Микродиахроническое исследование значений русских приставок методами дистрибутивной семантики

Автор: Елизавета Клыкова
Научный руководитель: Д. А. Рыжова

НИУ ВШЭ, Москва, 2022

Введение

Semantic Change Detection (SCD) – автоматическое выявление семантических сдвигов.

- активно развивается
- раньше – только на целых словах
- возможно ли на морфемах?

Введение

Сложность анализа семантики приставок:

- проблема аппроксимации значения
- только через лексемы, содержащие приставку
- → невозможно разработать золотой стандарт
- → невозможно оценить стандартным образом

Цель и задачи

Цель: подобрать компьютерный инструментарий для диахронического анализа семантики приставок.

Задачи:

0. Получить данные для моделей
1. Составить список приставок и глаголов с ними
2. Обучить модели и с их помощью получить представления нужных глаголов
3. Рассчитать изменения в семантике глаголов
4. Рассчитать изменения в семантике приставок (основываясь на пункте 3)
5. Сравнить результаты разных подходов
6. Оценить и проанализировать результаты

Существующие подходы

[Dubossarsky, 2018]: отсутствие золотого стандарта – большая проблема

[Rodina, Kutuzov, 2020]: RuSemShift – датасет из 140 русских слов, размеченный по методике Diachronic Usage Relatedness

[Kutuzov, Pivovarova, 2021] и **[Pivovarova, Kutuzov, 2021]**:

- RuShiftEval – датасет из 111 существительных, размеченный аналогичным образом
- первая задача SCD для русского языка, цель – ранжирование

Но: повторить для приставок невозможно

Материалы: данные для моделей

Диахронические датасеты НКРЯ (ок. 250 млн токенов)

Три периода:

- досоветский (1700–1916 гг.) – 72 млн токенов
- советский (1918–1991 гг.) – 93 млн токенов
- постсоветский (1992–2016 гг.) – 81 млн токенов

Материалы: данные для анализа

26 русских глагольных префиксов (с опорой на [Кронгауз, 1998]): *без-, вз-, в-, воз-, вы-, до-, за-, из-, на-, над-, недо-, низ-, о-, обез-, от-, пере-, по-, под-, полу-, пре-, пред-, при-, про-, раз-, с-, у-*

Для каждой приставки – список содержащих ее глаголов

Итог: датасет из 8434 глаголов и их частотностей в разные периоды

prefix	*lemma*	*abs_freq*	*abs_freq0*	*abs_freq1*	*abs_freq2*	*ipm_freq*	*ipm_freq0*	*ipm_freq1*	*ipm_freq2*
вз	взлетать	5316	485	2891	1940	21.555	6.717	31.075	23.835
вз	взлохмачивать	438	66	276	96	1.776	0.914	2.967	1.179
вз	взмаливаться	1342	264	584	494	5.441	3.657	6.277	6.069

Методы

Название	**Эмбеддинги**	**Леммати-зация**	**Особенности**
word2vec	статические	да	одно графическое представление = один вектор
fastText	статические	да	векторы последовательностей символов
ELMo	контекстуальные	нет	каждое вхождение слова (токен) = отдельный вектор
грам. профили [Janda, 2016]	—	нет	вектор глагола = набор грамматических разборов и их (относительных) частотностей {1.sg.praes: 4/7, ger.praes: 3/7}

Ранжированные списки

- Получаем векторы глаголов в разные периоды
- Считаем степень изменения (расстояние Манхэттена)
- Степень изменения приставки = средняя степень изменения глаголов с ней
- Ранжируем приставки по убыванию степени изменения

word2vec (леммы)	word2vec (леммы + части речи)	fastText	fastText (приставки)	грам. профили	ELMo
на-	*на-*	*полу-*	--	*на-*	*вы-*
из-	*из-*	*за-*	*о-*	*до-*	*у-*
за-	*за-*	*у-*	*до-*	*о-*	*по-*
о-	*о-*	*на-*	*пред-*	*пред-*	*под-*

