СИСТЕМНЫЙ АНАЛИЗ, УПРАВЛЕНИЕ И ОБРАБОТКА ИНФОРМАЦИИ / SYSTEM ANALYSIS, MANAGEMENT AND PROCESSING OF INFORMATION

DOI: https://doi.org/10.18454/itech.2024.2.3

МОДЕЛИ ПРЕДСТАВЛЕНИЯ ДАННЫХ В СИСТЕМАХ АВТОМАТИЧЕСКОЙ ОБРАБОТКИ ТЕКСТОВ

Обзор

Токолова А.А.[1, *], Пицхелаури С.Г.[2]
[1] ORCID : 0009-0003-9413-5087;
[2] ORCID : 0009-0005-6940-8148;
[1, 2] Московский авиационный институт, Москва, Российская Федерация
[2] МТС Диджитал, Москва, Российская Федерация

* Корреспондирующий автор (tokolovaa[at]gmail.com)

Аннотация

В статье рассматриваются современные подходы к обработке текстовых данных. Представлены уникальные модели, включая универсальную центроидно-контекстную модель (ЦКМ) и позиционную концептуальную модель представления понятий. Идея ЦКМ состоит в создании новых классов слов, ориентированных на схожесть грамматических признаков слов и их синтаксических функций в предложении. Тогда как модели, положенные в основу концептуального анализа текстов, позволяют выявить текстовую понятийную систему и установить смысловые отношения между ее элементами. Выявление системы понятий производится точными и предиктивными методами. Точное выделение понятий производится по эталонным концептуальным словарям (ЭКС). Предиктивное выделение понятий осуществляется по словарю концептуальных шаблонов. Обсуждаются модели представления данных и извлечения признаков из текстовых объектов с применением упорядоченных последовательностей и вектор-функций.

Ключевые слова: автоматическая обработка текстов, центроидно-контекстная модель, фразеологический концептуальный анализ, концептуальный анализ, метод лингвистической аналогии.

DATA REPRESENTATION MODELS IN AUTOMATIC TEXT PROCESSING SYSTEMS

Review article

Tokolova A.A.[1, *], Pitskhelauri S.[2]
[1] ORCID : 0009-0003-9413-5087;
[2] ORCID : 0009-0005-6940-8148;
[1, 2] Moscow Aviation Institute, Moscow, Russian Federation
[2] MTS Digital, Moscow, Russian Federation

* Corresponding author (tokolovaa[at]gmail.com)

Abstract

The article examines modern approaches to textual data processing. Unique models are presented, including the universal centroid-context model (CCM) and the positional conceptual model of concept representation. The idea behind the CCM is to create new word classes centred on the similarity of words' grammatical properties and their syntactic functions in a sentence. Whereas the models underlying the conceptual analysis of texts allow to identify the textual conceptual system and to establish semantic relations between its elements. The identification of the concept system is done by exact and predictive methods. Precise selection of concepts is performed using reference conceptual dictionaries (RCD). Predictive concept extraction is done by conceptual pattern dictionary. Models of data representation and feature extraction from text objects using ordered sequences and vector functions are discussed.

Keywords: automatic text processing, centroid-context model, phraseological conceptual analysis, conceptual analysis, method of linguistic analogy.

Введение

Проведенные исследования показали, что точные методы позволяют выявить понятийную систему научно-технических текстов по любой тематике на 70-85%, а предиктивные методы позволяют выявить оставшуюся часть (15-30%) с вероятностью 85-90% [2].

В настоящее время значительная часть разработчиков интеллектуальных систем обработки текстовых данных свои представления о технологиях искусственного интеллекта (ИИ) связывают с нейросетевым (НС) подходом. В основу НС подхода положена модель глубокого обучения, ориентированная на извлечение большого числа признаков из необработанных данных и решения сложных интеллектуальных задач [6]. Процессу автоматизированной обработки текстов (АОТ) предшествуют процессы обучения модели и проверки обученной модели. Для полноценного процесса обучения требуются размеченные данные больших объемов, при невозможности получения достаточных объемов данных полноценно обучить модель невозможно. Кроме того, в некоторых ситуациях возможны ограничения по использованию технологий НС [7].

