JUNLP at SemEval-2016 Task 13: A Language Independent Approach for Hypernym Identification

Promita Maitra and **Dipankar Das**
Department of Computer Science Engineering, Jadavpur University
Kolkata, India
promita.maitra@gmail.com and dipankar.dipnil2005@gmail.com

Abstract

This paper describes our approach to build a language-independent hypernym extraction system, based on two modules for the SemEval-2016 Task 13 on Taxonomy Extraction Evaluation (TExEval-2). This task focuses only on the hypernym-hyponym relation extraction from a list of terms collected from various domains and languages. The first module of our system is built on the state-of-the-art system using BabelNet while the second one deals with the parts found within terms and which are useful to establish a hierarchical relation among them. Our system performed well in terms of *recall* in most of the domains irrespective of the languages; however, the precision scores indicate a scope of improvement. In case of overall ranking, our present system stands fourth in monolingual (i.e. English) evaluation and second in multilingual (i.e. Dutch, Italian, French) setup.

1 Introduction

The rapid growth in terms of digitalized texts in the recent years (specially, in the fields including scientific, clinical, enterprise, legal, and personal information management) has made the management of textual information increasingly important. In order to fulfill the need of having more structured data, ontologies, taxonomy or hierarchical relations between ontological concepts are considered as useful tools for content organization, navigation, and retrieval, as well as to provide valuable input for semantically intensive tasks such as question answering and textual entailment. This task specifically focuses on the identification of hypernym-hyponym relation among terms in four different languages (English, Dutch, Italian and French) and different domains. Typically, taxonomy construction has three basic steps: entity or concept identification, discover different relations among different entities and taxonomy construction. The challenge organizers have made it easy by already providing us with a list of extracted entities (also called concepts/terms) for each of the domains in each of the languages. Our approach was to build a system that fits a multilingual setup and significantly reduces the computational time and complexity by taking existing resources into consideration.

The rest of the paper is organized as follows: Section 2 describes the task at hand in brief along with the main challenges. Section 3 gives a brief overview of the work already done in this field. Section 4 contains the details of the modules we have used to address the problem. Section 5 presents the analysis of results and finally, the conclusion and future scopes are discussed in Section 6.

2 Problem Description

SemEval-2016 Task 13: Taxonomy Extraction Evaluation (TExEval-2) has its main focus on hypernym-hyponym relation extraction from given lists of terms collected from multiple domains like Food, Environment and Science (Bordea et al., 2016). This year, the task organizers have extended the problem setup to address the multilingual structure. Along with English, there were terms in French, Dutch and Italian as well for all the domains. For this particular task, we did not have to go through the complexities of entity identification from a text

Proceedings of SemEval-2016, pages 1201–1205,
San Diego, California, June 16-17, 2016.

as the lists of terms were already given.

- i. One of the main challenges was that we were not provided with any annotated or plaintext corpus that we can use as training. However, the organizers suggested that it would be helpful if we explore the Wikipedia dump for the same.

- ii. Second big challenge was to develop a system that will work for languages we do not understand. Ontology development being such a task where some basic domain knowledge is inevitable, this multilingual setup was indeed a great concern for us.

- iii. We were specifically asked not to use the resources we most frequently use in this kind of tasks as they were used to construct the gold standard. The list of the resources that were prohibited is:

  - hypernym-hyponym relations from the WordNet [1],

  - skos:broader and skos:narrower relations from EuroVoc [2],

  - the Google product taxonomy [3],

  - the Taxonomy of fields and their subfields provided for the National Academies of Sciences, Engineering, and Medicine [4].

However, in contrast, we were free to add more terms if needed to the term lists that were provided by the organizers.

3 Related Work

Hypernym detection from text is one of the most popular hierarchical relation extraction tasks in ontology learning for which research work dates back to at least 1984 (Calzolari, 1984). Hypernym can be described as a linguistic term for a word whose meaning includes the meanings of other words, which are known as hyponyms. For instance, *flower* is a hypernym of *daisy* and *rose*. On the other hand, *daisy* and *rose* are some of the hyponyms of *flower*. In simple words, this relation deals with identifying the concepts and finding the particular superclass they fit in. Manually constructing these kind of relations from text is a time-consuming and labour-intensive procedure. Hence, the researchers felt the need to make this process automatic. The methods proposed can broadly be categorized into two: supervised and unsupervised. While the unsupervised methods can identify and extract semantic relations from plain text without the need of any pre-annotated text corpora, the supervised methods often find it difficult to find an annotated corpora in similar domain. A major part of the previous researches on automatic semantic classification of words was developed based on the method first proposed by Hearst, that the presence of certain lexico-syntactic patterns can indicate a particular semantic relationship between two noun phrases (Hearst, 1992). This paper introduced six basic lexical patterns. This rule based approach was further extended in subsequent works bringing out more valid patterns, either handcrafted or learned from training corpus for semantic relation extraction (Berland and Charniak, 1999) (Kozareva et al., 2008) (Widdows and Dorow, 2002). Pattern based results are effective and reliable, scores high on precision measure. However, these methods suffer in terms of recall. Later, a few distributional approaches were proposed by different authors making use of the large corpora present in (Guido Boella, 2014) (Navigli and Velardi, 2010). Machine learning based methods make use of features like term co-occurrence, semantic similarity or other syntantic information from text collection. Precision of these machine learning based approaches is lower compared to pattern based approach. The simple morpho-syntactic approach also proved to yield a decent result (Lefever, 2015) (Sang et al., 2011). In recent years, several researches are being carried out to extract semantic relations from texts in other languages.

