A Pretopological Framework for the Automatic Construction of Lexical-Semantic Structures from Texts

Guillaume Cleuziou
LIFO - Université d'Orléans
Orléans, France
guillaume.cleuziou@univ-orleans.fr

Davide Buscaldi
LIFO - Université d'Orléans
Orléans, France
davide.buscaldi@univ-orleans.fr

Vincent Levorato
LIFO - Université d'Orléans
Orléans, France
vincent.levorato@univ-orleans.fr

Gaël Dias
HULTIG - Universidade da Beira Interior
DLU - Université de Caen Basse-Normandie
ddg@hultig.di.ubi.pt

ABSTRACT

In this paper, we present a new approach for the automatic generation of lexical-semantic structures from texts. In particular, we propose a pretopological framework to formalize and combine various hypotheses on textual data in order to automatically derive a structure similar to common lexical-semantic knowledge bases such as WordNet. In addition, we define a new metric to intrinsically evaluate structures.

Categories and Subject Descriptors

I.2.4 [**Artificial Intelligence**]: Knowledge Representation Formalisms and Methods; I.2.7 [**Artificial Intelligence**]: Natural Language Processing

Keywords

Lexical-Semantic Structures; Pretopology; Evaluation

1. INTRODUCTION

Coding the semantic relationships between concepts of discourse into a lexical-semantic structure may enrich the reasoning capabilities of Information Retrieval and Natural Language Processing applications. However, their development is largely limited by the efforts required for their construction. To reduce the amount of work needed, many research have appeared in recent years to learn such structures from texts, fostering new surveys in the field [1, 5]. Learning lexical-semantic resources from texts instead of manually creating them has undeniable advantages. First, creating resources from texts within a domain may fit the semantic component neatly and directly, which will never be possible with general-purpose resources. Second, the cost per entry is greatly reduced, giving rise to much larger resources than an advocate of a manual approach could ever afford.

Different learning methods have been proposed to automatically build lexical-semantic structures. They can be grouped into three main classes: the similarity-based methods [12, 3], the set-theoretical approaches [9, 4] and the associative frameworks [13, 7]. In this paper, we aim at learning terminological ontologies following the associative framework but **relieving the frequency problem** evidenced in [13, 7]. For that purpose, we propose to analyze the topology of the graph structure between terms induced by associative measures. Within this context, we propose an **unsupervised methodology** based on Pretopology, which automatically learns lexical-semantic structures. Thus, from a given set of terms from potentially different domains and any domain corpus, we assess the asymmetric proximity between terms using asymmetric similarity measures. From the resulting proximity matrix, we present a Pretopological framework to obtain a non-triangular directed acyclic graph corresponding to the semantic structure of the domains.

The evaluation of learned structures is a rather complicated task. Indeed, [14] claims that there are several possibilities of conceptualizations for one domain that might differ in their usefulness for different groups of people, but not in their soundness and justification. In this paper, we propose an exhaustive intrinsic evaluation of the learned lexical-semantic structures by comparing them to the state-of-the-art approach [13]. For that purpose, we take as baseline the work done by [11] and propose an alternative solution to overcome some evidenced drawbacks as well as we present an original methodology for a "fair" comparison.

2. PRETOPOLOGICAL FRAMEWORK

The links between elements of a population can be modeled in several ways, e.g. Topology. However, topological axioms and properties are too restrictive to model a space in concrete terms. Instead, Pretopology models proximity in a more general way. So, we propose to use this theoretical framework to model a "lexical space" with pretopological relations and derive a structure with propagation strategies.

CIKM'11, October 24–28, 2011, Glasgow, Scotland, UK.

2.1 Notions of Pretopology

We can define a pretopological space by a family of neighborhoods. Let (E, a) be a pretopological space [2] where $a(.)$ is a pseudo-closure function and E a non-empty set. A neighborhood $N(x)$ of $x \in E$ is a subset of E containing x and a family of neighborhoods $\mathcal{N}(x)$ for x can be defined by the union of neighborhoods as $\mathcal{N}(x) = \{N \subseteq E | x \in N\}$. We then construct the pseudo-closure function based on the family of neighborhoods as $\forall A \in \mathcal{P}(E), a(A) = \{x \in E | \ \forall \ N \in \mathcal{N}(x), N \cap A \neq \emptyset\}$. Within our context, a pretopological space is defined by a vocabulary E (set of terms) and a pseudo-closure operator $a(.)$ supposed to model the propagation of semantic dependencies over term sets. The way to define the family of neighborhoods is crucial for the modeling. For example, the approach proposed by [13] can be instantiated in our pretopological framework by considering a family composed of two neighborhoods $\mathcal{N}_{OHC}(x) = \{N_O(x), N_{HC}(x)\}$ matching the two properties (high confidence and order respectively) used in the subsumption definition of [13] i.e. $N_{HC}(x) = \{y \in E | P(y|x) > t\}$ and $N_O(x) = \{y \in E | P(y|x) \geq P(x|y)\}$.

