Learning Pretopological Spaces for Lexical Taxonomy Acquisition

Guillaume Cleuziou[1] and Gaël Dias[2]

[1] Université d'Orléans, INSA Centre Val de Loire, LIFO EA 4022, France
[2] Université de Caen Basse-Normandie, GREYC UMR 6072, France

Abstract. In this paper, we propose a new methodology for semi-supervised acquisition of lexical taxonomies. Our approach is based on the theory of pretopology that offers a powerful formalism to model semantic relations and transforms a list of terms into a structured term space by combining different discriminant criteria. In order to learn a parameterized pretopological space, we define the Learning Pretopological Spaces strategy based on genetic algorithms. In particular, rare but accurate pieces of knowledge are used to parameterize the different criteria defining the pretopological term space. Then, a structuring algorithm is used to transform the pretopological space into a lexical taxonomy. Results over three standard datasets evidence improved performances against state-of-the-art associative and pattern-based approaches.

1 Introduction and Related Work

By coding the semantic relations between terms, lexical taxonomies (LTs) such as WordNet [9] have enriched the reasoning capabilities of applications in information retrieval and natural language processing. However, the globalized development of semantic resources is largely limited by the efforts required for their construction [6]. As a consequence, many research studies have been appearing to automatically learn LTs [4]. Instead of manually creating LTs, learning them from texts has undeniable advantages. First, they may fit the semantic component neatly and directly within a given domain. Second, the cost per entry is greatly reduced, giving rise to much larger resources.

The two main stages for the automatic construction of LTs are term extraction (TE) and term structuring (TS). A substantial amount of works exist on TE [10, 7], but the present study exclusively focuses on the TS stage. Within this context, similarity-based, pattern-based, set-theoretical and associative approaches have traditionally been proposed.

Similarity-based or clustering-based approaches [12, 11] hierarchically cluster terms based on similarities of their meanings usually represented by a vector of quantifiable features. They have the main advantage that they are able to discover relations which do not explicitly appear in text. They also avoid the problem of inconsistent chains by addressing the structure of a taxonomy globally from the outset. However, it is generally believed that these methods can not generate relations as accurate as pattern-based approaches. *Pattern-based*

strategies [6, 17] define lexical- syntactic patterns for semantic relations and use these patterns to discover instances of relations. They are known for their high accuracy in recognizing instances of relations if the patterns are carefully chosen, either manually or via automatic boostrapping. However, this approach suffers from sparse coverage of patterns in specific corpora, especially technical domain ones. Moreover, it may evidence inconsistent concept chains as instances are extracted in pairs and gathered to form taxonomy hierarchies. *Set-theoretic approaches* [3] use formal concept analysis that naturally structures terms with intensional inclusion relations within a concept lattice. Such term organization differs from usual lexical taxonomies that provide semantic relations between terms rather than inclusion relations between formal concepts. This strategy usually highlights low performance as contextual vector seldom overlap in large open uncontrolled domains. Finally, *associative frameworks* [14] use asymmetric similarities between terms to model the subsumption relation. For that purpose, distributions of terms over document collections are used to discover general/specific noun relationships. The main drawback of this approach is that the subsumption model implicitly hypotheses that general terms are always more frequent than their specific terms, which is not always satisfied in practice.

Note that these methodologies rely on one exclusive criterion to model the subsumption (is-a) relation and build the respective taxonomy. In order to take advantage of multiple criteria, two important works have been proposed [16, 18]. Both methodologies first learn an ontology metric, which models the is-a relation based on vectors of discriminant criteria (e.g. contextual, cooccurrence, syntactic dependency or patterns). This step is obtained by supervised learning over existing taxonomies. The logistic regression is used by [16] and [18] applies the ridge regression. Then, the ontology metric guides the incremental taxonomy acquisition process modeled as an optimization task: 1-objective for [16] and 2-objectives for [18]. The main advantage of these approaches is to model the is-a relation between terms based on multiple criteria, thus greatly avoiding data sparseness and low coverage. However, both proposals depend on a supervised learning stage that relies on large known ontologies such as WordNet or Open Directory Project. However, in real-world situations, this knowledge is not accessible and only partial (usually small) knowledge of the domain can be accessed. Moreover, note that these large resources are mainly available for the English general language. As such, language/domain/genre adaptability is not ensured.

In this paper, we propose a new semi-supervised multi-criteria strategy for taxonomy induction. The overall idea is (1) to learn a propagation metric[3] based on a set of relevant associative and pattern-based features constrained by small (yet accessible) pieces of knowledge of the domain and (2) to induce the taxonomy based on a pretopological framework which transforms the pretopological term space into a directed acyclic graph, the output taxonomy. To achieve these objectives, we consider pretopology on the multi-criteria analysis point of view, where criteria are statistical indices (associative approach) and linguistic patterns (pattern-based approach) retrieved from a corpus. In particular, we define

[3] As opposed to the ontology metric.

