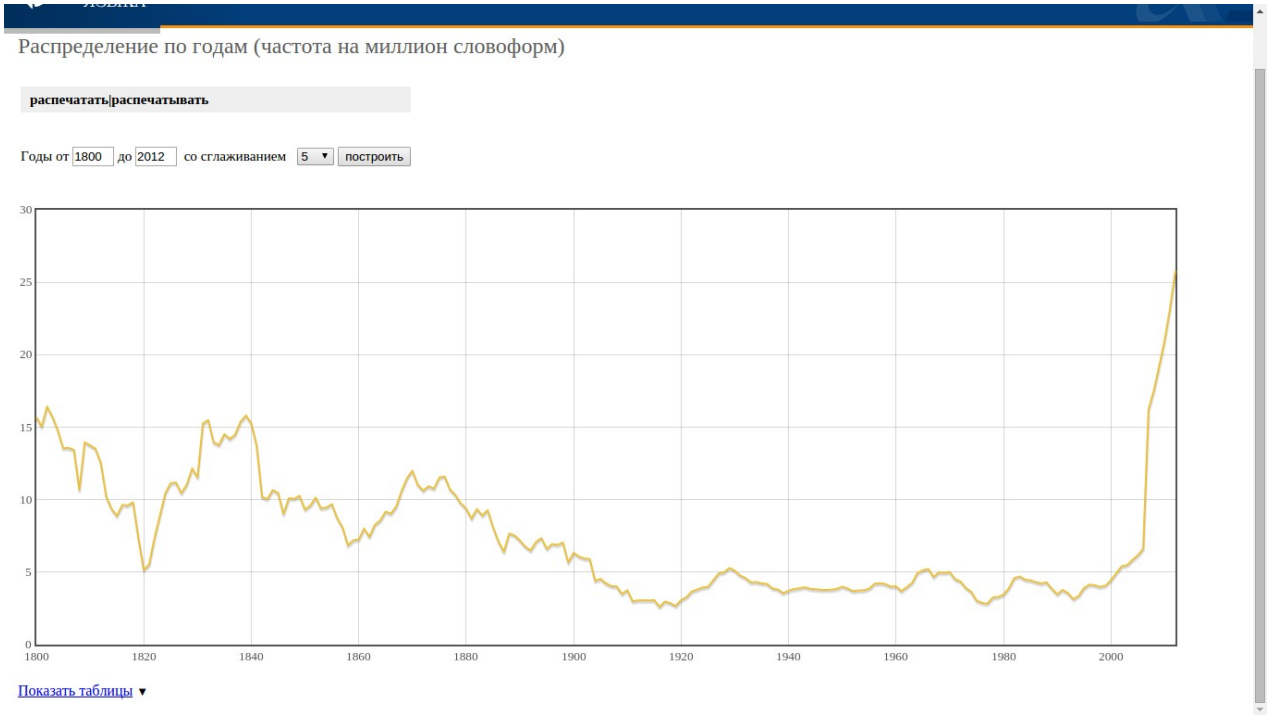
# Диахронический анализ значений слов с помощью дистрибутивно-семантических моделей

## Постановка задачи

Язык постоянно меняется, и одной из областью исследований в лингвистике является анализ этого процесса. В частотности, с помощью современных корпусных инструментов лингвист может проследить, например, изменения в использовании той или иной грамматической конструкции, проанализировав её частотные характеристики в разные временные периоды. Также с помощью корпуса можно понять, какие слова (лексемы) вышли из употребления или, наоборот, в определенный момент времени “вошли” в язык. Но лингвистам часто хотелось бы увидеть, как в диахронии изменяются значения слов, а не сами слова. В некоторых случаях лингвист может посмотреть на график частотного распределения слова по годам (в качестве примера привожу скриншот с сайта НКРЯ -- график для глагола *распечатать*), заметить резкое изменение, посмотреть соответствующие контексты в корпусе и убедиться, что резкое изменение частот связано с приобретением нового значения (для *распечатывать* это ‘печатать на принтере’).




Однако такой анализ ограничен тем, что при выборе слов для анализа лингвист основывается лишь на своей интуиции; тем самым

он вряд ли сможет охватить все слова, чьи значения подверглись заметному изменению.

Цель проекта заключается в автоматическом выявлении слов, которые с течением времени приобрели новое значение. Для этого строятся диструбутивно-семантические модели, обученные на корпусах разных временных периодов, далее для каждого слова или определенной группы слов сравниваются семантически близкие слова каждого из временных периодов (слова с косинусным расстоянием, близким к единице) [Kim *et al* 2014, Kulkarni *et al* 2015, Kutuzov & Kuzmenko 2016].

