



DOI: 10.22363/2313-2299-2026-17-2-459-473

EDN: LJHONM

Научная статья / Research Article

TextAnalyst и GPT: модель текста (модель текстового описания сущего) и семантика текста

А.А. Харламов  

Московский государственный лингвистический университет, *Москва, Российская Федерация*
Институт высшей нервной деятельности и нейрофизиологии РАН, *Москва, Российская Федерация*

Московский физико-технический институт (национальный исследовательский университет), *Долгопрудный, Российская Федерация*

Национальный исследовательский университет «Высшая школа экономики», *Москва, Российская Федерация*

 kharlamov@analyst.ru

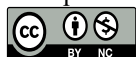
Аннотация. В статье описан один из подходов к представлению семантики (смысла) текста в виде однородной семантической сети со взвешенными вершинами и их связями. Это представление противопоставляется представлению смысла текста в современных искусственных нейронных сетях типа GPT, где текст также (если исследовать представление внимательно) представляется в виде сети в многомерном пространстве признаков, реализованном множестве искусственных нейронов, составляющих искусственную нейронную сеть. Различие заключается в невозможности разобраться с этим представлением в сети, использующей миллиарды параметров представления — в невозможности выявления в явном виде этой самой семантической сети. Потому многие механизмы обработки текстов (автоматическое аннотирование текстов, сравнение текстов по смыслу, классификация текстов) содержательно остаются на совести разработчиков таких инструментов из-за невозможности контролировать решения, принимаемые этими инструментами: слишком велик объем обрабатываемой информации и отсутствуют инструменты проверки качества работы.

Ключевые слова: смысл текста, однородная семантическая сеть текста

Конфликт интересов: автор заявил об отсутствии конфликтов интересов.

История статьи: дата поступления: 23.12.2025; дата приема в печать: 15.02.2026.

Для цитирования: Харламов А.А. TextAnalyst и GPT: модель текста (модель текстового описания сущего) и семантика текста // Вестник Российского университета дружбы народов. Серия: Теория языка. Семиотика. Семантика. 2026. Т. 17. № 2. С. 459–473. <https://doi.org/10.22363/2313-2299-2026-17-2-459-473> EDN LJHONM



TextAnalyst and GPT: Text Model (Text Description Model) and Text Semantics

Alexander A. Kharlamov  

Moscow State Linguistic University, *Moscow, Russian Federation*

Institute of Higher Nervous Activity and Neurophysiology, RAS, *Moscow, Russian Federation*

The Moscow Institute of Physics and Technology, *Dolgoprudny, Russian Federation*

National Research University “Higher School of Economics”, *Moscow, Russian Federation*

kharlamov@analyst.ru

Abstract. The article describes one of the approaches to representing the semantics (meaning) of a text in the form of a homogeneous semantic network with weighted vertices and their connections. This representation is contrasted with the representation of the meaning of the text in modern artificial neural networks such as GPT, where the text is also (if you examine the representation carefully) represented as a network in a multidimensional feature space implemented by a set of artificial neurons that make up an artificial neural network. The difference is that it is impossible to interpret this representation in a network that uses billions of representation parameters—the impossibility of explicitly identifying this very semantic network. Therefore, many text processing mechanisms (automatic annotation of texts, comparison of texts by meaning, classification of texts) meaningfully remain the responsibility of the developers of such tools due to the inability to control the decisions made by these tools: The volume of processed information is too large and there are no tools for checking the quality of work.

Keywords: meaning of the text, homogeneous semantic network of the text

Conflict of interests: The author declares no conflict of interest.

Article history: Received: 23.12.2025; Accepted: 15.02.2025.

For citation: Kharlamov, A.A. (2026). TextAnalyst and GPT: Text Model (Text Description Model) and Text Semantics. *RUDN Journal of Language Studies, Semiotics and Semantics*, 17(2), 459–473. (In Russ.). <https://doi.org/10.22363/2313-2299-2026-17-2-459-473> EDN LJHONM

Введение

Искусственные нейронные сети! GPT! Народ сошел с ума от восторга перед этими явлениями природы. Если не смотреть на обратную сторону медали — совершенно непохожую по качеству реакцию GPT на такое же к нему обращение (когда ответ GPT — полная чушь!) — то этот восторг объясним. Это похоже на то, как старенькая хозяйка, распутывая провод пылесоса, говорит ему: «подожди, миленький, сейчас я тебя освобожу!». Что хочу, то и вижу.

Далее не будет почти ничего по поводу GPT, только в конце мы попытаемся коснуться как самого явления, так и границ его применимости.

Далее во введении требуется показать предметную область со ссылками на наиболее характерные работы в ней. Как следует из названия, в статье представлена предметная область «Семантика». Объем этой предметной области очень велик, что можно понять хотя бы из различных определений этой предметной области, которые мы приводить не будем. Все дело в том, что под этим названием «Семантика» собралось много всего в разных предметных

областях от психологии до лингвистики, от систем распознавания речи до систем для анализа текстов.

