NEO: A Tool for Taxonomy Enrichment with New Emerging Occupations

Anna Giabelli[1,3], Lorenzo Malandri[1,3(✉)], Fabio Mercorio[1,3(✉)], Mario Mezzanzanica[1,3], and Andrea Seveso[2,3]

[1] Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Milan, Italy
[2] Department of Informatics, Systems and Communication, University of Milano-Bicocca, Milan, Italy
[3] CRISP Research Centre, University of Milano-Bicocca, Milan, Italy
{anna.giabelli,lorenzo.malandri,fabio.mercorio, mario.mezzanzanica,andrea.seveso}@unimib.it

Abstract. Taxonomies provide a structured representation of semantic relations between lexical terms, acting as the backbone of many applications. This is the case of the online labour market, as the growing use of Online Job Vacancies (OJVs) enables the understanding of how the demand for new professions and skills changes in near-real-time. Therefore, OJVs represent a rich source of information to reshape and keep labour market taxonomies updated to fit the market expectations better. However, manually updating taxonomies is time-consuming and error-prone. This inspired NEO, a Web-based tool for automatically enriching the standard occupation and skill taxonomy (ESCO) with new occupation terms extracted from OJVs. NEO - which can be applied to any domain - is framed within the research activity of an EU grant collecting and classifying OJVs over all 27+1 EU Countries.

As a contribution, NEO (i) proposes a metric that allows one to measure the pairwise semantic similarity between words in a taxonomy; (ii) suggests new emerging occupations from OJVs along with the most similar concept within the taxonomy, by employing word-embedding algorithms; (iii) proposes GASC measures (Generality, Adequacy, Specificity, Comparability) to estimate the adherence of the new occupations to the most suited taxonomic concept, enabling the user to approve the suggestion and to inspect the skill-gap. Our experiments on 2M+ real OJVs collected in the UK in 2018, sustained by a user-study, confirm the usefulness of NEO for supporting the taxonomy enrichment task with emerging jobs. A demo of a deployed instance of NEO is also provided.

The research leading to these results is partially supported within the EU Project AO/DSL/VKVET-GRUSSO/Real–time LMI 2/009/16 granted by the EU Cedefop Agency, in which some authors are involved as PI and researchers. *All authors equally contributed to this work.*

J. Z. Pan et al. (Eds.): ISWC 2020, LNCS 12507, pp. 568–584, 2020.
https://doi.org/10.1007/978-3-030-62466-8_35

1 Introduction and Motivation

Over the past several years, the growth of web services has been making available a massive amount of structured and semi-structured data in different domains. An example is the web labour market, with a huge number of Online Job Vacancies (OJVs)[1] available through web portals and online applications. The problem of processing and extracting insights from OJVs is gaining researchers' interest in the recent years, as it allows modelling and understanding complex labour market phenomena (see, e.g. [8,9,11,16,20,33,34]). At the same time, the ability to extract valuable knowledge from these resources strongly depends on the existence of an *up-to-date* target taxonomy. Those resources are essential for machine understanding and many tasks in natural language processing. In the European labour market, the key resource is ESCO.[2] Organisations and governments are making a great effort in keeping ESCO up-to-date with the labour market through expert knowledge. This challenging task needs an automated, scalable method capable of enriching ESCO with new terms from the job market.

Unlike the automated construction of new taxonomies from scratch, which is a well-established research area [31], the augmentation of existing hierarchies is gaining in importance, given its relevance in many practical scenarios (see, e.g. [30]). Human languages are evolving, and online contents are constantly growing. As a consequence, people often need to enrich existing taxonomies with new concepts and items, without repeating their whole construction process every time. To date, the most adopted approach to enrich or extend standard *de-jure* taxonomies lean on expert panels and surveys, that identify and validate which term has to be added to a taxonomy. The approach totally relies on human knowledge, with no support from the AI-side, and this makes the process costly, time-consuming, and unable to consider the peculiarities of country-specific labour markets. To extract semantic information from the OJVs, we resort to *distributional semantics*, a branch of linguistics based on the hypothesis that words occurring in similar context tend to have similar meaning [17]. Words are represented by semantic vectors, which are usually derived from a large corpus using co-occurrence statistics or neural network training, and their use improves learning algorithms in many NLP tasks. Semantic word vectors have empirically shown to preserve linguistic regularities [22], demonstrating their ability to enrich existing knowledge structures as well [12,25].

Contribution. In this paper we design and develop `NEO`, a novel system for enriching the ESCO taxonomy with mentions as they appear in the real labour market demand. A *novel occupation*, indeed, is a term that deserves to be represented within the taxonomy, that might represent either an emerging job (e.g.,

[1] An Online Job Vacancy (OJV, *aka*, job offers, job ads) is a document containing a *title* - that shortly summarises the job position - and a *full description*, usually used to advertise the skills a candidate should hold.

[2] European Commission: ESCO: European skills, competences, qualifications and occupations, available at https://ec.europa.eu/esco/portal/browse(2019).

SCRUM master) or a new alternative label characterising an existing job (e.g., *Android developer*). The novelty of NEO goes toward three directions:

- Define a domain-independent metric, i.e., the *Hierarchical Semantic Relatedness* (HSR) to measure the pairwise semantic similarity between words in a taxonomy;
- Synthesise and evaluate, with the supervision of the HSR, vector-space models for encoding the lexicon of both the taxonomy and the OJVs, in order to extract from the latter potential new emerging occupations, and define a set of measures, namely GASC (Generality, Adequacy, Specificity, Comparability) to estimate their suitability as entities of different taxonomic concepts;
- Provide to final users a Web-based tool to vote suggested mentions, supporting the experts in the taxonomy extension activity, and explaining the rationale behind each suggested new occupation through a skill-gap analysis.

