

ИССЛЕДОВАНИЕ АКТУАЛЬНОСТИ ОБРАЗОВАТЕЛЬНЫХ КОМПЕТЕНЦИЙ НА РЫНКЕ ТРУДА С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ ОБРАБОТКИ ЕСТЕСТВЕННОГО ЯЗЫКА

*С. Д. Белов*¹, *А. В. Ильина*^{1,2,*}, *В. В. Кореньков*^{1,2},
*В. А. Тарабрин*¹, *П. В. Зрелов*^{1,2}

¹ Объединенный институт ядерных исследований, Дубна

² Государственный университет «Дубна», Дубна, Россия

Представлены результаты развития методов сопоставления программ высшего образования и потребностей рынка труда. Данные методы разработаны авторами в рамках реализации аналитической платформы автоматизированного мониторинга рынка труда и интеллектуального анализа кадровых потребностей по номенклатуре специальностей вуза. Для анализа использовались тексты названий, уровней освоения и индикаторов достижения компетенций ИТ-профиля, а также векторные языковые модели с различными архитектурами и обучающими корпусами текстов. Приведено сравнение результатов, и показана устойчивость методов сопоставления образовательных компетенций и потребностей рынка труда.

The results of the development of methods for comparing higher education programmes and the needs of the labor market are presented. These methods were developed by the authors as part of the implementation of an analytical platform for automated monitoring of the labor market and intelligent analysis of personnel against university specialties. Texts of titles, levels of development and indicators of IT-profile competence achievement, as well as vector language models with various architectures and training text bodies were used for the analysis. The results are mutually compared and the stability of the methods of matching educational competencies and the needs of the labor market is shown.

PACS: 89.20.Ff; 07.05.Tp

ВВЕДЕНИЕ

В работах [1–3] представлена разработанная авторами аналитическая платформа, реализующая автоматизированный мониторинг и анализ рынка труда в Российской Федерации, а также мониторинг соответствия кадровых потребностей работодателей уровню подготовки специалистов.

* E-mail: annailina@jinr.ru

Информационной базой служат открытые источники — базы данных вакансий «HeadHunter», «Работа России» и «SuperJob», официальный перечень профессий Министерства труда и социальной защиты Российской Федерации, стандарты высшего образования.

В отличие от традиционных подходов сопоставления и анализа содержания образовательных программ, основанных на онтологических моделях и экспертных системах, в работах [1–3] анализ соответствия образовательных стандартов потребностям рынка труда впервые произведен на основе алгоритма перевода коротких предложений в векторное пространство с использованием нейросетевой модели Word2Vec [4].

В настоящей работе представлены результаты, связанные с развитием предложенных в [1–3] методов. Исходными данными со стороны системы высшего образования являются описания образовательных профессиональных компетенций. Однако в отличие от прежнего подхода используются не только тексты названий, но и тексты индикаторов достижения профессиональных компетенций, что потребовало определенных изменений в методике. В рамках нового подхода проведено сравнение ряда векторных нейросетевых моделей естественного языка, имеющих различную архитектуру и обученных на различных текстовых корпусах.

ОПИСАНИЕ ИСХОДНЫХ ДАННЫХ

Для целей исследования были взяты 14 профессиональных компетенций разных направлений и профилей из области информационных технологий уровня бакалавриата государственного университета «Дубна». Каждая компетенция представлена: 1) кодом и наименованием профессиональной компетенции, 2) кодом и наименованием индикатора достижения профессиональной компетенции, 3) основанием разработки профессиональной компетенции.

Подход к выбору исходных текстов данных со стороны рынка труда остался тем же, что и в работе [3]. Исследуемые данные были представлены тестовой выборкой объемом около 100 000 объявлений о вакансиях в сфере «Информационные технологии, интернет, телеком», размещенных на платформе HeadHunter в период с января по февраль 2022 г. Для анализа использовались тексты раздела «Обязанности».

