PIAAC2ESCO - An AI-driven classification of the PIAAC Background questionnaire onto the ESCO Skills Pillar

Technical Annex

*This project receives funding from the European Union's Horizon 2020 Resesarch and Innovation Programme under Grant Agreement No 101004703. Corresponding Author: Fabio Mercorio, fabio.mercorio@unimib.it

1. Contribution

We provide a novel mapping of ESCO skills to skill items elicited in the Programme for International Assessment of Adult Competencies (PIAAC) using AI algorithms of word embeddings. The approach here presented is built on top of the approach to bridge labour market taxonomy proposed by Giabelli et al. (2022).

2. Background and state of the art

2.1. Initiatives based on Online Job Ads

In recent years, the labor demand and supply conveyed through specialized Web portals and services have grown exponentially. This also contributed to introducing the term *Labor Market Intelligence* (LMI), which refers to the use and design of AI algorithms and frameworks to analyze Labor Market Information for supporting decision making (see, e.g., Turrell et al. (2018); Papoutsoglou et al. (2019); Javed et al. (2017); UK Commission for Employment and Skills (2015); Vinel et al. (2019); Giabelli et al. (2021a,c)).

In the LMI scenario, the problem of monitoring, analyzing, and understanding these labor market changes (i) timely and (ii) at a very fine-grained geographical level, has become practically significant in our daily lives. Recently, machine learning has been applied to compute and estimate the impact of robotization on the occupations of the US Labor Market Frey and Osborne (2017); Fleming et al. (2019) as well as to analyze skill relevance in the US standard taxonomy O*NET Alabdulkareem et al. (2018), just to cite a few.

In 2010 the European Commission issued the communication "A new impetus for European Cooperation in Vocational Education and Training (VET) to support the Europe 2020 strategy",[1] aimed at promoting education systems in general, and VET in particular. In 2016, the European Commission's highlighted the importance of Vocational and Educational activities, as they are "valued for fostering job-specific and transversal skills, facilitating the transition into employment and maintaining and updating the skills of the workforce according to sectoral, regional, and local needs".[2] In 2016, the EU and Eurostat launched the ESSnet Big Data project, involving 22 EU member states with the aim of "integrating big data in the regular production of official statistics, through pilots exploring the potential of selected big data sources and building concrete applications". In 2014

[1] Publicly available at https://goo.gl/Goluxo

[2] The Commission Communication "A New Skills Agenda for Europe" COM(2016) 381/2, available at https://goo.gl/Shw7bI

the EU Cedefop agency - aimed at supporting the development of European Vocational Education and Training - launched a call-for-tender for realizing a system able to collect and classify Web job vacancies from 5 EU Countries CEDEFOP (2014). Moreover, a further project has been initiated to extend and scale the ML-based system to the whole EU, including 28 EU country members and all the 32 languages of the Union CEDEFOP (2016), that is the project in which this work is framed within. Some results have been published in e.g. Boselli et al. (2018a, 2017); Colombo et al. (2019a); Boselli et al. (2018b); Colace et al. (2019).

In late 2020, EUROSTAT and Cedefop joined forces announcing a call for tender EuroStat (2020) aimed at establishing results from CEDEFOP (2014, 2016) fostering AI and Statistics to build up the European Hub of Online Job Ads.

As one might note, the use of classified online job advertisements and skills, in turn, enables several third-party research studies to understand and explain complex labour market phenomena. To give few recent examples, Colombo et al. (2019b) used online job advertisements for estimating the impact of AI in job automation and measuring the impact of digital/soft skills within occupations; In Giabelli et al. (2020b, 2021b) authors used classified online job advertisements to identify new emerging occupations. In Giabelli et al. (2021d) authors classified online job advertisements to build a recommendation system of skills for citizens, while in Giabelli et al. (2020a) the first graph database of the European labour market to be explored through graph-traversal queries has been realised.

Notably, on May 2020, the EU Cedefop Agency has been started using those online job advertisements to build an index named Cov19R that identifies workers with a higher risk of COVID-19 exposure, who need greater social distancing, affecting their current and future job performance capacity[3].

All these initiatives and research studies elucidate the importance of explaining the rationale behind the classification process for decision-makers.

2.2. Introducing Word Embeddings

Evaluating the intrinsic quality of vector space models, as well as their impact when used as the input of specific tasks (i.e., extrinsic quality), has a very practical significance (see, e.g. Turian et al. (2010); Camacho-Collados and Pilehvar (2018)), as this affects the believability[4] of the overall process or system in which they are used. In essence, we may

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

[4] Here the term *believability* is intended as "the extent to which data are accepted or regarded as true, real and credible"Wang and Strong (1996)

argue that the well-known principle *"garbage-in, garbage-out"* that characterise the data quality research in many domains also applies to word embedding, that is, *the lower the quality of the word embeddings, the lower the performance of the tasks that are based on them*.

Word embeddings are vector representations of words, based on the hypothesis that words occurring in a similar context tend to have a similar meaning. Words are represented by semantic vectors, which are usually derived from a large corpus using co-occurrence statistics, and their use improves learning algorithms in many NLP tasks. Two powerful methods to induce word embeddings are neural networks training Collobert and Weston (2008); Mikolov et al. (2013) and co-occurrence matrix factorisation Pennington et al. (2014); Levy and Goldberg (2014).

