И. В. Трофимов, Е. А. Сулейманова

Дистрибутивно-семантическая модель для выявления категориального сходства

Аннотация. В последнее время дистрибутивно-семантическое сообщество задается вопросом о целесообразности использования древесно-синтаксического контекста для моделирования дистрибутивной семантики. Построение синтаксических дистрибутивных моделей требует больших объемов аннотированных текстовых данных и сопряжено со значительными вычислительными затратами. Между тем, современные методы построения линейно-оконных моделей позволили им превзойти синтаксические модели на большинстве типовых задач.

Представленное в статье сравнительное исследование синтаксических и линейно-оконных моделей продемонстрировало, что при решении задач, связанных с различением категориального сходства и тематической ассоциации, использование синтаксических моделей вполне оправданно.

Ключевые слова и фразы: дистрибутивно-семантическая модель, модель дистрибутивной семантики, синтаксическая дистрибутивно-семантическая модель, категориальное сходство, синтаксическое отношение, word2vec, skipgram, RuSim1000.

Введение

В основе дистрибутивного моделирования семантики лежит идея представления значения языковой единицы (в частности, слова) через ее дистрибутивные свойства, получаемые из контекстов употребления единицы в больших объемах текстового материала. Центральным понятием дистрибутивной семантики выступает *дистрибутивное*

Работа выполнена по НИР «Исследование и разработка методов автоматического извлечения событийно-темпоральной информации из текстов» в рамках гос. задания, номер гос. регистрации AAAA-A17-117040610371-7.

- © И. В. Трофимов, Е. А. Сулейманова, 2018
- © Институт программных систем имени А. К. Айламазяна РАН, 2018
- © Программные системы: теория и приложения (дизайн), 2018

сходство — общность контекстов употребления языковых единиц, вычисляемая как сходство между векторными представлениями последних.

Долгое время, в соответствии с так называемой *дистрибутивной гипотезой*, считалось, что дистрибутивное сходство коррелирует со сходством (similarity) значений. Однако в последние годы научное сообщество постепенно приходит к согласию, что в общем случае степень дистрибутивного сходства между словами позволяет оценить, лишь насколько сильно их значения в языке связаны друг с другом. Другими словами, дистрибутивное сходство — это количественный параметр, который в каждом конкретном случае характеризует содержательно разные концептуальные отношения. Вслед за когнитивной психологией, различающей два вида знаний: знания о таксономическом, категориальном сходстве, основанном на общности физических или функциональных свойств, и знания, группирующие концепты на основе тематической, ситуативной близости, совместной встречаемости и т. п., всё разнообразие концептуальных отношений в корпусной семантике принято сводить к двум типам: (семантическое) *сходство* и (тематическая) *ассоциация*. Иногда сходство выделяют в особый тип отношений, противопоставляя его прочим видам семантической *связанности* (semantic relatedness). Вопросы терминологии довольно подробно рассматриваются в [1].

Несмотря на кажущуюся очевидность, понятие сходства оказывается настолько гибким, что с трудом поддается формальному определению даже в психологии, где является по существу одним из центральных теоретических конструкторов. В дистрибутивной семантике имеют хождение разные трактовки сходства в терминах лексико-семантических отношений — синонимии (квазисинонимии), гипо-гиперонимии, согипонимии, антонимии.

Способность дистрибутивно-семантической модели (ДСМ) отличать понимаемое тем или иным образом семантическое сходство от других случаев семантической связанности востребована в таких задачах, как автоматизированное построение словарей, тезаурусов, онтологий, разрешение лексической неоднозначности, извлечение информации, машинный перевод, автоматическое построение обучающих корпусов для различных задач и др. В частности, в [2] описан эксперимент, в котором на основе ДСМ подбирались замены для автоматического

аннотирования обучающих множеств при решении задачи разрешения событийно-несобытийной неоднозначности существительных.

