Научная статья
УДК 004.855
DOI: 10.14529/ctcr220302

СРАВНЕНИЕ НЕЙРОСЕТЕВЫХ МОДЕЛЕЙ НА АРХИТЕКТУРЕ ТРАНСФОРМЕРОВ В КОНТЕКСТЕ ЗАДАЧИ ОЦЕНКИ КОМПАКТНОСТИ ВЕКТОРНЫХ ПРЕДСТАВЛЕНИЙ СЕМАНТИЧЕСКИ БЛИЗКИХ ТЕКСТОВ ТРЕБОВАНИЙ ЕВРОПЕЙСКОЙ КЛАССИФИКАЦИИ НАВЫКОВ ESCO

И.Е. Николаев[1], *ivan_nikolaev@csu.ru, https://orcid.org/0000-0002-9686-2435*
А.В. Мельников[2], *MelnikovAV@uriit.ru*

[1] *Челябинский государственный университет, Челябинск, Россия*
[2] *Югорский научно-исследовательский институт информационных технологий, Ханты-Мансийск, Россия*

Аннотация. В процессе анализа коротких текстов требований вакансий российского рынка труда было выявлено, что одни и те же навыки могут иметь различные формулировки на естественном языке. В связи с этим актуальной задачей становится поиск нейросетевой модели, способной эффективно выделять семантически близкие группы текстов требований для дальнейшего формирования профилей навыков и профессий российского рынка труда. **Целью исследования** является разработка метода оценки нейросетевых моделей, построенных на архитектуре трансформеров, посредством сравнения компактности векторных представлений семантически близких коротких текстов навыков профессий из европейской классификации (European Skills, Competences, and Occupations). **Материалы и методы.** В статье приводится анализ для оригинальной модели европейской таксономии навыков ESCO на английском языке и текстов навыков, переведенных на русский язык сервисами автоматического перевода Yandex Переводчик и Google Translate. В статье также приводится сравнение различных методов получения вложений предложений (cls, mean, pooling, SentenceTransformers) для различных нейросетевых моделей, построенных на архитектуре трансформеров. **Результаты** исследования показывают, что с помощью предложенного метода можно эффективно осуществлять выбор нейросетевых моделей для задачи поиска групп семантически близких текстов требований из текстов онлайн-вакансий. **Заключение.** Предложенный метод позволил эффективно выбирать нейросетевые модели для задачи выделения компактных групп семантически близких текстов профессиональных навыков, что, в свою очередь, даст возможность выделять группы навыков при формировании профилей профессиональных навыков, включая семантически близкие формулировки, и профилей целых профессий. Такие инструменты позволят оперативно определять: ключевые изменения потребностей рынка труда на уровне отдельных компетенций позволят сформировать представление о динамике и наборах актуальных компетенций, повысят эффективность управленческих решений по созданию программ цифровой грамотности, переподготовки и повышения квалификации, позволят осуществлять оценку компетенций, помогут всем участникам рынка труда точнее оценивать существующие тенденции, предложение и спрос на рынке труда.

Ключевые слова: нейронные сети, кластерный анализ, профессиональные навыки, трансформеры, sentence transformer, ESCO, рынок труда

Для цитирования: Николаев И.Е., Мельников А.В. Сравнение нейросетевых моделей на архитектуре трансформеров в контексте задачи оценки компактности векторных представлений семантически близких текстов требований европейской классификации навыков ESCO // Вестник ЮУрГУ. Серия «Компьютерные технологии, управление, радиоэлектроника». 2022. Т. 22, № 3. С. 19–29. DOI: 10.14529/ctcr220302

Original article
DOI: 10.14529/ctcr220302

COMPARISON OF TRANSFORMER ARCHITECTURE NEURAL NETWORK MODELS BASED ON EVALUATING THE VECTOR REPRESENTATION COMPACTNESS OF SEMANTICALLY SIMILAR TEXTS IN THE EUROPEAN CLASSIFICATION SKILLS ESCO

I.E. Nikolaev[1], *ivan_nikolaev@csu.ru, https://orcid.org/0000-0002-9686-2435*
A.V. Melnikov[2], *MelnikovAV@uriit.ru*

[1] *Chelyabinsk State University, Chelyabinsk, Russia*
[2] *Ugra Research Institute of Information Technologies, Khanty-Mansiysk, Russia*