Поскольку автор чисто практически в основном работал с анализом текстов, понятие «Семантика» будет относиться именно к анализу текстов и касаться содержания (смысла) текста. Далее во введении будет рассмотрено несколько подходов к анализу семантики текста, что и представит предметную область, к которой относится данная работа. Будут рассмотрены подходы, где слова в тексте рассматриваются не связанными по смыслу (так называемый «мешок слов»), первым таким подходом был LSA (так называемый Латентный семантический анализ), будут рассмотрены подходы, где слова рассматриваются в цепочке некоторой длины (“word2vek”), наконец, будут рассмотрены подходы на основе графов.

Мешок слов. Первые подходы к анализу содержания текста были реализованы на основе так называемого «мешка слов», когда слова в тексте рассматриваются как несвязанные между собой элементы. Мешок слов (англ. *bag-of-words*) — упрощенное представление текста, которое используется в обработке естественных языков и информационном поиске. В этой модели текст (одно предложение или весь документ) представляется в виде мешка (множества) его слов без какого-либо учета грамматики и порядка слов, но с сохранением информации об их количестве. Это упрощенное представление текста, которое показывает, какие слова встретились в тексте, но при этом не учитывает их порядок. Такое представление легко запрограммировать, оно удобно для использования в задачах автоматической обработки текста. Несмотря на свою простоту, оно оказывается достаточно полезным и позволяет успешно решать такие задачи, как классификация текста, т.е. отнесение текста к определенной группе / категории.

Как устроен мешок слов? Представление мешка слов — это таблица с числами, в которой столбцы таблицы — уникальные слова, а строки — документы коллекции. В ячейках таблицы находится число вхождений слова в документ. Значит, в каждой строке получится набор чисел (он же вектор), характеризующий состав документа. В разных вариантах используются разные представления элементов этой таблицы (LSA, pLSA, LDA) [1, 2], но суть этой таблицы не меняется.

Word2vek. «Word-to-vec» — общее название для совокупности моделей на основе искусственных нейронных сетей, предназначенных для получения векторных представлений слов на естественном языке. Модель-вектор — включает, кроме анализируемого слова текста, еще n слов текста до и n слов после, то есть контекст слова. Используется для анализа семантики естественных языков, основанного на дистрибутивной семантике, машинном обучении и векторном представлении слов. Программное обеспечение под названием «word2vec» исходно было разработано группой исследователей Google в 2013 году [3].

Графы в анализе текстов. Графовый анализ — это метод обработки и визуализации данных, при котором все элементы системы представлены в виде узлов, а связи между ними — в виде рёбер. Такой подход позволяет увидеть, как именно элементы взаимодействуют: откуда происходили входы, какие элементы использовались для транзакций, где пересекаются цепочки элементов [4].

Как выяснилось, ни тот, ни другой, ни третий подходы не несут в себе почти ничего, что относится к содержанию (смыслу) текста, что можно было бы использовать практически при анализе текста, например, при сравнении текстов по смыслу. Но они содержат элементы представления, которые можно использовать, пытаясь выявить это самое «содержание» (смысл) текста.

Эти элементы рассматриваемого представления следующие: частота встречаемости элементов текста (слов), их объединение в цепочки, графическое представление «содержания» текста в виде графа.

Сейчас можно сказать несколько слов об этих понятиях, которые подробно будут рассмотрены позже. Частота встречаемости, пересчитанная в смысловой вес, — одно из основных понятий, характеризующих смысл текста; цепочки понятий (но не в тексте, а на семантической сети текста), характеризуют важность слов в тексте; а граф-отношений слов в тексте, взвешенный по вершинам и связям, является основной смысловой характеристикой текста.

Поговорим о семантической сети текста, которая почти исключительно характеризует содержание (смысл) текста.

Сеть как способ представления семантики текста

Семантическая сеть текста — понятие не новое. Хорошо известны так называемые неоднородные семантические сети [5], которые используются для представления содержания текста (сейчас одновременно с ними используются так называемые онтологии [6]). В них пытались соединить языковые и экстралингвистические представления. Если говорят о красном яблоке, то это и слово «красный», и 450 нм, которые характеризуют длину волны видимого спектра, соответствующую красному цвету.

Представляемая в семантической сети информация

Сначала поговорим о том, какая информация должна быть представлена в семантической сети, чтобы она характеризовала смысл текста. Дело в том, что семантика предложения текста в лингвистике представлена весьма хорошо, а семантика целого текста почти не рассматривается.

В семантике текста должно быть все, что касается языка, и ничего больше. Рассмотрим в качестве примера какой-то текст. Он состоит из пред-

ложений, содержание которых описывается некоторыми (хорошо известными в лингвистике) способами. Например, содержание предложения текста может быть представлено расширенной предикатной структурой (рис. 1).