The project - in which NEO is framed within - aims at realising a European system to collect and classify OJVs for the whole EU, including 27+1 EU country members and all the 32 languages of the Union [10] through machine learning. The use of classified OJVs and skills, in turn, enables a number of third-party research studies to understand complex labour market phenomena. Just to give a few examples, in [11] authors used OJVs to estimate the impact of AI in job automation and to measure the impact of digital/soft skills within occupations, validating theoretical results from [15]; in reaction to the COVID-19 emergency, the EU Cedefop Agency has been using those OJVs to build the *Cov19R* index for identifying workers with a higher risk of COVID exposure, who need greater social distancing, affecting their current and future job performance capacity.[3]

Though the ESCO taxonomy is a standard, it encodes neither the peculiarities nor the characteristics of countries' labour markets, that vary in terms of skills requested, the lexicon used for advertising similar jobs, and country-related economic conjunctures. This, as a consequence, sharply limits the usability of ESCO as a system for understanding and representing different labour markets.

The Relevance of NEO for the Labour Market. The following example should help in clarifying the matter. Figure 1 shows the structure of the ESCO occupation pillar: it is built on top of the ISCO taxonomy down to the fourth digit (i.e., 436 occupation codes). Then, it continues with 2,942 ESCO occupation concepts and up to 30,072 alternative labels as well. *How to maintain the taxonomy up-to-date to adhere to labour market lexicon? How to enrich the taxonomy with those mentions representing novel occupations? How to estimate similarities between occupations?* Those are just few questions with which economists and policymakers have to deal. The inspiring idea of NEO is to build a Web-based tool for supporting labour market specialists in enriching the taxonomy to better fit the labour market demand of new occupations.

The remainder of the paper is organised as follows. In Sect. 2 we survey related works and formalise the taxonomy enrichment problem and solution. In

[3] https://tinyurl.com/covid-r.

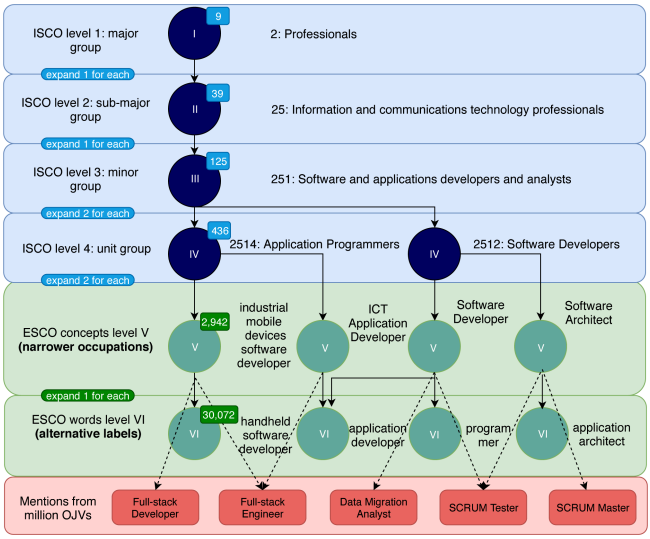


Fig. 1. *Motivating example.* Representation of the ESCO taxonomy, with new mentions representing novel jobs not yet included in ESCO as they emerge from the UK Web Labour Market Demand (2M+ vacancies processed in 2018).

Sect. 3 we describe `NEO` in three conceptual steps: (i) learn word embeddings while preserving taxonomic relations, (ii) suggest new entities for ESCO and (iii) evaluate the fitting of the new entities through GASC measures and by expert evaluation. Those steps are implemented in a real case scenario presented in Sect. 4. Finally, Sect. 5 concludes the paper and draws future work; a demo of `NEO` is also provided.

2 Related Work and Preliminaries

In the past literature, despite the automatic construction of taxonomies from scratch has received a considerable attention [31], the same cannot be said for augmentation of existing hierarchies. Most of the work in the area of automated taxonomy enrichment relies heavily on or domain specific knowledge [5,14] or lexical structures specific to an existing resource, like the WordNet synset [18,26,29] or Wikipedia categories [28]. In recent years few scholars tried to overcome those limitations developing methodologies for the automated enrichment of generic taxonomies. Wang et al. [32] use a hierarchical Dirichlet model to complete a hierarchy with missing categories. Then they classify elements of a corpus with the supervision of the complete taxonomy. For our purposes, this work has two shortcomings: the authors i) modify the structure of the taxonomy, while we want to preserve the ESCO framework and ii) do not update the hierarchical categories with new entities, which is the main goal of our tool. Other scholars [3,13] exploit semantic patterns between hypernyms and hyponyms in word

vector spaces. However, a primary challenge with those methods in semantic hierarchies learning is that the distributional similarity is a symmetric measure, while the hypernomy-hyponymy relation is asymmetric. For this reason, in this research we will focus on symmetric measures, like taxonomic similarity and *rca*.

Other researchers learn term embeddings of the taxonomic concepts and connect new concepts to the most similar existing concepts in the taxonomy. Vedula et al. [30] use word embeddings to find semantically similar concepts in the taxonomy. Then they use semantic and graph features, some of them coming from external sources, to find the potential parent-child relationship between existing concepts and new concepts from Wikipedia categories. Aly et al. [2] use the similarity between Poincaré term embeddings to find *orphans* (disconnected nodes) and *outliers* (child assigned to wrong parents) in a taxonomy. Finally, in [27] authors use a set of <query, anchor> concepts from an existing hierarchy to train a model to predict the parent-child relationship between the anchor and the query. Those methods learn a word vector representation of the taxonomy, without linking it to an external corpus of web data, while we incorporate taxonomic information into a word vector representation of an external text corpus. This allows drawing a semantic relation between a taxonomic concept and a mention.

2.1 Setting the Stage

In this section we introduce a formal definition of taxonomy and we formulate the problem of taxonomy enrichment, relying on the formalisation proposed by [19].

