The object of the study is context-aware phrase representations. The growing need to automate candidate recruitment and job recommendation processes has paved the way for the utilization of text embeddings. These embeddings involve translating the semantic essence of text into a continuous, high-dimensional vector space. By learning context-aware rich and meaningful representations of phrases within the human resource domain, the efficacy of similarity searches and matching procedures is enhanced, which contributes to a more streamlined and effective recruitment process. However, existing approaches do not take into account the context when modeling phrases. This necessitates the improvement of information technology analysis in this area. In this paper, it is proposed to mark the beginning and end of phrases in the text using special tokens. This made it possible to reduce the requirements for computing power by calculating all phrase representations present in the text simultaneously. The effectiveness of the improvement was tested on a new dataset to compare and evaluate the models in the task of modeling phrases in the field of human resources management. The proposed approach to modeling phrase representations with regard to context in the field of human resources management leads to an improvement in computational efficiency by up to 50 % and an increase in accuracy by up to 10 %. The architecture of the machine learning model for creating context-aware phrase representations is developed, which is characterized by the presence of blocks for taking into account phrase boundaries. Experiments and comparisons with existing approaches have confirmed the effectiveness of the proposed solution. In practice, the proposed information analysis technology can be used to automate the process of identifying and normalizing candidates' skills in online recruiting

Keywords: natural language processing, large language models, text embeddings, information retrieval

UDC 004.89

DOI: 10.15587/1729-4061.2024.313970

IMPROVING INFORMATION THEORY OF CONTEXT-AWARE PHRASE EMBEDDINGS IN HR DOMAIN

Maiia Bocharova
Corresponding author
PhD Student*
E-mail: bocharova.maiia@gmail.com
Eugene Malakhov
Doctor of Technical Sciences, Professor,
Head of Department*
*Department of Mathematical Support
of Computer Systems
Odesa I.I. Mechnikov National University
Vsevoloda Zmiienka str., 2. Odesa, Ukraine, 65082

Received date 09.07.2024
Accepted date 02.10.2024
Published date 30.10.2024

***How to Cite**: Bocharova, M., Malakhov, E. (2024). Improving information theory of context-aware phrase embeddings in HR domain. Eastern-European Journal of Enterprise Technologies, 5 (2 (131)), 53–60. https://doi.org/10.15587/1729-4061.2024.313970*

1. Introduction

Skills are pivotal in both the job market and human resources (HR) processes, where candidates seek relevant opportunities based on their skill set, and enterprises strive to ensure the future-proofing of their workforce's skills. However, the absence of structured skill information and the challenges associated with self- or manager-assessment approaches have impeded effective skill management.

Automated skill extraction techniques have emerged as promising solutions [1, 2], however normalization of skills and their mapping still remains a significant challenge. Skills are often expressed in different forms, terminologies, or levels of granularity, making it challenging to compare and categorize them accurately. For instance, a single skill can be described using multiple variations, synonyms, or formulations, leading to inconsistencies and difficulties in matching skills across different contexts. For example, "project management," "managing projects," or "project coordination" all refer to the same underlying concept but are expressed differently.

Existing resources such as O*NET [3] or official frameworks like the European Qualifications Framework (EQF), a part of the European Skills/Competences, Qualifications and Occupations commission [4], serve as public databases of skills. However, these databases have a significant drawback when it comes to detecting new skills or identifying known skills expressed in novel ways.

For instance, consider an applicant profile claiming proficiency in "onboarding new hires," which refers to their ability to mentor and guide new employees. Although skill databases like O*NET and ESCO include sections on teaching, training, coaching, and mentoring; they do not explicitly mention the term "onboarding". Consequently, existing skill extraction methods relying on these databases would be unable to identify this particular skill. Moreover, while ESCO updates its terminology annually, skill extraction methods leveraging this database could still be up to a year behind.

Text embeddings, which are continuous, low-dimensional representations of text, have been widely used in a range of downstream applications. These include tasks like information retrieval, question answering, and retrieval augmented generation (RAG) [5, 6]. By aligning extracted skills phrasx es with a common taxonomy and automatically expanding existing taxonomy with new concepts, it becomes possible to compare skills across different industries, job descriptions, or skill frameworks, improving the interoperability and transferability of skill-related information.

To address this challenge, various approaches can be employed to determine the similarity of skills and group the similar ones together. For example, by comparing vector representations of extracted skill phrases clusters of similar skills can be created, so that later the discovered clusters of related skills can be either mapped to the concepts which are already present in the taxonomy or integrated into the taxonomy as new categories.

Automated skills analysis can also help organizations determine what skills their employees already possess and highlight those that may need further development. Identification of new, in-demand competencies in the labor market also enables more targeted training initiatives, helping ensure that employees remain competitive.

English-language texts are particularly important for the study, as this language is the language of international communication, including in Ukraine [8].

Therefore, research on modeling context aware phrase representation for English is relevant.

2. Literature review and problem statement

The Transformer architecture [9] is the most widely used for solving natural language processing problems. The most common and promising modern models of this type of architecture will be discussed below.

