Методика создания семантических словарей: опросы респондентов и машинное обучение

В.Д. Соловьев

Казанский федеральный университет

Maki.solovyev@mail.ru

Аннотация

В докладе описывается технология создания словарей с рейтингами конкретности/ абстрактности и позитивности/негативности слов. Проблематика конкретности/ абстрактности является одной из ключевых в области репрезентации знаний и представляет серьезный вызов когнитивной науке [1]. Оценка слов как позитивных/негативных важна в такой хорошо известной прикладной задаче как сентимент-анализ. В обоих случаях соответствующие шкалы считаются непрерывными и требуются большие словари с рейтингами слов.

Ключевые слова: словари русского языка, опросы носителей языка, машинное обучение, конкретность, абстрактность, позитивность, негативность

Библиографическая ссылка: Соловьев В.Д. Методика создания семантических словарей: опросы респондентов и машинное обучение // Компьютерная лингвистика и вычислительные онтологии. Выпуск 6 (Труды XXV Международной объединенной научной конференции «Интернет и современное общество», IMS-2022, Санкт-Петербург, 23 – 24 июня 2022 г. Сборник научных статей). — СПб.: Университет ИТМО, 2022. С. 09-11. DOI: 10.17586/2541-9781-2022-6-09-11

В Казанском университете созданы словари обоих видов для русского языка и словарь с рейтингами конкретности/абстрактности для английского языка. Для их создания применены две технологии. Первая – опрос респондентов. Для русского языка словари обоих видов содержат по 1000 наиболее частотных слов. Для каждого слова получено не менее 40 оценок, что превышает аналогичный показатель основных словарей для других языков. Особое внимание было уделено контролю качества ответов респондентов. Система из 5 фильтров оценивала результаты и оценки недобросовестных респондентов удалялись. Подробное описание этого подхода можно найти в [2].

Словарь с рейтингами конкретности/абстрактности подробно проанализирован. Показано, что оценки не сильно зависят от возраста и уровня образования респондентов. Такое исследование проведено впервые в мире. Оценки респондентов находятся в хорошем согласии с данными Семантического словаря русского языка Н.Ю. Шведовой. Расхождения составляют менее 1,5%. Установлена степень эффективности морфологического критерия абстрактности, состоящего в том, что слова с определенными аффиксами (-изм и т.д.) являются абстрактными. Проведено сопоставление рейтингов слов русского языка и их переводных эквивалентов в английском языке. Коэффициент корреляции по Спирмену между ними весьма высокий – 0,78. Однако для некоторых слов различия в рейтингах слов весьма значительны. Например, слово *администрация* носителями русского языка трактуется как весьма конкретное. Видимо за этим понятием стоит некая типовая структура – определенная, известная человеку администрация, люди, с которыми приходилось общаться. А английское *administration* носители языка воспринимают как абстрактное. Так что имеют место и культурно-специфические феномены. Результаты изложены в [2].

Второй подход состоит в машинной экстраполяции оценок респондентов на большее количество слов. С применением наиболее современных методов глубокого обучения (модель нейронных сетей RuBERT) получены словари русского языка обоих видов, содержащие более 20 тыс. слов. Эта технология применена также для английского языка, в результате получен словарь рекордного качества – коэффициент корреляции по Спирмену с оценками респондентов составил 0,92. Лучший ранее опубликованный результат – 0,90. Подробное описание наших результатов по машинным словарям можно найти в [3]. Важный результат, приведенный в указанной работе, состоит в том, что для дообучения модели BERT не требуется много обучающих данных. При использовании для обучения всего 1000 слов уже можно получить словарь на 40 000 слов достаточно высокого качества.

Созданные словари использовались для проведения конкретных исследований. Одним из возможностей применения словаря конкретности/абстрактности является определение сложности текстов, что является важной задаче в сфере образования. Нами создана коллекция школьных учебников, и для каждого из них подсчитана доля абстрактных слов (по нашему словарю) в текстах. Как и следовало ожидать, оказалось, что чем старше класс, тем больше доля абстрактных слов в учебниках этого класса [2]. Таким образом, доля абстрактных слов может служить надежным предиктором сложности текстов.

Созданный нами словарь с рейтингами позитивности/негативности может использоваться различным образом. Одна возможность – пополнение аналогичных ранее созданных для сентимент-анализа словарей. Например, найдено значительное число слов, заведомо позитивных или негативных, которые отсутствуют в известном словаре Н.В. Лукашевич RuSentiLex (https://www.labinform.ru/pub/rusentilex/index.htm).

Другой пример использования – проверка гипотезы Поллианны о преимущественном использовании позитивной лексики в языке. Нами эта гипотеза подтверждена для русского языка как на материале словаря, так и материале корпусов текстов различных жанров, взятых из Национального корпуса русского языка.

Литература

[1] Соловьев В. Концепция конкретности/абстрактности: современное состояние // Достижения в области интеллектуальных систем и вычислений. 2021. Т. 1358. С. 275-283.

[2] Соловьев В., Вольская Ю., Андреева М., Заикин А. Словарь русского языка с индексами конкретности/абстрактности // Русское языкознание. 2022. Т. 26, № 2. С. 515–549.

[3] Соловьев В., Иванов В. Экстраполяция человеческих оценок конкретности/абстрактности слов нейронными сетями различной архитектуры // Прикладные науки (Швейцария). 2022. Т. 12. № 9. Ст. № 4750.

Methodology for Creating Semantic Dictionaries: Surveys of Respondents and Machine Learning

V. Solovyev

Kazan Federal University

The report describes the technology of creating dictionaries with ratings of concreteness/abstractness and positivity/negativity of words. The issue of concreteness/abstractness is one of the key issues in the field of knowledge representation and represents a serious challenge to cognitive science [1]. The assessment of words as positive/negative is important in such a well-known applied task as sentiment analysis.

In both cases, the corresponding scales are considered continuous and large dictionaries with word ratings are required.

Keywords: Russian language dictionaries, surveys of native speakers, machine learning, concreteness, abstractness, positivity, negativity

Reference for citation: Solovyev V.D. Methodology for creating semantic dictionaries: surveys of respondents and machine learning // Computational Linguistics and Computational Ontologies. Issue 6 (Proceedings of the XXV International Joint Scientific Conference "Internet and Modern Society", IMS-2022, St. Petersburg, June 23 - 24, 2022). - St. Petersburg: ITMO University, 2022. C. 09-11. DOI: 10.17586/2541-9781-2022-6-09-11

Reference

[1] Solovyev V. Concreteness/Abstractness Concept: State of the Art // Advances in Intelligent Systems and Computing. 2021. Vol.1358. P. 275-283.
[2] Solovyev V., Volskaya Yu., Andreeva M., Zaikin A. Russian dictionary with concreteness/ abstractness indices // Russian Journal of Linguistics. 2022. Vol. 26. № 2. P. 515–549.
[3] Solovyev V., Ivanov V. Extrapolation of Human Estimates of the Concreteness/Abstractness of Words by Neural Networks of Various Architectures // Applied Sciences (Switzerland). 2022. Vol. 12, № 9. Art. № 4750.