UOT-811.11-112
Səhifə: 70-76
https://doi.org/10.59849/2663-8967.2026.1.70

Гюнай Багирова
Азербайджанский Университет Языков,
доктор философии по филологии
Orcid: 0009-0003-8194-0255

ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ И ЛЕКСИКО-СЕМАНТИЧЕСКИЙ ПОДХОД К ЯЗЫКОВЫМ МОДЕЛЯМ

SÜNİ İNTELLEKT VƏ DİL MODELLƏRİNƏ LEKSİK-SEMANTİK YANAŞMA

XÜLASƏ

Məqalədə süni intellekt texnologiyalarının dil modellərinin qurulmasında və inkişafında rolu leksik-semantik yanaşma kontekstində araşdırılır. Müasir süni intellekt sistemləri, xüsusilə neyron şəbəkələrə əsaslanan böyük dil modelləri, leksik vahidlərin semantik strukturu, onların qarşılıqlı əlaqələri və kontekstual mənalarının emalı prinsipləri əsasında fəaliyyət göstərir. Tədqiqatda leksik-semantik sahə, semantik şəbəkə, konseptual struktur və mənanın çoxqatlılığı kimi anlayışların süni intellekt alqoritmlərində necə reallaşdığı təhlil edilir. Qeyd olunur ki, dil modellərinin effektivliyi yalnız statistik tezlik göstəricilərinə deyil, həm də semantik əlaqələrin dərin strukturuna əsaslanır.

Açar sözlər: *dilin modelləşdirilməsi, tokenləşdirmə, sözlər çantası, söz vektorları, kontekstual modellər, semantik şəbəkələr.*

Введение: В современном мире технологических нововведений искусственный интеллект занимает особое место. ИИ интегрирует во все области науки, в том числе и в область лингвистики. Ранее основывающиеся на традиционные лингвистические методы, исследования в современное время основываются на различные возможности информационных технологий. Такое развитие технологии изменило методы исследований и создало новые возможности. Цель выявления путей применения и научного потенциала искусственного интеллекта в науке лингвистики дает основу для изучения функций и перспективы технологических новшеств.

Воздействие ИИ на лингвистику явно выявляется не только в теоретической, но и в практической областях. Искусственный интеллект предлагает возможность моделировать с помощью систем ИИ человеческий язык математическими и статистическими методами. В образовании таких языковых моделей сформировались два основных, дополняющих друг друга, подхода: лексический (форма слова зачитывается основным) и семантический (значение слова зачитывается основным) подходы.

Лексический подход является первоначальным этапом моделирования языка в ИИ и моделирует слова, их формы и частоту в языке, рассматривая в основном форму слова и изучая его значение косвенно. Например, слово «cold» имеет различный оттенок значений в словосочетаниях «cold weather» и «cold behaviour», но лексический подход рассматривает его как одна единица. Этот подход тесно связан с структуральной лингвистикой и ранней корпус- лингвистикой. Основными свойствами лексического подхода являются рассмотрение слов как самостоятельных единиц, отсутствие контекста или его ограниченная форма, основание на статистику. При анализе лексических моделей текст делится на слова и сегменты (token) и исследуются статистические показатели. Хотя этот подход не различает многозначность и синонимичность, а также слабо моделирует

значение в контексте, он довольно эффективен в коллекциях небольших текстов и дает ясную статистическую интерпретацию.

Типическими методами лексического подхода являются Bag of Words, n-gram models, TF-IDF, Token. Модель Bag of Words определяет единицы текста независимо от контекста, основывается на показатели частоты слов в тексте и не рассматривает семантические отношения между ними. Суть его заключается в том, что, не рассматривая грамматическую структуру текста и порядок слов, вычисляется только наличие слов и частота их употребления. Сначала текст делится на токены (сегменты), потом создается векторное измерение каждого уникального слова, а в конце данные этого векторного измерения показывают частоту слов в тексте. Например, если взять два следующих текста «I read the book.» и «I love books.», то и модель Bag of Words можно представить в виде следующей таблицы:

Текст	I	read	the	book	Love
№ 1	1	1	1	1	0
№ 2	1	0	0	1	1

Преимуществами этой модели являются то, что она простая и быстрая и подходит для классических моделей ML. А недостатками являются то, что порядок слов и контекст теряются, синонимичность и многозначность не различаются.