Learning sentence representations is crucial for many natural language processing (NLP) tasks, including semantic parsing, machine translation and question answering. Out-of-the-box BERT [10] sentence embedding models often fall short when compared to simple baselines like averaging GloVe vectors in semantic textual similarity tasks [11]. This limitation arises from BERT's learning objectives, and to obtain accurate scores for pairs of sentences, both sentences need to be simultaneously processed. In response to this challenge, extensive research has been dedicated to adapting BERT embeddings to better represent sentences at the semantic level, which led to the development of models such as the Universal Sentence encoder and SentenceBERT [11]. These models were trained with supervision, leveraging large human-labeled datasets, in order to enhance their ability to generate meaningful sentence-level representations. However, the need to have a large amount of human-verified and labeled data limits the use of this approach. In addition, these approaches are not designed to model shorter lexical units, such as phrases.

Phrase-level representations play a crucial role in a wide range of tasks, including paraphrase detection, question answering, and topic modeling. The authors of [12] train BERT using synthetically generated paraphrases to detect lexically distinct texts. However, this approach relies on synthetic data and will be limited by the knowledge contained in the large language model used for generating the data.

Sentence similarity benchmarks [13] were created to allow comparing the ability of models to generate semantically meaningful representations of texts. This dataset consists of text pairs marked by humans in terms of their similarity. Although this approach to sentence comparison is widely adopted, using a similar technique for phrases has significant drawbacks. For example, phrases can change their meaning depending on the context, and their similarity can be misinterpreted without taking this context into account. These challenges indicate a need for a phrase similarity benchmark that provides contextual sentences for each phrase.

In the McPhraSy approach [14], the authors propose to use a 2-layer MLP that combines representations obtained from two separate pre-trained models. Thus, the basis is the representation of a phrase obtained from the PhraseBERT model and the representation of a mask in place of the same phrase in a random sentence obtained from the SpanBERT model. Based on these two representations, the final context-aware phrase representation is generated. This approach demonstrates effective results and relies only on real texts, without using large language models to generate data. However, it has the drawback of consuming excessive computational resources due to its reliance on two separate models.

Therefore, the importance of developing models for context-aware phrase representations cannot be overestimated.

In the human resources management domain, an important element is the study of the meaningful representation of job titles based on skills that appear in the same job description [15]. However, this approach has limitations when applied to individual skills. Grouping and normalizing skills which frequently co-appear could help address these challenges.

Skills grouping and normalization on the other hand is less researched. The task is mostly approached either as an Extreme Multi Label Classification (XMLC), whereas as a unit for classification in some research the sentence [16] is taken, and in some – the whole job description [17]. After that, the model is trained to evaluate the semantic similarity between the name of the skill and the sentence in which it appeared [16, 17]. However, this limits the ability of such architectures to adapt to other application areas, as they have a fixed number of classes that they can process.

Alternatively, a data-driven approach for skills grouping was introduced [18], where authors used a graph network. The approach described in their study conceptualizes each skill as a node within a network, with the connections or vertices between these nodes being established based on two key factors: the co-occurrence of skills in pairs, and the lexical similarity between different skills. This method offers a nuanced perspective on how skills relate to and influence each other in a professional context. However, the study faced a significant challenge in terms of computational demands. Graph network models, especially those with a large number of nodes and connections, require substantial computational resources. This limitation became evident in the study's scope, as the authors were constrained to analyze only 10,554 skills in their model, which is vastly inferior to the multitude of skill variations found in real-world environments.

The most promising recent approaches rely on the large language models (LLMs) generated synthetic data [19], showing that LLMs are capable of generating realistic data for further training of smaller models. As such, in [9] authors use bicoder for mapping skills from ESCO database to synthetically generated sentences, which contain those phrases into the shared embedding space, and the model is trained to bring such representation close in the vector space. However, models trained on synthetic data may struggle to generalize to unseen concepts or novel skill descriptions that differ substantially from the synthetic examples. Additional challenges can arise when discovering new skills, which the LLM has not seen due to the historical nature of its pre-training data. This can limit the model's ability to handle real-world skill variations effectively.

Thus, there is no single approach to the representation of phrases which take into account the context, which requires additional research on the adaptation of model architecture.

3. The aim and objectives of the study

The aim of the study is to improve the information technology for modeling contextually aware phrase representa-

tions. This will make it possible to effectively apply information models in the context of human resource management.

To achieve this aim, the following objectives must be solved:

– to develop a machine learning architecture for context-aware modeling of phrase representations;

– to establish the benchmark for the similarity of skill phrases containing contextual paragraphs, implement developments in the field of human resources management and evaluate the result;

– to conduct an experimental testing of the technology for analyzing context-aware phrase representations.

4. Materials and methods

The object of the study is the process of natural language processing used to analyze and interpret textual information in documents related to recruitment and personnel management, such as resumes and job postings. The main hypothesis is the possibility of using special tokens to mark phrase boundaries in order to improve the accuracy of modeling phrase representations in context. The development of the model was based on the assumption that it is possible to accurately extract important skill phrases from the text. For the sake of simplicity, models of BERT-base size were chosen for comparison.

