QASSIT at SemEval-2016 Task 13: On the Integration of Semantic Vectors in Pretopological Spaces for Lexical Taxonomy Acquisition

Guillaume Cleuziou
Université d'Orléans,
INSA Centre Val de Loire,
LIFO EA 4022, France
cleuziou@univ-orleans.fr

Jose G. Moreno
LIMSI, CNRS,
Université Paris-Saclay,
F-91405 Orsay
moreno@limsi.fr

Abstract

This paper presents our participation to the SemEval "Task 13: Taxonomy Extraction Evaluation (TExEval-2)" (Bordea et al., 2016). This year, we propose the combination of recent semantic vectors representation into a methodology for semisupervised and auto-supervised acquisition of lexical taxonomies from raw texts. In our proposal, first similarities between concepts are calculated using semantic vectors, then a pretopological space is defined from which a preliminary structure is constructed. Finally, a genetic algorithm is used to optimize two different functions, the quality of the added relationships in the taxonomy and the quality of the structure. Experiments show that our proposal has a competitive performance when compared with the other participants achieving the second position in the general rank.

1 Introduction

The task of automatic taxonomy extraction consists in the generation of hierarchical relations between pairs of terms from a given initial set of terms. In this paper, we describe the second participation of QASSIT team to the "Taxonomy Extraction Evaluation (TExEval)" task. In this opportunity, we were interested in re-examinating our proposed algorithm (Cleuziou and Dias, 2015)(Cleuziou et al., 2015), but with an improved range of information as input. Our algorithm is based on Pretopological Spaces combined with patterns and collocations counts. Patterns are a useful way to capture knowledge form huge corpus (Hearst, 1992), but its performance could be influence by the size of the corpus or the use of a fixed vocabulary. This is a main drawback of the previous version of the algorithm (Cleuziou et al., 2015), the requirement of non-zero mentions in a corpus of the input terms in the specific chosen patterns. For example, to an optimal performance our algorithm requires that the term "computational biology" appears at least once in one of the used patters. However, when searching in wikipedia with the query "*computational biology* is a ..." then we obtain results such as "PLOS **Computational Biology** is a peer-reviewed computational biology journal", "The Journal of **Computational Biology** is a monthly peer-reviewed scientific journal", "Cancer **computational biology** is a field that aims to determine the future mutations in" and "**Computational biology** is a specialized domain that often requires knowledge of computer programming"[1]. The first three results include the desired term as a part of another more specific term (in this case journal titles) and only the last result could be considered as relevant information. Moreover, neither "domain" or "computer programming" are part of the input terms for the taxonomy. This situation makes difficult the direct use of patterns with our algorithm for the taxonomy construction task.

In order to overcome this difficulty, we have redefined our algorithm to exclusively use as input information gathered from the corpus without the use of specific query patterns. Additionally, we have explored the integration of recent techniques in words representation known as semantic vectors (Mikolov

[1] Only journal titles are obtained when the pattern is replaced by "is a**n** ".

Proceedings of SemEval-2016, pages 1315–1319,
San Diego, California, June 16-17, 2016.

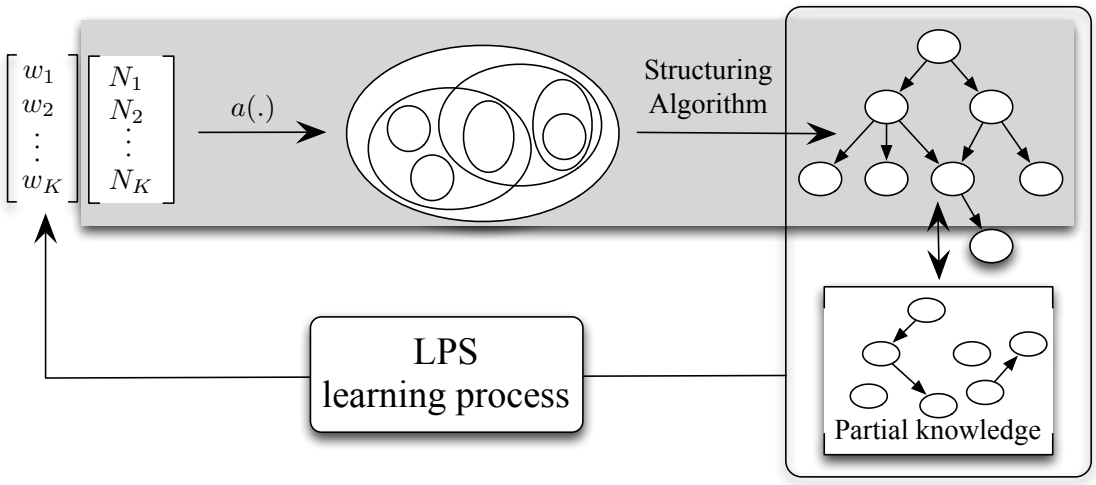

Figure 1: The LPS process uses partial knowledge on the expected structure in order to improve the parameterization of the pretopological space.

et al., 2013). This technique allows us to calculate similarities of the input terms if they are present in a corpus, avoiding the requirement of explicit patterns of the previous versions. Additionally, these vectors allow semantic similarity calculation of concepts that are not found together, but that their context does. The remainder of this paper includes a brief description of our pretopological spaces algorithm and their modifications for the integration of semantic vectors (Section 2). Experiments and results are presented in Section 3 and finally, discussion and conclusions are presented in Sections 4 and 5 respectively.