Кроме новых значений (в лингвистическом понимании этого термина), данный способ позволяет выявлять слова, характеризующиеся заметным изменением контекста употребления, которое связано с социально-культурными переменами в обществе. Так, в Gulordava & Baroni 2011 приводится пример контекстно-близких слов для слова *sleep* в 60-х годах и в 90-х: *deep, long, dreemless, much, cannot* и *disorders, deprivation, deep, disturbance, apnea* соответственно (очевидно, что само значение слова *sleep* не изменилось).

В рамках данного проекта я ограничилась анализом изменения значений слов в период с 2011 по 2014 г. на материале газетного подкорпуса НКРЯ.

**Данные**

В качестве обучающего корпуса я использовала часть газетного подкорпуса НКРЯ, а именно новостные статьи РИА-Новости, РБК, Известия, Комсомольская Правда.
Из-за ограниченных вычислительных возможностей и ошибочных первоначальных обучений в конечном итоге получилось две модели: первая модель была обучена на корпусе новостных статей, вышедших в 2011 году (6113257 токенов), вторая модель представляет собой первую модель, дообученную на корпусе новостных статей, вышедших в 2014 году (6247923 токенов).

Перед обучением модели все документы я разбила на предложения, каждое предложение токенизировала и затем лемматизировала с помощью Pymorphy2. Кроме этого, в процессе предобработки заменила все числа на ‘num’, удалила остатки html-разметки и стоп-слова. В качестве списка стоп-слов (802 леммы) использовался частотный список лемм служебных частей речи и частотный список

местоимений, взятых из Нового частотного словаря русского лексики [Ляшевская & Шаров 2009].

**Методы**
Для обучения word2vec модели использовалась библиотека gensim.

Алгоритм работы был следующим:

(1) Обучить word2vec модель на корпусе новостных статей 2011 года;
(2) Дообучить получившуюся word2vec модель на корпусе новостных статей 2014 года, предварительно расширив словарь первой модели с помощью корпуса новостных статей 2014 года (идею, как это сделать, я взяла из обсуждения в гугл группе gensim https://groups.google.com/forum/#!topic/gensim/CBPl4aXN7Ao);
(3) Составить выборку слов для анализа значений; для этого используется список слов (432 леммы) из словаря значимой газетно-новостной лексики [Ляшевская & Шаров 2009], такой что каждый его элемент содержится в словаре обеих моделей;
(4) Используя списки из 10 семантических близких слов первой и второй модели для каждого элемента описанной выше выборки, вычислить коэффициент Кендалла (если коэффициент близок к единице, то это значит, что списки семантически близких слов второй модели незначительно изменились по сравнению со списками семантически близких слов первой модели, если коэффициент близок к нулю, то -- обратное); выбор в пользу данного коэффициента был сделан на основании результатов, приведенных в [Kutuzov & Kuzmenko 2016].

Эксперименты на корпусе новостных статей 2011 года показали, что наилучшим образом интерпретируемые результаты для задачи поиска семантически близких слов дает алгоритм skip-gram с одновременным использованием негативного сэмплирования (15 слов) и  иерархического софтмакса. Остальные параметры были следующими: размерность векторов -- 300, размер окна -- 10, минимальная частота слов -- 5, суб-сэмплирование не использовалось. Конечно, очень спорно, что модель с такими параметрами наиболее адекватна, поскольку качество модели довольно сложно оценить, и в той или иной степени оно зависит от конечной цели. К сожалению, в оценке адекватности модели (оценивая списки 10 семантически близких слов для первых 10 существительных словаря значимой газетно-новостной лексики:

*президент, театр, год, спектакль, правительство, компания, страна, фильм, реформа, выборы*) я полагалась лишь на свою интуицию.

Однако мне не удалось дождаться дообучения модели, обученной на таких параметрах, так как модель корректно дообучается, только когда дообучение происходит без использования библиотеки Cython . В связи с этим первую модель пришлось обучать со следующими параметрами: алгоритм skip-gram, негативное сэмплирование -- 5 слов, размерность векторов -- 300, размер окна -- 10, минимальная частота слов -- 5, суб-сэмплирование не использовалось (алгоритм CBOW работает значительно быстрее, но и дает менее осмысленные результаты на небольшом корпусе); такая модель дообучилась за 8ч 49мин.

**Результаты**

При подсчете корреляции выборка слов сократилась с 432 до 241, поскольку в 191 одном случае оказалось, что списки ассоциатов имеют только по одному общему слову.

