Entity Linking as a Population Mechanism for Skill Ontologies: Evaluating the Use of ESCO and Wikidata

Lino González, Elena García-Barriocanal, and Miguel-Angel Sicilia

Computer Science Department, University of Alcalá, Polytechnic Building. Ctra. Barcelona km. 33.6, 28871 Alcalá de Henares, Madrid, Spain
{lino.gonzalez, elena.garciab, msicilia}@uah.es

Abstract. Ontologies or databases describing occupations in terms of competences or skills are an important resource for a number of applications. Exploiting large knowledge graphs thus becomes a promising direction to update those ontologies with entities of the latter, which may be updated faster, especially in the case of crowd-sourced resources. Here we report a first assessment of the potential of that strategy matching knowledge elements in ESCO to Wikidata using NER and document similarity models available at the Spacy NLP libraries. Results show that the approach may be effective, but the use of pre-trained language models and the short texts included with entities (labels and descriptions) does not result in sufficient quality for a fully automated process.

Keywords: Entity linking · ESCO · Skill ontologies

1 Introduction

Competence/skill databases and knowledge bases are an important component for different applications. Notably, matching training and job offers to candidate profiles requires some kind of expression of the available capacities and the competence or skill gap, that can be used as the basis for building models, e.g. models that match needs in projects [1].

While there are some mature and curated occupational databases that connect job positions to competence components, the lexical resources they contain require in some sectors a constant update to adapt to the changing job market as expressed in job offerings, since the latter are nowadays mostly posted and disseminated in the form of semi-structured text.

A promising approach for the update of those competences is that of reusing other, non-occupational or general-purpose open knowledge bases that are curated as crowd-sourced resources, as for example, Wikipedia-related projects. This could be useful in reducing update time, and enrich the databases, and may also support other applications or related functionality that could exploit the
knowledge graphs that such kind of general purpose resources provide. Entity linking techniques become thus a promising approach to complement expert curation in occupational databases with entities matched in open general purpose knowledge graphs. However, this requires an assessment of the effectiveness of those tools, with regards to the quality and usefulness of the links produced.

In this paper, we present the results of an experiment in entity linking for occupational databases. Concretely, we provide results of the use of state of the art entity linking algorithms between the large effort of ESCO, the *European Skills, Competences, Qualifications and Occupations* ontology [3] and Wikidata. The rationale for using ESCO is that the structure of skills includes fine-grained knowledge items, that are more likely to produce matches that may be useful in bringing more elements to the database. For example, a match of a programming language or some concrete industrial machine may be used to extract more potential knowledge items by traversing Wikidata relations, and hopefully, some of them would reflect novel or recent skills that have been incorporated to Wikidata as part of the continuous process of crowd-sourcing by volunteer curators.

The rest of this paper is structured as follows. Section 2 provides background information on occupational models and ESCO, and briefly surveys related research. Section 3 describes the materials and methods used and their rationale. Then, results are discussed in Sect. 3. Finally, conclusions and outlook are provided in Sect. 5.

2 Background

Occupational databases containing competences and skills have been developed in the last years, principally as a way to support statistics and policy on the labour market. These databases are typically of a national scope, and follow diverse schemas for the description of competences and skills.

The European Commission is developing ESCO (European Skills, Competences, Qualifications and Occupations) together with stakeholders as employment services, employer federations, trade unions, and professional associations. ESCO [3] is an attempt to provide vocabularies for the labour market, with concepts as subclasses of SKOS concepts\(^1\). It covers three different domains: occupations, knowledge, skills and competences, and qualifications. Here we are concerned with the second. Concretely, we deal with the definition of ESCO skills. The ESCO skills pillar distinguishes between (a) skill/competence concepts and (b) knowledge concepts by indicating the skill type. There is however no distinction between skills and competences. Since skill/competence concepts are usually described as short phrases describing some work-related ability or performance, we focus here on mapping only “knowledge”-type skills, since they contain in many cases proper nouns (e.g. names of computer languages, software tools or machines) that are better candidates for an unambiguous mapping to resources in publicly available knowledge graphs. A central use case for ESCO is matching job offers [2], and for that task, having a rich and updated list of concrete entities is critical.

---

\(^1\) [https://www.w3.org/TR/2009/REC-skos-reference-20090818/](https://www.w3.org/TR/2009/REC-skos-reference-20090818/)
While ongoing editorial work of the ESCO Reference Groups was the primary method for initial content creation, mining external resources is considered a potential method for update. Previous work has already combined ESCO with other models or assets for particular purposes. For example, Sibarani et al. [7] combine ESCO and Schema.org\(^2\) for the task of job market analysis. Shakya and Paudel [6] use ESCO in candidate matching as a schema to integrate disparate data. However, to the best of our knowledge, enriching ESCO with open knowledge graphs has not been addressed in previous work.