Классический лингвистический подход, основанный на изучении грамматики и синтаксиса, часто представляет язык через универсальные правила, игнорируя индивидуальные особенности и разнообразие коммуникативных ситуаций. Он, таким образом, ограничивает понимание того, как язык используется на практике в различных контекстах. Большой проблемой в рамках этого подхода являлась проблема вариативности представления смыслового

содержания текстовых конструкций. Традиционные лингвистические модели, которые опираются на жёсткие грамматические правила и ограниченные лексические ресурсы, сталкиваются с ограничениями в отображении всего богатства грамматических и семантических контекстов, которые могут присутствовать в тексте [10]. Одной из основных проблем таких моделей является их недостаточная гибкость, ограниченное пространство признаков и невозможность быстрой коррекции в случае несоответствия конкретным текстовым данным.

В последнее время наметился подход, базирующийся на теоретической концепции фразеологического концептуального анализа текста (ФКАТ) [1]. Этот подход в значительной мере является дальнейшим развитием традиционного лингвистического подхода.

Фразеологический концептуальный анализ текстов

Концепция ФКАТ базируется на уникальной машинной грамматике, традиционных лингвистических моделях дополненных рядом гибких динамических языковых моделей, в основу которых положены принципы и методы лингвистической аналогии [9], а также ориентацией на использование многомиллионных словарных ресурсов.

Основные положения концепции фразеологического концептуального анализа текстов:

• смысловое содержание текстов выражается с помощью единиц смысла;

• понятие – самая устойчивая единица смысла;

• объекты предложения обладают особыми признаками, выражающимися через предикатно-актантную структуру (ПАС) и набором отношений с другими объектами;

• сверхфразовые единства формируются из предложений и представляются в виде последовательностей предложений (связного текста).

Идея закладывать в модель флективных классов слов русского языка [1] строгое соответствие между их формой представления и грамматической информацией послужила основанием для созданияновых классов – тех, где слова имеют одинаковые наборы грамматических признаков, соответствующих их формам представления в сходных контекстах. Появление идеи формирования новых классов слов, ориентированных на схожесть грамматических признаков слов и схожесть их синтаксических функций в предложении, впервые было предложено в работах [1] и активно использовалась при составлении эталонных концептуальных словарей (ЭКС).

Позиционная концептуальная модель представления понятия

Модели, положенные в основу *концептуального анализа текстов*, позволяют выявить текстовую понятийную систему и установить смысловые отношения между ее элементами. Выявление системы понятий производится точными и предиктивными методами. Точное выделение понятий производится по ЭКС, включающего более двух млн. понятий. Предиктивное выделение понятий осуществляется по словарю концептуальных шаблонов, в котором в сжатом мнемоническом виде представлены обобщенные формы слов наименований понятий в их контекстном окружении.

В процессе концептуального анализа на основе данных, полученных с помощью морфологического и семантико-синтаксического анализа, из предложения выделяются смысловые конструкции с помощью эталонного концептуального словаря (в словарном концептуальном анализе) и шаблонов представления текстовой формы словосочетаний эталонного концептуального словаря (в предиктивном концептуальном анализе) [3], [4], [5].

Сначала предложение подается на обработку словарному концептуальному анализу. На вход словарному концептуальному анализу оно подается в виде множества слов $Pred = \{word_i\}_{i=1}^{dimPred}$.

$$fNS(word_i) = word.norm_i. \quad (1)$$

Далее предложение разбивается на фрагменты $\{word.frag.norm_i\}_{i=1}^{k}$ с помощью функции:

$$Frag : Pred \rightarrow \{word.frag.norm_i\}_{i=1}^{k}, k \geq dimPred \quad (2)$$

Разбиение на фрагменты реализуется по следующему правилу:

$$[word.norm_1, word.norm_2], ..., [word.norm_1, ..., word.norm_{dimPred}],$$
$$[word.norm_2, word.norm_3], ...$$

Здесь $word.frag_i = [\cdot]$.