Definition 1 (Taxonomy). *A taxonomy* $\mathcal{T}$ *is a 4-tuple* $\mathcal{T} = (\mathcal{C}, \mathcal{W}, \mathcal{H}^c, \mathcal{F})$. $\mathcal{C}$ *is a set of concepts* $c \in \mathcal{C}$ *(aka, nodes) and* $\mathcal{W}$ *is a set of words (or entities) belonging to the domain of interest; each word* $w \in \mathcal{W}$ *can be assigned to none, one or multiple concepts* $c \in \mathcal{C}$. $\mathcal{H}^c$ *is a directed taxonomic binary relation existing between concepts, that is* $\mathcal{H}^c \subseteq \{(c_i, c_j) | (c_i, c_j) \in \mathcal{C}^2) \wedge i \neq j\}$. $\mathcal{H}^c(c_1, c_2)$ *means that* c_1 *is a sub-concept of* c_2. *The relation* $\mathcal{H}^c(c_1, c_2)$ *is also known as* $IS-A$ *relation (i.e.,* c_1 $IS-A$ *sub-concept of* c_2*).* $\mathcal{F}$ *is a directed binary relation mapping words into concepts, i.e.* $\mathcal{F} \subseteq \{(c, w) | c \in \mathcal{C} \wedge w \in \mathcal{W}\}$. $\mathcal{F}(c, w)$ *means that the word* w *is an entity of the concept* c. *Notice* $\mathcal{T}$ *might be represented as a DAG.*

Given an existing taxonomy $\mathcal{T}$, the goal of NEO is to expand $\mathcal{T}$ with new *mentions* (entities) coming from a text corpus. Each mention is assigned to one or multiple candidate destination nodes of $\mathcal{T}$, along with a score value and a set of measures. More formally we have the following.

Definition 2 (Taxonomy Enrichment Problem (TEP)). *Let* $\mathcal{T}$ *be a taxonomy as in Definition 1, and let* $\mathcal{D}$ *be a corpus; a Taxonomy Enrichment Problem (TEP) is a 3-tuple* $(\mathcal{M}, \mathcal{H}^m, \mathcal{S})$*, where:*

- $\mathcal{M}$ *is a set of mentions extracted from* $\mathcal{D}$*, i.e.,* $m \in \mathcal{M}$*;*
- $\mathcal{S} : \mathcal{W} \times \mathcal{M} \rightarrow [0, 1]$ *is a scoring function that estimates the relevance of* m *with respect to an existing word* w*. Ideally, the scoring function might consider the frequency of* m*, as well as the similarity between* m *and* w *according to* $\mathcal{D}$*.*
- $\mathcal{H}^m \subseteq \{(c, m) | c \in \mathcal{C} \wedge m \in \mathcal{M}\}$ *is a taxonomic relation (edge) existing between a <concept,mention> pair. Intuitively,* $\mathcal{H}^m$ *models the pertinence of* m *to be an entity of the existing concept* c*;*

A solution to TEP computed over $\mathcal{D}$ *is a 7-tuple* $\mathcal{T}^{\mathcal{D}} = (\mathcal{C}, \mathcal{W}, \mathcal{H}^c, \mathcal{F}, \mathcal{M}, \mathcal{H}^m, \mathcal{S})$.

3 How Does NEO Work?

In this section we describe our overall approach to enriching ESCO with new emerging occupations, that is mainly composed by three steps: i) *learn word embeddings* ii) *suggest new entities* iii) *vote and enrich* as shown in Fig. 2.

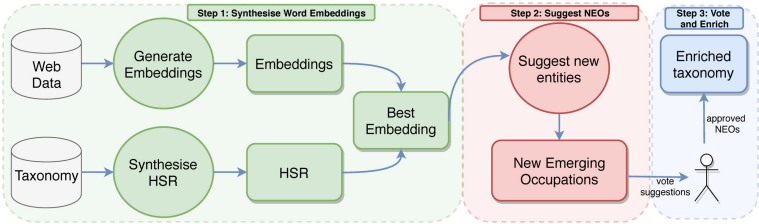


Fig. 2. A representation of the NEO workflow highlighting the main modules.

3.1 Step 1: Synthesise Word Embeddings

The first step requires to learn a vector representation of the words in the corpus to preserve the semantic relationships expressed by the taxonomy itself. To do that, we rely on three distinct sub-tasks, that are the following.

Step 1.1: Generation of embeddings. NEO employs and evaluates three of the most important methods to induce word embeddings from large text corpora. One, GloVe [23], is based on co-occurrence matrix factorisation while the other two, Word2Vec [21], and FastText [6], on neural networks training. Notice that FastText considers sub-word information, and this allows one to share information between words to represent rare words, typos and words with the same root.

Step 1.2: Hierarchical Semantic Relatedness (HSR). Semantic relatedness is a well-known measure for the intrinsic evaluation of the quality of word embeddings, developed in [4]. It evaluates the correlation between a measure of similarity between two terms used as the gold standard and the

cosine similarity between their corresponding word vectors. A common way to build the gold standard is by human evaluation [4]. In many cases, however, this task is difficult, time-consuming, and it might be inaccurate.
Below we introduce a measure of relatedness in a semantic hierarchy, based on the concept of information content. Intuitively, the lower the rank of a concept c which contains two entities, the higher the information content the two entities share. Building on [24], we can supplement the taxonomy with a probability measure $p : \mathcal{C} \rightarrow [0, 1]$ such that for every concept $c \in \mathcal{C}$, $p(c)$ is the probability of encountering an instance of the concept c. From this definition, p (i) is monotonic and (ii) decreases with the rank of the taxonomy, that is if c_1 is a sub-concept of c_2, then $p(c_1) \leq p(c_2)$. According to information theory, the self-information of a concept $c \in \mathcal{C}$ can be approximated by its negative log-likelihood defined as:

$$\mathcal{I}(c) = -\log p(c) \tag{1}$$

We can define the relatedness between concepts of the semantic hierarchy as:

$$sim(c_1, c_2) = \max_{c \in Z(c_1, c_2)} \mathcal{I}(c) = \mathcal{I}(\ell_{c_1, c_2}) \tag{2}$$

where $Z(c_1, c_2)$ is the set of concepts having both c_1 and c_2 as sub-concepts. Given (i), (ii) and Eq. 1, it is easy to verify that ℓ_{c_1, c_2} is the Lowest Common Ancestor of the concepts c_1 and c_2. To estimate the values of p, in [24] the author uses the frequency of the concepts in a large text corpus. Anyway, our purpose is to infer the similarity values intrinsic to the semantic hierarchy. Since we want to extend a semantic hierarchy built by human experts, we adopt those values as a proxy of human judgements. As a consequence, we use the frequencies of the concepts in the taxonomy to compute the values of p.

$$\hat{p}(c) = \frac{N_c}{N} \tag{3}$$

where N is the cardinality, i.e. the number of entities (words), of the taxonomy, and N_c is the sum of the cardinality of the concept c with the cardinality of all its sub-concepts. Note that $\hat{p}(c)$ is monotonic and increases with granularity, thus respects our definition of p. Now, given two words $w_1, w_2 \in W$, we define $Z(w_1)$ and $Z(w_2)$ as the sets of concepts containing w_1 and w_2 respectively, i.e. the *senses* of w_1 and w_2. Therefore, given a pair of words w_1, w_2, there are $N_{w_1} \times N_{w_2}$ possible combinations of their word senses, where N_{w_1} and N_{w_2} are the cardinality of $Z(w_1)$ and $Z(w_2)$ respectively. We refer to $\mathcal{L}$ as the set of all the Lowest Common Ancestor ℓ_{c_1, c_2} for all the combinations of $c_1 \in Z(w_1), c_2 \in Z(w_2)$.

