Сравнительный анализ ассоциаций в корпусах социальных сетей на основе дистрибутивно-семантических моделей для русского языка

Анна А. Антипенко, Ольга А. Митрофанова

***Аннотация*— Данная статья отражает результаты эксперимента по автоматическому извлечению ассоциативных связей из корпусов русскоязычных текстов социальных сетей Facebook и Pikabu с помощью алгоритмов и инструментов дистрибутивной семантики. Выбор материала социальных сетей обуславливается спецификой полилогического интернет-дискурса, совмещающего черты письменной и устной разговорной речи. Была высказана гипотеза о возможности воспроизведения методики ассоциативного эксперимента при работе с корпусами текстов социальных сетей на основе дистрибутивно-семантических моделей. Для лексем, выражающих ключевые понятия русскоязычной картины мира, автоматически извлечены ассоциаты с использованием нейросетевых архитектур Word2Vec (CBOW и Skip-gram). Был проведен сопоставительный анализ полученных данных и данных Русского ассоциативного словаря, Русской региональной ассоциативной базы данных (Сибирь и Дальний Восток) и Русского дистрибутивного тезауруса. Была разработана и реализована методика количественной оценки соответствий между результатами, полученными из разных источников. Была подтверждена специализация используемых словарных источников и дистрибутивно-семантических моделей в отношении парадигматических и синтагматических связей. Экспериментальные данные позволили провести лингвистический анализ языкового сознания современных пользователей социальных сетей и выявить тенденции в динамике его развития.**

***Ключевые слова*— дистрибутивная семантика, word2vec, ассоциативный эксперимент, социальные сети, русский язык**

I. Введение

Традиционные исследования языкового сознания в психолингвистике основываются на свободном ассоциативном эксперименте, при котором группе испытуемых предъявляются слова-стимулы и фиксируются реакции. Полученные данные используются для создания ассоциативных словарей, которые представляют собой отражение языковой картины мира «стандартного» носителя языка.

Как пишет Н.В. Уфимцева, по ассоциативным реакциям «можно судить о речевой синтагматике» [Уфимцева, 2011: 229]: велика вероятность совместной встречаемости ассоциатов в потоке речи. Можно предположить, что слова схожей семантики, встречающиеся в одних и тех же контекстах, связаны в языковом сознании человека, как и ассоциаты, а связи между ними сопоставимы со связями между стимулом и реакцией при ассоциативном эксперименте. Существующие инструменты анализа естественного языка позволяют выявлять семантически близкие слова, основываясь на векторных представлениях слов в дистрибутивно-семантических моделях.

Целью нашего исследования стало автоматическое выделение и сопоставление ассоциативных связей ключевых концептов русскоязычной картины мира в корпусах постов социальных сетей Facebook и Pikabu. Наше решение основано на методах корпусной лингвистики, психолингвистики и дистрибутивной семантики.

Главной особенностью нашей работы является адаптация методологии очной работы с испытуемыми к условиям исследования текстового общения носителей русского языка в социальных сетях, в получении новых данных о динамике языкового сознания носителей русского языка, основанных на корпусах Facebook и Pikabu, в верификации этих данных по данным лексикографических источников.

II. Методологические основания исследования

A. Дистрибутивно-семантические модели

Один из наиболее распространенных подходов к моделированию лексического значения слова и семантических отношений – это построение многомерных векторов в дистрибутивно-семантических моделях, отражающих значения слов, на основании совместной встречаемости в больших корпусах текстов [Baroni, Lenci 2010]. Векторные представления слов представляют интерес с точки зрения когнитивного моделирования, поскольку они обладают сходством с наборами нейронов [Rohde, Gonnerman, Plaut 2006]. С помощью векторных моделей решаются такие задачи, как извлечение отношений и фактов, идентификация конструкций, назначение семантических ролей, более

Статья получена 16 декабря 2019 года. Данная статья основана на материалах доклада авторов в рамках конференции «Интернет и современное общество 2019».

А.А. Антипенко – Just AI, Санкт-Петербург, Россия (e-mail: anne.morke@gmail.com).

О.А. Митрофанова – Санкт-Петербургский государственный университет, Санкт-Петербург, Россия (e-mail: o.mitrofanova@spbu.ru).

того, это наиболее распространенный метод вычисления семантической близости слов [Jurafsky, Martin 2017].

