Построение векторной семантической модели на основе русскоязычных текстов: первые эксперименты

Владислав Порицкий
БГУ, Минск
v.poritski@gmail.com

Оксана Волчек
БГУ, Минск
volchekoa@gmail.com

Аннотация

Векторные модели семантики языковых единиц (vector space models, VSM) известны компьютерным лингвистам уже более двух десятков лет, однако до настоящего времени их принято строить и оценивать преимущественно на материале английского языка. Мы обучаем векторные модели на двух небольших корпусах русскоязычных текстов (художественных и газетных) и оцениваем качество их предсказаний в задаче выбора пар синонимов. Лучший результат $F = 0.76$ был достигнут, когда семантические расстояния измерялись косинусом с PMI-взвешиванием при единичной ширине контекстного окна. Мера Йенсена-Шеннона, с помощью которой удалось добиться $F = 0.71$, допускает более радикальное снижение размерности, но усложняет масштабирование объёма корпуса.

1. Введение

Векторные модели семантики языковых единиц (VSM) известны компьютерным лингвистам уже более двух десятков лет, однако до настоящего времени их принято строить и оценивать преимущественно на материале английского языка. Мы попытались восполнить имеющийся пробел и построили VSM на двух корпусах русскоязычных текстов общим объёмом 17.4 млн словоупотреблений. Перед моделями ставилась задача разграничить синонимы и несинонимы. Разные комбинации параметров оценивались долей полученных совпадений с эталоном.

Эта статья сообщает о первых результатах наших экспериментов. В параграфе 2 даётся краткий обзор истории и теории VSM. Параграф 3 содержит описание нашей модели и процедуры её оценивания. В параграфе 4 показано, как качество предсказаний зависит от используемой метрики, выбираемого подмножества признаков и ширины контекстного окна. Статья завершается обсуждением итогов и перспектив исследования.

2. Из истории дистрибутивных семантических моделей

Мысль о замене референциальной семантики языкового знака на дистрибутивную связывают с именем Л. Витгенштейна, который высказал её в 1940-е гг., развивая концепцию «языковых игр»: «Для большого класса случаев... значение слова есть его употребление в языке» [1]. В те же годы и лингвисты структурального направления поняли, что о грамматических и семантических свойствах языковой единицы (морфемы или лексемы) можно многое узнать из её дистрибуции, не привлекая никаких дополнительных сведений. Эта идея, представленная в работах З. Харриса [13] и Дж. Р. Фёрса [11], стала известна как *дистрибутивная гипотеза*. В рамках семантики её формулируют так: слова, которые встречаются в одинаковых контекстах, склонны иметь близкие лексические значения.

Первые практические опыты дистрибутивной группировки слов, родственных по смыслу (в частности, синонимов), относятся к 1960-м гг. На Западе этим занималась К. Спэрк-Джонс [24], а в СССР – В. А. Москович [4] и, несколько позже, Б. А. Плотников [5]. Поскольку их исследования не имели большого резонанса, дистрибутивной семантике суждено было пережить второе рождение в начале 1990-х гг. Именно тогда стало ясно, что на множестве лексических единиц можно задать метрику (или полуметрику, квазиметрику), измерив для каждой пары слов «семантическое расстояние» с опорой на корпусные данные. После работ [22], [18] и [12] в качестве общепринятого утвердился следующий способ: собрать в большом корпусе сведения о совместной встречаемости лексических единиц и охарактеризовать каждое слово w вектором частот разных слов в его контекстных окружениях заданной ширины. Затем частоты могут домножаться на веса, приписанные словам-признакам (элементам базиса) в соответствии с тем, насколько тесна их связь с w. Тогда величина семантической близости меж-

ду w и произвольным w' вычисляется как расстояние в какой-либо метрике между векторами, описывающими их. При оценивании моделей широко используются эталонные человеческие суждения о семантической близости, полученные психолингвистами [19, 17], а также словари-тезаурусы. Таким путём идут, среди прочих, Л. Ли [16], Дж. Уидс [26] и Дж. Каррен [9], в чьих работах исследовано поведение целого ряда весовых функций и метрик относительно эталонов.

