NUIG-UNLP at SemEval-2016 Task 13: A Simple Word Embedding-based Approach for Taxonomy Extraction

Joel Pocostales
Insight Centre for Data Analytics
National University of Ireland, Galway
name.surname@insight-centre.org

Abstract

This paper describes the NUIG-UNLP system submitted to SemEval-2016, Task 13. We implement a semi-supervised method that extracts hypernym candidates for the terms provided in the test list. The main assumption of our system is that hypernyms may be induced by adding a vector offset to the corresponding hyponym word embedding. The vector offset is obtained as the average offset between 200 pairs of hyponym-hypernym in the same vector space. Our approach ranked second on connectivity (c.c.) and categorisation (i.i.) for the English taxonomy construction, and fifth on the overall ranking. Despite of these modest results, our system achieved comparable evaluations scores with the other participants.

1 Introduction

Hyponyms and hypernyms (sometimes called subordinate and superordinate terms, respectively) describe a type of relation which, in general, can be defined in terms of asymmetric entailment: given the hyponym of feline, cat, and its hypernym, feline, we can state that all cats are felines, but not that all felines are cats.

Likewise, the relations that hyponyms and hypernyms signal can also be characterized as a isa relation between a hyponym X and hypernym Y: for nouns, X is a Kind of Y or X is a type of Y (Saint-Dizier and Viegas, 1995). These particular *type / kind-of* relations form the backbone of the construction of Lexical Taxonomies and Ontologies (Buitelaar et al., 2004; Navigli et al., 2011), and those in turn plays a essential role in many Natural Language Processing applications: Question Answering, Textual Entailment, Natural Language Inference, or Text Summarization (Bordea et al., 2015).

In this regard, despite the fact that taxonomy construction can be addressed from a diversity of approaches, the lexico-syntactic patterns-based are still the most widely used. Nevertheless, in the last years some vector space-based approaches have emerged for learning semantic hierarchies (Saxe et al., 2013; Khashabi, 2013; Fu et al., 2014; Rei and Briscoe, 2014; Tan et al., 2015; Nayak, 2015). In the next sections we will mainly turn our attention to this type of approaches.

1.1 Task Definition

The five participating teams in SemEval-2016 Task 13 were provided with six datasets in four languages (English, Dutch, French and Italian)[1]. The datasets can be divided in three domains (science, environment and food). Additionally, this year the TExEval-2 task has a focus in four subtasks related to taxonomy construction:

1. Taxonomy construction
2. Hypernym identification
3. Multilingual taxonomy construction
4. Multilingual hypernym identification

However, due to lack of time, we decided to address only the English monolingual subtasks.

[1] The corresponding system description papers can be found in Cleuziou and Moreno (2016), Panchenko et al. (2016) and Tan (2016).

Proceedings of SemEval-2016, pages 1189–1193,
San Diego, California, June 16-17, 2016.

The key idea behind TExEval tasks is the creation and evaluation of systems capable of automatically extracting hierarchical relations from text and then constructing taxonomies. Following (Fu et al., 2014), ideally, the construction of those hierarchies can be seen as a directed acyclic graph DAG with a finite set of nodes (words) and edges representing the *asymmetric* and *transitive* hyponym-hypernym relations. This is formally defined by Fu et al. (2014) as follows:

- $\forall x, y \in L : x \xrightarrow{H} y \Rightarrow \neg \left(y \xrightarrow{H} x \right)$
- $\forall x, y, z \in L : \left(x \xrightarrow{H} z \wedge z \xrightarrow{H} y \right) \Rightarrow x \xrightarrow{H} y$

where in our case x, y and z denote the terms in the domain list $L_d \in L$, and the hyponym-hypernym relation is represented by $\xrightarrow{H}$. Therefore, the aim of the task was to return a list of pairs $x \xrightarrow{H} y$ for each term in the six different domains L_d.

2 Experimental Setup

We describe in this section our taxonomy extraction system.

2.1 Training Data

Since TExEval-2 organizers did not provide any specific corpus for the task, we used the latest Wikipedia dump[2]. We preprocessed it using the WikiExtractor tool[3], which generates a plain text from a Wikipedia database dump discarding markup tags and any other element different than text, such as tables, references, lists and images. On the other hand, in order to generate a single word embedding for each entry in the test list, we underscore all the entries containing open compound words:

```
civil engineering ⇒
 civil_engineering
```

2.2 Word Embeddings Generation

We use the log-bilinear model GloVe (Pennington et al., 2014) trained over the above-mentioned Wikipedia corpus to generate vector space representations of words. Following the analogy task results presented in their paper and some pre-experimental test, we set a windows size of 10 and 300 dimensions word embeddings. The number of iterations of the model was set to 20.

