NEO: A System for Identifying New Emerging Occupation from Job Ads

Anna Giabelli,[1,2] **Lorenzo Malandri,**[1,2] **Fabio Mercorio,**[1,2] **Mario Mezzanzanica,**[1] **Andrea Seveso**[2,3]
[1] Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Milan, Italy
[2] CRISP Research Centre, University of Milano-Bicocca, Milan, Italy
[3] Department of Informatics, Systems and Communication, University of Milano-Bicocca, Milan, Italy
{anna.giabelli, lorenzo.malandri, fabio.mercorio, mario.mezzanzanica, andrea.seveso}@unimib.it

Abstract

We demonstrate NEO, a tool for automatically enriching the European Occupation and Skill Taxonomy (ESCO) with terms that represents new occupations extracted from million Online Job Advertisements (OJAs). NEO proposes (i) a novel metric that allows one to measure the semantic similarity between words in a taxonomy, and (ii) a set of measures that estimate the adherence of new terms to the most suited taxonomic concept, enabling the user to evaluate the suggestions. To test its effectiveness, NEO has been evaluated over 2M+ 2018 UK job ads, along with a user-study to confirm the usefulness of NEO in the taxonomy enrichment task.

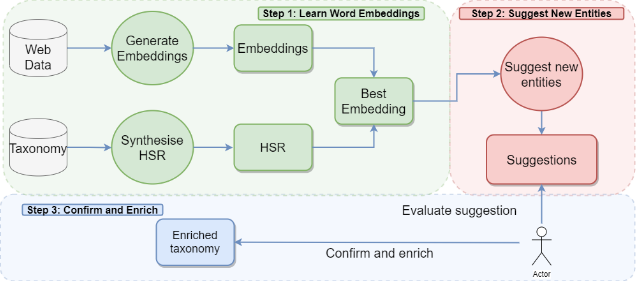


Figure 1: A representation of the NEO workflow highlighting the main modules. Taken from (Giabelli et al. 2020b)

Introduction

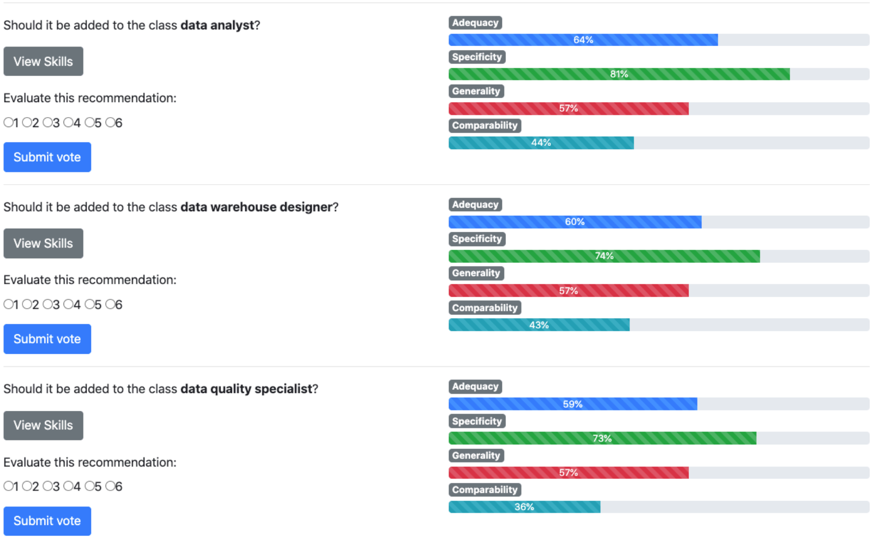
Unlike the automated construction of new taxonomies from scratch, which is a well-established research area (Wang, He, and Zhou 2017), the augmentation of existing hierarchies is gaining in importance, given its relevance in many practical scenarios (see, e.g. (Vedula et al. 2018; Malandri et al. 2021)). To date, the most adopted approach to enrich or extend standard *de-jure* taxonomies - that cannot be constructed from scratch - lean on expert panels, that identify and validate which term has to be added to a taxonomy. This process totally relies only on human knowledge, making it costly, time-consuming and error prone, besides suffering from sparse coverage. Those challenges need the development of automated methods for taxonomy enrichment.

Overview of NEO

NEO (Giabelli et al. 2020b) aims at enriching the European standard labour market taxonomy ESCO (European Commission 2019) with new potential occupations derived from real Online Job Advertisements (OJAs). It is developed as part of the research activity of an ongoing EU grant aimed at realising the first EU real-time labour market monitoring system, by collecting and classifying OJAs over all 27+1 EU countries and 32 languages (CEDEFOP 2016; Boselli et al. 2018a)[1]. *Novel occupations* are usually intended as *mentions* that deserve to be represented within the taxonomy, as they might represent either an emerging job (e.g., *SCRUM master*) or a new alternative label characterising an existing job (e.g., *Android developer*). This activity is crucial to allow economists and policy makers to observe up-to-date labour market dynamics using standard taxonomies as a *lingua franca*, overcoming linguistic boundaries (see, e.g. (Frey and Osborne 2017; Giabelli et al. 2020a; Colombo, Mercorio, and Mezzanzanica 2019)). NEO relies on distributional semantics to extract semantic information from the OJVs, exploiting the characteristics that words occurring in similar context tend to have a similar meaning. Our approach is composed of three steps: i) *synthesise word embeddings* ii) *suggest new entities* iii) *vote and enrich* (Fig. 1).

