УДК 159.9 **DOI: 10.34670/AR.2023.96.87.032**

Алгоритм эмоциональной разметки прилагательных русского языка

Мещанкин Андрей Вячеславович

Соискатель,
Институт психологии, социологии и социальных отношений,
Московский городской педагогический университет,
123022, Российская Федерация, Москва,
2-й Сельскохозяйственный проезд, 4;
e-mail: meschankin.andrey@physics.msu.ru

Аннотация

Основой психолингвистического анализа текста является работа с его различными структурными единицами. Мы сосредоточили свое внимание на изучении психолингвистических характеристик слова, так как оно является основной структурной единицей языка. Востребованные подходы к психолингвистическому анализу подразумевают работу с большим набором предобработанных, размеченных единиц текста – слов. Современные подходы дистрибутивной семантики дают возможность ученым проводить исследования на стыках наук. Совместив исследования в области компьютерной лингвистики и психологии, мы предлагаем алгоритм разметки прилагательных русского языка, позволяющий автоматически получать модель, включающую в себя данные об эмоциональном окрасе прилагательных русского языка. Подобные исследования являются необходимым этапом для дальнейшего понимания связи когнитивных особенностей человека и его письменной речью. Получено 26 113 размеченных прилагательных. Нами был предложен алгоритм эмоциональной разметки прилагательных русского языка, на основе знания о семантической близости, произведена автоматическая разметка эмоциями большинства прилагательных в русском языке, проведена валидация алгоритма.

Для цитирования в научных исследованиях

Мещанкин А.В. Алгоритм эмоциональной разметки прилагательных русского языка // Психология. Историко-критические обзоры и современные исследования. 2022. Т. 11. № 6А. С. 86-93. DOI: 10.34670/AR.2023.96.87.032

Ключевые слова

Психолингвистика, эмоции, дистрибутивная семантика, корпус слов, алгоритм распределения, эмотивность, прилагательные, распределение, кодирование, контент-анализ.

Характеристики слова как психолингвистической единицы

Основой психолингвистического анализа текста является работа с его различными структурными единицами [Yadollahi et al., 2018]. Мы сосредоточили свое внимание на изучении психолингвистических характеристик слова, так как оно является основной структурной единицей языка. Для того, чтобы понять природу слова как психометрической единицы контент-анализа, несколькими авторами уже были предприняты попытки предварительной разметки набора слов различными маркерами.

Первостепенной проблемой данного подхода является отсутствие общности, так, одни и те же слова, могут обладать различной эмоциональной разметкой, в зависимости от специфики изучаемого текста. Кроме того, разметка слов является крайне трудозатратной задачей [Mohammad, 2010; Murthy, 2021].

Еще одной проблемой является понятие репрезентативной эмотивности, то есть способность слова выражать психологические (эмоциональные) состояния и переживания человека. Кроме того, последние исследования показывают, что эмоциональное состояние и эмоциональные отношения могут быть представлены в языке и быть репрезентативными не только при прямой номинации (страх, гнев, любовь), но и непосредственным выражением.

Если исходить при определении эмотивности из понятия ситуации, представляющей эмоциональное состояние субъекта, то придется признать, что существуют разнообразные средства репрезентации эмоционального состояния и отношения в различных условиях общения и зависимости от намерений говорящего.

Хотя отсутствие контекста (специфики текста) значительно усложняет работу исследователей, тем не менее работа со словом дает нам возможность выбрать наиболее подходящую для этого часть речи, где проблема. Очевидно, что наибольшей репрезентативной эмотивностью обладает прилагательные русского языка, как часть речи, обозначающая непроцессуальный признак предмета.

Согласно Б.Г. Ананьеву, письменный язык по праву считается формой с максимально возможным влиянием на внутреннюю речь. А внутренняя речь, в свою очередь, формируется непосредственно из когнитивных процессов мозга, которые, в том числе, отражают наш эмоциональный отклик [Культурноисторическая психология, 2010]. Таким образом, письменная речь передает не только смысловую информации, но также, содержит в себе косвенную информацию об эмоциях.

Исходя из вышеперечисленного, мы видим перспективным выбором именно эмоций, как характеристик прилагательного, что в дальнейшем даст исследователям возможность для комплексного психолингвистического анализа текста [Murthy, 2021].