Hence, the hierarchical semantic relatedness between the words w_1 and w_2 is:

$$HSR(w_1, w_2) = \sum_{\ell \in \mathcal{L}} \frac{N_{<w_1, w_2> \in \ell}}{N_{w_1} \times N_{w_2}} \times \mathcal{I}(\ell) \tag{4}$$

where $N_{<w_1,w_2>\in\ell}$ is the number of pairs of senses of word w_1 and w_2 which have ℓ as lowest common ancestor.

Step 1.3. Word Embedding Selection. Finally, the performance of each word vector model generated in Step 1.1 is assessed by the Spearman Correlation of the HSR between all the pairs of words in the taxonomy with the cosine similarity between their vectors in the model space. The Spearman Correlation coefficient can be interpreted as a measure of fidelity of the vector model to the taxonomy.

3.2 Step 2: Suggest New Entities

Step 2 is aimed at extracting new occupation terms from the corpus of OJVs, and at suggesting the most suitable concepts under which they could be positioned in $\mathcal{T}$. To do this, `NEO` works in two distinct steps shown in PseudoCode 1: first, it extracts a set of mentions $\mathcal{M}$ from the corpus $\mathcal{D}$ of OJVs; then, it proposes a set of measures, namely GASC (Generality, Adequacy, Specificity, Comparability) to estimate the suitability of a mention $m \in \mathcal{M}$ as an entity of the concepts in $\mathcal{C}$.

PseudoCode 1 `NEO`

Require: $\mathcal{T}(\mathcal{C}, \mathcal{W}, \mathcal{H}^c, \mathcal{F})$ as in Def. 1; $\mathcal{D}$ dataset;
1: $\mathcal{E} \leftarrow$ best_embedding$(\mathcal{D}, \mathcal{T})$ — *Step*1
2: $\mathcal{M} \leftarrow \emptyset$ //init the set of mentions
3: **for all** $w \in \mathcal{W}$ **do**
4: $\quad \mathcal{M} \leftarrow \mathcal{M} \cup most\_similar(\overrightarrow{\mathcal{E}[w]}, \mathcal{S})$ //ordered according to $\mathcal{S}$ of Eq.5
5: **for all** $m \in \mathcal{M}$ **do** — *Step*2
6: $\quad \mathbf{G}_m \leftarrow$ compute Eq.6
7: $\quad$ **for all** $c \in \mathcal{C}$ **do**
8: $\quad\quad \mathbf{S}_{m,c}, \mathbf{A}_{m,c}, \mathbf{C}_{m,c} \leftarrow$ compute Eq.7, 8, 9
9: $\quad\quad \mathcal{H}^m \leftarrow$ user_eval$(m, c, \mathbf{G}_m, \mathbf{A}_{m,c}, \mathbf{S}_{m,c}, \mathbf{C}_{m,c})$ — *Step*3
10: **return** $(\mathcal{M}, \mathcal{H}^m)$

Step 2.1: Extract new mentions from the corpus. First we select a starting word w_0 from the taxonomy. Then we consider the top-5 mentions in $\mathcal{D}$ with associated the highest *score* value $\mathcal{S}$ with w_0, where the score $\mathcal{S}(m, w)$ of the mention m w.r.t. the generic word w is defined as:

$$\mathcal{S}(m, w) = \alpha \cdot cos\_sim(m, w) + (1 - \alpha) \cdot freq(m) \tag{5}$$

where $cos\_sim(m, w)$ is the cosine similarity between the mention m and the word w in the word embedding model $\mathcal{E}$ selected in Sect. 3.1, while $freq(m)$ is the frequency of the mention m in the corpus. We concentrate on the most important terms, (i) computing the score value only for the top-k most similar mentions, (ii) filtering out the words which are rarely used in the OJVs. To

do this, we compute the cumulative frequency of $freq(m)$ and we keep only the mentions determining the 95% of the cumulative.[4]

Step 2.2: Suggest the best entry concepts for the new mention. Once $\mathcal{M}$ is synthesised, the most suitable concepts are identified on the basis of four measures, namely GASC (*Generality*, *Adequacy* *Specificity*, and *Comparability*), that estimate the fitness of a concept c for a given mention m.

Generality and Specificity. The *Generality* (**G**) of a mention m measures to which extent the mention's embedding is similar to the embeddings of all the words in the taxonomy $\mathcal{T}$ as a whole, in spite of the concept. Conversely, the *Specificity* (**S**) between the mention m and the concept c measures to which extent the mention's embedding is similar to the embeddings that represent the words associated to concept c in $\mathcal{E}$. They are defined as follows.

$$\mathbf{G}_m = \frac{1}{|\mathcal{W}|} \sum_{w \in \mathcal{W}} \mathcal{S}(m, w) \quad (6) \qquad \mathbf{S}_{m,c} = \frac{1}{|\mathcal{F}(c,w)|} \sum_{w \in \mathcal{F}(c,w)} \mathcal{S}(m, w) \quad (7)$$

Adequacy. The *Adequacy* (**A**) between m and c estimates the fitting of the new mention m, extracted from the corpus, to the ESCO concept c, on the basis of the vector representation of m and the words $w \in \mathcal{F}(c, w)$, i.e. their use in the OJVs corpus. **A** is computed as:

$$\mathbf{A}_{m,c} = \frac{e^{\mathbf{S}_{m,c}} - e^{\mathbf{G}_m}}{e - 1} \in [-1, 1] \quad (8)$$