4 System Description

In the present challenge, we had to keep three main points in mind. We wanted to make a single system appropriate for a multilingual setup (Dutch, French, Italian and English). However, it became

[1] https://wordnet.princeton.edu/

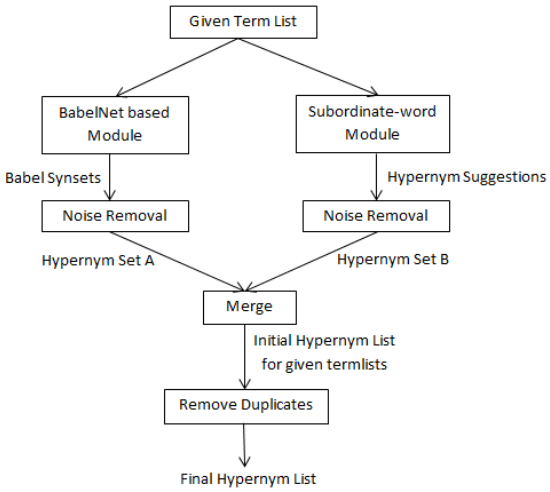
[2] http://eurovoc.europa.eu/

[3] https://www.google.com/basepages/producttype/taxonomy.en-US.txt

[4] http://sites.nationalacademies.org/PGA/Resdoc/PGA_044522

more difficult as we were not allowed to use any of the widely used resources like WordNet, EuroVoc, Google Product Taxonomy etc. Building a taxonomy which would provide structured information about semantic relations between words is an extremely slow and labor-intensive process. Therefore, we kept our focus on building a system which would be simple and significantly light in terms of computation time.

Figure 1: Basic system diagram



Our system has two main modules, as shown in Figure 1:

- i. Extracting semantic relations from BabelNet.
- ii. Analyzing the terms to find a subterm suitable to become a hypernym.

4.1 BabelNet Based Module

BabelNet[5] is an open source resource containing both multilingual dictionary with lexicographic and encyclopedic coverage of terms, and a network of concepts and named entities connected in a very large network of semantic relations, called Babel synsets (Navigli and Ponzetto, 2012). Each of the Babel synsets represents a given meaning and contains all the synonyms which express that meaning in a range of different languages. BabelNet 3.5 covers 272 languages, which also include our task related languages like English, French, Dutch and Italian.

[5]http://babelnet.org/

Finding out semantic relations from the entire Wikipedia dump with a pattern based approach proved to be quite a long process and computationally expensive as there can be numerous types of valid patterns that can hold a hypernym-hyponym relation. On the other hand, it would take days to initially start with a few patterns and then search for more with a bootstrapping approach. On the other hand, BabelNet already provides a variety of semantic relations for a large number of concepts using knowledge from various resources available including Wikipedia. So, our system execution time gets significantly reduced if we just use the semantic relation set available in BabelNet instead of extending the Wikipedia corpus and analyzing it for the pattern search. Secondly, we wanted to have a system that would fit into the multilingual setup that the task intends to have this year. The facts were that we do not have a satisfactory amount of knowledge required for identifying the valid patterns for hypernym-hyponym relations in languages other than English, and we also do not have an annotated training data to learn those patterns via a bootstrapping method for those languages. Therefore, it was essential for us to have a tool that could automatically extract such knowledge from corpus.

For each term that appears in each domain, we obtain a synset from the BabelNet for hypernym relations found over the Wikipedia articles in different languages. As the task organizers specifically asked not to use resources like WordNet, only the Wikipedia source is taken into account while obtaining hypernym relations, skipping the others like WordNet, VerbNet, Microsoft Terminology etc. This is done to make the system computionally light and reduce the huge time needed to process the Wikipedia articles searching for patterns. We consider the terms for their NOUN POS tag sense, with the language mentioned in the query. We only considered the NOUN POS tags because it was seen from our observation of term lists, that they contain terms which are mostly nouns. We get the synset for each term which contains a lot of noise such as repetitive sense words, out-of-domain senses, senses in different morphological form than the existing terms, etc. We fed the raw synset output to a cleansing module which would give us only the unique in-domain terms in their correct morphological form as

given in the term-list.

We further extended this module to find the synsets of the terms present in the cleansed output in order to obtain the entire hypernym tree for the given term which helps to increase the recall of our system.

4.2 Subordinate-word Module

This module deals with finding appropriate parts of given terms that can possibly be the hypernym of the original term. For example, *Fruit Custard* is a type of *Custard*. Now these subordinate-words which are potential hypernyms can be of the following two types.