Abstract. In the process of analyzing short texts of the requirements of the Russian labor market, it was revealed that the same skills may have different formulations in natural language. In this regard, the search for a neural network model capable of effectively identifying semantically similar groups of texts of requirements for further formation of profiles of the skills and professions of the Russian labor market becomes an urgent task. **The purpose of the study** is to develop a method for evaluating neural network models built on the architecture of transformers by comparing the compactness of vector representations of semantically close short texts of skills of professions from the European classification (European Skills, Competencies, and Occupations). **Materials and methods.** The article provides an analysis for the original model of the European taxonomy of ESCO skills in English, and the texts of skills translated into Russian by the Yandex Translator and Google Translate automatic translation services. The article also provides a comparison of various methods for obtaining sentence attachments (cls, mean, pooling, Sentence Transformers) for various neural network models built on the transformer architecture. The results of the study show that with the help of the proposed method, it is possible to effectively implement the choice of neural network models for the task of searching for groups of semantically similar texts of requirements from online job texts. **Conclusion.** The proposed method made it possible to effectively select neural network models for the task of identifying compact groups of semantically similar texts of professional skills, which in turn will make it possible to identify groups of skills when forming profiles of professional skills, including semantically similar formulations, and profiles of entire professions. Such tools will allow you to quickly identify: key changes in the needs of the labor market at the level of individual competencies, will allow you to form an idea of the dynamics and sets of relevant competencies, will increase the effectiveness of management decisions to create digital literacy programs, retraining and advanced training, will allow you to assess competencies, will help all participants in the labor market to more accurately assess the existing trends, supply and demand in the labor market.

Keywords: neural networks, cluster analysis, professional skills, transformers, sentence transformer, ESCO, labor market, Silhouette

For citation: Nikolaev I.E., Melnikov A.V. Comparison of transformer architecture neural network models based on evaluating the vector representation compactness of semantically similar texts in the European classification skills ESCO. *Bulletin of the South Ural State University. Ser. Computer Technologies, Automatic Control, Radio Electronics*. 2022;22(3):19–29. (In Russ.) DOI: 10.14529/ctcr220302

Введение

Современный рынок труда стремительно меняется. В первую очередь изменения обусловлены полномасштабным процессом цифровизации, а пандемия COVID-19 стала своего рода катализатором происходящих изменений. Изменение форм занятости, требований к соискателям, переход многих сфер деятельности в онлайн, изменение рабочих графиков, должностных инструкций и обязанностей, замена живого общения гаджетами и технологиями – все это предвестники глобальных, структурных изменений на рынке труда.

В этой связи особую актуальность и значимость приобретают исследования, направленные на создание инструментов оперативного мониторинга рынка труда на уровне отдельных компетенций, в условиях постоянных изменений требований работодателей, на основе анализа открытых данных систем онлайн-рекрутмента, в первую очередь текстовой информации из реальных онлайн-вакансий.

Такие инструменты позволят оперативно определять: ключевые изменения потребностей рынка труда на уровне отдельных компетенций, позволят сформировать представление о динамике и наборах актуальных компетенций, повысят эффективность управленческих решений по созданию программ цифровой грамотности, переподготовки и повышения квалификации, позволят осуществлять оценку компетенций, помогут всем участникам рынка труда точнее оценивать существующие тенденции, предложение и спрос на рынке труда.

Необходимость разработки таких инструментов также подтверждается на уровне федеральных проектов Российской Федерации: «Кадры для цифровой экономики» обозначена в национальной программе «Цифровая экономика Российской Федерации». В рамках рассматриваемого федерального проекта предполагается разработка и использование инструментов глобального мониторинга профилей компетенций человека в условиях цифрового развития, прогнозирование востребованности знаний и навыков специалистов.

Долгое время большинство исследований изменений рынка труда было связано со статистическим анализом различных экономических показателей за определенный период в той или иной отрасли: количество вакансий, заработная плата, данные по безработице. В рамках данного подхода ученые оперируют агрегированными «сухими», статистическими цифрами, пытаясь выявить изменения на стратегическом уровне. Это обстоятельство часто становится непреодолимым препятствием на пути оперативного выявления глубинных изменений в структуре и динамике изменений требований рынка труда на уровне анализа всех имеющихся данных онлайн-вакансий.

В последние годы резко вырос интерес к использованию методов искусственного интеллекта (ИИ) для анализа рынка труда. Данное направление даже получило отдельный термин – «разведка рынка труда» (LMI – labor market intelligence). Хотя единого определения LMI не существует, его можно рассматривать как разработку и реализацию алгоритмов и структур ИИ для анализа данных рынка труда в качестве поддержки для планирования политики и принятия решений [1–3].

Благодаря развитию инструментов для анализа текстов на естественном языке отдельными группами ученых делаются попытки анализировать изменения рынка труда по текстам вакансий из открытых источников [4–9]. Такой подход имеет ряд преимуществ, так как позволяет выявлять изменения на уровне отдельных профессий/специальностей, отдельных требований работодателей, например: позволяет осуществлять мониторинг онлайн-вакансий в отдельных регионах и странах в режиме реального времени; производить прогнозирование востребованности, мониторинг изменений отдельных навыков, компетенций и технологий в рамках отдельной профессии или целой отрасли, проводить оперативное сравнение аналогичных рынков труда в разных странах и регионах.