As the pseudo-closure function is not idempotent, its successive applications lead to the achievement of closed subsets. These closed subsets represent interdependent subsets related to the pseudo-closure function. As a consequence, a structure is induced by the elementary closed subsets and maximal closed subsets can be seen as the less homogeneous groups of E. The nature of these particular subsets is interesting in terms of space analysis, as we can consider an inclusion relation between them, leading to a structural analysis algorithm. Such a structure can be obtained with the pretopological algorithm proposed by [10]. In particular, we proposed a top-down version of this algorithm in [6]. So, when using $\mathcal{N}_{OHC}$, the algorithm provides a non-triangular directed acyclic graph that is exactly the final structure obtained by [13]. In the next section, we take advantage of this general framework to propose new neighborhoods relevant for lexical space modeling.

2.2 Lexical Space Modeling

Various pretopological neighborhoods may exist to model the proximity relations between elements in the vocabulary.

2.2.1 *K-Nearest Neighbors*

A k-Nearest Neighbors pretopological space (k-NN) consists in defining the neighborhood of an element x by the subset composed of the k elements having the highest proximity with x. For the lexical application, we choose as neighborhood for a term x, the terms y with the highest confidences using $P(y|x)$. As such, we define the following family of neighborhoods: $\mathcal{N}_{kNN}(x) = \{N_{kNN}(x), N_O(x)\}$ where $N_{kNN}(x) = \{y \in E | y \in kNN_E(x)\}$. The family $\mathcal{N}_{kNN}$ leads to a pseudo-closure operator $a(.)$ such that $a(x)$ is the set $\{x\}$ extended by its more general terms (N_O), which have x among its k best direct predecessors (N_{kNN}).

2.2.2 *Directed Relative Neighborhood (DRN)*

The second modeling is based on a statistical property observed on lexical structures. Given a benchmark structure $\mathcal{S}_r$ as reference and a corpus on the domain of $\mathcal{S}_r$, we performed an analysis on the distribution of the confidence values along the paths from a root to a leaf on the reference. Several intuitive hypotheses have been tested and one of them appeared to be statistically relevant. Let $x_1, x_2, \ldots, x_n$ be a path in the reference structure such that x_i subsumes x_{i+1}. We observed that a term x_i in the path has a higher minimal confidence with its predecessors than its successors have with their own predecessors. This statement can be formalized by the following property: $\forall i, \ min\{P(x_j|x_i)\}_{j=1}^{i-1} \geq min\{P(x_j|x_{i+1})\}_{j=1}^{i}$. By applying this property locally on a triplet (w, x, y), y is a neighbor of x if and only if any w satisfies the property to be a successor of x in the path (x, y) i.e. $N_{DRN}(x) = \{y \in E | \forall w \in E, \ P(y|x) \geq min\{P(x|w), P(y|w)\}\}$. As a consequence, a new family of neighborhoods is proposed $\mathcal{N}_{DRN}(x) = \{N_{DRN}(x), N_O(x)\}$. In particular, the pseudo-closure operator derived from $\mathcal{N}_{DRN}$ extends a term singleton $\{x\}$ with its more general terms (N_O) satisfying the extended ultrametric property (N_{DRN}). The interesting property of $\mathcal{N}_{DRN}$ is that it is free of parameter. However, it leads on practice to over-sized neighborhoods. A way to adjust the neighborhoods consists in introducing the high confidence parameter such that $\mathcal{N}_{HC\_DRN}(x) = \{N_{DRN}(x), N_O(x), N_{HC}(x)\}$. Another solution that avoids the threshold problem is presented in the next section.

2.2.3 *K-Nearest DRN*

As mentioned above, the *Directed Relative Neighborhood* produces over-sized neighborhoods. However, it could be used as a relevant "filter" and force other neighborhood functions to select elements satisfying a property observed on expected structures. In that sense, we define a new neighborhood that combines on the one hand the structural benefits and the simplicity of the parametrization of the kNN approach and on the other hand the statistical property of the DRN topology such that $N_{kN\_DRN}(x) = \{y \in E | y \in kNN_{N_{DRN}(x)}(x)\}$. Finally, the new family of neighborhoods is given by $\mathcal{N}_{kN\_DRN}(x) = \{N_{kN\_DRN}(x), N_O(x)\}$.