2 Pretopological Spaces for Lexical Taxonomy Acquisition

We used the learning pretopological spaces framework (LPS) proposed in our previous participation in the TExEval task and fully described in (Cleuziou and Dias, 2015). The general LPS framework is illustrated in Figure 1, and a brief description is presented below. This algorithm considers as input a set of non-symmetric binary relations $\{N_1, \ldots, N_K\}$ over a set of terms E and a partial knowledge S used as true partial information to structure E; LPS aims to find a $Space\ w$ which induces a good subsumption propagation function $a()$ for structuring E; the fitness function that guides the learning process is defined by:

$$Score(w, S) = F_{meas.}(w, S) \times I_{struct.}(w) \quad (1)$$

where F and I are two terms quantifying respectively the satisfactions about:

- the matching with the partial knowledge S and
- a taxonomy structural property expected as output (e.g. a tree-like structure).

The score defined in Equation 1 is used to guide the exploration of the space of solutions through a learning strategy based on a Genetic Algorithm (GA).

2.1 Sources used for LPS Taxonomy Acquisition

In our previous algorithm, we have used patterns and collocation measures as piece of information to model the subsumption propagation between terms. In this version, we abandon the use of patterns due the additional extra task it requires to avoid noisy information. However, we introduce another source of information with low manual requirements in order to have a more automatic version of our algorithm. The three main sources of information we use are described bellow.

First, we have integrated a new robust semantic representation called semantic vectors or word

Table 1: Comparative automatic evaluation of the proposed taxonomies.

Dataset (Domain)	Measure	Best	QASSIT	Rank
Environment (Eurovoc)	Fscore	**0.2992**	0.1725	4/5
	F&M	0.2384	**0.4349**	1/5
Science	Fscore	**0.3669**	0.2165	3/5
	F&M	0.3634	**0.5757**	1/5
Science (Eurovoc)	Fscore	**0.3118**	0.2431	3/5
	F&M	**0.3893**	**0.3893**	1/5
Science (Wordnet)	Fscore	**0.3776**	0.2384	4/5
	F&M	**0.2255**	**0.2255**	1/5

embedding (Mikolov et al., 2013). These vectors are extracted using an unsupervised framework from large amounts of text information. The main characteristic of these vectors are their ability to encode semantic similarities in a rather small vector (300 dimensions) for each word or multi-word present in the corpus. In our experiments we have used a set of pre-trained vectors over an open domain collection [2]. We have searched by each corresponding term and assigned only one vector to it. If the term is a composed term and it is not found in the pre-trained set them the sum of vectors related to each word that compose the term is used. Following this strategy near to 95% of the terms have an assigned vector. Similarities between vectors are computed using the cosine similarity.

Second, we have extracted the collocation values between terms, to do so we have used the english subpart of wikipedia.org for frequency counts extraction. For each pair of terms (x, y), we retrieve the number of wikipedia pages where both terms occur (hits(x, y)). For example, hits(memory, politics) is retrieved with the following query [“memory” AND “politics”].

Finally, the partial knowledge[3] has been obtained by first extracting a list of candidate subsumption pairs observing suffix matching and then by manually correcting the candidate list and/or adding new pairs of subsumptions with the aim to reach at least two hundreds subsumption relations into S.

Each of the three previous sources led to a couple of (non-symetric) binary relations over the set of terms E. Finally, sixteen binary relations are given as input of the LPS framework in order to learn a relevant pretopological space.

3 Experiments and results

Manual and automatic evaluation were performed using four datasets[4]. Further details could be found in (Bordea et al., 2016). The automatic evaluation is based on the comparison of the proposed taxonomy against the respective gold standard. Several automatic metric were used by the task organizers. However, we have focused on more robust metrics such as $Fscore = 2(P * R)/(P + R)$ and $F\&M$. Results for the best[5] participant (Best column) and ours (QASSIT column) are reported in Table 1. Similarly, manual quality evaluation is performed over a random selection of hundred ISA relationships found in the proposed taxonomies. Each of these hundred relationship is binary evaluated as relevant or not. The used metric is $P_m = |correctISA|/100$ which calculates the accuracy of the taxonomies. Results for the best participant (Best column) and ours (QASSIT column) are presented in Table 2. In both tables the column *Rank* correspond to the obtained rank when compared with the other participants.