В зависимости от типа используемых контекстов, ДСМ разделяются на линейно-оконные и древесно-синтаксические. Построение последних значительно более затратно как в смысле требуемых лингвистических ресурсов (им необходимы большие объемы синтаксически аннотированных текстов), так и в вычислительном отношении. Высказываются сомнения относительно того, насколько это оправданно в настоящее время, поскольку утверждается [3], что должным образом подобранные параметры позволяют современным линейно-оконным ДСМ превзойти синтаксические на большинстве типовых задач. С другой стороны, существует подкрепленное экспериментами мнение [4], что синтаксические ДСМ лучше моделируют семантическое сходство, тогда как линейно-оконные модели — тематическую ассоциацию. В [1] изложены результаты оценки русскоязычной синтаксической модели на специфической задаче различения случаев категориального сходства (возможно, в комбинации с тематической близостью) и чисто тематической ассоциации. Для оценки использовался оригинальный дейтасет RuSim1000, состоящий из 1000 пар семантически связанных существительных, 500 положительных примеров (существительные в паре категориально схожи) и 500 отрицательных примеров (существительные связаны ассоциативно, но не схожи). Представленная модель довольно успешно справилась с задачей.

Настоящее исследование предпринято со следующей целью. Во-первых, предложить метод построения синтаксических ДСМ на современной методологической базе и, во-вторых, провести сравнительный анализ древесно-синтаксических и линейно-оконных ДСМ на задаче выявления категориального сходства [1].

1. Метод построения дистрибутивно-семантической модели

В области разработки дистрибутивно-семантических моделей тенденцией последнего времени стало использование (в качестве методологической основы) нейросетевых решений cbow и skipgram, предложенных Миколовым [5]. Благодаря вычислительной эффективности и высокому качеству получаемых ДСМ, эти подходы получили широкое распространение и легли в основу многочисленных вариаций.

В частности, Леви и Голдберг [6] предложили использовать отдельные словари для *слов модели* и *контекстов* в рамках архитектуры skipgram. В этом случае словарь контекстов может кодировать произвольную информацию, в том числе синтаксическую. Для определения того, что является контекстом данного слова, Леви и Голдберг вместо традиционного линейного окна использовали синтаксическое окно (множество слов, находящихся в каком-либо синтаксическом отношении с данным словом). В свою очередь, для моделирования контекстов использовался кортеж вида `<dir, dep_type, word>`, где `dir` — направление синтаксической связи, `dep_type` — тип синтаксического отношения, `word` — контекстная лексическая единица. Для создания своей модели мы придерживались того же принципа, но внесли ряд изменений, о которых пойдет речь в дальнейшем.

В разработанном нами методе построения ДСМ можно выделить следующие крупные шаги:

- морфологический и синтаксический анализ текстового корпуса;
- формирование словарей: *словника* (словаря слов модели) и *словаря контекстов*;
- обучение нейросети skipgram, адаптированной для работы с отдельными словарями.

Поскольку специфика нашего метода не затрагивает вопросов обучения нейросети, ограничимся рассмотрением первых двух шагов.

1.1. Морфологический и синтаксический анализ корпуса

Для морфологической и синтаксической разметки текстового корпуса мы использовали аналитический конвейер¹, созданный по инициативе Шарова [7] и состоящий из токенизатора, теггера частей речи TreeTagger, морфологического анализатора CSTlemma и синтаксического анализатора MaltParser, обученного на корпусе SynTagRus [8]. Эмпирические исследования [9] показали, что MaltParser обеспечивает довольно высокую точность применительно к русскому языку: 89 % точность установления синтаксического родителя (UAS, unlabeled attachment score), 82 % точность установления и родителя, и типа синтаксического отношения (LAS, labeled attachment score). Причем для существительных, прилагательных, числительных и причастий

¹<http://corpus.leeds.ac.uk/mocky/>

эти показатели еще выше, но с поиском родителей для предлогов дело обстоит существенно хуже ($UAS = 79$, $LAS = 60$). Разработчики отмечают, что это третий по качеству синтаксический парсер для русского языка (по состоянию на 2012 год)².

