Исследование и разработка моделей векторного представления слов

Кемаев Юрий, 441 гр.

Научный руководитель: к.ф-м.н, доцент Турдаков Д. Ю.

ИСП РАН, 2017г.

Векторные представления слов

Определение

Векторные представления слов — распределенные векторы многомерного пространства, полученные в результате взаимно-однозначного отображения из лексикона естественного языка и отражающие семантические и синтаксические особенности слов.

Модель векторного представления слов — совокупность предпосылок и знаний, на основе которых осуществляется такое отображение.

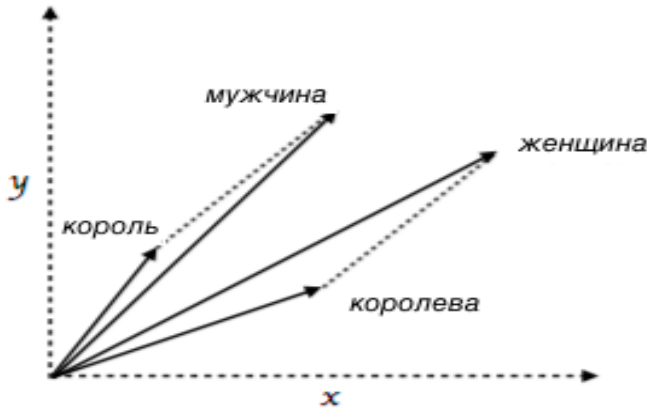
Ключевое свойство

Вектора аналогичных понятий обладают следующим свойством:

$$v_{\text{король}} - v_{\text{мужчина}} \approx v_{\text{королева}} - v_{\text{женщина}}$$

Откуда, зная три слова из такого отношения, можно алгебраически получить четвертое:

$$v_{\text{король}} = \underset{w \in \text{Словарь}}{\operatorname{argmin}} \; ||v_w - (v_{\text{королева}} - v_{\text{женщина}} + v_{\text{мужчина}})||$$



Зависимость между парами однородных аналогий

Зачем они нужны?

Представления выступают в роли ***признаков машинного обучения***, позволяя качественно решить такие задачи, как:

1. извлечение информации из текстов
2. распознавание речи
3. анализ тональности
4. тематическое моделирование
5. автоматическое реферирование
6. фильтрация контента
7. машинный перевод
8. генерирование текста
9. синтез речи
10. ...

Такие задачи возникают в области **информационного поиска** — процесса поиска в большой коллекции некоего неструктурированного материала, удовлетворяющего информационные потребности.

Цель работы

Цель

Исследование существующих и разработка новых моделей векторных представлений для эффективной и качественной векторизации слов русского языка

Этапы работы:

1. **Анализ существующих моделей** с целью выявления наиболее эффективных и применимых на практике
2. **Разработка и реализация новых моделей**
3. **Выбор данных и метрик для тестирования**
4. **Сравнительный анализ** разработанных моделей и существующих

Этапы работы

1. **Анализ существующих моделей**
2. Разработка и реализация новых моделей
3. Выбор данных и метрик для тестирования
4. Сравнительный анализ

