DOI: 10.51790/2712-9942-2024-5-4-09

МЕТОДЫ АНАЛИЗА ТЕКСТА НА ЕСТЕСТВЕННОМ ЯЗЫКЕ В ЗАДАЧЕ ЭКСТРАКТИВНОГО ИНФОРМАЦИОННОГО ПОИСКА

Ю. В. Перцев
Южно-Уральский государственный университет, г. Челябинск, Российская Федерация
✍ *me@pertsvuv.com*

Аннотация: в статье рассматривается лингво-математическое обеспечение интеллектуальных информационно-поисковых систем. Активно развивающиеся последнее десятилетие большие языковые модели, способные решать задачи информационного поиска, часто оказываются ресурсоемкими и имеют избыточный функционал при встраивании в специализированные информационные системы. Это создает необходимость разработки более легковесных методов обработки текста на естественном языке. Рассматривается экстрактивный подход к организации вопросно-ответного поиска, задача которого состоит в нахождении предложений, отвечающих на вопрос в заранее выбранном документе. В рамках организации этого подхода предлагаются методы анализа морфологии, синтаксиса и семантики естественного языка. Для реализации графового синтаксического анализа, основанного на взвешивании полного ориентированного графа искусственной нейронной сетью прямого распространения, собран корпус текстов на русском языке, содержащий 8800 предложений. Также этот корпус используется для получения набора синтаксически ориентированных векторных представлений слов, применяющегося на этапе семантического анализа, посредством модели, основанной на архитектуре непрерывного мешка слов. Механизм ранжирования предложений относительно вопроса основан на формализации семантики текста на естественном языке в виде сильно-связного ориентированного графа, выявляющего неявные содержательные закономерности языковых структур.

Ключевые слова: обработка естественного языка, компьютерная лингвистика, корпус текстов, синтаксис, семантика.

Для цитирования: Перцев Ю. В. Методы анализа текста на естественном языке в задаче экстрактивного информационного поиска. *Успехи кибернетики*. 2024;5(4):67–74. DOI: 10.51790/2712-9942-2024-5-4-09.

Поступила в редакцию: 07.10.2024. *В окончательном варианте: 05.11.2024.*

NATURAL LANGUAGE PROCESSING FOR EXTRACTIVE SEARCH

Yu. V. Pertsev
South Ural State University, Chelyabinsk, Russian Federation
✍ *me@pertsvuv.com*

Abstract: the paper presents the linguistic software for intelligent search systems. Large language models have been actively developing for the last decade. LLMs suitable for information search often require extensive resources and have redundant functionality when embedded into targeted information systems. Lightweight approaches to natural language processing are needed. We considered an extractive approach to a “question-answer” search intended to find sentences that answer a question in the specified document. For this, we proposed methods for analyzing the morphology, syntax, and semantics of the natural language. A corpus of Russian language texts containing 8,800 sentences was collected to implement graph-based syntax analysis with a weighting of a completely oriented graph by a forward-propagation artificial neural network. This corpus was also used to produce a set of syntax-oriented vector representations of words, applied in the semantic analysis by using a model based on a continuous bag of words architecture. The sentence ranking by relevance to the question is based on representing the semantics of the natural language text as a strongly connected directed graph, revealing implicit meaningful patterns within the language structures.

Keywords: natural language processing, computational linguistics, text corpora, syntax, semantics.

Cite this article: Pertsev Yu. V. Natural Language Processing for Extractive Search. *Russian Journal of Cybernetics*. 2024;5(4):67–74. DOI: 10.51790/2712-9942-2024-5-4-09.