После обработки корпуса конвейером Шарова выяснилось, что для заметного числа словоформ в тексте нормальные формы либо не построены, либо построены некорректно (отчасти это было обусловлено ошибками токенизации), при этом морфологические характеристики были предсказаны. Мы выполнили дополнительную обработку результатов нормализации следующим образом. Каждая русскоязычная словоформа обрабатывалась морфологическим анализатором, разработанным сообществом АОТ [10]. Затем среди полученных вариантов лемм отыскивались такие, для которых выполняется условие: в парадигме присутствует словоформа, совпадающая с исходной (текстовой) формой, а ее морфологические атрибуты совпадают с морфологическими характеристиками, предсказанными анализатором *CSTlemma* (сопоставление морфологических характеристик выполнялось частично и зависело от части речи).

В результирующей разметке сохранялись только те словоформы, для которых выполняется хотя бы одно из следующих условий:

- упомянутому выше условию соответствует ровно одна лемма;
- упомянутому выше условию соответствуют несколько лемм, но среди них имеется такая, что ее нормальная форма совпадает с предсказанной анализатором *CSTlemma*.

Для причастий был реализован специальный алгоритм восстановления нормальной формы по адъективной модели (анализатор АОТ приводит причастия к форме инфинитива соответствующего глагола).

Кроме перечисленного, были разработаны эвристические процедуры для обработки слов, содержащих дефисы. Мы полагаем, что они не оказали существенного влияния на результат, поэтому опускаем их детальное описание.

²На сайте <http://corpus.leeds.ac.uk/mocky/> говорится следующее:

In a competition of Russian dependency parsers in 2012 this simple parser produced fairly reliable results, ranking 3 out of 8 by the F-measure.

1.2. Формирование словарей

В рамках описываемого исследования мы ограничились построением дистрибутивно-семантической модели только для нарицательных существительных. Кроме того, нам хотелось создать *качественную* модель, избавленную от слов-опечаток, ошибок нормализации и прочих «шумов», поэтому список лексики (как для словника, так и для словаря контекстов) формировался вручную с группировкой по частям речи. Всего рассматривалось четыре части речи: существительное, прилагательное, глагол, причастие.

Для перечня существительных за основу был взят «Толковый словарь русских существительных»³, дополненный выявленными в корпусе частотными существительными. Все существительные, которые аналитический конвейер помечал как собственные имена, отфильтровывались. Лексика других частей речи отбиралась только по частотному порогу⁴.

Корректность отобранных лексических единиц верифицирована людьми. Исправлять приходилось также неточности аналитического конвейера. Например, названия геополитических образований и географических объектов конвейер причислял к именам нарицательным.

В ходе формирования словаря контекстов не обошлось без эвристических приемов. **MaltParser**, обученный на корпусе **SynTagRus**, порождает поверхностно-синтаксическую разметку с довольно детальной типизацией синтаксических отношений (СинтО). Чтобы ограничить объем словаря контекстов, было предпринято следующее.

Во-первых, **нормализация предлогов** на основе построенного вручную словаря предлогов. Разные предлоги, имеющие схожее значение независимо от управляемого падежа (например, *вроде*, *навряде*, *наподобие*, *подобно*, *сродни*, *а-ля*) приводятся к единому идентификатору. С другой стороны, одному и тому же предлогу, имеющему разное значение в зависимости от управляемого падежа (например, *на* с винительным и *на* с предложным падежом), соответствуют разные идентификаторы.

Во-вторых, **преобразование комбинаций**, содержащих некоторые типы СинтО.

³https://noun_ru.academic.ru

⁴В экспериментах порог составлял 1000 словоупотреблений.

Предложное СинтО. Словоформа, подчиняющая предлог по некоторому СинтО, напрямую связывается со словоформой, подчиненной предлогу по предложному СинтО. Имя полученной связи содержит имя СинтО и идентификатор предлога. Например:

преобразуется в

Ср. решение для предлогов в [6]: предлог по существу используется для «означивания» отношения, которым связываются подчиняющая предлог и подчиненная предлогу словоформы.

