A Human-AI Teaming Approach for Incremental Taxonomy Learning from Text

Andrea Seveso, Fabio Mercorio, Mario Mezzanzanica
University of Milano-Bicocca
andrea.seveso@unimib.it

Abstract

Taxonomies provide a structured representation of semantic relations between lexical terms, acting as the backbone of many applications. The research proposed herein addresses the topic of taxonomy enrichment using an ”human-in-the-loop” semi-supervised approach. I will be investigating possible ways to extend and enrich a taxonomy using corpora of unstructured text data . The objective is to develop a methodological framework potentially applicable to any domain.

1 Introduction

Over the past several years, the growth of web services has been making available a massive amount of structured and unstructured data in different domains. Much of this knowledge is available as unstructured text, therefore not easily interpretable by automated systems. The ability of extracting valuable knowledge from these resources often 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.

Unlike the automated construction of new taxonomies from scratch, which is a well-established research area [Wang *et al.*, 2017], the augmentation of existing hierarchies is gaining in importance, given its relevance in many practical scenarios (see, e.g. [Vedula *et al.*, 2018]). 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. This process relies only on human knowledge, making the process costly and time-consuming. Over time, just as languages and domains change and update themselves, so it is necessary to keep the taxonomies updated with respect to the reference field. Experts can be helped in their decision making process by semi-automatic systems, which offer them suggestions and explanations in order to improve their work in terms of both speed and quality. Since human validation is always required, we consider the taxonomy enrichment process as a semi-automatic system in a human-in-the-loop approach.

The main objective of the project is to study and define new methodologies and models based on Machine Learning for the automatic enrichment of taxonomies.

1.1 Objectives and Research Questions

The main research questions that will be addressed in this project are the following: (i) How can semi-supervised machine learning techniques help significantly in extracting terms / relations from large domain text corpus? (ii) How then can human intervention in the process of enriching and updating taxonomies be minimized? (iii) What is the best method to automatically suggest domain relationships to an experienced user, so that they can fully understand the motivations behind the suggestion?

1.2 Previous Work

Most of the work in the area of automated taxonomy enrichment relies heavily on or domain specific knowledge or lexical structures specific to an existing resource, like the WordNet synset [Toral and Monachini, 2008]. In recent years few scholars tried to overcome those limitations developing methodologies for the automated enrichment of generic taxonomies. Wang et al. [Wang *et al.*, 2014] 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. 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. [Vedula *et al.*, 2018] 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.

Those methods learn a word vector representation of the taxonomy, without linking it to an external corpus of web data, while I would like to 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 Methods

The intuition of the research project is to define a methodology that combines the potential of two techniques - not directly related to taxonomy enriching - but widely used in AI and data science: (i) word-embeddings [Schnabel *et al.*, 2015], which they allow to represent terms and relations in an n-dimensional word space, and to perform vector operations between them; (ii) eXplainable AI (XAI), to make the reasons behind the suggestion and / or learning understandable to the domain expert user;

The objective therefore is the creation of a methodology that allows the identification of terms and relations through the analysis of a large amount of textual data. Each of the automatically synthesized elements will then be subjected to *validation* by expert users, in a human-in-the-loop approach, with the relative explanation so that the user can approve the suggestion or correct the algorithm, having understood the underlying reasons. In fact, embedding representations have many advantages over the classic BOW (Bag of Words) representations: similar terms are close in the vector space, moreover, surprisingly also other semantic relationships such as geographical connections or masculine-feminine terms are easily recoverable through algebraic operations on vectors. In this approach, Machine Learning techniques are used to exploit latent features hidden in text data in order to enrich taxonomies, produce classifiers and recommendation systems.

Human interventions will be limited, and restricted to activities such as active learning, manually assigning labels and evaluating model predictions; actions not always necessary but often useful to improve training datasets.

As for the evaluation of the recommendation, it is essential to exploit the eXplainable AI [Guidotti *et al.*, 2018] techniques to make the reasons behind the suggestion understandable to the domain user. The use of XAI allows the domain expert to better understand the operation of the taxonomy enrichment system, increasing transparency and decreasing the effort required of the user.

3 Ongoing Research

3.1 NEO: A Tool for Taxonomy Enrichment with New Emerging Occupations

NEO [Giabelli *et al.*, 2020] is a taxonomy enrichment tool that aims to enrich the standard occupation and skill taxonomy (ESCO) with new occupation terms extracted from Online Job Vacancies (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, and has been deployed on a set of 2M+ real OJVs collected from UK in 2018 within the project.

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.

NEO synthesised and evaluated more than 240 vector space models, identifying 49 novel occupations, 43 of which (88%) were validated as novel occupations by a panel of 10 experts involved in the validation of the system. Two statistical hypothesis tests confirmed the positive correlation between the novel metrics proposed in NEO and the user judgements, and this suggests that the system is able to accurately identify novel occupations and to suggest an IS-A relation within the taxonomy.

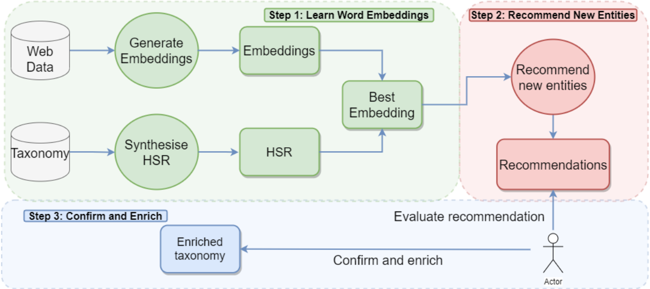

Figure 1: A representation of the NEO workflow highlighting the main modules.

4 Future Contributions

Future work is aimed at at generalizing the approach of NEO for any domain, transforming it into a general framework for enriching taxonomies. Another point of interest are knowledge graphs, that could be used to model the taxonomies. Graphs allow adding new properties to nodes and relationships between nodes, meaning that it is a flexible way for representing and enriching domain knowledge.

References

[Giabelli *et al.*, 2020] Anna Giabelli, Lorenzo Malandri, Fabio Mercorio, Mario Mezzanzanica, and Andrea Seveso. NEO: A tool for taxonomy enrichment with new emerging occupations. In *International Semantic Web Conference*. Springer, 2020.

[Guidotti *et al.*, 2018] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):93, 2018.

[Schnabel *et al.*, 2015] Tobias Schnabel, Igor Labutov, David Mimno, and Thorsten Joachims. Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 298–307, 2015.

[Toral and Monachini, 2008] Antonio Toral and Monica Monachini. Named entity wordnet. In *LREC*, 2008.

[Vedula *et al.*, 2018] Nikhita Vedula, Patrick K Nicholson, Deepak Ajwani, Sourav Dutta, Alessandra Sala, and Srinivasan Parthasarathy. Enriching taxonomies with functional domain knowledge. In *SIGIR*, 2018.

[Wang *et al.*, 2014] Jingjing Wang, Changsung Kang, Yi Chang, and Jiawei Han. A hierarchical dirichlet model for taxonomy expansion for search engines. In *WWW*, pages 961–970, 2014.

[Wang *et al.*, 2017] Chengyu Wang, Xiaofeng He, and Aoying Zhou. A short survey on taxonomy learning from text corpora: Issues, resources and recent advances. In *EMLP*, pages 1190–1203, 2017.