Diachronic Word Embeddings. Related work

Бакшандаева Дарья
Родина Юлия
Фомин Вадим
Куратор: Андрей Кутузова

Temporal Analysis of Language through Neural Language Models [Kim et al. 2014]

Самый первый и базовый подход

Датасет: Google Books

Алгоритм: нарезать на слайсы, дообучать модель с предыдущих слайсов на новом

Расстояние: косинусная близость (т. к. модели сравнимые)

Гранулярность: один год

Анализ: руками (наиболее + наименее изменившиеся и т. п.)

Word	Neighboring Words in	
	1900	2009
gay	*cheerful pleasant brilliant*	*lesbian bisexual lesbians*
cell	*closet dungeon tent*	*phone cordless cellular*
checked	*checking recollecting straightened*	*checking consulted check*
headed	*haired faced skinned*	*heading sprinted marched*
actually	*evidently accidentally already*	*really obviously nonetheless*

Table 2: Top 3 neighboring words (based on cosine similarity) specific to each time period for the words identified as having changed.

Temporal Analysis of Language through Neural Language Models [Kim et al. 2014]

Плюсы

+ Прост в идее и реализации
+ Поэтому легко использовать
+ Показал значимые изменения в семантике

Минусы

- Неустойчив к уменьшению корпуса
- Нет автоматической оценки качества

Statistically Significant Detection of Linguistic Change [Kulkarni et al. 2015]

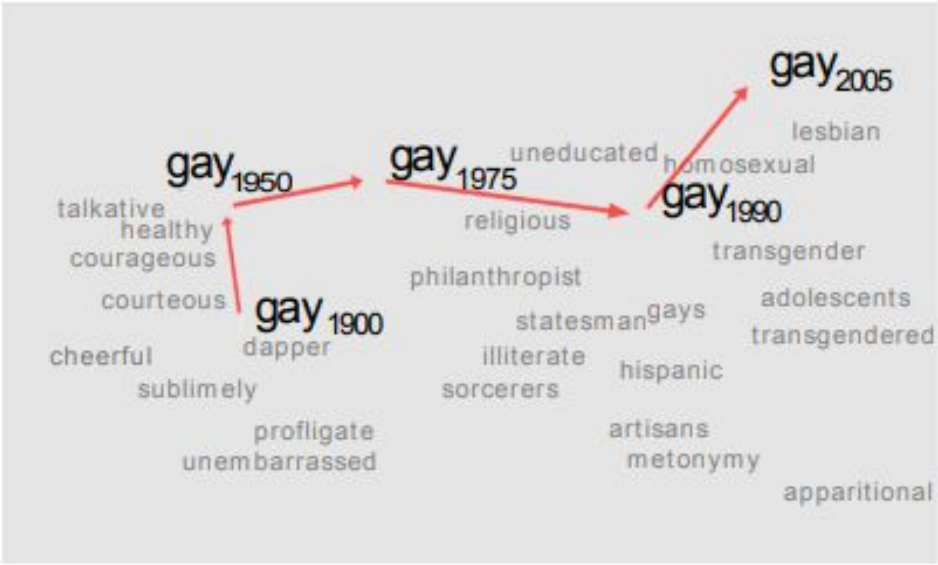


Figure 1: A 2-dimensional projection of the latent semantic space captured by our algorithm. Notice the semantic trajectory of the word **gay** transitioning meaning in the space.

	Google Ngrams	Amazon	Twitter
Span (years)	105	12	2
Period	5 years	1 year	1 month
# words	$\sim 10^9$	$\sim 9.9 \times 10^8$	$\sim 10^9$
$\|\mathcal{V}\|$	$\sim$50K	$\sim$50K	$\sim$100K
# documents	$\sim 7.5 \times 10^8$	$8. \times 10^6$	$\sim 10^8$
Domain	Books	Movie Reviews	Micro Blogging

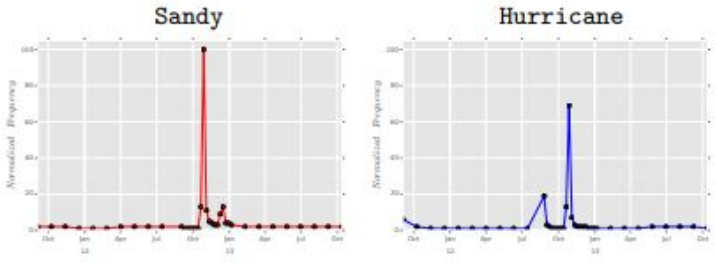
Table 1: Summary of our datasets

- Time series for each word.
- Three approaches: Frequency, Syntactic, and Distributional.