Original article submitted: 07.10.2024. *Revision submitted: 05.11.2024.*

Введение

В настоящее время перспективным направлением развития интеллектуальных информационно-поисковых систем являются программные комплексы, способные анализировать и синтезировать тексты на естественном языке, — системы с естественно-языковым интерфейсом. К информационно-поисковым системам с естественно-языковым интерфейсом относят интеллектуальные поисковые системы, вопросно-ответные сервисы, виртуальные собеседники и иные системы, работа которых сопряжена с интеллектуальной обработкой естественного языка при решении задач информационного поиска. Появление больших языковых моделей и реализация информационно-поисковых систем на их основе, способных эффективно решать задачи информационного поиска, приведут сначала к стремительному изменению ландшафта рынка поисковых систем, а затем и к реорганизации всей человеко-машинной коммуникации [1]. Однако большие языковые модели, представляющие собой массивные программные комплексы с сотнями миллионов настраиваемых параметров и стремящиеся к информационной универсальности и мультимодальности, с прикладной точки зрения зачастую оказываются ресурсоемкими инструментами с избыточным функционалом. Это обуславливает необходимость разработки легковесных методов обработки естественного языка.

Модель задачи экстрактивного поиска

Рассмотрим два концептуально разных подхода к организации вопросно-ответного информационного поиска. Пусть пользователь сформулировал некоторый информационно сложный поисковый запрос, то есть такой вопрос, что ответ на него не может быть заключен в единственном односложном предложении, а вопросно-ответной системе предстоит дать на этот вопрос наиболее полный и точный ответ. Первый подход к решению подобной задачи заключается в анализе нескольких документов и выделении наиболее релевантных участков текста, совокупность которых образует ответ на вопрос пользователя, — таким образом, задача фактически сводится к ранжированию связных частей текста и выбору некоторого количества наиболее релевантных вопросу. Другим подходом является генерация ответа, содержательно обобщающего участки текста одного документа или коллекции документов, то есть синтез нового текста, не представленного явно в исходном. Эти концептуально разные подходы — экстрактивный и генеративный — обуславливают особенности пользовательского опыта в различных поисковых ситуациях.

Задача экстрактивного вопросно-ответного поиска формулируется следующим образом: дана пара (d, Q), где Q — вопросительное предложение, представленное в виде мешка слов, $d \in D$ — документ, предварительно выбранный из некоторой коллекции так, чтобы подходить по смыслу к вопросу Q. Необходимо найти предложения, отвечающие на вопрос Q.

Структура вопросно-ответной системы

В качестве ядра вопросно-ответной системы рассмотрим лингвистический модуль, обеспечивающий интеллектуальную обработку текста. Эффективность системы всецело диктуется качеством реализации алгоритмической и программной составляющих лингво-математического обеспечения, то есть, с одной стороны, теоретической эффективностью методов обработки естественного языка, а с другой — непосредственным воплощением этих методов в программном коде.

Естественный язык, будучи семиотически наиболее сложной и неоднородной знаковой системой, состоит из нескольких взаимосвязанных уровней, — их принято рассматривать в виде следующих формальных структур (рисунок 1).

В контексте вычислительного языкознания лексический уровень рассматривает словари, отдельные слова и их свойства, синтаксис — грамматику сочетания слов и фраз, семантика — языковой смысл. На рисунке отсутствуют два уровня: морфологический, рассматривающий грамматику словоизменения, и прагматический — язык в контексте коммуникации, а в некоторых языковых теориях принято выделять подуровни, например поверхностные и глубинные в морфологии и синтаксисе; морфологический уровень на рисунке инкапсулирован лексическим уровнем.

Лингвистический модуль выполняет две задачи: получение формализованных представлений семантики текстов для последующей организации поиска и семантический анализ вопроса. Структура вопросно-ответной системы, соответствующая представленной формализации языка, показана на рисунке 2.

