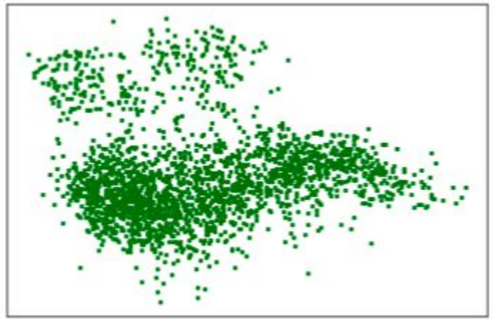


FIGURE 5.2: PCA projections of word ‘пионер’ representations in two time periods.

a) <...> на артиллерию английскую, вместе с туземной, положим 3 т. артиллеристов; затем 1000 человек инженеров, саперов и пионеров <...> ‘<...> for English artillery, together with the indigenous, we allocate 3 thousand artillerymen; then 1000 engineers, sappers and **pioneers** <...>’ [E. A. Egorov. Pohod russkoj armii v Indiyu (1855)]
b) <...> эти русские пионеры построят дорогу, они осядут вокруг этой дороги, они вдвинутся в край и вдвинут вместе с тем туда и Россию. ‘<...> these Russian **pioneers** will build a road, they will settle around this road, they will move to the region and the will move Russia there’ [P. A. Stolypin. Rech’ o sooruzhenii Amurskoj zheleznoj dorogi (1908)]
c) Словом, во всех областях ученой литературы Петрарка был настоящим пионером. ‘In all areas of scientific literature, Petrarch was a real **pioneer**.’ [A. K. Dzhivelegov. Nachalo ital’yanskogo Vozrozhdeniya (1908)]

TABLE 5.3: Usage examples from the pre-Soviet corpora for word ‘пионер’.

Заключение

- 2 новых датасета для диахронических исследований русского языка
- ELMo модели, обученные на диахронических корпусах с большой гранулярностью в 3-х вариантах
- Применены разные алгоритмы для оценки интенсивности семантических сдвигов слов во времени
- Оценена корреляция этих показателей с человеческими суждениями и показано, что для модели, обученной на всем корпусе существует статистически значимая корреляция с золотым стандартом

Направления дальнейших исследований

- Определение типа выявляемого сдвига (расширение значения, метафоризация, культурно/лингвистически обусловленный и др.)
- Выравнивание контекстуализированных моделей
- Другие части речи
- Сравнение со статичными моделями на том же датасете
- Доработка методологии разметки (чувствительна к полисемии, не связанной с диахроническими изменениями)
- Эксперименты с учетом частотности слов в разных периодах
- ...

Основные источники

- Kutuzov, Andrey and Elizaveta Kuzmenko (2018). “Two centuries in two thousand words: Neural embedding models in detecting diachronic lexical changes”. In: Quantitative Approaches to the Russian Language, pp. 95–112.
- Schlechtweg, Dominik, Sabine Schulte im Walde, and Stefanie Eckmann (June 2018). “Diachronic Usage Relatedness (DURel): A Framework for the Annotation of Lexical Semantic Change”. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). New Orleans, Louisiana: Association for Computational Linguistics, pp. 169–174.
- Hu, Renfen, Shen Li, and Shichen Liang (July 2019). “Diachronic Sense Modeling with Deep Contextualized Word Embeddings:An Ecological View”. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics, pp. 3899–3908.
- Giulianelli, Mario (July 2019). “Lexical Semantic Change Analysis with Contextualised Word Representations”. M.S. Thesis. University of Amsterdam.
- Giulianelli, Mario, Marco Del Tredici, and Raquel Fernández (2020).“Analysing Lexical Semantic Change with Contextualised Word Representations”. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Forthcoming. Association for Computational Linguistics.
- Kutuzov, Andrey and Mario Giulianelli (2020).UiO-UvA at SemEval-2020 Task 1: Contextualised Embeddings for Lexical Semantic ChangeDetection. Forthcoming.
- Martinc, Matej, Petra Kralj Novak, and Senja Pollak (2020). “Leveraging Contextual Embeddings for Detecting Diachronic Semantic Shift”. In: LREC.
- Martinc, Matej et al. (2020). “Capturing Evolution in Word Usage: Just Add More Clusters?” In: Companion Proceedings of the Web Conference 2020. WWW ’20. Taipei, Taiwan: Association for Computing Machinery, 343–349.

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Спасибо за внимание!

https://github.com/juliarodina/context-diachrony