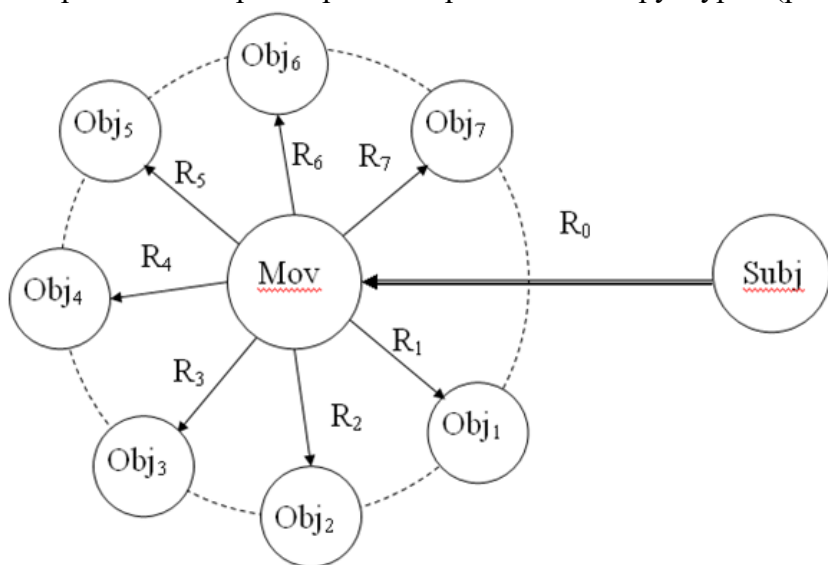


Рис. 1. Расширенная предикатная структура предложения (РПС).

Структура: Subj — субъект, Obj_i — актаны предиката, R_i — отношения предиката, R₀ — отношение «быть субъектом»

Источник: составлено А.А. Харламовым.

Figure 1. Extended Supply Predicate Structure (APF). Structure: Subj is the subject, Obj_i is the actant of the predicate, R_i is the relation of the predicate, R₀ is the relation to “being a subject”

Source: compiled by Alexander A. Kharlamov.

Предварительно сделаем одно замечание: введем в РПС как фрагмент будущей семантической сети направление от субъекта к главному объекту, ко второстепенным объектам, и к атрибутам. Что из расширенной предикатной структуры предложения должно перейти в семантическую сеть текста? Элементами предложений одного текста могут быть одинаковые сущности. Объединим расширенные предикатные структуры предложений текста по одинаковым сущностям. Получим сеть, которую (поскольку она состоит из РПС предложений текста) можем назвать семантической сетью этого текста.

Чтобы она действительно характеризовала смысл текста необходимо сделать еще один шаг и выполнить одно требование. Этот шаг заключается в том, что при объединении РПС в сеть возникают цепочки событий, которые характеризуют текст в целом. Что это значит? Не частота встречаемости слов в тексте, а их взаимосвязи в тексте (которые отражаются в семантической сети текста), характеризуют их важность в тексте [7]. И потому при выявлении важности понятия в тексте осуществляется пересчет частоты встречаемости понятий в их смысловой вес [8].

Однородная семантическая сеть

Требование же, о котором шла речь выше, заключается в том, что в прототипе искусственной системы для смыслового анализа текста (в мозгу человека), в котором осуществляются подобные представления, работают две сигнальные системы (первая — экстралингвистическая, и вторая — языковая), и, соответственно, реализованы две семантические сети — экстралингвистическая и языковая [8]. Они работают вместе, дополняя друг друга (это в экстралингвистической семантической сети существует параметр, характеризующий термин «красный», находящийся в языковой семантической сети — 450 нм).

Поэтому человек, глядя на представленную ему языковую семантическую сеть текста, дополняет ее экстралингвистической семантической сетью (в основном это сеть, реализованная на представлениях визуальной и соматосенсорной модальностей), имеющейся в нашем мозге. Поэтому в языковой семантической сети не нужны некоторые подробности: нам не обязательно знать точно предикат, достаточно знать, что субъект и объект взаимодействуют. Собственно, эта информация не является лишней, но работа с сетью, в которой ее нет, проще: ее можно реализовать автоматически. Дело в том, что работа с неоднородными семантическими сетями (где взаимодействие субъекта с главным и второстепенными объектами обозначено отношением) существенно сложнее, чем работа с однородной семантической сетью. И по этой причине эта работа выполняется по сю пору вручную (пока не существует алгоритмов автоматической работы с неоднородными семантическими сетями [5]).

Таким образом, использование однородных семантических сетей позволяет, с одной стороны, автоматизировать процесс обработки текстовой информации (сложный момент здесь — вычислить семантические веса понятий — перевзвесить вершины сети [7]), а с другой стороны, это упрощает структуру сети (разные словоформы одного слова приводятся к одной корневой основе), что делает алгоритм обработки более робастным.

Обработка текста в технологии TextAnalyst

Использование семантической сети для обработки текстов (кроме моментов, связанных с обработкой отдельного текста, ее можно использовать для обработки совместно нескольких текстов, например, для сравнения текстов) позволяет реализовать, весьма удобную для представления текста функциональность, которая используется, например филологами, в виде гипертекстовой структуры [7].