Methods [14, 17] suffer either from being restricted to a very small percentage of skills or from regarding the whole sentence as a minimum unit for classification, when in real-world vacancies this assumption is valid only for ~40 % of the vacancy sentences (Fig. 1). Arguably the model will also benefit from seeing more context than just one sentence (namely paragraph).

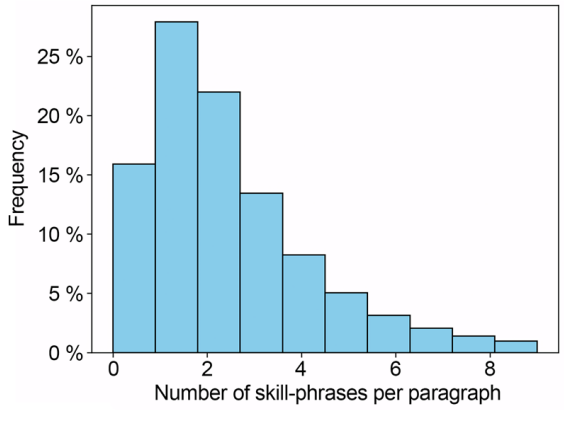

Fig. 1. Number of skill phrases in one paragraph in vacancies

As for the resumes, which mostly come from PDF documents, the challenge on its own lies in restoring sentence boundaries after parsing. This problem originates from the insertion of extra new line characters, which is inherent to processing such documents. Incorrectly splitting sentences by punctuation, which usually marks the end of the sentence, can result in the loss of valuable skills and contamination of context. Because of this, opting for longer text chunks when dealing with work histories extracted from resumes can be beneficial.

The subject of the study is context-aware methods of modeling phrase representation.

Multi-context Phrase Similarity [14] representation learning method is most closely related to our approach. Indeed, it leverages context information not only during training, but during inference, which enables it to adapt to phrases which change their meaning according to context. Additionally, it lets this approach to work seamlessly with phrases not seen during training. However, this method has two drawbacks: it uses two separate models, which slows down its speed; and during training, only the final layer is tuned, while the weights of the base model remain unchanged, which does not allow the model to fully adapt to new domains. In addition, as shown in Fig. 1, in HR-related texts (resumes and job postings), an average of 2.18 skills are present in one paragraph. Therefore, calculating the representations for each phrase by passing the same chunk of text with a different mask through the model, as well as calculating the representations for each phrase by a separate model, is very computationally expensive. A more efficient strategy is to reduce computational requirements by introducing special markers within the text. These markers, placed at the beginning and end of each skill phrase, enable the extraction of all vector representations in a single pass through the model.

Thus, the method of multi-context phrase similarity [14] needs to be improved. To achieve this goal, it is possible to use a single model trained using a context-aware phrase alignment strategy, where each phrase is marked with its own boundaries, and the vector corresponding to the left boundary token is taken as the phrase representation. The method of multi-contextual phrase similarity for phrase representation improved in this way is defined as SkillPhrase2Vec.

To mark the beginning and end of a phrase, two new tokens initialized from unused BERT tokens are used. All skills detected in the paragraph are marked this way.

Anchor skill phrase *ph* should be defined as any chosen phrase. Given such anchor skill phrase *ph*, belonging to a paragraph t_i, a random paragraph t_j, which contains a positive phrase *ph*+ – the same phrase as anchor for unsupervised setting or a phrase which was identified as a synonym phrase by a large pre-trained model, from a different job description is selected.

After that phrase *ph* in t_i is masked using a single "[MASK]" token, and the phrase *ph*+ in t_j which serves as a positive example is left unchanged.

Vector representation of the token, which marks the left boundary of the phrase, is used as a representation of the skill phrase.

The concept of contrastive learning, particularly with in-batch negatives, has become a prevalent technique in the research community for learning robust text embeddings. This approach allows for the efficient reuse of calculated embeddings during training for every possible pair within a batch. Building on this concept, the Multiple Negative Ranking Loss is utilized. This loss function operates by considering each pair (q_j, p_j) within a given batch, denoted as B. The size of this batch is defined as k. The Multiple Negative Ranking Loss is then calculated for each of these pairs as follows:

$$Loss_{q_j,p_j} = -\log_e \frac{e^{s(q_j,p_j)}}{\sum_{i=1}^{k} e^{s(q_j,p_i)}}. \quad (1)$$

where $s(q_j, p_j)$ is the cosine similarity of vectors q_j and p_j; q_j is the representation of the masked phrase at the j-th position in the batch; p_j is the representation of the unmasked

phrase at the j-th position in the batch; e is the base of the natural logarithm; $\log_e$ is the natural logarithm; k is the size of the batch; $Loss_{q_j,p_j}$ is the loss relative to the specified index for the pair q_j, p_j.

Subsequently, the model is trained to minimize the distance between the representations of the left phrase boundary token in positive pairs. This token should encapsulate the meaning of the particular phrase under consideration in both scenarios: masked (query q_j) and not masked (positive p_j). Simultaneously, divergence in the representations of the phrases which are not related is enforced, thereby facilitating the acquisition of impactful phrase representations.