Современное состояние дистрибутивной семантики отражено в обзоре [25]. Для векторных моделей уже решена проблема снижения размерности и достигнуты некоторые успехи в задаче разграничения разных типов семантических отношений. В последние годы активно идёт работа по интеграции дистрибутивной семантики с композиционной (compositional). Но вся эта деятельность разворачивается в основном на английском материале и в англоязычном научном пространстве. По-русски имеются только обзорная статья О. А. Митрофановой [3] и работа группы авторов из Киева [2] (на материале нескольких корпусов английского языка). Кроме того, в диссертации [7] методы, родственные VSM, используются для извлечения коллокаций из русскоязычных текстов.

Важно отметить, что векторные семантические модели выучиваются «без учителя». Поэтому применительно к ним нельзя говорить о «языконезависимости» (language independency) в том же смысле, в котором этот термин используют, говоря о статистических лемматизаторах и парсерах. Из любого корпуса текстов можно выучить векторную модель, но затем её придётся специфически параметризовать в зависимости от языка и от конкретного используемого эталона.

3. Наша модель

3.1. Исходные данные

Мы строили векторные семантические модели на основе двух корпусов, содержащих художественные и газетные тексты. Корпус художественных текстов объёмом 9.9 млн словоупотреблений включает три непересекающиеся коллекции: русская проза XX века и переводная проза XX века (по 2.5 млн словоупотреблений), русская проза 1990-2000-х гг. (около 5 млн). Корпус газетных текстов объёмом 7.5 млн словоупотреблений включает подшивки газет «Известия» за август 2008 и апрель 2009 г., «Комсомольская правда» за сентябрь и ноябрь 2008, январь 2009 г., «Культура» за промежуток с августа 2008 по июль 2009 г.

Оба корпуса были лемматизованы (ср. обсуждение в [20, p. 76]). Для этого использовалась связка из лемматизаторов MyStem [23] и TreeTagger с русским файлом параметров [21, 15]. Такой подход видится оправданным из двух соображений.

С одной стороны, MyStem не располагает механизмом дизамбигуации и возвращает для слов, содержащихся в словаре Зализняка, полные списки возможных лемм и грамматических разборов. Как правило, предлагаемые TreeTagger'ом лемма и разбор содержатся в этих списках (с поправкой на различную систему морфологических тегов), поэтому в базовом случае TreeTagger играет роль дизамбигуатора поверх выдачи MyStem'а.

С другой стороны, файл параметров для TreeTagger натренирован на подкорпусе НКРЯ со снятой грамматической омонимией объёмом 6 млн словоупотреблений, а этого относительно мало для полноценного лексического покрытия произвольного (open-domain) входного текста. Как следствие, TreeTagger склонен приписывать неправильные леммы формам слов, ни разу не встреченных в тренировочном массиве данных. MyStem имеет широкое лексическое покрытие и удачную эвристику нахождения лемм от незнакомых словоформ, что позволяет опираться только на его гипотезы при обработке исключительных случаев.

3.2. Расчёт

В базис векторной модели включались леммы, найденные связкой лемматизаторов. Таким образом, для каждой лексической единицы выучивался вектор частот различных лемм в её контекстных окружениях фиксированной ширины. Ширина контекстного окна бралась равной от 1 до 4, так что всего было выучено двенадцать моделей: по четыре для каждого корпуса и ещё четыре для объединённой коллекции текстов. Затем, по мере необходимости, векторы частот могли подвергаться PMI-взвешиванию, нормировке и лапласовскому сглаживанию.

PMI (pointwise mutual information, точечная взаимная информация) – это мера ассоциированности слов в текстах. Для леммы w и другой леммы w' из её контекстного окружения $\mathrm{PMI}(w, w') = \log_2 \frac{p(w, w')}{p(w)p(w')}$, где вероятности $p(w)$ и $p(w')$ понимаются как доли форм слов w и w' от общего объёма корпуса, а вероятность $p(w, w')$ – как доля контекстных окружений, содержащих оба слова, от общего числа контекстных окружений (это объём корпуса плюс небольшая константа, которой можно пренебречь). При PMI-взвешивании для каждой пары (w, w') число вхождений w' в контекстное окружение w заменяется на величину $\mathrm{PMI}(w, w')$.

Нормировка преобразует исходные векторы в векторы единичной длины, сохраняя направления. Пусть $v_w = (v_{w1}, \ldots, v_{wn})$ – n-мерный вектор,

представляющий лемму w. L_1-норма этого вектора $||v_w||_{L_1} = \sum_{i=1}^{n} v_{wi}$, а L_2-норма $||v_w||_{L_2} = \sqrt{\sum_{i=1}^{n} v_{wi}^2}$. При нормировке для каждой пары (w, w') число вхождений w' в контекстное окружение w делится на норму v_w; будет это L_1- или L_2-норма, зависит от метрики, которую предполагается использовать.

