Научная статья
Original article
УДК 004.032.2

АНАЛИЗ ПРОГРАММНЫХ МОДЕЛЕЙ ДЛЯ ОПРЕДЕЛЕНИЯ МЕРЫ СМЫСЛОВОЙ БЛИЗОСТИ ПРЕДЛОЖЕНИЙ ЕСТЕСТВЕННОГО ЯЗЫКА

ANALYSIS OF PROGRAM MODELS TO DETERMINE THE MEASURE OF SEMANTIC PROXIMITY OF NATURAL LANGUAGE SENTENCES

Смирнов Александр Андреевич, студент 2 курса магистратуры, факультет "Информатика и системы управления", МГТУ им. Н.Э. Баумана, Россия, г. Москва
Салып Богдан Юрьевич, студент 2 курса магистратуры, факультет "Информатика и системы управления", МГТУ им. Н.Э. Баумана, Россия, г. Москва
Научный руководитель: Ничушкина Татьяна Николаевна

Smirnov Alexander Andreevich, 2nd year master's student, Faculty of Informatics and Control Systems, Moscow State Technical University. N.E. Bauman, Russia, Moscow
Salyp Bogdan Yurievich, 2nd year master's student, Faculty of Informatics and Control Systems, Moscow State Technical University. N.E. Bauman, Russia, Moscow
Scientific adviser: Nichushkina Tatyana Nikolaevna

Аннотация: В статье проводится анализ программных моделей для определения меры смысловой близости предложений естественного языка.

Рассматриваются специфика, сферы применения и проблемы обработки естественного языка. Выявляются типы моделей семантического анализа текстов. Приводятся модели сравнения предложений, позволяющие определить меру их смысловой близости, такие как мера Жаккара, метод k-средних, косинусная мера близости и нейросетевая архитектура LSTM, и их ключевые свойства.

Abstract: The article analyzes program models to determine the measure of semantic similarity of natural language sentences. The specifics, scopes and problems of natural language processing are considered. Types of models of semantic analysis of texts are revealed. Models for comparison of sentences are presented that allow determining the measure of their semantic proximity, such as the Jaccard measure, the k-means method, the cosine measure of proximity and the LSTM neural network architecture, and their key properties.

Ключевые слова: естественный язык, смысловая близость предложений, программные модели, контекст, текстовый анализ.

Key words: natural language, semantic similarity of sentences, program models, context, text analysis.

Введение

По мере технологического развития задачи распознавания, описания и обработки естественного языка (далее – ЕЯ) с дальнейшей автоматизацией сопряжённых с ним процессов становятся всё более актуальными [1]. В настоящий момент наблюдается значительный прогресс в области обработки ЕЯ, однако, несмотря на это, уровень семантической обработки ЕЯ всё ещё остаётся недостаточным, что обуславливается неоднозначностью различных единиц языка. Для решения этой проблемы большое значение имеет анализ семантической близости текстов и смысловой близости предложений, осуществляемый методами компьютерной лингвистики и искусственного интеллекта [2]. Определение смысловой близости предложений может быть

реализовано посредством различных подходов, что делает актуальным их исследование и сопоставление.

Целью работы является анализ программных моделей для определения меры смысловой близости предложений ЕЯ. Для её достижения были использованы методы анализа и синтеза научных публикаций и литературных источников по рассматриваемой теме.

Обработка ЕЯ представляет собой общее направление математической лингвистики и искусственного интеллекта, изучающее проблемы компьютерного анализа и синтеза ЕЯ [3]. Обработка ЕЯ используется для решения следующих задач:

- машинный перевод;
- классификация текстов по категориям;
- сентиментный анализ;
- извлечение именованных сущностей.

Основной проблемой обработки ЕЯ является неоднозначность его единиц, отражающаяся на всех уровнях [4]. Выделяют лексическую, синтаксическую, структурную, семантическую и прагматическую неоднозначность. Эффективность решения проблемы неоднозначности в современных системах составляет порядка 60-70 %, что требует совершенствования моделей и методов обработки ЕЯ в целом и определения меры смысловой близости предложений в частности.