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) approximate the expected LT structure and (2) verify a given number of linguistic patterns constraints. In order to learn the parameters, we define a new *Learning Pretopological Spaces* (LPS) strategy based on genetic algorithms, which leads to induce a LT from an "optimized" P-space. The main advantages of the LT acquisition methodology presented in this paper, when compared to state-of-the-art methodologies are enunciated as follows:

(1) We learn a propagation metric, which directly models the is-a relation into the taxonomy induction process in contrast to [18] and [16] who propose a two-steps process,
(2) Linguistic patterns, which embody small (yet accessible) pieces of knowlegde of the domain constrain the semi-supervised learning process but are also used as relevant criteria,
(3) We deal with both general and specialized domains where linguistic patterns fail to retrieve any relation,
(4) Our framework is quasi-independent regarding to language as only few and simple linguistic patterns and raw texts are required.

In the remainder of this paper, we first define the required notions of our pretopological framework and its usage for multi-criteria analysis (Section 2). Then, in Section 3, we define the concept of parameterized pretopological space (P-space) and propose the learning pretopological spaces (LPS) strategy based on genetic algorithms in the context of taxonomy induction. In Section 4, we evaluate our framework on the LT reconstruction task, considering both general (i.e. WordNet) and specialized domains (i.e. UMLS). Finally, in Section 5, concluding statements are enunciated.

2 Pretopological Framework

Pretopology [1] is a theory that generalizes both topology and graph theories and is commonly used to model complex propagation phenomena thanks to a pseudo-closure function. Let's consider a non-empty set E and its powerset $\mathcal{P}(E)$. A pretopological space[4] is noted (E, a), where $a(.)$ is a pseudo-closure function described in Definition 1.

Definition 1 (Pseudo-closure). *A pseudo-closure is a function* $a(.) : \mathcal{P}(E) \rightarrow \mathcal{P}(E)$*, which respects the following three conditions:*

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

[4] Note that in this paper, we always consider V-type spaces, as they present good structuring properties.

So, the pseudo-closure function 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. Two other concepts are required to introduce our model: *elementary closed subset* and *maximal elementary closed subset* formalized in Definitions 2 and 3 respectively.

Definition 2 (Elementary Closed Subset). *An elementary closed subset $F_{\{x\}}$, is the closure of a singleton $\{x\}$ with $x \in E$.*

Definition 3 (Maximal Elementary Closed Subset). *A maximal elementary closed subset is an elementary closed subset, maximal in terms of inclusion.*

These definitions give us two key concepts on a structuring point of view: (1) an elementary closed subset $F_{\{x\}}$ refers to the subset of items reachable from x and (2) when $F_{\{x\}}$ is maximal, it means that x is only reachable from items y with an identical elementary closed subset ($F_{\{x\}} = F_{\{y\}}$), thus capturing a kind of equivalence class.

2.1 Pretopology and Multi-criteria Analysis

Pretopology can be used in the context of multi-criteria analysis since it allows complex but efficient aggregation of several criteria at the pseudo-closure function level. So, considering (1) a set of K criteria providing different views on the manner a discrete set E is structured and (2) each criterion defining one neighborhood relation on E and $N_k(x)$ the k^{th} neighborhood of x, the family of neigborhoods $\mathcal{N} = \{N_1, \ldots, N_K\}$ suggests a multi-criteria environment. Note that to be consistent with the formal definition of neighborhoods [1], we constrain any $N_k(x)$ to contain x itself:

$$\forall k = 1, \ldots, K, \quad \forall x \in E, \quad x \in N_k(x). \tag{1}$$

A usual pseudo-closure definition for neighborhood aggregation, which satisfies the V-type space conditions is given by

$$\forall A \in \mathcal{P}(E), a(A) = \{x \in E | \ \forall \ N_k \in \mathcal{N}, N_k(x) \cap A \neq \emptyset\}. \tag{2}$$

Such a pseudo-closure expands a subset A to an item x if and only if all neighborhoods (criteria) of x intersect A. It is important to note that when A is not reduced to a singleton, the agreement can be reached by intersections that concern different items of A. Thus, a complex propagation process is defined at the subset level rather than at the element level and there is no way to reproduce such a process on a single neighborhood structure that would result from the *a priori* aggregation of the different criteria[5].

[5] Proof of this statement is out of the scope of this paper.

2.2 Pretopology and LT Acquisition

It is well-established that known LTs such as WordNet or Cyc share some specific common structure. As a consequence, a learned LT should ideally satisfy the following two structural requirements:

(1) a non-binary tree-like structure: each node must be characterized by two disjoint sets of predecessors and successors with no cycles,
(2) aggregating nodes: each node must contain one or several terms from the vocabulary E.

Such a structure can be obtained based on a pretopological term space with the structuring algorithm proposed by [8]. In our specific case, we propose a top-down version of this algorithm. So, instead of considering minimal closed subsets, we consider maximal ones. The basic idea of the algorithm for LT induction is defined in Figure 1 and illustrated in Figure 2[6].

1. Determine elementary closed subsets associated to each element x of E giving rise to the family of closures $\mathcal{F}e(E, a)$.
2. Find the family of maximal closed subsets $\mathcal{F}M(E, a)$. This means enumerating all the maximal elementary closed subsets by inclusion in $\mathcal{F}e(E, a)$. Any element $F \in \mathcal{F}M(E, a)$ is then a core.
3. Within each core, recursively determine the largest elementary closed subsets of E in terms of inclusion, until no other can longer be found. The recursive process allows to generate, from each core, a set of homogeneous parts by successive reductions and outputs the final LT.