3 Materials and Methods

The method for linking skills consisted on three steps: entity recognition, entity linking and extracted candidate entities. A pre-trained Named Entity Recognition (NER) model (described below) was used for the first step. First, the entities obtained were filtered based on the manual inspection of each of the matchings and their type. Then, the filtered skills were matched against a Wikidata dump. Finally, the resulting Wikipedia resources matched were examined manually and a final extraction step involved a search for related instances that were candidates to be added to ESCO.

The NER model provided in SpaCy is based on state of the art neural models [8], that uses convolutional networks and built on GloVe vectors. We used the en\_core\_web\_lg model, based on the large Ontonotes 5\(^3\) corpus comprising various genres of text (news, conversational telephone speech, weblogs, usenet newsgroups, broadcast, talk shows). The NER f-score reported in Spacy documentation is 85.36.

The documents for the matching were the result of the concatenation of the fields preferredLabel, description and altLabels found in the file containing the skills obtained from the ESCO website (version 1.0.3), in English language. Only skills of ESCO type knowledge were used for the matching, since the departure assumption is that names of concrete entities appear in that kind of ESCO resources. A file with the matchings was produced, including information on the matched skill and the text and type of the entity identified.

Entity linking was carried out by disambiguation of entities in the large Wikibase knowledge graph [4].

We used SPARQL queries that match the strings of the terms found, we look for entities whose descriptions are similar to the preferredLabel and description fields extracted from ESCO. A direct FILTER query on labels is not feasible as the queries timeout, however, Wikibase provides a way of using the MediaWiki API, with all labels in words indexed as in the following query (where \(<\text{label}>\) is a parameter):

\[
\text{SELECT ?item ?label ?itemDescription} \\
\text{WHERE{}}
\]

\(^2\) https://schema.org/.
\(^3\) https://catalog.ldc.upenn.edu/LDC2013T19.
However, the above query produces many false positives that have to be disambiguated. The approach was based on document similarity using the same Spacy word embedding mentioned for the NER task. The documents compared were the skill preferred label concatenated with the description on one hand, and the Wikibase label concatenated with the description on the other (all in their English versions).

4 Results and Discussion

4.1 Recognition of Entities

The results of the NER step produced 2520 matches over 962 different skills. The average number of matches per skill is 2.6, with 5.0 as 90% quantile. The matches were distributed across reuse levels in the following way: around 20% cross-sector and 56% sector specific, near 15% transversal, and less than 1% occupation-specific.

Regarding Spacy entity types\(^4\), 51% correspond to ORG, i.e. “companies, agencies, institutions, etc.”, 14% to NOPR (“nationalities or religious or political groups.”) and 7% to PERSON. Others between 1% and 5% in descending order of appearance are PRODUCT (“objects, vehicles, foods, etc. – not services”), CARDINAL, LANGUAGE, GPE (“countries, cities, states.”), LAW (“named documents made into laws.”), LOC (“non-GPE locations, mountain ranges, bodies of water.”), DATE and ORDINAL.

The mappings were manually evaluated and labelled according to the following values:

- **RELEVANT** for matches that appear correct and relevant as skills.
- **IRRELEVANT** for matches that appear correct but are not relevant as skills. These include for example dates, cardinals or ordinals.
- **INCORRECT** for matches that appear errors of the NER model.

\(^4\) https://spacy.io/api/annotation#named-entities.
Table 1. Types of entities found and description

| Entity type | Percentage | Description                                      |
|-------------|------------|--------------------------------------------------|
| ORG         | 51%        | Companies, agencies, institutions, etc.          |
| NOPR        | 14%        | Nationalities or religious or political groups   |
| PERSON      | 7%         | Person                                           |
| OTHERS      | 1%-5%      | Product, cardinal, language, gpe, law, loc, date, ordinal |

Further, each of the correct matches was evaluated with regards to the entity type as CORRECT or INCORRECT.

A total of 683 matches were labelled as RELEVANT, 80% sector specific, 13% cross-sector and only 6% occupation specific. Of those, only 264 were unique matches.

4.2 Entities Linked

From the collection of relevant skills, the process of automated mapping with Wikibase was carried out. The manual inspection of the results showed correct matches for around 50% of the distinct entities relevant coming from the NER process.

