QASSIT: A Pretopological Framework for the Automatic Construction of Lexical Taxonomies from Raw Texts

Guillaume Cleuziou[1], **Davide Buscaldi**[2], **Gael Dias**[3] **Vincent Levorato**[4], **Christine Largeron**[5]
[1] LIFO - University of Orléans, France
[2] LIPN - University of Paris 13, France
[3] GREYC - University of Caen-Basse Normandie, France
[4] IRISE - CESI Orléans, France
[5] LHC - University of Saint-Etienne, France

Abstract

This paper presents our participation to the SemEval Task-17, related to "Taxonomy Extraction Evaluation" (Bordea et al., 2015). We propose a new methodology for semi-supervised and auto-supervised acquisition of lexical taxonomies from raw texts. Our approach is based on the theory of pretopology that offers a powerful formalism to model subsumption relations and transforms a list of terms into a structured term space by combining different discriminant criteria. In order to reach a good pretopological space, we define the Learning Pretopological Spaces method that learns a parameterized space by using an evolutionary strategy.

1 Introduction

Lexical Taxonomies (LTs) play an essential role in Information Retrieval (IR) and Natural Language Processing (NLP). By coding the semantic relations between terminological concepts, LTs can enrich the reasoning capabilities of applications in IR and NLP. However, the globalized development of semantic resources is largely limited by the efforts required for their construction (Kozareva and Hovy, 2010). As a consequence, instead of manually creating LTs, many research studies have been appearing to automatically learn such structures (Buitelaar et al., 2005; Biemann, 2005; Cimiano et al., 2009; Kozareva and Hovy, 2010; Velardi et al., 2013).

The two main stages for the automatic construction of LTs are Term Extraction and Term Structuring. The proposed approach is focused on the second stage, thus matching with the aim of the SemEval task, by inducing LTs from pre-existing lists of terms (provided by the organizers).

As starting point, we consider the work from (Cleuziou et al., 2011) which introduced a set of new statistically-based criteria (e.g. Nearest-Neighbor-like relations) and combined them using the theory of pretopology (Brissaud, 1975). This formalism offers a new framework to model the subsumption relation at the term set level rather than considering (binary) subsumption relations only between pairs of terms.

Based on the concepts of (pseudo-)closure and closed subsets they transform the list of terms into a semantic space. A structuring algorithm based on the work of (Largeron and Bonnevay, 2002) is then applied to transform the semantic space of terms into a LT i.e. an acyclic directed (non-triangular) graph.

This theory should allow to combine both associative- and pattern-based methods within a virtuous multi-criteria structuring process. To achieve this objective, we consider pretopology on the multi-criteria analysis point of view, where criteria are statistical indices and linguistic patterns retrieved from a corpus. In particular, we define the concept of *Parameterized pretopological space* (P-space), where parameters express the confidence that exists over each criterion. As such, LT induction can be viewed as learning the set of parameters (confidences), which best (1) approximates the expected LT structure and (2) verifies a given number of linguistic patterns constraints.

In order to learn the parameters, we define a new *Learning Pretopological Spaces* (LPS) method and

use an evolutionary strategy which leads to induce a LT from an "optimized" P-space.

In the remaining of this paper, we first introduce the new concept of P-Space in Section 2. Then, we present the general LPS learning process in Section 3. Finally, we describe in Section 4, the use of the LPS paradigm in the particular context of the SemEval Task-17 and discuss the obtained results.

2 Pretopology and P-Spaces

Pretopology is a theory introduced by (Brissaud, 1975) that generalizes both Topology and Graph theories. This formalism, as reviewed by (Belmandt, 2011) is commonly used to model complex propagation phenomena thanks to a pseudo-closure operator, recently employed in (Cleuziou et al., 2011) for LT acquisition.

Let us consider a non-empty set E, and its powerset $\mathcal{P}(E)$. A (V-type) pretopological space is noted (E, a), where $a(.)$ is a pseudo-closure function ($\mathcal{P}(E) \rightarrow \mathcal{P}(E)$) such that :

i) $a(\emptyset) = \emptyset$,

ii) $\forall A \in \mathcal{P}(E), A \subseteq a(A)$,

iii) $\forall A, B \in \mathcal{P}(E), A \subseteq B \Rightarrow a(A) \subseteq a(B)$.