Уже достаточно давно была разработана линейка продуктов TextAnalyst для смыслового анализа текстов [9], которая и реализует описанную выше технологию построения однородной семантической сети с перевзвешенными

в смысловой вес вершинами, функциональность которой по анализу текстов будет представлена ниже.

Разработанная система обработки текстовой информации основана на использовании структурных свойств языка и текста, которые могут быть выявлены с помощью статистического анализа, где статистика конкретного текста позволяет автоматически сформировать представление семантики текста в виде однородной семантической сети, на основе которого и реализуются функции организации текстовой базы в гипертекстовую структуру, автоматического реферирования, кластеризации и классификации текстов, а также функция смыслового поиска.

Этот подход позволяет на основе анализа статистики слов и их связей в предложениях текста реконструировать внутреннюю (смысловую) структуру текста. Статистический анализ фиксирует частоту появления элементов текста (слов или устойчивых словосочетаний), а также частоту появления взаимосвязей между выявленными элементами текста. Далее статистические показатели пересчитываются в семантические с помощью алгоритмов, подобных алгоритму перерасчета весов в сетях Хопфилда [10]. Идея подобных алгоритмов заключается в том, что при расчете какой-то характеристики элемента сети учитываются подобные характеристики элементов, с ним связанных, в виде численных показателей связей. После такого пересчета статистических характеристик текста в семантические характеристики понятий текста, которые мало соответствуют анализируемой предметной области, получают малый вес, а наиболее представительные — наделяются высокими показателями. Полученная семантическая сеть позволяет производить различные виды анализа текстовой информации. Сеть отражает внутреннюю структуру текста, значимость выделенных понятий, а также показывает степень связанности понятий в тексте. Такое представление текста может быть получено полностью автоматически. Так формируется семантическая сеть текста.

Семантическая сеть, представляющая текст, далее используется как для анализа исходного текста, так и для совместного анализа нескольких текстов анализируемого корпуса текстов. Анализируемый текст, входящий в гипертекстовую структуру вместе с построенной на основе этого текста семантической сетью, используется: (1) для представления смысла (структуры) этого текста в виде семантической сети; (2) для представления оглавления исходного текста в виде минимального древовидного подграфа этой сети; (3) вершины сети — слова текста — позволяют выявить предложения текста, содержащие эти слова; (4) предложения текста с наибольшим весом формируют аннотацию текста; наконец, (5) имеется возможность поиска элементов текста в этой гипертекстовой структуре с целью быстрого обращения к информации в тексте.

Возможность совместного анализа нескольких текстов позволяет: (1) выявлять степень смыслового пересечения текстов (пересечение их семантических сетей); (2) классифицировать тексты (выявлять степень смыслового пересечения семантической сети текста с семантическими сетями предметных областей — классов); (3) кластеризовать предметную область (корпус текстов) на кластеры.

Семантические веса элементов сети используются при расчете близости (релевантности) фрагментов текста к запросу со стороны пользователя системы. На их основе возможно выделение наиболее информативных участков текста. Использование ассоциативных связей элементов сети позволяет расширять поле поиска информации. Ответ на запрос пользователя в этом случае может содержать информацию, явно не указанную в запросе, но связанную с ней по смыслу.

Программная реализация технологии

Система TextAnalyst реализована как инструмент для автоматического формирования баз знаний на основе множества естественно-языковых текстов. Ядро системы выполнено как программный компонент (inproc server), соответствующий спецификации Component Object Model (COM) фирмы Microsoft [9].

Ядро системы реализует следующие функции: (1) нормализацию грамматических форм слов и вариаций словосочетаний; (2) автоматическое выделение базовых понятий текста (слов и словосочетаний) и их взаимосвязей с вычислением их относительной значимости в тексте; (3) формирование представления семантики текста (множества текстов) в форме семантической сети.

В состав ядра системы, помимо блока первичной обработки, входят следующие блоки (рис. 2): лингвистический процессор (2), блок выделения понятий текста (12), блок формирования семантической сети (10), блок хранения семантической сети (11).

Блок первичной обработки. Задачами этого блока являются извлечение текста из файла (входного потока данных) и подготовка его к обработке в лингвистическом процессоре. Подготовка текста заключается в очистке его от символов, неизвестных лингвистическому процессору, а также в корректной обработке таких единиц текста, как аббревиатуры, инициалы, заголовки, адреса, номера, даты, указатели времени.

Лингвистический процессор. Лингвистический процессор осуществляет предобработку входного текста (последовательности символов в определенной кодировке) на основе априорных лингвистических знаний, общих для выбранного языка (в настоящий момент поддерживаются несколько европейских языков, помимо русского и английского), и выполняет следующие функции: сегментацию предложений текста на основе знаков пунктуации и специальных грамматических слов и их фильтрацию; нормализацию слов и словосочетаний — фильтрацию флексий (окончаний) с сохранением только

корневых основ; фильтрацию в тексте семантически несущественных, вспомогательных слов: удаляются предлоги, числительные и самые общеупотребимые слова с широким значением; маркировку общеупотребимых слов.

Сегментация на предложения позволяет разбить текст на участки, которые могут содержать терминологические словосочетания предметной области, и избежать выделения неадекватных словосочетаний на стыках таких участков.