В дистрибутивной семантике разрабатываются как синтагматические, так и парадигматические модели [Sahlgren 2008]. Одним из важных параметров, который необходимо учитывать в таких моделях, является размер контекстного окна. Чем уже контекстное окно, тем вероятнее, что в нем будут регистрироваться синтагматические связи целевого слова. С расширением окна можно ожидать перехода к выявлению разнообразных парадигматических отношений.

На сегодняшний день самым востребованным типом дистрибутивно-семантических моделей является Word2Vec [https://code.google.com/archive/p/Word2Vec/] – нейросетевая архитектура, позволяющая получать векторные представления слов в многомерном векторном пространстве. Векторы слов расположены в этом пространстве таким образом, что слова, встречающиеся в одних и тех же контекстах, находятся на близком расстоянии.

Word2Vec использует два алгоритма обучения: Continious Bag of Words (CBOW), предсказывающий слово по его окружению, и Skip-gram, предугадывающий по слову его окружение. Архитектура CBOW работает быстрее, однако модели, основанные на архитектуре Skip-gram, позволяют получить более точный результат, особенно для редких слов. Алгоритмы CBOW и Skip-gram были предложены Т. Миколовым и коллегами [Mikolov, Chen, Corrado, Dean 2013]. Данные алгоритмы нацелены на минимизирование вычислительной сложности.

Существуют модели типа Word2Vec для русского языка, созданные в рамках проекта RusVectōrēs [https://rusvectores.org/ru/] на основе разнообразных корпусных источников (НКРЯ, Википедия, новостной корпус и т.д.) [Kutuzov, Kuzmenko 2017]. Наряду с этим, возможно построение аналогичных моделей на специализированных корпусах, семантические отношения в которых являются предметом экспериментального исследования, что и было осуществлено в нашей работе.

B. Оценка дистрибутивно-семантических моделей

Справедливо заметить, что сами дистрибутивно-семантические модели, которые в нашем случае используются для выявления ассоциативных отношений, подлежат качественной и количественной оценке. Существует несколько способов верификации моделей [Bakarov 2019].

С одной стороны, разработаны процедуры для проверки качества обучения дистрибутивно-семантических моделей и их способности к предсказанию того или иного типа отношений. Одним из таких тестов является псевдо-дизамбигуация [Panicheva, Protopopova, Bukia, Mitrofanova 2017].

С другой стороны, для оценки дистрибутивно-семантических моделей широко используются данные тезаурусов, среди которых первое место занимают тезаурусы типа WordNet; целесообразно использование тестов на определение регулярных ассоциативных отношений типа TOEFL [Rohde, Gonnerman, Plaut 2006]. В качестве альтернативного источника можно предложить использование ассоциативных словарей.

C. Ассоциативный эксперимент и ассоциативные словари

Наше исследование предполагает автоматизированную процедуру, воспроизводящую схему ассоциативного эксперимента, где в качестве стимулов рассматриваются слова, реализующие основные концепты русскоязычной картины мира, а реакции автоматически извлекаются из дистрибутивно-семантических моделей корпусов текстов социальных сетей [Кольцов, Кольцова, Митрофанова, Шиморина 2014]. В данном случае ассоциативные словари могут рассматриваться как источники эталонных данных об ассоциативных реакциях носителей русского языка.

Русский ассоциативный словарь (РАС) [РАС 2014–2016] представляет собой ассоциативно-вербальную сеть. При создании РАС ассоциативный эксперимент проводился среди студентов-носителей русского языка в возрасте от 17 до 25 лет. В исходный набор стимулов вошли 1277 единиц, взятых из следующих источников: «Словарь ассоциативных норм русского языка» [САНРЯ 1977], «Частотный словарь русского языка» [ЧСРЯ 1977], «Русский семантический словарь» [РСС 1983], для некоторых слов из списка были подобраны синонимы и антонимы, образованы дериваты, а также в список были включены некоторые предлоги, частицы, междометия, числительные, союзы, местоимения для наиболее системного отражения русской лексики.

Русская региональная ассоциативная база данных (Сибирь и Дальний Восток) [СИБАС 2008 – 2018] является результатом комплексных работ по изучению языкового сознания современных русскоговорящих. Проект был реализован в Институте филологии СО РАН и НГУ в сотрудничестве с ИЯз РАН и рядом академических институтов и вузов России. СИБАС объединяет в себе основной корпус и подкорпусы: подкорпус вербальных ассоциаций носителей русского языка в Казахстане (РКАС) и подкорпус вербальных ассоциаций военных (ПВАС), которые были подготовлены в период с 2015 по 2018 гг. Ресурс объединяет данные свыше 5000 анкет по 100 стимулов в каждой, общее число стимулов превышает 1000 единиц.