It is crucial to notice that $a(.)$ is not necessarily idempotent unlike in Topology (where $a(a(A)) = a(A)$). So, the pseudo-closure behaves as an expansion operator that enlarges any non-empty subset $A \subset E$. As a consequence, successive applications of $a(.)$ on A lead to a fix-point, called *closed subset* and noted F_A (or $F(A)$). At this stage, the reader has to consider E as a set of terms to structure and the pseudo-closure operator $a(.)$ modeling the propagation of the term domination (or subsumption) relation.

Let us also define the notions of *elementary closed subset* ($F_{\{x\}}$) that refers to the closure of a singleton that is maximal if $\nexists y,\ F_{\{x\}} \subset F_{\{y\}}$. In the scope of LT acquisition, these concepts will be used to model the domination/subsumption inheritance between terms, $F_{\{x\}}$ referring to a set of terms dominated by a term x that has no dominator when $F_{\{x\}}$ is maximal.

In order to perform the expansion process, we define a *P-Space* as a V-type pretopological space with a parameterized pseudo-closure function $a(.)$ defined for any $A \in \mathcal{P}(E)$ by

$$a(A) = \{x \in E \mid \sum_{N_k \in \mathcal{N}} w_k.\mathbb{1}_{N_k(x) \cap A \neq \emptyset} \geq w_0\} \quad (1)$$

with $\mathcal{N}$ a family of neighborhoods over E and such that $w_0 > 0$, $\sum_{k=1}^{K} w_k \geq w_0$ and $\forall k \neq 0,\ w_k \geq 0$.

Here, a neighborhood can be viewed as a statistical indice or a linguistic pattern retrieved from a corpus which identifies a subsumption relation between terms. In particular, each parameter w_k in (3) quantifies a kind of "reliability" on the k^{th} neighborhood and w_0 represents a global required confidence to expand the subset A. Thus, a subset A will be expanded to an element x only if the sum of the confidences on the criteria in agreement with the expansion exceeds the global required confidence w_0. The P-Space concept thus offers a wide range of neighborhood combinations by considering the set of any monotonic linear threshold functions.

Given a V-type pretopological space, (Largeron and Bonnevay, 2002) proposed an algorithm that structures the set E into a DAG (Directly Acyclic Graph).

3 Learning P-Spaces process (LPS)

We propose a learning pretoplogical spaces framework (LPS), illustrated in Figure 1. Considering a partial knowledge S providing a true partial structuring on E, LPS aims to find a P-Space - namely a function as in (3) and more concretely a set of parameters $\mathbf{w}$ - inducing a good structuring according to a fitness function defined by :

$$Score(\mathbf{w}, S) = F_{Measure}(\mathbf{w}, S) \times I_{structure}(\mathbf{w}) \quad (2)$$

with F and I, two terms quantifying respectively the satisfactions about :

(1) the constraints implied by the partial knowledge S and

(2) the expected structural properties of the output : a taxonomy-like structuring in the specific LT acquisition context.

The score (2) is used to guide the exploration of the space of solutions through a learning strategy based on a Genetic Algorithm (GA).

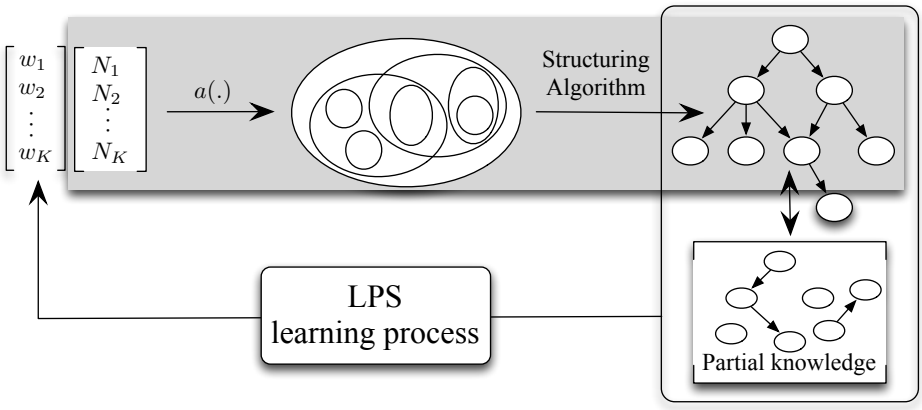

Figure 1: The LPS process uses partial knowledge on the expected structure in order to improve the parameterization of the pseudo-closure operator.

