TALN-UPF: Taxonomy Learning Exploiting CRF-Based Hypernym Extraction on Encyclopedic Definitions

Luis Espinosa-Anke and **Horacio Saggion** and **Francesco Ronzano**
Tractament Automàtic del Llenguatge Natural (TALN)
Department of Information and Communication Technologies
Universitat Pompeu Fabra
Carrer Tànger, 122-140
08018 Barcelona, Spain
{luis.espinosa, horacio.saggion, francesco.ronzano}@upf.edu

Abstract

This paper describes the system submitted by the TALN-UPF team to SEMEVAL Task 17 (Taxonomy Extraction Evaluation). We present a method for automatically learning a taxonomy from a flat terminology, which benefits from a definition corpus obtained by querying the BabelNet semantic network. Then, we combine a machine-learning algorithm for term-hypernym extraction with linguistically-motivated heuristics for hypernym decomposition. Our approach performs well in terms of vertex coverage and newly added vertices, while it shows room for improvement in terms of graph topology, edge coverage and precision of novel edges.

1 Introduction

Learning semantic relations out of flat terminologies is an appealing task due to its potential application in tasks like Question Answering (Cui et al., 2005; Boella et al., 2014), automatic glossary construction (Muresan and Klavans, 2002), Ontology Learning (Navigli et al., 2011) or Textual Entailment (Roller et al., 2014). Today, in the context of massive web-enabled data, hypernym (is-a) relations are the focus of much research, as they constitute the backbone of ontologies (Navigli et al., 2011). However, one challenge remains open in the automatic construction of knowledge bases that exploit this type of relation. It is unfeasible to have up-to-date semantic resources for each domain, as they are limited in scope and domain, and their manual construction is knowledge intensive and time consuming (Fu et al., 2014).

Given this rationale, Task 17 (Bordea et al., 2015) in the SEMEVAL 2015 set of shared tasks focuses on Taxonomy Extraction Evaluation, i.e. the construction of a taxonomy out of a flat set of terms belonging to one of the four domains of choice (food, chemical, equipment and science). These terms have to be hierarchically organized, and new terms are allowed to be included in the taxonomy. As for evaluation, for each domain, two taxonomies were used as gold standard: One created by domain experts; and one derived from the WordNet taxonomy rooted at the domain node, e.g. food[1]. Finally, evaluation is carried out from two standpoints: (1) The taxonomy topology and the rate of replicated nodes and edges are taken into account when compared to a gold standard taxonomy; and (2) Human experts validated as correct or incorrect a subset of the newly added edges.

In this paper we describe our contribution to this shared task. Our approach relies on a set of definitional sentences for each term, from which term→hypernym relations are extracted using a machine-learning classifier. In a second step, linguistically-motivated rules are applied in order to (1) extract a hypernym candidate when the confidence of the classifier was below a threshold, and (2) decompose multiword hypernyms in more general concepts (e.g. from *coca-cola→carbonated soft drink* to *carbonated soft drink→soft drink* and *soft drink→drink*).

[1]For our domain notation we simply use the name of the domain for manually constructed taxonomies (e.g. "food"), and add the prefix *wn_* for the WordNet taxonomies (e.g. "wn_food").

Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pages 949–954,
Denver, Colorado, June 4-5, 2015.

The remainder of the paper is structured as follows: Section 3 describes the modules of our approach, Section 4 presents and discusses the evaluation procedure as well as results, and finally Section 5 analyzes the performance of our system as well as the difficulties encountered, and suggests potential avenues for future work.

2 Background

Generally, taxonomy learning from text has been carried out either following rule-based or distributional approaches. In terms of rule-based methods reported in the literature, (Hearst, 1992) introduced lexico-syntactic patterns, which were exploited in subsequent work (Berland and Charniak, 1999; Kozareva et al., 2008; Widdows and Dorow, 2002; Girju et al., 2003). Distributional approaches, on the other hand, have become increasingly popular due to the availability of large corpora. Systems aimed at extracting hypernym relations from text have exploited hybrid patterns as word-class lattices (Navigli and Velardi, 2010), syntactic relations as features for an SVM classifier (Boella et al., 2014) or word-embedding-based semantic projections (Fu et al., 2014; Roller et al., 2014). Inspired by the reported success in the latter methods, we opted for combining syntactic patterns with machine learning to extract hypernyms from domain sentences.

3 Method

This section describes the main modules that constitute our taxonomy learning system.

