Using vector representations for matching tasks to skills

Miriam Amin[1,*], Jan-Peter Bergmann[1] and Yuri Campbell[1]

[1]Fraunhofer Center for International Management and Knowledge Economy (IMW), Neumarkt 9-19, 04109 Leipzig, Germany

Abstract

Science, Technology and Innovation (ST&I) companies as well as large research organizations are repeatedly facing the problem of matching an emerging task with the appropriate skill that is present somewhere in an organizational unit. Many organizations already have skill or competence taxonomies that can be useful in this regard. In this working paper, we present our experiments on automatically recommending suitable skills from the internal skill taxonomy of the Fraunhofer Society research organization to incoming research requests in order to support human decision making processes. We applied three different vector-based approaches for this end, one based on language models, one on word embeddings and one on a simple one-hot-encoding of keywords. Our results show that the language-model-based approach outperforms the other methods and is able to recommend skills to research requests with an MAP of 0.82. These first findings pave the way for further improvements of our method and for the transfer to other related problems.

Keywords

Recommender Systems, Knowledge Management, Skill Taxonomy, Competence Taxonomy, Task-Skill Matching

1. Introduction

Recommender Systems are widely used in Human Ressources, mainly in the processes of hiring and recruiting. A frequent field of application is the matching of suitable job seekers to a job vacancy. The methods applied for this end range from the application of LSTMs [1] over word embeddings [2] to state-of-the art language models. Many authors do not merely rely on pretrained language models like BERT, but fine-tune these models with resume and job vacancy data [3]. Although applications of AI and NLP in the field of Knowledge Management have recently been identified as promising [4], research in this area is just starting to gain traction.

Especially ST&I companies, but also large research organizations, may repeatedly face the problem of matching an emerging task with the matching skill required for completing that task in order to forward it to the appropriate organizational unit. Many organizations already have skill or competence taxonomies that can be useful in this regard. Skill taxonomies (also skill framework, competence taxonomy or competence framework) list and describe the skills that are present or desired in an organization, cluster them in a hierarchical manner and store them in a database.

Nevertheless, due to the often very extensive taxonomies alone, the manual matching of such queries with skills remains laborious. Systems that support experts by recommending a set of highly suitable skills can be very helpful in the human decision making process. We present an approach that automatically matches a research request with the most suitable technological skills and demonstrate the application of our method on real examples from the Fraunhofer Society. To our best knowledge, no comparable system has been presented so far.

The Fraunhofer Society is a German publicly funded research organization, operating 76 research institutes and units that are working in different areas. The more than 20.000 scientific, technical and administrative workers in 2020 cover a very broad spectrum of competences. With contract research as a main source of revenue, the organization regularly receives research requests, which must then be forwarded to the appropriate organizational units. In this working paper, we want to present our experiments on automatically recommending suitable skills from the Fraunhofer internal skill taxonomy to incoming research requests.

In the next section, we describe our datasets, the Fraunhofer skill taxonomy and the corpus of research requests in greater detail. After that, we dicuss the different methods that we used and compared to the end of matching the two datasets with each other. In the subsequent results chapter, we briefly present the results of our experiments. The article concludes with the discussion of our results and the methods and highlights the next steps we want to take with these approaches.

RecSys in HR'22: The 2nd Workshop on Recommender Systems for Human Resources, in conjunction with the 16th ACM Conference on Recommender Systems, September 18–23, 2022, Seattle, USA.

*Corresponding author.

miriam.amin@imw.fraunhofer.de (M. Amin); jan-peter.bergmann@imw.fraunhofer.de (J. Bergmann); yuri.cassio.campbell.borges@imw.fraunhofer.de (Y. Campbell)

0000-0001-6912-4122 (M. Amin); 0000-0001-8918-0551 (J. Bergmann); 0000-0002-4166-2081 (Y. Campbell)

Scientific discipline	Research field	Skill
Simulation, control and operational management of energy supply systems	Energy informatics	AI-based autonomous actions
Concatenated skill string		
Simulation, control and operational management of energy supply systems Energy informatics AI-based autonomous actions		

Table 1
Example of a skill hierarchy and the concatenated skill string.

Research request	We are searching for a solution to link a smart metering system of high-resolution electricity, gas and heat data with our intelligent cloud solution. In the cloud, we want to automatically process the data using machine learning to check for consistency and completeness and to enable load forecasts and cost optimization. We are also looking for the joint development of innovative business models.

