Automatic Taxonomy Generation: A Use-Case in the Legal Domain

Cécile Robin, James O'Neill, Paul Buitelaar

*Insight Centre for Data Analytics, NUI Galway,
Galway, Ireland
{cecile.robin,james.oneill,paul.buitelaar}@insight-centre.org

Abstract

A key challenge in the legal domain is the adaptation and representation of the legal knowledge expressed through texts, in order for legal practitioners and researchers to access this information more easily and faster to help with compliance related issues. One way to approach this goal is in the form of a taxonomy of legal concepts. While this task usually requires a manual construction of terms and their relations by domain experts, this paper describes a methodology to automatically generate a taxonomy of legal noun concepts. We apply and compare two approaches on a corpus consisting of statutory instruments for UK, Wales, Scotland and Northern Ireland laws.

1. Introduction

A quicker understanding and comprehension of legal documents is an imperative for practitioners in the legal sector, who have witnessed a steep increase in legislation since the financial crisis in 2008. This can result in law containing even more ambiguous and complex expressions, which can subsequently lead to damaging non-compliance problems for financial institutions. A fundamental way of arranging the knowledge in legal texts to mitigate such problems is by representing the domain in the form of a taxonomy of legal concepts.

Our proposed approach tackles this issue through the automatic construction of a legal taxonomy, directly extracted from the content of the corpus of legal texts analysed. The idea here is to be able to create a classification based on the field of application of any type of legal documents, and facilitating the maintenance of the versions. This approach would help to track changes in regulations and to keep up-to-date with new ones, making this information easily searchable and browsable.

We compare two systems for automatic taxonomy generation applied to a small corpus of legal documents. First, we provide related work on automatic taxonomy generation in general, and in the legal domain in particular. We then describe the two approaches chosen for our study. Next, we examine the experiments performed with both systems on a subset corpus of the UK Statutory Instruments, providing a comparative analysis of the results, before providing suggestions for future work.

2. Related Work

Generic domain approaches Taxonomy construction is a relatively unexplored area, however (Bordea et al., 2016) organised a related task in SemEval-2016: TExEval, where the aim was to connect given domain-specific terms in a hyperonym-hyponym manner (relation discovery), and to construct a directed acyclic graph out of it (taxonomy construction). Only one out of the 6 teams produced a taxonomy, focusing thus more on the relation discovery step. Most systems relied on WordNet (Fellbaum, 1998) and Wikipedia resources.

(Sujatha et al., 2011) did a structured review of all the main types of approaches involved in the task of automatic taxonomy construction. It includes the use of WordNet, Natural Language Processing (NLP) techniques, tags from Web resources, or large external corpora. However, WordNet is a generic lexical resource and is not fitted for the legal language whose definitions and semantic relations are very specific to the domain, as well as constantly evolving. As for external annotated data, these are often non available and also non dynamic resources, therefore not well suited for our task.

(Ahmed and Xing, 2012) use Dynamic Hierarchical Dirichlet Process to track topics over time, documents can be exchanged however the ordering is intact. They also applied this to longitudinal *Neural Information Processing Systems* (NIPS) papers to track emerging and decaying topics (worth noting for tracking changing topics around compliance issues).

(Pocostales, 2016) described a semi-supervised method for constructing an *is-a* type relationship (i.e hypernym-hyponym relation) that uses *Global Vectors for Word Representation* (GloVe) vectors trained on a Wikipedia corpus. The approach attempts to represent these relations by computing an average offset for a set of 200 hypernym-hyponym vector pairs (sampled from *WordNet*). This offset distance is then added to each term so that hypernym-hyponyms relations could be identified outside of the 200 pairs which are averaged.

In the legal domain Most work on taxonomy generation in the legal domain has involved manual construction of concept hierarchies by legal experts (Buschettu et al., 2015). This task, besides being both tedious and costly in terms of time and qualified human resources, is also not easily adaptable to changes. Systems for automatic legal-domain taxonomy creation have on the contrary received very low attention so far. Only (Ahmed et al., 2002) worked on a similar task, and developed a machine learning-based system for scalable document classification. They constructed a hierarchical topic schemes of areas of laws and used proprietary methods of scoring and ranking to classify documents. However, this work has been deposited as a patent and is not freely available.