3. INTRINSIC EVALUATION

Two benchmarks have been used as references to tackle two different semantic relationships: synonymy and meronymy. First, we used the UMLS[1] from which four distinct sub-domains have been selected (cardiovascular (CS), digestive (DS), respiratory (RS) and nervous (NS)). In particular, each sub-domain is represented by its own lexical structure present in the meta-thesaurus using the hypernym/hyponym relation. The second reference ontology was obtained from WordNet by considering all geographical places deriving from the concept "United States of America" by means of the meronymy relation. We call it GEO-WordNet. For each reference, we retrieved the proximities between terms from two different corpora. For the UMLS, we used (1) PubMed[2] and (2) BioMed[3]. For the GEO-WordNet, we exploited the Glasgow Herald (GH95) and Los Angeles Times (LAT94) both used in the GeoCLEF evaluation campaigns[4], where toponyms have been identified with the Stanford Named Entity Recognition (NER) [8] and disambiguated using a conceptual-density based method.

[11] proposed a way to compare ontologies at the structural level (the J_1 measure). Given a set of terms E and two

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

[2] http://www.ncbi.nih.gov/pubmed/

[3] http://www.biomedcentral.com/

[4] http://ir.shef.ac.uk/geoclef

ontologies $\mathcal{O}_1$ and $\mathcal{O}_2$ structuring E, the general principle is to compare for each entry $x \in E$ the matching between the super/subconcepts of x in $\mathcal{O}_1$ and the super/subconcepts of x in $\mathcal{O}_2$. This evaluation approach is also suitable in our context by quantifying the matching between the predecessors $Pred_{\mathcal{S}}(x)$ and successors $Succ_{\mathcal{S}}(x)$ of a term x in the two lexical structures $\mathcal{S}_1$, $\mathcal{S}_2$. However, the main drawback of J_1 is its insensitivity to the direction of the relations into the structures. Such, two structures with full inversion would have a perfect matching according. To avoid the inversion problem, we propose to consider separately predecessors and successors in the matching evaluation. A new J_2 matching index is proposed as the (geometric) mean of two Jaccard indices for which a perfect matching of 1 implies strictly identical structures as shown in Equation 1.

$$J_2(\mathcal{S}_1, \mathcal{S}_2) = \frac{1}{|X|} \sum_{x \in E} \left(\frac{|Pred_{\mathcal{S}_1}(x) \cap Pred_{\mathcal{S}_2}(x)|}{|Pred_{\mathcal{S}_1}(x) \cup Pred_{\mathcal{S}_2}(x)|} \right)^{1/2} \cdot \left(\frac{|Succ_{\mathcal{S}_1}(x) \cap Succ_{\mathcal{S}_2}(x)|}{|Succ_{\mathcal{S}_1}(x) \cup Succ_{\mathcal{S}_2}(x)|} \right)^{1/2} \qquad (1)$$

In order to evaluate the acquired structure with the intrinsic J_2 measures, we need to fix the parameter k or the threshold t depending on the pretopological space used. In our experiments, we noticed two generic phenomena: (1) the best structures are obtained with small-sized neighborhood and (2) the size of the neighborhoods must be comparable to be fair in the comparison of the different pretopological spaces. The first observation corroborates the intent of [13] to filter only very high confidence values by means of a strong threshold (e.g. $t = 0.8$). This is illustrated by Figure 1 that reports the J_2 scores obtained by the $\mathcal{N}_{OHC}$ pretopological space when the number of confidence values retained grows (i.e. threshold t decreases). Furthermore, such a threshold cannot be universal and must be adjusted for each corpus. For example, the proportion of confidence values greater than 0.8 is about 1% for the CS sub-domain of the UMLS on the BioMed corpus, but only 0.07% for the GEO-WordNet observed on the LAT94 corpus.

Based on these findings, we chose two heuristics for the parametrization of the pretopological spaces. Let n be the size of the vocabulary E to structure, the threshold t is adjusted in such a way that only n (first heuristic) and $2n$ (second heuristic) confidence values exceed t. The two heuristics are used for high confidence-based pretopological spaces (i.e. $\mathcal{N}_{OHC}$ and $\mathcal{N}_{HC\_DRN}$). For the nearest neighbors-like spaces (i.e. $\mathcal{N}_{kNN}$ and $\mathcal{N}_{kN\_DRN}$) neighbors with comparable sizes are obtained with parameters $k = 1$ and $k = 2$ respectively. Table 1 reports the structural evaluation of each acquired structure compared to the corresponding reference.

