УДК 378.147
2.3.1. Системный анализ, управление и обработка информации




А.М. Полянский, Т.Ф. Ковырзина
Вологодский государственный университет

ГИБРИДНЫЕ АЛГОРИТМЫ ОБРАБОТКИ ЕСТЕСТВЕННОГО ЯЗЫКА ДЛЯ ЗАДАЧ УЧЕБНО-МЕТОДИЧЕСКОГО ОБЕСПЕЧЕНИЯ: ОБЗОР И СРАВНИТЕЛЬНЫЙ АНАЛИЗ

В статье рассматривается проблема автоматизации процессов учебно-методического обеспечения в условиях цифровой трансформации высшего образования. Основное внимание уделяется анализу и сравнению методов обработки естественного языка (Natural Language Processing, NLP) для решения задач анализа рабочих программ дисциплин, контроля их соответствия образовательным стандартам, выявления дублирования и пробелов в содержании. Предлагается подход на основе гибридных алгоритмов, сочетающих статистические, семантические и экспертно-онтологические методы. В качестве примера применения подхода рассматривается задача семантического сопоставления компетенций и дисциплин. Приводятся практические рекомендации по внедрению гибридных моделей в образовательный процесс.

Обработка естественного языка, гибридные алгоритмы, учебно-методическое обеспечение, образовательные стандарты, семантический анализ, онтологии, машинное обучение.

Цифровая трансформация высшего образования актуализирует задачи автоматизации рутинных процессов учебно-методического обеспечения. К ним относятся анализ рабочих программ дисциплин (РПД), контроль их соответствия требованиям образовательных и профессиональных стандартов, выявление дублирования и пробелов в содержании. Традиционная экспертная оценка этих процессов трудоемка, субъективна и плохо масштабируется, особенно в условиях динамичного обновления IT-отрасли, задающей тон многим направлениям подготовки.

Методы обработки естественного языка (Natural Language Processing, NLP) представляют собой технологический ответ на эти вызовы. Однако наблюдается разрыв между развитием фундаментальных NLP-моделей и реальными условиями их внедрения в университетской среде, которая характеризуется ограниченными вычислительными ресурсами, отсутствием больших размеченных корпусов образовательных текстов, потребностью в интерпретируемости результатов и спецификой формальной, насыщенной нормативной терминологией русского языка.

Целью данной статьи является проведение критического обзора и сравнительного анализа спектра NLP-методов и обоснование выбора оптимального класса решений для задач учебно-методического обеспечения. Основной тезис состоит в том, что для данных задач наиболее эффективными являются не «чистые» state-of-the-art модели, а гибридные архитектуры, компенсирующие недостатки одних методов достоинствами других.

Классификация методов NLP
для анализа образовательных текстов

Для задач семантического сопоставления (например, «компетенция – дисциплина») применяется несколько семейств методов с принципиально разными возможностями.

1. Классические статистические методы (TF-IDF).

Метод TF-IDF (Term Frequency-Inverse Document Frequency) оценивает важность термина в документе относительно коллекции документов, преобразуя текст в числовой вектор [1]. Его преимущества для вузовского контекста – высокая интерпретируемость, вычислительная эффективность и работа на малых объемах данных. Главный недостаток – «семантическая слепота»: метод не учитывает синонимию и контекстную близость слов.

2. Методы распределенных представлений (Word2Vec, FastText).

Данные методы обучают векторные представления слов (эмбеддинги), размещая семантически близкие слова рядом в векторном пространстве [2]. Архитектура FastText, работающая на уровне субтокенов (символьных n-грамм), особенно эффективна для обработки морфологии русского языка [3]. Основное преимущество – способность улавливать семантические связи, недостаток – слабая интерпретируемость векторного сходства.

3. Контекстные нейросетевые модели (BERT и его адаптации).

Модели типа BERT генерируют контекстуализированные эмбеддинги, учитывающие значение слова в зависимости от всего предложения [4]. Для русского языка существуют специализированные адаптации, такие как RuBERT [5]. Хотя они демонстрируют высокую точность, их применение в условиях вуза ограничено высокой ресурсоемкостью, требовательностью к данным и сложностью интерпретации («черный ящик»).

Таблица

Сравнительный анализ методов NLP

Критерий / Метод	TF-IDF	FastText	BERT	Экспертная онтология
Потенциальная точность	Низкая	Средняя	Высокая	Низкая
Интерпретируемость	Высокая	Средняя	Низкая	Высокая
Ресурсоемкость	Низкая	Средняя	Высокая	Низкая
Требования к данным	Низкие	Средние	Высокие	Зависит от эксперта

4. Экспертно-онтологические компоненты.

Этот подход предполагает создание тезаурусов и онтологий, явно моделирующих связи между понятиями образовательных стандартов. Он не является самостоятельным алгоритмом машинного обучения, но выступает обязательным обогащающим слоем, обеспечивающим учет нормативной специфики предметной области.

Сравнительный анализ методов NLP

Сравнительный анализ методов по ключевым для вуза критериям (табл.) показывает, что ни один «чистый» подход не является удовлетворительным по всем параметрам одновременно. BERT точен, но ресурсоемок и непрозрачен; TF-IDF прозрачен и быстр, но семантически ограничен; FastText – компромиссный вариант, требующий интеграции с предметной областью.