Другая модель лексического подхода – n-gram определяет возможные отношения между последовательностями слов и частично отражает локальную структуру текста. Буква «n» в названии модели выражает количество последовательных элементов. Если составить n-gram модель предложения «I read them.», то она представится в следующей форме:

Unigram (1-gram): I

Bigram (2-gram): I read, read them

Trigram (3-gram): I read them

Преимуществами данной модели являются частичное сохранение порядка слов и эффективность в моделировании языка. А недостатками являются быстрое увеличение величины и проблематичность в редких n-граммах.

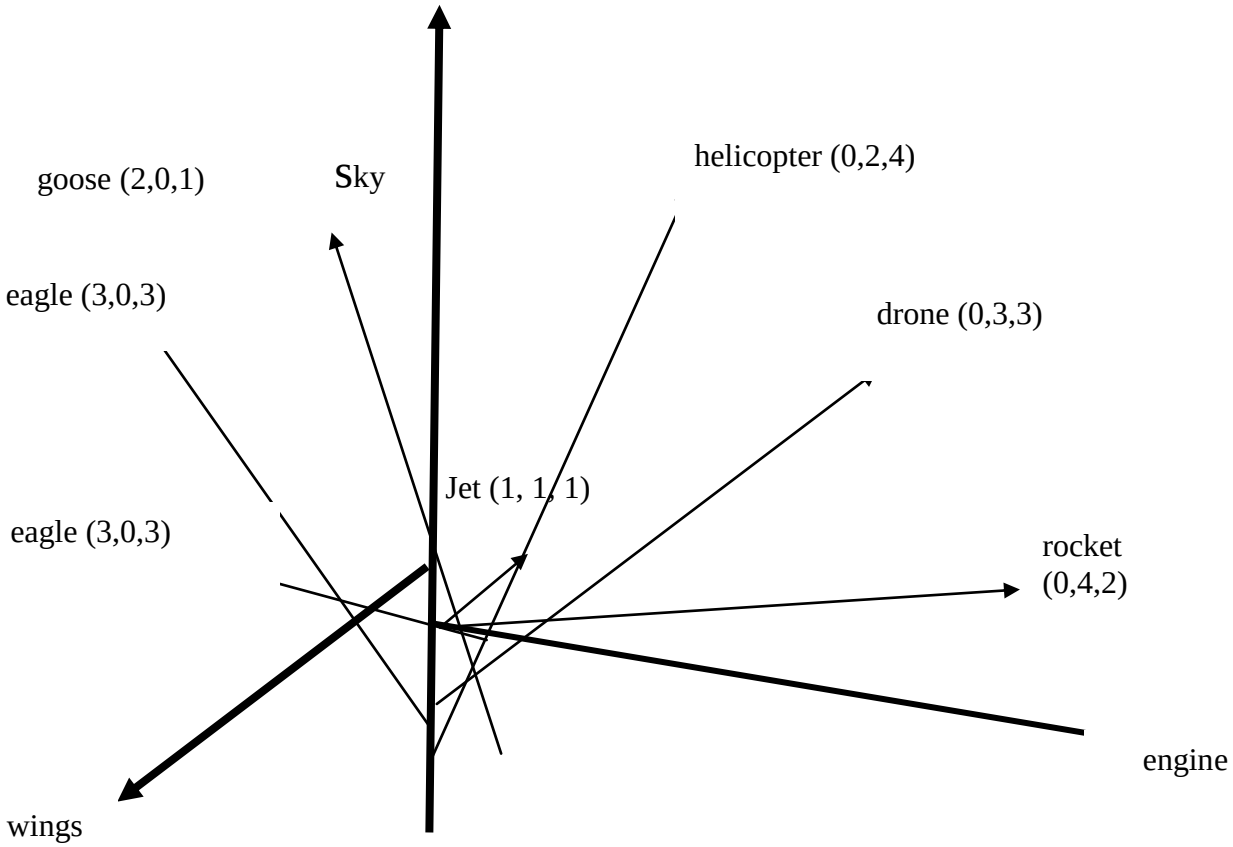
А модель TF-IDF (Term Frequency – Inverse Document Frequency) вычисляет не только частоту слов, но и степень их информативности для текста. Другими словами, TF (Term Frequency) показывает частоту употребления слов в тексте, а IDF (Inverse Document Frequency) показывает степень редкости слова во всем корпусе. Существует специальная формула вычисления: TF-IDF= TF ^{x} I. В результате анализа текста данной моделью выявляется уменьшение массы в нем таких слов как «and», «this», «a» и увеличение массы ключевых слов. Преимуществами этой модели являются большая информативность для представления текста и широкое пользование в поисковых системах. Однако, существуют некоторые ограничения: эта модель не может понимать контекст достаточно глубоко и не может моделировать синтаксические отношения.