We will now introduce our two chosen methodologies, based on NLP and clustering techniques.

3. Automatic Taxonomy Construction

This section describes the two presented bottom-up approaches to taxonomy generation. We begin with an overview of *Hierarchical Embedded Clustering*.

3.1. Hierarchical Embedded Clustering

Hierarchical Embedding Clustering (HEC) is an agglomerative clustering method that we have used for encoding noun phrase predict vectors (i.e Skipgram trained vectors). We first identify noun phrases in the text by extracting bigrams and retaining only the pairs that contain nouns, determined by the NLTK Maximum Entropy PoS tagger[1]. This is followed by a filtering stage, whereby the top *n=5377* noun phrases are chosen, based on the highest Pointwise Mutual Information (PMI) scores within a range chosen through a distributional analysis as shown in Figure 1. In this figure we present the scaled probability distribution (10^2) between noun phrase counts in the range $[10 - 100]$. The dashed line indicates the density, showing that most probability density is lying within the range $[10 - 60]$. This is a well established trend known as *Luhn's law* (Pao, 1978). Thus, we choose a filtering range between 10-150 to allow for good coverage with still some degree of specificity, resulting in $n = 5377$ filtered words and phrases.

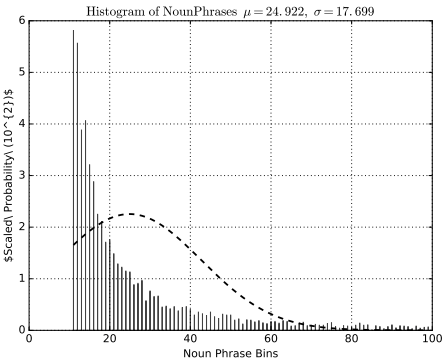

Figure 1: Noun Phrase Distribution

Once the noun phrases are selected, we obtain embedded vectors. Each word within a noun phrase is averaged column-wise, therefore representing a whole noun phrase as a single vector. This was carried out using both the corpus trained legal vector representations and the large scale pretrained vectors provided by GoogleNews[2]. However, we find that since the legal corpus was relatively small in comparison to Google vectors, it did not achieve the same coherency in grouping noun phrases in a hierarchical structure. Therefore, we focus only on results provided by Google's pretrained vectors. HEC is a bottom-up approach for creating taxonomies in the sense that each noun phrase is considered as its own cluster at the leaves, which are then incrementally merged until we arrive at the root.

[1] https://textblob.readthedocs.io/en/dev/_modules/textblob/classifiers.html

[2] https://code.google.com/archive/p/word2vec

3.2. Topical Hierarchy Generation

Saffron is a software tool[3] that aims to automatically construct a domain-specific topic hierarchy using domain modeling, term extraction and taxonomy construction.

Domain Modeling In order to define the domain of expertise of the corpus, *Saffron* first builds a domain model, i.e. a vector of words representing the highest level of generality in this specific domain (Bordea, 2013). Candidate terms are first extracted using feature selection: giving more weight on part-of-speech carrying meaning, and selecting single words (for genericity) represented in at least a 1/4 of the corpus (for enough specificity to the domain). In order to filter the candidate words, (Bordea, 2013) evaluates the coherence of a term within the domain based on (Mimno et al., 2011)'s work on topic coherency, following the assumption that domain terms are more general when related to many specific ones. The domain model created is then used in the next phase for the extraction of topics which will make up the taxonomy.

Term Extraction In the topic extraction phase, intermediate level terms of the domain are sought (as defined in (Bordea et al., 2013)). It involves two approaches: one looking for domain model words in the context of the candidate terms (within a defined span size), and the second using the domain model as a base to measure the lexical coherence of terms by PMI calculation. At the end of this phase, all domain-specific topics have been extracted from the corpus, ready to be included in the taxonomy.

Taxonomy Construction Building connections between the extracted topics is the next step toward the taxonomy construction. Edges are added in the graph for all pairs appearing together in at least three documents, and a generality measure allows to direct edges from generic concepts to more specific ones. A specific branching algorithm, successfully applied for the construction of domain taxonomies in (Navigli et al., 2011) trims the noisy directed graph. This produces a tree-like structure where the root is the most generic topic, and the topic nodes are going from broader parent concepts to narrower children.