Statistically Significant Detection of Linguistic Change [Kulkarni et al. 2015]

Почему частотный и синтаксический методы не очень?

- Резкое увеличение частоты не обязательно = изменение смысла, много false positive.
- Изменение смысла не обязательно = изменение ЧР (*мышка*) + нужен хорошо аннотированный корпус, много false negative.



(a) **Frequency method (Google Trends)**

(b) **Distributional method**

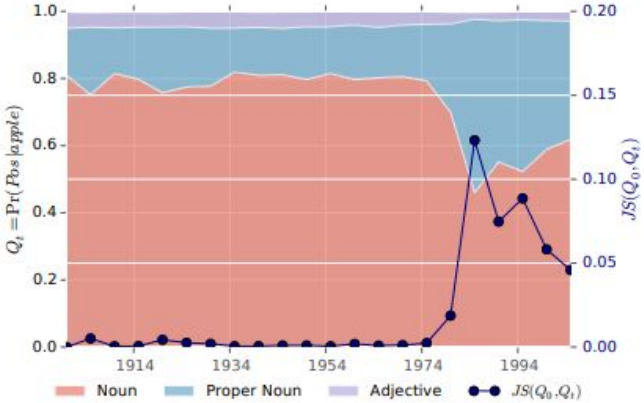


Figure 4: Part of speech tag probability distribution of the word `apple` (stacked area chart). Observe that the "Proper Noun" tag has dramatically increased in 1980s. The same trend is clear from the time series constructed using Jenssen-Shannon Divergence (dark blue line).

Statistically Significant Detection of Linguistic Change [Kulkarni et al. 2015]

Distributional Method

Это клёво, потому что хороший баланс FP и FN, не нужны лингвистические ресурсы.

- learning:
  - максимизация вероятностей слов контекста слова w;
  - минимизация negative log-likelihood;
  - оптимизация параметров с помощью стохастического градиентного спуска;
  - нормализация эмбеддингов по эвклидовой метрике;
  - я тоже ничего не поняла;
  - в экспериментах использовали реализацию skipgram в gensim.
- aligning:
  - чтобы все эмбеддинги были в единой системе координат;
  - кусочно-линейная регрессия.
- time series construction.

Statistically Significant Detection of Linguistic Change [Kulkarni et al. 2015]

Примеры результатов:

-- *Distributional better*

gay	1985	happy and gay	gay and lesbians
tape	1970	red tape, tape from her mouth	a copy of the tape
bitch	1955	nicest black bitch (Female dog)	bitch (Slang)

-- *Syntactic better*

bush	1989	Noun (bush and a shrub)	Proper Noun (George Bush)
apple	1984	Noun (apple, orange, grapes)	Proper Noun (Apple computer)
handle	1951	Noun (handle of a door)	Verb (he can handle it)

-- *Twitter & Amazon*

twilight	2009	twilight as in dusk	Twilight (The movie)
ray	2006	ray of sunshine	Blu-ray
snap	Dec 2012	snap a picture	snap chat
shades	Jun 2012	color shade, shaded glasses	50 shades of grey (The Book)

Statistically Significant Detection of Linguistic Change [Kulkarni et al. 2015]

+ **исходный код**;
+ масштабируемость на тексты разных предметных областей;
+ change point detection (точка в которой новое использование начинает доминировать);
+ есть применимый алгоритм оценки: искусственно созданные изменения; с использованием референса; с использованием людей.