In line with findings from the [20, 21] research, which underscored the benefits of calculating loss in a bidirectional manner, our method also incorporates this insight. Hence, let's extend our loss calculation to include (p, q) pairs:

$$Loss_{p_j,q_j} = -\log_e \frac{e^{s(p_j,q_j)}}{\sum_{i=1}^{k} e^{s(p_j,q_i)}}. \quad (2)$$

where $s(p_j, q_j)$ is the cosine similarity of vectors p_j and q_j; q_j is the representation of the masked phrase at the j-th position in the batch; p_j is the representation of the unmasked phrase at the j-th position in the batch; e is the base of the natural logarithm; $\log_e$ is the natural logarithm; k is the size of the batch; $Loss_{p_j,q_j}$ is the loss relative to the specified indices for the pair p_j, q_j.

To encapsulate the overall training objective, the final loss is defined by the sum of the losses for both (q_j, p_j) and (p_j, q_j) pairs across all indexes j in the batch:

$$Loss = \sum_{i=1}^{k} \left(Loss_{q_j,p_j} + Loss_{q_j,p_j} \right), \quad (3)$$

where $Loss$ is the final loss; k is the size of the batch.

The minimization of equation (3) objective is to force the embedding function to produce closely aligned embeddings for tokens representing the same (or synonym) phrases, and at the same time to ensure greater separation for embeddings of phrases that are in-batch negatives.

The new approach to learning phrase representations for HR-related texts was tested in two scenarios. First – unsupervised contrastive learning, where instances of the same phrase appearing in different contexts were utilized as positive examples. And second – in semi-supervised settings, where signals from a large pre-trained lexical similarity model were leveraged for supervision.

To prepare the training data for experiments, because of the inconsistent quality of parsed resumes (and to lesser extent vacancies) it becomes essential to establish a thorough filtering process. Following classical NLP preprocessing pipeline [20] the following preprocessing stages were implemented to refine the dataset.

Language identification and filtering: Since the study is focused on English language, a pre-trained fasttext language detector [22] was used to filter out non-English texts. All resumes and job postings where the model's confidence that the text was in English was below 80 % were filtered out. 6 % of vacancies and 3 % of resumes were filtered out at this stage.

When processing job postings, the first step was to remove paragraphs that were not related to the requirements for candidates (sections about company and a description of benefits in job postings). To do this, a proprietary model owned by Daxtra Technologies was used. This model works at the paragraph level and categorizes them according to their relevance for determining the candidate's portrait. A description of the architecture of this model is beyond the scope of this study. This way 2.8 million requirements-related sentences were obtained.

Resumes were segmented and only the paragraphs that belonged to the work histories and the summary section were taken. This resulted in 3.4 million paragraphs.

Next step was data deduplication.

Both resumes and job postings contain a lot of identical or almost identical text (because people submit the same resume several times and recruiters repost the same job posting on several websites). Template descriptions or requirements are often also used). Duplicates in the training data can significantly degrade the performance and accuracy of a language model. In particular, training language models on deduplicated data leads to a reduction in the number of training steps required to achieve the same or higher degree of accuracy [23].

To mitigate the issue, all duplicate samples were removed from the dataset. Given the large size of the dataset, hash values were used to identify and remove these duplicates. Before the procedure, whitespaces were normalized and the text was converted to lowercase. This way, 53 % of the data was filtered out.

To ensure diversity of the dataset, internal occupational sector taxonomy of Daxtra Technologies, which consists of 28 top levels was used. Collected ata was automatically classified in accordance with this taxonomy, and class balancing was undertaken. This was achieved by downsampling samples which belonged to most widely represented sectors (like IT)

The last step was detection of skill phrases from the paragraphs using the proprietary model of Daxtra Technologies. This model operates on token-level. Sentences and paragraphs that did not contain at least one skill phrase were filtered out.

This resulted in ~2.1 million valid samples for our task. Fig. 2 graphically illustrates the distribution of professional sectors in the final training set.

For model training, skills that appeared in the texts at least 10 times were selected.

In order to test the approaches, it is necessary to have a benchmark dataset that reflects the real variability and complexity of skills found in job and resume texts and allows for an adequate assessment of the accuracy and efficiency of the representation of such skill phrases. Creating a manually labeled dataset of sufficient size to normalize skills is not possible when using fine-grained taxonomies (ESCO has more than 13 thousand skills). Using coarser taxonomies such as O*NET, where there are only 35 skill groups, is not useful for real-world applications. But it is possible to create a benchmark that, similar to well-known test data sets for general sentence similarity training (STS, etc.), will allow comparing phrases, which describe skills, with each other.

From the prepared set, let's randomly select 50 thousand skill phrases that occurred at least 10 times, and then grouped them around the most frequent 10 thousand from this sample. The grouping was done using heuristics and several models, each focusing on different aspects such as lexical, morphological, or contextual similarity.

To select the pairs for annotation, certain heuristics were taken into account, including the predicted professional sector, to identify potential incorrect candidates for each of the selected query phrases.