Суть лапласовского сглаживания состоит в том, что ко всем элементам L_1-нормированного вектора добавляется небольшая положительная константа α, а затем изменённый вектор подвергается перенормировке по L_1-норме, так что его опять можно рассматривать как дискретное распределение, в котором, однако, каждому событию теперь уже приписана ненулевая вероятность. В наших расчётах использовалась $\alpha = 10^{-5}$.

Для измерения семантических расстояний использовались четыре функции, которые мы далее будем нестрого называть метриками: косинус, коэффициент Жаккара, мера Йенсена-Шеннона и симметризованная мера Кульбака-Лейблера.

Пусть даны векторы v_a и v_b одинаковой длины n. Косинус угла между ними $\cos(v_a, v_b) = \frac{\langle v_a, v_b \rangle}{||v_a||_{L_2} \cdot ||v_b||_{L_2}}$, где $\langle \cdot, \cdot \rangle$ – скалярное произведение. Косинус вычислялся на PMI-взвешенных данных, т. к. известно, что это улучшает результаты (см., например, [10]).

Для тех же векторов коэффициент Жаккара $J(v_a, v_b) = \frac{\sum_{i=1}^{n} \min(v_{ai}, v_{bi})}{\sum_{j=1}^{n} \max(v_{aj}, v_{bj})}$. Небольшой предварительный опыт показал, что оптимальным является вычислять коэффициент Жаккара на L_2-нормированных данных.

Пусть теперь v_a и v_b – L_1-нормированные сглаженные векторы. Дивергенция Кульбака-Лейблера $KL(v_a||v_b) = \sum_{i=1}^{n} v_{ai} \log_2 \frac{v_{ai}}{v_{bi}}$. Поскольку в общем случае $KL(v_a||v_b) \neq KL(v_b||v_a)$, мы вычисляли симметризованное расстояние как $\frac{KL(v_a||v_b) + KL(v_b||v_a)}{2}$.

Мера Йенсена-Шеннона вычисляется из тех же соображений, но не требует сглаживания. Пусть опять v_a и v_b L_1-нормированы и $\overline{v} = (\frac{v_{a1} + v_{b1}}{2}, \ldots, \frac{v_{an} + v_{bn}}{2})$ – их почленное арифметическое среднее. Тогда мера Йенсена-Шеннона $JS(v_a||v_b) = KL(v_a||\overline{v}) + KL(v_b||\overline{v})$.

Две популярные метрики: евклидово расстояние и коэффициент Дайса – не представлены в наших расчётах. Евклидово расстояние очень чувствительно к случаям, когда рассчитывается семантическая близость между словами с резко различающимися частотами. Поэтому вычислять его есть смысл лишь на L_2-нормированных данных, т. е. как расстояние между парой точек на единичной гиперсфере, которое легко выражается через косинус угла между соответствующими векторами и монотонно относительно него. Иначе говоря, евклидово расстояние не обещает ничего нового по сравнению с косинусом. Избыточным было бы рассматривать и коэффициент Дайса, который монотонен относительно коэффициента Жаккара [26, p. 53].

3.3. Оценивание

Эталон был построен на основе словаря [6], который в электронном виде содержится в программе Н. Кецариса «Рифмовник» (http://rifmovnik.ru) версии 2.01b. Выбирались пары слов-синонимов, каждое из которых в обоих корпусах имеет частоту от 10 до 100 на миллион словоупотреблений. Таких пар набралось 716, а после дополнительной вычитки их осталось 580: исключались пары, элементы которых синонимичны только в одном из ЛСВ, преимущественно переносном (*фрукт – тип*, *вагон – куча* и под.). Из этих 580 пар случайным перемешиванием было получено ещё 2900, в которых слова, согласно эталону, не только сами не являются синонимами, но и не имеют общих синонимов. Ожидается, что семантическая близость в таких парах должна быть в среднем существенно меньше.[1]

При оценивании все 3480 пар слов упорядочивались по убыванию семантической близости (т. е. для косинуса и коэффициента Жаккара – по убыванию, для мер Йенсена-Шеннона и Кульбака-Лейблера – по возрастанию). Подсчитывалась доля эталонных пар синонимов от первых 580 элементов списка. Легко убедиться, что это известная F-мера [26, p. 83]: среднее гармоническое точности (precision) и охвата (recall), которые в данном случае совпадают. Кроме того, мы пробовали оценивать качество модели суммой рангов эталонных пар синонимов в упорядоченном списке (ср. [9, p. 37]), но оказалось, что этот способ даёт почти такие же результаты, как и F-мера.