Important variations on the J_2 index are observed depending of the benchmark and the corpus. Very poor matching are obtained for example on the RS with PubMed where the scores are sometimes lower than 0.10 and strong promising matching are obtained for example on the NS sub-domain and the Geographical domain where the scores are closed or higher than 0.40. Such variations can be explained by the nature of the reference used and especially the kind of semantic relations into consideration. Some of the domains are structured based exclusively on the *Part-of* relation (e.g. Geo-WordNet) or the *Is-a* relation (e.g. NS sub-domain), while other references mix up both types of relations such as the CS sub-domain (and the global UMLS reference by extension). It seems to be incontestable that such an heterogeneity in the semantic structuring of the vocabulary is a problem that current approaches do not transcend.

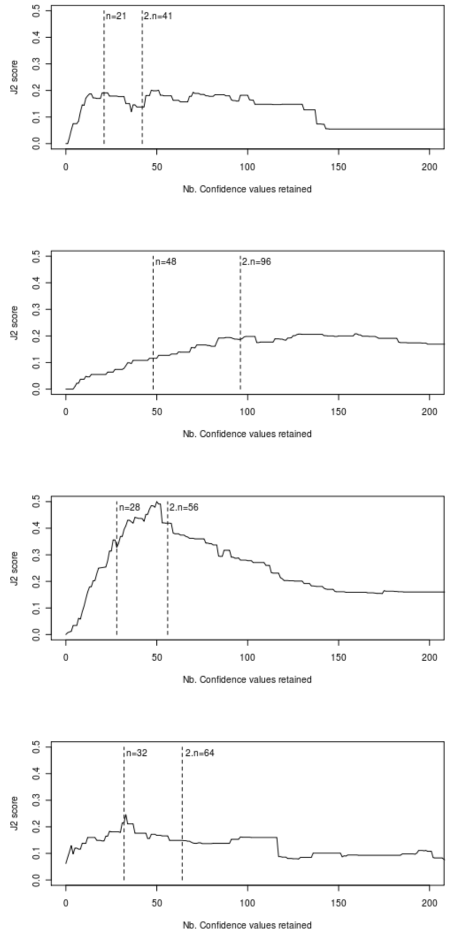

Figure 1: J_2 scores with the baseline approach of [13] on the BioMed corpus: CS, DS, NS and RS sub-domains in the top-down order.

The corpus used and the way to exploit the text collection also have a significant impact on the quality of the retrieved statistics and then on the acquired lexical structure. Indeed, the matching scores are lower overall when using the PubMed corpus compared to BioMed. It is mainly due to the fact that only abstracts of the scientific papers have been used for the statistics computation with PubMed whereas full text retrieval has been performed on BioMed, thus providing more confidence in the extracted statistics.

Table 1 also reports the comparison between the structures obtained from different pretopological spaces. Bold values in the table distinguish the modeling that leads to the best matching for a benchmark and a corpus. It is interesting to observe that even if the ultrametric topological space ($\mathcal{N}_{HC\_DRN}$) never leads to the best matching, the filter performed by this space is profitable to the nearest neighbors-like space since the $\mathcal{N}_{kN\_DRN}$ modeling obtains best results on four experiments. As a summary one can notice that the new proposed $\mathcal{N}_{kN\_DRN}$ modeling outperforms the baseline

Corpora	Domain	n	$\mathcal{N}_{OHC}$		$\mathcal{N}_{kNN}$		$\mathcal{N}_{HC\_DRN}$		$\mathcal{N}_{kN\_DRN}$	
			n	$2n$	k=1	k=2	n	$2n$	k=1	k=2
BioMed	Cardiovascular system	21	0.191	0.138	0.189	**0.304**	0.192	0.133	0.144	0.136
	Digestive system	48	0.116	**0.187**	0.104	0.137	0.116	0.175	0.110	0.116
	Nervous system	28	0.328	0.419	**0.428**	0.382	0.344	0.392	**0.428**	0.414
	Respiratory system	32	0.215	0.149	0.188	0.240	0.220	0.138	0.154	**0.251**
	UMLS (4 sub-domains)	128	0.172	**0.218**	0.151	0.162	0.184	0.213	0.180	0.173
PubMed	Cardiovascular system	21	0.100	**0.173**	0.133	0.147	0.097	0.162	0.133	0.166
	Digestive system	48	0.130	0.107	0.123	0.111	0.117	0.138	**0.243**	0.188
	Nervous system	28	0.196	0.258	**0.440**	0.401	0.208	0.257	0.429	0.381
	Respiratory system	32	0.095	0.119	**0.143**	0.139	0.092	0.127	0.131	0.101
	UMLS (4 sub-domains)	128	0.102	0.132	0.165	0.142	0.096	0.145	0.169	**0.171**
LAT94	GEO-WordNet (USA)	150	0.207	0.312	**0.392**	0.332	0.183	0.276	0.386	0.347
GH95		131	0.305	0.372	**0.399**	0.382	0.289	0.312	0.391	0.382