Корреляция ранжированных списков

	word2vec (леммы)	*word2vec (леммы + части речи)*	*fastText*	*fastText (приставки)*	*грамм. профили*	*ELMo*
word2vec (леммы)	1.000	0.765	0.058	-0.300	0.508	-0.171
word2vec (леммы + части речи)	0.765	1.000	0.080	-0.227	0.372	-0.188
fastText	0.058	0.080	1.000	-0.143	-0.008	-0.049
fastText (приставки)	-0.300	-0.227	-0.143	1.000	-0.126	0.239
грам. профили	0.508	0.372	-0.008	-0.126	1.000	-0.151
ELMo	-0.171	-0.188	-0.049	0.239	-0.151	1.000

Анализ частотности: способ 1

Суть: взять глаголы, встретившиеся во всех периодах; отсортировать по частотности; сравнить ранжирование (коэф. корреляции Спирмена)

Результаты:	досоветский → советский	советский → постсоветский	досоветский → постсоветский
наиболее изменившиеся	*вз-*, ***от-***, *с-*, *пере-*, *о-*	***от-***, *до-*, *вы-*, *у-*, *раз-*	*воз-*, *с-*, ***от-***, *о-*, *из-*
наименее изменившиеся	***под-****, *из-*, *при-*, *до-*, *воз-*	*из-**, ***под-***, *в-*, *на-**, *о-**	*до-*, ***под-****, *вз-*, *про-**, *на-**

Проблемы: для значимой корреляции необходимо большое число глаголов; низкий порог частотности глаголов → непоказательное ранжирование

Анализ частотности: способ 2

Суть: взять топ-100 наиболее частотных глаголов с каждой приставкой в каждом периоде; сравнить процент совпадений

Результаты:	досоветский → советский	советский → постсоветский	досоветский → постсоветский
наиболее изменившиеся	*о-, пере-, с-,* ***вз-, вы-***	*на-,* ***вы-, вз-****, за-, у-*	*у-,* ***вы-****, на-,* ***вз-****, при-*
наименее изменившиеся	*за-,* ***в-****, из-, раз-,* ***по-***	*про-, с-, о-,* ***в-, по-***	*про-,* ***в-****, за-,* ***по-****, от-*

Проблемы: невозможность анализа менее продуктивных приставок

Word2vec: кластеризация (на примере приставки *по-*)

[Мустайоки, Пуссинен, 2008]: экспансия приставки *по-*; в словарях нет глаголов типа *поразбросать, понавыдумывать, пооборвать, понавесить*

В наших данных:

- в постсоветском периоде выделяется кластер “*повыдергать, повыдергивать, повылазить, повылезать, повыскакивать, повыходить*”
- точность кластеризации выше в более поздних периодах: так, в постсоветском периоде выделяется кластер “*поесть, позавтракать, покушать, пообедать, поужинать*” (в советском он оказывался вместе с глаголами *побегать* и *помыться*)
- семантика корней наиболее значима

FastText

Два варианта:

- анализ приставок на основе эмбеддингов глаголов (как с word2vec)
- анализ непосредственно приставок

Не коррелируют между собой; fastText на приставках имеет слабую отрицательную корреляцию с word2vec и слабую положительную с ELMo

Проблема: много шума при анализе приставок как отдельных единиц (пересекаются с предлогами)

Грамматические профили + ELMo: кластеризация

1. Берем 5 наиболее изменившихся глаголов по методу грамматических профилей (*заметать, постановлять, изготавливаться, наследовать, претерпеть*)
2. Выбираем 100 случайных контекстов каждого глагола в каждом периоде
3. Получаем эмбеддинги глаголов с помощью ELMo
4. Кластеризуем с помощью иерархической кластеризации

Грамматические профили + ELMo: *изготавливаться*

В досоветском периоде типичны контексты типа (1):

(1) *Употребив следующий за тем день на отдание грустного долга нашему общему любимцу, мы* ***изготовились*** *в ночь и выступили на зорьке в дальнейший путь.* [НКРЯ]

В советском и постсоветском периодах все больше распространяются случаи типа (2):

(2) ***Изготавливается*** *коптильня из листовой стали или листового железа толщиной 0,8–1 мм.* [НКРЯ]

Грамматические профили + ELMo: *претерпеть*

В досоветском – почти всегда в значении *претерпеть лишения*:

(3) *В темной глубине его сердца все ярче разгорался злой огонь ревности, раздуваемый самолюбием человека,* ***претерпевшего*** *много унижений и обид.* [НКРЯ]

В советском и постсоветском распространяются употребления со значением *претерпеть изменения*:

(4) *Как сообщили в кредитном отделе банка, скорее всего, изменения* ***претерпит*** *сама схема выдачи ипотечных ссуд.* [НКРЯ]

Грамматические профили + ELMo: *заметать*

Расширение семантики глагола *заметать* в постсоветском периоде:

(5) *Было дело, **замели** его на пятнадцать суток – еще в империи, так он, гуляя по двору вытрезвителя, какому-то случайному прохожему продал по дешевке казенный мотоцикл […].* [НКРЯ]

(6) *Отец для паузы деловито откашлялся и, глазом не моргнув, набрехал, что Аня сегодня не сможет спуститься с неба, уж слишком снег густой, погода нелетная, но вот завтра – **заметано**!* [НКРЯ]

Результаты

1. Датасет из 8434-х глаголов с 26-ю глагольными приставками (+ частотности глаголов в трех периодах)
2. Эксперименты с разными архитектурами и их сравнение
3. Подходы к анализу изменений: анализ частотности и сочетаемости, кластеризация
4. Набор case studies, иллюстрирующих изменения в сочетаемости приставок и семантике глаголов

Выводы

Применение стандартных методов SCD осложняется нестандартностью задачи

Сочетаемость приставок с корнями, частотность глаголов и их разделение на кластеры – индикаторы диахронических сдвигов

Word2vec и fastText – обобщенная картина, ELMo – детальный анализ каждого случая

Метод грамматического профилирования и кластеризации эмбеддингов ELMo коррелирует с реальными изменениями

Итог: не система для решения компьютерной задачи, а компьютерный инструментарий для теоретических исследований

Полезные ссылки

Github-репозиторий дипломной работы

Датасет глаголов и приставок

Избранная литература

Кронгауз М. А. Приставки и глаголы в русском языке: семантическая грамматика. Москва: Школа «Языки русской культуры», 1998. 286 с.

Мустайоки А., Пуссинен О. Об экспансии глагольной приставки ПО в современном русском языке // *Инструментарий русистики: корпусные подходы*. Helsinki: Helsinky University Press, 2008. С. 247–275.

Dubossarsky H. Semantic change at large: A computational approach for semantic change research: PhD thesis. Hebrew University of Jerusalem, 2018. 75 p.

Hamilton W. L., Leskovec J., Jurafsky D. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change // *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. 2016.

Janda L. A. Linguistic profiles: A quantitative approach to theoretical questions // *Język i metoda*. 2016. № 3. P. 127–145.

Kutuzov A., Pivovarova L. Three-part diachronic semantic change dataset for Russian // *Proceedings of the 2nd International Workshop on Computational Approaches to Historical Language Change 2021*. Online: Association for Computational Linguistics, 2021. P. 7–13.

Pivovarova L., Kutuzov A. RuShiftEval: a shared task on semantic shift detection for Russian // *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2021".* 2021. P. 533–545.

Rachinskiy M., Arefyev N. Zero-shot Crosslingual Transfer of a Gloss Language Model for Semantic Change Detection // *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2021"*. 2021. P. 578–586.

Ryzhova A., Ryzhova D., Sochenkov I. Detection of Semantic Changes in Russian Nouns with Distributional Models and Grammatical Features // *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2021"*. 2021. P. 597–606.

Ответы на замечания: отсутствие золотого стандарта

Как разметить руками?

Если все же разметить, какой смысл в автоматизированном подходе?

Ответы на замечания: расстояние Левенштейна

Список = слово, глаголы = буквы

Количество перестановок, произошедших в ранжированных по частотности списках

Оценка на материале RuShiftEval

Модель	Период	Корреляция с зол. стандартом
word2vec (леммы)	досоветский → советский	-0.004 (p=0.964)
word2vec (леммы + части речи)		-0.040 (p=0.674)
fastText		-0.0348 (p=0.717)
word2vec (леммы)	советский → постсоветский	0.037 (p=0.696)
word2vec (леммы + части речи)		-0.027 (p=0.776)
fastText		-0.078 (p=0.413)
word2vec (леммы)	досоветский → постсоветский	-0.138 (p=0.149)
word2vec (леммы + части речи)		-0.116 (p=0.226)
fastText		0.005 (p=0.960)

Почему?