Фрагмент $word.frag_i$ сравнивается с нормализованными понятиями из эталонного концептуального словаря (ЭКС), при совпадении он добавляется $word.frag_i$ во множество $CONCEPT = \{concept_i\}_{i=1}^{dimConcept}$ и ему в соответствие ставится совпавшее понятие из ЭКС.

При несовпадении исходное предложение в виде последовательности слов $Pred = \{word_i\}_{i=1}^{dimPred}$ подается на вход предиктивному концептуальному анализу.

Каждому $word_i$ ставится в соответствие набор признаков (определенных на этапе МА) $p_i = \{Pb_i, Pa_i\}$, $\{OS_i, GK_i, GF_i\} \subset Pb_i \cup Pa_i$.

Полученная синтагма разбивается на фрагменты $\{word.frag_i\}_{i=1}^{k}$ с помощью функции:

$$Frag : Pred \rightarrow \{word.frag_i\}_{i=1}^{k}, k \geq dimPred. \quad (3)$$

Разбиение на фрагменты происходит по следующему правилу:

$$[word_1, word_2], \ldots, [word_1, \ldots, word_{dimPred}], [word_2, word_3], \ldots$$

Здесь $word.frag_i = [\cdot]$.

Имеется множество шаблонов $TEMP = \{t_i, \dots t_n\}$, $t_i = \{OS_i, GK_i, GF_i\}$ представления текстовой формы словосочетаний ЭКС (форма каждого слов 4 сим.: OS(FK+OK)=2сим., GK=1 сим., GF=1 сим.).

Проверяем вхождение признаков p_i в $TEMP$ для каждого набора фрагментов $word.frag_i$, если $\exists j : p_i = t_j$, то добавляем текстовое представление наименования понятия во множество $CONCEPT = \{concept_i\}_{i=1}^{dimConcept}$ и ставим ему в соответствие шаблон t_j .

Существует словарь стоп-слов $STOP = \{stop_i\}_{i=1}^{dimSTOP}$ (в нем содержатся служебные слова). Если $concept_i \in STOP$, то это понятие исключается из множества $CONCEPT$:

$$CONCEPT = CONCEPT \backslash \{concept_i\}. \quad (4)$$

Центроидно-контекстная модель представления данных

При создании технологий смысловой обработки текстов был разработан ряд моделей представления данных. Каждому уровню иерархии смысловых единиц текста соответствовала та модель, которая наиболее адекватно отображала понятийную систему признаков каждого текстового объекта. Но основной моделью, позволяющей реализовать возможность однозначного разрешения множества языковых ситуаций в их контекстных окружениях, является универсальная центроидно-контекстная модель (ЦКМ) [8].

На вход ЦКМ подается упорядоченное множество объектов $Sem = \{s_i\}_{i=1}^{n}$. Каждому объекту ставится в соответствие элемент упорядоченного множества $Pr = \{p_i\}_{i=1}^{n}, s_i \leftrightarrow p_i$, где p_i – свойства объекта. Задается радиус n , выбирается целевой объект (центроид) s_k , составляется позиционная модель, которая задается упорядоченной последовательностью:

$$PM = \{s_k, s_{k+1}, s_{k-1}, s_{k+2}, s_{k-2}, \dots, s_{k+n}, s_{k-n}\}. \quad (5)$$