These techniques consider each word as a distinct vector and ignore the morphological similarity among them. More recently Bojanowski et al. (2017) developed a version of the continuous skip-gram model Mikolov et al. (2013) which considers subword information. This architecture, called fastText, is an extension of word2vec for scalable word representation and classification. One of its major improvements to word2vec is to consider sub-word information by representing each word as the sum of its character n-gram vectors. Formally, given a word w, and a dictionary of size G, G_{w} is the set of n-grams of size G appearing in w. Denoted as z_g vector representation of the n-gram g, w will be represented as the sum of the vector representation of its n-grams and the score associated to the word w as:

$$\int(\mathrm{w}, \mathrm{c}) = \sum_{g \in G_{\mathrm{w}}} z_g^{\intercal} v_{\mathrm{c}} \tag{1}$$

where v_{c} is the vector representing the context. This simple representation allows one to share information between words, and this makes it useful to represent rare words, typos, and words with the same root.

Other embedding models have been evaluated along with fastText. Nevertheless, none of them fit our conditions. Neither classical embedding models Mikolov et al. (2013); Pennington et al. (2014) nor embeddings specifically designed to fit taxonomic data consider subword information. Moreover, they cannot be easily bonded with external sources in their generation phase, and this would reduce the flexibility of . Regarding hyperbolic and spherical embeddings like HyperVec Nguyen et al. (2017) or JoSe Meng et al. (2019), we discarded them since (i) they also donât consider subword information, which is important for short text and many words with the same root (e.g. engineer-engineering,developer-developing) like online job ads, and (ii) HyperVec uses hypernym-hyponym relationships

for training, while we train our models on a text corpus which has not such relationships. Finally, we considered context embeddings (see e.g. Devlin et al. (2018)). However, contextual embeddings represent words based on their context, thus capturing the uses of words across varied contexts. This is not suitable for our case, where we aim to compare words in a corpus and their similarity with words of taxonomy, with a given sense.

2.3. State of the art using AI to bridge taxonomies

In recent literature, taxonomy alignment has received considerable attention. Different approaches have been proposed to create an efficient mapping.

In one of the first approaches, Euzenat et al. (2004) compute the similarity between entities through a system of quasi-linear equations, which start from lexical similarity derived by WordNet 2.0 and gradually include contributions from structure comparing functions. Avesani et al. (2005) uses both a syntactic and a semantic score of taxonomic similarity, called COMA and S-match respectively. COMA exploits both element and structure level syntactic similarity, while S-match uses Wordnet 2.0 to derive semantic similarity between words. In Wu et al. (2017), an approach based on Wikipedia-matching and keywords is considered to perform document classification without employing standard occurrence methods. Despite being relevant and widely used, those methods are built on specific lexical resources (WordNet, Wikipedia) thus are not suitable for different domains. In Jung (2008) the authors use LSA (Latent Semantic Analysis) to group sets of related entities based on their co-occurrence matrix and TF-IDF, also considering the description of the taxonomic concepts. In Wu et al. (2016), authors train a bilingual topic model on contextual text extracted from the web to build semantic vectors of the topics of two multi-lingual taxonomies. The cosine similarity between those vectors represents the relevance of each concept in the source taxonomy and its candidate-matched categories. Each candidate entity is then evaluated through syntactic similarity. Those kinds of approaches make use of contextual information and learning algorithms. However, none of them neither considers the vertical structure of the taxonomy to match entities nor employ distributional semantics, which has shown to be beneficial in several NLP applications in the last years.

More recently, Giabelli et al. (2022) proposed the `WETA` (**W**eb **T**axonomy **E**mbedding **A**lignment) approach that exploits distributional semantic and context information to perform taxonomy alignment, blending a hierarchical approach based on cosine similarity and a machine learning classification task that uses the embeddings as input features. Moreover, it performs an intrinsic evaluation of the selected embedding model based on the structure of the taxonomy itself.

3. Data

Our contribution is structured in two steps, that require different input data. The sources that we use are the ESCO classification, the Survey of Adult Skills (PIAAC) and the OJV data. The first step of our analysis, presented in Section 4, requires the whole set of skills provided in the ESCO classification, Skills Pillar, and the list of selected items of the PIAAC Questionnaire. The second stage, which encompasses the application of the bridge to white-collar workers and is presented in Section **??**, leverages on answers to PIAAC items and OJV data.

3.1. ESCO, Skill Pillar

ESCO (European Skills, Competences, Qualifications and Occupations) is the European classification of skills, competencies and occupations. It provides a multilingual dictionary of occupations and skill requirements organised along two main pillars. The first is the Occupation pillar, which is referenced in the ISCO08 standard. The second is the Skills pillar which lists and describes competencies/skills which are linked to occupations. ESCO provides a list of occupations and related skills, organised as a network; nonetheless, it gives no information on the importance that skills have in the considered occupation.