Аналитическое, присвязочное и пассивно-аналитическое СинтО. Узел, из которого наряду с предикативным СинтО выходит присвязочное СинтО, игнорируется, а оба его зависимых соединяются фиктивной связью «предик+присвяз». Направление связи сохраняет подчиненность предикативного зависимого. Аналогичные операции выполняются для комбинаций предикативного СинтО с аналитическим и пассивно-аналитическим СинтО (строятся связи «предик+аналит» и «предик+пасс-анал»). Например:

Сочинительное СинтО, а также сочетание его с *сочинительно-союзным СинтО* заменяется фиктивной ненаправленной сочинительной связью. Кроме того, в цепочке из трех и более однородных членов такая равноправная сочинительная связь строится для каждой пары.

Наконец, еще одна возможность ограничить многообразие синтаксических контекстов состоит в **укрупнении типов СинтО**. Заметим, что вопросы использования и возможной ревизии в интересах построения ДСМ такой детальной системы поверхностно-синтаксических отношений, какой является используемая в **SynTagRus** разработка Лаборатории компьютерной лингвистики ИППИ РАН (порядка 75 типов

отношений [11]), заслуживают отдельного исследования. Возможно, целесообразнее было бы воспользоваться набирающим популярность межъязыковым способом описания синтаксиса, основанным на языке универсальных зависимостей [12]. Как бы то ни было, для экспериментов, представленных в настоящей статье, мы используем всего 23 СинтО, сгруппированных в 6 сводных типов связей: агентивно-предикативная, 1-комплетивная, комплетивно-обстоятельственная, атрибутивно-определятельная, композитная, сочинительно-кратная.

2. Эксперименты и результаты

Оценка эффективности построенных дистрибутивно-семантических моделей выполнялась на задаче выявления категориального сходства. В ходе экспериментального исследования проводилось сопоставление ДСМ, полученных нашим методом (далее *ДСМ-синт*), с моделями, созданными при помощи традиционных линейно-оконных подходов. Мы исследовали возможности:

- cbow- и skipgram-моделей, обученных на том же текстовом корпусе, который использовался для построения ДСМ-синт; обозначим их *ДСМ-лин*;
- общедоступных cbow- и skipgram-моделей, созданных на базе других корпусов.

Корпус, который использовался для обучения нашей модели, представляет собой смесь новостных сообщений и художественной литературы. Общий объем корпуса около 2 млрд. токенов (из них 0.7 млрд. приходится на новостные сообщения). При помещении в корпус каждый текст разбивался на предложения; предложения-дубликаты исключались. Таким образом, уместно говорить не о корпусе текстов, а о корпусе уникальных предложений русского языка (устранение дубликатов влияет на частоты употребления слов). Для линейно-оконных экспериментов данное множество предложений преобразовывалось в последовательность токенов без знаков препинания и границ предложений.

В качестве инструментария для создания ДСМ-лин мы использовали оригинальную утилиту `word2vec`⁵. За базовую конфигурацию был принят следующий набор параметров (таблица 1).

⁵<https://code.google.com/archive/p/word2vec/>

ТАБЛИЦА 1. Базовый набор параметров для построения ДСМ-лин

Размер обучающего множества	≈2 млрд. токенов
Частотный порог для слов в словаре (min-count)	100
Размерность эмбединга (size)	300
Размер окна (window)	2
Частотное подавление (sample)	1e-4
Вычислительная оптимизация	negative sampling (3)
Эпох обучения (iter)	3

Для поиска наилучшей ДСМ-лин мы варьировали параметры и оценивали модели на дейтасете для выявления категориального сходства RuSim1000. Следует оговориться, что мы использовали сокращенную версию RuSim1000 по следующей причине. Объем словника моделей ДСМ-синт был небольшой. Из-за неполноты словника оказалось невозможно оценить 86 пар из оригинального RuSim1000. Чтобы уравнивать возможности моделей, мы сократили RuSim1000 до 876 пар. В результате для ДСМ всех трех сопоставляемых групп число неоцененных пар было в пределах 10.