Эмоции

Еще Аристотель определял эмоции как то, что приводит к трансформации сознания субъекта, что влияет на его суждение. Согласно исследованиям Пола Экмана, и классическим взглядом на эмоции, они вызываются подсознательным процессом оценки чего-либо, что имеет значение для данного субъекта. Также, согласно Панксеппу, эмоции характеризуются поведенческими, экспрессивными, когнитивными и физиологическими изменениями [Pober, 2018].

Согласно взглядам А.Н. Леонтьева и С.Л. Рубинштейна, а также исходя из анализа работ Л.С. Выготского, можно сказать, что в отечественной психологии существует два основных аспекта эмоций:

1) Аспект отражения – эмоции являются специфической формой отражения значимости предметов и событий действительности для субъекта. Эмоции – это особый класс психических процессов и состояний, связанный с инстинктами, потребностями и мотивами, отражающихся в форме непосредственного переживания значимости действующих на индивида явлений и ситуаций для осуществления его жизнедеятельности;
2) Аспект отношения – эмоции выражают субъективное отношение человека к миру.

Таким образом, в мировой практике до сих пор не существует единого определения эмоций, но большинство исследователей сходятся во мнении, что эмоции – некий субъективный социальный объект, информацию о котором невозможно предать каким-либо другим способом, как опытным путем. Это означает, что как субъективный объект, эмоции зависят напрямую от предыдущего опыта человека и, в тоже время, как социальный – формируются согласно общепринятому поведению в обществе. Выходит, что эмоции несут в себе понятия и как о нашей индивидуальности, и как о социальной среде, окружающей нас.

Базовые эмоции

Хотя в современной психологии нет доминирующей теории механизма возникновения эмоций, все они сходятся в едином мнении, что так или иначе существует набор слов, отражающих наиболее общие для данной культуры эмоции.

Исторически сложилось, что данный набор эмоций называются базовыми, поскольку изначально исследователи предполагали, что они являются тем базисом из которого формируются все остальные эмоции.

Биологическое происхождение базовых эмоций основываются на исследованиях систематизирующих поведение животных и предполагают наличие определенных закрепленных нейронных путей за каждым из эмоциональных примитивов. К сожалению, исследования, проведенные на людях, оказались ненадежными [Lindquist, 2012].

Гнев, страх, отвращение, радость, удивление, доверие – наиболее часто используются как текстовая проекция базовых эмоций на письменную речь [Yuta Bann, 2012].

Текстовое отображение эмоций в письменной речи

Эмоциональная составляющая письменной речи является важной интерпретационной задачей для человека. Распознавание эмоций в письменной речи часто является субъективной особенностью, формирующейся за счет предыдущего опыта человека в его социальной среде.

Возможность подобной интерпретации можно свести к двум составляющим:

1) Субъективная эмоциональная составляющая,
2) Социальная эмоциональная составляющая

Другими словами, возможность субъектов понимать язык друг друга обосновывается также и тем, что социальная эмоциональная составляющая является непосредственной частью самого языка.

Формально, задача поиска социальной составляющей письменной речи сводится к ряду экспериментов, которые за счет многократного отображения эмоциональной составляющей на текстовые единицы письменной речи от различных субъектов и дальнейшего усреднения их, уменьшают значимость субъективной составляющей [Lupea, 2019]. Для такого рода отображения важно правильно определить единицы измерения как эмоционального пространства, так и пространства письменной речи.

Нами были выбраны слова как наименьшие единицы языка. Метриками эмоционального пространства – базовые эмоции. Функция отображения метрик эмоционального пространства на письменную речь – сложный когнитивный процесс, неосознанно присущий человеку.

Исходя из этого, выявление социальной эмоциональной составляющей языка становится возможным посредством проведения опросов.

Таким образом, мы смогли максимально уменьшить проблему общности эмоциональной разметки текста, но пока никак не снизили трудозатратность разметки слов.

Эмоциональная близость слов

Благодаря стремительному развитию интернета и популярности задач, связанных с автоматической обработкой естественных текстов, на сегодняшний день исследователями разработано несколько подходов, позволяющих находить близкие по смыслу слова.

Наиболее эффективно себя показали алгоритмы, работающие на основе дистрибутивного анализа больших корпусов текста [Lenci, 2008].

Наличие смысловой близости слов, дает возможность предположить и их эмоциональную близость.