Fig. 1. LT induction algorithm.

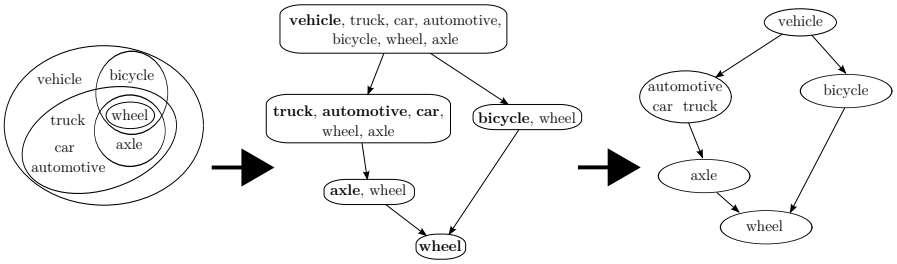


Fig. 2. Top-down structuring inducing a DAG from a pretolopological term space.

2.3 Current Limitations

Despite its interesting properties for multi-criteria analysis as evidenced in [5] for LT induction, in its current form, the pretopological LT process evidences two main limitations that make it under-efficient:

[6] More details can be found in [5].

(1) it is sensitive to unreliable criteria,
(2) it only allows a limited number of criteria to combine.

Both issues are due to the definition of the pseudo-closure operator itself that requires that all criteria must satisfy the intersection property in order to start the propagation process from elementary sets.

3 Learning Pretopological Spaces

To overcome previous limitations, we propose in this paper a new learning pretoplogical spaces (LPS) framework based on a more flexible pseudo-closure definition. It is illustrated in Figure 3.

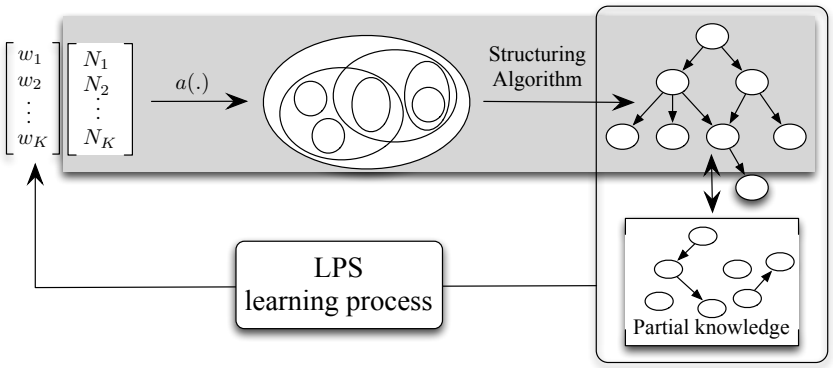


Fig. 3. The LPS process uses partial knowledge on the expected structure in order to improve the parameterization of the pseudo-closure operator.

This new framework relies on the one-pass process from [5] that first computes a unique pretopological space from a family of criteria using the pseudo-closure defined in (2) and then applies the top-down variant of the structuring algorithm from [8]. But, rather than providing the resulting structure as output, the LPS framework consists in comparing the built structure to some partial knowledge and modifying the pseudo-closure operator in order to improve the final structuring. This is achieved by an iterative semi-supervised learning process. Such a framework requires to introduce new concepts into the pretopology theory, especially the concept of *parameterized pretopological space* (P-Space).

3.1 Parameterized Pretopological Space

To relax the constraint that requires the agreement on all criteria to allow the propagation process, we propose to introduce a parameter p that indicates a requirement on the minimum number of neighborhoods that must intersect a subset A in order to expand it. Its formalization is given in Equation 3 with

$p \in \{1, \ldots, K\}$ and $\mathbb{1}_{N_k(x) \cap A \neq \emptyset} = 1$ if the neighborhood $N_k(x)$ intersects A and 0 otherwise.

$$\forall A \in \mathcal{P}(E), a(A) = \{x \in E \mid \sum_{N_k \in \mathcal{N}} \mathbb{1}_{N_k(x) \cap A \neq \emptyset} \geq p\} \tag{3}$$

To express the combination model as a learning model, we define the notion of *parameterized pretopological space* (P-Space) that introduces supplementary parameters to manage the reliability of the criteria in Definition 4.

Definition 4 (P-Space). *A P-space* $(E, a, \mathbf{w})$ *is a* V*-type pretopological space with a parameterized pseudo-closure* $a(.)$ *defined by*

$$\forall A \in \mathcal{P}(E), a(A) = \{x \in E \mid \sum_{N_k \in \mathcal{N}} w_k . \mathbb{1}_{N_k(x) \cap A \neq \emptyset} \geq w_0\} \tag{4}$$

such that (1) $w_0 > 0$, (2) $\sum_{k=1}^{K} w_k \geq w_0$ *and* (3) $\forall k,\ w_k \geq 0$.

