TALN at SemEval-2016 Task 14: Semantic Taxonomy Enrichment Via Sense-Based Embeddings

Luis Espinosa-Anke, Horacio Saggion and **Francesco Ronzano**
TALN - Department of Information and Communication Technologies
Universitat Pompeu Fabra
Carrer Tànger 122-134, 08018 Barcelona (Spain)
{luis.espinosa, horacio.saggion, francesco.ronzano}@upf.edu

Abstract

This paper describes the participation of the TALN team in SemEval-2016 Task 14: Semantic Taxonomy Enrichment. The purpose of the task is to find the best point of attachment in WordNet for a set of Out of Vocabulary (OOV) terms. These may come, to name a few, from domain specific glossaries, slang or typical jargon from Internet forums and chatrooms. Our contribution takes as input an OOV term, its part of speech and its associated definition, and generates a set of WordNet synset candidates derived from modelling the term's definition as a sense embedding representation. We leverage a BabelNet-based vector space representation, which allows us to map the algorithm's prediction to WordNet. Our approach is designed to be generic and fitting to any domain, without exploiting, for instance, HTML markup in source web pages. Our system performs above the median of all submitted systems, and rivals in performance a powerful baseline based on extracting the first word of the definition with the same part-of-speech as the OOV term.

1 Introduction

Semantic knowledge, expressed in terms of concepts and relations holding among them, is an essential enabling component of NLP applications (Jurgens and Pilehvar, 2015). One of the best known semantic repository is WordNet (Miller et al., 1990), a manually created semantic network with a coverage of over 200k English senses and 155k word forms. However, as knowledge domains advance and expand, novel terms beyond the coverage of WordNet are coined on a regular basis. Thus, the need for automatic approaches which are able to gather information from unstructured online sources and structure them is in high demand. In this context, WordNet has become the reference semantic network in previous attempts to formalize novel knowledge. The SemEval-2016 Task 14: Semantic Taxonomy Enrichment (Jurgens and Pilehvar, 2016) aims at providing an experimental ground on the task of, given an Out-of-Vocabulary (OOV) term, and an associated definition and part of speech, find the best point of attachment in WordNet (its most similar synset). This is a challenging problem because an OOV term may be defined without any explicit mention to its closest WordNet synset. For instance, for the OOV term (from training data) "lectionary", the associated definition is "The book that contains all the readings from the Scriptures for use in the celebration of the liturgy", and the best point of attachment is sacred_text#n#01. However, if one was to simply obtain the first head word of the definition and retrieve its first sense, the result would be book#n#01, with a score of 0.125/1 according to the official evaluation metric (see Section 4).

In this paper we describe our approach to such task. We base our method on the intuition that unseen terminology may be *understood* in terms of how terms are defined. Hence, we propose an algorithm which, for each definition, performs part-of-speech tagging and parsing, generates a set of noun and verb phrases, and then takes advantage of a vector space representation of word and phrase senses for modelling the definition. Our algorithm produces several WordNet attachment candidates as

Proceedings of SemEval-2016, pages 1332–1336,
San Diego, California, June 16-17, 2016. ©2016 Association for Computational Linguistics

the modelling takes place, e.g. by obtaining the definition's centroid and mapping it to WordNet, or by parsing the OOV term and searching for a mapping between its head and WordNet. For this task, we submitted the three runs of our system with the highest scores on the training data. Moreover, the task organizers provided two baselines: One which selected a random WordNet synset, and one which extracted the first word in the definition with the same part-of-speech, and assigned it its first sense in WordNet (*BaselineFirst*). The experimental results suggest that the baseline was a very strong competitor, ranking way above the median of the participating systems. Our best run was 3 points below the baseline and 5 points above the median according to a metric designed to compute the distance between two nodes in a hierarchical cluster (see Section 4).

The rest of this paper is structured as follows: After presenting an overview of related work (Section 2), we describe in detail the methodology followed for our submission (Section 3), then we provide evaluation results as compared with baseline systems and the participants' median score, along with a discussion of a few interesting cases. Finally, we outline directions of future work in a novel and exciting task, which opens a very interesting research problem to be tackled in the future.