3.2. PIAAC

PIAAC is a survey conducted by the Organisation for Economic Co-operation and Development (OECD) and contains information on key cognitive and workplace skills of adults aged 16-65 across OECD countries. The main aim of the PIAAC survey is to assess literacy, numeracy and problem solving skills in technology-rich environments using tests in each of these domains We focus on the "skill use at work" items. These items elicit the frequency of skills used in various domains of job tasks on a Likert scale. For instance, subjects are asked to which extent they e.g. teach people, use calculators, or read financial statements. The response scale is a 5-item Likert scale, ranging from 1 - never, to 5 - every day[5]. In total, the PIAAC survey contains 64 items eliciting skill use at work.

We use the latest wave of the PIAAC data which comprises representative samples of working-age individuals of 24 OECD countries and were collected between August 2011 and March 2012. In total, the sample comprises 250,000 observations, with sample sizes typically ranging from 4000-8000 observations in each country.

[5] 1 - Never, 2 - Less than once a month, 3 - Less than once a week but at least once a month, 4 - At least once a week but not every day, 5- Every day

3.3. Online job-advertisements data

Online job-advertisements data are obtained from the European Center for the Development of Vocational Training (CEDEFOP). The agency launched a tender in 2016 to develop a system to collect online job advertisements to analyse vacancies and emerging skill requirements across all EU Member States[6]. The system has become fully functional in the last quarter of 2018. We use data from 2019, the first full year available, and extract the universe of online job advertisements posted in UK, corresponding to 4,335,640 observations. The choice of the English language is due to the richness and novelty in jobs' descriptions.

4. Methods: Bridging PIAAC and ESCO Skill Pillar

In this section, we introduce a formal definition of taxonomy and formulate the problem of taxonomy alignment, relying on the formalisation proposed by Maedche and Staab (2001). Then, we summarise how `WETA` (Giabelli et al., 2022) bridges taxonomies through word embeddings.

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 $c \in \mathcal{C}$ (i.e., nodes) that can be classified in p different hierarchical levels: $\mathcal{C}_1, \ldots, \mathcal{C}_p$;*
- *$\mathcal{W}$ is a set of words (or entities, or leaf concepts) 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 between concepts, that is $\mathcal{H}^c \subseteq \{(c_i, c_j) \mid (c_i, c_j) \in \mathcal{C}^2) \wedge i \neq j\}$. $\mathcal{H}^c(c_1, c_2)$ means that c_1 is a sub-concept, or hyponym, of c_2, while c_2 is the hypernym of c_1, meaning c_2 has a broader meaning and constitutes a category into which c_1 falls. 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. $\mathcal{F} \subseteq \{(c, w) \mid c \in \mathcal{C} \wedge w \in \mathcal{W}\}$. $\mathcal{F}(c, w)$ means that the word w is an entity of the concept c.*

$\mathcal{T}$ could be represented as a Directed Acyclic Graph (DAG), therefore the concepts at the most specific level have an in-degree of 0, i.e. they don't have any incoming edge.

[6] https://www.cedefop.europa.eu/en/about-cedefop/public-procurement/real-time-labour-market-information-skill-requirements-setting-eu

We refer to those concepts as leaf concepts, which are the concepts representing different entities, or words. Note that in several taxonomies the terms representing leaf concepts are also item words, while concepts at a higher level are not.

Given an origin taxonomy $\mathcal{T}_o$ (i.e., PIAAC) and a destination taxonomy $\mathcal{T}_d$ (i.e., ESCO skill pillar), the goal of `WETA` is to suggest one or more concepts $c \in \mathcal{T}_d$ for each word $w \in \mathcal{T}_o$. More specifically, for each $w \in \mathcal{T}_o$, n possible $c \in \mathcal{T}_d$ are suggested based on the scoring function $\mathcal{S}$. More formally:

Definition 2 (Taxonomy Alignment Problem (TAP))**.** *Let $\mathcal{T}_o$ and $\mathcal{T}_d$ be respectively an origin and a destination taxonomy as in Def.1. A Taxonomy Alignment Problem (TAP) is a 3-tuple $(\psi, h, \mathcal{S})$, where:*

- $\psi : \mathcal{W}^o \times \mathcal{C}^d \rightarrow [0, 1]$ *is a scoring function that estimates the relevance of $c \in \mathcal{T}_d$ with respect to a word $w \in \mathcal{T}_o$ considering the prediction scores of a multi-class classification task;*
- $h : \mathcal{W}^o \times \mathcal{C}^d \rightarrow [0, 1]$ *is a scoring function that estimates the relevance of $c \in \mathcal{T}_d$ with respect to a word $w \in \mathcal{T}_o$ considering the semantic similarity of w with c and all its hypernyms;*
- $\mathcal{S}(\psi, h) \subseteq \{(w, c) \mid w \in \mathcal{W}^o \wedge c \in \mathcal{C}^d\}$ *is the score of an alignment relation existing between a word in $\mathcal{T}_o$ and a concept in $\mathcal{T}_d$, blending the results of the above mentioned scoring functions.*