В статьях [7, 8] предложен набор инструментов для анализа РТ с применением методов машинного обучения (embeddings и кластеризация) к онлайн-вакансиям на итальянском РТ. Подход позволяет рассчитать для каждой профессии разные типы требуемых навыков. Авторами предложена методика определения профессиональных (hard) (отдельно анализируется группа цифровых навыков) и общепрофессиональных (soft) навыков, а также мера их востребованности на рынке. Оценено влияние автоматизации на социальные и цифровые навыки в профессии. Предложена методика и меры изменения профессиональной терминологии во времени, используемой при описании профессий, а также представлен инструмент для обнаружения новых навыков и новых профессий через анализ онлайн-вакансий.

В статье [9] предлагается подход на основе построения индекса цифровизации по роду занятий, используя данные из онлайн-вакансий. Этот индекс позволяет анализировать уровни и изменения спроса на цифровые навыки в Германии в период 2014–2018 гг. Показано, что доля, требующая хотя бы одного цифрового навыка, выросла с 38,1 % в 2014 году до 47,5 % в 2018 году. Показано, что высококвалифицированные должности требуют большего владения цифровыми навыками, чем низкоквалифицированные (94 % против 62 %). Показан также различный уровень проникновения цифровизации по профессиям и отраслям. Помимо индустрии информации и коммуникаций цифровизация получила распространение в сфере финансовых услуг и страхования, а также среди людей, предоставляющих профессиональные, академические и технические услуги. И наоборот, в индустрии туризма и здравоохранения требуется относительно мало цифровых навыков, как и в сфере социальных услуг.

В процессе анализа коротких текстов требований онлайн-вакансий российского рынка труда было выявлено, что одни и те же навыки могут иметь различные формулировки на естественном языке. В этой связи ключевым этапом анализа текстов требований становится этап выделения компактных групп семантически близких навыков с условием, что семантически близкие тексты должны принадлежать одной компактной группе, а семантически разные компактные группы текстов должны быть максимально удалены друг от друга.

Для решения этой задачи в работе предлагается использовать инструмент эмбеддингов для цифрового представления текстовый информации. Эмбеддинги позволяют перейти от текстовой информации к числовым векторам, способным сохранять необходимые семантические свойства естественного языка.

Первые модели получения эмбеддингов, такие как BOW [10], TFIDF [11], Word2Vec [12], «понимали» (улавливали) смысл текста только на уровне отдельных слов, без учета контекста.

Сегодня одним из наиболее перспективных и популярных подходов к анализу естественного языка и пониманию смысла текста являются нейросетевые модели, использующие механизм внимания – способность поиска взаимосвязей между различными частями текста [13], и построенные на архитектуре так называемых трансформеров. Первой такой моделью стала в 2018 году модель BERT [14], представленная компанией Google. Появление BERT произвело настоящую революцию в компьютерной лингвистике. Особенность BERT заключается в том, что он способен генерировать векторное представление, учитывает контекст для всех слов и способен лучше справляется с долгосрочными зависимостями в тексте, иными словами, дольше удерживает информацию о контексте для каждого слова [15]. В настоящее время BERT (и его производные) показывает state-of-the-art на большинстве NLP (natural language processing) и NLU (natural language understanding) задач и превосходит нейросетевые модели предыдущего поколения, такие как word2vec, LSTM и др.

В настоящее время подход на основе построения моделей естественного языка, на основе обучения глубоких нейронных сетей является наиболее эффективным. Появляются новые подходы и модели: BERT, RoBERTa [16], GPT [17], T5 [18], XLM [19] и их модификации. Эти модели универсальны и способны извлекать из текста признаки, полезные для решения множества задач текстового анализа.

Последние достижения в компьютерной лингвистике позволили перейти к эффективным векторным представлениям для целых предложений и абзацев текста. В работе [20] описан проект SentenceTransformer (SBERT), который представляет собой технологию модификации предварительно обученной сети BERT. В работе используются сиамские и триплетные сетевые структуры для получения семантически значимых векторов предложений. Это позволяет дообучать модель на задаче определения семантически близких текстов. Модели дообучают таким образом, что векторы, вычисленные ими, сохраняли смысловые отношения между фразами – похожие по смыслу предложения кодируются в близкие по метрике векторы. Для таких моделей удобно применять метод instance-based learning (обучение на основе экземпляров-примеров). В настоящее время для SentenceTransformer на сайте разработчиков опубликованы ссылки на множество предобученных моделей, в том числе и для русского языка.

В контексте задачи поиска семантически близких навыков рынка труда отдельной задачей становится поиск подходящего размеченного датасета, в котором были бы собраны и размечены семантически близкие формулировки требований рынка труда.