2.3 Offset Model

Mikolov et al. (2013) and subsequently Levy and Golberg (2014) demonstrated that word embeddings generated by neural nets (and also other traditional distributional methods) preserve some syntactic and semantic information. Some of this encoded information, such as relational similarities between pairs of words, can be recovered by simple vector offsets between the vector embeddings of each word. Thus, as Mikolov et al. (2013) and Levy and Golberg (2014) showed, given two pairs of words that share a relation, $a : a^*$, $b : b^*$, the relation between those two words can be represented by their vector offset, as follows:

$$a^* - a \approx b^* - b \tag{1}$$

Therefore, the vector of the word b^* should be similar to the proxy vector y'

$$y' = b - a + a^* \tag{2}$$

where y', ideally, corresponds to the vector representation of b^*. Since y' will rarely match the exact position of the word b^*, different similarity measures may be applied to find the most similar word to y'. In this paper we will focus only in Cosine similarity (4) and Euclidean distance (5):

$$\cos(b^*, y) = \frac{b^* \cdot y'}{\parallel b^* \parallel \parallel y' \parallel} \tag{3}$$

maximizing the function:

$$\underset{b^* \in V}{\arg\max} \left(\cos(b^*, y') \right) \tag{4}$$

where V is the vocabulary.

And given the Euclidean distance formula, subsequently we obtain the following function:

$$d(b^*, y') = \parallel b^* - y' \parallel^2 \;\Rightarrow\; \underset{b^* \in V}{\arg\min} \parallel b^* - y' \parallel^2 \tag{5}$$

[2] We used the English snapshot of 17-Nov-2015

[3] https://github.com/attardi/wikiextractor

2.4 Offset Model for Hypernym-Hyponym Relation

Mikolov et al. (2013) and Levy and Golberg (2014) have only tested the vector offset method for simple symmetric relations such a capital-country, gender inflections, adjective-to-adverbs, etc. However, as Rei and Briscoe (2014) pointed out, hypernym-hyponyms relations are conceivable much more difficult to represent by simple vector offsets, as their relations rarely are symmetric.

Rei and Briscoe (2014) in their paper first assess how word embeddings perform in hypernym-hyponym detection and generation, and second, propose a new directional similarity measure (WeightedCosine) based on two new properties to detecting these relations.

In our submitted system, though, we finally decided not to implement this new measure due to lack of time.

2.5 Offset Model for the Hypernym-Hyponym Relation

We first generate a random list of 200 pairs of hypernym-hyponyms. This training list was extracted from the trial data provided in Bordea et al. (2015) and WordNet (Miller, 1995; Fellbaum, 1998) covering different domains.

Using the Gensim library[4] (Řehůřek and Sojka, 2010) we compute the vector offset as the average offset of all the pairs generated in the above-mentioned training data (Mikolov et al., 2013; Nayak, 2015):

$$v_{offset} = \frac{1}{n}\sum_{i=1}^{n}\left(v_{hyper(i)} - v_{hypo(i)}\right) \qquad (6)$$

where $n = 200$, as the number of pairs of hypernym-hyponym in our training data.

Once the v_{offset} has been obtained, we add it to the target terms in the test list:

$$y' \approx v_{term} + v_{offset} \qquad (7)$$

where we assume that the addition of the vectors v_{offset} and v_{term} projects y' close enough to the hidden hypernym representation b^*. Thus, we apply either the measure similarity (4) or (5) and we rank either the top 10 or 5 candidates, discarding those terms not included in the test list. We also implement a substring inclusion approach based on regexp (Nevill-Manning et al., 1999) so that

$$\texttt{civil engineering} \xrightarrow{H} \texttt{engineering}$$
$$\texttt{social psychology} \xrightarrow{H} \texttt{psychology}$$

In other words, given an open compound word such *civil engineering*, we assume that the second term *engineering* is the most likely hypernym of *civil engineering*.

3 Evaluation Metrics

In this section we present the results obtained in the second task on Taxonomy Extraction Evaluation as part of SemEval-2016. The metrics correspond to the structural analysis and the comparison against the Gold Standard effectuated by Bordea et al. (2016). The best results among all the systems appear in a bold font (note that Euclidean 5 and Cosine 5 have been excluded as they were not submitted on time).