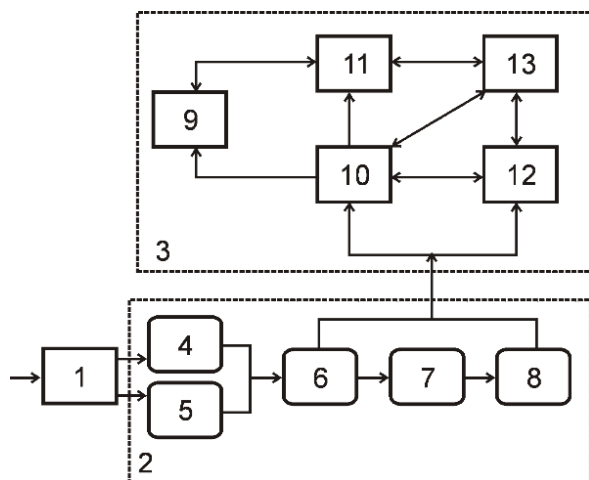


Рис. 2. Система анализа текстов содержит блок первичной обработки (1), лингвистический и семантический процессоры. Лингвистический процессор (2) состоит из словарей: (4) слов — разделителей, (5) служебных слов, (6) общеупотребимых слов, а также (7) флективных и (8) корневых морфем. Семантически процессор (3), в свою очередь, содержит: (9) блок отсылок в текст, (10) блок формирования семантической сети, (11) блок хранения семантической сети, (12) блок выделения понятий и (13) блок управления

Источник: составлено А.А. Харламовым.

Figure 2. A text analysis system comprises a primary processing unit (1), linguistic and semantic processors. The linguistic processor (2) consists of dictionaries: (4) divisional words, (5) service words, (6) common words, and (7) a phletical and (8) root morphs. Semantically, the processor (3), in turn, contains: (9) a unit of references into text, (10) a unit for forming a semantic network, (11) a unit for storing a semantic network, (12) a unit for separating concepts, and (13) a control unit.

Source: compiled by Alexander A. Kharlamov.

В результате предобработки семантически близкие словосочетания приводятся к одинаковой форме (нормализуются). Маркировка общеупотребимых слов необходима с целью исключения их выделения как самостоятельных терминов при дальнейшем анализе.

База общих языковых знаний лингвистического процессора содержит словари, по одному для реализации каждой из четырех функций: словарь слов-разделителей предложения, словарь вспомогательных слов, словарь флексий и словарь общеупотребимых слов.

Блок выделения понятий. Блок выделения ключевых понятий предметной области (слов и словосочетаний) реализует алгоритмы автоматического

формирования частотного словаря текста. Он реализован как иерархическая структура пар представлений, где первое представление в паре — это символ алфавита или номер предыдущей пары. Число уровней в иерархической структуре определяет априорно заданную максимально допустимую длину понятия предметной области и равняется двадцати.

На первом уровне иерархической структуры представлен словарь двухбуквенных специальных слов предметной области — слов, пропущенных через все фильтры лингвистического процессора и не отнесенных к общеупотребимым, а также двухбуквенных сочетаний из слов этого словаря. Там же хранятся двухбуквенные слова общеупотребимой лексики, входящие в устойчивые словосочетания, и их начальные двухбуквенные фрагменты. На втором уровне иерархической структуры хранятся словари трехбуквенных слов и сочетаний букв из словарей специальных и общеупотребимых слов, встреченных в тексте, в виде индексов элементов соответствующих словарей первого уровня, дополненных еще одной буквой. На последующих уровнях представление информации полностью однородно: там хранятся индексы элементов хранения более низкого уровня иерархии, дополненные одной буквой.

В процессе формирования представления информации в иерархической структуре подсчитывается частота встречаемости каждого сочетания букв в соответствующих элементах иерархии. Частота слов (сочетаний букв, не имеющих продолжения на следующем уровне) используется для последующего анализа: если есть скачок частоты — слово завершено.

Сформированное таким образом представление лексики текста подвергается затем пороговому преобразованию по частоте встречаемости. Порог отражает степень детальности описания текста. В процессе статистического анализа в иерархической структуре выделяются устойчивые термины и терминологические словосочетания, которые служат далее в качестве элементов для построения семантической сети. При этом общеупотребимые слова, а также словосочетания, содержащие только общеупотребимые слова, опускаются.

Блок формирования семантической сети. Блок формирования семантической сети реализован как база данных, в которой представляются семантические связи понятий предметной области. Поскольку типы семантических связей [11] в системе не определяются, такие связи представляют собой просто ассоциативные связи.

В качестве критерия для определения наличия семантической связи между парой понятий используется частота их совместной встречаемости в предложениях текста. Превышение частотой некоторого порога позволяет говорить о наличии между понятиями ассоциативной (семантической) связи, а совместные вхождения понятий в предложения с частотой меньше порога считаются просто случайными.