Table 2
Example research request

2. Methods

2.1. Data and Preprocessing

The Fraunhofer Society combines a wide variety of specialized institutes under one umbrella. To handle this variety of skills contained in the different institutions, the Fraunhofer Society developed an overview of its already existing competences as well as prospective ones. It is planned that employees will be able to subscribe to the skills and topics that interest them, i.e. skills are not automatically assigned to employees. Based on their individual selections, employees can then receive relevant messages and notifications about incoming research requests. These skills are hierarchically structured in a taxonomy with a tree-like structure with four levels: the root, the first level: scientific disciplines, the second level: their research fields, and finally, the skills are the leaves in this skill-tree.

The entire dataset includes approximately 1.000 skills that are either written in German, English or mixed English and German. Moreover, disciplines and research fields as well have similar language composition in their description. That means, even when a leaf is described in the English language, as Machine Learning, its research field *Künstliche Intelligenz* can be written in German, and vice versa. In order to give more contextual information to single skills, we concatenate skill, research field and scientific discipline to build one textual representation for every skill in this way. These preprocessed skills have an average length of 128 characters. Table 1 shows an example of a skill hierarchy and the concatenated skill string. In this specific case, all levels are in English.

On the other side, research requests are short texts of approximately 1.112 characters in length. Since they come from different authors, they are very diverse both structurally and stylistically. Also, they cover a large variety of research fields and can be German or English, but mainly German. Our experimental corpus of research requests conveys approximately 100 documents. Table 2 shows an example of such a research request.

2.2. Methods

In order to support the matching between research requests and in-house skills in large organizations, we propose a vector-based approach, which draws from recent Transfer Learning advances in Natural Language Processing. Firstly, we represent the skills in the taxonomy with a vector model. Then, with the same vector representation approach, we transform the requests and project them into the same vector space. Finally, every research request acts as a query for which we retrieve matching documents. In this Information Retrieval setting, we return the $n-$closest skill-vectors to a specific query vector as matches for that request.

In this framework, we test three distinct approaches to create useful vector representations for the task at hand. They are Keyword-Binarizer (KB), Keyword-Embedding (KE) and Language Model (LM). In the KB approach, we extract keywords using the keyword extraction algorithm YAKE! [5] from the text description of skills and requests, then a binary vector is constructed in an one-hot-encoding manner for all skills and requests. It is important to note here that YAKE! extracts keywords as well as keyphrases (the combination of two or more words). From now on in the text, we will refer to both as keywords only.

In the KE vector model, the texts undergo the same keyword extraction procedure as in KB. However, the final step for the construction of the vector representation is different. Here, given a skill or a request, we create the corresponding vector representation by averaging the Word2Vec embeddings of the keywords belonging to that skill/request. We use Word2Vec word-embeddings, which were trained by Deepset[1] on the whole German Wikipedia corpus. In cases where the vector representations for a specific word is not found in the embedding dictionary, we apply compound splitting and a vector retrieval is attempted for the resulting components. This procedure is specially useful for German, since many German words have a compositional structure, for ex-

[1]https://www.deepset.ai/german-word-embeddings

ample *Forschungsprojekt* = *Forschung* (research) + *Projekt* (project). Words for which no representation can be found receive a 0−vector, which practically cancels any impact they might have on the average representation.

Finally, in the LM approach, we use a multilingual language model which is fine-tuned on the task of semantic similarity. More precisely, we use the model `paraphrase-multilingual-mpnet-v2`, provided by Sentence-Transformers [2]. This model is suitable for creating vector representations of sentences and paragraphs for information retrieval, clustering or sentence similarity tasks[3]. The model `paraphrase-multilingual-mpnet-v2` is the multilingual version of the original model `all-mpnet-base-v2`. The model `paraphrase-multilingual-mpnet-v2` is trained via multilingual knowledge distillation [6]. In other words, a smaller multilingual model, in this case XLM-RoBERTa [7], is used as the student model, while a bigger MPNET [8] monolingual model is used to guide the multilingual vector representations of translated pairs by means of a double mean squared error loss on the generated representations for the multilingual training pair. The pre-trained monolingual teacher model MPNET was fine-tuned with SBERT-like objective [9] on more than 1 billion pairs of sentences/paragraphs[4]. The pre-training objective of the teacher model is an usual contrastive learning objective. That means, for a given pair of sentences, or paragraphs or sentence-paragraph, the model predicts which, out of a set of randomly constructed pairs with at least one component of the original pair, were actually paired in the billion dataset. In our use case, just the trained student model is used in order to create multilingual vector representations for skills and requests. Both require no further pre-processing steps before as the XLM-RoBERTa model has SentencePiece as its base tokenizer and it was previously pre-trained in many languages, among them English and German as well.