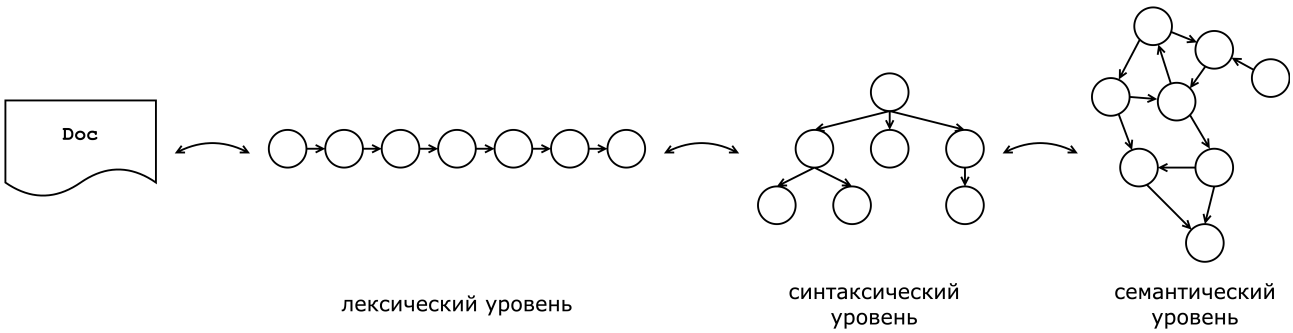

Рис. 1. *Высокоуровневое устройство естественного языка*

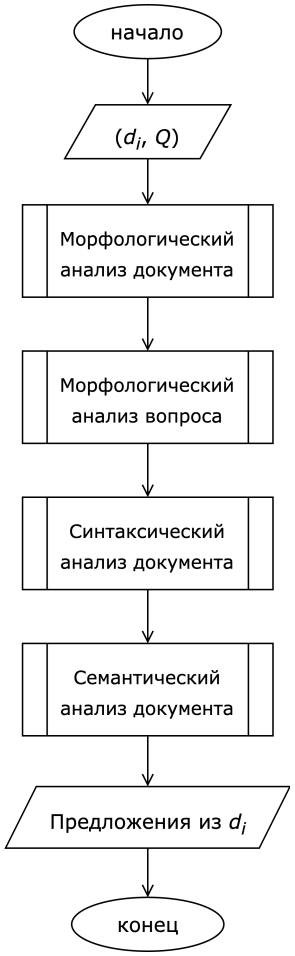

Рис. 2. *Вопросно-ответная система*

Заметим, что в рассматриваемой вопросно-ответной системе производится исключительно языковой анализ текстов, без использования того, что называется «концептуальным представлением», то есть без использования внешних по отношению к языку знаний о мире.

Анализ морфологии текста

Методы компьютерной морфологии позволяют анализировать и синтезировать слова естественного языка, однако обычно под морфологическим анализом понимается определение канонической формы слова и грамматических свойств заданной словоформы. Если требуется создать нужную грамматическую форму слова, то данная задача называется порождением словоформы. Перед морфологическим анализом проводится предварительная обработка текста: разбитие его на предложения и выделение в них отдельных элементов, таких как слова, числа, знаки препинания и другие атомарные последовательности символов, что называется токенизацией. После этого в словаре, специальным образом представленном в памяти компьютера, по словоформе осуществляется нахождение набора

граммем. Граммемы соответствуют принятым наборам в проекте OpenCorpora [2].

Анализ синтаксиса текста

Синтаксическим анализом называется процесс выявления грамматической структуры предложения. Рассматривая предложение как цепочку слов, мы можем представить информацию о его грамматическом строении как набор сведений о «главенствовании» одних точек цепочки над другими. Задать такой набор — значит задать некоторое дерево на множестве точек цепочки. В работе рассматривается нейросетевой метод построения грамматики зависимостей на материале русского языка.

Можно сказать, что всякая новая нотация, описывающая синтаксис, или всякий новый корпус текстов представляет собой новый формализм описания синтаксических отношений в некоторой грамматике, в данном случае — грамматики зависимостей. Наиболее популярными стандартами описания русского языка в терминах деревьев синтаксического подчинения являются Universal Dependencies и СинТагРус. В основу принимаемого в данной работе формата синтаксической структуры лег стандарт СинТагРус с некоторыми изменениями [3]. Так, например, в корпусе СинТагРус эллиптические конструкции явно восстанавливаются внутри синтаксических деревьев, а в формате зависимостей, принятом в данной работе, восстановление эллиптических конструкций не предполагается.