Анализ текстов профессиональных стандартов РФ показал, что они не содержат формулировок навыков в профессиональных терминах, которые используются в текстах требований реального рынка труда, что делает их использование совместно с нейросетевыми подходами (см. выше) нецелесообразным.

В процессе анализа англоязычных ресурсов были определены два проекта: европейская классификация навыков ESCO (далее ESCO) [21] и ONET [22, 23].

ESCO разрабатывается Европейской комиссией в Европейском центре развития профессионального образования Cedefop с 2010 года. Классификация ESCO определяет и классифицирует навыки, компетенции, квалификации, профессии и связи между ними, имеющие отношение к рынку труда ЕС, образованию и профессиональной подготовке кадров. ESCO была разработана в открытом ИТ-формате, доступна для бесплатного использования всеми и доступна через сервисную

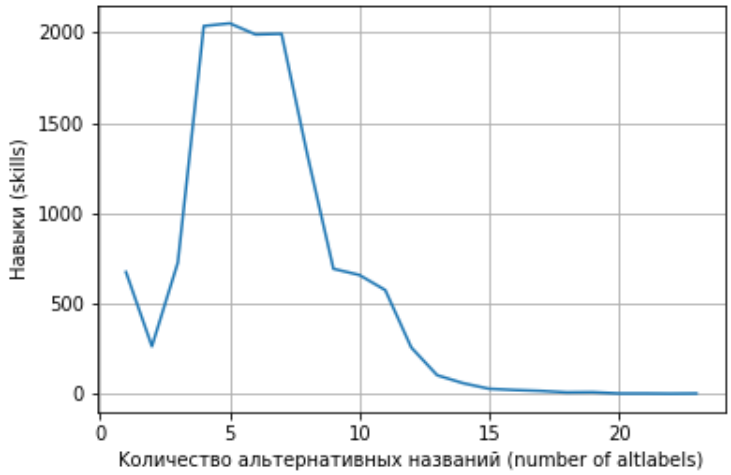
платформу. ESCO находится в постоянном развитии, базы регулярно обновляются и дополняются актуальной информацией. Результатами проекта регулярно пользуются все 28 стран – членов ЕС.

Система O*NET является американским аналогом системы ESCO. Центральным элементом проекта является база данных O*NET, содержащая сотни стандартизованных и специфичных для занятий дескрипторов почти для 1000 профессий, охватывающих всю экономику США. База данных, которая доступна для общественности бесплатно, постоянно обновляется на основании информации, поступающей от широкого круга работников каждой профессии.

В работе используется классификация навыков ESCO, так как она превосходит по количеству описанных навыков модель O*NET, а также имеет компонент, содержащий различные формулировки описания одного и того же навыка.

1. Описание датасета

В качестве основного датасета для эксперимента был выбран компонент ESCO, в котором представлен список навыков различных профессий на английском языке. Данный список содержит 13485 шт. концепций навыков, структурированных в иерархию. Каждый навык имеет приоритетное название (prefered_label) и альтернативные формулировки названий (alt_labels) (2–15 шт.), что увеличивает общую базу текстов навыков до 97 574 шт. Распределение количества альтернативных названий по навыкам представлено на рисунке.



Распределение количества альтернативных названий навыков
Distribution of the number of alternative named skills

Отдельно стоит отметить, что формулировки текстов навыков в базе ESCO представлены на естественном языке, что является большим плюсом, в контексте анализа текстов требований вакансий реального рынка труда. Также в датасете для каждого навыка присутствует поле описания (description), что может быть полезным для обогащения коротких текстов навыков дополнительной информацией в дальнейших исследованиях.

Пример названий навыков и их альтернативных названий из базы навыков ESCO представлен в табл. 1.

Таблица 1

Примеры текстов навыков ESCO

Table 1

Examples of ESCO Skill Texts

№	preferredLabel	altLabels
1	manage musical staff	manage staff of music, coordinate duties of musical staff, manage music staff, direct musical staff
49	aid customers	support clients, aid clients recommend to clients, recommend to customers, help clients, support customers, help customers
67	bend staves	bending staves,shape staves, form staves, curve staves, bow staves, bend stave

Окончание табл. 1
Table 1 (end)

№	preferredLabel	altLabels
87	perform toxicological studies	apply toxicological testing methods, perform toxicological tests, perform toxicological study, carry out toxicological studies
134	identify available services	establish available services, determine rehabilitation services, analyse rehabilitation services, establish rehabilitation services, determine available services, analyse available services, classify available services, classify rehabilitation services
145	ensure coquille uniformity	making sure coquille is uniform, ensure uniformity of coquille, ensuring coquille uniformity, checking that coquille is uniform, ensuring uniformity of coquille, ensuring coquille is uniform, ensure coquille is uniform, ensuring of coquille uniformity, make sure coquille is uniform, check that coquille is uniform
201	manufacture ingredients	ingredients manufacture, assemble ingredients, manufacture of ingredients, produce ingredients, construct ingredients, manufacture of an ingredient, fabricate ingredients

Помимо основного датасета навыков ESCO средствами автоматического перевода Yandex ('https://translate.api.cloud.yandex.net) и Google ('https://cloud.google.com/translate) были получены датасеты навыков для русского языка.