Элементы семантической (ассоциативной) сети и их связи имеют числовые характеристики, отражающие их относительный вес в данной предметной области — семантический вес. При достаточно представительном множестве текстов, описывающих предметную область, значения частот встречаемости понятий действительно отражают соответствующие семантические (субъективно оцениваемые) веса. Однако для небольших выборок, в частности при анализе отдельного текста, не все частотные характеристики соответствуют действительным семантическим весам — весам важности понятий в тексте. Для более точной оценки семантических весов понятий используются веса всех связанных с ними понятий, т.е. веса целого семантического сгущения. В результате такого анализа наибольший вес приобретают понятия, обладающие мощными связями и находящиеся как бы в центре семантических сгущений.

Основные функции системы «TextAnalyst»

На основе результатов работы модуля индексации реализованы следующие функции обработки текстовой информации: (1) функция формирования гипертекстовой структуры; (2) навигации по базе знаний; (3) формирования тематического дерева; (4) реферирования текстов; (5) автоматической кластеризации множества текстов; (6) сравнения текстов (автоматической классификации текстов); и, наконец, (7) функция формирования ответа на смысловой запрос пользователя — формирование тематического реферата.

После формирования семантической сети исходный текст, объединенный гиперссылками с семантической сетью, становится гипертекстовой структурой. Семантическая сеть в этом случае оказывается удобным средством навигации по тексту. Она позволяет исследовать основную структуру текста, переходя от понятия к понятию по ассоциативным связям. Пользуясь гиперссылками, пользователь может быстро найти множество предложений текста, содержащих эти понятия. С помощью тех же гиперссылок он может перейти от любого предложения непосредственно к его контексту в тексте. С этой же целью пользователь может пользоваться минимальным древовидным подграфом семантической сети — тематическим деревом. В нем оказываются иерархически представленными основные и соподчиненные понятия сети, причем понятия нижнего уровня объясняют содержание понятий более высокого уровня. Тематическим деревом также можно пользоваться для навигации по базе знаний — оно напоминает оглавление текста.

Семантическая сеть с числовыми характеристиками ее компонентов — понятий и их связей — позволяет вычислить вес каждого предложения в тексте. Множество предложений текста, выбранных в порядке их появления в тексте, вес которых превысил некоторый пороговый уровень, можно считать рефератом текста.

Семантическая сеть исследуемого текста (или группы текстов) может быть разбита на подсети удалением из нее слабых связей. Каждая такая подсеть группируется вокруг некоторого понятия с максимальным весом в данной подсети. Это понятие обозначает тему части текста или отдельных текстов, которые оказываются сгруппированными в данной подсети. Такая автоматическая кластеризация позволяет разбить множество текстов на рубрики, а также визуализировать динамику развития этих рубрик во времени.

Используя числовые характеристики семантической сети, можно сравнивать сети двух текстов с точки зрения вычисления их пересечения (общей части). То есть можно сравнивать степень совпадения текстов по смыслу. Если в качестве одного из текстов берется целая рубрика, то имеется возможность оценить степень принадлежности исходного текста к данной рубрике, то есть автоматически классифицировать тексты.

Система для смыслового анализа текстов позволяет реализовать также смысловой поиск (сформировать тематический реферат). Функция смыслового поиска, основываясь на ассоциативном иерархическом представлении содержания информации в базе, функциях кластеризации и классификации, осуществляет выборку информации, соответствующей запросу пользователя, и структурирует ее в соответствии с близостью к запросу.

Смысловой поиск, используя ассоциации, способен выдавать пользователю информацию, явно не указанную в тексте запроса, но связанную с ней по смыслу. Использование такого подхода ведет не к увеличению выдаваемой пользователю информации, а к ее тщательному отбору и анализу по главному критерию — смысловой близости к запросу.

Представление текста в технологии GPT

Как же устроена база знаний трансформеров, так стремительно ворвавшихся в нашу жизнь? Если отбросить подробности представления и использования информации в технологии GPT (как наиболее известного представителя этой когорты), то (опять же независимо от объема обрабатываемой информации) представление корпусов текстов в технологии GPT сводится к довольно длинным цепочкам текстовой информации заданной длины, представленным в искусственных нейросетевых структурах. Эти текстовые цепочки представляются в некоем многомерном пространстве, где возможно вычисление расстояния до такой цепочки (посредством механизма многоголового внимания [12]), с выбором искомой цепочки с наименьшим до нее расстоянием. Манипулирование этой информацией сводится к заданию вопроса системе и к синтезу ответа искусственной нейронной сети (ИНС) в виде связного текста. Вы — вопрос, Вам — ответ, где синтезируемый ответ оптимально приближен в исходном многомерном пространстве к Вашему вопросу с точки зрения упомянутого механизма многоголового внимания.