On one side, the *Adequacy* of a mention m to the concept c is directly proportional to the similarity with the other words $w \in \mathcal{F}(c, w)$ (i.e., the *Specificity* to the concept c). On the other side, the *Adequacy* is also inversely proportional to the similarity of m with all the words $w \in W$ (i.e., its *Generality*). The *Adequacy* is defined to hold the following:

$$\mathbf{A}_{m,c} \begin{cases} \gtreqless 0 & \text{if } \mathbf{S}_{m,c} \gtreqless \mathbf{G}_m \\ > \mathbf{A}_{m_2,c_2} & \text{if } \mathbf{S}_{m,c} - \mathbf{G}_m = \mathbf{S}_{m_2,c_2} - \mathbf{G}_{m_2} \wedge \ \mathbf{S}_{m,c} > \mathbf{S}_{m_2,c_2} \end{cases}$$

The first property guarantees zero to act as a threshold value, that is, a negative value of **A** indicates that the mention is related more to the taxonomy, rather than that specific concept c. Conversely, a positive **A** indicates the mention m might be a sub-concept of c. The second property guarantees that given two pairs of concepts and mentions - e.g. (m, c) and (m_2, c_2) - if the difference between their *Specificity* and *Generality* values is the same, then the pair having the higher *Specificity* will also have a higher value of *Adequacy*, still allowing `NEO` to distinguish between the two.

Comparability. To better investigate the comparability of the new mention m with the existing ESCO concepts, we consider their required skills. The skills

[4] k is set to $1,000$ whilst α is set to 0.85 to weight the frequency less than the similarity.

are identified in the context of [10] in the OJVs' descriptions, and classified using the ESCO skills/competencies pillar. Let us consider a set K_c of skills associated to the occupations belonging to the concept c in the OJVs, and a set K_m of skills associated to the mention $m \in \mathcal{M}$ in the OJVs. Given the set $K_U = K_c \cup K_m$ of the L skills associated with at least one out of m and c, we define two L-dimensional vectors $\mathbf{t_c} = (t_{c1}, \ldots, t_{cL})$ and $\mathbf{t_m} = (t_{m1}, \ldots, t_{mL})$ where the generic elements t_{cl} and t_{ml} represent the *revealed comparative advantage* (*rca*) [5] of skill k_l for c and m respectively. If $k_l \notin K_c, t_{cl} = 0$, and similarly if $k_l \notin K_m, t_{ml} = 0$. Given these vectors $\mathbf{t_c}$ and $\mathbf{t_m}$, the *Comparability* ($\mathbf{C}$) between the concept c and the mention m is defined as:

$$\mathbf{C}_{m,c} = \frac{\sum_{l=1}^{L} \min(t_{ml}, t_{cl})}{\sum_{l=1}^{L} \max(t_{ml}, t_{cl})} \tag{9}$$

The *Comparability* represents a method to assess the similarity between an ESCO occupation and a potentially new one not on the basis of their vector representation, but on the basis of their characteristics in the domain of interest.

3.3 Step 3: Vote and Enrich

Finally, we engage labour market experts to validate the outcome of Sect. 3.1 and Sect. 3.2, which are fully automated. The user evaluation is composed of two questions. We ask to evaluate *Q1*) if the mentions extracted from the corpus in Step 2.1 are valid emerging occupations and *Q2*) to which extend the concepts suggested as entry for a new mention are appropriate for it, basing on the name of the mention and the concepts and their skill-gap. We recall that a *novel occupation* is a term that deserves to be represented within the taxonomy, as it might represent either an emerging job or a new alternative label characterising an existing job. For *Q1* the user is asked to give a yes/no answer, while *Q2* is evaluated using a 1–6 Likert scale (from 1: *Completely disagree*, to 6: *Completely agree*). The user feedback is used as a judgement of response quality, meaning that a high evaluation of the best proposed suggestion implies a high quality of suggestion. In the study, we select 12 of the most representative ESCO ICT occupations, i.e. taxonomic entities. For each of them `NEO`, according to Step 2.1, suggests 5 new mentions, for a total of 60, and the expert evaluates whether they can be considered terms representing emerging occupations or not (*Q1*). Then, for each one of the 60 suggestions, `NEO` proposes three candidate concepts where to place the new mention. The first is the concept of starting word, and the other two are those with the highest *Adequacy*, as computed in Step 2.2, among the remaining. The experts evaluate the appropriateness of the proposed mentions for those three concepts (*Q2*).

[5] The $rca \in [0, +\infty]$ was introduced in 2018 in [1] to assess the relevance of skills in the US taxonomy O*Net. We adapted the *rca* to work on ESCO as well.

4 Experimental Results

Experimental Settings. The corpus contains 2,119,025 OJVs published in the United Kingdom during the year 2018, referring to the ESCO ICT positions reported in Table 1, and classified as we specified in [7,9]. OJV's titles were preprocessed applying the following pipeline: (1) tokenisation, (2) lower case reduction, (3) punctuation and stopwords removal (4) n-grams computation.

We deployed `NEO` over the UK dataset following the workflow of Sect. 3.1.

Table 1. OJVs collected from UK in 2018. Only Information and Communication Technology (ICT) occupation codes are shown.