Следующей моделью лексического подхода является токенизация (Tokenization), которая рассматривает наименьшие единицы отдельно взятыми. Токенизация является процессом деления текста на токены, т.е. наименьшие сегменты. Этим может быть слово (success), морфема (success-ful-ly), под-слово (suc, cess) или символ (s, u, c, e, f, l, y). Как и все остальные модели, токенизяция тоже отличается определенными преимуществами: она является первым этапом всех моделей и считается очень важным для морфологически богатых языков.

Ограничения и недостатки лексического подхода создали основу для формирования семантического подхода. В отличие от лексического подхода, семантический подход моделирует язык на основе контекста и связи между значениями. Анализ семантических моделей опирается на принцип семантики распределения. В этих моделях слова представляются в многомерных векторах в зависимости от их контекстуального

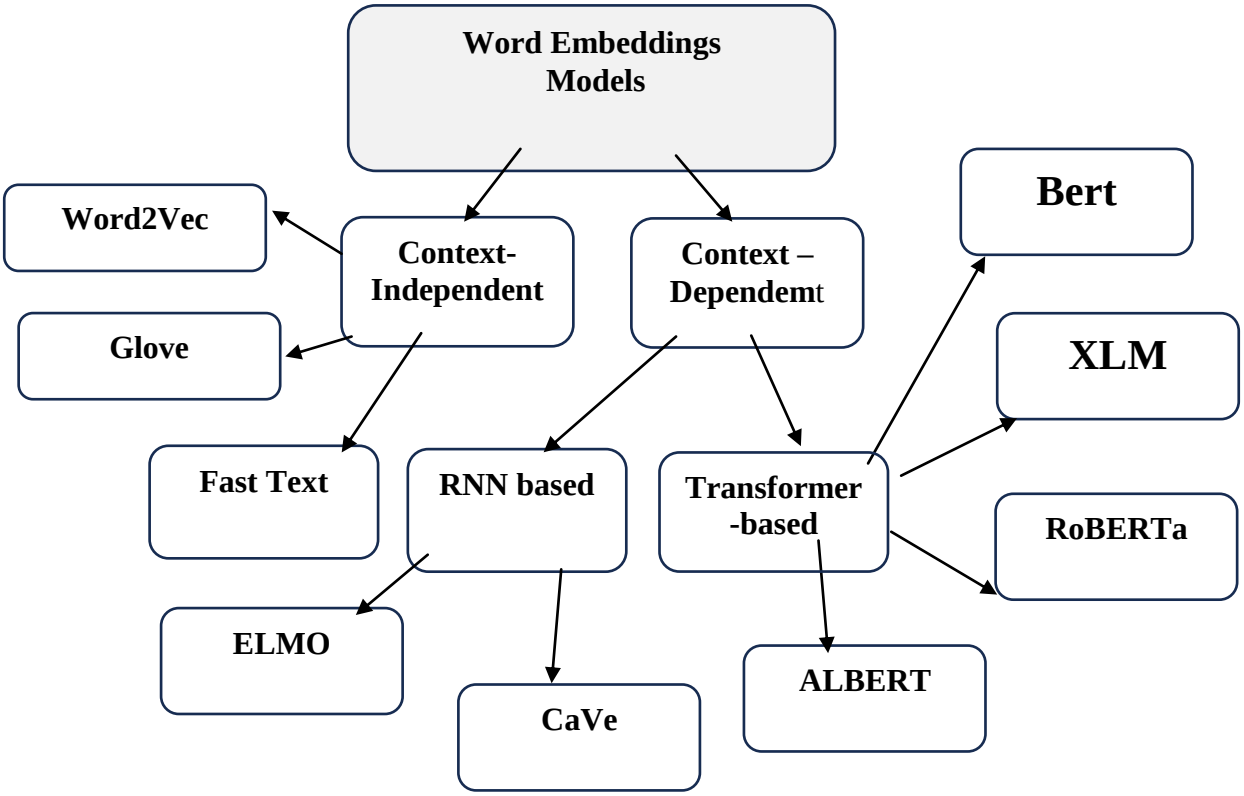
использования. Например, этот подход в зависимости от контекста выбирает наиболее подходящее значение слова «bank»: финансовое учреждение или берег реки. Основными свойствами семантического подхода являются то, что слова представляются в векторном пространстве (т.е. близкие по смыслу слова располагаются на малом расстоянии друг от друга в векторном пространстве), контекст играет важную роль, рассматривается близость между значениями. Этот подход имеет достаточно практические положительные свойства: различает многозначность по контексту, различает синонимичные и близкие по значению слова, очень продуктивен при переводе, образовании текста и опросе.