В силу того, что длина цепочки в такой ИНС может быть достаточно большой, создается впечатление, что такая ИНС самостоятельно генерирует

осмысленный связный текст. На самом деле это впечатление обманчиво: (1) суть этого текста сводится к ответу на вопрос; и (2) сгенерирован может быть только текст, который по смыслу содержится в базе данных ИНС (хотя бы собранный из разных частей). Попробуйте пообщаться с такой ИНС (хоть и с GPT) по поводу статьи Н.А. Козырева «Причинная или несимметричная механика в линейном приближении» [13]. Мы пытались, но ничего путного у ИНС выяснить не удалось (только Вы должны хорошо знать предметную область, чтобы корректно интерпретировать ответы ИНС): разговор с ИНС был очень длинным — несколько дней, поэтому привести его не представляется возможным.

Заключение

В заключение можно сказать следующее. Современные искусственные нейронные сети продвинулись далеко в реализации алгоритмов обработки информации разных видов: устных и письменных текстов, изображений, даже запахов. Но речь идет фактически только о распознавании.

В настоящий момент реализованы достаточно изоцированные механизмы анализа (до сих пор не реализованы алгоритмы распознавания сцен — совокупностей объектов — разработчики пока не догадались, что для этого надо совместить глубокие нейронные сети с сетями Хопфилда). Алгоритмов синтеза информации как целенаправленного поведения, реализовано до сих пор не было.

Дело в том, что целенаправленное поведение предполагает достижение цели, которая до сих пор ставится только человеком. Целенаправленное поведение по существу является деятельностью, направленной на удовлетворение потребностей (неважно каких). А удовлетворение потребностей искусственных систем может противопоставить искусственные системы естественным интеллектуальным системам (человеку). И хоть до возникновения сугубо искусственного интеллекта еще очень далеко (разве можно сравнить результаты проекта “Blue Brain” с естественным мозгом человека?), а возможно, искусственного интеллекта, сравнимого с человеческим, никогда и не появится; даже мысленно их желательно не противопоставлять.

Однако интеллектуальные системы, которые решают поставленные человеком задачи лучше него, необходимо разрабатывать и ими пользоваться. И известный прогресс в этом направлении явно есть. Просто его не следует преувеличивать.

Список литературы

1. *Deerwester S., Dumais S.T., Landauer T.K., Furnas G.W., Harshman R.* Indexing by Latent Semantic Analysis // *Journal of American Society of Information Science*. 1990. Vol. 41. No. 6. P. 391–407. [https://doi.org/10.1002/\(SICI\)1097-4571\(199009\)41:6<391::AID-AS11>3.0.CO;2-9](https://doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-AS11>3.0.CO;2-9)
2. *Blei D.M., Ng A.Y., Jordan M.I.* Latent Dirichlet Allocation // *The Journal of Machine Learning Research*. 2003. № 3. P. 993–1022.

3. Mikolov T., Chen K., Corrado G., Dean J. Efficient Estimation of Word Representations in Vector Space. In: *Proceedings of Workshop at ICLR, Scottsdale, 2–4 May 2013*. Scottsdale : AZ, Scottsdale, U.S.A.: Poisoned Pen Press. 2013. P. 1–11.
4. Григорьева Е.Г., Клячин В.А. Исследование статистических характеристик текста на основе графовой модели лингвистического корпуса // Известия Саратовского ун-та. Новая Серия Математика. Механика. Информатика. 2020. Т. 20. Вып. 1. С. 116–126. <https://doi.org/10.18500/1816-9791-2020-20-1-116-126> EDN: KNPYYV
5. Голенков В.В., Гулякина Н.А. Графодинамические модели параллельной обработки знаний: принципы построения, реализации и проектирования // Открытые семантические технологии проектирования интеллектуальных систем. 2012. № 2. С. 3–13. EDN: YUCQXM
6. Гаврилова Т.А., Хорошевский В.Ф. Базы знаний интеллектуальных систем. СПб. : Питер, 2001.
7. Kharlamov A.A. Network N-gram Model of the Text. Topic Tree of the Text — Minimal Tree Subgraph of the Semantic Network. In book: *Neuroinformatics and Semantic Representations: Theory and Applications*. Cambridge: Cambridge Scholars Publishing, 2020. P. 140–152.
8. Kharlamov A.A. Text and Quasi-Text Analysis: Upper Levels of Speech Information Processing. In: *Integral Robot Technologies and Speech Behavior*. A.A. Kharlamov, M. Pilgun (Eds.). Collective Monography. Chapter Three. Cambridge: Cambridge Scholars Publishing, 2023. P. 88–129.
9. Hopfield J.J. Neural networks and physical systems with emergent collective computational abilities // *Proc. Natl. Acad. Sci.* 1982. No. 79. P. 2554–2558
10. Осипов Г.С. Приобретение знаний интеллектуальными системами: Основы теории и технологии. М. : Наука. Физматлит, 1997. EDN: QOWVDX
11. Radford A., Wu J., Child R., Luan D., Amodei D., Sutskever I. Language Models are Unsupervised Multitask Learners. OpenAI Blog, 1, 9. 2019.
12. Козырев Н.А. Причинная или несимметричная механика в линейном приближении. Акад. наук СССР. Глав. астроном. обсерватория: Пулковое, 1958.