ISCO code	Occupation description	OJVs number
1330	ICT service managers	176,863
2511	Systems analysts	402,701
2512	Software developers	740,112
2513	Web and multimedia developers	225,784
2514	Applications programmers	30,383
2519	Software and applications developers and analysts	44,339
2521	Database designers and administrators	42,305
2522	Systems administrators	45,542
2523	Computer network professionals	15,411
2529	Database and network professionals	110,210
3511	ICT operations technicians	44,585
3512	ICT user support technicians	168,705
3513	Computer network and systems technicians	55,222
3514	Web technicians	5,708
3521	Broadcasting and audiovisual technicians	11,121

4.1 Step 1: Synthesise Word Embeddings

We trained space vector models using various architectures: *Word2Vec*, *GloVe* and *FastText*, generating 260 models. Hyperparameter selection for each architecture was performed with a grid search over the following parameter sets:

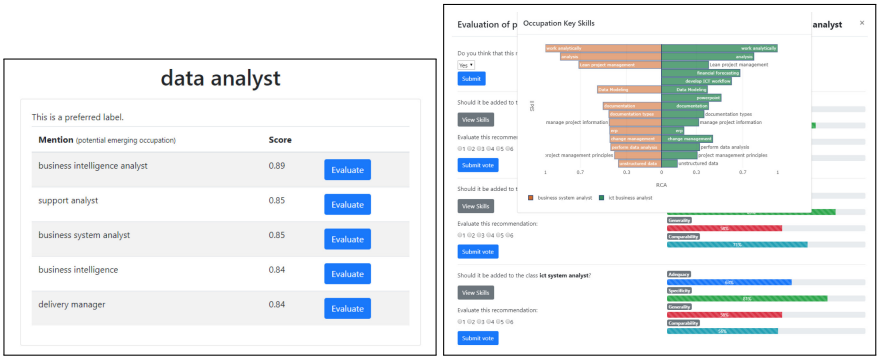
- *Word2Vec* (80 models): Algorithm $\in$ {SG, CBOW} $\times$ HS $\in$ {0, 1} $\times$ embedding size $\in$ {5, 20, 50, 100, 300} $\times$ number of epochs $\in$ {10, 25, 100, 200};
- *GloVe* (20 models): embedding size $\in$ {5, 20, 50, 100, 300} $\times$ number of epochs $\in$ {10, 25, 100, 200};
- *FastText* (160 models): Algorithm $\in$ {SG, CBOW} $\times$ embedding size $\in$ {5, 20, 50, 100, 300} $\times$ number of epochs $\in$ {10, 25, 100, 200} $\times$ learning rate $\in$ {0.01, 0.05, 0.1, 0.2}

Average training times (with std) in seconds were 890 ± 882 for *Word2Vec*, 55 ± 74 for *GloVe* and 246 ± 333 for *fastText*, running on an Intel i-7 CPU equipped with 32GB RAM. An intrinsic evaluation - as detailed in Step 1.3 - has been performed to select the embedding that better preserves taxonomic relations, by computing the Spearman correlation of the cosine similarity between each couple of occupations and their corresponding HSR. The model with highest correlation, with $\rho = 0.29$ and $p\_value \simeq 0$, has the following parameters: architecture = *fastText*, algorithm = CBOW, size = 300, epochs = 100, learning rate = 0.1. Figure 5 provides a scatter plot produced over the best embedding model - as emerges from Table 1 - generated by means of UMAP. Each icon is assigned to one ISCO level 4 group, as in Fig. 1. The ESCO concepts and words belonging to each group are showed, distinguishing between narrower occupations (shallow shape) and alternative labels (filled shape). Focusing on Fig. 5 one might observe that though a *data engineer* and a *data scientist* were designed to be sub-concepts in ESCO, as they belong both to the ▼ *2511: System Analyst* ISCO group, their meaning is quite different in the real-labour market, as any computer scientist knows. The former indeed shares much more with ■ *2521: Database designers and administrators* rather than its theoretical group. Conversely, in many practical cases, the taxonomy perfectly adheres to the real labour market demand for occupations. This is the case of ◆ *3521: Broadcasting and audio-visual technicians*, that composes a very tight cluster in the map, showing a perfect match between de-facto and de-jure labour market occupations. This also applies to ∗ *3513: Computer network and systems technicians*, even though in a lesser extent.[6]

4.2 Step 2: Suggest New Emerging Occupations

As a first step, the user selects the starting word w_0 among the occupations already in ESCO (*data analyst* in the example in Fig. 3). Then, `NEO` prompts the 5 mentions with associated the highest *score* with w_0 (Fig. 3a). The user can therefore select a mention m (*business system analyst*) to evaluate to which extent the mention fits as an entity of the starting word's ESCO concept c_j and as an entity of other two ESCO concepts $c_l, c_k \in \mathcal{C} \setminus c_j$, that are those with associated the highest value of *Adequacy* with the mention m (*ict business analyst* and *ict system analyst* in Fig. 3). For each one of these three pairs mention m and ESCO concept, `NEO` provides the GASC measures (see Fig. 3b). For each pair `NEO` provides comparison of the *rca* of skills for both the mention and the concept (Fig. 3b). These skills, together with the GASC measures, support the domain expert in evaluating if the suggested entry is appropriate or not as an entity of a concept, as thoroughly explained in Sect. 4.3.

[6] Both best/worst embeddings are available at https://tinyurl.com/worst-neo and https://tinyurl.com/best-neo respectively.



(a) New mentions from OJVs starting from the word *data analyst*.

(b) Vote the best class for the new mention selected (*business system analyst*).

Fig. 3. NEO suggests new mentions from the OJV corpus.

4.3 Step 3: Vote and Enrich with User Evaluation

In order to evaluate the effectiveness of NEO we recruited 10 among ML engineers and labour market experts involved in the development of the ML-based system within [10], but not in this research activity. We asked to ten experts to evaluate the system as detailed in Sect. 3.3.

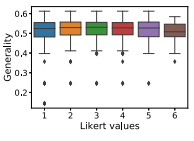
***Q1*: Does NEO suggest valid new emerging occupations?** In *Q1* we ask to the voters whether a suggested mention can be considered an occupation or not. Out of 60 proposed mentions, 11 are repeated starting from different words. For the remaining 49 unique mentions, 6 of them were evaluated to not be proper occupations, according to the majority of the votes. This means that 88% of the occupations were successfully evaluated to be new occupations. Though 6 out of 49 mentions did not pass the test, they are strongly related to the starting concept, referring to skills requested by those job profiles.[7] Figure 4e shows the new occupations found by NEO and the median of Likert scores of experts along with the ESCO concept suggested by NEO and approved by experts.