Вручную собрана и размечена коллекция текстов веб-страниц на русском языке, снабженная морфологической разметкой, указывающей нормальную форму слова и набор значений грамматических характеристик в измененной нотации OpenCorpora; синтаксической разметкой, задающей строй предложений в виде грамматики зависимостей; разметкой вида синтаксических отношений. Всего в корпусе содержится 8800 предложений, 210152 токена, из них 173943 входят в размеченные синтаксические структуры.

На малый корпус размером в 2200 предложений, представленный в формате SQLite, получено свидетельство о государственной регистрации базы данных [4].

Построение синтаксической структуры проходит в три этапа. На первом этапе из всех токенов, участвующих в синтаксическом анализе, формируется полный ориентированный граф, дополняющийся корневой вершиной (показан на рисунке 3). Затем граф взвешивается посредством нейронной сети. После этого в получившемся графе производится поиск ориентированного остовного корневого дерева максимального веса посредством алгоритма Чу-Лью-Эдмондса — это дерево и есть искомая синтаксическая структура.

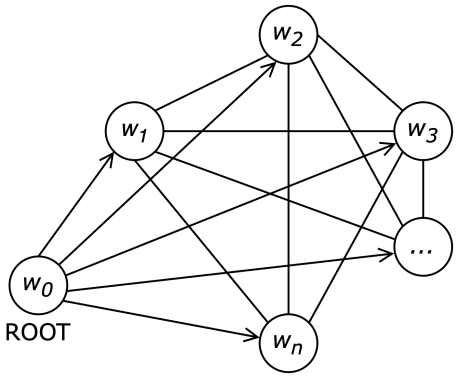

Рис. 3. *Репрезентация предложения в виде графа*

Искусственная нейронная сеть прямого распространения обучена на собранном корпусе текстов. Так как предложение в процессе анализа представляется в виде полного ориентированного графа с дополнительной корневой вершиной, нейронной сети необходимо оценить все дуги этого графа. Таким образом, для правильного ее обучения из каждой синтаксической структуры исходного корпуса формируется полный ориентированный граф, дополненный корневой вершиной; если исходное дерево содержит n зависимостей, то после преобразований получившийся граф содержит уже n^2 зависимостей.

Входной слой нейронной сети состоит из 429 узлов, второй слой — из 512 узлов, третий — из 128 узлов, четвертый — из 2 узлов. В качестве функции активации на промежуточных слоях используется ReLU, на выходном слое — функция Softmax. В качестве функции потери используется

бинарная кросс-энтропия. В качестве оптимизатора используется переработанный Adafactor с предварительно настроенными параметрами [5]. Функция оценки точности и функция потерь представлены на рисунке 4.