2. Описание моделей и способов получения векторных представлений коротких текстов навыков

Для проведения эксперимента были отобраны следующие базовые нейросетевые модели, построенные на архитектуре трансформеров: для английского модели представлены в табл. 2; для русского языка модели представлены в табл. 3.

Таблица 2

Базовые модели для английского языка

Table 2

Basic Models for English

Название модели	Слоев / Головы	Скрытых измерений	Размер словаря	Параметры	Данные
bert-base-cased	12/12	768	28 996	109M	BookCorpus, Wikipedia
bert-large-cased	24/16	1024	28 996	336M	BookCorpus, Wikipedia
bert-base-multilingual-cased	12/12	768	119 547	110M	Wikipedia на 104 языках

Таблица 3

Базовые модели для русского языка

Table 3

Basic models for the Russian language

Название модели	Слоев / Головы внимания	Скрытых измерений	Размер словаря	Количество параметров	Данные
DeepPavlov/rubert-base-cased	12/12	768	119 547	180M	*Получена из мультиязычной версии BERT путем transfer learning [25]
sberbank-ai/ruBert-base	12/12	768	120 138	178M	16 млрд токенов из различных датасетов 300 GB
sberbank-ai/ruBert-large	24/16	1024	120 138	427M	
sberbank-ai/ruT5-base	12/12	768	32 101	222M	

Для базовых моделей сравнивались основные способы получения векторов предложений:

• CLS_vector: первый вектор последнего скрытого слоя BERT. Считается, что CLS токен кодирует в себя всю репрезентативную информацию обо всех токенах предложения с помощью процедуры многоуровневого кодирования. Представление CLS индивидуально в разных предложениях. Этот вектор часто используют в задачах классификации предложений.

• MEAN_vector: средний вектор по всем векторам всех токенов на последнем скрытом слое BERT.

• POOLER_vector: состояние первого маркера последовательности (маркер классификации), прошедший обработку через слои, используемые для вспомогательной задачи предварительного обучения. Например, для моделей семейства BERT возвращается маркер классификации после обработки через линейный слой и функцию активации tanh. Веса линейного слоя обучаются на основе задачи прогнозирования (классификации) следующего предложения во время предварительной подготовки.

Помимо базовых моделей в эксперимент были отобраны топ 5 английских моделей из рейтинга, указанного на сайте разработчиков SentenceTrasnformer для английского языка (табл. 4). Полный список моделей приведен по ссылке: https://www.sbert.net/docs/pretrained_models.html#model-overview. Рейтинг моделей основан на средней производительности кодирования предложений для 14 различных задач NLP из разных предметных областей. Все модели являются универсальными и оптимизированы для многих сценариев использования.

Для русского языка были отобраны две модели, обученные для SentenceTransformer от научных групп DeepPavlov и Sberbank-AI (описание моделей представлено в табл. 5). Вектора, полученные с помощью этой библиотеки и предобученных моделей, в эксперименте отмечены SENTENCE_vector.

Таблица 4

SentenceTransformer модели для английского языка

Table 4

SentenceTransformer models for English

Название модели	Слоев / Головы внимания	Скрытых измерений	Размер словаря	Параметры	Данные
microsoft/mpnet-base	12/12	768	30 527	–	Reddit comments (15–18) S2ORC Citation pairs (Abstracts) WikiAnswers Duplicate question pairs and etc. 1B + training pairs
distilroberta-base	12/12	768	30 527	–	
microsoft/MiniLM-L12-H384-uncased	12/12	384	30 522	–	
nreimers/MiniLM-L6-H384-uncased	6/12	384	30 522	–	
sentence-transformers/multi-qa-distilbert-cos-v1	6/12	768	30 522	–	215 млн пар (вопрос, ответ) из разных источников

Таблица 5

SentenceTransformer модели для русского языка

Table 5

SentenceTransformer models for the Russian language

Название модели	Слоев / Головы внимания	Скрытых измерений	Размер словаря	Количество параметров	Данные
DeepPavlov/rubert-base-cased-sentence	12/12	768	119 547	180M	SNLI [26], переведенный на русский язык, и русский XNLI dev set [27]
sberbank-ai/ sbert_large_nlu_ru	24/16	1024	120 138	426.9M	16 млрд токенов из различных датасетов 300 GB

3. Описание эксперимента и методика оценки

Основная задача всего эксперимента заключается в определении нейросетевой модели, способной формировать кластеры, по принципу: семантически близкие тексты требований, относящиеся к одному навыку, должны располагаться ближе друг к другу, и при этом группы текстов разных навыков должны быть удалены друг от друга. Исходя из этого в качестве основной меры оценки получаемых кластеров была выбрана оценка Силуэт [24].