(Step 1) Synthesise Word Embeddings resorts to Deep Learning to learn a vector representation of words in the corpus, preserving the semantic relationships expressed by the taxonomy itself. To select the best representing vectors, we rely on three distinct sub-tasks, that are the following: **T1.1**: train three different word embedding models (Word2Vec, GloVe, FastText), **T1.2**. construct measure of pairwise semantic similarity between taxonomic elements, namely Hierarchical Semantic Relatedness (*HSR*) (Giabelli et al. 2020b). Compared with *HSR*, state-of-the-art metrics for semantic similarity (see Aouicha, Taieb, and Hamadou (2016) for a survey) suffer of two main drawbacks. First, when a word has multiple senses, those methods compute a value of similarity for each word sense and then consider only the highest, which is the self-information of the least frequent lowest common ancestor. As a consequence, more specific senses will have a higher value of similarity, but this

[1] https://tinyurl.com/skillovate


Should it be added to the class **data analyst**?

Evaluate this recommendation:

○1 ○2 ○3 ○4 ○5 ○6

Should it be added to the class **data warehouse designer**?

Evaluate this recommendation:

○1 ○2 ○3 ○4 ○5 ○6

Should it be added to the class **data quality specialist**?

Evaluate this recommendation:

○1 ○2 ○3 ○4 ○5 ○6

Figure 2: Evaluation of the ESCO concept candidate where the mention *business intelligence analyst* should be added

does not reflect the use of words in advertising job positions; second, though they consider the structure of the taxonomy (i.e., the relationship between concepts) they do not take into account the number of child entities (i.e., words) belonging to those concepts. This is crucial in our case as ESCO includes generic concepts that, in turn, contain many different occupations. On the contrary, some very specific concepts can be represented by a few occupations which are highly informative. The aim of *HSR* is to overcome these limitations to work with the ESCO taxonomy. **T1.3**. Evaluate the embeddings in terms of correlation between the *HSR* and the cosine similarity between pair of terms in the taxonomy.

(Step 2) Suggest New Entities is aimed at extracting new occupation terms from the corpus of OJAs, and to suggest the most suitable concepts under which they could be added in the taxonomy $\mathcal{T}$. First we select a starting word w_0 from $\mathcal{T}$. Then we consider the top-5 mentions in the corpus of documents $\mathcal{D}$ with associated the highest *score* value $\mathcal{S}$ with w_0, where is $\mathcal{S}$ a function that quantifies how similar a new mention m is in relation to w_0. The most suitable concepts for m are identified on the basis of four measures, namely GASC (*G̲enerality*, *A̲dequacy* *S̲pecificity*, and *C̲omparability*, formally defined in Giabelli et al. (2020b)), that estimate the fitness of a concept c for a given m. *Generality* quantifies how similar a mention is to all the other unrelated concepts; *Adequacy* quantifies how much the mention is overall a good candidate as an entity of the concept, considering both Specificity and Generality; *Specificity* quantifies how similar a mention is to the related concept; *Comparability* quantifies how similar two occupations are in term of the number of skills shared between them.

(Step 3) Vote and Enrich allows validating the outcome of the previous steps - which is fully automated - by asking: (*Q1*) whether the mentions extracted from the corpus are valid emerging occupations and (*Q2*) to what extent the concepts suggested as entry for a new mention are appropriate for it (by looking at skills-gap).

Experimental Results on 2M+ UK Job Ads

Experimental settings. The corpus, a subset of the data collected for the project (CEDEFOP 2016), contains 2,119,025 OJAs published in the United Kingdom in 2018. The best performing model is the following: architecture=*fastText*, algorithm=CBOW, size=300, epochs=100, learning rate=0.1.

As a first step, the user selects the starting word w_0 among the occupations already in ESCO. Then, NEO prompts the 5 mentions with associated the highest *score* with w_0. The user can therefore select a new mention m (*business intelligence analyst*)[2] to evaluate to which extent it fits as an entity of the starting word's ESCO concept, and as an entity of other two ESCO concepts selected (i.e., *data warehouse designer*. For each one of these three pairs mention $< m, concept >$ NEO provides the GASC measures (Fig. 2), along with a comparison of the *rca* (Alabdulkareem et al. 2018) of skills for both the mention and the concept (Fig. 3). These skills, together with the GASC measures, support the user in evaluating if the suggested entry is appropriate as an entity of a concept. A user evaluation involving 10 final users reveladed that 88% of the mentions (43 out of 49) were successfully evaluated to be new occupations, while the correlation (Spearman's ρ and Kendall's τ) between the Likert values and GASC values is positive and statistically significant.