2 Related Work

Expanding current semantic knowledge is a task that may be approached either by expanding WordNet with knowledge derived from structured or semi-structured repositories, or by learning hyponym-hypernym hierarchies separately in order to obtain novel knowledge. Within the former group, we find projects like BABELNET (Navigli and Ponzetto, 2012), a very large multilingual semantic network, or pairwise lexical resources alignments (Miller and Gurevych, 2014; Pilehvar and Navigli, 2014). These approaches, however, are not designed to capture novel terminology due to the fact that, for an alignment to occur, there must exist a direct correspondence between lemmas. This is addressed in projects belonging to the latter group, starting from (Snow et al., 2006), and following more recently with extraction of information from the web (Velardi et al., 2013; Luu Anh et al., 2014) or from BABELNET's glosses (Espinosa-Anke et al., 2016). These approaches, however, do not perform a direct mapping against WordNet, but rather separate taxonomies with their own hierarchical structure.

3 Method

Our approach to finding the best match in WordNet for an OOV term is based on transforming the associated textual definition into a representative vector. However, performing this action over plain text would introduce an ambiguity problem due to the occurrence of polysemic entities and specific lexical and syntactic formulations within the definition. Therefore, we leveraged SENSEMBED (Iacobacci et al., 2015), a state of the art vector space representation of word senses designed with BABELNET as a reference sense inventory. In SENSEMBED, each vector includes a concatenation of lemma and BABELNET synset. We exploit the mapping BABELNET→WordNet in order to find the best candidate for the given OOV term.

Let t be an OOV term, and d its associated definition, our sense-finding algorithm aims at extracting a suitable WordNet sense for t. To attain this goal, we perform a set of steps which generate a candidate set $\mathcal{C}$, which we will afterwards rank according to certain criteria. The different candidate rankings that we obtain constitute the three runs of our submission. Let us explain how each candidate is obtained.

Candidate Retrieval

First, we parse both t and d using the parser described in (Bohnet, 2010). Note that t may be a multiword term including prepositional phrases (e.g. "margin of error") and hence simply tagging and retrieving the last token would not suffice. We traverse the parse trees obtained in both cases in a depth-first-fashion until a lemma with a WordNet synset with the same part-of-speech as t is found. This generates two candidate WordNet synsets c_t^{wn}, and c_d^{wn}, which correspond to the first sense of the matching lemma. After completing this first step, $\mathcal{C} = \{c_t^{wn}, c_d^{wn}\}$.

The process of retrieving these candidates is inspired by the baseline *BaselineFirst* provided by the task organizers, which retrieved the first word in the definition with the same part-of-speech as the novel term. We evaluated a (non-submitted) *baseline* version of our system, which only considers c_t^{wn} and c_d^{wn}. We obtained slightly worse performance than the organizers' baseline on both trial and training data, mostly due to parsing errors.

The second step of the algorithm aims at finding optimal candidate synsets for t by modeling d in terms of its vector representation. We proceed as follows. First, after stopword removal, we apply a shallow parsing technique over part-of-speech tags in order to isolate both single and multiword NPs and VPs. Specifically, we group all textual chunks that match the following regular expressions:

```
NP = <JJ|NN.*>+
VP = <VB.*><NP>?
```

This step results in d_{ch}, the set of noun and verb phrases identified in a definition. We did not consider multiword expressions including a prepositional phrase, as these introduced a considerable amount of noise. For example, for d = "Genetic drift is a mechanism of evolution", d_{ch} = {genetic_drift, drift, mechanism, evolution}. Then, for each $ch \in d_{ch}$, we obtain, where possible, all its available senses in SENSEMBED and store them in two separate sets. The first one, denoted as D_{Senses}, contains all senses extracted from all $ch \in d_{ch}$ without distinguishing the original chunk from which they come from. It is simply a set of senses. The second set, denoted as $D_{chSenses}$, isolates all available senses corresponding to each chunk in an individual set, and thus constitutes a set of sets.

We use these two sets for obtaining (1) A *centroid sense* μ over the definition's associated senses D_{Senses}, and (2) The set containing only the *best sense* for each chunk. For the former case, $\mu(D_{Senses})$ is obtained as follows:

$$\mu = \frac{1}{|D_{Senses}|} \sum_{s \in D_{Senses}} \frac{s}{\|s\|} \quad (1)$$

As for the latter, *best sense* refers to the closest sense to μ by cosine score obtained from the set of senses retrieved for each chunk, namely $d_{chSen} \in D_{chSenses}$, which is computed as follows:

$$bestS(d_{chSen}) = \text{argmax}_{s \in d_{chSen}} \frac{s \cdot \mu}{\|s\| \, \|\mu\|} \quad (2)$$

The set of best senses per chunk is simply $D_{bestS} = \{bestS\,(d_{chSen})\,, d_{chSen} \in D_{chSenses}\}$, and will be used to generate the last candidate in our process.