Note that conditions (1), (2) and (3) over the set of parameters $\mathbf{w}$ are defined to respectively ensure the three conditions *i*), *ii*) and *iii*) expressed in Definition 1 over the V-type spaces. In particular, each parameter w_k in Equation (4) quantifies the reliability on the k^{th} criteria and w_0 represents a global required confidence to expand the subset. 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 evidences two strong advantages: (1) it overcomes the limitations about reliability and multiplication of the criteria and (2) it extends significantly the possibilities of combination, passing from a single conjunctive decision rule to a set of logical decision rule (without negation). But the noticeable improvement on the model makes a new challenging question to appear: **How to parameterize a P-Space?**

3.2 Semi-supervised Learning of P-Spaces

We propose a semi-supervised strategy to learn the parameters of a P-Space. So, if S is a given source providing a true **partial** structuring on E[7] and considering that a V-type pretopological space induces a unique DAG, the Learning P-Space (LPS) process aims to find a P-Space inducing a DAG that satisfies:

(1) the constraints implied by the partial knowledge S and
(2) a taxonomy-like structuring.

The following $Score(.,.)$ quantifies such a satisfaction:

$$Score(\mathbf{w}, S) = F_{Measure}(\mathbf{w}, S) \times I_{taxonomy}(\mathbf{w}). \tag{5}$$

[7] Note that in the context of LT acquisition S is usually a small number of "evident" subsumption relations between terms.

The $F_{Measure}$ is the usual external validation index [13] that, in our context, combines *precision* and *recall* calculated over the pairs of elements linked in the partial knowledge S only. More precisely, given a DAG $D_{\mathbf{w}}$ induced by the P-Space with parameters $\mathbf{w}$ and the partial knowledge S also formalized as a (more sparse) DAG, we first operate a closure operation on both graphs (resulting in $\bar{D}_{\mathbf{w}}$ and $\bar{S}$) in order to make any implicit (indirect) edge to emerge before computing *precision*, *recall* and $F_{Measure}$. Metrics are defined in Equations 6 where $\bar{S}^t$ denotes the graph opposite to $\bar{S}$, which must be considered in order to count the false positive relations.

$$\begin{aligned} precision &= \frac{|\{(x,y) \in \bar{D}_{\mathbf{w}} \cap \bar{S}\}|}{|\{(x,y) \in \bar{D}_{\mathbf{w}} \cap (\bar{S} \cup \bar{S}^t)\}|} \\ recall &= \frac{|\{(x,y) \in \bar{D}_{\mathbf{w}} \cap \bar{S}\}|}{|\{(x,y) \in \bar{S}\}|} \\ F_{Measure} &= \frac{2.precision.recall}{precision+recall} \end{aligned} \tag{6}$$

The $I_{taxonomy}$ term is used to control the structural properties of the induced DAG $D_{\mathbf{w}}$ (independently to S). Although, the structure of a taxonomy is not formally defined, one can observe that a taxonomy usually looks like more to a tree (with one parent per node - except for the root) than to a lattice structure (for example). In order to favor tree-like structures, we compute on $D_{\mathbf{w}}$ its average ascendant degree (i.e. average number of parents per node) $Ad(D_{\mathbf{w}})$, and we use it to penalize a DAG moving away from a tree-like structure. This constraint is formalized in Equation 7.

$$I_{taxonomy}(\mathbf{w}) = e^{-(Ad(D_{\mathbf{w}})-1)^2} \in [0,1] \tag{7}$$

The final satisfaction measure $Score(\mathbf{w}, S)$ reaches a maximum value of 1 for a DAG that (1) fits exactly to the knowledge source S and (2) structures the elements with an average ascendant degree of 1 (taxonomy).

This measure is used to guide the exploration of the space of solutions through a learning strategy based on a Genetic Algorithm (GA). GAs are stochastic exploration methods inspired from the natural selection principle [15]. Given a $fitness(.)$ function over the solution space, they simulate a natural evolution process by iteratively (1) generating populations of solutions (with mutation and crossover operators) and then (2) selecting the ones with highest fitness. The LPS approach uses such a stochastic exploration process based on the following fitness function:

$$fitness(\mathbf{w}) = \begin{cases} Score(\mathbf{w}, S) & \text{if } \mathbf{w} \text{ satisfies Def. (4)} \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

Figure 4 gives an overview of the general LPS process. Each iteration of the algorithm leads to an ensemble of P-Spaces, evaluated and selected as regard to their ability to induce a taxonomy-like structure satisfying partial knowledge requirements. Ultimately, only the best P-Spaces are returned. Let us notice

that, in addition to the expected taxonomies the final P-Spaces allow to induce, the returned P-Spaces are themselves of high interest since they hold a learned propagation process that could be reused in an incremental context[8].

1. Given:
   a set of elements E,
   a family of criteria $\mathcal{N} = \{N_1, \ldots, N_K\}$,
   a partial knowledge S,
   a maximum number of iterations t_{max}.
2. Build an initial population $\mathbf{W}^0 = \{\mathbf{w} \in [0,1]^{K+1}\}$.
3. $t \leftarrow 0$
4. For each solution $\mathbf{w} \in \mathbf{W}^t$
   - Build the induced DAG $D_{\mathbf{w}}$,
   - Evaluate its $fitness(\mathbf{w})$
5. Select the best P-Spaces
6. if $(t < t_{max})$ then
   $t \leftarrow t + 1$,
   - Generate a new population $\mathbf{W}^t$ by mutation and crossover,
   - GoTo step 4
7. else return the selection.

Fig. 4. LPS general algorithm.