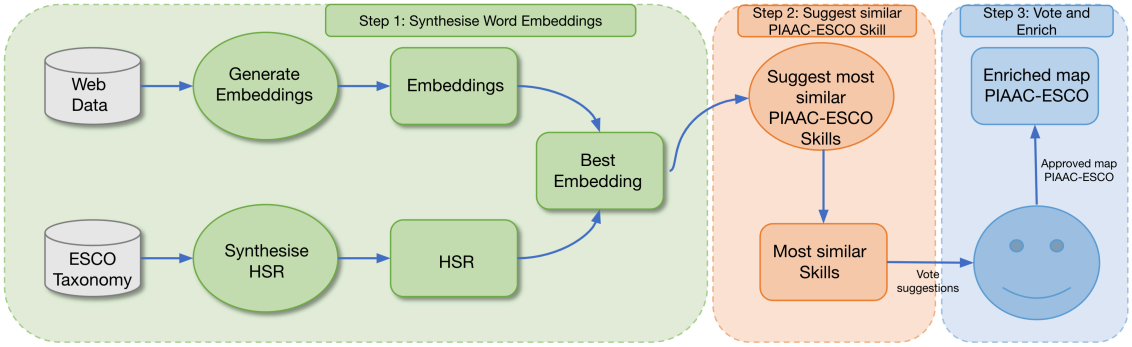
A solution to TAP computed over $\mathcal{T}_o$ and $\mathcal{T}_d$ is a 3-tuple $\mathcal{T}_{o,d} = (\mathcal{W}^o, \mathcal{C}^d, \mathcal{S})$

Given a hierarchical level x, we define $\mathcal{C}^o_x$ and $\mathcal{C}^d_x$ as the set of all the $c \in \mathcal{T}_o$ and $c \in \mathcal{T}_d$ respectively, that are classified in level x.

4.1. Bridging PIAAC on ESCO in a nutshell

This section aims to describe the global approach used to aligning the taxonomies $\mathcal{T}_o$ and $\mathcal{T}_d$. The first step allows us to train and select the best word embedding model, which is then used in the second step to suggest for each leaf concept $w_o \in \mathcal{W}^o$ n possible alignments $c_d \in \mathcal{C}^d_p$. The last step consists of the validation of the suggestions because the utility of `WETA` is the help it provides to the domain experts, narrowing the choices for the alignment that would otherwise be done by scratch. The approach is depicted in Fig. 1.

Figure 1: Graphical overview of the approach proposed



4.2. Step 1: Generate and evaluate embeddings

The main goal of the first step of `WETA` is to induce a vector representation of taxonomic terms that represent as much as possible the similarity of words within the taxonomy. To accomplish this, we perform three distinct tasks. We (i) generate word embeddings through a state of the art method; we (ii) compute the `HSR` of terms in $\mathcal{T}_o$ and $\mathcal{T}_d$, and finally, we (iii) select the embeddings for which the correlation between the cosine similarity between taxonomic terms and their `HSR` is maximised for both $\mathcal{T}_o$ and $\mathcal{T}_d$.

4.2.1. Embeddings generation

For the generation of the word embedding models, `WETA` employs the state of the art method FastText Bojanowski et al. (2017), a word embedding method that considers sub-word information and can deal with out-of-vocabulary words.

4.2.2. Selection of the best word embedding

To select the best embeddings model, we perform an intrinsic evaluation following Baroni et al. (2014). The authors select the word vectors model which has a maximum correlation between their cosine similarity and a benchmark value of semantic similarity. In Baroni et al. (2014) the authors use a handcrafted dataset of pairwise semantic similarity between common words as the gold benchmark. However, those resources usually have low coverage, especially in specific domains like the labour market. For this reason, we resort to a measure of semantic similarity in taxonomies developed in Malandri et al. (2021) and refined in Malandri et al. (2020), which measures semantic similarity in a taxonomy based on the structure of the hierarchy itself without using any external resource, thus, in a sense, preserving the semantic similarity intrinsic to the taxonomy. The `HSR` has proven to be useful in the selection of embeddings for several applications,

like taxonomy enrichment Giabelli et al. (2020c, 2021a) and job-skill mismatch analysis in the field of labour market Giabelli et al. (2021c).

Since in Malandri et al. (2020) the authors want to encode semantic information from a semantic hierarchy built by human experts, they adopt those values as a proxy of human judgements. Therefore, they compute:

$$\hat{p}(c) = \frac{N_c}{N} \tag{2}$$

where N is the cardinality of the taxonomy, and N_c is the cardinality of the concept c and all its hyponyms. Note that $\hat{p}(c)$ is monotonic and increases with granularity.

Intuitively, given two words w_1, w_2 in the taxonomy, $c_1 \in s(w_1)$ and $c_2 \in s(w_2)$ are defined as all the concepts containing w_1 and w_2 respectively, i.e. the *senses* of w_1 and w_2. Therefore, there are $S_{w_1} \times S_{w_2}$ possible combinations of their word senses, where S_{w_1} and S_{w_2} are the cardinality of $s(w_1)$ and $s(w_2)$ respectively. $\mathcal{L}$ is the set of all the lowest common ancestor for all the combinations of $c_1 \in s(w_1)$ and $c_2 \in s(w_2)$. The hierarchical semantic similarity between w_1 and w_2 is defined by the authors of Malandri et al. (2020) as:

$$sim_{\texttt{HSR}}(w_1, w_2) = \sum_{\ell \in \mathcal{L}} \hat{p}\left(\ell = LCA \mid w_1, w_2\right) \times I(LCA) \tag{3}$$