3.3. Model Comparison

The two approaches show similarities and dissimilarities. While both systems use a basic term extraction approach for the selection of candidate noun phrases, and PMI for ranking and filtering them, their approach is different. *Saffron* applies PMI to calculate the semantic similarity of the terms to a domain model, while *HEC* uses the outcome of Luhn's cut analysis instead. As for taxonomy construction, both methods construct abstract and loosely related connections for the taxonomy hierarchy, instead of the traditional *is-a* relation type. However, *Saffron* defines a global generality measure using PMI to calculate how closely related a term is to other terms from the domain, following the assumption that generic terms are most often used along with a large number of specific terms. On the contrary, *HEC* relies on agglomerative clustering to detect these relations among embedded vector noun phrases, using cosine similarity as similarity measure. For this step,

[3] http://saffron.insight-centre.org/

Saffron focuses rather on the hierarchy structure at the document level across the texts, whereas *HEC* works directly on all texts within the corpus. This results in abstract concepts at the intermediary levels of the clustering algorithm, and groupings of noun phrases at the leaves. In contrast, *Saffron* provides expressions from these groups at all levels, from the root to the tree.

4. Experimental Setup

This section gives a brief overview of the corpus used in our experiments. The experiments described here are a first step toward the larger objective of generating a taxonomy for legal corpora over a long time scale. We chose to test the two aforementioned approaches first on a subset of the available Statutory Instruments of Great Britain[4]. 41,518 documents have been produced between 2000 and 2016, each year being split in between UK, Scotland, Wales and Northern Ireland. For this experiment, we refine the analysis by selecting the most recent texts (i.e. 2016) of the UK Statutory Instruments (UKSI), that is 838 documents. We don't consider metadata (such as subject matters, directory codes) as they are not always available in legal texts. Furthermore, there is no agreed standard schema definition yet for describing legal documents across different jurisdictions. Our main goal is to compare the results provided by the two different techniques and determine which is the most suitable for the needs previously described, and focusing on the 2016 UK Statutory Instruments corpus eases the comparison towards that objective.

5. Results

In this section we analyse the noun concepts retrieved from each approach and the relations created in the automatically-built hierarchical taxonomy.

Hierarchical Clustering Approach Figure 2 displays the overall results in the form of a greyscale heatmap where noun phrases (rows) and embedding dimension values (columns) are displayed. A filtering phase is performed on the corpus for the HEC approach to clean potential noisy legal domain syntax (such as the references to regulations e.g. "*Regulation EC No. 1370/2007 means Regulation 1370/2007 ...* " which is not meaningful in our case. Noun phrases which appear less than ten times and in less than five documents are also excluded from the analysis, as they are considered too specific and sparse.

The embedded dimensions are reduced representations of the words in an embedding space. Therefore, if the same dimensions of a noun phrase pair both have positively or negatively correlated values in particular dimensions, it means their context is similar in those elements of the vector, meaning that the two noun phrases are related within that given context. From this figure, it can be identified that some noun phrases are merged due to a small number of dimensions being highly correlated in the embedding space, and not necessarily that all dimensions correlated consistently. This means that the noun phrases are very related only in certain contexts, based on the *GoogleNews* corpus

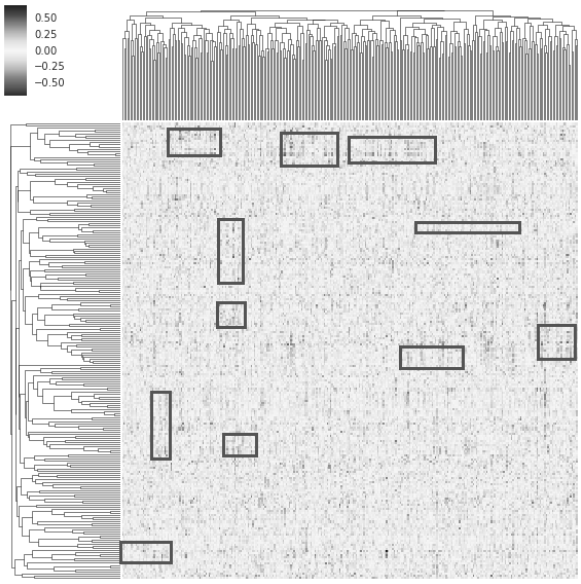

Figure 2: Heatmap of Noun Phrase Vectors (dendrogram)