Для этих целей проводится семантический анализ текстов, являющийся этапом в последовательности действий алгоритма их автоматического понимания и заключающийся в выделении семантических отношений между смысловыми единицами текстов [5]. В рамках данного анализа могут быть построены следующие типы моделей, соответствующие описываемым признакам [6]:

- лексическая, рассматривающая текст как множество лексем без учёта их отношений между собой;

• грамматические, рассматривающие текст как грамматическую структуру с учётом отношений элементов текста, заданных в некоторой грамматике;

• статистические, рассматривающие текст как множество элементов и их комбинаций со статистическими свойствами и использующие в качестве основного атрибута частотно-вероятностные характеристики элементов текста;

• семантические, рассматривающие текст как совокупность некоторых элементов, обладающих смыслом.

Модели сравнения предложений, позволяющие определить меру их смысловой близости, относятся к последней категории. В качестве основного критерия отбора таких моделей рассматривается их применение в сфере анализа текстовых данных в целом и использования в целях классификации документов и нахождения некоторой меры различия между ними в частности.

Одним из алгоритмов определения смысловой близости предложений ЕЯ является мера Жаккара, измеряющая меру сходства как пересечение, разделяемое объединением объектов [7]. Её формула (1) имеет следующий вид:

$$K = \frac{A \cap B}{A + B - A \cap B}, \quad (1)$$

где A – множество слов первого предложения, B – множество слов второго предложения, $A \cap B$ – множество слов из пересечения множеств A и B, $A + B - A \cap B$ – множество всех различных слов обоих предложений, K – мера Жаккара.

Для предложения мера Жаккара сравнивает вес суммы общих слов с весом суммы слов, присутствующих в любом из предложений, но не являющихся общими словами. Данная модель отличается простотой реализации и небольшой вычислительной сложностью, однако обладает существенным недостатком – не учитывает порядок слов в предложении, вследствие чего при увеличении размера сравниваемых текстов множество

общих слов будет увеличиваться по отношению к общему множеству всех слов, и количество ложноположительных результатов будет возрастать [8].

Другим подходом является метод иерархической кластеризации, в рамках которого каждый объект в начале анализа рассматривается как отдельный кластер [9]. На последующих этапах осуществляется слияние наиболее близких кластеров, продолжающееся до тех пор, пока подобное слияние возможно. Одним из наиболее популярных алгоритмов кластеризации является метод *k*-средних, отличающийся высокой скоростью и достаточно высокой степенью эффективности [10]. Целью данного метода является минимизация полной внутриклассовой дисперсии, что отражено в формуле (2):

$$V = \sum_{i=0}^{k} \sum_{X_j \in C_i} \left(X_j - \mu_i\right)^2, \quad (2)$$

где X_j – векторы характеристик, k – число кластеров, C_i – кластеры, μ_i – центры кластеров.

Реализация метода осуществятся по следующему алгоритму:

1. Выбор количества кластеров.

2. Случайный выбор из заданного предложения k слов, которые на первом шаге будут считаться центрами кластеров.

3. Распределение всех слов по кластерам, вычисление расстояния от слова до каждого центра кластера и причисление слова к кластеру, расстояние до центра которого будет наименьшим.

4. Распределение всех слов по кластерам и пересчёт центров кластеров. В качестве нового центра выбирается среднее арифметическое всех принадлежащих кластеру слов.

5. Повторение пунктов 3 и 4 до выполнения одного из условий:

- стабилизация кластерных центров;
- равенство числа итераций максимальному числу итераций.

Описанный алгоритм анализа предложений не гарантирует нахождения лучшего решения, поскольку отличается чувствительностью к размерности

данных. Для устранения данного недостатка используются различные способы оптимизации.

Также для определения меры смысловой близости предложений ЕЯ может использоваться косинусная мера близости, являющаяся оптимальной для отображения семантического сходства предложений [11]. Данный метод выражается следующей формулой (3):

$$\cos(\alpha) = \frac{\sum A_i B_i}{\|A\|\|B\|}, \quad (3)$$

где A, B – векторы для первого и второго предложения соответственно, α – угол между векторами A и B.

Метод косинусной меры близости отличается простотой, однако имеет сильную зависимость от выбранного метода векторизации предложений, поэтому в случае неверного выбора способа векторизации результат может обладать крайне низкой точностью.

Распространённым и очень эффективным методом определения меры смысловой близости предложений ЕЯ являются рекуррентные нейронные сети [12]. Однако базовый вариант данной модели подвержен проблеме исчезающих и взрывающихся градиентов, поэтому на практике применяется усложнённая архитектура – долгосрочная кратковременная память (Long Short-Term Memory, LSTM).