Основа моделей векторизации слов

Дистрибутивная гипотеза

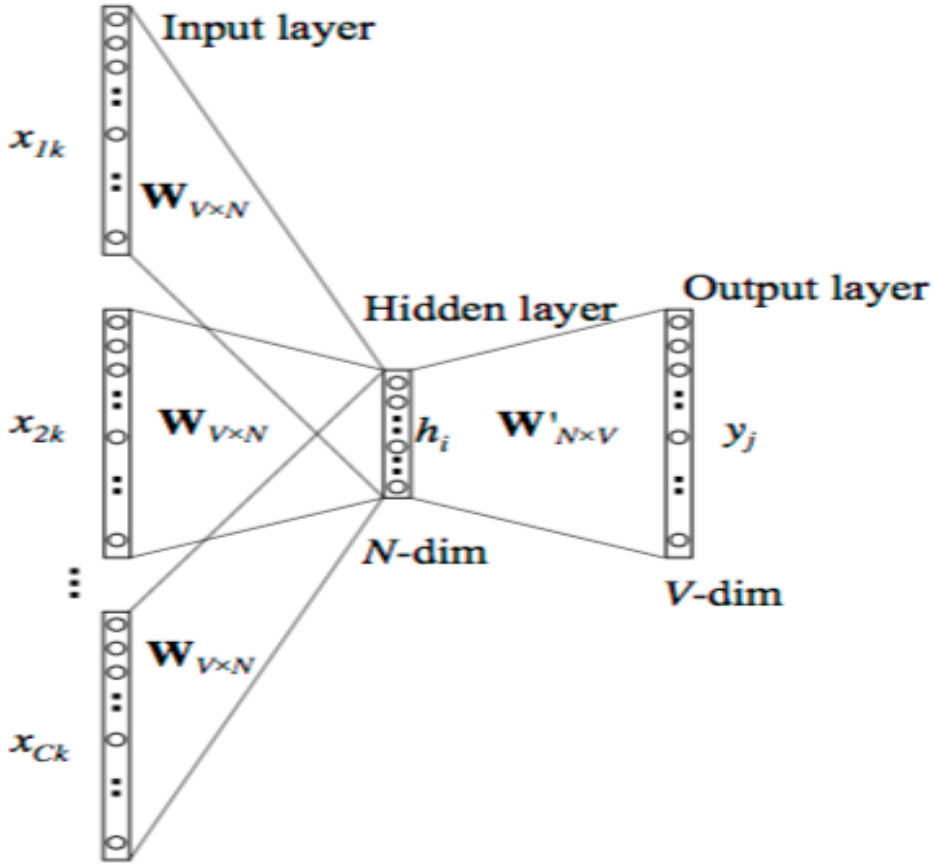
Лингвистические единицы, встречающиеся в схожих контекстах, имеют близкие значения

Harris, Z. (1954) Distributional structure

Word2Vec Continuous Bag-of-Words (2013)

Цель

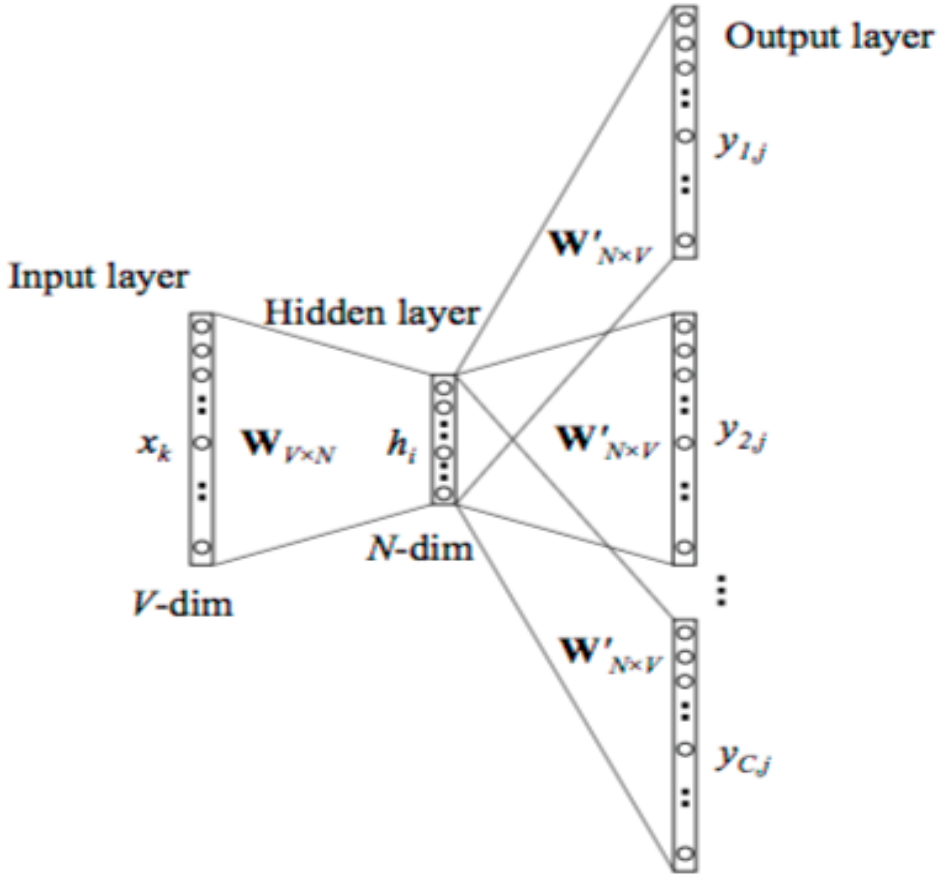
Модель *Continuous Bag-of-Words* предсказывает **слово** w_t **по его контексту** $w_{t-k}, \ldots, w_{t-1}, w_{t+1}, \ldots, w_{t+k}$ (Mikolov et. al, 2013)