4 LPS for the SemEval Task 17

Let us recall that, in addition to the list of terms E to structure, the LPS system requires as input : a family of neighborhoods $\mathcal{N}$ over E and a partial knowledge S.

Three kinds of associative criteria served as basis neighborhoods :

N_{kSand} corresponds to the subsumption relation modeled by (Sanderson and Croft, 1999) : $y \in N_{kSand}(x)$ iff $P(y|x) \approx \frac{hits(x,y)}{hits(x)} \geq \sigma_k \wedge P(y|x) > P(x|y)$.

N_{kNP} associates to each term x its k Nearest Parents in the sense of $P(y|x)$: $y \in N_{kNP}(x)$ iff $P(y|x)$ is one of the k best $\{P(z|x)\}_{z \in E}$.

N_{kNC} associates to each term x its k Nearest Children: $y \in N_{kNC}(x)$ iff $P(y|x)$ is one of the k best $\{P(y|z)\}_{z \in E}$.

All criteria depend of the parameter k that controls the number of selected relations. In particular, we adjust the thresholds σ_k in such a way that N_{kSand} selects as many relations as the two other criteria for a same value of k (i.e. $k.|E|$ relations). So, each type of criterion provides several effective criteria depending of the parameter k. We considered three different values for k ($\{1 \dots 3\}$) leading to nine neighborhood, plus the partial knowledge (that can also serve as a neighborhood).

The english subpart of wikipedia.org has been used as corpus for frequency counts extraction. For each pair of terms (x, y), we retrieve the number of wikipedia pages where both terms occur ($hits(x, y)$) in the corresponding sub-domain of wikipedia. Sub-domains are artificially identified by introducing the root term of the taxonomy into the wikipedia query. For example, $hits(memory, politics)$ is retrieved with the following query ["memory" AND "politics" AND "science"] as *memory* and *politics* are two terms contained into the *wn_science* list of terms to structure.

The partial knowledge as been obtained by first extracting a list of candidate subsumption pairs observing linguistic patterns into a corpus 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. The 10 linguistic patterns used, from (Kozareva and Hovy, 2010; Snow et al., 2004), are the following : {*X are Y that - X is a Y that - X is an Y that - Y such as X - Y including X - Y like X - X and other Y - X or other Y - such Y as X - Y, specially X*}

For any pairs of terms (x, y) from the list E, each pattern is tested on en.wikipedia.org and each time a pattern is observed between x and y, an edge $x \rightarrowtriangle y$ (x subsumes y) is added to S (after manual validation). A quantitative summary of the partial knowledge construction for each considered domain is reported in Table 1.

The LPS process has been applied on the four first lists of terms : *wn_science*, *science*, *wn_equipment* and *equipment* of limited sizes (less than 1,000).

Table 1: Quantitative summary of semi-automatic acquisition of the partial knowledges S.

List of terms	Nb. terms	Nb. candidate pairs	Nb. selected pairs	Nb. added pairs	Size of S
WN_Science	370	341	272	0	272
Science	462	347	230	0	230
WN_Equipment	475	296	162	133	295
Equipment	612	83	38	169	207
WN_Food	1485	2130	200	52	252
Food	1555	1630	144	83	227
WN_Chemical	1350	1908	227	0	227
Chemical	17,584		*not processed*		

GA was parameterized so that it iterates crossings and mutations on a population of 200 P-Spaces and finally selected the one maximizing the score (2). For example, on the *science* list, the P-Space acquired induces a LT reaching a score of 0.948, with a matching of 0.98 with S (the F term) and a structuring term (I) of 0.97. The underlying parameters $\mathbf{w}$ can be interpreted as a logical propagation rule combining neighborhoods from the given family $\mathcal{N}$; the obtained rule is

$$\begin{array}{c} \delta_S(x) \vee (\delta_{N_{1NS}}(x) \wedge \delta_{N_{2NF}}(x)) \\ \vee(\delta_{N_{3NS}}(x) \wedge \delta_{N_{1NF}}(x) \wedge \delta_{N_{1Sand}}(x)) \\ \vee(\delta_{N_{3NS}}(x) \wedge \delta_{N_{2NF}}(x)) \end{array} \quad (3)$$

formalizing the extension of a subset A to an element x when either :

- the neighborhood $N_S(x)$ intersects A (*i.e.* x is dominated by a term $y \in A$ according to the partial knowledge S) or,
- both neighborhoods $N_{1NS}(x)$ and $N_{2NF}(x)$ intersect A or,
- neighborhoods $N_{3NS}(x)$ and $N_{1NF}(x)$ and $N_{1Sand}(x)$ intersect A or,
- neighborhoods $N_{3NS}(x)$ and $N_{2NF}(x)$ intersect A.