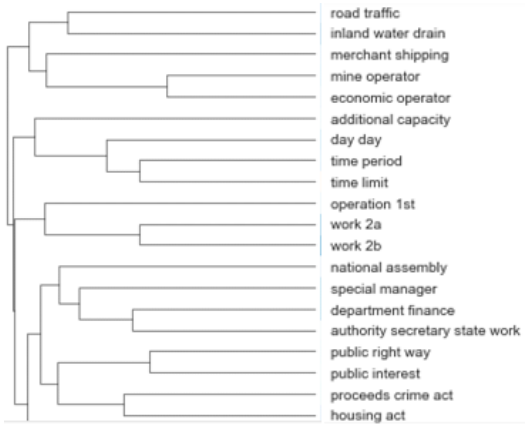

Figure 3: Sample of Hierarchical Clustering of Noun Phrase Embeddings for UK Statutory Instruments

which these vectors have been trained on, but not necessarily appearing together in other contexts. The rectangles within the heatmap aims at pointing out areas within the graph where this is particularly evident.

Figure 3 shows a snapshot of the results obtained from the previous visualization. Here we can see some interesting groups based on semantic relatedness. The *crime act* and *housing act* have merged with *public interest* and *right of way*, which illustrates a topic within the corpus. Likewise, *mine operator* and *economic operator* have been combined with *merchant shipping*. This appears to show an organized relationship of these two noun concepts.

Saffron Approach We visualize the representation of the taxonomy using an open source software platform, Cytoscape [5]. Nodes are topics, and the size of the nodes relates to the number of connections each topic shares with others. Figure 4 illustrates the whole taxonomy generated by *Saffron* for the corpus. Based on this representation, we detect the topics that are the most prominent in the 2016 UK Statutory Instruments, with four major themes

