JoTA: Aligning Multilingual Job Taxonomies through Word Embeddings (Student Abstract)

Anna Giabelli[2, 3], **Lorenzo Malandri**[1, 3], **Fabio Mercorio**[1, 3], **Mario Mezzanzanica**[1, 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, Univ. of Milano-Bicocca, Milan, Italy
{anna.giabelli, lorenzo.malandri, fabio.mercorio, mario.mezzanzanica}@unimib.it

Abstract

We propose JoTA (Job Taxonomy Alignment), a domain-independent, knowledge-poor method for automatic taxonomy alignment of lexical taxonomies via word embeddings. JoTA associates all the leaf terms of the origin taxonomy to one or many concepts in the destination one, employing a scoring function, which merges the score of a hierarchical method and the score of a classification task. JoTA is developed in the context of an EU Grant aiming at bridging the national taxonomies of EU countries towards the European Skills, Competences, Qualifications and Occupations taxonomy (ESCO) through AI. The method reaches a 0.8 accuracy on recommending top-5 occupations and a wMRR of 0.72.

Introduction

Lexical taxonomies are a natural way of expressing the semantic relationships between words and concepts through IS-A relations, and they are the mainstay of several downstream applications. Usually, within a single domain, there are multiple taxonomies, thus the problem of mapping them, when needed, is of primary importance.

To develop our automated taxonomy alignment method, we resort to *distributed semantics*, a branch of linguistics based on the hypothesis that words occurring in similar contexts tend to have a similar meaning. In distributed semantics, words are represented by semantic vectors that are derived from a text corpus, using neural network training. Semantic word vectors have empirically shown to capture linguistic regularities from texts (Mikolov et al. 2013), demonstrating their ability to enrich existing knowledge structures as well (Ristoski and Paulheim 2016). The contribution of this work is twofold:

- We propose a novel method for taxonomy alignment using word embeddings and domain knowledge;
- We apply JoTA in the context of a real-world project aimed at aligning the European national resources with the European labor market taxonomy ESCO (EURES 2019).

As far as we know, JoTA is the first approach that exploits distributional semantic and context information to perform taxonomy alignment. Moreover, we perform an intrinsic evaluation of the selected embedding model based on the structure of the taxonomy itself.

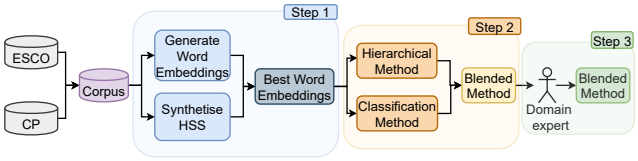

Figure 1: A representation of the JoTA workflow.

Formalization. A taxonomy $\mathcal{T}$ is a 4-tuple $\mathcal{T} = (\mathcal{C}, \mathcal{W}, \mathcal{H}^c, \mathcal{F})$, where $\mathcal{C}$ is a set of concepts $c \in \mathcal{C}$ (aka, nodes); $\mathcal{W}$ is a set of words belonging to the domain, and 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; finally, $\mathcal{F}$ is a directed binary relation mapping words into concepts. Given an origin taxonomy $\mathcal{T}_o$ and a destination taxonomy $\mathcal{T}_d$, the goal of JoTA 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 $\mathcal{S}$.

How Does JoTA Work?

The approach used to align $\mathcal{T}_o$ and $\mathcal{T}_d$ is mainly composed by the steps shown in Fig.1. 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 $c_d \in \mathcal{C}_p^d$. The last step consists of validating the suggestions because the utility of JoTA is the help it provides to the domain experts, narrowing the choices for the alignment.

Step 1: Generate and Evaluate Embeddings. The main goal of the first step of JoTA is to induce a vector representation of taxonomic terms that represent as much as possible the similarity of words within the taxonomy.

To accomplish that, we resort to HSS, a measure of pairwise semantic similarity in taxonomies developed in (Giabelli et al. 2021), which measures semantic similarity in a taxonomy based on the structure of the hierarchy itself, preserving the semantic similarity intrinsic to the taxonomy.

In this first step, we (i) generate word embeddings through a state of the art method; we (ii) compute the HSS of terms

in $\mathcal{T}_o$ and $\mathcal{T}_d$, and we (iii) select the embeddings for which the correlation between the cosine similarity between taxonomic terms and their HSS is maximized for $\mathcal{T}_o$ and $\mathcal{T}_d$.

Step 2: Taxonomy Alignment Method. Our methodology 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 specialized concepts in $\mathcal{T}_d$, i.e. those which are at the lowest level p, $\mathcal{C}_p^d$. 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.

Hierarchical approach. For each $w_o \in \mathcal{W}^o$, the set of words of the origin taxonomy, we create a list that contains the cosine similarity between w_o and each element in $w \in \mathcal{W}^d$. Then, we select the n words with the highest similarity for each $w_o \in \mathcal{W}^o$, and lastly we exploit their hierarchical concepts in $\mathcal{C}_p^d$, selecting the n concepts with the highest similarity.