***Q2*: To which extent the new mentions fit the suggested taxonomic concepts?** To assess the significance of our GASC measures, we use two well-known hypothesis tests, the Spearman's ρ and the Kendall's τ coefficient, that proved to be effective in the labour market domain (see, e.g. [33]). The correlation values are shown in Table 2 while the distribution of the GASC measures grouped according to values of the Likert scale is shown in Fig. 4(a–d). The association between the Likert values and the corresponding *Adequacy*, *Specificity*, and *Comparability* is positive, and hypothesis tests indicate that it is statistically significant. The strongest correlation is between Likert values

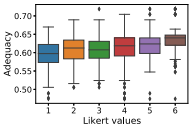
[7] The mentions evaluated not to be proper occupations are: data analytics, business intelligence, penetration testing, operation, data management, drupal.

and *Comparability*, indicating this is the measure on which the experts relied more. Conversely, the association between the Likert values and the *Generality* isn't statistically significant, coherently with the nature of *Generality* that does not aim to rank concepts with respect to a mention.

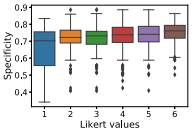
In summary, our results - sustained by an expert user study - show that NEO is able (i) to accurately identify novel occupations and (ii) to put them in the right place within the taxonomy. This, in turn, makes NEO a tool for supporting the process of identification of new emerging occupations enriching the taxonomy accordingly, taking into account the real labour market demand.



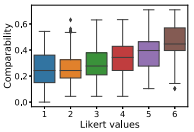
(a) *Generality*



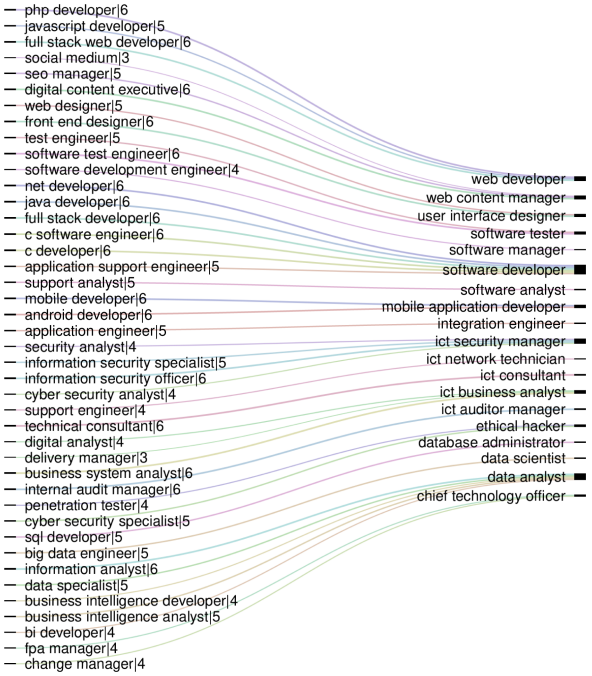
(b) *Adequacy*



(c) *Specificity*



(d) *Comparability*



(e) *Neo Emerging Occupations identified*

Fig. 4. *(left-side)* Box-plots representing the distribution of *Generality*, *Adequacy*, *Specificity* and *Comparability* grouped for each value of the Likert scale. *(right-side)* Alluvial diagram showing the mentions recognised as New Emerging Occupations with the median of Likert values (i.e., neo|score) and the corresponding ESCO concept suggested by NEO and validated by experts.

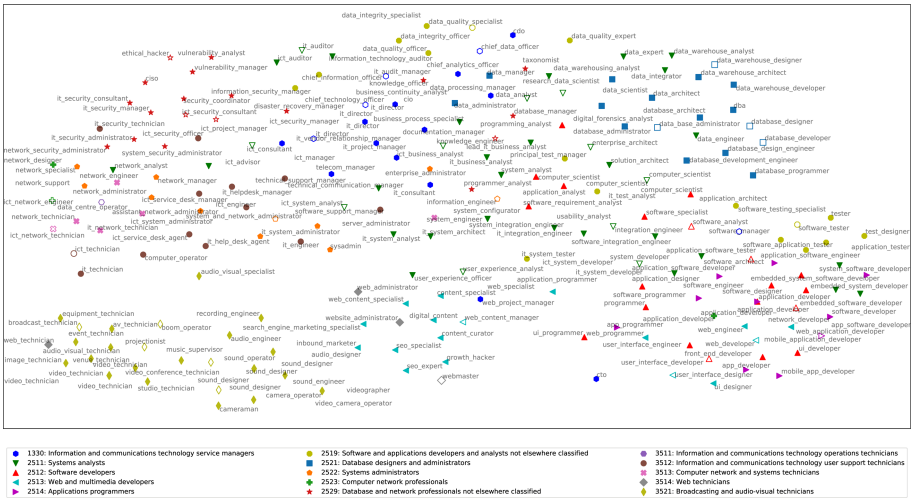


Fig. 5. UMAP plot of the **best** word-embedding model resulting Step 1, that is FastText, CBOW algorithm, Learning rate = 0.1, embedding size = 100, epochs = 100. Each icon is assigned to one ISCO level 4 group, as in Fig. 1. The ESCO concepts and words belonging to each group are shown distinguishing between narrower occupations (shallow shape) and alternative labels (filled shape). The image is also available at https://tinyurl.com/best-neo for a better visualisation.

Table 2. The results of correlation analysis between GASC and Likert values.

Measure	Kendall's τ	*p-value* ($H_0 : \tau = 0$)	Spearman's ρ	*p-value* ($H_0 : \rho = 0$)
Generality	−0.03	0.14	−0.04	0.13
Adequacy	0.20	2.48×10^{-21}	0.27	1.61×10^{-21}
Specificity	0.14	1.59×10^{-11}	0.19	2.59×10^{-11}
Comparability	0.34	2.21×10^{-60}	0.45	8.33×10^{-62}

5 Conclusion and Expected Outlook

In this paper, we proposed NEO, a tool framed within the research activities of an ongoing EU grant in the field of Labour Market Intelligence. NEO has been deployed on a set of 2M+ real OJVs collected from UK in 2018 within the project. NEO synthesised and evaluated more than 240 vector space models, identifying 49 novel occupations, 43 of which were validated as novel occupations by a panel of 10 experts involved in the validation of the system. Two statistical hypothesis tests confirmed the correlation between the proposed GASC metrics of NEO and the user judgements, and this makes the system able to accurately identify novel occupations and to suggest an IS-A relation within the taxonomy.