[4] http://www.legislation.gov.uk

[5] http://www.cytoscape.org/

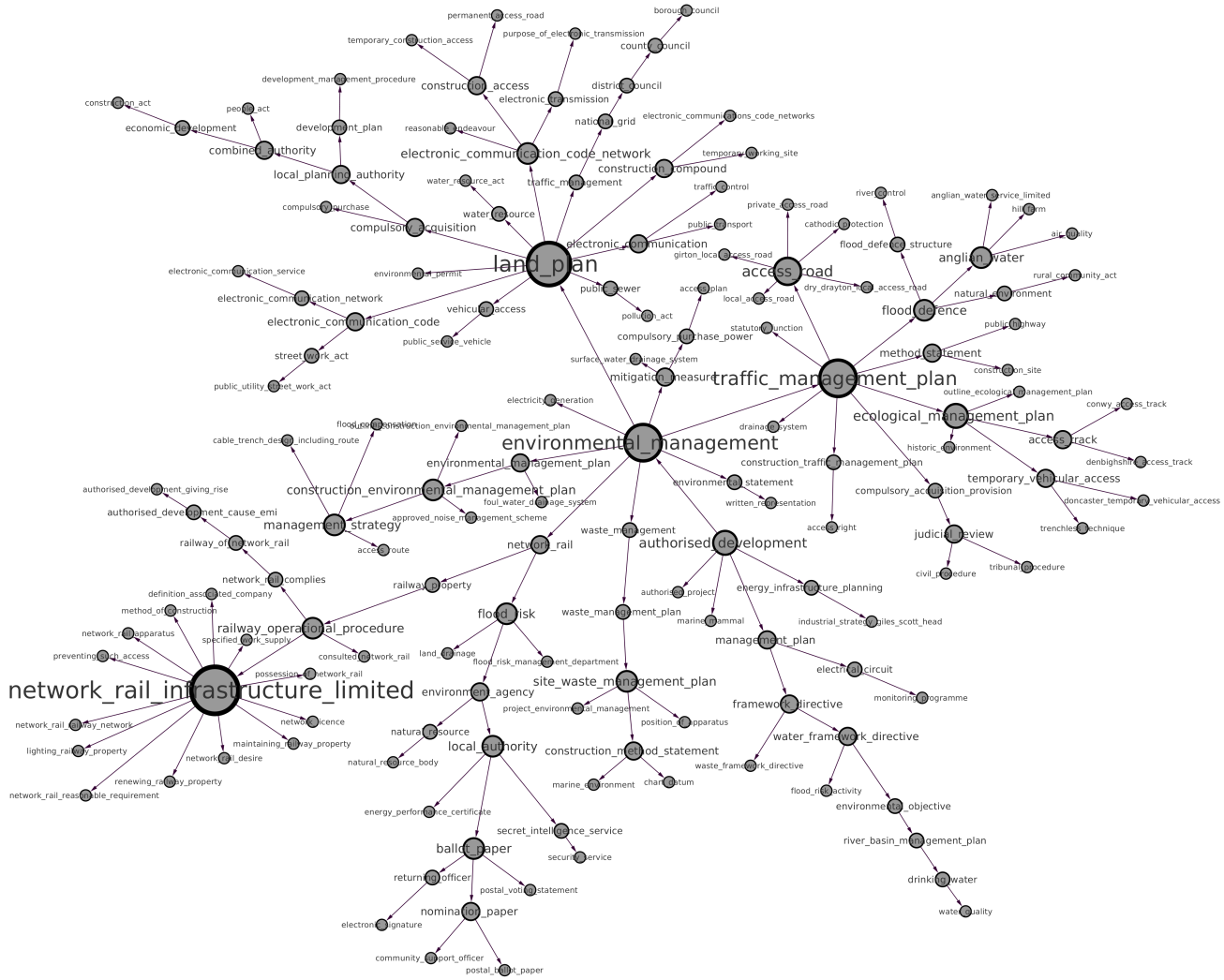

Figure 4: Saffron Taxonomy for the 2016 UK Statutory Instruments

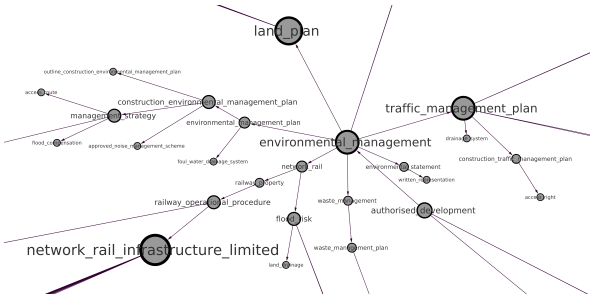

Figure 5: Main Topics from the 2016 UK Ireland Statutory Instruments

Figure 6: *Traffic Management Plan* topic within the UK Statutory Instruments

shown in more detail in Figure 5 (*network rail infrastructure limited*, *land plan*, *environmental management* and *traffic management plan*), included in their clusters of related topics. The proposed approach clearly shows the advantages of the hierarchical structure of the graph, which semantically merges topics from generic concepts to more specific ones, like in the *environmental management* node linking to *environmental management plan*, itself redirecting to *construction environmental management plan*, as we can see in Figure 5.

There is also a clear interest arising from connecting topics that appear together across the documents. This enables relations between concepts which might not be otherwise obvious to a legal practitioner. It would require carrying out an extensive amount of reading within a particular jurisdiction, while still being able to track links between various documents. For example, in Figure 6, we observe that *traffic management plan* is connected within

some regulation about *drainage system*, the accessibility of the road (*access road)*, and is mentioned through the documents alongside with concepts of *ecological management plan*, and *flood defense*. This clearly shows the potential of such semantic processing as an assistance to legal practitioners to identify topics surrounding certain legal issues or for summarizing a whole jurisdiction.

Comparison In the two approaches, both representations display connections between concepts in a different way. The approach of *Saffron*, involving the generation of a domain model, seems to retrieve concepts closer to a topic level than the HEC approach. However, the latter approach also brings up relations between terms worth being further considered. One can argue that both systems highlight different aspects of the legal domain from the same corpus, and allow to detect different behaviours, different relations and can be both useful to a domain expert. Furthermore, both methodologies show the importance of a hierarchical structure compared to a flat representation, as well as the usage of multi-word expressions as opposed to single word ones, which are more ambiguous and too broad for a practical use in such specific domain.