However, the results included a significant amount of false positive links, for a cosine similarity threshold of 0.8 (this was adjusted manually). These include for example:

- Resources about the skill, as in the case of books or articles. An example is Wikidata Q56114813 which is an article titled “A principled approach to operating system construction in Haskell” and has as the date of this writing almost no descriptive elements in Wikidata. This is a match for the programming skill in Haskell contained in ESCO.
- Errors in disambiguation of entities, as in the case of the match for Haskell to Oxford’s PhD thesis Avenues of art history: recent developments in English art history, with special reference to the works of Francis Haskell and their possible application to the study of Chinese art history Q59574327.

The inspection of cases shows that most of incorrect matches are to Wikidata entities lacking description or with short description texts. This suggests some filtering of potential matches based on the amount of lexical descriptions available to increase accuracy.

4.3 Extracted Candidate Entities

The inspection of the entities correctly linked resulted in the following categories of elements, when traversing a generalization up in Wikidata. The path to more general classes is traversing up the instance-of property (P31) or the subclass of property (P279).
- **programming language** (Q9143), as Erlang, Haskell or Java.
- **methodology** (Q185698). It should be noted that this is defined generically as “scientific method of accumulating data” as Kanban.
- **software** (Q7397) as Apache Tomcat.

However, the way of incorporating new instances is also not straightforward, since it can follow different paths. For example, the Haskell language is an instance of eight other terms, some of them controversial as interpreted language or lazy evaluation.

The lack of consistent semantic use of Wikidata is thus a problem that makes needs to be addressed separately. Figure 1 shows the graph representation of the instance of relations for Haskell. The graph illustrates the semantic consistence problems of the resource.

![Fig. 1. Graph with the instance of relations of entity Haskell](image)

Semantic inconsistencies hamper the possibility of automatically populating the ESCO skill model with terms obtained by property traversal. Other knowledge graphs with more constrained semantics may be an alternative, or some kind of manual pre-selection of traversal paths need to be used to guarantee the quality of the updates. This aspect of the accuracy and trustworthiness of Wikidata is under-researched according to [5], and is of special importance in tasks that attempt to bridge manually curated resources as ESCO with community-curated resources.

### 5 Conclusions and Outlook

We have reported the use of pre-trained language models for the task of extracting and then linking entities from the ESCO database of concrete skills to Wikidata terms for the specific purpose of extracting potential additional knowledge items to be integrated in ESCO. The performance of state of the art pre-trained
models produces a small set of relevant matches for the use case, that have to be manually filtered.

While the results produce a limited number of mappings and is prone to errors of different kinds, the extraction of related knowledge elements are still potentially relevant, and may be considered as a promising mechanism to keep the skill database updated, as a complement to manually curated additions. However, this also requires some kind of quality control on the entities being extracted from Wikidata, since its crowdsourced nature entails a degree of errors that prevent the direct inclusion of new knowledge items.

Future work may proceed in different directions. The limited success of the results in the NER task suggest that an application-specific NER model may be required for the domain of the labour market. In a different direction, the errors in the entity linking with Wikidata may be mitigated by exploiting the relations inside the knowledge graph as a lexical resource.

References

1. Bandini, S., Mereghetti, P., Merino, E., Sartori, F.: Case–based support to small–medium enterprises: the Symphony Project. In: Basili, R., Pazienza, M.T. (eds.) AI*IA 2007. LNCS (LNAI), vol. 4733, pp. 483–494. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-74782-6_42
2. Colombo, E., Mercorio, F., Mezzanzanica, M.: AI meets labor market: Exploring the link between automation and skills. Inf. Econ. Policy 47, 27–37 (2019)
3. Le Vrang, M., Papantoniou, A., Pauwels, E., Fannes, P., Vandensteen, D., De Smedt, J.: ESCO: boosting job matching in Europe with semantic interoperability. Computer 47(10), 57–64 (2014)
4. Mora-Cantallops, M., Sánchez-Alonso, S., García-Barriocanal, E.: A systematic literature review on Wikidata. Data Technol. Appl. 53(3), 250–268 (2019)
5. Piscopo, A., Simperl, E.: What we talk about when we talk about Wikidata quality: a literature survey. In: Proceedings of the 15th International Symposium on Open Collaboration, pp. 1–11 (2019)
6. Shakya, A., Paudel, S.: Job-candidate matching using ESCO ontology. J. Inst. Eng. 15(1), 1–13 (2019)
7. Sibarani, E.M., Scerri, S., Morales, C., Auer, S., Collarana, D.: Ontology-guided job market demand analysis: a cross-sectional study for the data science field. In: Proceedings of the 13th International Conference on Semantic Systems, pp. 25–32 (2017)
8. Strubell, E., Verga, P., Belanger, D., McCallum, A.: Fast and accurate entity recognition with iterated dilated convolutions. arXiv preprint arXiv:1702.02098 (2017)