References

1. Deerwester, S., Dumais, S.T., Landauer, T.K., Furnas, G.W., & Harshman, R. (1990) Indexing by Latent Semantic Analysis. *Journal of American Society of Information Science*. 41(6). 391–407. [https://doi.org/10.1002/\(SICI\)1097-4571\(199009\)41:6<391::AID-ASII>3.0.CO;2-9](https://doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASII>3.0.CO;2-9)
2. Blei, D.M., Ng, A.Y., & Jordan, M.I. (2003) Latent Dirichlet Allocation. *The Journal of Machine Learning Research*. 3. 993–1022.
3. Mikolov T., Chen K., Corrado G., & Dean J. (2013) Efficient Estimation of Word Representations in Vector Space. In: *Proceedings of Workshop at ICLR, Scottsdale, 2–4 May 2013*. Scottsdale : AZ, Scottsdale, U.S.A.: Poisoned Pen Press. 1–11.
4. Grigorieva, E.G., & Klyachin, V.A. (2020) The Study of The Statistical Characteristics of The Text Based on the Graph Model of the Linguistic Corpus. *Izvestiya of Saratov University. New Series. Mathematics. Mechanics. Informatics*. 20(1). 116–126. (In Russ.) <https://doi.org/10.18500/1816-9791-2020-20-1-116-126> EDN: KNPYYV
5. Golenkov, V.V., & Guliakina, N.A. (2012) Graphodynamical Models of Parallel Knowledge Processing. *Open Semantic Technologies for Intelligent Systems Design*. 2. 3–13. (In Russ.) EDN: YUCQXM
6. Gavrilova, T.A., & Khoroshevsky, V.F. (2001). *Knowledge Bases of Intelligent Systems*. St. Petersburg: Piter. (In Russ.)
7. Kharlamov, A.A. (2020) Network N-gram Model of the Text. Topic Tree of the Text — Minimal Tree Subgraph of the Semantic Network. In book: *Neuroinformatics and Semantic Representations: Theory and Applications*. Cambridge: Cambridge Scholars Publishing. 140–152.

8. Kharlamov, A.A. (2023) Text and Quasi-Text Analysis: Upper Levels of Speech Information Processing. In: *Integral Robot Technologies and Speech Behavior*. A. A. Kharlamov, M. Pilgun (Eds.). Collective Monography. Chapter Three. Cambridge: Cambridge Scholars Publishing, 88–129.
9. Hopfield, J.J. (1982) Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci.* 79. 2554–2558.
10. Osipov, G.S. (1997). *Knowledge Acquisition by Intelligent Systems: Fundamentals of Theory and Technology*. Moscow: Nauka. Fizmatlit. (In Russ.) EDN: QOWVDX
11. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019) *Language Models are Unsupervised Multitask Learners*. *OpenAI Blog*, 1, 9.
12. Kozyrev, N.A. (1958). *Causal or Asymmetric Mechanics in Linear Approximation*. Leningrad: Acad. Sci. USSR, Main Astronomical Observatory. (In Russ.)

Сведения об авторе:

Харламов Александр Александрович, доктор технических наук, профессор, профессор кафедры прикладной и экспериментальной лингвистики, Московский государственный лингвистический университет (119034, Российская Федерация, г. Москва, ул. Остоженка, д. 38 с. 1); заведующий кафедрой интеллектуальных информационных систем и технологий, Московский физико-технический институт (национальный исследовательский университет) (141701, Российская Федерация, Московская область, г. Долгопрудный, Институтский переулок, д. 9); старший научный сотрудник лаборатории нейроонтогенеза, Институт высшей нервной деятельности и нейрофизиологии РАН (117485, Российская Федерация, Москва, ул. Бутлерова, д. 5А); профессор департамента программной инженерии, Национальный исследовательский университет «Высшая школа экономики», (101000, Российская Федерация, Москва, Мясницкая ул., 20); *сфера научных интересов*: нейроинформатика; *e-mail*: kharlamov@analyst.ru
ORCID: 0000-0003-2942-5101; Scopus Author ID: 57193909855; ResearcherID: E-3760-2014; SPIN-код: 6959-6240.

Information about the author:

Alexander A. Kharlamov, Dr.Sc. (Technics); Prof. at Applied and Experimental Linguistics Department, Moscow State Linguistic University (38 build 1 Ostozhenka Str., Moscow, Russian Federation, 119034); Senior Researcher at Neuroontogenesis Laboratory, Institute of Higher Nervous Activity and Neurophysiology, RAS (5A Butlerova Str, Moscow, Russian Federation, 117485); Intellectual Information Systems and Technology Department Chair, The Moscow Institute of Physics and Technology (9 Institutskiy per., Dolgoprudny, Moscow Region, Russian Federation, 141700); Prof. at Program Engineering Department of HSE, National Research University “Higher School of Economics” (20 Myasnitskaya Str., Moscow, Russian Federation, 101978); *Research interests*: Neuroinformatics; *e-mail*: kharlamov@analyst.ru
ORCID: 0000-0003-2942-5101; Scopus Author ID: 57193909855; ResearcherID: E-3760-2014; SPIN code: 6959-6240.