Методология гибридного подхода

Выходом из этой дилеммы является переход к гибридным (ансамблевым) архитектурам. Принцип гибридизации заключается во взвешенной композиции нескольких разнородных методов. Исследования в области образовательных технологий подтверждают эффективность таких подходов при работе с гетерогенными данными [6].

На основе анализа предлагается обобщенная трехканальная архитектура:

Статистический канал (TF-IDF): обеспечивает интерпретируемость и работу с ключевыми терминами.

Семантический канал (FastText или адаптированный RuBERT): отвечает за улавливание смысловой близости и синонимии.

Экспертно-онтологический канал: гарантирует учет специфики образовательных стандартов.

Выходные данные каналов объединяются в модуле агрегации по формуле взвешенной суммы, что позволяет гибко балансировать между точностью, прозрачностью и предметной релевантностью.

Пример применения гибридного алгоритма для соотнесения компетенций и дисциплин

Рассмотрим задачу соотнесения компетенций и содержания дисциплины «Компьютерная графика» для направления подготовки 09.03.04 Программная инженерия.

1. Подготовка данных.

Тексты компетенций и описания дисциплин приводятся к единой форме (лемматизация, удаление стоп-слов).

Создается экспертный словарь терминов образовательного стандарта.

2. Применение гибридной модели.

TF-IDF-канал: выделяет ключевые термины, такие как «программное обеспечение», «алгоритмы визуализации».

FastText-канал: оценивает семантическую близость слов «графика» и «визуализация», «программа» и «инструмент».

Онтологический канал: проверяет соответствие терминов нормам ФГОС и профессионального стандарта.

3. Результаты агрегации.

Система выдает степень соответствия компетенций и тем дисциплины с указанием весов каждого канала, что позволяет методистам принимать обоснованные решения.

Практические рекомендации для внедрения в вузе

Для первого этапа прототипирования рекомендуется начинать с канала на основе TF-IDF, дополненного простым экспертный словарем.

Для рабочего инструмента методической поддержки целесообразно реализовать трехканальную гибридную архитектуру. В качестве семантического ядра, в зависимости от ресурсов, можно выбрать предобученную модель FastText или облегченную версию трансформера. Ключевое значение имеет итеративная разработка экспертного компонента при активном участии методистов.

Технологический стек может включать: для лингвистической предобработки – pymorphy3 (лемматизация) или более комплексные инструменты, такие как stanza [7]; для машинного обучения – scikit-learn (TF-IDF), gensim/fasttext (эмбеддинги), transformers (для работы с моделями типа RuBERT).

Проведенный анализ подтверждает, что для задач учебно-методического обеспечения, где интерпретируемость и учет предметной специфики не менее важны, чем точность, оптимальным решением являются гибридные алгоритмы, комбинирующие статистические, семантические и экспертно-онтологические подходы.

Перспективными направлениями дальнейших исследований являются: автоматическая оптимизация весов в гибридной модели, интеграция облегченных трансформеров для избирательного повышения точности, а также расширение анализа на мультимодальный образовательный контент.

Литература

1. Manning, C. D. Introduction to Information Retrieval / C. D. Manning, P. Raghavan, H. Schütze. – Cambridge : Cambridge University Press, 2008. – 482 p.

2. Mikolov, T. Efficient Estimation of Word Representations in Vector Space / T. Mikolov, K. Chen,

G. Corrado, J. Dean // Proceedings of the International Conference on Learning Representations (ICLR 2013). – 2013.

3. Bojanowski, P. Enriching Word Vectors with Subword Information / P. Bojanowski, E. Grave, A. Joulin, T. Mikolov // Transactions of the Association for Computational Linguistics. – 2017. – Vol. 5. – P. 135–146.

4. Devlin, J. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding / J. Devlin, M.-W. Chang, K. Lee, K. Toutanova // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. – 2019. – Vol. 1. – P. 4171–4186.

5. Kuratov, Y. Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language / Y. Kuratov, M. Arkhipov // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2019". – 2019.

6. Chen, X. Twenty Years of Personalized Language Learning: Topic Modeling and Knowledge Mapping / X. Chen, D. Zou, H. Xie, G. Cheng // Educational Technology & Society. – 2021. – Vol. 24, № 1. – P. 205–222.

7. Qi, P. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages / P. Qi, Y. Zhang, Y. Zhang, J. Bolton, C. D. Manning // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. – 2020. – P. 101–108.

8. Полянский, А. М. Методика соотнесения содержания образовательной компетенции и учебной дисциплины / А. М. Полянский, Т. Ф. Ковырзина // Вестник Вологодского государственного университета. Серия: Технические науки. – 2025. – № 2(28). – С. 40–41.

A.M. Polyansky, T.F. Kovyrzina
Vologda State University

HYBRID NATURAL LANGUAGE PROCESSING ALGORITHMS FOR EDUCATIONAL AND METHODOLOGICAL SUPPORT TASKS: REVIEW AND COMPARATIVE ANALYSIS

The article addresses the problem of automating educational and methodological support processes in the context of digital transformation of higher education. The main focus is on the analysis and comparison of natural language processing (NLP) methods for solving problems of analyzing discipline work programs, monitoring their compliance with educational standards, identifying duplication and gaps in content. An approach based on hybrid algorithms combining statistical, semantic and expert-ontological methods is proposed. As an example of the application of the approach, the task of semantic matching of competencies and disciplines is considered. Practical recommendations for the implementation of hybrid models in the educational process are provided.

Natural language processing, hybrid algorithms, educational and methodological support, educational standards, semantic analysis, ontologies, machine learning.