4 Experiments on LT Acquisition

The objective of the present proposal is to combine associative and pattern-based methods for LT acquisition by applying our multi-criteria LPS algorithm. Two situations have been considered in the following experiments:

When the linguistic patterns succeed in retrieving (maybe few) accurate relations. This is usually the case for generic domains. In this case, the set of term relations automatically extracted from a given set of patterns plays the role of the partial knowledge S. LT acquisition is thus performed in an *auto-supervised* context since no expert intervention is needed,

When the linguistic patterns fail to provide any reliable piece of knowledge that could guide the structuring process. This situation frequently occurs for specialized domains and makes most of the existing pattern-based approaches [17, 6] totally inoperative. An expert is so required to give at least a couple of term relations (S) as in a usual *semi-supervised* learning context.

4.1 Experimental Setups

For each LT construction experiment, the list of terms to structure E comes from a reference R (in english) and the acquired taxonomies are compared to

[8] This is out of the scope of this paper.

this reference using the $F_{Measure}$ as defined in (6) but with R that stands in for S. Indeed, S is only a set of term relations that helps the learning process and R is the complete gold standard reference taxonomy, to which the induced LT must be compared.

The linguistic patterns used are limited to the following list of four simple and usual ones [6]: "X such as Y", "X including Y", "X like Y" and "Y are X that". For any pair of terms (x, y) from the list, each pattern is tested on en.wikipedia.org and each time a pattern is observed between x and y, an edge $x \leftarrow y$ (x subsumes y) is added to S.

The english subpart of wikipedia.org (i.e. en.wikipedia.org) is also 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, $hits(cars, trucks)$ is retrieved with the following query ["cars" AND "trucks" AND "vehicle"], *vehicle* being the root of the taxonomy to reconstruct.

From the frequency counts, three kinds of associative criteria are built in order to serve as basis neighborhoods for the P-Spaces:

N_{kSand} corresponds to the subsumption relation modeled by [14] : $y \in N_{kSand}(x)$ 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 threshold 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. In the following experiments, each criterion will be used with three different values ($k \in \{1 \dots 3\}$) leading to families containing respectively three, six and nine effective criteria.

Let us notice that, unlike the two first criteria, N_{kNC} has a strong weakness as it tends towards a non-taxonomic structure. In particular, it will be used to illustrate the behaviour of our approach in the context of an existing unreliable criterion when compared to previous studies.

To conclude on the preliminaries, let us mention the following operational details. The LPS algorithm has been implemented using the R package "GA" [15] with default configurations for crossover and mutation operators. We fixed the size of the population in the range $\{25 \dots 1000\}$ depending of the number of terms to structure and a maximum number of iterations to 25. As GAs are stochastic methods, we select in the coming results the best learned P-Space (in terms of $fitness(\mathbf{w})$) over a set of 5 runs.

4.2 LT Acquisition with Auto-supervision

The LT construction task is experimented on three domains extracted from WordNet, *Vehicles*, *Plants* and *Food*, with respectively 108, 554 and 1485 terms. The first two datasets are usually used as gold standards on LT induction [17, 6] and the *Food* dataset is provided by the recent SEMEVAL 2015 contest [2].

Table 1 reports in the three top parts, the scores obtained (and the corresponding best parameters k) using purely associative approaches (without LPS), with or without aggregation of two or three statistical criteria.

Table 1. Quantitative evaluation of reconstructed lexical taxonomies on the domains *Vehicles*, *Plants* and *Food*.

	Vehicles				**Plants**				**Food**			
Criteria	Prec.	Rec.	FM.	k	Prec.	Rec.	FM.	k	Prec.	Rec.	FM.	k
[14] $N_{kSand.}$	0.75	0.35	0.48	2	0.55	0.32	0.40	2	0.28	0.20	0.23	4
N_{kNP}	0.44	0.45	0.44	2	0.57	0.29	0.38	1	0.50	0.23	0.31	1
N_{kNC}	0.06	0.26	0.10	2	0.04	0.02	0.03	1	0.01	0.03	0.01	10
2-Criteria Combinations (without LPS)												
$N_{kSand.} \wedge N_{kNP}$	0.77	0.34	0.47	2	0.70	0.31	0.43	2	**0.72**	0.19	0.30	8
$N_{kSand.} \vee N_{kNP}$	0.42	0.46	0.44	2	0.57	0.29	0.38	1	0.43	0.23	0.30	1
[5] $N_{kSand.} \diamond N_{kNP}$	0.77	0.34	0.47	2	0.70	0.31	0.43	2	**0.72**	0.19	0.30	8
3-Criteria Combinations (without LPS)												
$\wedge$ Combination	0.31	0.41	0.36	14	0.15	0.07	0.10	15	0.26	0.03	0.05	20
$\vee$ Combination	0.33	0.37	0.35	1	0.28	0.30	0.29	1	0.24	0.23	0.24	1
[5] $\diamond$ Comb.	0.45	0.36	0.40	6	0.16	0.34	0.22	14	0.26	0.03	0.05	20
LPS based on associative criteria only												
3 criteria	0.77	0.34	0.47	2	0.95	0.25	0.40	1	0.50	0.23	0.31	1
6 criteria	0.77	0.34	0.47	1..2	0.58	0.32	0.41	1..2	0.49	0.23	0.31	1..2
9 criteria	0.76	0.36	0.49	1..3	0.64	0.32	0.43	1..3	0.44	0.25	**0.32**	1..3
LPS based on associative criteria + the linguistic criteria S												
4 criteria	**0.84**	0.37	0.52	1	**0.96**	0.31	0.47	1	0.50	0.23	0.31	1
7 criteria	0.77	0.42	0.55	1..2	0.58	**0.40**	0.47	1..2	0.49	0.23	0.31	1..2
10 criteria	0.74	**0.48**	**0.58**	1..3	0.62	**0.40**	**0.49**	1..3	0.43	**0.27**	**0.32**	1..3