4 Discussion

In the general ranking our algorithm achieves the second position over the five participants in the Taxonomy Extraction Evaluation task; full results are

[2]Publicly available at https://code.google.com/archive/p/word2vec/ .

[3]Note that this is an expensive manual task. In future version we plan to eliminate this stage.

[4]The task include six datasets, however we have participated only in four of them.

[5]Best from the results obtained by (Tan et al., 2016), (Pocostales, 2016), JUNLP and (Panchenko and Biemann, 2016).

Table 2: Comparative manual evaluation over 100 randomly selected candidates.

Dataset (Domain)	Best	QASSIT	Rank
Environment (Eurovoc)	**0.22**	0.07	4/5
Science	**0.71**	0.07	4/5
Science (Eurovoc)	0.04	**0.05**	1/5
Science (Wordnet)	**0.47**	0.22	2/5

included in (Bordea et al., 2016), but it can be partially observed in Table 1. Our results outperform other participants in terms of F&M, but fails to obtain a similar performance in terms of $Fscore$. Note that the average of the column rank is 2,25 indicating that if only these metrics where considered our position remains the same. Indeed, in the evaluation many others factor were evaluated such as: Cyclicity, Categorisation (i.i.), Connectivity (c.c.) and domains. Our results are also good in all them where we obtain first or second best performance, except for Categorisation (i.i.) where we are the least performing team. Indeed, this can be explained by our choice of these objective functions that are driven by the partial knowledge. Due to the cost required to generate this partial knowledge, the manual annotators tend to relate one concept to many concepts which force some flatness in our taxonomy. However, the $I_{structure}$ objective criteria tends to force the acquisition of a hierarchical structure. In future experiments we plan to include additional objectives to avoid this situation.

In terms of manual evaluation, we obtain good results only for the *Science (Eurovoc)* dataset where our algorithm get the best performance. One explanation is the random selection of only hundred relationships to evaluate. Note that this sample is small compared with the actual number of terms in each dataset. However, only one of the participants manage to get good results in the manual evaluation. For the *Science dataset*, the ordered results of all participants are: 0.71, 0.14, 0.09, 0.07 and 0.06. Note that our performance, 0.07, is not very far from 3th and 5th position, but clearly far form the first one. This situation is quite similar for the *Enviromment (Eurovoc)* dataset, where the missing values of Table 2 are 0.02, 0.08 and 0.11. Note that, again, the best performance is clearly far from the other participants. A deeper analysis over the selection of the sample for manual annotation or a full manual evaluation[6] must be performed to grasp a better understanding of these differences.

5 Conclusion

In this paper we presented the participation of the QASSIT team in the "Taxonomy Extraction Evaluation (TExEval-2)" task. Our strategy is based on a semi-supervised pretopological framework that learns a subsumption propagation process over a set of terms described by association measures and semantic vectors as input. Our results achieve the best performances in terms of $F\&M$ metric over the four datasets. In the general ranking, our algorithm achieved the second position. In terms of the manual evaluation, we obtained the first position for the Science (Eurovoc) dataset, second position for the Science (Wordnet) and fourth position for the remaining two datasets. Results encourage us to continue the exploration of strategies based on the theory of pretopoly and their combination with external resources. However, in future versions of our algorithm we plan to eliminate or automatize the partial knowledge extraction to have a fully automatic taxonomy construction framework.

References

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.

Guillaume Cleuziou and Gaël Dias. 2015. Learning pretopological spaces for lexical taxonomy acquisition. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2015, Porto, Portugal, September 7-11, 2015, Proceedings, Part II*, pages 493–508.

Guillaume Cleuziou, Davide Buscaldi, Vincent Levorato, Gaël Dias, and Christine Largeron. 2015. QASSIT: A Pretopological Framework for the Automatic Construction of Lexical Taxonomies from Raw Texts. In *International Workshop on Semantic Evaluation (SEMEVAL 2015)*, Denver, United States.

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the*

[6] This clearly increases the annotation cost for the task organizers.

14th Conference on Computational Linguistics - Volume 2, COLING '92, pages 539–545.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119.

Stefano Ruppert Eugen Remus Steffen Naets Hubert Fairon Cedrick Ponzetto Simone Paolo Panchenko, Alexander Faralli 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*. Association for Computational Linguistics.

Joel Pocostales. 2016. NUIG-UNLP at SemEval-2016 Task 13: A Simple Word Embedding-based Approach for Taxonomy Extraction. In *Proceedings of the 10th International Workshop on Semantic Evaluation*. 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.