- сложная математика :-(;

Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change [Hamilton et al. 2016]

1. PPMI - positive point-wise mutual information.

$$\mathbf{M}_{i,j}^{\text{PPMI}} = \max\left\{\log\left(\frac{\hat{p}(w_i, c_j)}{\hat{p}(w)\hat{p}(c_j)}\right) - \alpha, 0\right\}$$

2. SVD - singular value decomposition, сингулярное разложение.

$$\mathbf{w}_i^{\text{SVD}} = (\mathbf{U}\boldsymbol{\Sigma}^{\gamma})_i$$

3. SGNS - skip-gram with negative sampling.

$$\hat{p}(c_i|w_i) \propto \exp(\mathbf{w}_i^{\text{SGNS}} \cdot \mathbf{c}_j^{\text{SGNS}})$$

Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change [Hamilton et al. 2016]

Данные

Name	Language	Description	Tokens	Years	POS Source
ENGALL	English	Google books (all genres)	8.5×10^{11}	1800-1999	(Davies, 2010)
ENGFIC	English	Fiction from Google books	7.5×10^{10}	1800-1999	(Davies, 2010)
COHA	English	Genre-balanced sample	4.1×10^{8}	1810-2009	(Davies, 2010)
FREALL	French	Google books (all genres)	1.9×10^{11}	1800-1999	(Sagot et al., 2006)
GERALL	German	Google books (all genres)	4.3×10^{10}	1800-1999	(Schneider and Volk, 1998)
CHIALL	Chinese	Google books (all genres)	6.0×10^{10}	1950-1999	(Xue et al., 2005)

Table 1: Six large historical datasets from various languages and sources are used.

Гиперпараметры: окна = 4, эмбеддинги = 300.
Выравнивание: ортогональный Прокрустов анализ (?? wut)
Измерение:

$$s^{(t)}(w_i, w_j) = \text{cos-sim}(\mathbf{w}_i^{(t)}, \mathbf{w}_j^{(t)}) \qquad \text{cos-dist}(\mathbf{w}_t, \mathbf{w}_{t+\Delta}),$$

Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change [Hamilton et al. 2016]

Evaluation

- *Synchronic Accuracy*: SVD performed best (ρ = 0.739), followed by PPMI (ρ = 0.687) and SGNS (ρ = 0.649).
- *Diachronic Validity* (28 known shifts and the top-10 “discovered” shifts by each method)
  - Detecting known shifts (приближаются или отдаляются слова)
    awful | → *disgusting, mess* | ← *impressive, majestic* (<1800)
    Почти все методы идеально определили направление, но хуже определяли статистическую значимость изменений.
  - Discovering shifts from data
    Разметка топ-10 изменений, как подлинных, пограничных или артефактов.
    SGNS 70%, 40% for SVD, 10% for PPMI.:

Method	Top-10 words that changed from 1900s to 1990s
PPMI	know, got, would, decided, think, stop, remember, **started**, must, wanted
SVD	harry, **headed**, **calls**, **gay**, wherever, male, **actually**, special, cover, naturally
SGNS	**wanting**, **gay**, **check**, **starting**, **major**, **actually**, touching, harry, **headed**, romance

Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change [Hamilton et al. 2016]

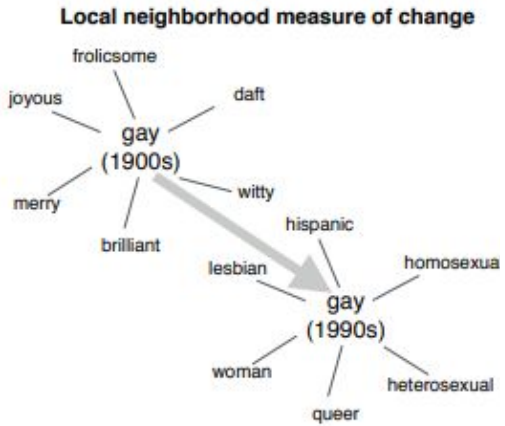
+ **исходный код**;
+ не только английский;
+ разные алгоритмы построения эмбеддингов;
+ laws of semantic change: conformity & innovation

- сейчас есть более прикольные подходы.

Cultural Shift or Linguistic Drift? Comparing Two Computational Measures of Semantic Change [Hamilton et al. 2016]

- **linguistic** *promise* (“I promise.”→“It promised to be exciting.”);
- **global**;
- >**verbs**;
- **global** measure ;

- **cultural** *cell* (“prison cell” → “cell phone”);
- **local**;
- >**nouns**;
- **local neighborhood** measure;



Dynamic Word Embeddings [Bamler & Mandt 2017]

Датасет: Google Books, обращения президентов США, корпус твиттера

Алгоритм: обычный скипграм → байесовский скипграм → байесовский скипграм с латентной переменной времени
Фильтрация (видим прошлое) vs. сглаживание (мостик между прошлым и будущим)