The N_{2Sand} criterion corresponding to the methodology of [14] clearly outperforms all other single criteria in terms of *precision* while N_{kNP} evidences increased *recall* compared to all other criteria for the *Vehicles* domain. Note that this situation is reversed for the *Plants* and *Food* domains, which indicates that the subsumption relation can be described differently depending on the studied domain. This is an important issue when compared to [18, 16], who suppose that the is-a relation can universally be learned from WordNet. Expectedly, N_{kNC} shows poor performance due to its non-taxonomic nature.

When the best two criteria are joined into a non-guided (without LPS) aggregation strategy, results show similar performance (with slight improvements for

the *Food* domain), especially for the conjunctive ($\wedge$) and pretopological ($\diamond$) [5] aggregations. However, the disjunctive ($\vee$) aggregation operator leads to worst results as the subsumption definition is not enough constrained. Note that the disjunctive and conjunctive aggregations consist in generating one new criterion from initial ones by considering respectively their union and their intersection i.e. the neighborhood family is thus reduced to a single neighborhood.

Finally, when the three criteria (including a non-performant one, i.e. N_{kNC}) are gathered in the multi-criteria framework without LPS, all aggregations fail and performance drastically drops. The difference is even higher for the *Plants* and *Food* domains, which are known to be well-structured. These experiments clearly show the incapacity of this previous model to handle unreliable criteria. The next experiments aim to evidence the superiority of the LPS strategy.

As pattern-based methods succeed in extracting reliable relations from the three domains, we performed our LPS approach in an *auto-supervised* way. In particular, for *Vehicles*, 93 relations were found corresponding to a recall of 17.6% and with a rate of 78.5% in precision as regards the reference. For *Plants*, 332 relations were foundwith a recall of 10.2% and a precision of 61.9%, and for *Food* only 244 relations were extracted, resulting in a low recall (3%) and with a small precision (36%). So, the fourth sub-table of Table 1 shows how the LPS methodology allows to learn new P-Spaces by selecting and combining more efficiently three, six or nine associative criteria and reaches slightly improved results to the ones presented by [14], which are the best up to now in terms of associative frameworks. Interestingly, higher precision is obtained to the detriment of recall, especially for the *Plants* domain.

Let us mention that if S can be used as a partial supervisor in the LPS method, it can also be used, without reserve, as a new criterion in the family of neighborhoods $\mathcal{N}$. In Table 1, the bottom part reports the scores obtained by introducing the pattern-based feature S as a supplementary criterion to consider in the construction of the combination rule. This experiment evidences the efficiency of LPS with such a mixture of pattern-based and associative criteria that makes reachable new P-Spaces inducing strongly improved taxonomies (e.g. up to 9% $F_{Measure}$ for *Vehicles* and 6% for *Plants*). To illustrate the P-Space learned by the four criteria (N_{1Sand}, N_{1NP}, N_{1NC} and S) on the *Vehicles* domain, we derive the DNF rule from the final parameters $\mathbf{w}$ and we obtain the following (simplified) expansion strategy:

$$\delta_S \vee (\delta_{N_{1Sand}} \wedge \delta_{N_{1NP}}) \vee (\delta_{N_{1Sand}} \wedge \delta_{N_{1NC}}), \tag{9}$$

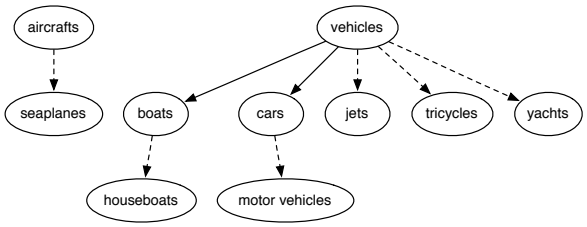
which means that a subset A can be expanded with an element x if one of the following properties are satisfied:

(1) x is in relation with at least one element from A in the partial knowledge S,
(2) both neighborhoods $N_{1Sand}(x)$ and $N_{1NP}(x)$ intersect A,
(3) both neighborhoods $N_{1Sand}(x)$ and $N_{1NC}(x)$ intersect A.

It is also interesting to visualize well-chosen results in order to understand the different behaviours between approaches. In Figure 5, we present LT subparts

respectively resulting from the best configurations of the methodology proposed by [14] and the 10 criteria LPS learning approach. In particular, continuous edges are present in S and dashed ones are learned relations for *Vehicles*.