3.1 Results

Table 1 shows the structural analysis of our system and the corresponding results when compared to Gold Standard. We were only able to submit the first system (Euclidean 10), and we had to exclude the food domains due to time limitations. However, we will also present here metrics beyond the official system submission, i.e, Euclidean 5 and Cosine 5 covering all domains provided on the test data.

The hyphen (-) is used in cases when the number of cycles could not be computed due to hardware limitations. This outcome should be interpreted as negative, as the presence of cycles goes against Directed Acyclic Graph (DAG) definition.

As per the structural analysis, the main goal is to evaluate the number of correct nodes and edges in comparison with the Gold Standard. Thus, the quantifying metrics in the left block (|V| ... i.n.) cannot really be considered aside of the Golden Standard evaluation. Therefore, in this section we will mainly focus on the qualifying metrics instead of the quantifying ones.

We observe that, likely, due to the restrictions imposed in our algorithm not allowing hypernym can-

[4] http://radimrehurek.com/gensim/

Euclidean 10	**|V|**	**|E|**	**#c.c**	**cycles**	**i.n.**	**#VC**	**%VC**	**:VN**	**#EC**	**%EC**	**:EN**
Enviro_Eu.	312	456	58	347	**176**	221	0.8467	0.2917	72	0.2758	1.4713
Science	596	**1656**	99	-	**409**	422	0.9336	0.2907	163	0.3505	3.2107
Science_Eu.	97	218	13	269	72	83	0.6620	0.1443	29	0.2339	1.5242
Science_WN	251	**929**	9	-	195	241	0.6513	0.0398	163	0.3602	1.6947
Euclidean 5	**|V|**	**|E|**	**#c.c**	**cycles**	**i.n.**	**#VC**	**%VC**	**:VN**	**#EC**	**%EC**	**:EN**
Enviro_Eu.	329	944	58	310	192	221	0.8467	0.3283	134	0.5134	3.1034
Science	594	1366	94	-	396	417	0.9226	0.2980	132	0.2839	2.6538
Science_Eu.	98	212	13	239	72	83	0.6640	0.1531	26	0.2097	1.5000
Science_WN	246	748	9	1175180	188	236	0.6378	0.0406	127	0.2810	0.1374
Cosine 5	**|V|**	**|E|**	**#c.c**	**cycles**	**i.n.**	**#VC**	**%VC**	**:VN**	**#EC**	**%EC**	**:EN**
Enviro_Eu.	246	294	48	71	136	177	0.6781	0.2804	56	0.2145	0.9118
Food	909	888	206	84	460	679	0.4383	0.2530	190	0.1197	0.4398
Food_WN	983	1160	181	157	520	813	0.5471	0.1729	332	0.2106	0.5253
Science	403	835	45	1173	281	316	0.6991	0.2159	118	0.2538	1.5419
Science_Eu.	76	106	14	13	46	63	0.5040	0.1710	22	0.1774	0.6774
Science_WN	200	353	16	89	137	190	0.5135	0,0500	99	0.2190	0.5619

Table 1: Official Evaluation metrics for Euclidean measure top 10, 5 and Cosine measure 5 with substring inclusion. Number of nodes and edges |V|), |E|, connected components (c.c.), cycles, intermediate nodes (i.n.), number of vertices, vertices coverage and vertex novelty (#VC, %VC, VN), number of edges, edge coverage and edge novelty (#EC, %EC and EN)

	Euclidean 10			***Euclidean 5***			***Cosine 5***		
	Precision	**Recall**	**F-Score**	**Precision**	**Recall**	**F-Score**	**Precision**	**Recall**	**F-Score**
Enviro_Eu.	0.1579	**0.2759**	0.2008	0.1419	0.5134	0.2224	0.1905	0.2145	0.2018
Food	N/A	N/A	N/A	N/A	N/A	N/A	0.2140	0.1197	0.1535
Food_WN	N/A	N/A	N/A	N/A	N/A	N/A	0.2862	0.2106	0.2427
Science	0.0984	**0.3505**	0.1537	0.0967	0.2839	0.0144	0.1413	0.2537	0.1815
Science_Eu.	0.1330	0.2339	0.1695	0.1226	0.2097	0.1548	0.2075	0.1774	0.1913
Science_WN	0.1754	0.3606	0.2360	0.1698	0.2810	0.2117	0.2804	0.2190	0.2460
AVERAGE	0.1412	0.3052	0.1900	0.1327	0.3220	0.1508	0.2200	0.1100	0.2028

Table 2: Precision, Recall and F-Score for Euclidean Distance 10, 5 and Cosine 5

didates out of the test list, there are no significant differences between Euclidean top 10 and 5.