At this stage, we obtain three more candidates to our candidate set, namely (1) The closest vector to the centroid from of all available senses in SENSEMBED, μ^{wn}; (2) The sense $bestS\,(d_{chSen})$ appearing at the first position of the sentence, denoted as $firstBest$; and (3) The sense in D_{Senses} with highest cosine similarity with μ, denoted as $highestBest$. In our sample sentence, μ^{wn} = branching_ratio$_{bn}$[1], $firstBest$ = drift$_{syn}$, and $highestBest$ = drift$_{syn}$.

Finally, we introduce one last additional candidate to our candidate pool. We observed that μ may not be an optimal centroid for the sentence as it is computed by considering all possible senses for each extracted chunk. For this reason, we take advantage of the fact that we have an available set of best senses for each chunk, namely D_{bestS}, and create a *sense-aware centroid*, denoted as μ_{best}^{wn}. Finally, $\mathcal{C} = \mathcal{C} \cup \{\mu^{wn}, firstBest, highestBest, \mu_{best}^{wn}\}$.

Ranking Candidates

Having obtained a set of candidate WordNet synsets $\mathcal{C}$ for an OOV term t, we propose a ranking scheme aimed at providing three possible outcomes for our system. Note that we may not have all candidates available from all the previously described steps, as there may be BABELNET synsets without a direct WordNet mapping, definitions may be very short (even one word), or there may be no available candidate with the same part-of-speech as t's head.

We submitted three runs of our system, which we describe as follows:

- **RunHeads** The baseline *BaselineFirst*, provided by the organizers, performed well, i.e. it found the exact correct attachment for almost half of the training instances. Based on this observation, in this submission we prioritized those cases in which there was a direct mapping between the first sense of the definition root and a WordNet synset (c_d). If

[1] We denote with subscript *syn* those WordNet senses obtained via direct mapping from SENSEMBED, and with subscript *bn* those only found in BABELNET.

this was not the case, we searched for the mapping of t's head (c_t), as we found that in many cases, when t was a multiword expression, an optimal point of attachment could be found thanks to its head word. If none of these attempts were successful, we opted for selecting candidates in the following order: μ^{wn}_{best}, μ^{wn}, $firstBest$, and $highestBest$.

- **RunSenses** This run first incorporated a voting scheme based on candidates in C providing one same answer. If this was the case across three or more candidates, this was the synset selected by the system. However, if agreement across candidates was below three, priority was given to candidates coming from the definition modeling via SENSEMBED, and if no candidate was found, the system searched for mappings in c^{wn}_d and c^{wn}_t.

- **RunHyps** This run of our system was the most conservative approach. From all possible candidates extracted either via SENSEMBED, or from c_d and c_t, we computed the number of hyponyms each candidate had in the WordNet hierarchy. Then, our system selected as answer the most general synset, i.e. the one with the highest number of hyponyms. Our hypothesis was that by including the most generic terms possible (upper in the hierarchy) we would be less likely to incur in wrong senses of highly specific terms, where different senses may be placed more sparsely than in more generic ones.

4 Evaluation

Evaluation is performed over several criteria. First, the distance between the selected point of attachment and the gold standard is computed via the Wu & Palmer similarity (**W&P**) (Wu and Palmer, 1994). Second, a Recall measure (**R**), aimed at allowing systems to not provide an answer in doubtful cases, where the chances of an incorrect attachment would be high. It is simply computed as the percentage of items answered by the system. Finally, a score on *lemma match* (**LM**) is provided, in order to cover cases where the system identifies a correct lemma but selects the wrong sense of the lemma.

4.1 System Performance

We observe in the results shown in Table 1 that the three runs performed similarly, with only a few noticeable differences which we comment as follows. First, *RunHeads* and *RunSenses* seem to perform similarly. This may be due to the fact that in some cases no candidate was obtained from the sense-based modeling and hence the system resorted to direct matching from definition and term heads[2]. Another reason can be found in cases in which the definition is semantically very compact, i.e. all its extracted chunks belong to a similar semantic field and therefore the centroid vector we obtain is the same as the definition head, and also the first synset of the definition. This is the case, for instance, the term "complex disease", with definition "A complex disease is caused by the interaction of multiple genes and environmental factors". Here, the gold synset was *disease#n#01*, and this answer was provided by c_t, $firstBest$, and μ_{best}.