- The subordinate-word can itself be an independent term present in the term-list given. For example, if we have both the terms *Biochemistry* and *Chemistry* in the term-list, we can just analyze the term *Biochemistry* and identify *Chemistry* as its possible hypernym.

- There might be multiple terms for which no common part is an independent term but significant overlap exists among those, even more than once. In such cases we have introduced that overlapped part as our new term in the term-list. For example, we have *Chocolate Pudding* and *Vanilla Pudding* as two terms in our list but no entry for *Pudding*. Since we get overlapping in previous two terms with *Pudding*, we can consider *Pudding* as the possible hypernym of *Chocolate Pudding* and *Vanilla Pudding*.

However, the problem is that we were getting some noise in the input due to the stopwords present in the list. For example, University of PlaceA and University of PlaceB will have University and of as the subordinate-word hypernyms. Of cannot be a hypernym to some term. So we remove those subordinate-words which has only stopwords in them. Again, we had to deal with different morphological forms of the same word as hypernyms, for example science and sciences. For such instances, we checked if any one form is the part of our term list. If yes, we keep that form and remove others or we keep the lemmatized form otherwise.

5 Analysis of Results

Just as construction of suitable ontology from text, evaluation of an extracted ontology is not a simple task either. For this particular task, structural evaluation was done which includes the presence of cycles, the number of intermediate nodes compared to leaf nodes, and the number of over generic relations with the root node. The output relations were also evaluated against collected gold standards collected from WordNet and other well known, openly available taxonomies using evaluation measures like standard precision, recall and Fscore.

Language	Precision	Recall	Fscore
English	0.15	0.30	0.20
Dutch	0.16	0.22	0.19
French	0.17	0.25	0.20
Italian	0.13	0.20	0.19

Table 1: Average Precision, Recall, Fscore for Gold standard evaluation across all domains.

Table1 shows the average result of our system with respect to the gold standard evaluation for each language taking an average over all the domains. We had our focus on generating a hypernym tree for each term by providing the hypernyms of a term as next input to the system. This resulted in better recall but the precision of our system showed a visible decline compared to the baseline system for all the languages.

Table 2 shows the structural evaluation of the output produced by our system for different domains in English and other languages. The structural measures used for the evaluation are as follows:

- V: number of distinct vertices;

- E: number of distinct edges;

- c.c.: number of connected components;

- i.i.: intermediate nodes = V - L where L is the set of leaves

- cycles: YES = the taxonomy contains cycles, NO = the taxonomy is a Directed Acyclic Graph (DAG)

- Cumulative Fowlkes and Mallows Measure (FM): cumulative measure of the similarity of two taxonomies.

As we can see, though we have cycles present in relations of English language, all other language-output is a DAG. We achieved better score in categorization due to high number of distinct vertices, edges and intermediate nodes obtained by our system.

Measure	English	Multilingual
Cyclicity	3	0
Structure (FM)	0.1498	0.0155
Categorisation (i.i.)	377	178.22
Connectivity (c.c.)	53.17	34.89

Table 2: Structural Evaluation for English and other languages.

6 Conclusion and Future Scope

In this paper we have discussed in brief our approach to address the Taxonomy Extraction and Evaluation task of SemEval 2016. We have kept our focus mainly on keeping the computation time optimal and building a system suitable for multilingual setup. We took a completely unsupervised approach without directly considering a text corpora. Both of our modules need only the terms to be fed as input and generates possible hypernym candidate for each term in a short span of time compared to the huge amount of time and knowledge needed to manually craft the lexico-syntactic pattern/regular expressions in each domain and then analyzing the large corpora for possible matches. The results were good in terms of recall, but the precision score suffered due to our approach of generating as many possible in-domain hypernyms of a term as we can.

We can try to improve our system's performance by making use of information available with Wikipedia dump other than the article texts such as infobox properties, redirect links, article titles, categories or other meta-information available. Also, provided a training set, we believe a bag-of-word model constructed within a specific context window can yield better overall results.

References

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

Nicoletta Calzolari. 1984. Detecting patterns in a lexical data base. In *10th International Conference on Computational Linguistics and 22Nd Annual Meeting on Association for Computational Linguistics*, page 170. Association for Computational Linguistics.

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

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proc. of the Fourteenth International Conference on Computational Linguistics*, Nantes, France.

Zornitsa Kozareva, Ellen Riloff, and Eduard Hovy. 2008. Semantic class learning from the web with hyponym pattern linkage graphs. In *In Proceedings of ACL-08*, volume 8, page 1048. Association for Computational Linguistics.

Els Lefever. 2015. Lt3: A multi-modular approach to automatic taxonomy construction. In *In Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, page 944. Association for Computational Linguistics.

Roberto Navigli and Simone Paolo Ponzetto. 2012. The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Elsevier*, page 217.

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

Erik Tjong Kim Sang, Katja Hofmann, and Maarten de Rijke. 2011. *Extraction of Hypernymy Information from Text*.

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