3.1 Definition corpus compilation

We benefit from BabelNet, a very large multilingual semantic network that combines, among other resources, Wikipedia and WordNet (Navigli and Ponzetto, 2010). We get a set of BabelNet synsets associated to each term and for each synset, we extract its definition. In this step we assume that a term's definition appears in the first sentence of its Wikipedia article, which is a regular practice in the literature (see (Navigli and Velardi, 2010) or (Boella et al., 2014)). This step allowed us to compile a domain corpus of definitional knowledge, and thus maximizing the number of relevant terms definitions. However, noise is also introduced in our corpus. For example, given the term *botifarra* (a Catalan type of sausage), we add two definitions to our corpus:

> **Relevant**: Botifarra is a type of sausage and one of the most important dishes of the Catalan cuisine.
>
> **Noisy**: Botifarra is a point trick-taking card game for four players in fixed partnerships played in Catalonia.

3.2 Hypernym Extraction

Given a set of definitional text fragments where the definiendum[2] term is known, i.e. can be extracted from the url of the Wikipedia page, our goal is to tag the tokens of the definition that correspond to one or more hypernyms. To this end, we train a Conditional Random Fields (Lafferty et al., 2001) classifier[3] with the WCL Dataset (Navigli and Velardi, 2010). We argue that CRFs are a valid approach for sequential classification, and particularly for this task, due to their potential to capture prior and posterior token features on the current iteration. The WCL dataset includes near 2000 definitional sentences with terms and hypernyms manually annotated. We preprocess and parse the WCL dataset with a dependency parser (Bohnet, 2010), and then train our classifier with the following set of features.

> **surface**: A word's surface form.
>
> **lemma**: The lemma of the word.
>
> **pos**: The word's part-of-speech.
>
> **head_id**: The id of the word to which the current token depends in a dependency syntactic tree.
>
> **deprel**: Syntactic function of the current word in relation to its head.
>
> **def—nodef**: Whether the current token appears before or after the first verb of the sentence.

[2]The classic components lexicographic *genus-et-differentia* definition are (1) Definiendum (concept being defined); (2) genus (hypernym or immediate superordinate that describes the definiendum); and (3) definiens or cluster of words that differentiate a definiendum from others of its kind.

[3]https://code.google.com/p/crfpp/

term—noterm: Whether the token is part of the definiendum term or not.

Our CRF classifier learns the above word-level features in a word window of $[-2, 2]$. The prediction the classifier must learn follows the classic BIO format, i.e. whether a word is at the beginning of a hypernym phrase, inside or outside. We evaluate this hypernym extraction module on the WCL dataset (Navigli and Velardi, 2010) performing 10-fold cross-validation. It achieves and F-score of 79.86, outperforming existing state-of-the-art systems described in the literature (Navigli and Velardi, 2010; Boella et al., 2014).

Despite the good performance of this module, we observe two potential drawbacks in terms of its fitness for the taxonomy learning task. Firstly, we aim at recovering hypernym candidates even in cases in which they are predicted with low confidence at the classification step. We build on the assumption that all encyclopedic definitions are very likely to include a hypernym, and hypothesize that it will help increasing recall while keeping precision at a reasonable rate. Secondly, when a multiword hypernym is retrieved by our module, it might not match exactly a term from the seed terminology (e.g. *original_term→soft drink*, and *retrieved_term→carbonated soft drink*). Therefore, we aim at decomposing it by dropping one modifier at a time and creating new arcs recursively. These two steps are described in more detail in the following subsection.

Post-classification Heuristic

Our recall-enhancing strategy consists in a post-classification heuristic inspired by Flati et al. (2014): (1) We exploit the tree-like dependency structure of a parsed sentence in order to find the most likely token to be the head of a hypernymic phrase. We look for definitions where no hypernym was identified. Then, we find the node with the *Predicative Complement* (PRD) syntactic function. If such node is not a *stop-hypernym* (such as *type*, *class*, *family* or *kind*), we consider it a valid head of a hypernymic-phrase[4]. Then, we collect all its noun and adjective children with the syntactic function *Modifier of Nominal* (NMOD). If, however, such node is a stop-hypernym, we go down the syntactic tree one level and look for a direct *Preposition* node with syntactic function NMOD. Then, we extract this preposition's adjective and noun children if they have the syntactic function *Modifier of Prepositional* (PMOD).

For example, consider the following sentence: "Whisky or whiskey is a type of distilled alcoholic beverage made from fermented grain mash". Here, *type* is the Predicative Complement node but it is an uninformative word for describing the term *whisky*. Therefore, our algorithm goes one level down the syntactic tree and identifies the token *beverage* as the direct child of the preposition and therefore extracts this token as hypernym.