Анализ: визуализация (см. след. слайд) и эвалюейшн

Эвалюейшн: случайно выбираем пары “слово / контекст”, просим предсказать год, считаем по формулам теории вероятности:

$$\frac{1}{|n_t^{\pm}|} \log p(n_t^{\pm} | \tilde{U}_t, \tilde{V}_t).$$

(Взвешиваем по числу всех пар “слово / контекст” за этот год)
Затем сравниваем с бейслайнами (см. след слайд)

Dynamic Word Embeddings [Bamler & Mandt 2017]

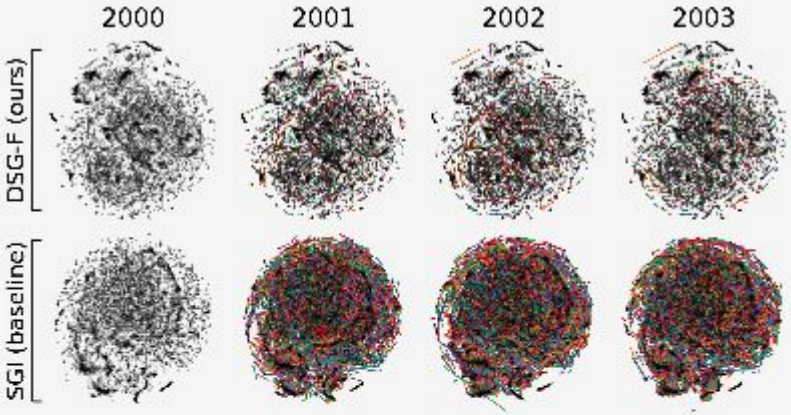


Figure 3. Word embeddings over a sequence of years trained on Google books, using DSG-F (proposed, top row) and compared to the static method by Hamilton et al. (2016) (bottom). We used dynamic t-SNE (Rauber et al., 2016) for dimensionality reduction. Colored lines in the second to fourth column indicate the trajectories from the previous year. Our method infers smoother trajectories with only few words that move quickly. Figure 4 shows that these effects persist in the original embedding space.

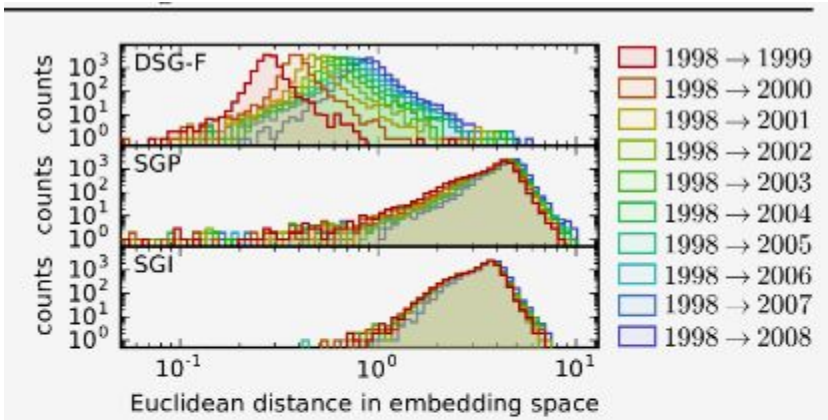


Figure 4. Histogram of distances between word vectors in the year 1998 and their positions in subsequent years (colors). DSG-F (top panel) displays a continuous growth of these distances over time, indicating a directed motion. In contrast, in SGP (middle) (Kim et al., 2014) and SGI (bottom) (Hamilton et al., 2016), the distribution of distances jumps from the first to the second year but then remains largely stationary, indicating absence of a directed drift; i.e. almost all motion is random.

Dynamic Word Embeddings [Bamler & Mandt 2017]