Модели Word embeddings являются представлением слов как математических объектов в высоко-масштабном пространстве, где главной идеей является семантика распределения. Основными моделями Word embeddings считаются Word2Vec, которая в свою очередь включает в себя CBOW (прогноз слова на основе контекста) и Skip-gram (прогноз контекста на основе слов), основывающаяся на глобальную статистику модель GloVe, а также FastText, которая рассматривает под-словесные единицы (морфемы). Преимуществами моделей Word embeddings являются то, что они создают возможность определить семантическую близость (cosine similarity) и аналогии между словами (king – man + woman = queen). Эти модели показывают высокие результаты в выявлении синонимичных отношений и тематической схожести.



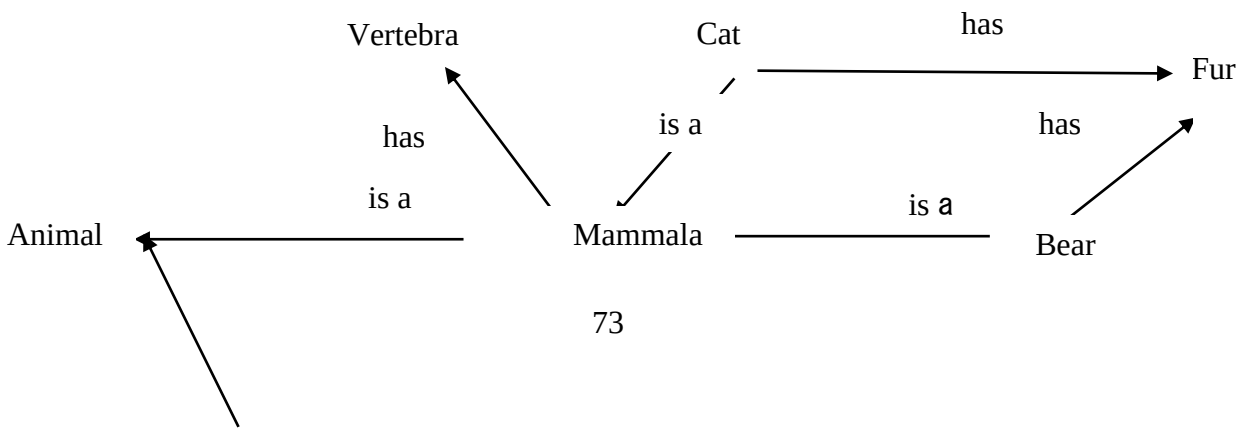
Контекстуальные модели, основанные на трансформер системы, дифференцируют значения слов в зависимости от конкретного его использования и создают возможность анализировать общую семантическую структуру текста. Особенно, контекстуальные модели представляют значения слов как динамические единицы, меняющиеся в зависимости от конкретного пользования. Это оценивается как синтез классических семантических теорий с современными технологиями ИИ. Основными моделями данного подхода являются ELMo (двунаправленный LSTM), BERT (двусторонний контекст, основанный на трансформер), а также GPT, который является последовательной (unidirectional) языковой моделью. В этих случаях вектор слова меняется в зависимости от предложения и решение полисемии дает точные результаты. Например, слово «bank» в