Another interesting discussion can be derived from observing the difference between scores in **LM** by the first two systems, on one hand, and *RunHyps* on the other. While *RunHyps* ranks 2nd in our local rank, its **LM** score is more than 10 points below its closest competitor. This is due to the fact that, since *RunHyps* is a conservative approach more aimed at providing generic answers rather than matches, it shows a reasonable performance in terms of **W&P**, but falls short in **LM** as it almost never will find the correct point of attachment at the top of WordNet's hierarchy.

In terms of comparative evaluation, all our submitted systems rank below the baseline *Baseline-First*, which shows that a simpler approach based on majority class (selecting the first WordNet synset for a given lemma, which usually corresponds to the most generic or widespread sense) is difficult to beat. However, in comparison with the median of all participating systems, all our submissions achieve **F1** scores between 4 and 5 points higher.

[2] In general we were reasonably satisfied with the coverage of SENSEMBED, as with our approach, we found at least one candidate BABELNET synset with WordNet mapping in the vast majority of cases. For instance, our run *RunHeads* missed only two nouns and four verbs. In these cases we set a default point of attachment *entity#n#01* and *breathe#v#01* respectively.

Subm	W&P	LM	R	F1
RunHeads	0.476	0.36	1	0.645
RunSenses	0.463	0.353	1	0.635
RunHyps	0.471	0.24	1	0.641
BaselineFirst	0.513	0.415	1	**0.678**
Median	-	-	-	0.590

Table 1: Evaluation summary for our submission

5 Future Work

We computed an *upper bound* on the training data, i.e. the score we would have achieved if, for each case, we had selected the best of the candidates proposed by our system. This artificial upper bound reached a **W&P** score of almost 0.70, which suggests that our approach has clear and well defined room for improvement. We are currently investigating potential ways to address this issue in order to surpass the baseline scores.

Acknowledgments

This work was partially funded by Dr. Inventor (FP7-ICT-2013.8.1611383), and by the Spanish Ministry of Economy and Competitiveness under the Maria de Maeztu Units of Excellence Programme (MDM-2015-0502).

References

Bernd Bohnet. 2010. Very High Accuracy and Fast Dependency Parsing is not a Contradiction. In *COLING*, pages 89–97.

Luis Espinosa-Anke, Horacio Saggion, Francesco Ronzano, and Roberto Navigli. 2016. ExTaSem! Extending, Taxonomizing and Semantifying Domain Terminologies. In *AAAI*.

Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. 2015. SENSEMBED: Enhancing Word Embeddings for Semantic Similarity and Relatedness. In *Proceedings of ACL*, Beijing, China, July. Association for Computational Linguistics.

David Jurgens and Mohammad Taher Pilehvar. 2015. Reserating the awesometastic: An automatic extension of the WordNet taxonomy for novel terms. In *Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics: Human language technologies, Denver, CO*, pages 1459–1465.

David Jurgens and Mohammad Taher Pilehvar. 2016. Semeval-2016 Task 14: Semantic Taxonomy Enrichment. In *Semeval 2016*.

Tuan Luu Anh, Jung-jae Kim, and See Kiong Ng. 2014. Taxonomy Construction Using Syntactic Contextual Evidence. In *EMNLP*, pages 810–819, October.

Tristan Miller and Iryna Gurevych. 2014. WordNet - Wikipedia - Wiktionary: Construction of a Three-Way Alignment. In *LREC*, pages 2094–2100.

George A. Miller, R.T. Beckwith, Christiane D. Fellbaum, D. Gross, and K. Miller. 1990. WordNet: an Online Lexical Database. *International Journal of Lexicography*, 3(4):235–244.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. *Artificial Intelligence*, 193:217–250.

Mohammad Taher Pilehvar and Roberto Navigli. 2014. A Robust Approach to Aligning Heterogeneous Lexical Resources. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*, pages 468–478.

Rion Snow, Daniel Jurafsky, and Andrew Y Ng. 2006. Semantic taxonomy induction from heterogenous evidence. In *Proceedings of COLING/ACL 2006*, pages 801–808. Association for Computational Linguistics.

Paola Velardi, Stefano Faralli, and Roberto Navigli. 2013. Ontolearn Reloaded: A graph-Based Algorithm for Taxonomy Induction. *Computational Linguistics*, 39(3):665–707.

Zhibiao Wu and Martha Palmer. 1994. Verbs Semantics and Lexical Selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 133–138. Association for Computational Linguistics.