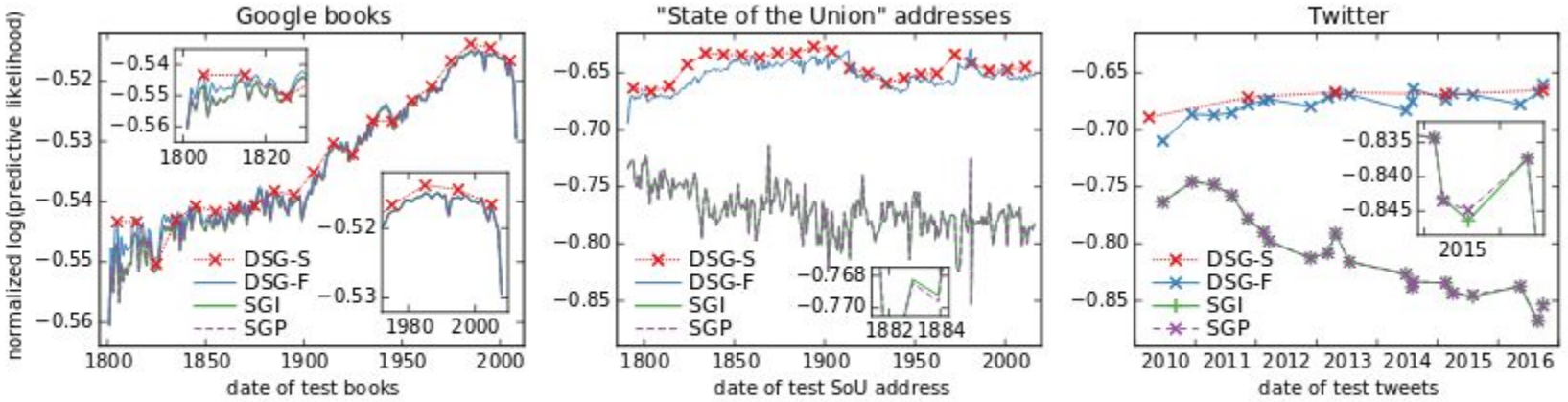


Figure 6. Predictive log-likelihoods (Eq. 16) for two proposed versions of the dynamic skip-gram model (DSG-F & DSG-S) and two competing methods SGI (Hamilton et al., 2016) and SGP (Kim et al., 2014) on three different corpora (high values are better).

Dynamic Word Embeddings [Bamler & Mandt 2017]

Плюсы

+ Устойчив к отсутствию данных по отдельным слайсам (годам / месяцам /...)
+ Качественный эвалюейшн, работает хорошо и это доказано
+ Гранулярность почти не ограничена снизу

Минусы

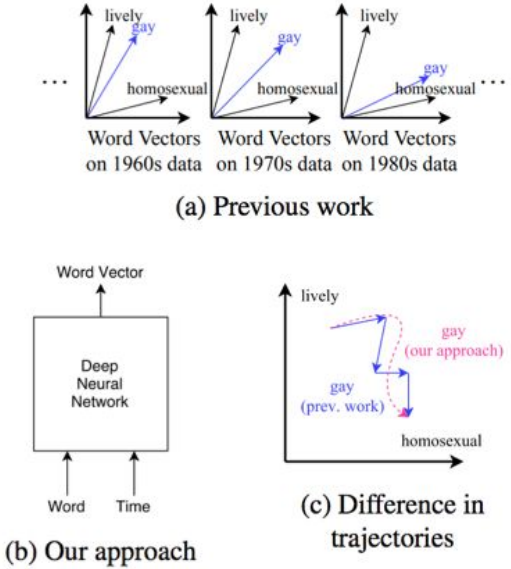
- Сложная математическая сторона дела
- Нет открытого исходного кода, разобраться так, чтобы написать самим, в ближайшем будущем невозможно

Deep Neural Models of Semantic Shift [Rosenfeld & Erk 2018]

Проблема:

большие слайсы → грубая, ‘крупнозернистая’ репрезентация диахронических лексических изменений

маленькие слайсы → модели обучаются на данных, которых недостаточно



Решение:

использующая глубокие нейронные сети модель, в которой время представлено непрерывной переменной, а употребление слов моделируется как функция времени

Deep Neural Models of Semantic Shift [Rosenfeld & Erk 2018]

Датасет: секция английской художественной литературы из Google Books ngram corpus, произведения, написанные в период 1900-2009 гг.