The final external evaluation (comparison against the gold standard) revealed that the LT induced by the previous P-Space obtains the best score (0.523) using the cumulative Fowlkes&Mallows measure (Fowlkes and Mallows, 1983).

Due to time limitations, learning P-Spaces with LPS was not possible for the domains *wn_food*, *food* and *wn_chemical*. For these domains, we computed the neighborhoods and rather than *learning* a combination rule fitting to the dataset, we tested the four combination rules acquired from the four previous domains, computed their ability to induce a good LT (by computing the score (2)) and finally we kept the best one. Table 2 finally summarizes the results obtained by the team QASSIT and its relative positioning into the task.

Table 2: Results obtained on the 8 domains in terms of fitness and gold standard evaluation ; symbol * indicates domains for which a learning stage has been performed.

Domains	internal score (2)	F&M measure	Best F&M	rank
WN_Science*	0.97	0.29	0.54	3/6
Science*	0.95	0.52	0.52	1/6
WN_Equip.*	0.63	0.36	0.69	2/6
Equipment*	0.34	0.49	0.49	2/6
WN_Food	0.56	0.32	0.59	3/6
Food	0.73	0.34	0.45	2/6
WN_Chemical	0.54	0.39	0.39	1/6
Chemical		*not processed*		

5 Conclusion

The automatic evaluation against the gold standards has been then completed by a manual analysis that revealed lower comparative results for the seven taxonomies acquired with the LPS approach. But the main lesson to learn from this second type of evaluation is the high discrepancy between the taxonomies obtained with a learning stage (at least 0.20 of F-measure each time) and the the ones obtained by reusing combination rules (less than 0.10 each time). These results encourages future researches toward the scalability of the LPS learning process and various improvements in terms of statistical neighborhoods enhancement and linguistic patterns selection.

References

Z.T. Belmandt. 2011. *Basics of pretopology*. Hermann.

Chris Biemann. 2005. Ontology learning from text: a survey of methods. *LDV-Forum*, 20(2):75–93.

Georgeta Bordea, Paul Buitelaar, Stefano Faralli, and Roberto Navigli. 2015. Semeval-2015 task 17: Taxonomy extraction evaluation. In *Proceedings of the 9th International Workshop on Semantic Evaluation*. Association for Computational Linguistics.

Marcel Brissaud. 1975. Les espaces prétopologiques. *Compte-rendu de l'Académie des Sciences*, 280(A).

Paul Buitelaar, Philipp Cimiano, and Bernardo Magnini, editors. 2005. *Ontology Learning from Text: Methods, Evaluation and Applications*, volume 123 of *Frontiers in Artificial Intelligence and Applications*. IOS Press.

Philipp Cimiano, Alexander Mädche, Stephen Staab, and Johanna Völker. 2009. Ontology learning. In *Handbook of Ontologies*, pages 245–267. Springer Verlag.

Guillaume Cleuziou, Davide Buscaldi, Vincent Levorato, and Gaël Dias. 2011. A pretopological framework for the automatic construction of lexical-semantic structures from texts. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM)*, pages 2453–2456.

E. B. Fowlkes and C. L. Mallows. 1983. A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association*, 78(383):553–569.

Zornitsa Kozareva and Eduard Hovy. 2010. A semi-supervised method to learn and construct taxonomies using the web. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1110–1118.

Christine Largeron and Stéphane Bonnevay. 2002. A pretopological approach for structural analysis. *Information Sciences*, 144:169–185, July.

Mark Sanderson and Bruce Croft. 1999. Deriving concept hierarchies from text. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 206–213.

Rion Snow, Daniel Jurafsky, and Andrew Y Ng. 2004. Learning syntactic patterns for automatic hypernym discovery. *Advances in Neural Information Processing Systems 17*.

Paola Velardi, Stefano Faralli, and Roberto Navigli. 2013. Ontolearn reloaded: A graph-based algorithm for taxonomy induction. *Computational Linguistics*, 39(3):665–707.