НЕЙРОСЕТЕВЫЕ ЯЗЫКОВЫЕ МОДЕЛИ

В исследовании использовались три нейросетевые языковые модели:

- модель семейства Word2Vec [4], предварительно обученная на Национальном корпусе русского языка (НКРЯ) за 2017 г. и размещенная на сервисе RusVectōrēs под названием «ruscorpora_1_300_10» в 2017 г.*

* Ко времени публикации данной работы описание модели доступно в веб-архиве страницы «Модели» сервиса RusVectōrēs за 2017 г. [5].

- модель семейства FastText [6], предварительно обученная на русскоязычном веб-корпусе за 2018 г.*

- модель семейства BERT [8] под названием «sbert_large_mt_nlu_ru», применяемая в решении широкого спектра задач обработки и понимания естественного языка, среди которых: распознавание намерений, выделение именованных сущностей, сентимент-анализ, поиск схожих запросов. Модель BERT имеет архитектуру трансформера и состоит из двух основных компонентов — кодировщика и декодировщика. Кодировщик представляет собой последовательность слоев трансформеров, которые принимают входной текст и генерируют его векторное представление. Декодер используется для генерации выходного текста на основе этого векторного представления.

Так же, как и в работе [3], определялась семантическая близость текстов вакансий, представляемых полем «обязанности», и текстов наименований (названий) компетенций образовательных стандартов. В дополнение в качестве альтернативного подхода вместо названий компетенций использовались их индикаторы. В качестве метрики сходства применялось косинусное расстояние между векторами в семантическом пространстве — чем оно больше, тем выше степень семантической близости текстов, и, следовательно, более актуальными представляются данные образовательных компетенций на рынке труда.

ПОДГОТОВКА ИСХОДНЫХ ДАННЫХ

Необходимость реализации различных алгоритмов предварительной обработки входных текстовых данных обусловлена использованием моделей различных архитектур, обученных на разных текстовых корпусах.

Для модели семейства Word2Vec исходные текстовые данные были предварительно лемматизированы и размечены в соответствии с частями речи по стандарту UPOS**.

Для модели семейства FastText исходные данные лемматизировались, но, в отличие от Word2Vec, уже без разметки в соответствии с частями речи.

Для использования модели BERT не требовалось никакой предобработки.

ВИЗУАЛИЗАЦИЯ РЕЗУЛЬТАТОВ

Одним из возможных подходов к визуализации результатов является создание взвешенного графа, отражающего связи между профессиональными компетенциями и потребностями рынка труда [1]. В таком графе образовательные компетенции и требования рынка труда представляются

* Модель доступна на сервисе RusVectōrēs под названием «aganeum_pone_fasttextskipgram_300_5_2018» [7].

** Universal POS tags (<https://universaldependencies.org/u/pos/>).

вершинами разных типов. Вершины, относящиеся к образовательным компетенциям, маркируются индексом компетенции. Ребра графа связывают только вершины разных типов и показывают наличие и степень связи между ними, определяемой косинусным расстоянием. Важным преимуществом этого подхода является возможность фильтрации связей между вершинами с помощью изменения порогового значения. Порог позволяет отбирать только такие ребра, вес которых превышает его значение.

Выбор (подбор) порогового значения важен, если надо отобразить только определенную часть ребер графа, например, с целью сравнения результатов разных методов в одинаковых условиях. Так, если мы хотим сравнить результаты, зная, что конкретный вуз готовит специалистов только для определенной части рынка труда, «претендую», например, лишь на 80% вакансий, то и все оценки следует проводить с учетом этого фактора.

АНАЛИЗ РЕЗУЛЬТАТОВ

На рис.1 представлены графы соответствия профессиональных компетенций программ IT-направлений государственного университета «Дубна» вакансиям рынка труда в сфере «Информационные технологии,