Word2Vec Skipgram (2013)

Цель

Модель *Skipgram* предсказывает **контекст** $w_{t-k}, \ldots, w_{t-1}, w_{t+1}, \ldots, w_{t+k}$ **по слову** w_t (Mikolov et. al, 2013)



Недостатки существующих моделей

Основные недостатки

- Существующие модели слабо учитывают морфологию языка (**критично для русского языка!**)
- **. . . и никак не используют накопленные априорные знания о нем в процессе обучения**

FastText (2016)

Цель

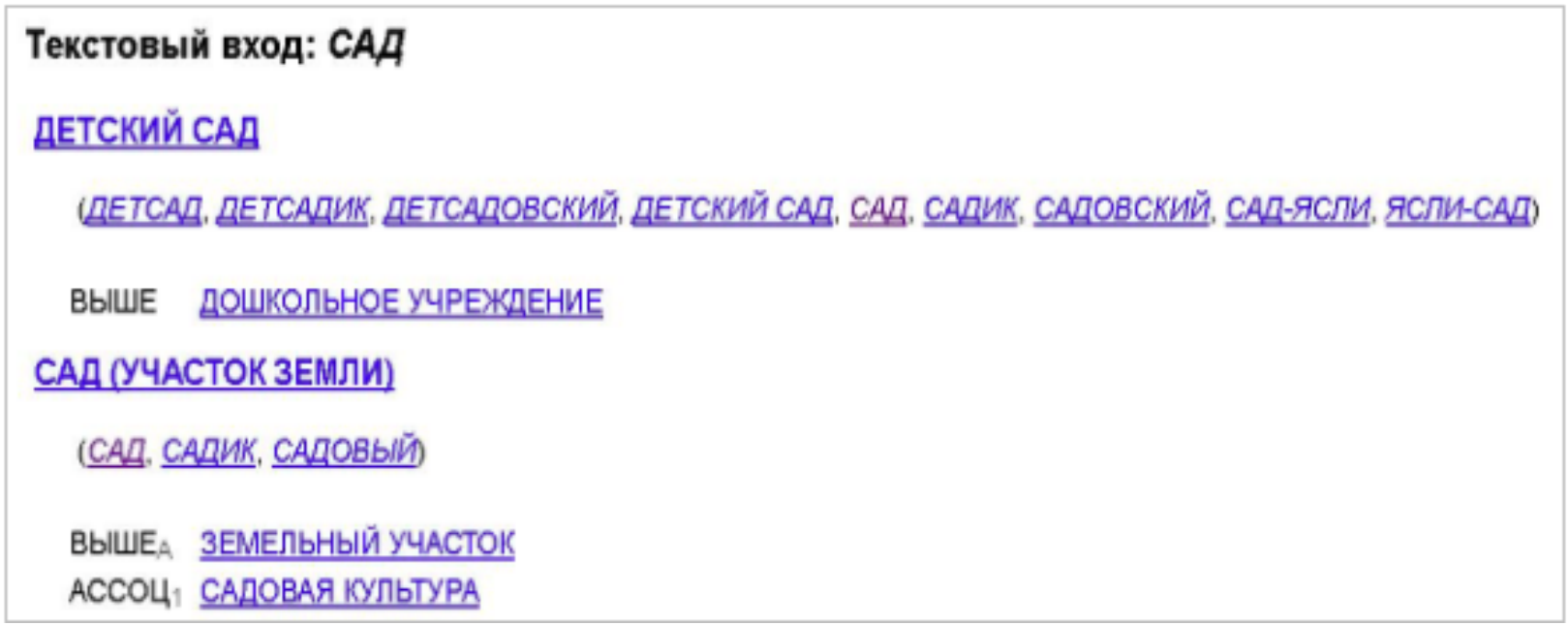
Модель *FastText* предсказывает **контекст** $w_{t-k}, \ldots, w_{t-1}, w_{t+1}, \ldots, w_{t+k}$ **по слову** w_t и его n-граммам (Joulin et. al, 2016)