3.3 Hypernym Decomposition

This step is aimed at generating deeper paths from a term and its hypernym by recursively decomposing a candidate hypernym. For example, consider the previous example's *term→hypernym* relation if the hypernym's modifiers are taken into account: *whisky→distilled alcoholic beverage*. Our objective is to generate the following set of relations: *distilled alcoholic beverage→alcoholic beverage* and *alcoholic beverage→beverage*. In this way, we improve the taxonomy since, in taxonomy learning, longer hypernymy paths should be preferred (Navigli et al., 2011), and we enable other potential *distilled alcoholic beverages* to be connected with *alcoholic beverage* rather than the more generic term *beverage*.

We achieve this by performing a similar algorithm as in the post-classification heuristic, i.e. exploiting head and modifier relations in a dependency tree.

3.4 Graph Generation

At this stage, we have a dataset of *term→hypernym* pairs, and from here populating the taxonomy is a trivial task. For each pair, if neither *term* nor *hypernym* exist in the graph, add both nodes and connect them. If *term* exists in the graph, only add the *hypernym* and connect the existing *term* node with it. If on the contrary, only the *hypernym* is found in the graph, connect the term to the existing hypernym node. Finally, we go back to the initial flat terminology and, if no path is found between a term node and the root node, add a direct edge between them. This last step guarantees that the taxonomy will preserve

[4]The full list of stop-hypernyms is available at www.wibitaxonomy.org

	chem	wn_chem	equip	wn_equip	food	wn_food	sci	wn_sci
VC	1	0.997	1	1	0.8695	1	0.9977	0.8624
EC	0.0004	0.093	0.1577	0.0453	0.0359	0.0782	0.0172	0.1111
RNE	0.7089	0.9531	0.9235	6.903	0.9527	0.9315	3.4731	0.78
F&M	0.2225	0.2787	0.4482	0.0901	0.3267	0.3091	0.2202	0.2126
Cycles	no	yes	yes	yes	no	yes	yes	no
Precision	0.0006	0.0889	0.1458	0.0287	0.0363	0.0775	0.0733	0.1246
Recall	0.0004	0.093	0.1577	0.2	0.0359	0.0782	0.2559	0.1111
F-Score	0.0005	0.0909	0.1515	0.0503	0.0361	0.0778	0.1139	0.1175

Table 1: Summary of the results obtained with our approach in the structural evaluation in terms of vertex coverage (VC), edge coverage (EC), ratio of novel edges (RNE), cumulative Fowlkes and Mallows Measure (F&M), whether the taxonomy contains cycles (Cycles), and Precision, Recall and F-Score against gold standard taxonomies.

the vast majority of the initial terms (if not all, as can be seen in Table 1).

4 Evaluation

Evaluation is carried out considering the structural properties of the taxonomy, as well as its quality when compared to gold-standard (see Table 1). These gold taxonomies can be either the subgraphs rooted at one relevant WordNet term (chemical, food, equipment or science), or taxonomies manually crafted by domain experts.

These results suggest that the approach described in this paper can be safely followed to construct a taxonomy from a flat terminology as input, provided major issues like domain-specificity or WSD are addressed. Our approach strongly depends of available definitions of terms in Wikipedia, which was not the case in very specific domains (such as the *chemical* terminology). On the other hand, however, the hypernym extraction pass worked well and thus we are encouraged to work in this direction, stressing the importance of an appropriate domain dataset from which definitional knowledge can be extracted.

In order to compare the system and reference taxonomies, the evaluation consists in computing node and edge coverage by taking into account the number of nodes and edges in common and the sizes of the taxonomies. In addition, the results of a structural metric are also provided, such metric being the Fowlkes&Mallows measure (Fowlkes and Mallows, 1983), a method for comparing hierarchical clusters. The results show poor performance of our system in inferring relations among concepts at deeper levels in the taxonomy. One of the reasons this might be due to is the fact that the lexicalization of a term does not necessarily have to be exact between a BabelNet synset and an associated Wikipedia definition.

Regarding the manual evaluation of the quality of newly acquired edges, our system is unsurprisingly weak (P=10.2%)[5] due to the inherent term ambiguity which makes our system retrieve noisy definitions at each step. We hypothesize that our results might be higher in the chemical domain, since terminology would be less prone to be polysemous. However, this domain was not considered for this evaluation measure. These negative results together with the good performance of the hypernym-extraction module stress the need to retrieve valid domain specific definitional sentences for our approach to work well.

