Deftor at SemEval-2016 Task 14: Taxonomy Enrichment using Definition Vectors

Hristo Tanev
Joint Research Centre, EC
via Enrico Fermi 2749
Ispra, Italy
hristo.tanev@jrc.ec.europa.eu

Agata Rotondi
Universita’ Ca’ Foscari
Venezia, Italy
rotondiagata@gmail.com

Abstract

In this paper we describe the participation of the Joint Research Centre, EC, in task 14 - Semantic Taxonomy Enrichment at SemEval 2016. The algorithm which we propose transforms each candidate definition into a term vector, where each dimension represents a term and its value is calculated by TF.IDF. We attach the candidate term as a hyponym to the WordNet synset with the most similar definition. The results we obtained are encouraging, considering the simplicity of our approach. The obtained F measure is below the average, but above one of the baselines.

1 Introduction

In this paper we describe the participation of the Joint Research Centre, EC, in task 14 - Semantic Taxonomy Enrichment at SemEval 2016. We particpate for the first time in a similar task. We opted for a relatively simple method for searching of relevant synsets, which does not exploit any external dictionary or another semantic resource. We called our system Deftor (DEFinition vecTOR). Deftor is a system which represents the definitions (glosses) as lexical vectors and finds the most similar one for each new lemma.

Automatic enrichment of taxonomies and knowledge bases is very important especially for rapidly changing domain. The taxonomy enrichment task is quite challenging, mostly because of the many possibilities when attaching a new term to an existing taxonomy: First, a new word can be attached as a hyponym to different concepts, which describe it at different levels of abstraction. For example, in WordNet the *hurricane* is a hyponym of *cyclone*, which is a hyponym of *windstorm*, which itself is a hyponym of *storm* and the *storm* is a hyponym of *atmospheric phenomenon*. It is not always easy to decide where to attach a concept: in the above mentioned case the definition of *storm* and *windstorm* are not very different. In this case, it is also difficult to decide if a new concept should be merged with a similar concept from WordNet or it should be attached as a hyponym. Another problem are the multiple aspects from which a concept can be perceived. For example, one can consider *hurricane* to be a *natural disaster*. It is also a *weather condition* or *cause of death*. All these considerations unfortunately make taxonomy enrichment task quite ambiguous and difficult to tackle. In some cases, the right attachment of a new concept will be difficult also for a human expert. Although the task is quite complicated, as we have pointed out, we applied a simple approach which is based on comparison of the lexical content of the definitions of the new concepts and the WordNet synsets.

Our approach to the taxonomy enrichment task represents each synset from WordNet and the candidate new terms as word vectors from their definitions and then attaches each new term as a hyponym to the synset for which the cosine similarity of its definition vector and the definition vector of the new term is the highest.

The algorithm which we propose transforms each candidate definition into a *definition vector*, a term vector, where each dimension represents a term and its weight is calculated by TF.IDF. In this way we

Proceedings of SemEval-2016, pages 1342–1345,
San Diego, California, June 16-17, 2016. ©2016 Association for Computational Linguistics

represent each WordNet definition, as well as the definitions of the new terms for inclusion in WordNet. Moreover, we expanded each definition vector with the definitions of the words from this vector.

We then calculated the cosine similarity between each WordNet synset definition and the definition of the candidate term whose place in the WordNet hierarchy is to be identified. Then, we attach the candidate term as a hyponym to the synset with the most similar definition.

Using this method we obtained results which are encouraging, considering the simplicity of our method - it relies solely on WordNet and no additional dictionaries or other resources were used. The results we obtained are somehow below the average, but above the weaker baseline.

2 Related Work

There are plenty of scientific papers, which address the taxonomy/ontology enrichment task and in particular the automatic enrichment of WordNet. See among the others (Haridy et al., 2010), (Navigli et al., 2004) and (Nimb et al., 2013). Existing work falls in one of the following categories: 1. Adding new semantic relations in an existing ontology (Montoyo et al., 2001). 2. Adding new senses for existing terms, e.g. (Nimb et al., 2013) 3. Adding new terms (Jurgens and Pilehvar, 2015). The new terms which are added may belong to already existing terminology (Stankovic et al., 2014), to a particular domain (e.g. biomedical (Poprat et al., 2008), medical (Smith and Fellbaum, 2004), or architectural (Bentivogli et al., 2004)), or they can belong to one well defined class like in (Toral and Monachini, 2008) who adds proper nouns to WordNet. The new terms may be taken from dictionaries or extracted from a corpus. In several cases the exploited resource is Wikipedia, like (Ruiz-Casado et al., 2005) and (Ponzetto and Navigli, 2009). Most of the work based on Wikipedia are limited mainly to the noun concepts; to overcome this limitation (Jurgens and Pilehvar, 2015) proposed to extend WordNet with novel lemmas from Wiktionary. For the WordNet enrichment task different resources have been exploited and different approaches have been experimented: distributional similarity techniques (Snow et al., 2006), structured based approaches (Ruiz-Casado et al., 2005), creation of a new ontology and its merging with the existing one by alignment based methods (Pilehvar and Navigli, 2014) or considering the attributes distribution (Reisinger and Paşca, 2009).

The taxonomy enrichment task can be considered as a specific case of the ontology learning and population task, (Buitelaar and Cimiano, 2008), whose purpose is automatic learning of semantic classes and relations. Ontology population is about finding instances, which belong to certain ontological classes, like *Paris* is an instance of the class *city*.