Ключевая идея

Неявное использование морфологических особенностей языка путем представления слова как среднего от векторов своих n-грамм

Краткое описание

РуТез-2.0 — тезаурус для русского языка, содержит ≈ 35 тыс. понятий, связанных 4-мя типами отношений

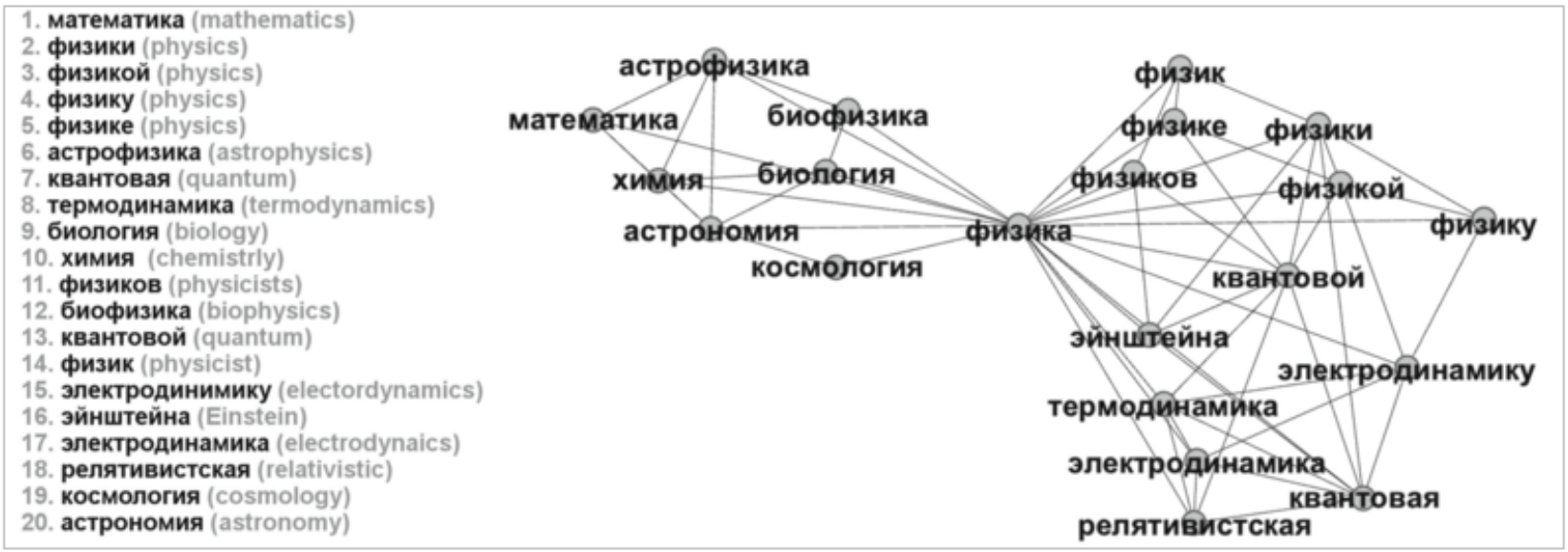


Информация о слове “сад”