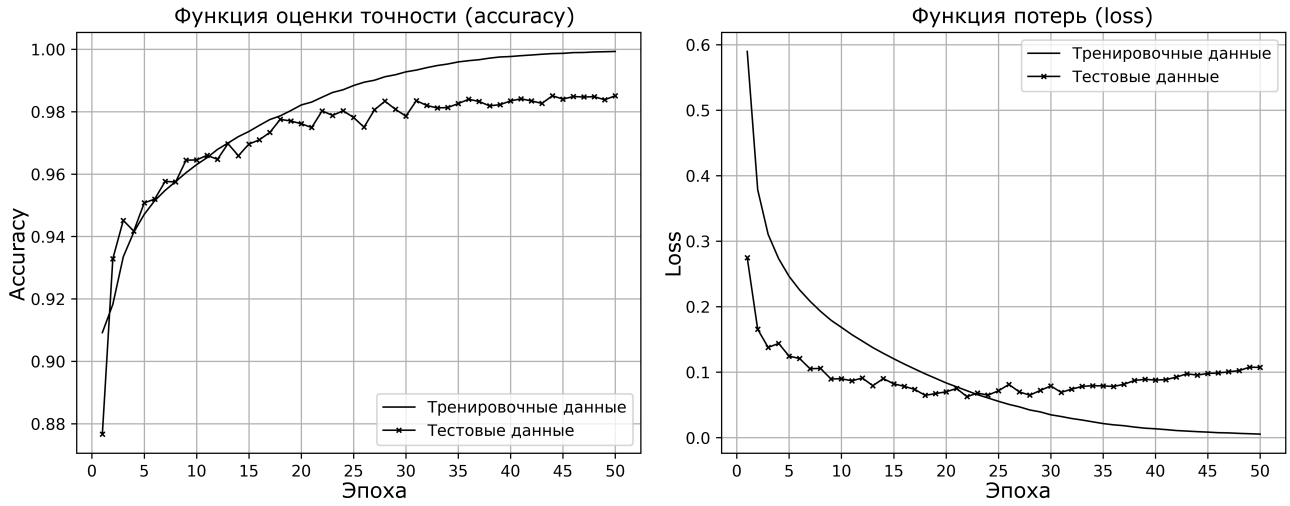

Рис. 4. *Функция оценки точности и функция потерь на 50 эпохах обучения*

На программу анализа текста с выделением синтаксических зависимостей получено свидетельство о государственной регистрации [6].

Для представления слов в виде единообразных объектов, удобных для машинных вычислений и наиболее полно отражающих языковой смысл, используются векторные пространства слов, получаемые с помощью обработки корпусов текстов специальными методами. Одним из основных методов является семейство word2vec [7, 8]. В работе [9] излагается идея метода построения синтаксически ориентированного векторного пространства слов на основе архитектуры Skip-Gram. Таким образом, контекст предсказываемого слова формируется не линейно, а согласно окрестности узла слова внутри синтаксического дерева. Очевидным преимуществом использования грамматики зависимостей в качестве контекста является инвариантность синтаксического дерева относительно порядка слов. Разработан метод синтеза синтаксически ориентированных векторных представлений на основе архитектуры непрерывного мешка слов [10]. Полученное таким образом векторное пространство используется для конструирования функции семантического сходства (1).

Анализ синтактико-семантической структуры текста

Идея формального описания семантики естественного языка основана на моделировании процесса чтения текста. Для начала рассмотрим этот процесс с точки зрения человека. Текст состоит из предложений, предложения — из слов — символьных цепочек, обладающих внутренней структурой — морфологией и структурой сочетаемости — синтаксисом. Мы вправе ожидать, что порядок слов в языке не случаен, а подчиняется некоторым неявным закономерностям: определение всегда стоит после определяемого слова или всегда перед ним, связанные друг с другом слова располагаются рядом, а не раскиданы в разные концы предложения и т.п. Однако можно заметить, что процесс чтения и восприятия различных частей предложения или целого текста может значительно варьироваться: например, в процессе чтения мы можем пропустить причастный оборот, некоторую вводную конструкцию или даже часть текста, — такие приемы чтения могут быть осуществлены без особого ущерба для понимания.

Пусть текст представлен ориентированным графом, причем таким, что структура входящих в него предложений формируется на основе деревьев синтаксических зависимостей. Эту структуру назовем семантическим графом. На графе определен некоторый агент, произвольно перемещающийся от одного слова к другому. Агент способен перейти ко всякому слову из любого другого слова за конечное число шагов (рисунок 5).

Если агент достигает слова, от которого зависит только одно слово, он может вернуться к предыдущему ветвлению — это позволяет ему выбрать более одного следующего слова для «чтения» (рисунок 6).