В общем случае рекуррентная нейросеть на каждом временном шаге t принимает вектор входного примера x_t и обновляет собственное скрытое состояние h_t с применением логистической функции, зависящей от x_t и предыдущего состояния сети [13]. В LSTM-сетях с целью решения проблемы затухания сигнала ошибки в состав модели добавляется нейрон сети дополнительной переменной памяти c_t, значение которой на каждом временном шаге обновляется по следующей формуле (4):

$$c_t = f_t \circ c_{t-1} + i_t \circ \tanh(W_{cx} x_t + W_{ch} h_t - 1 + b_c), \quad (4)$$

где $\circ$ – произведение Адамара, i_t – канал, определяющий долю функции активации в текущем значении c_t от текущего входного вектора, f_t – канал,

определяющий долю предыдущего значения c_{t-1}, сохраняющуюся в текущем c_t, при этом

$$i_t = \sigma(W_{ix}x_t + W_{ix}h_{t-1} + W_{ic}c_t + b_i),$$
$$f_t = \sigma(W_{fx}x_t + W_{fh}h_{t-1} + W_{fc}c_{t-1} + b_f),$$

где σ – логистическая функция,

$$h_t = o_t \circ \tanh(c_t),$$
$$o_t = \sigma(W_{ox}x_t + W_{oc}h_{t-1} + W_{oc}c_{t-1} + b_o).$$

С целью увеличения ёмкости и производительности нейросети в её состав могут быть последовательно включены несколько LSTM-слоёв [14]. LSTM-сети хорошо приспособлены к обучению на задачах классификации, обработки и прогнозирования и могут успешно применяться для сравнения смысла предложений, поскольку способны запоминать информацию на долгий период времени и учитывать контекст предложения.

Проведённый анализ программных моделей для определения меры смысловой близости предложений ЕЯ показал, что лучшего результата обработки можно достичь посредством применения нейронной сети архитектуры LSTM. Несмотря на сложность реализации данной модели и ресурсоёмкости процесса обучения, она не обладает серьёзными ограничениями в области анализа текста и учитывает контекст слов. Остальные рассматриваемые модели – косинусная мера, мера Жаккара и метод *k*-средних – отличаются большей простотой реализации, однако не учитывают контекст или порядок слов, что ограничивает возможности их применения для решения задачи определения меры смысловой близости предложений ЕЯ.

Список литературы

1. Поречный А.С. Построение семантико-синтаксической модели текстов для определения их смысловой близости // Информатика: проблемы, методы, технологии: материалы XXI Международной научно-методической конференции. – Воронеж, 2021. – С. 1488-1495.

2. Глазкова А.В. Статистическая оценка информативности признаков для задачи поиска семантически близких предложений // Программные системы и вычислительные методы. – 2020. – № 1. – С. 8-17.
3. Бабаев А.М. Основные принципы обработки естественного языка // Дневник науки. – 2019. – № 12 (36). – С. 14-21.
4. Цитульский А.М., Иванников А.В., Рогов И.С. NLP – обработка естественных языков // StudNet. – 2020. – Т. 3, № 6. – С. 467-475.
5. Корешкова Т. Семантический анализ для автоматической обработки естественного языка [Электронный ресурс] // Научно-технический центр ФГУП «ГРЧЦ». – 2021. – URL: https://rdc.grfc.ru/2021/09/semantic_analysis/ (дата обращения: 18.04.2022).
6. Шишаев М.Г. Нейросетевые модели в задачах семантического анализа текстов на естественном языке // Труды Кольского научного центра РАН. – 2020. – Т. 11, № 8 (11). – С. 91-100.
7. Порошкина В.В. Меры подобия в рекомендательных системах // Аллея науки. – 2019. – Т. 1, № 4 (31). – С. 909-915.
8. Давыдова Ю.В. Алгоритм нечёткого текстового поиска в виртуальных социальных сетях // International Journal of Open Information Technologies. – 2018. – Т. 6, № 5. – С. 21-27.
9. Коннов И.В., Кашина О.А., Гильманова Э.И. Решение задачи кластеризации методами оптимизации на графах // Учёные записки Казанского университета. Серия: Физико-математические науки. – 2019. – Т. 161, № 3. – С. 423-437.
10. Котелина Н.О., Матвийчук Б.Р. Кластеризация изображения методом k-средних // Вестник Сыктывкарского университета. Серия 1: Математика. Механика. Информатика. – 2019. – № 3 (32). – С. 101-112.
11. Каряева М.С., Браславский П.И., Соколов В.А. Векторное представление слов с семантическими отношениями: экспериментальные наблюдения // Моделирование и анализ информационных систем. – 2018. – № 25 (6). – С. 726-733.