Оценка Силуэт (англ. Silhouette)

Оценка Силуэт была разработана Питер Дж. Руссеу для интерпретации и проверки кластерного анализа.

Значение Силуэта показывает, насколько объект похож на свой кластер по сравнению с другими кластерами.

Оценка для всей кластерной структуры:

$$Sil(C) = \frac{1}{N}\sum_{c_k \in C}\sum_{x_i \in c_k}\frac{b(x_i,c_k)-a(x_i,c_k)}{\max\{a(x_i,c_k),b(x_i,c_k)\}}, \quad (1)$$

где

$$a(x_i, c_k) = \frac{1}{|c_k|}\sum_{x_j \in c_k}\|x_i - x_j\|; \quad (2)$$

$$b(x_i, c_k) = min_{c_l \in C \setminus c_k}\{\frac{1}{|c_l|}\sum_{x_j \in c_l}\|x_i - x_j\|\}; \quad (3)$$

$a(x_i, c_k)$ – среднее расстояние от $x_i \in c_k$ до других объектов из кластера c_k (компактность) (2);

$b(x_i, c_k)$ – среднее расстояние от $x_i \in c_k$ до объектов из другого кластера $c_l: k \neq l$ (отделимость) (3).

Оценка *Sil*(*C*) лежит в пределах от –1 до 1. Чем ближе данная оценка к 1, тем лучше.

Итоговый алгоритм оценки качества кластеров семантически близких коротких текстов навыков для компонента классификации навыков ESCO:

1) получение векторных представлений для текстов навыков из нейросетевых моделей методами CLS, MEAN, POOLER, SentenceTransformer;

2) вычисление оценки Силуэт для всех моделей и для всех методов получения векторных представлений текстов навыков.

4. Описание результатов

Результаты сравнения качества векторных представлений в задаче формирования кластеров семантически близких текстов требований для разных нейросетевых моделей по оценке Силуэт представлены в таблицах:

- для английского языка (табл. 6);
- для русского языка (табл. 7).

Таблица 6

Результаты сравнения нейросетевых моделей по оценке Силуэт для английского языка

Table 6

Comparison results of neural network models according to the Silhouette assessment for the English language

Name model	CLS	MEAN	POOLER	SENTENCE
bert-base-cased	0,0118	0,1111	–0,2887	
bert-large-cased	–0,1431	0,0955	–0,5180	
bert-base-multilingual-cased	–0,0111	0,0900	–0,0745	
distilroberta-base	0,1034	0,0883	0,1005	
microsoft/MiniLM-L12-H384-uncased	–0,0517	0,0386	–0,0584	0,0386
microsoft/mpnet-base	0,0575	0,0566	0,0564	0,0566
nreimers/MiniLM-L6-H384-uncased	0,0022	0,1139	–0,0025	0,1139
roberta-base	0,0642	0,0727	0,0607	0,0727
sentence-transformers/multi-qa-distilbert-cos-v1	0,1588	0,1578	0,1556	**0,2585**

Таблица 7

Результаты сравнения нейросетевых моделей по оценке Силуэт для русского языка

Table 7

Comparison results of neural network models according to the Silhouette assessment for the Russian language

Name model	CLS		MEAN		POOLER		SENTENCE	
	Google	Yandex	Google	Yandex	Google	Yandex	Google	Yandex
DeepPavlov/rubert-base-cased	–0,110	–0,145	0,067	0,055	–0,403	–0,389		
sberbank-ai/ruBert-base	0,030	0,026	0,116	0,111	–0,109	–0,116		
sberbank-ai/ruBert-large	0,011	–0,002	0,093	0,089	–0,27	–0,27		
sberbank-ai_ruT5-base	–0,199	–0,211	–0,06	–0,043				
sberbank-ai_ruT5-large	–0,157	–0,218	–0,020	0,001				
DeepPavlov/rubert-base-cased-sentence	0,123	0,13	0,138	0,143	0,076	0,08	**0,138**	**0,143**
sberbank-ai/ sbert_large_nlu_ru	–0,019	–0,0539	0,0633	0,036	–0,129	–0,177	0,063	0,036

Особенности перевода текстов навыков с английского языка на русский с использованием сервисов автоматического перевода Яндекс и Google. В результате автоматического перевода количество уникальных текстов после перевода от Яндекса и Google снизилось с 97 574 уникальных текстов до 75 693 (22 %), и до 77 946 (20 %) соответственно. Данное обстоятельство можно объяснить тем, что мы работаем с группами текстов, изначально очень близких по смыслу, и сервисы автоматического перевода в ряде случаев просто не могу уловить особенности перевода с английского на русский и поэтому переводят часть текстов одинаково, что и приводит к снижению количества уникальных текстов.