К упорядоченной последовательности PM применяется вектор-функция fPR, которая каждому объекту из PM ставит в соответствие признаки из множества Pr:

$$\mathrm{fPR}\,(\mathrm{s_i}) = \begin{cases} \mathrm{p_i}, \mathrm{s_i} \in \mathrm{Sem}, \\ \mathrm{p_0}, \ \mathrm{s_i} \notin \mathrm{Sem}. \end{cases} \quad (6)$$

Здесь p_0 – фиктивный признак фиктивного элемента. В результате применения функции получается упорядоченная последовательность:

$$fPR(PM) = PMp = \{p_k, p_{k+1}, p_{k-1}, p_{k+2}, p_{k-2}, \dots, p_{k+n}, p_{k-n}\}. \quad (7)$$

Пусть существует функция $fSim(similarfunction)$, которая проверяет, насколько два вектора схожи между собой:

$$fSim(a, b) = i, \text{ если } \forall k \in \{1, \dots, i\} a_k = b_k \text{ и } a_{i+1} \neq b_{i+1}. \quad (8)$$

Задается порог совпадения начальной части модели bv и во множестве MOD шаблонов синтагм ЦКМ из словаря шаблонов ищется наиболее схожая структура в смысле максимума процента покрытия $CovPer$:

$$\mathrm{PMpMod} = \underset{\mathrm{meMOD}}{\mathrm{argmax}}\ \mathrm{CovPer}(\mathrm{PMp}, \mathrm{m})$$
$$= \begin{cases} 0\%, & \mathrm{fSim}(\mathrm{PMp}, \mathrm{m}) < \mathrm{bv} \\ \frac{\mathrm{l} - \mathrm{fHam}(\mathrm{PMp}, \mathrm{m})}{\mathrm{l}} \cdot 100\%, & \mathrm{fSim}(\mathrm{PMp}, \mathrm{m}) \geq \mathrm{bv} \end{cases}, \quad (9)$$

где $fHam(\cdot, \cdot)$ – функция, которая вычисляет расстояние Хэмминга (количество несовпадающих элементов), $1 = 2n + 1$ – количество элементов синтагмы.

Множество MOD включает в себя шаблоны признаков ЦКМ из словаря шаблонов. Каждый элемент множества имеет признак, принимающий значение 0 или 1.

Пусть функция $ftakePR$ – функция взятия признака элемента $m \in MOD$.

В результате ЦКМ мы находим ответ на вопрос: *является ли однозначное решение для анализируемой языковой ситуации именно таким, как описано в соответствующем словарном ресурсе. Этот ответ можно получить следующим образом:*

$$\mathrm{answer} = \begin{cases} \text{да}, & \mathrm{ftakePR}\ (\mathrm{PMpMod}) = 1, \\ \text{нет}, & \mathrm{ftakePR}\ (\mathrm{PMpMod}) = 0. \end{cases} \quad (6)$$

Представленная модель ЦКМ реализована в ряде процедур автоматической обработки текстов (АОТ), примером ее применения может служить процедура семантико-синтаксического анализа, в котором модель ЦКМ используется:

а) для разрешения омонимии словоформ по их контекстному окружению;

б) для установления границ простых предложений в составе сложного;

в) для определения элементов предикатно-актантной структуры;

г) для построения бинарных отношений между элементами каркаса предложения и элементами его синтаксических конструкций.

Заключение

В настоящей статье рассмотрены различные подходы к обработке текстовых данных, начиная от использования нейросетевых методов, завершая уникальными моделями, основанными на теоретической концепции фразеологического концептуального анализа текста (ФКАТ). Отмечается, что технологии искусственного интеллекта и классические лингвистические модели имеют свои ограничения и преимущества, которые могут быть учтены при разработке интеллектуальных систем обработки текстов.

Исследования, проведенные в данной области, отмечают, что точные и предиктивные методы выявления текстовой понятийной системы могут обеспечить высокий уровень точности в определении смысловых отношений между элементами текста. Это позволяет с достаточной вероятностью выделить как точные, так и предиктивные концептуальные связи, обогатив тем самым понимание содержания текстовых данных.