Медиана значений коэффициента Кендалла -- 0.63, среднее -- 0,62 (при подсчете каждого коэффициента брала его абсолютно значение, об этом см. ниже). Если считать, что в выше описанном 191 случае, значение коэффициента равно 0, то получаем, что медиана равна 0.19, среднее -- 0.34. Это говорит о том, что контекст большинства слов заметно изменился.

Коэффициент Кендалла, как и другие статистические тесты, имеет значение в диапазоне от -1 до 1 (при отрицательной или положительной корреляции). В данной работе для случаев, когда ассоциаты слова одной модели хорошо коррелируют с ассоциатами другой, я пренебрегаю знаком коэффициента коррелиции (т.е. считаю значения коэффициента близкие к -1 и 1 равнозначными), поскольку в таких случаях точно не наблюдается изменения значения или контекста. То же верно и для значений коэффициента близких к 0, с той лишь разницей, что как раз эти случаи и представляют особый интерес. Поэтому использовала абсолютные значение коэффициента (если этого не делать, медиана -- 0.33, среднее -- 0.22 (с учетом 191 случая медиана -- 0, среднее -- 0.12), что вряд ли о чем-то говорит).

Что касается списка из 191 слова (т.е. слов, ассоциаты которых не имеют общих элементов), то для всех них характерно изменение именно контекста употребления, а не появление нового значения (а

на таком небольшом корпусе, как используемый в работе, изменение контекста не носит “глобальный” характер, не отражает изменений в обществе, оно обуславливается доминированием текстов определенной тематики). Приведём пример одного из таких слов:

10 семантически близких слов первой и второй модели (левый и правый столбец соответственно) для слова *номинация*:

| | |
|---|---|
| *император* | *лауреат* |
| *теннисистка* | *кинотавр* |
| *сотникова* | *оскар* |
| *юниор* | *каннский* |
| *фигуристка* | *удостоить* |
| *дебютировать* | *приз* |
| *репетиция* | *номинант* |
| *мисс* | *евровидение* |
| *леди* | *кинофестиваль* |
| *дэвис* | *полнометражный* |

При анализе результатов, стало понятно, что коэффициент Кендалла для данной задачи далеко не всегда показателен. Приведем сначала пример, в котором коэффициент, как кажется, соответствует действительности, а затем рассмотрим сложности, связанные с использованием данного коэффициента.

10 семантически близких слов первой и второй модели (левый и правый столбец соответственно) для слова *авиакомпания,* коэффициент Кендалла равен 0.42:

| | |
|---|---|
| *перевозчик* | *перевозчик* |
| *лицензия* | *аэрофлот* |
| *сертификат* | *авиаперевозчик* |
| *рейс* | *трансаэро* |
| *росавиация* | *рейс* |
| *аэропорт* | *чартерный* |
| *трансаэро* | *аэропорт* |
| *континент* | *таможня* |
| *авиановый* | *авиабилет* |
| *чартерный* | *ютэйр* |

Общие ассоциаты: *перевозчик, рейс, аэропорт, трансаэро, чартерный*

Во-первых, коэффициент Кендалла некорректно применять для слов, ассоциаты которых имеют по два общих элемента -- значение всегда будет либо -1, либо 1.

Во-вторых, там, где справедливо ожидать высокое значение коэффициента, имеется близкое к 0, и наоборот. Это связано с тем, что при подсчете коэффициента Кендалла учитывается только порядок следования рангов, а не их величина (если так можно сказать). Рассмотрим два слова (ассоциаты обоих имеют по три общих элемента): *фонд* и *инвестор*. Для первого слова получаем следующие последовательности рангов для ассоциатов первой и второй модели соответственно: [1, 2, 4] и [1, 3, 2], для второго -- [4, 8, 9] и [1, 2, 8]. Коэффициент Кендалла для первого слова 0.33, для второго 0.99. Что, как можно заметить, совсем не то, что хотелось бы видеть -- очевидно, что контекст для слова *инвестор* в большей степени изменился, по сравнению с контекстом слова *фонд*.

10 семантически близких слов первой и второй модели (левый и правый столбец соответственно) для слова *фонд,* коэффициент Кендалла равен 0.33:

| | |
|---|---|
| *нпф* | *нпф* |
| *негосударственный* | *пенсионный* |
| *вклад* | *негосударственный* |
| *пенсионный* | *вэб* |
| *грант* | *накопление* |
| *страхование* | *благосостояние* |
| *внебюджетный* | *паевой* |
| *благотворительный* | *венчурный* |
| *некоммерческий* | *пожертвование* |
| *пфр* | *офшор* |