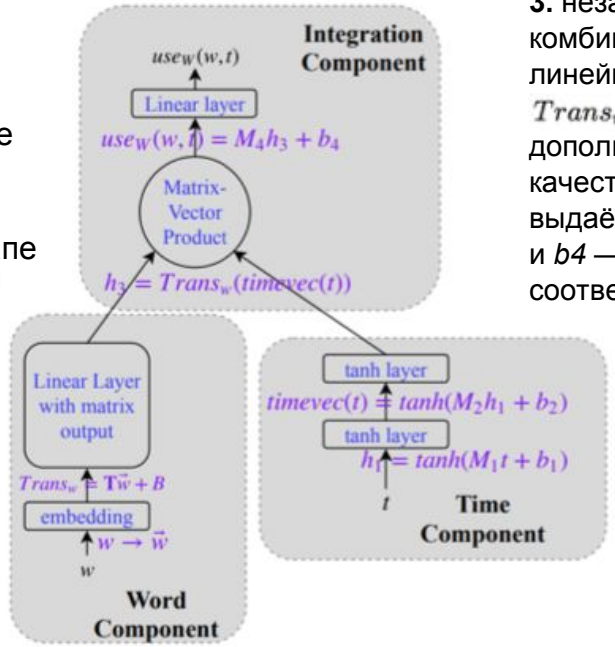
DiffTime Model: модификация алгоритма ***SGNS*** (создание дифференцируемой функции $use_W(w,t)$, которая возвращает целевой эмбеддинг для целевого слова *w* во время *t*, и дифференцируемой функции $use_C(c,t)$, которая возвращает контекстный эмбеддинг для контекстного слова *c* во время *t*).

3 компонента:

- *time component* — принимает на вход время и выдаёт эмбеддинг, который характеризует данную точку во времени
- *word component* — принимает на вход слово и выдаёт независимый от времени словесный эмбеддинг, преобразующийся затем в набор параметров, которые могут модифицировать временной эмбеддинг
- *integration component* — комбинирование временного и словесного эмбеддингов

Deep Neural Models of Semantic Shift [Rosenfeld & Erk 2018]

Архитектура *функции*
$use_W(w,t)$;
соответствующая ей ***функция***
$use_C(c,t)$
для контекстных слов имеет такую же структуру (в том числе делит веса с функцией целевого слова), за исключением того, что на втором этапе используется свой набор векторов $\vec{c}$



2. каждое целевое слово w получает векторную репрезентацию $\vec{w}$; с помощью модифицированного линейного слоя, в котором веса представлены трёхмерным тензором, а значения смещения — матрицей, вектор $\vec{w}$ преобразуется в матрицу $Trans_w$

3. независимый от слова эмбеддинг $timevec(t)$ комбинируется с независимым от времени линейным преобразованием $Trans_w$: $Trans_w$ применяется к $timevec(t)$; дополнительный линейный слой используется в качестве выходного — он принимает на вход h_3 и выдаёт получившуюся функцию $use_W(w,t)$, где *M4* и *b4* — веса и значения смещения выходного слоя соответственно

1. двухслойная нейронная сеть прямого распространения, с гиперболическим тангенсом в качестве активационной функции; на вход принимается время t (временная точка масштабируется до значения от 0 до 1, где 0 — 1900 г., 1 — 2009 г.), выход — временной эмбеддинг $timevec(t)$; *M1* и *M2* — веса первых двух слоёв, *b1* и *b2* — значения смещения

Deep Neural Models of Semantic Shift [Rosenfeld & Erk 2018]

Модель дифференцируема по времени → получаем производную функции $use_W(w,t)$ по *t*, чтобы смоделировать, как меняется употребление слова *w* в момент времени *t*, с какой скоростью → исследуем связь между скоростью лексических изменений слова и его ближайшими соседями