where $I(c)$ is the self-information of the concept c and $\hat{p}(\ell = LCA \mid w_1, w_2)$ is the probability of LCA being the lowest common ancestor of w_1, w_2, and it can be computed as follows:

$$\hat{p}\left(\ell = LCA \mid w_1, w_2\right) = \frac{\frac{S_{<w_1,w_2>\in\ell}}{|descendants(\ell)|^2} \times \frac{N_\ell}{N}}{\sum_{k\in\mathcal{L}} \frac{S_{<w_1,w_2>\in k}}{|descendants(k)|^2} \times \frac{N_k}{N}} \tag{4}$$

where $S_{<w_1,w_2>\in\ell}$ is the number of pairs of senses of words w_1 and w_2 that have ℓ as lower common ancestor, the term $|descendants(\ell)|$ represents the number of sub-concepts of ℓ, and N_ℓ is the cardinality of ℓ and all its descendants. For more details see Malandri et al. (2020, 2021).

4.3. Step 2: Taxonomy alignment method

`WETA` proposes a methodology for taxonomy alignment that suggests, for each word, or leaf concept, $w_o \in \mathcal{T}_o$, a set of n possible destination concept in $\mathcal{T}_d$. The destination concepts are selected among most specialised concepts in $\mathcal{T}_d$, i.e. those which are at the lowest level p, that is $\{c_1, \ldots, c_n\} \in \mathcal{C}_p^d$.

To do this, we perform two different processes that lead to independent results, and then we blend their suggestions to obtain a robust mapping between taxonomies.

4.3.1. Hierarchical approach

For each $w_o \in \mathcal{W}^o$, the set of words of the original taxonomy, we create a list that contains the cosine similarity between w_o and each element in $w \in \mathcal{W}^d$:

$$L_H^{(w_o)} = \{(w, sim(w_o, w)) \mid w \in \mathcal{W}^d\} \; \forall w_o \in \mathcal{W}^o \tag{5}$$

where sim is the cosine similarity between the vector representations of the two inputs in the best word embedding model, see Sec. 4.2.

Given the i-th pair $(w, sim)_i$, we can refer to its similarity as sim_i to define the function W as follows:

$$W(sim_i) = (sim_i)^2 \cdot (sim_i - sim_{i+1}) \tag{6}$$

where to transform each similarity score we consider the next similarity in the ordered list. Thanks to W, we can highlight the situations in which, for example, the similarity score between w_o and the first word in $L_H^{(w_o)}$ is significantly higher than the other scores in the ordered list, rather than a situation where all the elements have a high similarity with w_o. Now, we exploit the hierarchical concepts: for each $w \in L_H^{(w_o)}$ we extract its respective hypernym at the level p. We define $L_{Hp}^{(w_o)}$ as the list that contains all the level p hypernyms of every $(w, sim) \in L_H^{(w_o)}$, we order them, and we keep the n with associated the highest similarity. More formally:

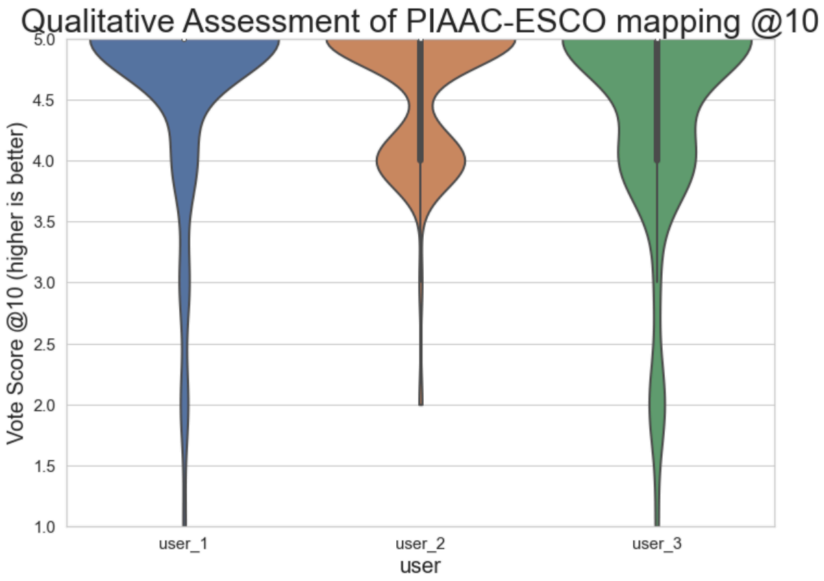
$$\begin{aligned} L_{Hp}^{(w_o)} &= \{(c_p, h_p) \mid \exists (w, sim) \in L_H^{(w_o)} : \mathcal{F}(c_p, w) \wedge c_p \in \mathcal{C}_p^d\} \; \forall w_o \in \mathcal{W}^o \\ &\text{with } h_p = \max_{sim \in S_p} sim, \; S_p = \{sim | \; (w, sim) \in L_H^{(w_o)}\} \end{aligned} \tag{7}$$

We keep only the n pairs with the highest similarity score h_p.