Общие ассоциаты: *нпф, негосударственный, пенсионный*

10 семантически близких слов первой и второй модели (левый и правый столбец соответственно) для слова *инвестор,* коэффициент Кендалла равен 0.99:

| | |
|---|---|
| *инструмент* | *вложение* |
| *актив* | *инвестиция* |
| *вкладывать* | *инвестиционный* |
| *вложение* | *госбанк* |
| *облигация* | *эмитент* |
| *держатель* | *вложиться* |
| *баланс* | *инвестировать* |
| *инвестиция* | *вкладываться* |
| *вкладываться* | *покупатель* |
| *институциональный* | *скупка* |

Общие ассоциаты: *вложение инвестиция вкладываться*

В-третьих, когда ассоциаты слова имеют довольно много общих слов (7 и более из 10), логично предположить, что семантического сдвига не произошло. Однако коэффициент Кендалла в некоторых случаях говорит об обратном -- он имеет близкое к 0 значение. Это объясняется тем, что порядок следования общих ассоциатов в одной модели не коррелирует с порядком следования общих ассоциатов в другой. Так, для *избиратель*, ассоциаты которого имеют 7 общих элементов, коэффициент равен 0.048 (порядок следования рангов общих ассоциатов:  [1, 2, 3, 4, 5, 6, 7] и [3, 6, 1, 5, 7, 2, 4] для первой и второй модели соответственно).

10 семантически близких слов первой и второй модели (левый и правый столбец соответственно) для слова *избиратель,* коэффициент Кендалла равен 0.048:

| | |
|---|---|
| *проголосовать* | *явка* |
| *выборы* | *голосовать* |
| *явка* | *проголосовать* |
| *праймериз* | *мандат* |
| *голос* | *праймериз* |
| *голосовать* | *выборы* |
| *мандат* | *голос* |
| *бюллетень* | *электорат* |
| *кандидат* | *выдвижение* |
| *единоросс* | *голосование* |

Общие ассоциаты: *проголосовать, выборы, явка, праймериз, голос, голосовать, мандат.*

В силу описанных недостатков коэффициента Кендалла, для оценки сходства ассоциатов двух моделей был применен коэффициент Жаккара. Здесь нет необходимости исключать слова, ассоциаты которых не имеют общих слов (коэфициент для них просто будет равен 0), поэтому можно использовать полную выборку слов. Подсчет показал, что теперь медиана равна 0.11, среднее -- 0.13 (ранее 0.19 и 0.34 соответственно). Теперь с бОльшей уверенностью можно утверждать, что большинство слов выборки изменило контекст употребления.

Что касается значений коэффициента Жаккара для рассмотренных выше “проблемных” примеров, то теперь *инвестор* и *фонд* закономерно имеют одинаковое значение 0.174, а *избиратель* -- более реалистичное 0.54. Таким образом, коэффициент Жаккара нивелирует разницу в порядке следования общих ассоциатов, однако

он не учитывает близость общих ассоциатов к их исходному слову (совпадение первых и последних ассоциатов даст то же значение, что и совпадение первых в обоих моделях), что иллюстрирует пример с *инвестором* и *фондом.*

**Выводы**

1) Полученные результаты не соответствуют ожиданиям. Алгоритм выявляет слова, изменившие в первую очередь контекст употребления, нежели значение. Такие результаты объясняются скорее некорректным подбором обучающих данных -- размер корпусов, временные периоды, тип текстов, чем выбранным методом (способ измерения сходства ассоциатов однако нуждается в существенной доработке). Кажется, что то, что не удалась обучить модель с желаемыми параметрами, не так повлияло на результат, как выбор обучающих данных.
2) Относительно обучающих данных стоит также заметить, что стемминг и рассмотрение именованных сущностей как целостных единиц может повысить качество моделей.

**Литература**

Ляшевская, О. Н., Шаров, С. А. 2009. Частотный словарь современного русского языка (на материалах Национального корпуса русского языка). М.: Азбуковник, 2009.

Gulordava, K., Baroni, M. 2011. A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus. Proceedings of the EMNLP 2011 Geometrical Models for Natural Language Semantics (GEMS 2011) Workshop, East Stroudsburg PA: ACL, 67-71.

Kim, Y., Chiu, Y., Hanaki, K., Hegde, D., Petrov, S. 2014. Temporal analysis of language through neural language models. In Proceedings of the 52nd Annual Meeting of the Asso-
ciation for Computational Linguistics, Baltimore, USA.

Kulkarni, V., Al-Rfou, R., Perozzi, B., Skiena, S. 2015. Statistically significant detection of linguistic change. Proceedings of the 24th International Conference on World Wide Web, 625-635.

Kutuzov A., Kuzmenko, E. 2016. Cross-lingual Trends Detection for Named Entities in News Texts with Dynamic Neural Embedding Models.