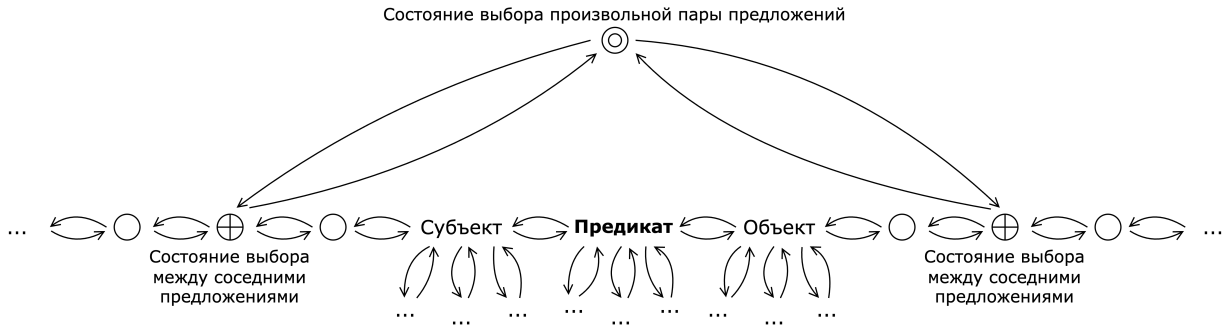

Рис. 5. *Строение предложения в семантическом графе*

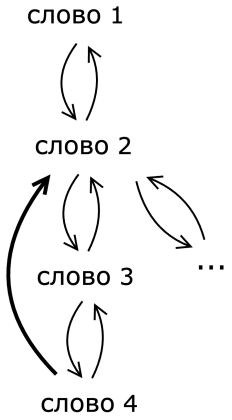

Рис. 6. *Обратный переход на участке семантического графа*

Предложение, как правило, описывает некоторую ситуацию с одним или несколькими «участниками». Организующим началом всякого предложения является предикат: ему в грамматике зависимостей подчинены участники ситуации, то есть субъект действия и объект действия. Иногда объект, субъект или предикат могут отсутствовать, но их можно восстановить посредством обращения к контексту. Эта цепочка — субъект, предикат, объект — является центральной конструкцией всего предложения, и если эллипсис в ней игнорируется на уровне синтаксиса, то на уровне семантики все три элемента присутствуют, — пропущенное слово цепочки заменяется знаком ∅. Все остальные конструкции предложения зависят от слов главной цепочки. Агент, «читающий» текст, волен выбирать, какой участок предложения ему интересен, а какой участок он готов пропустить.

Длина предложения зачастую обусловлена объединением нескольких грамматических основ в одном предложении, таким образом в нем соседствуют несколько главных конструкций. В этом случае из предложения выделяются главные цепочки с сохранением синтаксических зависимостей от них.

Агент одновременно с каждым переходом от одного слова w_i к w_j обращается к некоторому внешнему контексту — вопросу, заданному к тексту (рисунок 7). Вопрос представлен в виде мешка слов $Q = \{q_1, ..., q_m\}$.

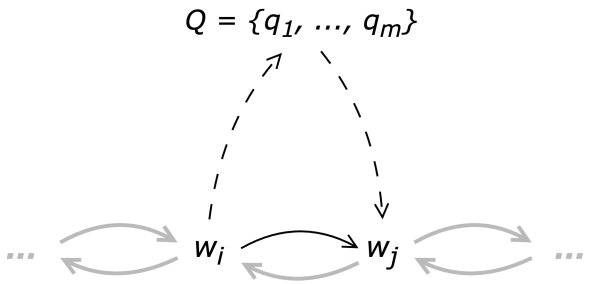

Рис. 7. *Переход в семантическом графе от слова w_i к слову w_j с учетом контекста вопроса*

Пусть задан словарь — множество лексем $Dict$, охватывающий все слова семантического графа и все слова вопроса, $w_i, w_j \in Dict$. Введем функцию семантического сходства двух слов sim: $Dict \times Dict \to (0,\ 1]$:

$$\mathrm{sim}(w_i, w_j) = \begin{cases} 1, & i = j \\ \frac{1}{2}, & w_i = \varnothing \ \vee \ w_j = \varnothing \\ r, & \frac{1}{2} < r < 1 \\ s, & 0 < s < \frac{1}{2} \end{cases}. \tag{1}$$

Функция $\mathrm{sim}(w_i, w_j)$ равна r, если слова w_i и w_j семантически близки. Иначе, функция равна s, если слова w_i и w_j семантически далеки.