4. Результаты

4.1. Отбор признаков в базис

Размерность полученного пространства составила порядка 10^5 (для художественных текстов – 171830 лемм в базисе, для газетных – 169005, для совмещённого корпуса – 282932). Это избыточно много для эффективного расчёта мер близости на

[1] Наш эталон наряду с программами, которые использовались при расчётах, доступен для скачивания на странице https://github.com/poritski/RuVSM.

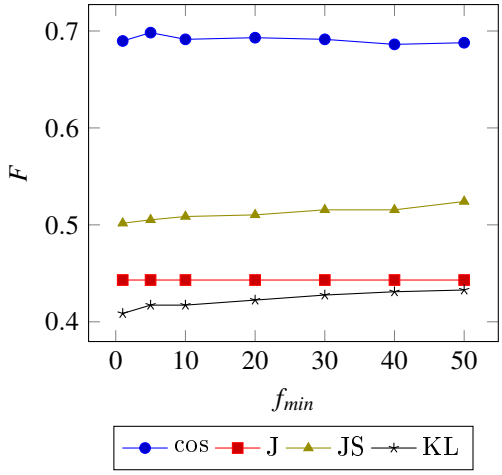

Рис. 1. Зависимость качества оценивания от нижнего частотного порога, f_{max} не ограничена

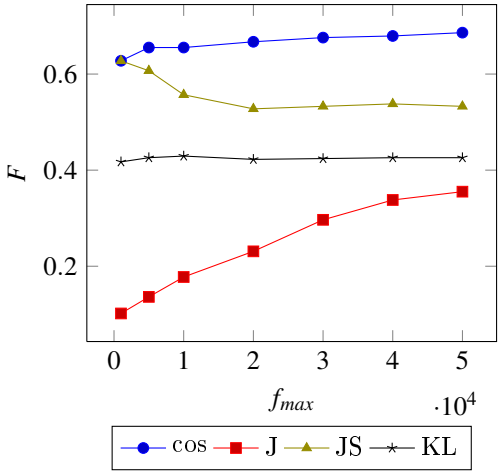

Рис. 2. Зависимость качества оценивания от верхнего частотного порога, $f_{min} = 50$

векторах, т. к. время вычисления линейно зависит от их длины. В нашей ситуации «проклятие размерности» смягчается тем фактом, что эталон содержит всего несколько тысяч пар слов. Но чтобы стало возможным использовать систему в практических целях, размерность нужно снизить. Простейший способ сделать это – отбор признаков (лемм) в базис в зависимости от ряда параметров: частота слова в корпусе, его длина, стандартное отклонение по соответствующему столбцу и др.

Мы взяли модель, обученную на корпусе художественных текстов с шириной контекстного окна 2, и исследовали поведение разных метрик при отборе признаков в базис с ограничением по максимальной и минимальной частоте. Полученные результаты представлены на рис. 1 и 2.

На рис. 1 по оси абсцисс отложены значения нижнего частотного порога f_{min}: испытывались значения, равные 1 (отбраковки нет), 5, 10, ..., 50. По оси ординат отложена F-мера. Оказывается, что при отсутствии верхней частотной границы нижнюю границу имеет смысл поднимать, т. к. это резко снижает размерность модели (до порядка 10^4 признаков) и почти не меняет качество оценивания. На коэффициенте Жаккара подъём нижнего порога никак не сказался, в теоретико-информационных метриках обеспечил улучшение на 1...3%, косинус меняется труднопредсказуемо и при $f_{min} = 40$ имеет минимум, который мало (около 1%) отличается от максимума, достигнутого при $f_{min} = 5$. Таким образом, гапаксы и вообще низкочастотные слова можно не включать в базис.