6. Conclusion and Future Work

This work has presented a comparison of two fully automated approaches for identifying and relating salient noun concepts in a taxonomy for the legal domain. The results show coherent groupings of words into legal concepts in both approaches, providing highlights on the emerging topics within the legal corpus. This motivates further research for automatic taxonomy construction to assist legal specialists in various applications. This kind of content management in the legal domain is essential for compliance, tracking change in law and terminology and can also assist legal practitioners in search.

Although both approaches seem to show interesting results in automatic taxonomy construction, there is a considerable difficulty in evaluating such systems in a quantitative way, due to the lack of benchmarks to evaluate taxonomies created for specific domains, and the low agreement between experts on fast changing areas. In (Bordea et al., 2016), the authors evaluated expert agreement on the hierarchical relations between terms. The lowest was shown to be in the Science domain, highlighting the difficulty for experts to get a good overview of a domain which is subject to constant changes. Moreover, their approach to automatically evaluate the resulting hierarchies uses a gold standard taxonomy strictly extracted from *WordNet* (Fellbaum, 1998). This resource is too generic for the intermediate level of terms, on which we are focusing in this approach, specifically to the legal domain (eg. "notice of appeal", "housing allowance", "pension scheme"). However, we plan on carrying out further studies towards a formal representation of concepts within a domain, undertaken by domain experts. This kind of benchmark would establish an evaluation dataset for this domain, where the generated taxonomies are evaluated with taxonomy matching and alignment measures. We also consider establishing an expert user study to evaluate the generated results, with the idea to get legal domain practitioners' views on the practicability of such representation, and the pertinence of the relations established.

Acknowledgements This work has been funded in part by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289 (INSIGHT) and by Enterprise Ireland and the IDA under the Technology Centre Programme [Grant TC-2012-009]

7. References

Ahmed, Amr and Eric P. Xing, 2012. Timeline: A dynamic hierarchical Dirichlet process model for recovering birth/death and evolution of topics in text stream. *CoRR*, abs/1203.3463.

Ahmed, S., T.L. Humphrey, X.A. Lu, J.T. Morelock, J.M. Peck, and J.S. Wiltshire, 2002. System and method for classifying legal concepts using legal topic scheme. EP Patent App. EP20, 000, 952, 140.

Bordea, Georgeta, 2013. *Domain adaptive extraction of topical hierarchies for Expertise Mining*. Ph.D. thesis.

Bordea, Georgeta, Paul Buitelaar, and Tamara Polajnar, 2013. Domain-independent term extraction through domain modelling. In *Proceedings of the 10th International Conference on Terminology and Artificial Intelligence*.

Bordea, Georgeta, Els Lefever, and Paul Buitelaar, 2016. Semeval-2016 task 13: Taxonomy extraction evaluation (texeval-2). In *Proceedings of SemEval-2016*. Association for Computational Linguistics.

Buschettu, Alberto, Giulio Concas, Filippo Eros Pani, and Daniele Sanna, 2015. *A Kanban-Based Methodology to Define Taxonomies and Folksonomies in KMS*. Berlin, Heidelberg: Springer Berlin Heidelberg, pages 539–544.

Fellbaum, Christiane, 1998. *WordNet*. Wiley Online Library.

Mimno, David, Hanna M. Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum, 2011. Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11. Stroudsburg, PA, USA: Association for Computational Linguistics.

Navigli, Roberto, Paola Velardi, and Stefano Faralli, 2011. A graph-based algorithm for inducing lexical taxonomies from scratch. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Vol. 3*, IJCAI'11. AAAI Press.

Pao, Miranda Lee, 1978. Automatic text analysis based on transition phenomena of word occurrences. *Journal of the Association for Information Science and Technology*, 29(3):121–124.

Pocostales, Joel, 2016. Nuig-unlp at semeval-2016 task 13: A simple word embedding-based approach for taxonomy extraction. In *SemEval@ NAACL-HLT*.

Sujatha, R., R. Bandaru, and R. Rao, 2011. Taxonomy construction techniques - issues and challenges. Civil Eng., 2.