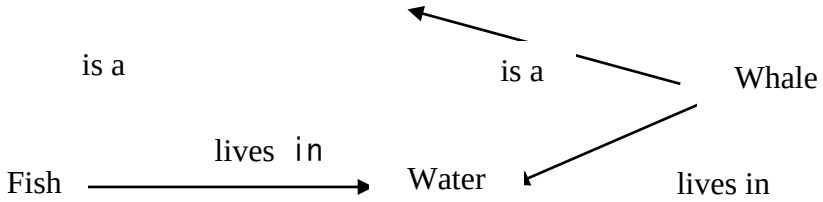
каждом из следующих предложений выражается различными векторами: «I sat by the river bank.», «I went to the bank to withdraw money».



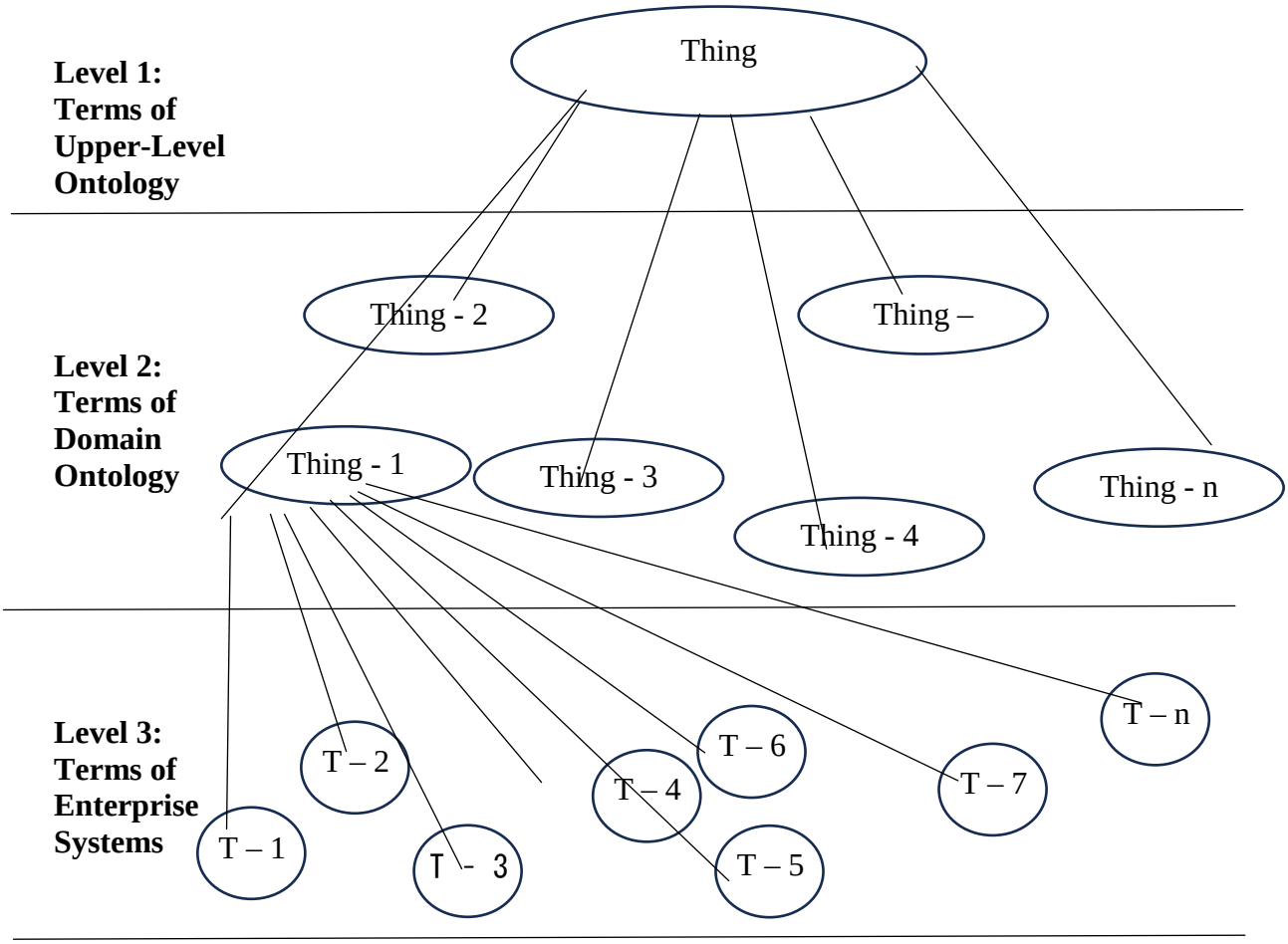
Transformer же является архитектурой всех современных языковых моделей, таких как BERT, GPT, T5 и другие. Основными компонентами данного подхода являются модель само-внимания (self-attention), вычисляющая воздействие слов в предложении друг на друга, модель многоголовочного внимания (multi-head attention), рассматривающая различные семантические отношения параллельно, блоки кодирования и декодирования (encoder/decoder) и, наконец, кодирование позиционной информации (positional encoding), которое сохраняет порядок слов. Преимущества архитектуры трансформера то, что она ловит отдаленные семантические отношения, проводит параллельное вычисление с высокой скоростью и качеством, рассматривает глубокую контекстуальную семантику.