Значит, для фиксированного слова в пространстве отображения эмоций на письменную речь предполагается выполнение выражения:

$$\vec{e}_k = \sum_{\mathrm{i},\mathrm{i}\neq k}^{N} \vec{e}_i \cdot \alpha_{\mathrm{ki}} \quad (1),$$

где $\vec{e}_k$– эмоциональный вектор фиксированного (к-ого) слова, $\vec{e}_i$– эмоциональный вектор близкого по смыслу (i-ого) слова, α_{ki}– коэффициент близости между фиксированным словом и близкому ему по смыслу.

Методология эксперимента. Экспериментальные данные

Для проведения эмоциональной разметки слов русского языка, нами были отфильтрованы только прилагательные, как наиболее простая часть речи для идентификации эмоций, и данный список был ранжирован по убыванию частоты употребления в устной речи, в соответствии «Новому частотному словарю русской лексики» под редакцией О.Н. Ляшевскаой и С.А. Шарова. Таким образом, мы выделили те слова, которые наиболее употребимы людьми и, предположительно, эмоциональная идентификация их будет наиболее простой.

На специализированном веб-сайте пользователям предлагалось разместить эти прилагательные, в соответствующие эмоциям поля. Прилагательные отбирались так, чтобы каждому уникальному пользователю не попадались одинаковые слова и выдавались в количестве 10 штук. Если исходя из разметки нескольких пользователей для слова достигалась статистическая достоверность эмоциональной разметки, то оно переставало участвовать в ранжировании и более не показывалось, уступая свое место еще не размеченному слову.

Таким образом, мы получили для каждого слова набор данных, представленных в таблице 1.

Таблица 1 - Пример набора сырых данных

Слово	userId	anger	fear	joy	disgust	trust
радостный	id1	0	0	1	0	0

Всего в опросе приняло участие 156 уникальных пользователей, и было размечено 6060 слов.

Для каждого уникального слова из размеченной выборки, была проведена агрегация слов по их значению, при этом суммировались как значения эмоций, так и количество уникальных пользователей, размечавших это слово (табл.2).

Таблица 2 - Пример агрегации сырых данных

Слово	userCount	Sum_anger	Sum_fear	Sum_joy	Sum_disgust	Sum_trust
радостный	20	2	1	17	0	0

Таким образом, проведя нормировку значений каждой эмоции на количество уникальных пользователей, были получены эмоциональные вектора, как раз отражающий социальную часть эмоций на данное слово, для выборки из 505 наиболее часто употребляемых в русском языке прилагательных.

Данная выборка, по своей сути, может быть представлена как известные вершины графа, представляющего собой семантическую сеть прилагательных, связанных мерой их близости.

Имея 505 слов изначальной выборки, мы смогли автоматически разметить эмоциональными векторами, отражающими социальную часть эмоций, оставшиеся 26 113 прилагательных, присутствующих в семантической сети прилагательных такой-то.

Поскольку валидация всех автоматически размеченных прилагательных представляет собой времязатратную, трудную задачу то нами было принято решение рассмотреть не все размеченные слова, а взять репрезентативную выборку.

$$n=\frac{\frac{Z^2pq}{\Delta^2}}{1+\frac{\frac{Z^2pq}{\Delta^2}-1}{N}} \quad (2),$$

Объем репрезентативной выборки n, выбранной случайным образом из генеральной совокупности определяется согласно выражению (2), где Z- коэффициент доверительного интервала, N- объем генеральной совокупности, p- уровень значимости, q- уровень доверия, Δ- предельная ошибка выборки.

Согласно выражению 2, достаточным объемом репрезентативной выборки будет 379 прилагательных. Рассмотрим распределение суммарной ошибки для них.