2.3. Validation

In order to validate the three approaches described in the preceding section, we took two different samples of the request corpus, retrieved the top five skill recommendations from each method and assessed the relevance.

For the experiments at hand, we needed to conduct the relevance assessment manually. In the near future, however, a completely expert-labeled ground truth dataset will be at our hand, recording all relevant skills for each request. We labeled a request-skill-pair with the relevance value '2' when the skill was completely relevant

[2]https://www.sbert.net/docs/pretrained_models.html
[3]https://huggingface.co/sentence-transformers/all-mpnet-base-v2
[4]https://huggingface.co/sentence-transformers/all-mpnet-base-v2

Method	Sampling method: Top Similarities		Expert		Mean	
	NDCG	MAP	NDCG	MAP	NDCG	MAP
LM	**0.70**	**0.89**	**0.63**	**0.76**	**0.67**	**0.82**
KE	0.25	0.37	0.13	0.16	0.19	0.27
KB	0.28	0.39	0.15	0.28	0.21	0.33

Table 3
NDCG@5 and MAP values for the three vectorization methods and the two sampling methods. LM - Language Model, KE - Keyword-Embeddings, KB - Keyword-Binarizer

for the request, '1' when it was not completely relevant, but also not irrelevant and '0' when it was completely irrelevant.

We took two samples of ten requests, each with a different sampling method. In the sampling method 'expert', we selected ten requests in which the authors of this paper themselves have expert knowledge of the required skills - resulting in ten IT and AI related requests. For the sampling method 'top similarities', we considered the top five skills with the highest similarity scores for each request. We then took the mean of these top five similarity scores. For each vectorization method, we then selected the top ten request with the highest mean similarities. Note that the 'top similarities' sample sets differ among the methods. In addition, we calculated the mean value from the 'expert' and the 'top similarities' sample.

With 20 relevance assessments for each method, we were able to calculate the Normalized Discounted Cumulative Gain@5 (NDCG@5) and the Mean Average Precision (MAP) for each system. In order to calculate these measures despite the missing ground truth, we assumed that there are five matching skills for each request. In order to calculate the MAP, which requires a binary relevance, we considered the relevance labels '1' and '2' as relevant and '0' as irrelevant.

3. Results

The purpose of our experiment was to find out which NLP method yields the best results for the task of recommending skills from a standardised skill ontology to a specific task or request. Table 3 shows an overview of the NDCG@5 and the MAP scores obtained during our experiments.

From the data, it is apparent that the language model-based method yielded by far the best results. Over all samples, the language model achieved an impressive MAP of 0.82 and and NDCG of 0.67. The other two methods are far behind.

To illustrate the findings of these first experiments, we show the top five skill recommendations of each method

Before prompt engineering	Simulation, control and operational management of energy supply systems Field of competence energy informatics AI-based autonomous actions
After prompt engineering	We work on AI-based autonomous actions. Our field of competence is energy informatics, within the research field of simulation, control and operational management of energy supply systems

Table 4
An example of the prompt engineering of a skill string

for one request in table 5.

4. Discussion

The results of these preliminary experiments are very satisfactory. We have shown that our language-model-based method in particular performed very well for matching skills to specific tasks. That was somewhat surprising against the background that the skills have a comparatively short text length and thus do not provide much context for the language model to compute semantic similarities. Equally surprising was that the word-embedding-based method (KE), which were supposed to perform well even without context, showed such poor performance. We suspect that this is due to the rather technical vocabulary in both the skills and the requests that is not present in our word embedding vocabulary. Our attempt to counteract this with the compound splitting described above does not seem to have achieved the expected results.

Nevertheless, we are convinced that the performance - particularly that of the LM approach - can still be improved by further tuning. In future work, we want to experiment with further text preprocessing and prompt engineering methods. For example, we are interested how the transformation of the skill string into a real-language sentence impacts the performance. For this, a sentence template with slots for the hierarchical elements of the skill string can be used. Table 4 shows an example of such a transformed string. With such a transformation, we hope to provide even more context to the Language Model, especially to the attention mechanism. Moreover, LMs are trained and optimized on whole natural sentences, not on syntaxless word groups.