Russian Distributional Thesaurus

Краткое описание

Russian Distributional Thesaurus (RDT) — тезаурус для русского языка, содержит ≈ 932 тыс. понятий и граф их подобия



Ближайшие соседи слова “физика”

Этапы работы

1. Анализ существующих моделей
2. **Разработка и реализация новых моделей**
3. Выбор данных и метрик для тестирования
4. Сравнительный анализ

Лингвистические характеристики слов языка

Каждое слово обладает множеством свойств в разных областях языкознания (в синтаксическом, семантическом и др. срезах).
Примеры областей и свойств:

- морфемный состав — свойствами слова являются его морфемы
- граф аналогий языка — свойствами слова являются его связи с другими словами
- n-граммый состав — свойствами словя являются его n-граммы

Представление слова

Каждому слову из лексикона ставится в соответствие

$$F^w = \{(i,j) \mid i = 1,\ \ldots,\ k;\ j = 1,\ \ldots,\ S_i;\ p_j^i(w) = 1\}$$

— набор индексов свойств из каждой области языкознания.

Ключевая идея

Переход от **представлений слов** к **представлениям свойств**

Представление слова

Теперь слово w представляется как усреднение по векторам **всех своих свойств**:

$$v(w) = \frac{1}{|F^w|} \sum_{(i,j) \in F^w} v(i,j)$$

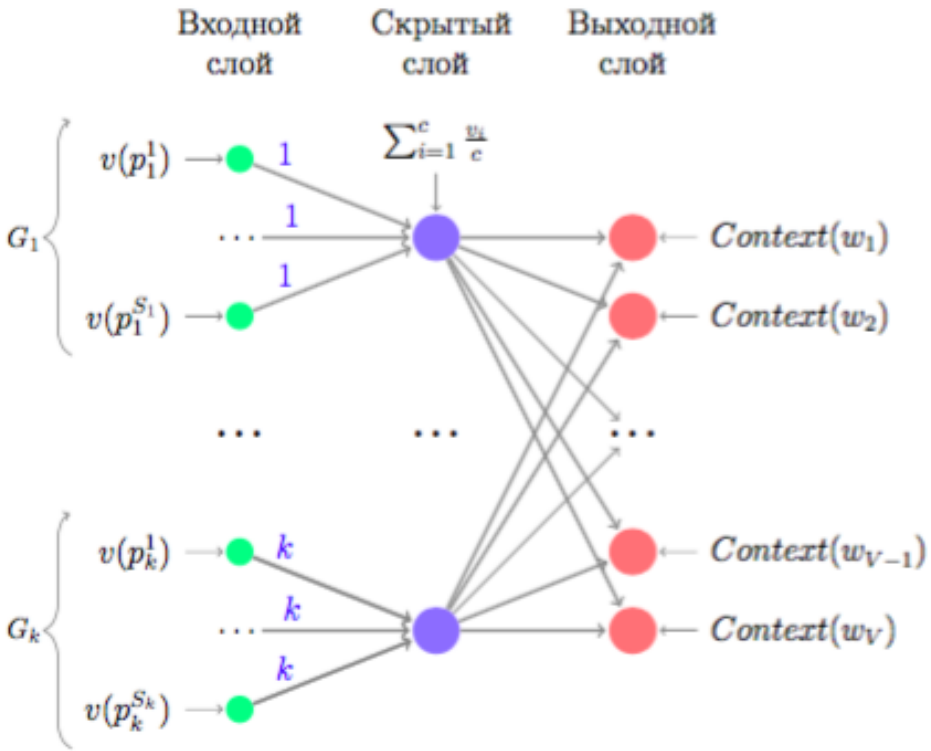
Модификации архитектуры нейронной сети

Выделенное семейство моделей позволяет ввести некоторые **модификации архитектуры** оригинальной сети **для максимально эффективного использования** информации, предоставляемой каждой областью языкознания.

Semi-boosting сеть

Ключевая идея

Обучить k моделей, каждая из которых исправляет коллективную ошибку предшествующих (по аналогии с *градиентным бустингом*).