We also note that the number of cycles was considerably higher on the euclidean approaches, in fact, exceeding the computer memory capacities for some domains, namely, Science, Science_WordNet (see the surprisingly high figure for Euclidean 5, Science_WN). On the other hand, unlike our initial assumptions, the cosine approach did not perform much better than the Euclidean ones. Our system obtained comparable recall values with the other systems, at the expenses, though, of the precision. Therefore, the results achieved by our systems are in general modest, especially taking into consideration that our algorithm also included a substring inclusion module (as described in section 2.5).

3.2 Conclusion and Discussion

Although there is still room for improvement in our system, we conclude that the diversity involved in the complex hypernym-hyponym relations cannot easily be captured by a simple vector offset mean. As direction for future work, it might be worth considering domain specific vectors as well as incrementing the number of training pairs for the vector offset mean.

Our system ranked second on connectivity (c.c.) and categorisation (i.i.) for the English taxonomy construction, and fifth on the overall ranking (see Bordea et al. (2016) for further details on the evaluation metrics).

Acknowledgments

This work has been funded by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289 (INSIGHT).

References

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.

Georgeta Bordea, Els Lefever, and Paul Buitelaar. 2016. Semeval-2016 task 13: Taxonomy Extraction Evaluation (TExEval-2). In *Proceedings of the 10th International Workshop on Semantic Evaluation*. Association for Computational Linguistics.

Paul Buitelaar, Philipp Cimiano, and Bernardo Magnini. 2004. Ontology Learning from Text : An Overview. *Learning*, pages 1–10.

Guillaume Cleuziou and Jose G. Moreno. 2016. QASSIT at SemEval-2016 Task 13: On the integration of Semantic Vectors in Pretopological Spaces for Lexical Taxonomy Acquisition. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*. Association for Computational Linguistics.

Christiane Fellbaum. 1998. WordNet: An Electronic Lexical Database.

Ruiji Fu, Jiang Guo, Bing Qin, Wanxiang Che, Haifeng Wang, and Ting Liu. 2014. Learning Semantic Hierarchies via Word Embeddings. *Association for Computational Linguistics*, pages 1199–1209.

Daniel Khashabi. 2013. In *On the Recursive Neural Networks for Relation Extraction and Entity Recognition Authors*. Technical report, may.

Omer Levy and Yoav Goldberg. 2014. Linguistic Regularities in Sparse and Explicit Word Representations. In *Proceedings of the 18th Conference on Computational Natural Language Learning (CoNLL 2014)*, pages 171–180.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, number June, pages 746–751.

George A. Miller. 1995. WordNet: A Lexical Database for English. *Commun. ACM*, 38(11):39–41, November.

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

Neha Nayak. 2015. In *Learning Hyperonyms over Word Embeddings*. Student technical report.

Craig G. Nevill-Manning, Ian H. Witten, and Gordon W. xc Paynter. 1999. Lexically-generated subject hierarchies for browsing large collections. *International Journal on Digital Libraries*, 2(2-3):111.

Alexander Panchenko, Stefano Faralli, Eugen Ruppert, Steffen Remus, Hubert Naets, Cedrick Fairon, Simone Ponzetto, Paolo, and Chris Biemann. 2016. TAXI at SemEval-2016 Task 13: a Taxonomy Induction Method based on Lexico-Syntactic Patterns, Substrings and Focused Crawling. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May. ELRA.

Marek Rei and Ted Briscoe. 2014. Looking for Hyponyms in Vector Space. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 68–77.

Patrick Saint-Dizier and Evelyne Viegas. 1995. Computational lexical semantics. *Studies in natural language processing*, pages ix, 447 p.

Andrew M. Saxe, James L McClelland, and Surya Ganguli. 2013. Learning hierarchical category structure in deep neural networks. *Proceedings of the 35th annual meeting of the Cognitive Science Society*, pages 1271–1276.

Liling Tan, Rohit Gupta, and Josef van Genabith. 2015. USAAR-WLV: Hypernym Generation with Deep Neural Nets. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 932–937, Denver, Colorado. Association for Computational Linguistics.

Liling Tan, Francis Bond, and Josef van Genabith. 2016. USAAR at SemEval-2016 Task 13: Hyponym Endocentricity. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*. Association for Computational Linguistics.