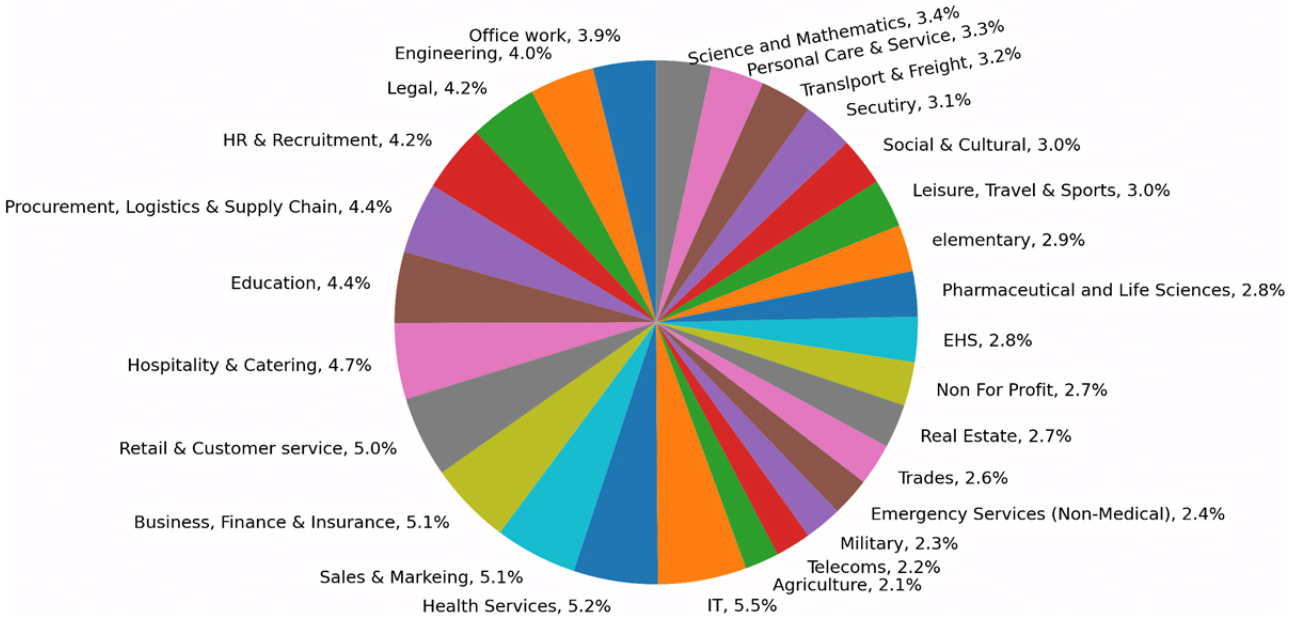

Fig. 2. Distribution of professional sectors in training data

The benchmark was constructed with the focus on the following:

– adversarial lexical match: phrases that have matching words, but are unrelated ("flower planting"→"plant flow"; "microsoft outlook"→"strategic outlook") were included. Models which are using lexical or morphological similarity are expected to struggle with such cases;

– words and phrase disambiguation: certain words/phrases can have multiple meanings depending on context. For example, the abbreviation "ER" can refer either to "employee relations" or to the "emergency room" depending on the surrounding context words;

– hard negatives: phrases where individual models did not agree between themselves were selected and annotated with great care to make the benchmark more challenging;

– implementation details: BERT-base was fine-tuned on an NVIDIA T4 GPU for 1 epoch when using batch size of 24 and on NVIDIA A100 GPU when using batch size 128. Initial 10 % of steps are used for warm-up in both settings, following which the learning rate is linearly increased to its maximum value. Let's utilize the AdamW optimizer with a learning rate of 2e-5 was utilized for both the T4 GPU and A100 GPU

5. Results of research on modeling the context-aware representation of phrases in the field of personnel management

5. 1. Development of the architecture of the model and procedure for learning context-dependent phrase representations

Fig. 3 shows the context-aware phrase representation modeling architecture. Unlike the classical architecture of the Transformer model [9], the proposed architecture includes additional blocks that allow defining phrase boundaries (Data Transformation block in Fig. 3). Thus, the proposed architecture is a modification of the classical Transformer architecture.

The experimental results confirm the dependence of the quality of the learned representations on the number of negative examples in the training batch (Fig. 4), which is in correlation with the data [24]. The results of the "all negative pairs" approach (Fig. 4) indicate a decrease in metrics on the validation set. Therefore, it is not advisable to recommend it for use.