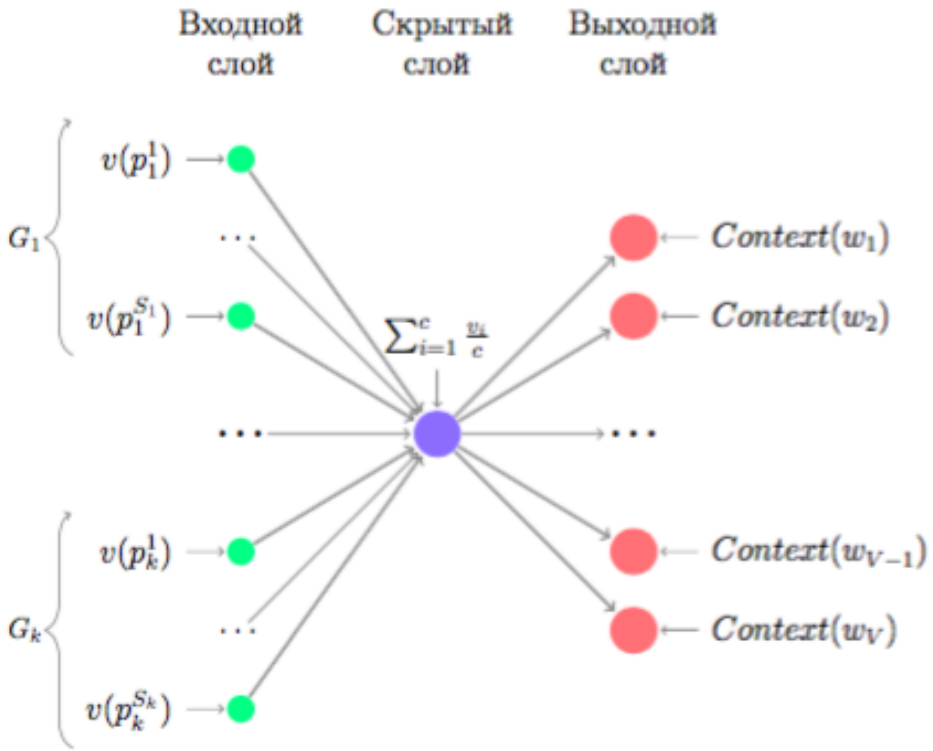


Архитектура semi-boosting сети

Полносвязная сеть с dropout

Ключевая идея

Получить усреднение параметров нейронной сети по всевозможным 2^N архитектурам, где N — количество нейронов входного слоя в сети.



Архитектура полносвязной сети с dropout

Реализация

Особенности реализации

Реализация моделей имеет обладает следующими основными свойствами:

- реализация выполнена на языке *C++*
- архитектура модели выбирается пользователем
- параметры модели задаются пользователем
- обучение происходит параллельно на нитях (threads)
- *вертикальная* и *горизонтальная* масштабируемость
- расширяемость — можно подключать любые предметные области (требуется соблюсти формат)

Этапы работы

1. Анализ существующих моделей
2. Разработка и реализация новых моделей
3. **Выбор данных и метрик для тестирования**
4. Сравнительный анализ

Оценка качества

Качество выходных представлений моделей тестировалось с использованием комплекта задач и данных **RUSSE**, созданного специально с целью оценки оценки векторов слов русского языка.

Данное решения было принято по следующим причинам:

1. данный комплект позволяет всесторонне оценить качество векторов, так как включает в себя несколько тестовых задач (6 штук)
2. в комплекте представлены почти все существующие на данный момент задачи для русского языка
3. в свободном доступе имеется обертка, написанная на *Python*

RUSSE, данные

RUSSE содержит 6 наборов данных для оценки:

1. русифицированный набор **WordSim353 Rel** — 250 кортежей
2. русифицированный набор **WordSim353 Sim** — 202 кортежа
3. **RusseHJ** — набор понятий и оценок, выставленных людьми, — 333 кортежа
4. датасет с кортежами связанных отношениями разной природы слов — 9549 кортежей
5. два датасета с когнитивными ассоциациями по 1952 и 3003 кортежа соответственно

Данные для обучения и предобработка

Входными данными для обучения были взяты ≈ 12 **млн сообщений** из различных открытых сообществ социальной сети “Вконтакте”.