На рис. 2 по оси абсцисс отложены значения верхнего частотного порога f_{max}, равные 1 тыс., 5 тыс., 10 тыс., ..., 50 тыс. В этих вычислениях f_{min} была закреплена равной 50. Как видно, при снижении верхней частотной границы (при довольно высоко зафиксированной нижней) качество ухудшается во всех метриках, кроме Йенсена-Шеннона. Особенно впечатляюще падает качество оценивания коэффициентом Жаккара. При расчётах с его использованием (а также и с использованием косинуса) высокочастотные слова следует оставлять в базисе, тем более что отказ от них слабо снижает размерность модели. Напротив, мера Йенсена-Шеннона требует устанавливать верхний частотный порог достаточно низко: максимум F-меры приходится на $f_{max} = 10^3$.

4.2. Ширина контекстного окна

Из рис. 1 и 2 видно, что стабильно лучший результат в нашей задаче показывает косинус, а на втором месте оказывается мера Йенсена-Шеннона. Коэффициент Жаккара и мера Кульбака-Лейблера, по крайней мере при испытанных комбинациях параметров, не так хорошо отличают синонимы от несинонимов. Мы исключили их из дальнейшего рассмотрения и исследовали, как поведение двух оставшихся метрик зависит от выбора исходного корпуса и ширины контекстного окна при обучении модели (см. рис. 3 и 4).

Было найдено, что газетные тексты хуже подходят для обучения модели, чем художественные. Как кажется, это обусловлено не только объёмом корпуса, но и стилевой спецификой (более клишированный язык, ограниченный тематический репертуар). На объединённой коллекции текстов косинус и мера Йенсена-Шеннона ведут себя почти одинаково, но с одним тонким различием.

Косинус, представленный на рис. 3, позволяет совсем не менять настройки модели при переходе от небольшого одностилевого корпуса к сравнительно

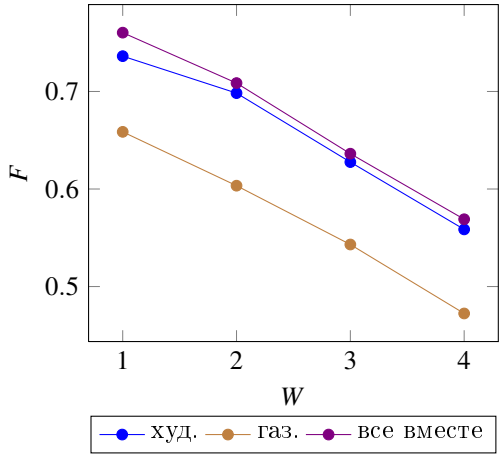

Рис. 3. Зависимость качества оценивания от ширины окна, cos + PMI, $f_{min} = 5$

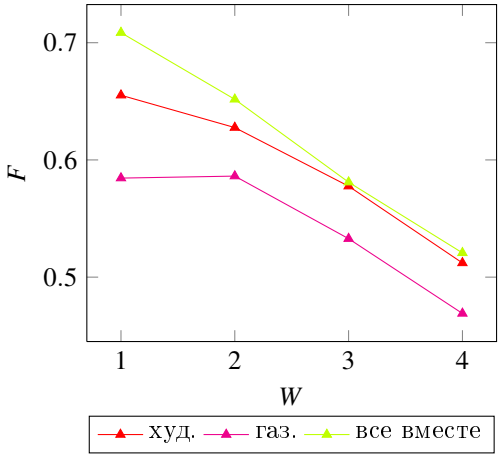

Рис. 4. Зависимость качества оценивания от ширины окна, JS + L_1, $f_{min} = 50\,/\,100$, $f_{max} = 1000\,/\,2000$

большому и стилистически разнородному: на газетных текстах достигнуто $F = 0.659$, на художественных $F = 0.736$ и на сводном корпусе $F = 0.760$, всё это при нижнем частотном пороге $f_{min} = 5$ и ширине контекстного окна 1 (с расширением окрестности качество предсказаний падает почти линейно).

Мера Йенсена-Шеннона, представленная на рис. 4, тоже оставляет возможность масштабировать исходный корпус, правда, чуть более сложным образом. Когда отбор признаков в базис происходит по частотным порогам $f_{min} = 50$, $f_{max} = 1000$, удаётся достичь $F = 0.586$ на газетных текстах и $F = 0.655$ на художественных текстах. Качество предсказаний падает с ростом ширины окна, хотя и не так последовательно, как в случае косинуса. Переход к сводному корпусу сколько-нибудь улучшает результаты, только если частотные пороги пропорционально меняются. При $f_{min} = 100$, $f_{max} = 2000$ на объединённой коллекции текстов удалось получить $F = 0.709$.