(a) Obtained LT with $N_{kSand.}$ best configuration ($F_{Measure} = 0.48$ and $k = 2$)



(b) Obtained LT with 10 criteria best configuration ($F_{Measure} = 0.58$)

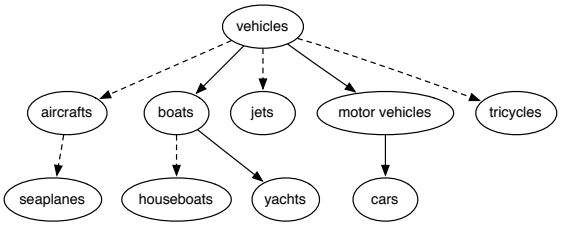


Fig. 5. Examples of induced LT with $N_{kSand.}$ and LPS for *Vehicles*.

Finally, in order to propose a meaningful evaluation, we summarize in Table 2 the comparative results with most reproducible state-of-the-art approaches[9]: [14] for the associative paradigm, [5] for the initial pretopological framework and [6] for the pattern-based approach.

Table 2. Comparison of LT acquisition methodologies on *Vehicles*, *Plants* and *Food*.

	Vehicles			**Plants**			**Food**		
Method/Approach	Prec.	Rec.	FM.	Prec.	Rec.	FM.	Prec.	Rec.	FM.
[14] associative	0.75	0.35	0.48	0.55	0.32	0.40	0.28	0.20	0.23
[5] pretopological	0.45	0.36	0.40	0.16	0.34	0.22	0.26	0.03	0.05
[6] pattern-based	**0.79**	0.18	0.29	**0.62**	0.10	0.18	0.36	0.03	0.05
LPS framework	0.74	**0.48**	**0.58**	**0.62**	**0.40**	**0.49**	**0.43**	**0.26**	**0.32**

Let us mention that in order to compare these methods within similar conditions, the final structuring approach from [6] has been performed on the partial knowledge S extracted from en.wikipedia.org. Results obtained by [6] on the same datasets are higher but they are obtained using the (non-free) Yahoo!Boss

[9] Note that [18] evidence a $F_{Measure}$ of 0.82 on a non-available WordNet dataset, where training and testing are performed over the same data, thus invalidating any conclusion. Note also that [16] only present experiments for populating an existing ontology and direct comparison cannot be evidenced.

search engine. Moreover, note that results of [6] are very similar to the ones of [17] who use different evaluation metrics in their paper. As a consequence, only results of [6] are reported here.

Results in Table 2 clearly reveal the benefice of mixing both statistical information and linguistic patterns within a unified learning process. The pattern-based approach obtains, as expected, better precision but significantly fails to retrieve most of the relations, whereas the LPS framework outperforms any other method on *recall* without drastic loss in *precision* so evidencing high $F_{Measure}$.

4.3 LT Acquisition with Semi-supervision

To deal with the acquisition of specialized LTs, we take as reference the concatenation of four sub-domains from the *Unified Medical Language System*[10] (UMLS) produced by the U.S. National Institutes of Health. The selected sub-domains are cardiovascular system, digestive system, nervous system and respiratory system. Their concatenation results in a list of 128 specialized terms like *upper gastrointestinal tract* or *blood-retinal barrier*.

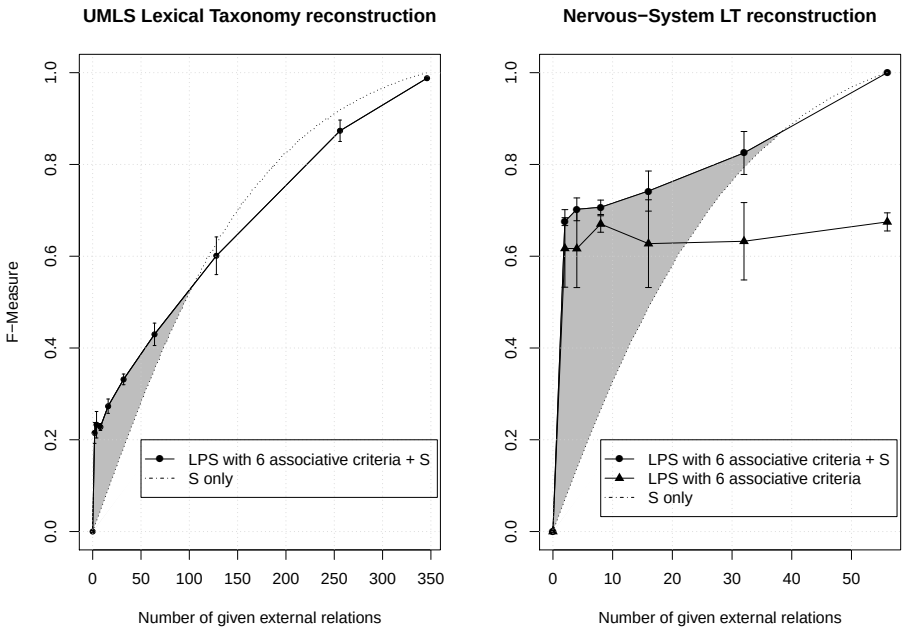


Fig. 6. Quantitative evaluation of reconstructed lexical taxonomies on the medical domain (UMLS).

Over this term list, none of the four considered lexical patterns retrieved any relation on en.wikipedia.org, whereas terms actually occur with an average of 2225 counts per term. A pattern-based extraction test performed on the more

[10] http://www.nlm.nih.gov/research/umls/

specialized corpus PubMed[11] has led to the same statement. Thus, we used our LPS methodology in a semi-supervised context. In particular, we simulated the input of expert knowledge S by randomly extracting relations from the reference.