Следующая модель семантического подхода – семантические сети (Semantic Networks) – представляет граф-структуры понятий в семантических сетях (узел + связь), т.е. в узле отражаются понятия и слова, а со стороны показываются их семантические отношения (is-a, part-of, cause-of). Сети набора синонимов (synsets) и гиперонимных /гипонимных (hiperonim/hiponim) отношений являются известными примерами данной модели. Несмотря на то, что они статические и создают сложности при масштабировании больших текстов, эти сети преимущественны потому, что близки к человеческой логике и отражают объяснимую семантику.





Последней моделью семантического подхода являются онтологии, которые описывают понятия в виде моделей формальных и иерархических знаний. Они имеют следующие составные части: классы (classes), образцы (instances), особенности (properties), утверждения (axioms). Эти онтологии широко пользуют в семантических сетях OWL и RDF, экспертных системах и в ИИ, основанных на знания. В отличие от Embedding, которая дает имплицитное статистическое знание, онтология выражает эксплицитное знание. Если сравнить все семантические модели, станет ясно, что самыми контекстуальными и динамическими методами считаются контекстуальные модели и модель трансформера.



Таким образом, лексическая модель больше подходит для анализа поверхностных структур языка, а семантическая модель – для более глубоких слоев значений. Лексический анализ эффективен для терминологических и статистических исследований, а семантический анализ – для исследования дискурса, значения и контекста. Поэтому интегративная модель является наиболее оптимальной для охвата комплексной природы языка. Такая разница между этими двумя моделями с лингвистической точки зрения может оцениваться как проявление лексико-семантических отношений в компьютерной сфере.