Таким образом, современные методы обработки текстовых данных продолжают эволюционировать, предоставляя более интеллектуальные и гибкие подходы к анализу языка, отражая его многообразие и контекстуальные особенности.

Конфликт интересов

Не указан.

Рецензия

Все статьи проходят рецензирование. Но рецензент или автор статьи предпочли не публиковать рецензию к этой статье в открытом доступе. Рецензия может быть предоставлена компетентным органам по запросу.

Conflict of Interest

None declared.

Review

All articles are peer-reviewed. But the reviewer or the author of the article chose not to publish a review of this article in the public domain. The review can be provided to the competent authorities upon request.

Список литературы / References

1. Белоногов Г.Г. Компьютерная лингвистика и перспективные информационные технологии. Теория и практика построения систем автоматической обработки текстовой информации / Г.Г. Белоногов, Ю.П. Калини, А.А. Хорошилов — Москва: Русский мир, 2004. — 246 с.

2. Колин К.К. Искусственный интеллект в технологиях машинного перевода / К.К. Колин, А.А. Хорошилов, Ю.В. Никитин и др. // Социальные новации и социальные науки. — 2021. — 2. — с. 64-72. — DOI: 10.31249/snsn/2021.02.05.

3. Хорошилов А.А. Автоматическое выявление и классификация информационных событий в текстах СМИ / А.А. Хорошилов, Р.Р. Мусабаев, Я.Д. Козловская и др. // Научно-техническая информация. Серия 2: Информационные процессы и системы. — 2020. — 7. — с. 27-38. — DOI: 10.36535/0548-0027-2020-07-4 .

4. Хорошилов А.А. Автоматическое создание формализованного представления смыслового содержания неструктурированных текстовых сообщений СМИ и социальных сетей / А.А. Хорошилов, Ю.В. Никитин, А.А. Хорошилов и др. // Системы высокой доступности. — 2014. — 3. — с. 36-51.

5. Хорошилов А.А. Определение тональности сообщений СМИ методом их концептуального анализа / А.А. Хорошилов, Я.Д. Козловская, Р.Р. Мусабаев и др. // Моделирование и анализ данных. — 2019. — 4. — с. 67-79.

6. Пекунов В.В. Извлечение информации из нейронных сетей прямого распространения в виде простых алгебраических моделей / В.В. Пекунов // Информационные технологии. — 2017. — 1. — с. 76-80.

7. Большакова Е.И. Автоматическая обработка текстов на естественном языке и анализ данных / Е.И. Большакова, К.В. Воронцов, Н.Э. Ефремова и др. — Москва: НИУ ВШЭ, 2017. — 269 с.

8. Хорошилов А.А. Контекстное разрешение омонимии на основе центроидно-контекстной модели / А.А. Хорошилов, Ю.В. Никитн, А.В. Кан и др. // Труды ИСП РАН. — 2022. — 5. — с. 171-182. — DOI: 10.15514/ISPRAS-2022-34(5)-11.

9. Аблов И.В. Средства машинной грамматики русского языка (по Г.Г. Белоногову) / И.В. Аблов, В.Н. Козичев, А.А. Хорошилов и др. // Научно-техническая информация. Серия 2: Информационные процессы и системы. — 2018. — 6. — с. 32-46.

10. Калини Ю.П. Современные технологии автоматизированной обработки текстовой информации / Ю.П. Калини, А. А. Хорошилов, А.А. Хорошилов // Системы высокой доступности. — 2015. — 2. — с. 19-34.