Наибольшей точности (0.8952) на дейтасете удалось добиться ДСМ-лин с параметрами, показанными в таблице 2.

ТАБЛИЦА 2. Набор параметров ДСМ-лин, достигшей наиболее высокой средней точности

Частотный порог для слов в словаре (min-count)	100
Размерность эмбединга (size)	300
Размер окна (window)	1
Вычислительная оптимизация	negative sampling (20)
Эпох обучения (iter)	3

Сначала мы выяснили, как на качество результата влияет выбор частотного порога при формировании словаря (рисунок 1а).

Выбор размера окна значительно влияет на результат (рисунок 1б). Наилучшей точности удалось добиться на окне размера 1. Это согласуется с результатами, полученными ранее другими исследователями на схожих задачах (см., например, рис. 3 в [13]). В дальнейшем мы ограничились исследованием моделей с окнами размера 1 и 2.

Количество эпох обучения (рисунок 1б) не оказывает существенного влияния на качество модели. В свою очередь, выбор способа оптимизации оказался значимым (рисунок 1г).

РИСУНОК 1. Зависимость эффективности ДСМ-лин от частотного порога для словаря, размера окна, количества эпох обучения и параметров оптимизации

Так как выбор частотного порога (в исследованных пределах) не оказывает существенного влияния на результат, для дальнейших экспериментов мы ограничились исследованием ДСМ-лин с частотным порогом 100.

Результаты оценки общедоступных линейно-оконных моделей, построенных на базе архитектур cbow и skipgram (третья группа исследуемых ДСМ), приводятся в таблице 3. Как показывают результаты, увеличение объема корпуса само по себе не приводит к росту средней точности на дейтасете RuSim1000.

Теперь перейдем к рассмотрению ДСМ-синт.

ТАБЛИЦА 3. Эффективность общедоступных линейно-оконных ДСМ

Модель и ее параметры		Средняя точность, RuSim1000 (876 пар)
taiga_upos_skipgram_300_2_2018		
источник	RusVectōrēs ^a	0.87874
корпус	≈5 млрд. слов	
частотный порог	200	
архитектура	skipgram	
размерность эмбединга	300	
размер окна	2	
ruwikiruscorpora_upos_skipgram_300_2_2018		
источник	RusVectōrēs ^a	0.85156
корпус	≈0.6 млрд. слов	
частотный порог	40	
архитектура	skipgram	
размерность эмбединга	300	
размер окна	2	
news_upos_cbow_600_2_2018		
источник	RusVectōrēs ^a	0.80996
корпус	≈5 млрд. слов	
частотный порог	200	
архитектура	cbow	
размерность эмбединга	600	
размер окна	2	
all.norm-sz500-w10-cb0-it3-min5		
источник	RDT ^b	0.82322
корпус	≈12.9 млрд. слов	
частотный порог	5	
архитектура	skipgram	
размерность эмбединга	500	
размер окна	10	

^a<https://rusvectors.org/ru/models/>^bhttps://nlpub.ru/Russian_Distributional_Thesaurus

Объёмы списков, полученных в результате отбора лексики, показаны в таблице 4.

Для формирования словника использовался только список существительных. Для формирования словаря контекстов — все четыре списка.

Таблица 4. Результат отбора лексики для построения модели

существительных	19088
прилагательных	13912
причастий	992
глаголов	8593
итого	42585

При создании словаря контекстов мы установили частотный порог 5. Это означает, что для попадания в словарь кортеж `<dir, dep_type, word>` должен встречаться не менее 5 раз в качестве контекста существительных (из списка существительных).

Варьируемыми параметрами в исследовании стали: размерность эмбединга, количество эпох обучения и параметры оптимизации. Их начальные значения — 300, 3 и `negative sampling` (3) соответственно. Результаты представлены на рисунках 2а–2в.

Рисунок 2. Зависимость эффективности ДСМ-лин от количества эпох обучения, размерности эмбединга и параметров оптимизации

Наилучшая из ДСМ-синт достигла величины средней точности 0.95039 при следующих параметрах: размерность эмбединга 300, количество эпох обучения 3, оптимизация методом negative sampling (20 случайных сэмплов на обучающий пример).