Современные системы ИИ, особенно системы, основанные на архитектуру трансформера, объединяя эти два подхода, на первом этапе делят текст на лексические сегменты, а потом распределяя эти сегменты по семантическим векторам, обрабатывают их на контекстуальном уровне. Такой интегративный подход дает возможность моделировать

одновременно и структуру, и смысловые слои языка, а также обеспечивает более разумный подход ИИ к языку. В результате интеграции этих двух подходов в современных системах ИИ сформировались более функциональные и гибкие модели языка. С точки зрения лингвистики такой подход к языку создает ряд возможностей: исследуется лексико-семантические отношения слов на математическом уровне, моделируются, основанные на значения, природные свойства языка на цифровом уровне, тестируются классические лингвистические теории.

Заключение: В итоге, развитие языковых моделей в ИИ отражает целый эволюционный путь перехода с лексического уровня на семантическую глубину. Если лексический подход сыграл базовую роль в образовании первичных языковых моделей, то семантический подход значительно расширил функциональные возможности этих моделей. Современные языковые модели, объединяя теоретические и практические потенциалы обоих подходов, создают условия для более точного и гибкого анализа языка. Они синтезируют оба подхода с целью получения наиболее близких человеческому языку результатов.

Литература

1. Abdullayev, A. (2010). Linqvistik Stilistika. Bakı: Elm.
2. Əliyev, A. (2020). Kompüter dilçiliyinə giriş. Bakı: Bakı Universiteti nəşri.
3. Quliyeva, S. (2022). Süni intellekt və dilçilik problemləri. Dil və Ədəbiyyat, 4, 45-52.
4. Баранов, А.Н. (2001). Введение в прикладную лингвистику. Москва: Эдиториал УРСС.
5. Гальперин, И.Р. (1981). Текст как объект лингвистического исследования. Москва: Наука.
6. Захаров, В.П. (2014). Компьютерная лингвистика. Санкт-Петербург: СпбГУ
7. Halliday, M.A.K., & Hasan, R. (2014). Cohesion in English. London: Routledge.
8. Jurafsky, D., & Martin, J.H. (2023). Speech and Language Processing. Pearson.
9. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. Proceedings of ACL, 1-12.
10. Goldberg, Y. (2017). Neutral Network Methods for Natural Language Processing. Morgan & Claypool.
11. Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global vectors for word representation. EMNLP, 1532-1543.
12. Widdowson, H.G. (2007). Discourse Analysis. Oxford: Oxford University Press.

Гюнай Багирова

Искусственный интеллект и лексико-семантический подход к языковым моделям

Аннотация

В статье рассматривается роль искусственного интеллекта в формировании языковых моделей с позиции лексико-семантического подхода. Современные системы искусственного интеллекта, в частности крупные языковые модели, основанные на нейронных сетях, функционируют на основе обработки семантической структуры лексических единиц и их контекстуальных связей. В исследовании анализируются такие понятия, как лексико-семантическое поле, семантическая сеть, концептуальная структура и многозначность, а также их реализация в алгоритмах искусственного интеллекта. Подчеркивается, что эффективность языковых моделей определяется не только статистическими параметрами, но и глубинной организацией семантических связей.

Ключевые слова: моделирование языка, токенизация, мешок слов, словесные векторы, контекстуальные модели.

Gunay Bagirova

Artificial intelligence and the lexical-semantic approach to language models

Abstract

The article examines the role of artificial intelligence in the development of language models from the perspective of the lexical-semantic approach. Modern AI systems, particularly large neural network-based language models, operate through the processing of semantic structures of lexical units and their contextual relations. The study analyzes key concepts such as lexical-semantic fields, semantic networks, conceptual structures, and polysemy, and explores how these are implemented within AI algorithms. It is argued that the effectiveness of language models depends not only on statistical frequency patterns but also on the deep organization of semantic relationships.

Key words: *language modelling, Tokenization, Bag of Words, Word Embeddings, contextual models, Semantic Networks.*

Çapa tövsiyə edən: ***Azərbaycan Dillər Universiteti***

Rəyçilər: ***professor M.Mahmudov***
dosent E.Əliyeva