Again, we should address that the sample size of this experiment is still rather small and results need to be confirmed as soon as the entire dataset of research requests was labelled with the matching skills.

We also hope to make further improvements to our approach with such a ground truth at hand. Not only would this allow us to calculate more evaluation measures, such as precision@k and F1@k, we could also fine-tune the vector-space model. With contrastive learning, we could optimize the vector space in a way that requests move closer to the matching skills and further away from the mismatching skills, hoping that this new vector space is transferable to unknown requests.

Last, and maybe most importantly, we want to explore the transferability of our method to other, related problems. These are, e.g., recommending skills for more general tasks and work assignments or even finding the worker or team with the optimal skill set for requests, tasks and work assignments.

However, it remains very important to mention that such recommender systems are only useful and properly utilized when they are designed to support an essentially human-driven decision-making process.

References

[1] C. Qin, H. Zhu, T. Xu, C. Zhu, L. Jiang, E. Chen, H. Xiong, Enhancing person-job fit for talent recruitment, in: K. Collins-Thompson (Ed.), The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, ACM Conferences, ACM, New York, NY, 2018, pp. 25–34. doi:10.1145/3209978.3210025.

[2] J. Zhao, J. Wang, M. Sigdel, B. Zhang, P. Hoang, M. Liu, M. Korayem, Embedding-based recommender system for job to candidate matching on scale, 2021. URL: https://arxiv.org/pdf/2107.00221.

[3] D. Lavi, V. Medentsiy, D. Graus, consultantbert: Fine-tuned siamese sentence-bert for matching jobs and job seekers, in: The Workshop on Recommender Systems for Human Resources (RecSys in HR 2021), 2021.

[4] M. H. Jarrahi, D. Askay, A. Eshraghi, P. Smith, Artificial intelligence and knowledge management: A partnership between human and ai, Business Horizons (2022). doi:10.1016/j.bushor.2022.03.002.

[5] R. Campos, V. Mangaravite, A. Pasquali, A. Jorge, C. Nunes, A. Jatowt, Yake! keyword extraction from single documents using multiple local features, Information Sciences 509 (2020) 257–289. doi:10.1016/j.ins.2019.09.013.

[6] N. Reimers, I. Gurevych, Making monolingual sentence embeddings multilingual using knowledge distillation, 2020. URL: https://arxiv.org/abs/2004.09813. doi:10.48550/ARXIV.2004.09813.

[7] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, 2019. URL: https:

Research Request	We are searching for a solution to link a smart metering system of high-resolution electricity, gas and heat data with our intelligent cloud solution. In the cloud, we want to automatically process the data using machine learning to check for consistency and completeness and to enable load forecasts and cost optimization. We are also looking for the joint development of innovative business models.			
	Rank	**Research field**	**Skill**	**Assessment Label**
Language Model	1	Energy Information Technology	Data Science, Statistics, Time Series Analyses, AI/ML	2
	2	Energy Information Technology	Data Management	2
	3	Economic and regulatory assessment	Energy system analyses	2
	4	Energy Information Technology	AI-based methods of optimized, predictive network operation management	1
	5	Energy Information Technology	Standards and interfaces for interoperable communication	2
Keyword-Embedding	1	Storage & storage systems	Integration of new storage systems	0
	2	Lightweight construction technologies	Functional integration in lightweight construction	0
	3	Power grids	Modeling of power grids	0
	4	Artificial Intelligence Methods	Generation of Synthetic Training Data	0
	5	Artificial Intelligence Methods	AI Technologies in Production & Logistics	1
Keyword-Binarizer	1	Module manufacturing/ integration	Packaging for RF and analog mixed-signal modules	0
	2	Process Technologies	Epitaxy	0
	3	Component manufacturing	High- and ultra-high-frequency components (High-Frequency Devices)	0
	4	Component manufacturing	Actuators, MEMS actuators	0
	5	Component packaging, module manufacturing/ integration	Display, RFID packaging	0

Table 5
Top 5 recommendations of all three methods for one example research request

//arxiv.org/abs/1911.02116. doi:10.48550/ARXIV.1911.02116.

[8] K. Song, X. Tan, T. Qin, J. Lu, T.-Y. Liu, Mpnet: Masked and permuted pre-training for language understanding, 2020. URL: https://arxiv.org/abs/2004.09297. doi:10.48550/ARXIV.2004.09297.

[9] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, 2019. URL: https://arxiv.org/abs/1908.10084. doi:10.48550/ARXIV.1908.10084.