Процесс предобработки состоит из следующих шагов:

1. Замена знаков препинания на пробелы
2. Приведение слов в нижний регистр
3. Замена всех числительных на “1”
4. Лемматизация с помощью морфологического анализатора **Yandex Mystem 3**
5. Создание словаря, содержащего **100 тыс. наиболее частотных уникальных понятий**
6. Фильтрация слов корпуса, которые не вошли в сформированный словарь (размер корпуса уменьшился на 1.3%)

Области языкознания

Использовались 4 области знаний:

1. n-граммы слов, полученные с помощью морфологического анализатора **pymorphy2** (*понадобиться: понадоб, понадоби, понадо, надо, онадоби, ...*)
2. морфемы слов, полученные эвристически (*разглашать: раз-глаш-ать*)
3. аналогии, полученные из тезауруса **Russian distributional thesaurus** (*производная: производный, константа, функция, компонента, ...*)
4. синонимы, полученные из тезауруса **РуТез-2.0** (*публикование: обнародование, опубликовывать, помещать, печататься, ...*)

Параметры тестируемых моделей

Общие для всех моделей параметры были взяты по умолчанию (**предложенные авторами базовых моделей**), а именно:

- размерность выходных представлений – 300
- размер скользящего окна – 5
- количество шумных слов для данного – 5
- длина n-грамм – от 3 до 6
- learning rate – 0.025
- количество эпох – 1
- линейное подавление learning rate через каждые 100 шагов скользящего окна

Каждая модель из разработанных использует **не более 10 своих самых частотных свойств** из каждой предметной области.

Этапы работы

1. Анализ существующих моделей
2. Разработка и реализация новых моделей
3. Выбор данных и метрик для тестирования
4. **Сравнительный анализ**

Результат, часть 1

Модель	*WS353 rel*	*WS353 sim*	*HumJudge*
CBoW (2013)	0.5951	0.6869	0.6782
SGNS (2013)	0.6327	0.7249	0.6967
FastText (2016)	0.6021	0.7264	0.6896
fully-conn.	0.6344	0.7607	0.7237
f-c (no t.)	0.6359	0.7557	0.7240
f-c + dropout	0.5972	0.7154	0.7175
f-c + dropout (no t.)	0.5809	0.6362	0.6832
semi-boosting	**0.6594**	**0.7855**	0.7642
semi-boosting (no t.)	0.6538	0.7196	**0.7704**
mean-prior.	0.5907	0.6695	0.6632
mean-prior. (no t.)	0.5847	0.6564	0.6248

Таблица: Оценка (Spearman's correlation с экспертной разметкой) выходных векторных представлений слов

Результат, часть 2

Модель	*RuTes sem*	*AssocThes*	*AssocOnline*
CBoW (2013)	0.7120	0.5184	0.8554
SGNS (2013)	0.7241	0.5256	0.8581
FastText (2016)	0.7130	0.5287	0.8434
fully-conn.	0.7315	0.5315	0.8604
f-c (no t.)	0.7240	0.5255	0.8584
f-c + dropout	0.7160	0.5183	0.8481
f-c + dropout (no t.)	0.6718	0.5134	0.8215
semi-boosting	**0.7405**	**0.5420**	**0.8701**
semi-boosting (no t.)	0.7260	0.5400	0.8634
mean-prior.	0.7009	0.5010	0.8315
mean-prior. (no t.)	0.6843	0.5051	0.8241

Таблица: Оценка (accuracy) выходных векторных представлений слов

Выводы

В рамках настоящей выпускной квалификационной работы были получены следующие результаты:

1. Проведен анализ существующих методов построения векторных представлений слов, составлен их подробный обзор
2. Введено семейство моделей векторного представления слов, обобщающее идеи существующих (моделей).
3. Разработаны и реализованы несколько методов векторизации, показывающих одни из лучших результатов в задачах обработки русского языка

Спасибо за внимание