Figure 6 (left) shows the performances in reconstructing the UMLS subpart with LPS as regards to the number of given external relations. The means and standard deviations on 5 trials (different sets of randomly extracted relations) are reported and the area in gray represents the benefice of the LPS process with respect to the structuring based on the external knowledge only.

First, we can notice that the obtained $F_{Measures}$ are much lower than the scores obtained on the previous general domains. This statement reveals the complexity of the task that is reinforced by the fact that UMLS is not structured with only is-a relations but also with part-of subsumptions. Despite that, we clearly observe that the acquired taxonomies take advantage of the proposed semi-supervised LPS methodology, especially in situations where the given external knowledge is lacking. This situation matches with a more realistic practical context of use.

Finally, we performed the same experiment on the Nervous system subdomain that contains 28 terms mainly structured with is-a relations. We can observe in Figure 6 (right) the same tendency but with strongly increased benefits although on a small dataset.

5 Conclusions

In this paper, we proposed a new learning strategy to efficiently combine linguistic and statistical features for lexical taxonomy acquisition. This methodology uses the pretopological formalism into which we defined the new concept of P-Space that relies on a parameterized pseudo-closure operator formalized in a multi-criteria analysis context. Then, we developed a semi-supervised strategy called LPS to learn P-Spaces in the taxonomy induction perspective. Experiments confirmed our expectations on both general and specialized domains. In particular, significant $F_{Measure}$ improvements are obtained for the auto-supervised context when compared to recent works [6]. Moreover, where pattern-based methodologies [6, 17] fail to learn LTs due to the absence of pattern evidences (usually for specialized domains), the introduction of external knowledge combined with statistical features allows the construction of LTs with reasonable accuracy.

References

1. Z.T. Belmandt. *Basics of pretopology.* Hermann, 2011.
2. Georgeta Bordea, Paul Buitelaar, Stefano Faralli, and Roberto Navigli. Semeval-2015 task 17: Taxonomy extraction evaluation (texeval). In *Proceedings of the 9th International Workshop on Semantic Evaluation.* Association for Computational Linguistics, 2015.

[11] http://www.nlm.nih.gov/research/pubmed/

3. Philipp Cimiano, Andreas Hotho, and Steffen Staab. Learning concept hierarchies from text corpora using formal concept anaylsis. *Journal of Artificial Intelligence Research*, 24:305–339, 2005.
4. Philipp Cimiano, Alexander Mädche, Stephen Staab, and Johanna Völker. Ontology learning. In *Handbook of Ontologies*, pages 245–267. Springer Verlag, 2009.
5. Guillaume Cleuziou, Davide Buscaldi, Vincent Levorato, and Gaël Dias. A pretopological framework for the automatic construction of lexical-semantic structures from texts. In *20th ACM International Conference on Information and Knowledge Management (CIKM)*, pages 2453–2456, 2011.
6. Zornitsa Kozareva and Eduard Hovy. Tailoring the automated construction of large-scale taxonomies using the web. *Language Resource Evaluation*, 47(3), 2013.
7. Zornitsa Kozareva, Ellen Riloff, and Eduard Hovy. Semantic class learning from the web with hyponym pattern linkage graphs. In *46th Annual Meeting of the Association for Computational Linguistics: Human Language Technology (ACL-HLT)*, pages 1048–1056, 2008.
8. Christine Largeron and Stéphane Bonnevay. A pretopological approach for structural analysis. *Information Sciences*, 144:169–185, July 2002.
9. George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. Introduction to wordnet: An on-line lexical database. *International Journal of Lexicography*, 3(4):235–244, 1990.
10. Roberto Navigli and Paola Velardi. Learning domain ontologies from document warehouses and dedicated websites. *Computational Linguistics*, 30(2):151–179, 2004.
11. Gerhard Paaß, Jörg Kindermann, and Edda Leopold. Learning prototype ontologies by hierachical latent semantic analysis. In *Workshop on Knowledge Discovery and Ontologies at the joint European Conferences on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD)*, 2004.
12. Fernando Pereira, Naftali Tishby, and Lillian Lee. Distributional clustering of english words. In *31st Annual Meeting on Association for Computational Linguistics (ACL)*, pages 183–190, 1993.
13. C. J. Van Rijsbergen. *Information Retrieval*. Butterworth-Heinemann, Newton, MA, USA, 2nd edition, 1979.
14. Mark Sanderson and Bruce Croft. Deriving concept hierarchies from text. In *22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 206–213, 1999.
15. Luca Scrucca. Ga: A package for genetic algorithms in r. *Journal of Statistical Software*, 53(4):1–37, 2013.
16. Rion Snow, Daniel Jurafsky, and Y. Andrew Ng. Semantic taxonomy induction from heterogenous evidence. In *21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 801–808, 2006.
17. Paola Velardi, Stefano Faralli, and Roberto Navigli. OntoLearn Reloaded: A Graph-Based Algorithm for Taxonomy Induction. *Computational Linguistics*, 39(3):665–707, 2013.
18. Hui Yang and Jamie Callan. A metric-based framework for automatic taxonomy induction. In *Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 271–279, 2009.