В заключение приведем сводную таблицу различных моделей и их оценки на дейтасете для категориального сходства (таблица 5).

Таблица 5. Сравнение синтаксической и линейно-оконных ДСМ

Модель и ее параметры	Средняя точность, RuSim1000 (876 пар)
лучшая из ДСМ-синт	
корпус	≈2 млрд. слов
архитектура	skipgram
размерность эмбединга	300
	0.95039
лучшая из ДСМ-лин	
корпус	≈2 млрд. слов
частотный порог	100
архитектура	cbow
размерность эмбединга	300
размер окна	1
	0.8952
taiga_upos_skipgram_300_2_2018	
корпус	≈5 млрд. слов
частотный порог	200
архитектура	skipgram
размерность эмбединга	300
размер окна	2
	0.87874

3. Заключение

Целью исследования было построение дистрибутивно-семантической модели, учитывающей синтаксический контекст, и сравнение ее

с линейно-оконными моделями на задаче различения категориального сходства и тематической ассоциации.

В сравнении участвовали:

- синтаксические модели, построенные в соответствии с изложенным в статье методом;
- линейно-оконные модели, построенные на том же корпусе, что и синтаксические;
- несколько общедоступных линейно-оконных моделей.

Исследовалось влияние параметров моделей на качество решения поставленной задачи.

Эксперименты продемонстрировали превосходство синтаксических моделей над линейно-оконными. Среди последних наилучшие результаты показали модели с минимальным размером окна.

Список литературы

- [1] I. V. Trofimov, E. A. Suleymanova. “A syntax-based distributional model for discriminating between semantic similarity and association”, *Computational Linguistics and Intellectual Technologies*, **1**:16, Proceedings of the International Conference “Dialogue 2017” (2017), pp. 349–359. [URL](#) [↑]_{444, 445}
- [2] И. В. Трофимов, Е. А. Сулейманова, Н. А. Власова, А. В. Подобреев. «Разрешение событийно-несобытийной неоднозначности существительных», *Программные системы: теория и приложения*, **9**:4 (2018), с. 3–33. [URL](#) [↑]₄₄₄
- [3] G. Lapesa, S. Evert. Large-scale evaluation of dependency-based DSMs: Are they worth the effort? *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, Short Papers*. V. 2 (April 3–7, 2017, Valencia, Spain), Association for Computational Linguistics, 2017, pp. 394–400. [URL](#) [↑]₄₄₅
- [4] F. Hill, R. Reichart, A. Korhonen. “SimLex-999: evaluating semantic models with (genuine) similarity estimation”, *Computational Linguistics*, **41**:4 (2015), pp. 665–695. [DOI](#) 10.1017/S0022268915000237 [URL](#) [↑]₄₄₅
- [5] T. Mikolov, K. Chen, G. Corrado, J. Dean. *Efficient estimation of word representations in vector space*, 2013. arXiv: 1301.3781v3 [URL](#) [↑]₄₄₅
- [6] O. Levy, Y. Goldberg. “Dependency-based word embeddings”, *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Short Papers* (June 23–25, 2014, Baltimore, Maryland, USA), 2014, pp. 302–308. [URL](#) [↑]_{446, 449}