Table 1: Structural matching of each acquired structure with its reference using the J_2 index.

proposed by [13] in two thirds of the experimented contexts and sometimes with strong improvements as for example on the geographical benchmark with a score increased by 87% at the very most.

4. CONCLUSIONS

In this paper, we presented a new framework to automatically build terminological ontologies based on the formalism of Pretopology. In particular, Pretopology proposes a well-founded mathematical framework to model the degree of generality/specificity as well as the semantic closeness between terms based on an asymmetric proximity measure. Unlike similarity-based and set-theoretic approaches, we deal with asymmetry in NLP based on the associative framework, which allows domain and language independency and opens new research directions. In particular and compared to the work of [13], we proposed to focus on the topology of the structure obtained from the proximity measure thus avoiding isolated terms and simplifying the parametrization. We also proposed an exhaustive intrinsic evaluation of the learned lexical-semantic structures based on a new metric called the J_2 index, which proposes a solution to the ontology inversion problem untreated by [11]. We validated our model based on two benchmarks: the UMLS referential medical ontology over the PubMed and BioMed corpora, and the GEO-WordNet referential over two news stories collections. We compared it to the best-so-far state-of-the-art methodology proposed by [13]. The results showed that the pretopological structuralist formalism outperforms the methodology of [13] in the majority of the cases.

5. REFERENCES

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

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

[3] P. Cimiano, A. Hotho, and S. Staab. Comparing conceptual, partitional and agglomerative clustering for learning taxonomies from text. In *Proceedings of the European Conference on Artificial Intelligence (ECAI 2004)*, pages 435–439, 2004.

[4] P. Cimiano, A. Hotho, and S. Staab. Learning concept hierarchies from text corpora using formal concept anaylsis. *Journal of Artificial Intelligence Research*, 24:305–339, 2005.

[5] P. Cimiano, A. Mädche, S. Staab, and J. Völker. Ontology learning. In *Handbook of Ontologies*, pages 245–267. Springer Verlag, 2009.

[6] G. Cleuziou, G. Dias, and V. Levorato. Acquisition de structures lexico-sémantiques à partir de textes : un nouveau cadre de travail fondé sur une structuration prétopologique. In *Proceedings of the 11ème Conférence Internationale Francophone sur l'Extraction et la Gestion des Connaissances*, 2011.

[7] G. Dias, R. Mukelov, and G. Cleuziou. Unsupervised graph-based discovery of general-specific noun relationships from web corpora frequency counts. In *Proceedings of the 12th International Conference on Natural Language Learning (CoNLL 2008)*, 2008.

[8] J. Finkel, T. Grenager, and C. Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics on Computational Linguistics (ACL 2005)*, pages 363–370, 2005.

[9] B. Ganter and R. Wille. *Formal Concept Analysis: Mathematical Foundations.* 1998.

[10] C. Largeron and S. Bonnevay. A pretopological approach for structural analysis. *Information Sciences*, 144:169–185, 2002.

[11] A. Maedche and S. Staab. Measuring similarity between ontologies. In *Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web (EKAW 2002)*, pages 251–263, 2002.

[12] G. Paaß, J. Kindermann, and E. Leopold. Learning prototype ontologies by hierachical latent semantic analysis. In *Proceedings of the 15th Joint Conference European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD 2004)*, 2004.

[13] M. Sanderson and D. Lawrie. Building, testing, and applying concept hierarchies. In W. B. Croft, editor, *Advances in Information Retrieval*, pages 235–266. Kluwer Academic Publishers, 2000.

[14] B. Smith. Ontology. In *The Blackwell Guide to Philosophy of Computing and Information*, pages 155–166. Malden: Blackwell, 2004.