Заключение

Рассмотренный в статье метод оценки нейросетевых моделей на основе сравнения компактности векторных представлений позволил эффективно ранжировать нейросетевые модели для задачи выделения компактных групп семантически близких текстов профессиональных навыков.

- По результатам из табл. 6 и 7 модели sentence как для английского, так и для русского языка превосходят базовые модели. Можно сделать вывод, что процесс дообучения моделей на задаче определения семантически близких предложений увеличивает качество формируемых кластеров профилей навыков из классификации ESCO.
- Среди моделей для английского языка наилучший результат по оценке Силуэт получила модель sentence-transformers/multi-qa-distilbert-cos-v1 (sentence silhouette 0,2585), это может быть обусловлено особенностями датасетов, на которых обучалась изначальная базовая модель, и наличием в них специфической лексики, что помогает модели лучше справляться в задаче формирования кластеров профилей навыков из классификации ESCO. Высокий результат таже показали модели nreimers/MiniLM-L6-H384-uncased (sentence silhouette 0,1139) и bert-base-cased (mean silhouette 0,1111).
- Среди моделей для русского языка также наилучший результат показали модели на основе sentence_transformer. Модель от группы DeepPavlov rubert-base-cased-sentence по оценке Силуэт значительно превосходит остальные модели для русского языка.
- На примере наилучшей модели для русского языка можно заметить небольшую, но стабильную по всем методам получения векторов предложений разницу в пользу датасета автоматического перевода от Яндекс. Этим может быть обусловлено тем обстоятельством, что перевод от Яндекс привел к снижению уникального количество текстов, что, в свою очередь, привело к сужению пространства возможных значений векторов, что в итоге дало чуть более высокий конечный результат.

Список литературы/References

1. Комиссия Великобритании по трудоустройству и профессиональным навыкам, «Важность LMI». 2015. [Электронный ресурс]. URL: https://www.gov.uk/government/publications/the-importance-of-labour-market-intelligence (дата обращения: 30.05.2022). [*UK Commission for Employment and Skills, "The importance of LMI"*. 2015. Available at: https://www.gov.uk/government/publications/the-importance-of-labour-market-intelligence (accessed 30.05.2022). (In Russ.)]

2. Mezzanzanica M., Mercorio F. Big data enables labor market intelligence. In: *Encyclopedia of Big Data Technologies*; 2019. P. 226–236. DOI: 10.1007/978-3-319-63962-8_276-1

3. *Concept paper on Labour Market Information System*. 2012. Available at: http://www.cgsc.in/Concept_Paper_LMIS.pdf (accessed 30.05.2022).

4. Vinel M., Ryazanov I., Botov D., Nikolaev I. Experimental Comparison of Unsupervised Approaches in the Task of Separating Specializations Within Professions in Job Vacancies. In: *Conference on Artificial Intelligence and Natural Language*. Springer, Cham; 2019. P. 99–112. DOI: 10.1007/978-3-030-34518-1_7

5. Nikolaev I., Ryazanov I., Botov D. The Comparison of Distributive Semantics Models Applied to the Task of Short Job Requirements Clustering for the Russian Labor Market. In: *8th Scientific Conference on Information Technologies for Intelligent Decision Making Support (ITIDS 2020).* Atlantis Press; 2020. P. 295–301. DOI: 10.2991/aisr.k.201029.056

6. Giabelli A., Malandri L., Mercorio F., Mezzanzanica M. GraphLMI: A data driven system for exploring labor market information through graph databases. In: *Multimedia Tools and Applications*; 2020. P. 1–30. DOI: 10.1007/s11042-020-09115-x

7. Colombo E., Mercorio F., Mezzanzanica M. Applying machine learning tools on web vacancies for labour market and skill analysis. In: *The Economics and Policy Implications of Artificial Intelligence*; 2018.

8. Boselli R., Cesarini M., Marrara S., Mercorio F., Mezzanzanica M., Pasi G., Viviani M. WoLMIS: a labor market intelligence system for classifying web job vacancies. *Journal of intelligent information systems*. 2018;51(3):477–502. DOI: 10.1007/s10844-017-0488-x

9. Kane L.O., Narasimhan R., Burning J.N., Taska B. *Digitalization in the German Labor Market: Analyzing Demand for Digital Skills in Job Vacancies*. Bertelsmann Stiftung; 2020.