3 Algorithm

As we pointed out earlier our method is based on similarity of *definition vectors*. In order to create a definition vector for a word sense, we perform part-of-speech tagging of its gloss and we represent each gloss as a list of lemmas of its non-stop words. Words are downcased. After that, as a second step, each definition vector is being expanded with the lemmas from the glosses of its words, obtained on the first step. For example, if the WordNet definition for *computer* is *a machine for performing calculations automatically*, then our algorithm creates a first version of the definition vector with the non-stop lemmas *machine, perform, calculation* and the TF.IDF values of these words. Then, the algorithm takes the glosses of all the WordNet senses of the words in the first version of the vector. In this particular case, we will add to the definition vector of *computer* the words from the glosses of all the senses of *machine*, *perform*, and *calculation*. Moreover, part-of-speech tags of these words are known, since we perform part of speech tagging of the glosses. As an additional step of pre-processing we extract the genus from the gloss - usually the first word which defines the more generic concept under which is the defined term

We have processed all the WordNet synsets, where each synset is represented as a definition vector. Then an inverted index was created for each definition vector, in which a word points to the definition vectors in which it appears. Then, for each new term t, we do the following

1. Find the definition vector d of t.

2. For each word w from d we find via the inverted index all the synsets whose definition vectors contain w and whose part-of-speech is the same as the one of t. Let's denote the set of definition vectors of these synsets as D.

3. We find the similarity of d and each vector $d_i \in D$. The similarity is being calculated as $d.d_i.cos(d, d_i)$, this formula was empirically derived from the training data.

4. If the part of speech of t is verb, we add to the above-calculated similarity score the similarity of the glosses of the genus of t and the genus of the synset under consideration

5. The synset with highest similarity is taken and then the new term is attached as its hyponym. If the similarity of the best synset is found to be under a certain threshold, then we do not attach the new term and we skip it

4 Evaluation

The evaluation shows that our system can be improved significantly, but still results are encouraging considering the simplicity of our approach. The F1 of our only run was found to be 0.5132, which is below the baseline First word, first sense, but much above the baseline Random synset, which shows the feasibility of our approach.

It goes without saying that our results can be improved, nevertheless we propose an approach whose main advantage is its simplicity and it is independent of external resources. Our method is potentially multilingual and can be applied on taxonomies in languages other than English. Our system Deftor does not rely on any external knowledge and it uses only part-of-speech tagging and lemmatization as a pre-processing step. Since P.O.S. taggers and lemmatizers are available in different languages, we can easily adapt Deftor between languages and between domains, since the exploited algorithm does not depend on the taxonomy domain. Moreover, the simplicity of our approach make it quite efficient and easy to implement.

References

Luisa Bentivogli, Andrea Bocco, and Emanuele Pianta. 2004. Archiwordnet: integrating wordnet with domain-specific knowledge. In *Proceedings of the 2nd International Global Wordnet Conference*, pages 39–47.

Paul Buitelaar and Philipp Cimiano. 2008. *Ontology learning and population: bridging the gap between text and knowledge*, volume 167. Ios Press.

Shaimaa Haridy, Nagwa L Badr, Omar H Karam, and Tarek F Gharib. 2010. Enriching ontologies using coarse-grained word senses in the case of: Wordnet. *Egyptian Computer Science Journal*, 34(2).

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.

Andrés Montoyo, Manuel Palomar, and German Rigau. 2001. Wordnet enrichment with classification systems. In *Proceedings of WordNet and Other Lexical Resources: Applications, Extensions and Customisations Workshop.(NAACL-01) The Second Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 101–106.

Roberto Navigli, Paola Velardi, Alessandro Cucchiarelli, Francesca Neri, and Ro Cucchiarelli. 2004. Extending and enriching wordnet with ontolearn. In *Proc. 2nd Global WordNet Conf.(GWC)*, pages 279–284.

Sanni Nimb, Bolette S Pedersen, Anna Braasch, Nicolai H Sørensen, and Thomas Troelsgård. 2013. Enriching a wordnet from a thesaurus. *Lexical Semantic Resources for NLP*, page 36.

Mohammad Taher Pilehvar and Roberto Navigli. 2014. A robust approach to aligning heterogeneous lexical resources. *A A*, 1:c2.

Simone Paolo Ponzetto and Roberto Navigli. 2009. Large-scale taxonomy mapping for restructuring and integrating wikipedia. In *IJCAI*, volume 9, pages 2083–2088.

Michael Poprat, Elena Beisswanger, and Udo Hahn. 2008. Building a biowordnet by using wordnets data formats and wordnets software infrastructurea failure story. *Software Engineering, Testing, and Quality Assurance for Natural Language*, page 31.

Joseph Reisinger and Marius Paşca. 2009. Latent variable models of concept-attribute attachment. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the*

AFNLP: Volume 2-Volume 2, pages 620–628. Association for Computational Linguistics.

Maria Ruiz-Casado, Enrique Alfonseca, and Pablo Castells. 2005. Automatic assignment of wikipedia encyclopedic entries to wordnet synsets. In *Advances in web intelligence*, pages 380–386. Springer.

Barry Smith and Christiane Fellbaum. 2004. Medical wordnet: a new methodology for the construction and validation of information resources for consumer health. In *Proceedings of the 20th international conference on Computational Linguistics*, page 371. Association for Computational Linguistics.

Rion Snow, Daniel Jurafsky, and Andrew Y Ng. 2006. Semantic taxonomy induction from heterogenous evidence. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 801–808. Association for Computational Linguistics.

Staša Vujicic Stankovic, Cvetana Krstev, and Duško Vitas. 2014. Enriching serbian wordnet and electronic dictionaries with terms from the culinary domain. *Volume editors*, page 127.

Antonio Toral and Monica Monachini. 2008. Named entity wordnet. In *In Proceedings of the 6th International Conference on Language Resources and Evaluation*. Citeseer.