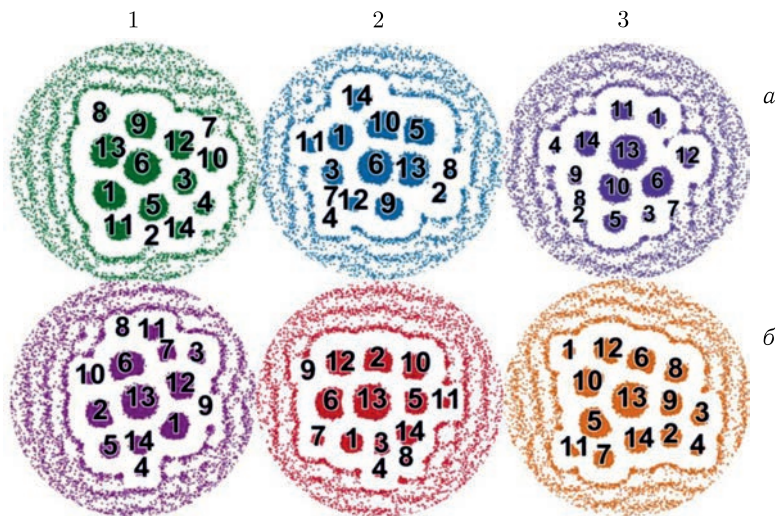


Рис. 1. Сравнение графов «Профессиональная компетенция — вакансия», отображающих результаты исследований с использованием различных уровней детализации формулировок профессиональных компетенций — названий (а) и индикаторов достижения (б), а также различных моделей — Word2Vec (1), FastText (2), BERT (3) для набора профессиональных компетенций программ IT-направлений государственного университета «Дубна»

интернет, телеком» (см. раздел «Описание исходных данных») для двух типов исходных данных и трех языковых моделей.

При сравнении работы алгоритмов пороговые значения для каждой модели выбирались таким образом, чтобы доля отобранных ребер была одинаковой и составляла 80%. Так, для модели Word2Vec порог был установлен равным 0,8 (как в случае использования только названий компетенций, так и в случае использования индикаторов достижения). Для модели FastText порог равнялся 0,77 (в случае названия компетенций) и 0,78 (в случае индикаторов достижения). Для модели BERT — 0,72 и 0,74 соответственно.

Хорошо видно, что все 14 компетенций, независимо от типа их представления и используемой модели, отражены в центре графа, в области наиболее сильных связей.

На рис. 2 представлены результаты анализа графов, которые позволяют сделать ряд выводов о закономерностях, наблюдаемых во взаимосвязи системы высшего образования и рынка труда.

Независимо от типа исходных данных и выбранной модели наиболее частое отражение на рынке труда имеют компетенции с идентификаторами 6 (наименование компетенции «Способен выполнять проектную деятельность по разработке и созданию (модификации) ИС, разработке технико-экономического обоснования проектов для улучшения бизнес-процессов и ИТ-инфраструктуры предприятия») и 13 (наименование

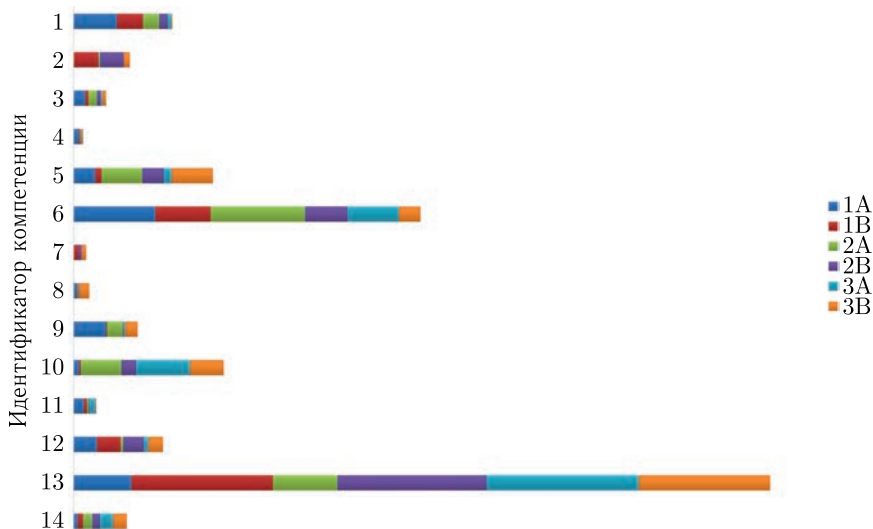


Рис. 2. Столбчатая диаграмма с накоплением, отображающая количество связей «Профессиональная компетенция — вакансия» для каждой профессиональной компетенции в процентах от общего количества связей