- [7] S. Sharoff, J. Nivre. “The proper place of men and machines in language technology. Processing Russian without any linguistic knowledge”, *Computational Linguistics and Intellectual Technologies*, Papers from the Annual International Conference “Dialogue”, RGGU, 2011, pp. 657–670. ⁴⁴⁶
- [8] Ю. Д. Апресян, И. М. Богуславский, Б. Л. Иомдин, Л. Л. Иомдин, А. В. Санников, В. З. Санников, В. Г. Сизов, Л. Л. Цинман. «Синтаксически и семантически аннотированный корпус русского языка: современное состояние и перспективы», *Национальный корпус русского языка: 2003–2005. Результаты и перспективы*, Индрик, М., 2005, с. 193–214. ⁴⁴⁶
- [9] J. Nivre, I. M. Boguslavskii, L. L. Iomdin. “Parsing the SynTagRus Treebank of Russian”, *Proceedings of the 22nd International Conference on Computational Linguistics*, Coling 2008 (August 2008, Manchester, UK), 2008, pp. 641–648. ⁴⁴⁶
- [10] А. В. Сокирко. «Морфологические модули на сайте www.aot.ru», *Компьютерная лингвистика и интеллектуальные технологии*, Труды международного семинара «Диалог’2004», 2004, 7 с. ⁴⁴⁷
- [11] I. Boguslavsky. “SynTagRus — a deeply annotated corpus of Russian”, *Les motions dans le discours — Emotions in Discourse*, eds. P. Blumenthal, I. Novakova, D. Siepmann, Peter Lang, 2014, pp. 367–380. ⁴⁵⁰
- [12] J. Nivre, M.-C. de Marneffe, F. Ginter, Y. Goldberg, J. Hajič, Ch. D. Manning, R. McDonald, S. Petrov, S. Pyysalo, N. Silveira, R. Tsarfaty, D. Zeman. “Universal Dependencies v1: A multilingual treebank collection”, *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, LREC 2016, 2016, pp. 1659–1666. ⁴⁵⁰
- [13] A. Kutuzov, I. Andreev. “Texts in, meaning out: neural language models in semantic similarity task for Russian”, *Computational Linguistics and Intellectual Technologies*, Papers from the Annual conference “Dialogue”, RGGU, М., 2015, pp. 133–144. ⁴⁵¹

Поступила в редакцию 12.11.2018

Переработана 05.12.2018

Опубликована 30.12.2018

Рекомендовал к публикации

к.т.н. Е. П. Куршев

Пример ссылки на эту публикацию:

И. В. Трофимов, Е. А. Сулейманова. «Дистрибутивно-семантическая модель для выявления категориального сходства». *Программные системы: теория и приложения*, 2018, **9**:4(39), с. 443–460.

10.25209/2079-3316-2018-9-4-443-460

http://psta.psiras.ru/read/psta2018_4_443-460.pdf

Об авторах:

Игорь Владимирович Трофимов

Старший научный сотрудник Исследовательского центра искусственного интеллекта ИПС им. А. К. Айламазяна. Специалист по технологиям автоматической обработки текстов на естественном языке, автоматического планирования.

0000-0002-6903-4730

e-mail: itrofimov@gmail.com

Елена Анатольевна Сулейманова

Научный сотрудник Исследовательского центра искусственного интеллекта ИПС им. А. К. Айламазяна, одна из разработчиков технологии построения систем извлечения информации

0000-0002-0792-9651

e-mail: yes2helen@gmail.com

CSCSTI 16.31.21
UDC 004.85

Igor Trofimov, Elena Suleymanova. *A dependency-based distributional semantic model for identifying taxonomic similarity.*

ABSTRACT. Are dependency-based distributional semantic models worth the computational cost and the linguistic resources they require? As our evaluation study suggests, the answer should be "yes" if the task in hand involves distinguishing between feature-based similarity and pure association. After extensive parameter tuning, window-based models still fall behind dependency-based ones when evaluated on our Russian-language similarity/association dataset. (*In Russian*).

Key words and phrases: distributional semantic model, dependency-based DSM, taxonomic similarity, feature-based similarity, word2vec, skipgram, RuSim1000.