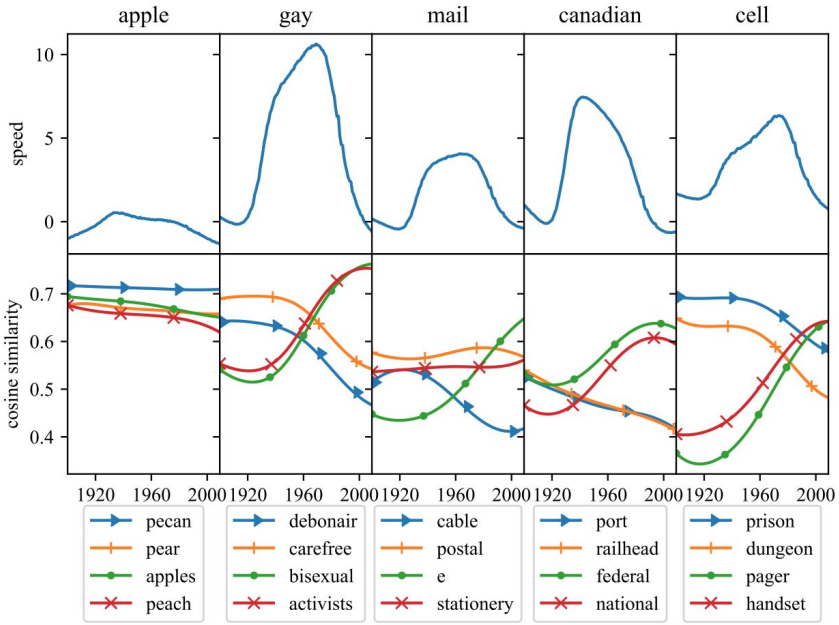


Figure 4: Speed and nearest neighbors over time of selected words. The top graphs as the speed at which a word changes usage according to our model. The bottom graphs are selected nearest neighbors for those words. Each of the chosen nearest neighbors appear as a top 10 nearest neighbor to the word at some year.

Deep Neural Models of Semantic Shift [Rosenfeld & Erk 2018]

Оценка качества модели:

- семантическая близость слов (MEN), сравнение с некоторыми другими моделями
- искусственная задача: создание ‘синтетических’ слов, которые со временем меняют своё значение по сигмоидальному пути; искусственные слова — комбинация реальных слов (*banana∘lobster*), относящихся к различным классам в базе данных BLESS

  Слабости: качество модели оценивается на искусственных словах — нельзя точно сказать, насколько она успешна применительно к реальным; генерируются слова, которые меняют своё значение от одного к другому — не учитываются другие изменения (например, сужение/расширение значения); непрерывные модели, включающие сигмоидальную функцию в свою архитектуру, могут оказаться в привилегированном положении

Method	Time	Spearman's ρ
LargeBin	1990s bin	0.615
SmallBinPreInit	1995 bin	0.489
SmallBinReg	1995 bin	0.564
DiffTime	start of 1995	**0.694**

Table 1: Synchronic accuracy of the methods. Time is the point of time we use as our synchronic model.

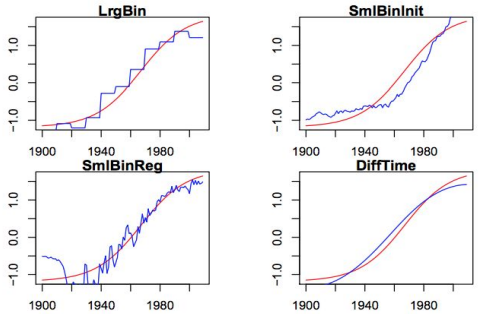


Figure 3: Graph Comparisons between $shift(t; r_1 \circ r_2)$ (red) and $rec(t; r_1 \circ r_2,)$ (blue) for the synthetic word $pistol \circ elm$. The x-axis are the years and the y-axis are the values of $shift(t; r_1 \circ r_2)$. $rec(t; r_1 \circ r_2, mod)$ and $shift(t; r_1 \circ r_2)$ have been Z-scaled.

Deep Neural Models of Semantic Shift [Rosenfeld & Erk 2018]

Плюсы

+ бóльшая реалистичность: если рассматривать время как непрерывную переменную, появляется возможность фиксировать постепенный характер лексических изменений
+ позволяет более точно и полно представлять причины, лежащие в основе того или иного семантического сдвига
+ можно использовать, чтобы предсказать скорость изменения значения того или иного слова
+ показывает лучшие результаты в сравнении с некоторыми другими моделями

Минусы

- метод оценки требует дальнейшей разработки и улучшения
- отсутствие гиперпараметров нейронной сети → необходимость настраивать её самостоятельно (при отсутствии опыта)

Evaluation

Intrinsic evaluation — методы оценки, при которых дистрибутивные репрезентации слов сравниваются с эмпирически полученными оценками людей и делается вывод об их схожести:

- *семантическая близость слов (word semantic similarity, word similarity)*: существующие наборы данных для оценки: SimVerb-3500, MEN, RW; один из самых популярных и самых старых; проблема вариативности оценок
- *аналогии слов (word analogy, linguistic regularities, word coherence)*: метрики: 3CosAdd (3CosMul), PairDir, Analogy Space Evaluation; размеченные наборы данных: BATS, Google Analogy, WordRep; проблема — отсутствие единой метрики для оценки отношений между словами в векторном пространстве
- *семантическая встраиваемость (thematic fit)*: наборы данных для оценки: MSTNN, GDS, F-Inst & F-Loc; проблема — нужен корпус текстов с разметкой частей речи и семантических ролей существительных, неоднозначность
- *категоризация слов (word categorization)*: определение качестве кластеризации векторов слов — проблема в выборе подходящего способа кластеризации, адекватной метрики оценки

Evaluation

- *определение синонимов (synonym detection)*: преимущество перед методом семантической близости — относительность оценок; наборы размеченных данных: TOEFL Synonym Questions, ESL Synonym Questions

- *оценка в себе (evaluation through cross-match test)*: сравнивать можно с любыми моделями, которые определяются как ‘качественные’, ‘хорошие’

- *семантическое различие слов (semantic difference)*: в основном предполагается, что рассматриваемые слова -- конкретные существительные; наборы данных — базы данных, в которых хранятся слова вместе с атрибутами (BLESS, Feature Norms Dataset)

- *графематический анализ (evaluation through the form of a linguistic sign)*: идея фоносемантических паттернов была подтверждена не только для латиницы, но и для кириллицы; открытых наборов данных нет

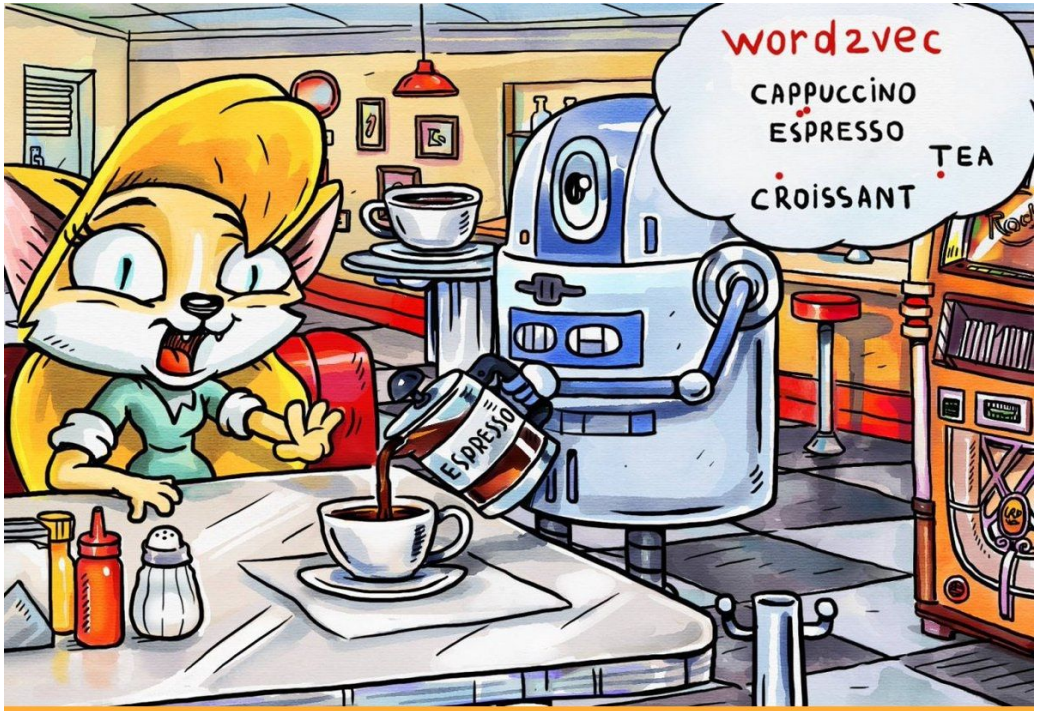
Проблема: наборы данных для оценки разработаны для английского языка

Summary

Модели:
простые vs. сложные, с исходным кодом и без

Идеальная модель:
1) сложная с исходным кодом / такая простая, что можно самому сделать и модифицировать
2) устойчивая к уменьшению датасета и большой гранулярности

Evaluation:
Количественный + качественный



— Эспрессо?! Но я же заказывала капучино!

— Не переживайте. Косинусное расстояние между ними насколько мало, что это почти что одно и тоже.

Спасибо за внимание!