12. Богомолов Ю.А. Обзор моделей нейронных сетей для обработки естественного языка // StudNet. – 2020. – Т. 3, № 4. – С. 203-217.
13. Сбоев А.Г. Нейросетевое моделирование и машинное обучение на основе экспериментальных и наблюдательных данных: дис. … д-ра физ.-мат. наук: 05.13.18 / Александр Георгиевич Сбоев; Рос. науч. центр «Курчат. ин-т». – М., 2021. – 389 с.
14. Куратов Ю.М. Специализация языковых моделей для применения к задачам обработки естественного языка: дис. … канд. физ.-мат. наук: 05.13.17 / Юрий Михайлович Куратов; Моск. физ.-тех. инст. – М., 2020. – 121 с.

Bibliography

1. Porechny A.S. Building a semantic-syntactic model of texts to determine their semantic similarity // Informatics: problems, methods, technologies: materials of the XXI International scientific and methodological conference. - Voronezh, 2021. - S. 1488-1495.
2. Glazkova A.V. Statistical assessment of the information content of features for the problem of searching for semantically close sentences // Program systems and computational methods. - 2020. - No. 1. - P. 8-17.
3. Babaev A.M. Basic principles of natural language processing // Diary of science. - 2019. - No. 12 (36). - S. 14-21.
4. Tsitulsky A.M., Ivannikov A.V., Rogov I.S. NLP - natural language processing // StudNet. - 2020. - V. 3, No. 6. - S. 467-475.
5. Koreshkova T. Semantic analysis for automatic processing of natural language [Electronic resource] // Scientific and technical center of the Federal State Unitary Enterprise "GRChTs". – 2021. – URL: https://rdc.grfc.ru/2021/09/semantic_analysis/ (date of access: 04/18/2022).
6. Shishaev M.G. Neural network models in the problems of semantic analysis of texts in natural language // Proceedings of the Kola Scientific Center of the Russian Academy of Sciences. - 2020. - T. 11, No. 8 (11). - S. 91-100.

7. Poroshkina V.V. Similarity measures in recommender systems // Alley of Science. - 2019. - V. 1, No. 4 (31). - S. 909-915.
8. Davydova Yu.V. Fuzzy text search algorithm in virtual social networks // International Journal of Open Information Technologies. - 2018. - V. 6, No. 5. - S. 21-27.
9. Konnov I.V., Kashina O.A., Gilmanova E.I. Solution of the clustering problem by optimization methods on graphs. Uchenye zapiski Kazanskogo universiteta. Series: Physical and mathematical sciences. - 2019. - T. 161, No. 3. - S. 423-437.
10. Kotelina N.O., Matviychuk B.R. Image clustering by k-means // Bulletin of the Syktyvkar University. Series 1: Mathematics. Mechanics. Informatics. - 2019. - No. 3 (32). - S. 101-112.
11. Karyaeva M.S., Braslavsky P.I., Sokolov V.A. Vector representation of words with semantic relations: experimental observations // Modeling and analysis of information systems. - 2018. - No. 25 (6). - S. 726-733.
12. Bogomolov Yu.A. Overview of neural network models for natural language processing // StudNet. - 2020. - V. 3, No. 4. - S. 203-217.
13. Failures A.G. Neural network modeling and machine learning based on experimental and observational data: dis. … Dr. phys.-math. Sciences: 05.13.18 / Alexander Georgievich Sboev; Ros. scientific Center "Kurchat. in-t. - M., 2021. - 389 p.
14. Kuratov Yu.M. Specialization of language models for application to natural language processing problems: dis. … cand. Phys.-Math. Sciences: 05.13.17 / Yuri Mikhailovich Kuratov; Moscow fiz.-tech. inst. - M., 2020. - 121 p.

© Смирнов А.А., Салып Б.Ю., 2022 Научно-образовательный журнал для студентов и преподавателей «StudNet» №5/2022.

Для цитирования: Смирнов А.А., Салып Б.Ю. АНАЛИЗ ПРОГРАММНЫХ МОДЕЛЕЙ ДЛЯ ОПРЕДЕЛЕНИЯ МЕРЫ СМЫСЛОВОЙ БЛИЗОСТИ ПРЕДЛОЖЕНИЙ ЕСТЕСТВЕННОГО ЯЗЫКА// Научно-образовательный журнал для студентов и преподавателей «StudNet» №5/2022.