компетенции «Способен проектировать и создавать программное обеспечение, соответствующее требованиям заказчика, включая разработку программного интерфейса, проектирование структур данных, модулей и компонентов программного обеспечения, выполнять процедуры их сборки и проводить оценку работоспособности программного продукта»). Наименьшее отражение имеют компетенции под идентификаторами 4 (наименование компетенции «Способен проектировать и разрабатывать компоненты корпоративных информационных систем и информационных систем электронного бизнеса») и 7 (наименование компетенции «Способен разрабатывать требования и проектировать программное обеспечение»).

Наибольшая востребованность компетенций наблюдается для вакансий со следующими специализациями: веб-программирование для бизнеса, техническая поддержка, управление IT-проектами. Наибольшее количество связей относится к вакансиям с требованием опыта работы от 1 года до 3 лет. Около 10% связей относится к вакансиям без опыта работы. Наиболее часто наблюдаются связи для вакансий с зарплатами в интервале от 100 000 руб. (8%) до 150 000 руб. (8%). Регионы востребованности компетенций: г. Москва (в среднем около 45%) и г. Санкт-Петербург (в среднем — 14%). Исследуемые компетенции оказались менее востребованы в других регионах (менее 3%).

ЗАКЛЮЧЕНИЕ

Разработанный в [1–3] с использованием семантической модели Word2Vec подход к определению востребованности на рынке труда профессиональных компетенций высшего образования показал близкие результаты и при использовании других нейросетевых языковых моделей (FastText и BERT), а также относительную устойчивость к изменению исходных данных в рамках компетентностной образовательной модели. Это свойство указывает на возможность практического применения разработанного подхода для исследования востребованности образовательных компетенций на рынке труда. Сравнение проводилось на примере данных образовательных программ IT-направлений государственного университета «Дубна». В результате проведенной работы представлен ряд выводов о закономерностях, наблюдаемых во взаимосвязи системы высшего образования и рынка труда.

СПИСОК ЛИТЕРАТУРЫ

1. Зрелов П. В., Кореньков В. В., Кутовский Н. А., Петросян А. Ш., Румянцев Б. Д., Семенов Р. Н., Филозова И. А. Мониторинг потребностей рынка труда в выпускниках вузов на основе аналитики с интенсивным использованием данных // Тр. XVIII Междунар. конф. «Аналитика и управление данными в областях с интенсивным использованием данных» (DAMDID/RCDL'2016),

- Ершово, Россия, 11–14 окт. 2016 г. М.: Федер. исслед. центр «Информатика и управление» РАН, 2016. С. 124–131.
2. *Валентей С. Д., Зрелов П. В., Кореньков В. В., Белов С. Д., Кадочников И. С.* Мониторинг соответствия профессионального образования потребностям рынка труда // *Общест. науки и современность*. 2018. № 3. С. 5–16.
 3. *Belov S. et al.* Methods and Algorithms of the Analytical Platform for Analyzing the Labor Market and the Compliance of the Higher Education System with Market Needs // *Proc. Sci. 2022. Conf. DLCP2022*. P.028; doi: <https://doi.org/10.22323/1.429.0028>.
 4. *Mikolov T. et al.* Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781. 2013.
 5. *Кутузов А., Кузьменко Е.* RusVectōrēs: модели. web.archive.org. May 19, 2017. <https://web.archive.org/web/20170519222814/https://rusvectors.org/ru/models/> (дата обращения: 09.09.2023).
 6. *Joulin A., Grave E., Bojanowski P., Mikolov T.* Bag of Tricks for Efficient Text Classification. arXiv:1607.01759. 2016.
 7. *Kutuzov A., Kuzmenko E.* RusVectōrēs: Models, RusVectors, 2023. Available at: <https://rusvectors.org/ru/models/> (accessed 09.09.2023).
 8. *Devlin J., Chang M.-W., Lee Kenton, Toutanova K.* BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805. 2018.