4.4. Step 3: Evaluation of the suggestions

The usefulness of `WETA` is that it provides a limited number of suggestions to the domain experts to simplify their work of taxonomy alignment that otherwise would be all manual. The last step consists of the validation of the suggestions provided to complete the alignment procedure. Tab. 2 provides the list of PIAAC questions that are to be aligned with ESCO. Let us consider the PIAAC question (G_Q05g) that asks workers to what extent they *"use a programming language to program or write computer code"*. Our

Figure 2: Violin Plots of the validation phase of the PIAAC to ESCO mapping. Votes are expressed in the Likert scale: the higher, the better



approach automatically suggests a list of ESCO skills that are likely to be related to the PIAAC question. In this example, our approach proposes (limited to top-5 matches):

- computer programming
- Python (computer programming)
- Java (computer programming)
- software and applications development and analysis
- database and network design and administration

Fig.2 reports the results of the validation phase, made by involving experts of the PILLARS consortium. Each member was asked to vote if - and to what extent - the ESCO suggestions were relevant and consistent with the PIAAC questions using a Likert scale. We reported the mean and standard deviation @10 for each PIAAC skill for users. The mean value is over 4,6 and the mean of the standard deviation demonstrates that is a quite stable result. From this plot, it is clear that vote scores are concentrated in the upper part of the graph so for most skills there is at least one suggestion with a high level of agreement.

The selected PIAAC questions are presented in 1.

Table 1: Selected PIAAC questions and metadata

PIAAC item	Label	Set	Value scheme
F_Q02b	Teaching people	General	Frequency (time units)
F_Q02d	Selling	General	Frequency (time units)
F_Q04a	Influencing people	General	Frequency (time units)
F_Q05a	Simple problems	Problem solving	Frequency (time units)
G_Q01b	Read letters memos or mails	Literacy	Frequency (time units)
G_Q01g	Read financial statements	Literacy	Frequency (time units)
G_Q01h	Read diagrams maps or schematics	Literacy	Frequency (time units)
G_Q02a	Write letters memos or mails	Literacy	Frequency (time units)
G_Q03b	Calculating costs or budgets	Numeracy	Frequency (time units)
G_Q03c	Use or calculate fractions or percentages	Numeracy	Frequency (time units)
G_Q03d	Use a calculator	Numeracy	Frequency (time units)
G_Q03g	Use simple algebra or formulas	Numeracy	Frequency (time units)
G_Q03h	Use advanced math or statistics	Numeracy	Frequency (time units)
G_Q04	Experience with computer in job	ICT	Yes (1) / No (2)
G_Q05a	For mail	ICT - Internet	Frequency (time units)
G_Q05d	Conduct transactions	ICT - Internet	Frequency (time units)
G_Q05e	Spreadsheets	ICT - Computer	Frequency (time units)
G_Q05f	Word	ICT - Computer	Frequency (time units)
G_Q05g	Programming language	ICT - Computer	Frequency (time units)
I_Q04d	Like learning new things	Learning strategies	Extents
I_Q04l	Figure out how different ideas fit together	Learning strategies	Extents

]

Table 2: PIAAC questions used to bridge skills over ESCO

PIAAC code	PIAAC question
F_Q02a	sharing work-related information with co-workers?
F_Q02b	instructing, training or teaching people, individually or in groups?
F_Q02c	making speeches or giving presentations in front of five or more people?
F_Q02d	selling a product or selling a service?
F_Q02e	advising people?
F_Q03a	planning your own activities?
F_Q03b	planning the activities of others?
F_Q03c	organising your own time?
F_Q04a	persuading or influencing people?
F_Q04b	negotiating with people either inside or outside your firm or organisation?
F_Q05a	The next question is about "problem solving" tasks you ^DoDid in your ^JobLastjob. Think of "problem solving" as what happens when you are faced with a new or difficult situation which requires you to think for a while about what to do next. How often ^AreWere you usually faced by relatively simple problems that ^TakeTook no more than 5 minutes to find a good solution?
F_Q05b	And how often ^AreWere you usually confronted with more complex problems that ^TakeTook at least 30 minutes to find a good solution? The 30 minutes only refers to the time needed to THINK of a solution, not the time needed to carry it out.
F_Q06b	working physically for a long period?
F_Q06c	using skill or accuracy with your hands or fingers?
F_Q07a	Do you feel that you have the skills to cope with more demanding duties than those you are required to perform in your current job?
G_Q01a	read directions or instructions?
G_Q01b	read letters, memos or e-mails?
G_Q01c	read articles in newspapers, magazines or newsletters?
G_Q01d	read articles in professional journals or scholarly publications?
G_Q01e	read books?
G_Q01f	read manuals or reference materials?
G_Q01g	read bills, invoices, bank statements or other financial statements?