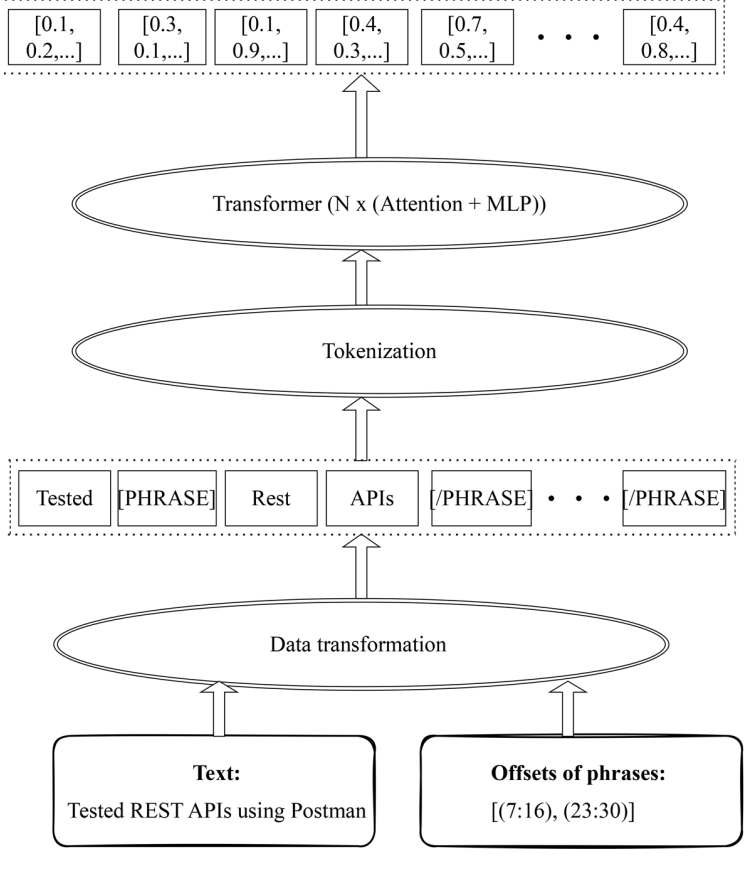

Fig. 3. Context-aware phrase representations modeling architecture

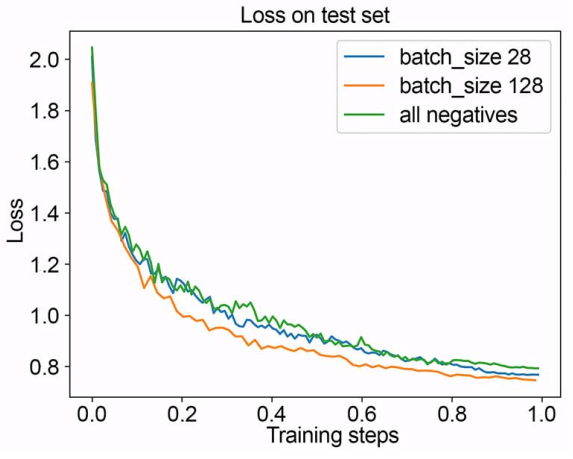

Fig. 4. Metrics during model training

The obtained results indicate the feasibility of using large data batches (128 pairs) for contrastive training. Thus, it is reasonable to use a suitably large batch size for practical purposes. The applied architecture, which is based on BERT with the addition of special tokens that mark phrase boundaries, shows positive results.

5. 2. Justification of the phrase similarity benchmark, practical application and evaluation of results

The obtained benchmark comprises a set of 5,513 unique phrases, which form 6,634 negative and 6,723 positive pairs, resulting in 13,357 annotated pairs.

The statistics of the benchmark dataset are shown in Table 1.

Table 1

Statistics of the benchmark dataset

Average number of words per phrase	2.8
Number of unique words	2,329
Percentage of words overlap in positive pairs	9.6 %
Percentage of words overlap in negative pairs	6 %
Average number of words in the context sentence or paragraph	23

Although the benchmark dataset is structured as the pairwise classification problem, the phrase representations are modeled independently and then the cosine similarity distance between them is calculated. For each model, the optimal threshold for binary classification was determined, which allowed to calculate the average accuracy, which serves as the main metric for evaluation. Table 2 below shows the results obtained with the state-of-the-art text representation models, as well as the results achieved with the SkillPhrase2vec method presented in this paper.

Table 2

Accuracy on the Skill STS benchmark achieved by the models which were trained without supervision

Model	Accuracy of the model, trained on general data	Accuracy of the model, trained on HR data
FastText	0.655	0.661
SimCSE	0.694	0.697
McPhraSy	not applicable	0.706
SkillPhrase2Vec unsup	not applicable	0.812

Evaluation of the quality of vector representations, produced by models trained without supervision, on the Skill STS benchmark introduced in this paper, as shown in Table 2, emphasizes the effectiveness of the proposed unsupervised learning approach. In particular, the results underscore the importance of including contextual information in the process of modeling skill phrase representations.

Accuracy achieved on the Skill STS benchmark by models trained with distant supervision is shown in Table 3.

Table 3

Accuracy achieved on the Skill STS benchmark by models trained with distant supervision

Model	Accuracy
Phrase-BERT	0.642
E5 base	0.696
GTE base	0.726
SimLM	0.753
Jina v1	0.796
GPT sentences aug	0.832
SkillPhrase2Vec sup	0.931

As can be seen from Table 3, the presented model demonstrates the best performance, significantly outperforming the accuracy of solutions presented by other researchers. In addition, the results show the benefits of using a trained model that models lexical similarity of texts when learning contextually aware phrase representations using the proposed approach. Indeed, the accuracy of SkillPhrase2Vec sup is 14 % higher than that of the similar unsupervised model (0.931 vs. 0.812).

5. 3. Experimental testing of the technology for analyzing context-aware phrase representations

The proposed technology for analyzing context-aware phrase representations has been tested at Daxtra Technologies. The company's products are aimed at automating and improving the recruitment process for recruitment agencies, corporate HR departments, and job boards.