2010 *Mathematics Subject Classification:* 68T50, 68T05

References

- [1] I. V. Trofimov, E. A. Suleymanova. "A syntax-based distributional model for discriminating between semantic similarity and association", *Computational Linguistics and Intellectual Technologies*, **1**:16, Proceedings of the International Conference "Dialogue 2017" (2017), pp. 349–359. [URL](#) [↑]_{444,445}
- [2] I. V. Trofimov, Ye. A. Suleymanova, N. A. Vlasova, A. V. Podobryayev. "Disambiguation between eventive and non-eventive meaning of nouns", *Program Systems: Theory and Applications*, **9**:4 (2018), pp. 3–33 (in Russian). [DOI](#) [URL](#) [↑]₄₄₄
- [3] G. Lapesa, S. Evert. Large-scale evaluation of dependency-based DSMs: Are they worth the effort? *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, Short Papers. V. 2 (April 3–7, 2017, Valencia, Spain), Association for Computational Linguistics, 2017, pp. 394–400. [URL](#) [↑]₄₄₅
- [4] F. Hill, R. Reichart, A. Korhonen. "SimLex-999: evaluating semantic models with (genuine) similarity estimation", *Computational Linguistics*, **41**:4 (2015), pp. 665–695. [DOI](#) [URL](#) [↑]₄₄₅
- [5] T. Mikolov, K. Chen, G. Corrado, J. Dean. *Efficient estimation of word representations in vector space*, 2013. arXiv [arXiv](#): 1301.3781v3 [↑]₄₄₅
- [6] O. Levy, Y. Goldberg. "Dependency-based word embeddings", *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, Short Papers (June 23–25, 2014, Baltimore, Maryland, USA), 2014, pp. 302–308. [URL](#) [↑]_{446,449}

- [7] S. Sharoff, J. Nivre. “The proper place of men and machines in language technology. Processing Russian without any linguistic knowledge”, *Computational Linguistics and Intellectual Technologies*, Papers from the Annual International Conference “Dialogue”, RGGU, 2011, pp. 657–670. ⁴⁴⁶
- [8] Yu. D. Apresyan, I. M. Boguslavskiy, B. L. Iomdin, L. L. Iomdin, A. V. Sannikov, V. Z. Sannikov, V. G. Sizov, L. L. Tsinman. “A syntactically and semantically tagged corpus of Russian: state of the art and prospects”, *Natsional’nyy korpus russkogo yazyka: 2003–2005. Rezul’taty i perspektivy*, Indrik, M., 2005, pp. 193–214 (in Russian). ⁴⁴⁶
- [9] J. Nivre, I. M. Boguslavskii, L. L. Iomdin. “Parsing the SynTagRus Treebank of Russian”, *Proceedings of the 22nd International Conference on Computational Linguistics*, Coling 2008 (August 2008, Manchester, UK), 2008, pp. 641–648. ⁴⁴⁶
- [10] A. V. Sokirko. “Morphological modules on the web”, *Komp’yuternaya lingvistika i intellektual’nyye tekhnologii*, Trudy mezhdunarodnogo seminara “Dialog’2004”, 2004 (in Russian), 7 pp. ⁴⁴⁷
- [11] I. Boguslavsky. “SynTagRus — a deeply annotated corpus of Russian”, *Les motions dans le discours — Emotions in Discourse*, eds. P. Blumenthal, I. Novakova, D. Siepmann, Peter Lang, 2014, pp. 367–380. ⁴⁵⁰
- [12] J. Nivre, M.-C. de Marneffe, F. Ginter, Y. Goldberg, J. Hajič, Ch. D. Manning, R. McDonald, S. Petrov, S. Pyysalo, N. Silveira, R. Tsarfaty, D. Zeman. “Universal Dependencies v1: A multilingual treebank collection”, *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, LREC 2016, 2016, pp. 1659–1666. ⁴⁵⁰
- [13] A. Kutuzov, I. Andreev. “Texts in, meaning out: neural language models in semantic similarity task for Russian”, *Computational Linguistics and Intellectual Technologies*, Papers from the Annual conference “Dialogue”, RGGU, M., 2015, pp. 133–144. ⁴⁵¹

Sample citation of this publication:

Igor Trofimov, Elena Suleymanova. “A dependency-based distributional semantic model for identifying taxonomic similarity”. *Program Systems: Theory and Applications*, 2018, **9**:4(39), pp. 443–460. (In Russian).

10.25209/2079-3316-2018-9-4-443-460

http://psta.psisras.ru/read/psta2018_4_443-460.pdf