4.3. «Аутсайдеры»

Таблицы семантических расстояний, измеренных косинусом и мерой Йенсена-Шеннона при указанных оптимальных параметрах, были рассмотрены нами более подробно. Упорядочив списки по убыванию семантической близости, мы обратили внимание на два момента:

- какие пары, отсутствующие в эталоне, попадают в первые 580 записей, т. е. оцениваются как синонимичные;
- какие пары синонимов, напротив, оказываются в конце списка, в числе заведомых несинонимов.

Оказалось, что в обеих метриках высокую степень близости часто демонстрировали пары слов, смысловое сходство которых трудно проинтерпретировать: *драма – бытие*, *тайна – счастье* и др. Тем не менее, в верхушку списка попало и несколько вполне осмысленных групп лексики:

- видовые пары (*перевести – переводить*);
- синонимы, отсутствующие в эталоне (*неплохой – отличный*, *уничтожать – разбирать*);
- антонимы (*умный – глупый*, *подтверждать – отрицать*, *забирать – класть*) и иные парадигматически связанные слова (*светлый – тьма*, *потрясти – ударить* и др.);
- синтагматически связанные слова (*обитать – планета*, *проявить – слабость* и др.).
- ассоциативно связанные слова, которые могут описывать одну ситуацию (*захватывать – криминальный*, *последствие – участь* и др.);

Видовые пары глаголов и вся парадигматика (в том числе антонимы и синонимы, не отмеченные в эталоне) почти одинаково выделяются обеими метриками как семантически близкие, а состав остальных групп различается. Можно заметить, что мера Йенсена-Шеннона чаще перемещает вверх синтагматически связанные слова, нередко разной частеречной принадлежности: *материальный – исправлять*, *духовный – соединять* и т. п.

При измерении семантической близости косинусом пары синонимов попадают в нижнюю часть таблицы очень редко. Среди них оказываются модальные слова (*итак – следовательно*), а также пары с омонимом или полисемантом, который вступает в синонимические отношения только одним своим лексическим значением, обычно менее частотным (*здорово – привет*, *согласный – стройный*). При использовании меры Йенсена-Шеннона обе эти особенности

проявляются ещё отчётливее: пары типа *видимо – очевидно*, *очевидно – вероятно* или *задний – обратный*, *править – распоряжаться* чаще попадают в число дистрибутивно несхожих.

Вычисление семантических расстояний в каждой паре лемм, для которых в модели выучены векторы, указало бы, видимо, много новых пар семантически близких слов, отсутствующих не только в эталоне, но и в целом словаре [6]. В сущности, получился бы автоматически построенный тезаурус, как в работах [12] и [9]. Но эта задача имеет сложность по времени $O(r^2c)$, где $r \sim 10^5$ – количество попарно сравниваемых лемм, $c \sim 10^4$ – количество элементов базиса. Такого рода опыт пока выходит за рамки наших возможностей.

5. Выводы и перспективы

Подводя итоги наших первых экспериментов, надо отметить следующие важные факты.

1. Формировать исходный корпус при построении VSM лучше всего на основе художественных текстов. В ряде работ последних лет ([8], [14] и др.) VSM обучаются на англоязычной Википедии. Наши предварительные результаты косвенно говорят о том, что это, возможно, не самый плодотворный путь, т. к. по клишированности язык Википедии сближается с языком газетных текстов.

2. Модели с единичной шириной контекстного окна лучше всего отличают синонимы от несинонимов. Это согласуется с результатами О. Ферре [10], а также М. Зальгрена, обнаружившего, что парадигматические связи в лексике наиболее ярко проявляются при малой ширине контекстного окна [20, p. 106-107].

3. От низкочастотных элементов базиса можно избавляться, а утрата высокочастотных, наоборот, снижает качество модели (за исключением меры Йенсена-Шеннона). Это противоречит общепринятой практике удаления высокочастотных стоп-слов [25, p. 154].

4. Лучшие результаты показывает косинус с PMI-взвешиванием, а мера Йенсена-Шеннона чуть уступает ему по качеству предсказаний, хотя и допускает более радикальное снижение размерности.