Conclusion

We demonstrated NEO, a system developed within the research activities of an ongoing EU tender for monitoring EU labour market (see, (Boselli et al. 2018a,b)).

NEO allows identifying potential new occupations as they emerge from job ads through deep learning, suggesting the suitable concept to enrich the European standard occupation taxonomy by means of (i) a novel semantic similarity metric

[2] Remember those occupations are not included in the ESCO taxonomy though requested by the market

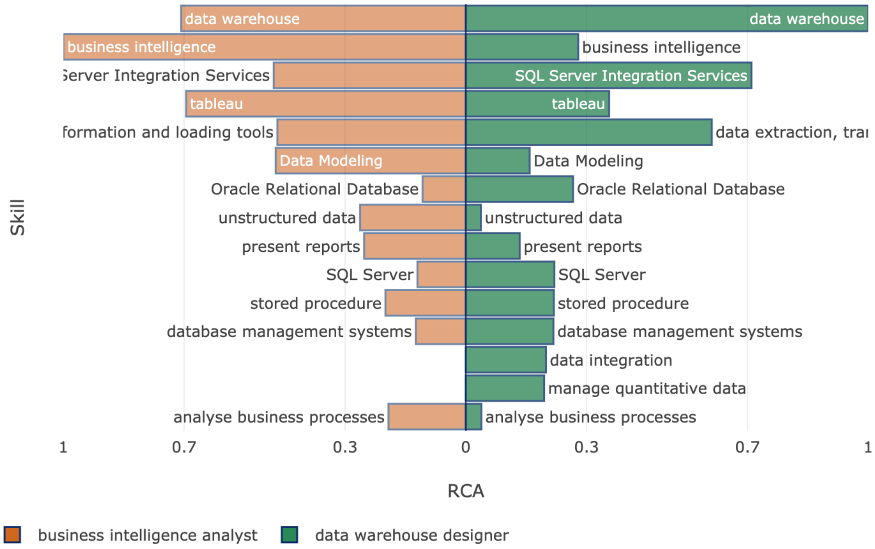


Figure 3: Skill relevance between the new term *business intelligence analyst* and the parent ESCO concept suggested *data warehouse designer*

and (ii) a set of measures that estimate the adherence of new terms to the concept. Though NEO can be used with any OJA dataset and EU language, here it has been trained on a 2M+ 2018 UK vacancies identified 49 novel occupations, 43 of which were validated as novel occupations by a panel of 10 experts as final users. Two statistical hypothesis tests confirmed the correlation between the proposed GASC metrics of NEO and the user judgements.
A demo is provided at https://tinyurl.com/NEO-aaai2021.

References

Alabdulkareem, A.; Frank, M. R.; Sun, L.; AlShebli, B.; Hidalgo, C.; and Rahwan, I. 2018. Unpacking the polarization of workplace skills. *Science advances* 4(7): eaao6030.

Aouicha, M. B.; Taieb, M. A. H.; and Hamadou, A. B. 2016. Taxonomy-based information content and wordnet-wiktionary-wikipedia glosses for semantic relatedness. *Applied Intelligence* 45(2): 475–511.

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

Boselli, R.; Cesarini, M.; Mercorio, F.; and Mezzanzanica, M. 2018b. Classifying online Job Advertisements through Machine Learning. *Future Generation Computer Systems* 86: 319–328.

CEDEFOP. 2016. Real-time Labour Market information on Skill Requirements: Setting up the EU system for online vacancy analysis, available at https://goo.gl/5FZS3E. Last accessed 19/11/2020.

Colombo, E.; Mercorio, F.; and Mezzanzanica, M. 2019. AI meets labor market: Exploring the link between automation and skills. *Information Economics and Policy* 47: 27–37.

European Commission. 2019. ESCO: European Skills, Competences, Qualifications and Occupations, available at https://ec.europa.eu/esco/portal/browse. Last accessed 19/11/2020.

Frey, C. B.; and Osborne, M. A. 2017. The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting and Social Change* 114: 254–280. ISSN 0040-1625.

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

Giabelli, A.; Malandri, L.; Mercorio, F.; Mezzanzanica, M.; and Seveso, A. 2020b. NEO: A Tool for Taxonomy Enrichment with New Emerging Occupations. In *International Semantic Web Conference*, 568–584. Springer.

Malandri, L.; Mercorio, F.; Mezzanzanica, M.; and Nobani, N. 2021. MEET-LM: A method for embeddings evaluation for taxonomic data in the labour market. *Computers in Industry* 124: 103341.

Vedula, N.; Nicholson, P. K.; Ajwani, D.; Dutta, S.; Sala, A.; and Parthasarathy, S. 2018. Enriching taxonomies with functional domain knowledge. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 745–754.

Wang, C.; He, X.; and Zhou, A. 2017. A short survey on taxonomy learning from text corpora: Issues, resources and recent advances. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 1190–1203.