Continued on next page

Table 2 – continued from previous page

PIAAC code	PIAAC question
G_Q01h	read diagrams, maps or schematics?
G_Q02a	write letters, memos or e-mails?
G_Q02b	write articles for newspapers, magazines or newsletters?
G_Q02c	write reports?
G_Q02d	fill in forms?
G_Q03b	calculate prices, costs or budgets?
G_Q03c	use or calculate fractions, decimals or percentages?
G_Q03d	use a calculator - either hand-held or computer based?
G_Q03f	prepare charts, graphs or tables?
G_Q03g	use simple algebra or formulas?
G_Q03h	use more advanced math or statistics such as calculus, complex algebra, trigonometry or use of regression techniques?
G_Q04	^DoiDid you use a computer in your ^JobLastjob?
G_Q05a	use email?
G_Q05c	use the internet in order to better understand issues related to your work?
G_Q05d	conduct transactions on the internet, for example buying or selling products or services, or banking?
G_Q05e	use spreadsheet software, for example Excel?
G_Q05f	use a word processor, for example Word?
G_Q05g	use a programming language to program or write computer code?
G_Q05h	participate in real-time discussions on the internet, for example online conferences, or chat groups?
H_Q01a	read directions or instructions?
H_Q01b	read letters, memos or e-mails?
H_Q01c	read articles in newspapers, magazines or newsletters?
H_Q01d	read articles in professional journals or scholarly publications?
H_Q01e	read books, fiction or non-fiction?
H_Q01f	read manuals or reference materials?
H_Q01g	read bills, invoices, bank statements or other financial statements?
H_Q01h	read diagrams, maps, or schematics?
H_Q02a	write letters, memos or e-mails?

Continued on next page

Table 2 – continued from previous page

PIAAC code	PIAAC question
H_Q02b	write articles for newspapers, magazines or newsletters?
H_Q02c	write reports?
H_Q02d	fill in forms?
H_Q03b	calculate prices, costs or budgets?
H_Q03c	use or calculate fractions, decimals or percentages?
H_Q03d	use a calculator - either hand-held or computer based?
H_Q03f	prepare charts, graphs or tables?
H_Q03g	use simple algebra or formulas?
H_Q03h	use more advanced math or statistics such as calculus, complex algebra, trigonometry or use of regression techniques?
H_Q04a	Have you ever used a computer?
H_Q05a	use email?
H_Q05c	use the internet in order to better understand issues related to, for example, your health or illnesses, financial matters, or environmental issues?
H_Q05d	conduct transactions on the internet, for example buying or selling products or services, or banking?
H_Q05e	use spreadsheet software, for example Excel?
H_Q05f	use a word processor, for example Word?
H_Q05g	use a programming language to program or write computer code?
H_Q05h	participate in real-time discussions on the internet, for example online conferences or chat groups?
I_Q04b	When I hear or read about new ideas, I try to relate them to real life situations to which they might apply
I_Q04d	I like learning new things
I_Q04h	When I come across something new, I try to relate it to what I already know
I_Q04j	I like to get to the bottom of difficult things
I_Q04l	I like to figure out how different ideas fit together
I_Q04m	If I don't understand something, I look for additional information to make it clearer

References

Alabdulkareem, A., Frank, M.R., Sun, L., AlShebli, B., Hidalgo, C., Rahwan, I., 2018. Unpacking the polarization of workplace skills. Science Advances 4.

Avesani, P., Giunchiglia, F., Yatskevich, M., 2005. A large scale taxonomy mapping evaluation, in: International Semantic Web Conference, Springer. pp. 67–81.

Baroni, M., Dinu, G., Kruszewski, G., 2014. Donât count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors, in: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 238–247.

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.

Boselli, R., Cesarini, M., Marrara, S., Mercorio, F., Mezzanzanica, M., Pasi, G., Viviani, M., 2018a. Wolmis: a labor market intelligence system for classifying web job vacancies. J. Intell. Inf. Syst. 51, 477–502.

Boselli, R., Cesarini, M., Mercorio, F., Mezzanzanica, M., 2017. Using machine learning for labour market intelligence. ECML PKDD 2017: Machine Learning and Knowledge Discovery in Database , 330–342.

Boselli, R., Cesarini, M., Mercorio, F., Mezzanzanica, M., 2018b. Classifying online job advertisements through machine learning. Future Generation Computer Systems , 319–328.

Camacho-Collados, J., Pilehvar, M.T., 2018. From word to sense embeddings: A survey on vector representations of meaning. JAIR 63.

CEDEFOP, 2014. Real-time labour market information on skill requirements: feasibility study and working prototype". https://goo.gl/qNjmrn.

CEDEFOP, 2016. Real-time labour market information on skill requirements: Setting up the eu system for online vacancy analysis. https://goo.gl/5FZS3E.

Colace, F., Santo, M.D., Lombardi, M., Mercorio, F., Mezzanzanica, M., Pascale, F., 2019. Towards labour market intelligence through topic modelling, in: Proceedings of the 52nd Hawaii International Conference on System Sciences (HICSS), pp. 5256–5265. URL: http://hdl.handle.net/10125/59962.

Collobert, R., Weston, J., 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning, in: The 25th International Conference on Machine Learning, p. 160â167.

Colombo, E., Mercorio, F., Mezzanzanica, M., 2019a. Ai meets labor market: Exploring the link between automation and skills. Information Economics and Policy 47. URL: http://www.sciencedirect.com/science/article/pii/S0167624518301318, doi:https://doi.org/10.1016/j.infoecopol.2019.05.003.

Colombo, E., Mercorio, F., Mezzanzanica, M., 2019b. AI meets labor market: exploring the link between automation and skills. Information Economics and Policy 47.

Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 .

EuroStat, 2020. Towards the european web intelligence hub â european system for collection and analysis of online job advertisement data (wih-oja), available at https://tinyurl.com/y3xqzfhp.