Список литературы на английском языке / References in English

1. Belonogov G.G. Komp'juternaja lingvistika i perspektivnye informatsionnye tehnologii. Teorija i praktika postroenija sistem avtomaticheskoj obrabotki tekstovoj informatsii [Computer Linguistics and Perspective Information Technologies. Theory and Practice of Building Systems of Automatic Processing of Text Information] / G.G. Belonogov, Ju.P. Kalini, A.A. Horoshilov — Moskva: Russkij mir, 2004. — 246 p. [in Russian]

2. Kolin K.K. Iskusstvennyj intellekt v tehnologijah mashinnogo perevoda [Artificial Intelligence in Machine Translation Technologies] / K.K. Kolin, A.A. Horoshilov, Ju.V. Nikitin et al. // Social Innovations and Social Sciences. — 2021. — 2. — p. 64-72. — DOI: 10.31249/snsn/2021.02.05. [in Russian]

3. Horoshilov A.A. Avtomaticheskoe vyjavlenie i klassifikatsija informatsionnyh sobytij v tekstah SMI [Automatic Detection and Classification of Information Events in Media Texts] / A.A. Horoshilov, R.R. Musabaev, Ja.D. Kozlovskaja et al. // Scientific and Technical Information. Series 2: Information Processes and Systems. — 2020. — 7. — p. 27-38. — DOI: 10.36535/0548-0027-2020-07-4 . [in Russian]

4. Horoshilov A.A. Avtomaticheskoe sozdanie formalizovannogo predstavlenija smyslovogo soderzhanija nestrukturirovannyh tekstovyh soobschenij SMI i sotsial'nyh setej [Automatic Construction of a Formalized Representation of

Semantic Contents of Unstructured Texts of Mass-media and Social Networks] / A.A. Horoshilov, Ju.V. Nikitin, A.A. Horoshilov et al. // Highly Available Systems. — 2014. — 3. — p. 36-51. [in Russian]

5. Horoshilov A.A. Opredelenie tonal'nosti soobschenij SMI metodom ih kontseptual'nogo analiza [Determine the Tonality of News Media Reports by Conceptual Analysis] / A.A. Horoshilov, Ja.D. Kozlovskaja, R.R. Musabaev et al. // Modelling and Data Analysis. — 2019. — 4. — p. 67-79. [in Russian]

6. Pekunov V.V. Izvlechenie informatsii iz nejronnyh setej prjamogo rasprostranenija v vide prostyh algebraicheskih modelej [The Derivation of the Information from Artificial Feed-Forward Neural Networks in the Form of the Simple Algebraic Models] / V.V. Pekunov // Information Technology. — 2017. — 1. — p. 76-80. [in Russian]

7. Bol'shakova E.I. Avtomaticheskaja obrabotka tekstov na estestvennom jazyke i analiz dannyh [Automatic Natural Language Processing and Data Analysis] / E.I. Bol'shakova, K.V. Vorontsov, N.E. Efremova et al. — Moskva: NIU VShE, 2017. — 269 p. [in Russian]

8. Horoshilov A.A. Kontekstnoe razreshenie omonimii na osnove tsentroidno-kontekstnoj modeli [Context Resolution of Homonymy Based on a Centroid-context Model] / A.A. Horoshilov, Ju.V. Nikitn, A.V. Kan et al. // Proc. ISP RAS. — 2022. — 5. — p. 171-182. — DOI: 10.15514/ISPRAS-2022-34(5)-11. [in Russian]

9. Ablov I.V. Sredstva mashinnoj grammatiki russkogo jazyka (po G.G. Belonogovu) [Means of Machine Grammar of the Russian Language (by G.G. Belonogov)] / I.V. Ablov, V.N. Kozichev, A.A. Horoshilov et al. // Scientific and Technical Information. Series 2: Information Processes and Systems. — 2018. — 6. — p. 32-46. [in Russian]

10. Kalini Ju.P. Sovremennye tehnologii avtomatizirovannoj obrabotki tekstovoj informatsii [Modern Technologies of Automated Text Iinformation Processing] / Ju.P. Kalini, A. A. Horoshilov, A.A. Horoshilov // Highly Available Systems. — 2015. — 2. — p. 19-34. [in Russian]