Русский дистрибутивный тезаурус (РДТ) представляет собой первый свободно доступный дистрибутивный тезаурус русского языка [https://nlpub.ru/Russian_Distributional_Thesaurus]. РДТ основан на модели Skip-gram (Word2Vec), обученной на корпусе книг, собранных в электронной библиотеке lib.rus.ec, объемом 12.9 млрд с/у. Были созданы модели с различными параметрами, их точность оценивалась вручную на основе краудсорсинга. В соответствии с ручной оценкой, точность лучшей модели составляет 97.1% для первых пяти соседей и 91.2% для первых двадцати соседей. РДТ представляет собой граф подобия слов, который содержит почти 932 тыс. входов, 4,5 млн выходов и 194 млн семантических отношений. Полученные данные находятся в свободном доступе и могут быть использованы для различных лингвистических исследований [Panchenko 2017].

III. Лингвистические данные

Материалом данного исследования является корпус текстов русскоязычного сегмента социальной сети Facebook и корпус русскоязычного сегмента социальной сети Pikabu, оба корпуса без метаданных о пользователях. Объем необработанных данных составил 28953525 и 6361954 с/у соответственно.. Лемматизация корпусов производилась морфоанализатором MyStem [https://tech.yandex.ru/mystem/].

В предобработке был использован словарь стоп-слов объемом 1104 слова, составленный на основе словарей служебной лексики и оборотов НКРЯ [Митрофанова 2015]. После удаления стоп-слов объемы корпусов существенно сократились и составили 15552981 и 1304696 с/у соответственно.

Сначала было принято решение отбирать слова на основании ключевых концептов, составляющих ядро языкового сознания носителей русского языка, выделенных Н.В. Уфимцевой [Уфимцева 2009]. Однако наиболее важным критерием отбора было наличие отобранных лексем во всех рассматриваемых источниках.

Был составлен частотный список слов в корпусах социальных сетей, после предварительной обработки и отобрана первая тысяча лексем. Для слов из начала списка проверялась частота по «Частотному словарю современного русского языка» [Ляшевская, Шаров 2009]. Частота в данном словаре указана в ipm (instances per million – количество употреблений на миллион слов корпуса). Для двух отобранных слов – имен собственных *Россия* и *Москва* – данные о частоте в частотном словаре отсутствуют, но их можно восстановить по Национальному корпусу русского языка (НКРЯ). В НКРЯ есть данные об абсолютных частотах, чтобы получить частоту в ipm, нужно разделить абсолютные частоты на 92 – сумма орфографических слов корпуса, принятая за единицу вычисления ipm в словаре ЧССРЯ. Таким образом, получается, что все отобранные лексемы являются достаточно частотными, диапазон частот в ipm – от 89,5 (*находить*) до 3727,5 (*год*), средняя частота составляет 618. Также проверялось наличие всех лексем из списка в РАС и РДТ, были оставлены те лексемы, которые есть во всех источниках.

В итоге для ассоциативного эксперимента был составлен список из 96 слов:

1) 61 существительное (*человек, год, день, жизнь, друг, время, Россия, мир, ребенок, слово, дело, работа, страна, место, дом, город, любовь, рука, вопрос, женщина, душа, бог, утро, глаз, сила, народ, Москва, история, деньги, земля, война, ночь, голова, сердце, власть, лицо, конец, сторона, час, свет, мама, праздник, случай, вода, мужчина, отношение, часть, фильм, книга, семья, проблема, школа, вечер, путь, имя, право, результат, мысль, дорога, сын, язык*),
2) 29 глаголов (*знать, говорить, хотеть, давать, понимать, любить, жить, думать, видеть, начинать, оставаться, приходить, иметь, смотреть, получать, находить, проходить, принимать, написать, ждать, пойти, стоять, помогать, решать, читать, отвечать, помнить, сидеть, называть*),
3) 5 прилагательных (*новый, хороший, большой, добрый, нужный*),
4) 1 наречие (*сегодня*).