Euzenat, J., Loup, D., Touzani, M., Valtchev, P., 2004. Ontology alignment with ola, in: Proc. 3rd ISWC2004 workshop on Evaluation of Ontology-based tools (EON), No commercial editor.. pp. 59–68.

Fleming, M., Clarke, W., Das, S., Phongthiengtham, P., Reddy, P., 2019. The future of work: How new technologies are transforming tasks .

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

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

Giabelli, A., Malandri, L., Mercorio, F., Mezzanzanica, M., 2022. Weta: Automatic taxonomy alignment via word embeddings. Computers in Industry 138, 103626.

Giabelli, A., Malandri, L., Mercorio, F., Mezzanzanica, M., Seveso, A., 2020b. NEO: A tool for taxonomy enrichment with new emerging occupations, in: The Semantic Web - ISWC 2020 - 19th International Semantic Web Conference, Springer. pp. 568–584. URL: https://doi.org/10.1007/978-3-030-62466-8_35, doi:10.1007/978-3-030-62466-8_35.

Giabelli, A., Malandri, L., Mercorio, F., Mezzanzanica, M., Seveso, A., 2020c. Neo: A tool for taxonomy enrichment with new emerging occupations, in: International Semantic Web Conference, Springer. pp. 568–584.

Giabelli, A., Malandri, L., Mercorio, F., Mezzanzanica, M., Seveso, A., 2021a. Neo: A system for identifying new emerging occupation from job ads, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 16035–16037.

Giabelli, A., Malandri, L., Mercorio, F., Mezzanzanica, M., Seveso, A., 2021b. Neo: A system for identifying new emerging occupation from job ads, in: The 35th AAAI Conference on Artificial Intelligence - Demo Track.

Giabelli, A., Malandri, L., Mercorio, F., Mezzanzanica, M., Seveso, A., 2021c. Skills2job: A recommender system that encodes job offer embeddings on graph databases. Applied Soft Computing 101, 107049.

Giabelli, A., Malandri, L., Mercorio, F., Mezzanzanica, M., Seveso, A., 2021d. Skills2job: A recommender system that encodes job offer embeddings on graph databases. Appl. Soft Comput. 101, 107049. URL: https://doi.org/10.1016/j.asoc.2020.107049, doi:10.1016/j.asoc.2020.107049.

Javed, F., Hoang, P., Mahoney, T., McNair, M., 2017. Large-scale occupational skills normalization for online recruitment, in: Twenty-Ninth IAAI Conference.

Jung, J.J., 2008. Taxonomy alignment for interoperability between heterogeneous virtual organizations. Expert Systems with Applications 34, 2721–2731.

Levy, O., Goldberg, Y., 2014. Neural word embedding as implicit matrix factorization, in: Advances in Neural Information Processing Systems, pp. 2177–2185.

Maedche, A., Staab, S., 2001. Ontology learning for the semantic web. IEEE Intelligent systems 16, 72–79.

Malandri, L., Mercorio, F., Mezzanzanica, M., Nobani, N., 2020. Meet: A method for embeddings evaluation for taxonomic data, in: 2020 International Conference on Data Mining Workshops (ICDMW), IEEE. pp. 31–38.

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

Meng, Y., Huang, J., Wang, G., Zhang, C., Zhuang, H., Kaplan, L., Han, J., 2019. Spherical text embedding, in: Advances in Neural Information Processing Systems, pp. 8208–8217.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J., 2013. Distributed representations of words and phrases and their compositionality, in: Advances in neural information processing systems, pp. 3111–3119.

Nguyen, K.A., Köper, M., Walde, S.S.i., Vu, N.T., 2017. Hierarchical embeddings for hypernymy detection and directionality. arXiv preprint arXiv:1707.07273 .

Papoutsoglou, M., Ampatzoglou, A., Mittas, N., Angelis, L., 2019. Extracting knowledge from on-line sources for software engineering labor market: A mapping study. IEEE Access .

Pennington, J., Socher, R., Manning, C., 2014. Glove: Global vectors for word representation, in: EMNLP, pp. 1532–1543.

Turian, J., Ratinov, L., Bengio, Y., 2010. Word representations: a simple and general method for semi-supervised learning, in: ACL.

Turrell, A., Speigner, B., Djumalieva, J., Copple, D., Thurgood, J., 2018. Using job vacancies to understand the effects of labour market mismatch on uk output and productivity .

UK Commission for Employment and Skills, 2015. The importance of LMI, available at https://goo.gl/TtRwvS.

Vinel, M., Ryazanov, I., Botov, D., Nikolaev, I., 2019. Experimental comparison of unsupervised approaches in the task of separating specializations within professions in job vacancies, in: Conference on Artificial Intelligence and Natural Language, Springer. pp. 99–112.

Wang, R.Y., Strong, D.M., 1996. Beyond accuracy: What data quality means to data consumers. Journal of management information systems 12, 5–33.

Wu, T., Qi, G., Wang, H., Xu, K., Cui, X., 2016. Cross-lingual taxonomy alignment with bilingual biterm topic model., in: AAAI, pp. 287–293.

Wu, Z., Zhu, H., Li, G., Cui, Z., Huang, H., Li, J., Chen, E., Xu, G., 2017. An efficient wikipedia semantic matching approach to text document classification. Information Sciences 393, 15–28.