10. Harris Z.S. Distributional structure. *Word*. 1954;10(2-3):146–162.

11. Jones K.S. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*. 2004;60(5):493–502. DOI: 10.1108/00220410410560573

12. Mikolov T., Chen K., Corrado G., Dean J. *Efficient estimation of word representations in vector space*. 2013. Available at*:* https://arxiv.org/pdf/1301.3781.pdf (accessed 30.05.2022).

13. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Polosukhin I. Attention is all you need. In: *Advances in neural information processing systems*; 2017. P. 5998–6008.

14. Devlin J., Chang M.W., Lee K., Toutanova K. *BERT: Pre-training of deep bidirectional transformers for language understanding*. 2018. Available at*:* https://arxiv.org/pdf/1810.04805.pdf (accessed 30.05.2022).

15. Ezen-Can A. *A Comparison of LSTM and BERT for Small Corpus*. 2020. Available at*:* https://arxiv.org/pdf/2009.05451.pdf (accessed 30.05.2022).

16. Liu Y., Ott M., Goyal N., Du J., Joshi M., Chen D., Stoyanov V. *RoBERTa: A robustly optimized bert pretraining approach*. 2019. Available at*:* https://arxiv.org/pdf/1907.11692.pdf (accessed 30.05.2022).

17. Brown T., Mann B., Ryder N., Subbiah M., Kaplan J.D., Dhariwal P., Amodei D. *Language models are few-shot learners*. Available at*:* https://arxiv.org/pdf/2005.14165.pdf (accessed 30.05.2022).

18. Raffel C., Shazeer N., Roberts A., Lee K., Narang S., Matena M., Liu P.J. *Exploring the limits of transfer learning with a unified text-to-text transformer*. 2019. Available at*:* https://arxiv.org/pdf/1910.10683.pdf (accessed 30.05.2022).

19. Lample G., Conneau A. *Cross-lingual language model pretraining*. 2019. Available at*:* https://arxiv.org/pdf/1901.07291.pdf (accessed 30.05.2022).

20. Reimers N., Gurevych I. *Sentence-bert: Sentence embeddings using siamese bert-networks*. 2019. Available at*:* https://arxiv.org/pdf/1908.10084.pdf (accessed 30.05.2022).

21. *ESCO: European skills, competences, qualifications and occupations*. 2018. Available at: https://ec.europa.eu/esco/portal (accessed 30.05.2022).

22. Peterson N.G., Mumford M.D., Borman W.C., Jeanneret P., Fleishman E.A. An occupational information system for the 21st century: The development of O* NET. In: *American Psychological Association*; 1999. DOI: 10.1037/10313-000

23. Peterson N. G., Mumford M. D., Borman W. C., Jeanneret P.R., Fleishman E.A., Levin K.Y., Dye D.M. Understanding work using the Occupational Information Network (O* NET): Implications for practice and research. *Personnel psychology*. 2001;54(2):451–492. DOI: 10.1111/j.1744-6570.2001.tb00100.x

24. Rousseeuw P.J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*. 1987;20:53–65. DOI: 10.1016/0377-0427(87)90125-7

25. Kuratov Y., Arkhipov M. *Adaptation of deep bidirectional multilingual transformers for Russian language*. 2019. Available at*:* https://arxiv.org/pdf/1905.07213.pdf (accessed 30.05.2022).

26. Bowman S. R., Angeli G., Potts C., Manning C.D. *A large annotated corpus for learning natural language inference*. 2015. Available at*:* https://arxiv.org/pdf/1508.05326.pdf (accessed 30.05.2022).

27. Conneau A., Lample G., Rinott R., Williams A., Bowman S.R., Schwenk H., Stoyanov V. *XNLI: Evaluating cross-lingual sentence representations*. 2018. Available at*:* https://arxiv.org/pdf/1809.05053.pdf (accessed 30.05.2022).

Информация об авторах

Николаев Иван Евгеньевич, старший преподаватель кафедры информационных технологий и экономической информатики, Челябинский государственный университет, Челябинск, Россия; ivan_nikolaev@csu.ru.

Мельников Андрей Витальевич, д-р техн. наук, проф., директор, Югорский научно-исследовательский институт информационных технологий, Ханты-Мансийск, Россия; MelnikovAV@uriit.ru.

Information about the authors

Ivan E. Nikolaev, Senior Lecturer, Department of Information Technologies and Economic Informatics, Chelyabinsk State University, Chelyabinsk, Russia; ivan_nikolaev@csu.ru.

Andrey V. Melnikov, Dr. Sci. (Eng.), Prof., Director, Ugra Research Institute of Information Technologies, Khanty-Mansiysk, Russia; MelnikovAV@uriit.ru.

Статья поступила в редакцию 30.05.2022
The article was submitted 30.05.2022