IV. Извлечение ассоциатов методами дистрибутивной семантики

A. Инструментарий

В данной работе использовалась реализация Word2Vec в библиотеке gensim [https://radimrehurek.com/gensim/]. Программа обработки ассоциативных отношений была реализована на языке программирования Python в среде разработки Spyder. Для извлечения ассоциатов было создано две модели при помощи инструмента Word2Vec: CBOW (Continious Bag of Words) и Skip-gram. При создании обеих моделей игнорировались слова, которые встретились в корпусе менее пяти раз. Размерность векторов составила 100, поскольку, несмотря на то, что большие значения могут давать более точный результат, для них требуется очень большой объем исходных данных. Использовалось контекстное окно размера 5, что обусловлено объемом корпуса, а также характером текстов – для текстов социальных сетей характерно изложение мыслей в краткой, сжатой форме.

Word2Vec позволяет сохранять созданные модели и использовать их для таких задач, как оценка сходства между словами в запросе, исключение лишнего слова из запроса, поиск наиболее близких по значению слов. В качестве запроса использовались слова, отобранные ранее, извлекались первые 20 близких слов. Таковы, например, ассоциаты, полученные для стимула *свет*, в моделях CBOW и Skip-gram:

Свет CBOW: *неучение, тьма, луч, лучик, светить, гаснуть, сиять, освещать, погасать, темнота, свеча, кромешный, солнце, озарять, тень, фонарь, зеркальце, лампа, зажигаться, мрак*;

Свет Skip-gram: *негасимый, коптилка, освещать, луч, неучение, зажигаться, тьма, озарять, мерцание, сживать, неяркий, гаснуть, тень, озаряться, темнота, тускнеть, небосклон, светляк, мглистый, непроглядный*.

B. Статистический и лингвистический анализ экспериментальных данных

Для анализа результатов все полученные данные (близкие слова из моделей CBOW и Skip-gram для двух корпусов, а также первые 20 слов-ассоциатов из РАС, СИБАС и РДТ) были собраны в сравнительную таблицу, в которой производился подсчет совпадений между ассоциатами из различных источников. Наборы слов сравнивались попарно. Для всех лексем была проведена оценка точных совпадений, также были вычислены коэффициенты, отражающие степень сходства между двумя наборами слов, для первых пятидесяти лексем.

Мы использовали следующую схему расчета весовых коэффициентов в парах «стимул – реакция»:

1) точные совпадения – 5;

2) дериваты (не являющиеся антонимами) – 4;
3) синонимы – 3;
4) гипонимы, гиперонимы, согипонимы, холонимы, меронимы (слова, связанные отношениями род/вид либо часть/целое) – 2;
5) антонимы – 1.

Было подсчитано среднее число точных совпадений (ТС1 – диапазон значений 1,7…3), средний процент точных совпадений (ТС2 – диапазон значений 8,3…14,6). Те совпадения, которые были зарегистрированы, отражают наиболее устойчивые синтагматические связи, как правило, отражающие регулярную лексическую сочетаемость. Учитывая условия, описанные выше, а также наличие довольно сильных связей при учете парадигматических отношений, можно сделать вывод, что совпадения не случайны.

Наибольшее число совпадений – 20−30% – наблюдается для лексем *сын* (*дочь, отец, мать, старший, блудный, брат*), *результат* (*итог, оценка, анализ, положительный*), *сидеть* (*стоять, лежать, диван, кресло*), *нужный* (*необходимый, важный, ненужный, полезный*).

Для значительной части лексем обнаруживается совпадение на 10–15% (*время, знать, говорить, работа, жить, страна, город, оставаться, душа, война, сердце, лицо, стоять*, *человек, друг, хороший, сегодня, ребенок, слово, большой, любовь, женщина, бог, утро, сила, история* и др.).

В то же время, наблюдаемые различия между данными ассоциативных словарей и корпусов социальных сетей свидетельствуют о динамике языкового сознания русскоговорящих.

Было выявлено 14 пар «стимул – реакция», зарегистрированных во всех четырех источниках: *год – месяц, Россия – страна, хороший – плохой, страна – Россия, большой – огромный, женщина – мужчина, смотреть – глядеть, ночь – утро, стоять – сидеть, конец – начало, час – минута, отвечать – спрашивать, мама – папа, мужчина – женщина.*

Можно предположить, что эти ассоциативные связи находятся в ядре языкового сознания среднего носителя русского языка, так как при эксперименте использовались данные разноплановых источников: два ассоциативных словаря и два корпуса текстов социальных сетей. В то же время, возможность синхронизации данных из этих источников свидетельствует о совместимости двух исследовательских процедур – ассоциативного эксперимента и автоматического извлечения ассоциативных отношений на основе дистрибутивно-семантических моделей.