Classification approach. This approach relies on a multiclass classification task: we have $\mathcal{C}_p^d$ as the target variable, which is the more specific concept level, and the word embeddings associated to each $w_o \in \mathcal{W}^o$ as the independent variable. For each word w_o, at the end of the classification process, we consider the prediction scores and select the n top-scored level p concepts.

Blended approach. This part consists in blending the results obtained respectively from the hierarchical method and the classification one. First, for each $w_o \in \mathcal{W}^o$, we store the shared matches of the two methods. Then we complete the lists of n suggestions (with $n = 5$) considering some recommendations from the hierarchical approach and some from the classification one, with a preference for the latter since it obtains better results (see Tab.1).

Step 3: Evaluation of the Suggestions. The usefulness of JoTA 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. Thus, the last step consists of the validation of the suggestions provided to complete the alignment procedure.

Results and Concluding Remarks

We apply the taxonomy alignment process to bridge $\mathcal{T}_o$ - the Italian taxonomy CP[1] - to $\mathcal{T}_d$ - the European taxonomy ESCO[2]. JoTA employs FastText for training for embeddings[3], while the classification task is performed employing a 2-layer neural network[4]. For each $c_o \in \mathcal{T}_o$, we examine their suggested matches with concepts in $\mathcal{T}_d$ to assess the correctness of the method in comparison with the mapping between $\mathcal{T}_o$ and $\mathcal{T}_d$ validated by a group of domain experts.

For the evaluation, we consider the top-5 Accuracy and the MRR (Mean Reciprocal Rank) because the taxonomy alignment solution is seen as a ranking problem:

$$MRR = \frac{1}{|\mathcal{W}^o|} \sum_{i=1}^{|\mathcal{W}^o|} \cdot \begin{cases} \frac{1}{rank_i} & \textit{if i has a correct suggestion} \\ 0 & \textit{otherwise} \end{cases}$$

We also use the $wMRR$ (Bar-Yossef and Kraus 2011), a weighted version of the MRR that considers also the hypernyms of the suggestions:

$$wMRR = \frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \cdot \begin{cases} \frac{1}{rank_i} & \textit{if i has a correct suggestion} \\ \frac{3}{4}\frac{1}{rank_i} & \textit{if i has a correct level 3 hypernym} \\ \frac{2}{4}\frac{1}{rank_i} & \textit{if i has a correct level 2 hypernym} \\ \frac{1}{4}\frac{1}{rank_i} & \textit{if i has a correct level 1 hypernym} \\ 0 & \textit{otherwise} \end{cases}$$

Method	top-5 Accuracy	MRR	$wMRR$
Hierarchical approach	0.76	0.63	0.69
Classification approach	0.77	0.64	0.71
Blended approach	0.8	0.66	0.72

Table 1: The results of top-5 Accuracy , MRR, wMRR.

The results of the evaluation are shown in Tab. 1. The blended method achieves the best performances.

Conclusion and Future Outlook. This research is framed in the context of an EU grant that aims at aligning the ESCO taxonomy with national labor market taxonomies through AI (EURES 2019). We created JoTA, a methodology for automatic taxonomy alignment of lexical taxonomies through word embeddings. JoTA associates all the leaf terms of the origin taxonomy to one or many concepts in the destination one, merging the results of a hierarchical method based on cosine similarity and the results of a classification task. We applied JoTA in the context of a real-world project, aligning the Italian taxonomy CP to the European Taxonomy ESCO.

The proposed approach is implemented in the labor market domain but is domain-independent. Our results show that JoTA reached a 0.8 accuracy on recommending top-5 occupations and a wMRR of 0.72.

[1]http://professioni.istat.it/cp2011/

[2]https://tinyurl.com/sv4squr

[3]Best model found through HSS with: *algorithm*=cbow, *size*= 150, *epochs*= 100 and *learning rate*= 0.05.

[4]5-folds cross-validation, 2-layer NN with a ReLU activation function for the hidden layers, categorical cross-entropy as loss function, RMSprop as an optimizer, 100 epochs, batch size of 64.

References

Bar-Yossef, Z.; and Kraus, N. 2011. Context-sensitive query auto-completion. In *WWW*, 107–116.

EURES. 2019. A Data Driven Bridge Towards ESCO using AI Algorithms, granted by EURES (call EaSI-EURES VP/2019/010).

Giabelli, A.; Malandri, L.; Mercorio, F.; Mezzanzanica, M.; and Seveso, A. 2021. Skills2Job: A recommender system that encodes job offer embeddings on graph databases. *Applied Soft Computing*, 101: 107049.

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

Ristoski, P.; and Paulheim, H. 2016. Rdf2vec: Rdf graph embeddings for data mining. In *ISWC*, 498–514.