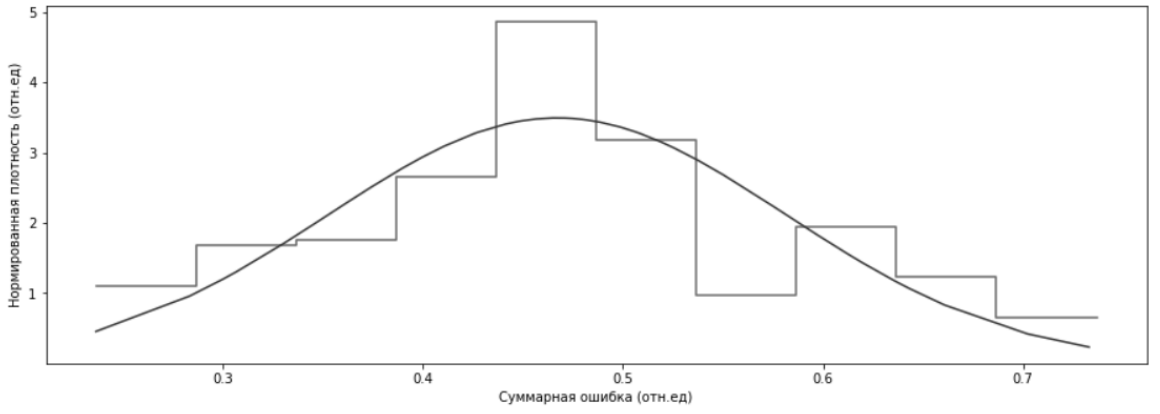

Рисунок 1 - Распределение суммарной ошибки выборки

На основании распределения на графике 1, мы видим, что наиболее вероятная суммарная ошибка прогнозирования для прилагательных русского языка равна 0,46.

Хотя суммарная ошибка для данной выборки оказалась достаточно велика, мы связываем это с довольно маленьким набором изначальной выборки. Чем больше вершин графа в семантической сети изначально известно, тем более точна работа алгоритма.

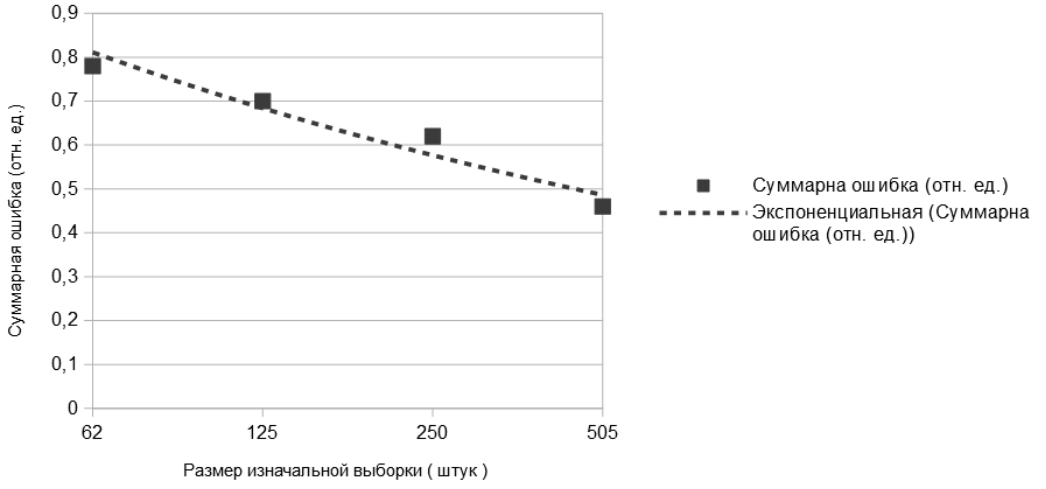

Рисунок 2 - Зависимость суммарной ошибки от размера выборки

Заключение

Нами был предложен алгоритм эмоциональной разметки прилагательных русского языка, на основе знания о семантической близости, произведена автоматическая разметка эмоциями большинства прилагательных в русском языке, проведена валидация алгоритма.

Библиография

1. Верани А. Роль внутренней речи в высших психических процессах // Культурно-историческая психология. 2010. Т. 6. № 1. С. 7-17.
2. Еголаева Е.С, Карнаухов В.А. Эмоции как проблема исследования в отечественной психологии // Научная инициатива в психологии. 2021. С. 55-58.
3. Lenci A. Distributional semantics in linguistic and cognitive research // Rivista di Linguistica. 2008. Vol 20. № 1. P. 1-31.
4. Lindquist K. The brain basis of emotion: A meta-analytic review // Behav Brain Sci. 2012. 35 (3). P. 121-143.
5. Lupea M. Studying emotions in Romanian words using Formal Concept Analysis // Computer Speech & Language. 2019. Vol. 57.
6. Mohammad S. Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon // Workshop on Computational Approaches to Analysis and Generation of Emotion in Text. 2010. P. 26-34.
7. Murthy A. A Review of Diferent Approaches for Detecting Emotion from Text // Annual International Conference on Data Science, Machine Learning and Blockchain Technology. 2021. Vol. 1110.
8. Pober J. What Emotions Really Are (In the Theory of Constructed Emotions) // Philosophy of Science. 2018. Vol. 85. № 4. P. 640–659.
9. Yadollahi A. et al. Current State of Text Sentiment Analysis from Opinion to Emotion Mining // ACM Computing Surveys. 2018. Vol. 50. Issue 2. 25. P. 1-33.
10. Yuta Bann E. Discovering Basic Emotion Sets via Semantic Clustering on a Twitter Corpus. 2012. URL: https://arxiv.org/abs/1212.6527