О динамике языкового сознания свидетельствует наличие в данных корпусов социальных сетей Facebook и Pikabu новой лексики, обусловленное изменениями, происходящими в языке, и возникновением новых реалий и понятий: *искать – авито, получать – шенген, деньги – ипотека, друг – роднуля, работа – офис, клиент, дом – съемный, город – мегаполис, вопрос – респондент, получать – перечислять, принимать – пролонгация, написать – эссе, власть – оппозиция, олигарх, коррупционер, фильм – трейлер, язык – субтитр*.

V. Заключение

В данной работе было проведено исследование по автоматическому извлечению ассоциативных связей для слов, выражающих ключевые концепты русскоязычной картины мира, из корпусов текстов социальных сетей Facebook и Pikabu. Для извлечения близких по значению слов использовалась нейросетевая архитектура Word2Vec. Было реализовано два варианта алгоритма Continious Bag of Words и Skip-gram. Полученные данные сравнивались с данными РАС, СИБАС и РДТ.

Было выявлено существенное расхождение данных, полученных из различных источников, наряду с этим, зарегистрированы совпадения, являющиеся информативными. В условиях эксперимента можно считать, что расхождения обусловлены не только методологическими различиями, но и динамикой языкового сознания носителей русского языка.

В ходе исследования были получены и обработаны эмпирические данные, свидетельствующие о динамике языкового сознания носителей русского языка и о ядре языкового сознания современных пользователей социальных сетей.

Библиография

[1]Bakarov A. A Survey of Word Embeddings Evaluation Methods URL: https://www.academia.edu/35845049/A_Survey_of_Word_Embeddings_Evaluation_Methods

[2]Baroni M., Lenci A. Distributional Memory: A General Framework for Corpus-Based Semantics // Computational Linguistics. Vol. 36(4). 2010. P. 673–721.

[3] Jurafsky D., Martin H. Speech and Language Processing (Third Edition Draft). 2017. URL: https://web.stanford.edu/~jurafsky/slp3/

[4]Kutuzov A., Kuzmenko E. WebVectors: A Toolkit for Building Web Interfaces for Vector Semantic Models // D. Ignatov et al. (eds.) Analysis of Images, Social Networks and Texts. AIST 2016. Communications in Computer and Information Science. Vol. 661. Springer, Cham. 2017.

[5]Mikolov T., Chen K., Corrado G., Dean J. Efficient Estimation of Word Representations in Vector Space // Proceedings of Workshop at ICLR. 2013a.

[6] Panchenko A. et al. Human and Machine Judgements for Russian Semantic Relatedness // D. Ignatov et al. (eds.) Analysis of Images, Social Networks and Texts: AIST–2016. Communications in Computer and Information Science. Vol. 661. Springer, Cham, 2017.

[7]Panicheva P., Protopopova E., Bukia G., Mitrofanova O. Evaluating Distributional Semantic Models with Russian Noun-Adjective Compositions // Communications in Computer and Information Science (CCIS). vol. 661. Analysis of Images, Social Networks and Texts: 5th International Conference, AIST 2016, Yekaterinburg, Russia, April 7–9, 2016, Revised

Selected Papers. Springer, Cham, 2017. P. 236–247.

[8] Panicheva P., Erofeeva A., Ledovaya Ja. Semantic Feature Aggregation for Gender Identification in Russian Facebook // Artificial Intelligence and Natural Language 6th Conference, AINL 2017, St. Petersburg, Russia, September 20–23, 2017, Revised Selected Papers. Communications in Computer and Information Science. Vol. 789. Springer, 2017. P. 3–15.

[9] Rohde D., Gonnerman L., Plaut D. An Improved Model of Semantic Similarity Based on Lexical Co-occurrence // Communications of the ACM. № 8. 2006. P. 627–633.

[10] Sahlgren M. The Distributional Hypothesis. From Context to Leaning // Distributional Models of the Lexicon in Linguistics and Cognitive Science (Special Issue of the Italian Journal of Linguistics). Rivista di Linguistica. 2008. Vol. 20(1). P. 33−53.