The testing process included the integration of the developed architecture into existing candidate analysis and search systems, such as Daxtra Parser and Daxtra Search Nexus, to improve the quality of recruitment. A token classifier model was trained to extract key skill phrases from job postings and resumes. The model was trained with 35 thousand annotated sentences and achieved 87 % quality in token classification.

Comparison of the obtained results with the results of the previously applied approach (Jina v1) does not contradict the data in Table 3. Comparison of the obtained results with the results of the expert evaluation (by a specialist) of the extracted and grouped skill phrases revealed 92 % agreement. The expert noted an improvement in grouping in cases where the context of the phrases was important for their normalization.

6. Discussion of the results on the theory and practice of modeling context-aware phrase representations

The experimental table (Table 2) shows that the quality of PhraseBERT model [12], which is specially designed for phrase embeddings learning, is significantly lower than that

of general sentence representation models (E5 base, GTE base, SimLM, Jina v1). Indeed, the accuracy rate (Table 2) for the PhraseBERT model is the lowest compared to the deep neural network models as described above. The data obtained are in line with previous research findings [14].

Comparing results of PhraseBERT to a shallow neural network (Fasttext [22]), it can be seen that PhraseBERT shows worse metrics. This can be attributed to several factors, including the varying pre-training scales of the models (PhraseBERT was trained on synthetically generated paraphrases and phrases with contexts extracted from Book3 corpus), while other general models used diverse large-scale datasets, which include different real-world data like internet QA forums (question-answer pairs), social media (title-short summary), scientific papers, knowledge bases and manually annotated sentence similarity datasets (NLI, SNLI, etc), and as such are more robust to domain shifts. Reliance on frozen representation from the PhraseBERT model is possibly the reason why the McPhraSy model, implemented using the architecture described in the original paper, also shows mediocre performance. This has been taken into account and corrected in the theoretical justification of the SkillPhrase2Vec model.

GPT sentences aug [19], trained on synthetically generated sentences which contain specific skills shows the promise of using Large Language Models in generating realistic data and potential in distilling knowledge from them. However, training on real-world data provides a more diverse and nuanced understanding of language, encompassing the complexity and variability inherent in genuine contexts.

The application of the developed architecture (Fig. 3), due to the ingestion of special phrase boundary markers, avoids the disadvantages of the above models. The presented architecture allows the model to take into account the context during inference by combining textual information from both sides of the phrases. This facilitates efficient modeling of new phrases that were not present in the training data.

All experiments were conducted using models of the BERT-base size. Despite the effectiveness of the proposed solution in this category of models, it is important to investigate the impact of model size and increased training data on the quality of the learned representations. The limitation of the architecture is the need to predefine "key phrases".

Recently promising approaches based on the paradigm of Masked-Auto-Encoders for unsupervised pre-training of sentence embeddings [25], which force the model to condense the meaning of the sentence in the special token's representation, were introduced. Adapting this kind of unsupervised pre-training could be beneficial for phrase in context representation learning. Exploring the extension of this approach to a multilingual setting presents an exciting avenue, as it addresses the growing demand for cross-cultural and diverse applications in the field.

7. Conclusions

1. Architecture for context-aware phrase representation modeling that differs from the classical Transformer-type architecture by the presence of an additional block for determining phrase boundaries has been developed. This architecture, in combination with a new approach to learning with context-aware phrase alignment, leads to a significant improvement in the quality of phrase representations by 9.9–10.6 %, depending on the specific scenario.

2. SkillPhrase2Vec has been shown to work seamlessly with new skills not seen during training. The new semantic similarity benchmark for skills extracted from real HR texts, together with the corresponding contextual paragraphs, allows for reliable comparison of skill phrase representation models.

3. The proposed information analysis technology for context-aware phrase representation modeling was experimentally tested at Daxtra Technologies. 92 % agreement with the expert evaluation conducted by a specialist, which is 15 % higher than the previously used approach (Jina v1) was achieved. The increase in agreement with expert opinion can be attributed to the use of context in modeling phrase representation.

Conflict of interests

The authors declare that they have no conflict of interest in relation to this study, including financial, personal, authorship, or other, that could affect the study and its results presented in this article.

Funding

The study was conducted without financial support.

Data availability

Manuscript has related data in the data warehouse [https://github.com/maiiabocharova/skill_phrase2vec].

Use of artificial intelligence tools

The authors confirm that they did not use artificial intelligence technologies in the creation of the presented work.

Acknowledgments

We thank Daxtra Technologies for providing data and a platform for testing the information technology of context-aware phrase representations.