5 Conclusions and Discussion

We have described a system designed for constructing a taxonomy from a flat list of terms. It is based on a module that queries BabelNet for Wikipedia definitions in order to obtain definitional knowledge

[5]Full results for all systems are reported in http://alt.qcri.org/semeval2015/task17/index.php?id=evaluation.

for each term. Then, a machine-learning algorithm is trained with a manually annotated dataset with hypernym relations in definitional sentences, and applied to our definition dataset. Different post-classification heuristics are afterwards incorporated to the pipeline with a two-fold objective: (1) Extract a candidate hypernym in cases where the classifier lacked the confidence to tag one or more tokens as possible hypernyms, and (2) Decompose candidate hypernyms exploiting the syntactic relation between their head and its modifiers in a syntactic dependency tree. Finally, with a set of *term→hypernym* pairs we populate a domain taxonomy by connecting terms and hypernyms, and finally by fixing disconnected nodes from the root.

We have demonstrated that our approach has very high vertex coverage, and on the other hand is flawed in capturing deep taxonomic relations among entities. The hypernym extraction module achieves state-of-the-art performance and due to the simplicity of the features used is open for improvements, either by incorporating semantic similarity among tokens, frequencies in domain corpora, or a token's position in the syntactic tree.

We observe a clear room for improvement in the domain corpus compilation part, and for the future we are investigating the potential of the Wikipedia Categories Graph in order to gather domain definitions from pages that are in recurrent categories in the BabelNet synset list.

Acknowledgments

We would like to express our gratitude to the anonymous reviewers for their helpful comments. This work is partially funded by the SKATER project, TIN2012-38584-C06-03, Ministerio de Economía y Competitividad, Secretaría de Estado de Investigación, Desarrollo e Innovación, España; and project Dr. Inventor (FP7-ICT-2013.8.1 611383).

References

Matthew Berland and Eugene Charniak. 1999. Finding parts in very large corpora. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 57–64. Association for Computational Linguistics.

Guido Boella, Luigi Di Caro, Alice Ruggeri, and Livio Robaldo. 2014. Learning from syntax generalizations for automatic semantic annotation. *Journal of Intelligent Information Systems*, pages 1–16.

Bernd Bohnet. 2010. Very high accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 89–97, Stroudsburg, PA, USA. Association for Computational Linguistics.

Georgeta Bordea, Paul Buitelaar, Stefano Faralli, and Roberto Navigli. 2015. Semeval-2015 task 17: Taxonomy extraction evaluation (texeval). In *Proceedings of the 9th International Workshop on Semantic Evaluation*. Association for Computational Linguistics.

Hang Cui, Min-Yen Kan, and Tat-Seng Chua. 2005. Generic soft pattern models for definitional question answering. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 384–391. ACM.

Edward B. Fowlkes and Colin L. Mallows. 1983. A method for comparing two hierarchical clusterings. *Journal of the American statistical association*, 78(383):553–569.

Ruiji Fu, Jiang Guo, Bing Qin, Wanxiang Che, Haifeng Wang, and Ting Liu. 2014. Learning semantic hierarchies via word embeddings. In *Proceedings of the 52th Annual Meeting of the Association for Computational Linguistics: Long Papers*, volume 1.

Roxana Girju, Adriana Badulescu, and Dan Moldovan. 2003. Learning semantic constraints for the automatic discovery of part-whole relations. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 1–8. Association for Computational Linguistics.

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics-Volume 2*, pages 539–545. Association for Computational Linguistics.

Zornitsa Kozareva, Ellen Riloff, and Eduard H. Hovy. 2008. Semantic class learning from the web with hyponym pattern linkage graphs. In *ACL*, volume 8, pages 1048–1056. Citeseer.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

A Muresan and Judith Klavans. 2002. A method for automatically building and evaluating dictionary resources. In *Proceedings of the Language Resources and Evaluation Conference (LREC*.

Roberto Navigli and Simone Paolo Ponzetto. 2010. Babelnet: Building a very large multilingual semantic network. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 216–225. Association for Computational Linguistics.

Roberto Navigli and Paola Velardi. 2010. Learning word-class lattices for definition and hypernym extraction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 1318–1327, Stroudsburg, PA, USA. Association for Computational Linguistics.

Roberto Navigli, Paola Velardi, and Stefano Faralli. 2011. A graph-based algorithm for inducing lexical taxonomies from scratch. In *IJCAI*, pages 1872–1877.

Stephen Roller, Katrin Erk, and Gemma Boleda. 2014. Inclusive yet selective: Supervised distributional hypernymy detection. In *Proceedings of the Twenty Fifth International Conference on Computational Linguistics (COLING-14), Dublin, Ireland*.

Dominic Widdows and Beate Dorow. 2002. A graph model for unsupervised lexical acquisition. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–7. Association for Computational Linguistics.