[11] Кольцов С.Н., Кольцова О.Ю., Митрофанова О.А., Шиморина А.С. Интерпретация семантических связей в текстах русскоязычного сегмента Живого Журнала на основе тематической модели LDA // Технологии информационного общества в науке, образовании и культуре: сборник научных статей. Материалы XVII Всероссийской объединенной конференции «Интернет и современное общество» IMS–2014, Санкт-Петербург, 19–20 ноября 2014 г. СПб., 2014. С. 135–142.

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

[13] Митрофанова О.А. Вероятностное моделирование тематики русскоязычных корпусов текстов с использованием компьютерного инструмента GenSim // Труды международной конференции «Корпусная лингвистика−2015». СПб.: Издательство Санкт-Петербургского университета, 2015. С. 332–343.

[14] РАС – Русский ассоциативный словарь: в 4 т. / Ю.Н. Караулов [и др.]. М., 1994–1996. Т. 1.

[15] Русский семантический словарь: Опыт автоматического построения тезауруса: от понятия к слову / Ю.Н. Караулов [и др.]. М.: Наука, 1983.

[16] САНРЯ – Словарь ассоциативных норм русского языка / А.А. Леонтьев [и др.]. М.: Издательство Московского университета, 1977.

[17] СИБАС – Русская региональная ассоциативная база данных (2008 – 2018) (авторы-составители И.В.Шапошникова, А.А.Романенко) URL:http://adictru.nsu.ru

[18] Уфимцева Н.В. Образ мира русских: системность и содержание // Язык и культура. М., 2009. С. 98–111.

[19] Уфимцева Н.В. Языковое сознание: динамика и вариативность. М.: Институт языкознания РАН, 2011.

[20] ЧСРЯ – Частотный словарь русского языка / под ред. Л.Н. Засориной. М.: Русский язык, 1977.

Comparative Study of Word Associations in Social Networks Corpora by means of Distributional Semantics Models for Russian

Anna A. Antipenko, Olga A. Mitrofanova

Abstract— **The paper discusses results of the experiment on automatic extraction of associative relations from corpora of Russian texts from Facebook and Pikabu social networks by means of distributional semantic models. The choice of linguistic data for analysis, namely, social networks texts, is determined by the specificity of polylogic internet-discourse which combines traits of written and colloquial speech. We put forward the hypothesis on the possibility of reproduction of associative test technique in the experiments with distributional semantic models. Experiments were carried out with the help of algorithms and tools of Distributional Semantics. We extracted associations for lexemes expressing key concepts of Russian-specific world view. The procedure was performed by means of Word2Vec (CBOW and Skip-gram) neural network architectures. We carried out linguistic analysis of the output data and compared it with the associations described in the Russian Associative Dictionary, Russian regional association database (Siberia and Fare East) and the Russian Distributional Thesaurus. Results achieved in course of experiments allow to make conclusions on the dynamic of Russian-specific language consciousness of contemporary social network users. We worked out and implemented the procedure of quantitative evaluation of data extracted from different sources. We found evidence on the specialization of lexicographic resources and distributional semantic models as regards paradigmatic and syntagmatic relations. Experimental data allowed to carry out linguistic analysis of contemporary Russian-specific world view of social networks users and to reveal tendencies in its development.**

Keywords— **Distributional Semantics, Corpus Linguistics, social networks, associative experiment, language consciousness, Russian**

Anna Andreevna Antipenko
BA in Computational Linguistics obtained at the Department of Mathematical Linguistics, Faculty of Philology, Saint-Petersburg State University, Russia (https://spbu.ru/)
Linguist Researcher, JustAI (https://just-ai.com/)
e-mail: anne.morke@gmail.com

Olga Alexandrovna Mitrofanova
PhD, Associate Professor, Department of Mathematical Linguistics, Faculty of Philology, Saint-Petersburg State University, Russia (https://spbu.ru/)
e-mail: o.mitrofanova@spbu.ru
elibrary: authorid=4169-6068
scopus.com: authorId=36932407200
ORCID: orcidid=0000-0002-3008-5514