We are working to scale NEO over multiple country-datasets and occupations, as well as to apply the approach proposed by NEO to other domains.

DEMO. A demo video is provided at https://tinyurl.com/neo-iswc-demo.

References

1. Alabdulkareem, A., Frank, M.R., Sun, L., AlShebli, B., Hidalgo, C., Rahwan, I.: Unpacking the polarization of workplace skills. Sci. Adv. **4**(7), eaao6030 (2018)
2. Aly, R., Acharya, S., Ossa, A., Köhn, A., Biemann, C., Panchenko, A.: Every child should have parents: a taxonomy refinement algorithm based on hyperbolic term embeddings. arXiv preprint arXiv:1906.02002 (2019)
3. Anh, T.L., Tay, Y., Hui, S.C., Ng, S.K.: Learning term embeddings for taxonomic relation identification using dynamic weighting neural network. In: EMNLP (2016)
4. Baroni, M., Dinu, G., Kruszewski, G.: Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In: ACL (2014)
5. Bentivogli, L., Bocco, A., Pianta, E.: ArchiWordNet: integrating wordnet with domain-specific knowledge. In: International Global Wordnet Conference (2004)
6. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. TACL **5**, 135–146 (2017)
7. Boselli, R., et al.: WoLMIS: a labor market intelligence system for classifying web job vacancies. J. Intell. Inf. Syst. **51**(3), 477–502 (2018). https://doi.org/10.1007/s10844-017-0488-x
8. Boselli, R., Cesarini, M., Mercorio, F., Mezzanzanica, M.: Classifying online job advertisements through machine learning. Future Gener. Comput. Syst. **86**, 319–328 (2018)
9. Boselli, R., Cesarini, M., Mercorio, F., Mezzanzanica, M.: Using machine learning for labour market intelligence. In: Altun, Y., et al. (eds.) ECML PKDD 2017. LNCS (LNAI), vol. 10536, pp. 330–342. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-71273-4_27
10. CEDEFOP: Real-time labour market information on skill requirements: setting up the EU system for online vacancy analysis (2016). https://goo.gl/5FZS3E
11. Colombo, E., Mercorio, F., Mezzanzanica, M.: AI meets labor market: exploring the link between automation and skills. Inf. Econ. Policy **47**, 27–37 (2019)
12. Efthymiou, V., Hassanzadeh, O., Rodriguez-Muro, M., Christophides, V.: Matching web tables with knowledge base entities: from entity lookups to entity embeddings. In: d'Amato, C., et al. (eds.) ISWC 2017. LNCS, vol. 10587, pp. 260–277. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-68288-4_16
13. Espinosa-Anke, L., Camacho-Collados, J., Delli Bovi, C., Saggion, H.: Supervised distributional hypernym discovery via domain adaptation. In: EMNLP (2016)
14. Fellbaum, C., Hahn, U., Smith, B.: Towards new information resources for public health–from WordNet to MedicalWordNet. J. Biomed. Inform. **39**(3), 321–332 (2006)
15. Frey, C.B., Osborne, M.A.: The future of employment: how susceptible are jobs to computerisation? Technol. Forecast. Soc. Change **114**, 254–280 (2017)
16. Giabelli, A., Malandri, L., Mercorio, F., Mezzanzanica, M.: GraphLMI: a data driven system for exploring labor market information through graph databases. Multimed. Tools Appl. (2020). https://doi.org/10.1007/s11042-020-09115-x. ISSN 1573-7721

17. Harris, Z.S.: Distributional structure. Word **10**(2–3), 146–162 (1954)
18. Jurgens, D., Pilehvar, M.T.: Reserating the awesometastic: an automatic extension of the WordNet taxonomy for novel terms. In: ACL, pp. 1459–1465 (2015)
19. Maedche, A., Staab, S.: Ontology learning for the semantic web. IEEE Intell. Syst. **16**(2), 72–79 (2001)
20. Mezzanzanica, M., Boselli, R., Cesarini, M., Mercorio, F.: A model-based approach for developing data cleansing solutions. J. Data Inf. Qual. (JDIQ) **5**(4), 1–28 (2015)
21. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
22. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: NIPS (2013)
23. Pennington, J., Socher, R., Manning, C.: GloVe: global vectors for word representation. In: EMNLP, pp. 1532–1543 (2014)
24. Resnik, P.: Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. JAIR **11**, 95–130 (1999)
25. Ristoski, P., Paulheim, H.: RDF2Vec: RDF graph embeddings for data mining. In: Groth, P., et al. (eds.) ISWC 2016. LNCS, vol. 9981, pp. 498–514. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46523-4_30
26. Schlichtkrull, M., Alonso, H.M.: MSejrKu at SemEval-2016 Task 14: taxonomy enrichment by evidence ranking. In: SemEval, pp. 1337–1341 (2016)
27. Shen, J., Shen, Z., Xiong, C., Wang, C., Wang, K., Han, J.: TaxoExpan: self-supervised taxonomy expansion with position-enhanced graph neural network. In: WWW, pp. 486–497 (2020)
28. Sumida, A., Torisawa, K.: Hacking Wikipedia for hyponymy relation acquisition. In: IJCNLP (2008)
29. Toral, A., Monachini, M.: Named entity wordnet. In: LREC (2008)
30. Vedula, N., Nicholson, P.K., Ajwani, D., Dutta, S., Sala, A., Parthasarathy, S.: Enriching taxonomies with functional domain knowledge. In: SIGIR (2018)
31. Wang, C., He, X., Zhou, A.: A short survey on taxonomy learning from text corpora: issues, resources and recent advances. In: EMLP, pp. 1190–1203 (2017)
32. Wang, J., Kang, C., Chang, Y., Han, J.: A hierarchical dirichlet model for taxonomy expansion for search engines. In: WWW, pp. 961–970 (2014)
33. Xu, T., Zhu, H., Zhu, C., Li, P., Xiong, H.: Measuring the popularity of job skills in recruitment market: a multi-criteria approach. In: AAAI (2018)
34. Zhang, D., et al.: Job2Vec: job title benchmarking with collective multi-view representation learning. In: CIKM, pp. 2763–2771 (2019)