References

1. Green, T., Maynard, D., Lin, C. (2022). Development of a benchmark corpus to support entity recognition in job descriptions. Proceedings of the Thirteenth Language Resources and Evaluation Conference. Available at: https://aclanthology.org/2022.lrec-1.128/
2. Zhang, M., Jensen, K., Sonniks, S., Plank, B. (2022). SkillSpan: Hard and Soft Skill Extraction from English Job Postings. Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. https://doi.org/10.18653/v1/2022.naacl-main.366
3. O*NET OnLine. Available at: https://www.onetonline.org/

4. European Skills/Competences, Qualifications and Occupations (ESCO). Available at: https://ec.europa.eu/social/main.jsp?catId=1326&langId=en
5. Malakhov, E., Shchelkonogov, D., Mezhuyev, V. (2019). Algorithms of Classification of Mass Problems of Production Subject Domains. Proceedings of the 2019 8th International Conference on Software and Computer Applications, 149–153. https://doi.org/10.1145/3316615.3316676
6. Prykhodko, S., Prykhodko, N. (2022). A Technique for Detecting Software Quality Based on the Confidence and Prediction Intervals of Nonlinear Regression for RFC Metric. 2022 IEEE 17th International Conference on Computer Sciences and Information Technologies (CSIT), 499–502. https://doi.org/10.1109/csit56902.2022.10000532
7. Gotthardt, M., Mezhuyev, V. (2022). Measuring the Success of Recommender Systems: A PLS-SEM Approach. IEEE Access, 10, 30610–30623. https://doi.org/10.1109/access.2022.3159652
8. Про застосування англійської мови в Україні. Документ 3760-IX. Прийняття від 04.06.2024. Available at: https://zakon.rada.gov.ua/laws/show/3760-20#Text
9. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N. et al. (2017). Attention is all you need. arXiv. https://doi.org/10.48550/arXiv.1706.03762
10. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North. https://doi.org/10.18653/v1/n19-1423
11. Reimers, N., Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). https://doi.org/10.18653/v1/d19-1410
12. Wang, S., Thompson, L., Iyyer, M. (2021). Phrase-BERT: Improved Phrase Embeddings from BERT with an Application to Corpus Exploration. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. https://doi.org/10.18653/v1/2021.emnlp-main.846
13. Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., Specia, L. (2017). SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation. Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). https://doi.org/10.18653/v1/s17-2001
14. Cohen, A., Gonen, H., Shapira, O., Levy, R., Goldberg, Y. (2022). McPhraSy: Multi-Context Phrase Similarity and Clustering. Findings of the Association for Computational Linguistics: EMNLP 2022, 3538–3550. https://doi.org/10.18653/v1/2022.findings-emnlp.259
15. Decorte, J.-J., Van Hautte, J., Demeester, T., Develder, C. (2021). JobBERT: Understanding job titles through skill. International workshop on Fair, Effective And Sustainable Talent management using data science (FEAST) as part of ECML-PKDD 2021. arXiv. https://doi.org/10.48550/arXiv.2109.09605
16. Decorte, J.-J., Van Hautte, J., Deleu, J., Develder, C., Demeester, T. (2022). Design of negative sampling strategies for distantly supervised skill extraction. 2nd Workshop on Recommender Systems for Human Resources (RecSys in HR 2022) as part of RecSys 2022. arXiv. https://doi.org/10.48550/arXiv.2209.05987
17. Bhola, A., Halder, K., Prasad, A., Kan, M.-Y. (2020). Retrieving Skills from Job Descriptions: A Language Model Based Extreme Multi-label Classification Framework. Proceedings of the 28th International Conference on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.513
18. Djumalieva, J., Sleeman, C. (2018). An Open and Data-driven Taxonomy of Skills Extracted from Online Job Adverts. Developing Skills in a Changing World of Work, 425–454. https://doi.org/10.5771/9783957103154-425
19. Decorte, J.-J., Verlinden, S., Van Hautte, J., Deleu, J., Develder, C., Demeester, T. (2020). Extreme Multi-Label Skill Extraction Training using Large Language Models. International workshop on AI for Human Resources and Public Employment Services (AI4HR&PES) as part of ECML-PKDD 2023. arXiv. https://doi.org/10.48550/arXiv.2307.10778
20. Günther, M., Mastrapas, G., Wang, B., Xiao, H., Geuter, J. (2023). Jina Embeddings: A Novel Set of High-Performance Sentence Embedding Models. Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023), 8–18. https://doi.org/10.18653/v1/2023.nlposs-1.2
21. Mashtalir, S. V., Nikolenko, O. V. (2023). Data preprocessing and tokenization techniques for technical Ukrainian texts. Applied Aspects of Information Technology, 6 (3), 318–326. https://doi.org/10.15276/aait.06.2023.22
22. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T. (2017). Enriching Word Vectors with Subword Information. Transactions of the Association for Computational Linguistics, 5, 135–146. https://doi.org/10.1162/tacl_a_00051
23. Lee, K., Ippolito, D., Nystrom, A., Zhang, C., Eck, D., Callison-Burch, C., Carlini, N. (2022). Deduplicating Training Data Makes Language Models Better. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). https://doi.org/10.18653/v1/2022.acl-long.577
24. Wang, L., Yang, N., Huang, X., Jiao, B., Yang, L., Jiang, D. et al. (2022). Text embeddings by weakly-supervised contrastive pre-training. arXiv. https://doi.org/10.48550/arXiv.2212.03533
25. Xiao, S., Liu, Z., Shao, Y., Cao, Z. (2022). RetroMAE: Pre-Training Retrieval-oriented Language Models Via Masked Auto-Encoder. Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. https://doi.org/10.18653/v1/2022.emnlp-main.35