Определим оценку контекста:

$$\varepsilon_{sim}^{ij} = \frac{1}{M} \frac{\sum\limits_{m=1}^{M} \mathrm{sim}(w_j, q_m)}{\mathrm{sim}(w_i, w_j)}. \tag{2}$$

Определим оценку перехода от слова w_i к слову w_j, учитывая контекст и встречаемость предложений со словом w_j (S — общее количество предложений в тексте, S_{w_j} — количество предложений, содержащих слово w_j, $\alpha \geqslant 1$, $\beta > 0$):

$$\Delta_{sim} = \left(\frac{S}{S_{w_j}} + \alpha \right)^{\beta \cdot \varepsilon_{sim}^{ij}}. \tag{3}$$

Особенность грамматики зависимостей заключается в том, что она позволяет рассматривать только бинарные отношения между словами, игнорируя отношения между словосочетаниями. Таким образом, агент, обходящий семантический граф, сформированный из бинарных наборов подчинения слов, знает только о предыдущем и следующем словах и «забывает» все ранее прочитанные слова. Определим вероятность перехода:

$$p_{ij} = \frac{\left(\frac{S}{S_{w_j}} + \alpha \right)^{\beta \cdot \varepsilon_{sim}^{ij}}}{\sum\limits_{k=1}^{K} \left(\frac{S}{S_{w_k}} + \alpha \right)^{\beta \cdot \varepsilon_{sim}^{ik}}}. \tag{4}$$

Решим задачу вопросно-ответного поиска, вычисляя соответствие предложения S_k вопросу Q (π_i — величина стационарного распределения для слова w_i):

$$\mathrm{score}(S_k) = -\frac{1}{|S_k|} \sum_{i \in S_k} \pi_i \sum_j p_{ij} \ln p_{ij}. \tag{5}$$

Заключение

Методы и программы, представленные в данной работе, могут служить основой для разработки и совершенствования информационно-поисковых систем с естественно-языковым интерфейсом и иных программных комплексов, реализующих обработку естественного языка. Корпус текстов может служить основой для обучения моделей обработки естественного языка.

ЛИТЕРАТУРА

1. *Could ChatGPT Pose a Threat to Google's Dominance in Search?* Режим доступа: https://www.entrepreneur.com/science-technology/could-chatgpt-pose-a-threat-to-googles-dominance-in-search/449033.
2. *Проект «Открытый корпус»*. Режим доступа: http://opencorpora.org.
3. Дяченко П. В., Иомдин Л. Л., Лазурский А. В. и др. Современное состояние глубоко аннотированного корпуса текстов русского языка (СинТагРус). *Труды института русского языка им. В. В. Виноградова*. 2015;6:272–300. EDN: VJQBEX.
4. Перцев Ю. В., Япарова Н. М. *Синтаксически аннотированный корпус веб-текстов русского языка*. Свидетельство о государственной регистрации базы данных № 2023621467 от 02.05.2023.

5. Shazeer N., Stern M. *Adafactor: Adaptive Learning Rates with Sublinear Memory Cost*. DOI: 10.48550/arXiv.1804.04235.
6. Перцев Ю. В., Япарова Н. М. *Программа анализа русскоязычных текстов с выделением некоторых синтаксических зависимостей*. Свидетельство о государственной регистрации программы для ЭВМ № 2022681794 от 10.11.2022.
7. Mikolov T., Yih W., Zweig G. Linguistic Regularities in Continuous Space Word Representations. *Proceedings of NAACL-HLT*. 2013. P. 746–751.
8. Mikolov T., Chen K., Corrado G. S., Dean J. Efficient Estimation of Word Representations in Vector Space. *International Conference on Learning Representations*. DOI: 10.48550/arXiv.1301.3781.
9. Levy O., Goldberg Y. Dependency-Based Word Embeddings. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* 2014;2:302–308. DOI: 10.3115/v1/P14-2050.
10. Перцев Ю. В., Япарова Н. М. *Программа синтеза синтаксически ориентированных векторных представлений слов*. Свидетельство о государственной регистрации программы для ЭВМ № 2024617697 от 01.04.2024.