В дальнейшем будет интересно оснастить наши корпуса зависимостной синтаксической разметкой (например, с помощью парсера Сергея Протасова – http://sz.ru/parser) и перестроить VSM с учётом синтаксических связей. Предстоит также поэкспериментировать с одно- и разностилевыми корпусами большего размера, порядка 10^8 словоупотреблений, и с техниками снижения размерности векторных представлений (такими, как SVD). Некоторые популярные схемы взвешивания признаков (бинаризация, tf-idf, t-тест) и метрики (мера Лина, α-skew) не были испытаны в наших экспериментах и ещё ожидают анализа на русском материале. Наконец, очень полезно будет пополнить эталон парами антонимов, когипонимов и иных парадигматически связанных слов, чтобы стало возможным перейти к задаче разграничения типов семантических отношений по дистрибутивным данным.

Список литературы

[1] Л. Витгенштейн, *Философские исследования*. Новое в зарубежной лингвистике, вып. XVI, М., 1985, 79–128.

[2] И. С. Мисуно, Д. А. Рачковский, С. В. Слипченко, *Векторные и распределённые представления, отражающие меру семантической связи слов*. Математичні машини і системи (2005), №3, 50–66.

[3] О. А. Митрофанова, *Измерение семантических расстояний как проблема прикладной лингвистики*. Структурная и прикладная лингвистика, вып. 7, СПб., 2008, 92–101.

[4] В. А. Москович, *Статистика и семантика (опыт статистического анализа семантического поля)*. М., 1969.

[5] Б. А. Плотников, *Дистрибутивно-статистический анализ лексических значений*. Минск, 1979.

[6] *Словарь синонимов: Справочное пособие*. Под ред. А. П. Евгеньевой. М., 1975.

[7] М. В. Хохлова, *Исследование лексико-синтаксической сочетаемости в русском языке с помощью статистических методов (на базе корпусов текстов)*. АКД, СПб., 2010.

[8] G. Boleda, E. M. Vecchi, M. Cornudella & L. McNally, *First-order vs. higher-order modification in distributional semantics*. Proceedings of EMNLP-CoNLL'2012, 1223–1233.

[9] J. R. Curran, *From distributional to semantic similarity*. Ph.D. thesis, University of Edinburgh, 2004.

[10] O. Ferret, *Testing semantic similarity measures for extracting synonyms from a corpus*. Proceedings of LREC'2010, 3338–3343.

[11] J. R. Firth, *A synopsis of linguistic theory 1930–1955*. Studies in Linguistic Analysis (Blackwell, Oxford, 1957), 1–32.

[12] G. Grefenstette, *Explorations in automatic thesaurus discovery*. Kluwer Academic Publishers, 1994.

[13] Z. Harris, *Distributional structure*. Word (1954), 10 (23), 146–162.

[14] A. Herbelot, *What is in a text, what isn't, and what this has to do with lexical semantics*. Proceedings of IWCS'2013, 321–327.

[15] M. Kopotev, S. Sharoff, T. Erjavec, A. Feldman & D. Divjak, *Designing and evaluating Russian tagset*. Proceedings of LREC'2008, 279–285.

[16] L. Lee, *Similarity-based approaches to natural language*

processing. Ph.D. thesis, Harvard University, 1997.

[17] G. A. Miller & W. G. Charles, *Contextual correlates of semantic similarity*. Language and Cognitive Processes (1991), 6(1), 1–28.

[18] F. Pereira, N. Tishby, & L. Lee, *Distributional clustering of english words*. Proceedings of ACL'93, 183–190.

[19] H. Rubenstein & J. Goodenough, *Contextual correlates of synonymy*. Computational Linguistics (1965), 8, 627–633.

[20] M. Sahlgren, *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. Ph.D. thesis, Stockholm University, 2006.

[21] H. Schmid, *Probabilistic part-of-speech tagging using decision trees*. Proceedings of international conf. on new methods in language processing (Manchester, 1992).

[22] H. Schütze, *Dimensions of meaning*. Proceedings of the conf. on supercomputing (Minneapolis, 1992), 787–796.

[23] I. Segalovich, *A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine*. Proceedings of MLMTA'2003, 273–280.

[24] K. Spärck Jones, *Synonymy and semantic classification*. Doctoral diss., University of Cambridge, 1964.

[25] P. D. Turney & P. Pantel, *From frequency to meaning: vector space models of semantics*. Journal of Artificial Intelligence Research (2010), 37, 141–188.

[26] J. Weeds, *Measures and applications of lexical distributional similarity*. Ph.D. thesis, University of Sussex, 2003.