References

[1] Ufimtseva N.V. Jazykovoje soznanije: dinamika i variativnost. M.: Institut jazykoznanija RAN, 2011.
[2] Koltsov S.N., Koltsova O.Ju., Mitrofanova O.A., Shimorina A.S. Interpretatsija semanticheskih svyazej v tekstah russkojazychnogo segmenta Jzivogo Zhurnala na osnove tematicheskoj modeli LDA // Tehnologii informatsionnogo obschestva v nauke, obrazovanii i culture: sbornik nauchnyh statej. Materialy XVII Vserossijskoj objedinennoj konfrencii «Internet i sovremennoje obschestvo» IMS–2014, Sankt-Peterburg, 19–20 nojabrya 2014 r. SPb, 2014. S. 135–142.
[3] Slovar assotsiativnyh norm russkogo jazyka / A.A. Leontjev [i dr.]. M.: Izdatelstvo Moskovskogo universiteta, 1977.
[4] Russkij assotsiativnyj slovar: v 4 t. / Ju.N. Karaulov [i dr.]. M., 1994–1996. T. 1.
[5] Chastotnyj slovar ruskogo jazyka / pod red. L.N. Zasorinoj. M.: Russkij jazyk, 1977.
[6] Ruskij semanticheskij slovar: Opyt avtomaticheskogo postrojenija tezaurusa: ot ponyatija k slovu / Ju.N. Karaulov [i dr.]. M., Nauka, 1983.
[7] Panchenko A. et al. Human and Machine Judgements for Russian Semantic Relatedness // D. Ignatov et al. (eds.) Analysis of Images, Social Networks and Texts: AIST–2016. Communications in Computer and Information Science. Vol. 661. Springer, Cham, 2017.
[8] Baroni M., Lenci A. Distributional Memory: A General Framework for Corpus-Based Semantics // Computational Linguistics. Vol. 36(4). 2010. P. 673–721.
[9] Rohde D., Gonnerman L., Plaut D. An Improved Model of Semantic Similarity Based on Lexical Co-occurrence // Communications of the ACM. № 8. 2006. P. 627–633.
[10] Jurafsky D., Martin H. Speech and Language Processing (Third Edition Draft). 2017. URL: https://web.stanford.edu/~jurafsky/slp3/
[11] Sahlgren M. The Distributional Hypothesis. From Context to Leaning // Distributional Models of the Lexicon in Linguistics and Cognitive Science (Special Issue of the Italian Journal of Linguistics). Rivista di Linguistica. 2008. Vol. 20(1). P. 33−53.
[12] Mikolov T., Chen K., Corrado G., Dean J. Efficient Estimation of Word Representations in Vector Space // Proceedings of Workshop at ICLR. 2013.

[13] Kutuzov A., Kuzmenko E. WebVectors: A Toolkit for Building Web Interfaces for Vector Semantic Models // D. Ignatov et al. (eds.) Analysis of Images, Social Networks and Texts. AIST 2016. Communications in Computer and Information Science. Vol. 661. Springer, Cham. 2017.

[14] Panicheva P., Erofeeva A., Ledovaya Ja. Semantic Feature Aggregation for Gender Identification in Russian Facebook // Artificial Intelligence and Natural Language 6th Conference, AINL 2017, St. Petersburg, Russia, September 20–23, 2017, Revised Selected Papers. Communications in Computer and Information Science. Vol. 789. Springer, 2017. P. 3–15.

[15] Panicheva P., Protopopova E., Bukia G., Mitrofanova O. Evaluating Distributional Semantic Models with Russian Noun-Adjective Compositions // Communications in Computer and Information Science (CCIS). vol. 661. Analysis of Images, Social Networks and Texts: 5th International Conference, AIST 2016, Yekaterinburg, Russia, April 7–9, 2016, Revised Selected Papers. Springer, Cham, 2017. P. 236–247.

[16] Mitrofanova O.A. Verojatnostnoje modelirovanije tematiki russkojazychnyh korpusov tekstov s ispolzovanijem kompjuternogo instrumenta GenSim // Trudy mezhdunarodnoj konferencii «Korpusnaja lingvistika−2015». SPb.: Izdatelstvo Sankt-Peterburgskogo universiteta, 2015. S. 332–343.

[17] Ufimtseva N.V. Obraz mira russkih: sistemnost i soderzhanije // Jazyk i kultura. M., 2009. S. 98–111.

[18] Lyashevskaja O.N., Sharov S.A. Chastotnyj slovar sovremennogo russkogo jazyka (na materialah Natsionalnogo korpusa russkogo jazyka). M.: Azbukovnik, 2009.