[Pivovarova, Kutuzov, 2021] – значимая корреляция на baseline-методе (эмбеддинги word2vec + ближайшие соседи):

Досоветский → советский	**Советский → постсоветский**	**Досоветский → постсоветский**	**В среднем**
0.314*	0.302*	0.381*	0.332*

[Ryzhova et al., 2021] – низкое качество метода “word2vec + косинусное расстояние”:

Досоветский → советский	**Советский → постсоветский**	**Досоветский → постсоветский**	**В среднем**
0.141	0.246*	0.330*	0.239*

Почему?

А у нас?

- другой препроцессинг (MyStem)
- другая метрика (расстояние Манхэттена, по [Rachinskiy, Arefyev, 2021])

Метрика	**Досоветский → советский**	**Советский → постсоветский**	**Досоветский → постсоветский**
расстояние Манхэттена	-0.004 (p=0.964)	0.037 (p=0.696)	-0.138 (p=0.149)
косинусное расстояние	-0.025 (p=0.795)	0.197* (p=0.038)	0.063 (p=0.513)

Ответы на замечания: достоверность ранжирования

Ранжирование спорно: оценить его невозможно, а оценка на RuShiftEval низка

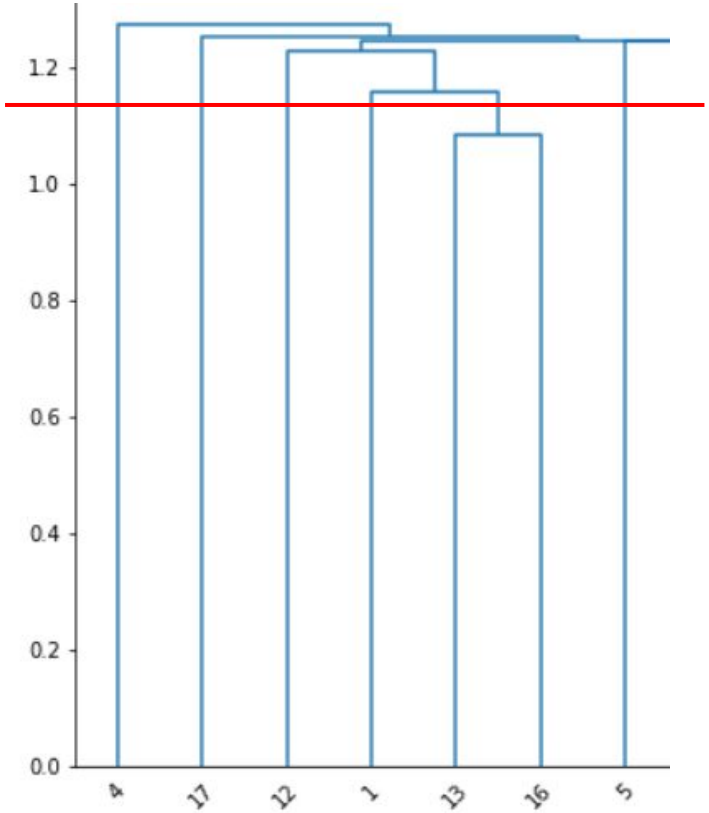
Но: степень “согласия” моделей – косвенный указатель на достоверность

А также: любая оценка (на RuShiftEval или на глаголах) опосредована

Ответы на замечания: иерархическая кластеризация

В работе показана не вся иерархия,
а срез при определенном пороге

Смысл подхода в том,
чтобы не задавать число кластеров



Существующие подходы

[Hamilton et al., 2016] – общие принципы диахронических изменений:

- law of conformity – более частотные слова меняются медленнее
- law of innovation – более многозначные слова меняются быстрее

[Dubossarsky et al., 2017]:

- многозначность тесно связана с частотностью
- выдвинутые принципы – отчасти артефакты моделей