Algorithm for emotional labeling of adjectives in the Russian language

Andrei V. Meshchankin

Applicant,
Institute of Psychology, Sociology and Social Relations,
Moscow City Teachers Training University,
129226, 4, 2nd Selskokhozyaystvennyi driveway,
Moscow, Russian Federation;
e-mail: meschankin.andrey@physics.msu.ru

Abstract

The basis of psycholinguistic text analysis is the work with its various structural units. We focused our attention on the study of the psycholinguistic characteristics of the word, since it is the main structural unit of the language. Popular approaches to psycholinguistic analysis involve working with a large set of preprocessed, marked-up text units, which are words. Modern approaches of distributive semantics enable scientists to conduct research at the intersection of sciences. Combining research in the field of computational linguistics and psychology, we propose an algorithm for labeling Russian adjectives that allows you to automatically obtain a model that includes data on the emotional coloring of Russian adjectives. Such studies are a necessary step for further understanding the relationship between human cognitive features and written speech. To carry out the emotional markup of words in the Russian language, we filtered out only adjectives, as the simplest part of speech for identifying emotions, and this list was ranked in descending order of frequency of use in oral speech. Through the analysis made we have received 26,113 tagged adjectives. We have proposed an algorithm for the emotional marking of adjectives in the Russian language, based on knowledge of semantic proximity, automatic marking of most adjectives in the Russian language with emotions, and validated the algorithm.

For citation

Meshchankin A.V. (2022) Algoritm emotsional'noi razmetki prilagatel'nykh russkogo yazyka [Algorithm for emotional labeling of adjectives in the Russian language]. *Psikhologiya. Istoriko-kriticheskie obzory i sovremennye issledovaniya* [Psychology. Historical-critical Reviews and Current Researches], 11 (6A), pp. 86-93. DOI: 10.34670/AR.2023.96.87.032

Keywords

Psycholinguistics, emotions, distributive semantics, corpus of words, distribution algorithm, emotiveness, adjectives, distribution, coding, content analysis.

References

1. Egolaeva E.S, Karnaukhov V.A. (2021) Emotsii kak problema issledovaniya v otechestvennoi psikhologii [Emotions as a research problem in domestic psychology]. In: *Nauchnaya initsiativa v psikhologii* [Scientific initiative in psychology].
2. Lenci A. (2008) Distributional semantics in linguistic and cognitive research. *Rivista di Linguistica*, 20, 1, pp. 1-31.
3. Lindquist K. (2012) The brain basis of emotion: A meta-analytic review. *Behav Brain Sci*, 35 (3), pp. 121-143.
4. Lupea M. (2019) Studying emotions in Romanian words using Formal Concept Analysis. *Computer Speech & Language*, 57.
5. Mohammad S. (2010) Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion

Lexicon. In: *Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*.
6. Murthy A. (2021) A Review of Diferent Approaches for Detecting Emotion from Text. *Annual International Conference on Data Science, Machine Learning and Blockchain Technology*, 1110.
7. Pober J. (2018) What Emotions Really Are (In the Theory of Constructed Emotions). *Philosophy of Science*, 85, 4, pp. 640–659.
8. Verani A. (2010) Rol' vnutrennei rechi v vysshikh psikhicheskikh protsessakh [The role of inner speech in higher mental processes]. *Kul'turno-istoricheskaya psikhologiya* [Cultural-historical psychology], 6, 1, pp. 7-17.
9. Yadollahi A. et al. (2018) Current State of Text Sentiment Analysis from Opinion to Emotion Mining. *ACM Computing Surveys*, 50, 2. 25, pp. 1-33.
10. Yuta Bann E. (2012) *Discovering Basic Emotion Sets via Semantic Clustering on a Twitter Corpus*. Available at: https://arxiv.org/abs/1212.6527 [Accessed 08/08/2022]