INVESTIGATING NATURAL LANGUAGE PROCESSING TECHNIQUES FOR A RECOMMENDATION SYSTEM TO SUPPORT EMPLOYERS, JOB SEEKERS AND EDUCATIONAL INSTITUTIONS

Koen Bothmer, Tim Schlippe

1. Overview

Goal of this work
- **Who:** Support employers, job seekers, educational institutions
- **What:** Close gap between skills required in the job market, skills of job seekers, and skills taught in education (Palmer, 2017)
- **Why:** In line with UN sustainable development goal 4: “Ensure inclusive and equitable quality education and promote lifelong learning opportunities for all” (UN 2021)

2. Related Work

Usually, employers, job seekers, and educational institutions use AI systems in isolation of each other:

1. **Employers** want to automatically rank CVs
   - by semantic matching of skills from LinkedIn profiles and skills from job description, using a taxonomy of skills (Folalioglu et al, 2014)
   - (Fernández-Reyes & Shinde, 2019) use word embeddings to match CVs to job postings
   - (Wang, Alouache & Joubert 2021) combine a knowledge graph and BERT to rank CVs
2. **Job seekers** want to know how they fit to job postings,
   - Job recommendation systems have been researched by (Singh et al., 2012), (Aotaibii, 2012), (Hong et al, 2013), etc.
3. **Educational institutions** want to advise potential students and fit curricula to the job market’s demands.
   - (Deepani et al, 2021) give a systematic review of recent publications on course recommendation and report a growing popularity of data mining techniques

→ Our recommendation system
   - supports all: employers, job seekers, and educational institutions
   - does not need a taxonomy of skills as it uses an unsupervised learning approach

3. NLP Pipeline to Extract, Vectorize, Cluster and Compare Skills which processes skills from job postings, learning curricula, and CVs

**Extracting skill sets**
- We collected 2.6k job postings for the job title: Data Scientist
- Employers tend to put skills in bullet points.
- We extracted 21.5k bullet points likely to be skills.
- We deal with outliers in our NLP pipeline later.

**Pre-processing skill sets**
- We evaluated various vectorization methods.
- Based on practical experiments and the Silhouette score of the final pipeline, we selected Sentence-BERT (Reimers & Gurevych, 2020) with the pre-trained model ‘all_distilroberta_v1’
  You have scripting experience with Python and on R and SQL
  At least 2 years of relevant experience costing in Python and SQL
  You are skilled in the communication with stakeholders

Inspired by (Alammar, 2019)

**Removing outliers**
- Not all bullet points are skills.
- We combined UMAP (McInnes et al., 2020) and DBSCAN (Ester et al., 1996) to detect and remove outliers.
- Our 21.5k bullet points reduced to 18.8k skills.

**Clustering skill sets**
- We evaluated several clustering methods based on literature and practical experiments.
- We selected K-means (MacQueen et al., 1967) and determined the optimal number of 31 clusters by Silhouette score.

**Application: Skill Scanner**
- Our pipeline was trained on 18.8k skills and manually evaluated on 100 unseen skills from job postings, learning curricula, and CVs.
- Skill Scanner’s accuracy to assign unseen skills to the correct cluster is 83%.
- With the clustering approach, Skill Scanner is able to deal with synonyms and different abstraction levels and thus produce comparable recommendations.

4. Conclusion and Discussion

**Conclusion**
- The job market dictates what job seekers should learn and educational institutions should teach.
- Our system processes skills in job postings, CVs, and curricula.
- It outputs recommendations for employers, job seekers, and educational institutions.
- Based on present and missing skills and their importance to employers.

**Follow-up**
- We conducted a user study to collect feedback from potential users (Bothmer & Schlippe, 2022)
- who generally agreed on Skill Scanner’s potential to carry out processes faster, effectively, autonomous, explainable, and in a more supported manner.

**Future work**
- Apply our pipeline to other job positions
- Use fine-tuned Sentence-BERT instead of ‘all_distilroberta_v1’

**Application: Skill Scanner**
- Our pipeline was trained on 18.8k skills and manually evaluated on 100 unseen skills from job postings, learning curricula, and CVs.
- Skill Scanner’s accuracy to assign unseen skills to the correct cluster is 83%.
- With the clustering approach, Skill Scanner is able to deal with synonyms and different abstraction levels and thus produce comparable recommendations.