Concept Building with Non-Hierarchical Relations Extracted from Text – Comparing a Purely Syntactical Approach to a Semantic one

Sílvia Maria Wanderley Moraes, Vera Lúcia Strube de Lima and Luis Otávio Furquim

Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS)
Faculdade de Informática
Av. Ipiranga, 6681 Prédio 32, Porto Alegre, RS, Brazil
silvia.moraes@pucrs.br, vera.strube@pucrs.br, luisfurquim@gmail.com

Abstract. In this paper, we introduce Semantic Lattice (SemLat), a method that allows the construction of concept lattices from lexical-semantic information extracted from PropBank-style labelled texts. We apply SemLat to Tourism and Finances domain texts from Wikicorpus 1.0 through case studies that are examined in detail. We compare conceptual structures generated by SemLat, that makes use of semantic relations, to structures generated from purely syntactic relations. We intrinsically evaluate the structures using a semantic-similarity based structural measure. We also analyse, in a qualitative approach, the contribution of semantic roles in concept formation. We claim that conceptual structures generated by SemLat produce richer concepts as they provide intentional descriptions that are more informative, from a semantic point of view.

Keywords: Formal Concept Analysis, Semantic Role, Concept Lattices.

1 Introduction

Conceptual structures such as terminologies, thesauri, taxonomies and ontologies are important resources for information systems. Since building and maintaining such structures is costly, automatic and semi-automatic approaches have been proposed to minimize the effort of extracting concepts and semantic relations from texts. We are interested in exploring the potential of the semantic roles in the learning of conceptual structures. A semantic role expresses the meaning of an argument in a situation described by the verb in a sentence. With the use of semantic roles, we can identify, for example, the agentive entity of an action, even if it appears in diverse syntactical positions through the text. In this paper, we present the Semantic Lattice (SemLat) - a simple method to generate concept lattices from semantic relations extracted from texts, exploring the benefits of Formal Concept Analysis (FCA) as a conceptual clustering method. We intrinsically evaluate the conceptual lattices built, using a structural measure based on semantic similarity. We qualitatively analyse the contribution of semantic roles in the formation of concepts. Results show that conceptual structures created by SemLat generate richer concepts, as they provide intentional descriptions that are more informative, from a semantic point of view.

This paper is organized in 6 sections. In Section 2 we study related work. Section 3 shortly introduces semantic roles and FCA. Section 4 briefly describes the SemLat method. Section 5 presents the studies concerning SemLat and Section 6 brings our conclusions.

2 Related Work

The idea of combining the FCA method with semantic roles is not new. Kamphuis and Sarboin [1] propose to represent a sentence in natural language, associating FCA to semantic roles. They deal with two types of linguistic relations: minor (nouns to adjectives and adverbs) and major (verbs to nouns). Differently from that work, we extract relations from linguistically tagged texts namely the major ones. Rudolf Wille [2] also presents examples of FCA structures combined with semantic roles. He combines conceptual graphs with FCA structures, aiming the formalization of useful logic to representation and processing. There are no comments, in his work, on the processing of information present in the conceptual graphs, so we understand that neither the construction of these graphs nor their mapping into FCA structures, were performed automatically. Our study deals with the automatic extraction of information from texts (to generate representation structures) and we analyse the limits of our approach. The FCA method was aheady combined with semantic roles, as in [3], where efforts turn to the linguistic analysis as a purpose for representing FrameNet through concept lattices. Distinct from our work, the authors do not use FCA as a support to build ontological structures from texts. Instead, we use textual information and PropBank annotation to identify the roles. Although the approaches in [1,2,3] seemed promising at the time they have been proposed, they were little explored probably due to the difficulties with the text annotation process, since the appearance of automatic semantic role annotators is more recent. Even with thorough literature review, we did not find, to date, studies that explore the use of semantic roles in conjunction with the FCA method to support construction of ontological structures from texts. We address this issue in our research.

3 Semantic Roles and FCA

Semantics roles are “roles within the situation described by a sentence” [4]. Although there is no consensus on a single list of semantic roles, some are widely accepted [5] such as: Agent, Patient, Instrument, Theme, Source and Destination. The barrier regarding the definition of roles has been circumvented by assigning numerical labels (A0, A1, A2, ...) to the arguments of the verbs. This is the case for PropBank[1] corpus, which has been extensively used to train semantic role taggers for the English language. The F-EXT-WS tool used to tag

[1] http://www.cis.upenn.edu/˜ace

the corpora in the present study, also adopts these labels [6]. For the English language, it provides Part-of-Speech (POS) tagging, syntactical annotation and semantic roles tagging. F-EXT-WS uses the tags defined for PennTreeBank [2].

FCA was introduced by Rudolf Wille in the 80's as a method for data analysis [2]. A key element in FCA is the formal context, characterized by the triple (G, M, I), where: G is the set of domain entities, called formal objects; M consists of the features of these entities, their formal attributes; and I is the binary relation on $G \times M$, called the incidence relation, which associates a formal object to its attributes. The formal concepts are built from the formal context. A formal concept is determined by the pair (O, A) if and only if $O \subset G$ and $A \subset M$. Once the concepts have already been defined, the concept lattice is created [7].

4 The SemLat Method

The SemLat method is the result of several studies, including Relational Concept Analysis [8], in the interest of how to include semantic roles in lattices [9,10]. SemLat comprises 3 stages shown in Fig. 1. The SemLat input is a corpus annotated with lexical-semantic information, lemmatized. From this corpus we create the conceptual structure.

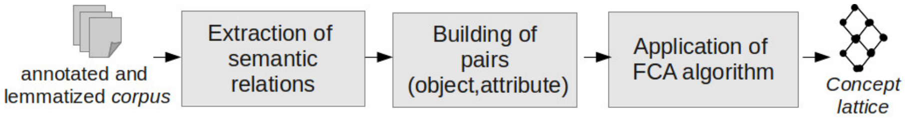
annotated and
lemmatized *corpus*
Extraction of semantic relations
Building of pairs (object,attribute)
Application of FCA algorithm
Concept lattice

Fig. 1. SemLat stages

The 'Extraction of semantics relations' stage consists of the building of tuples containing, for a certain verb, its arguments and the semantic roles associated with these arguments. Aiming to build a conceptual structure, relevant noun phrases are extracted from the arguments. The steps to build tuples are:

1. To analyze the sentences, identifying and extracting verbs and respective arguments and associated semantic roles.
2. To identify the noun phrases in the verb arguments discarding those formed by proper nouns (as we have not included an instance level in the ontological structure).
3. To form tuples, using information extracted from sentences in steps 1 and 2. Each tuple must contain noun phrases and their correspondent semantic roles. Tuples are in the following format: (np_1,sr_1,np_2,sr_2) where np_i and sr_i correspond, respectively, to the noun phrase and its semantic role.

Let's consider the following sentence from PropBank: "The financial-services company will pay 0.82 share for each Williams share." After annotating (Fig. 2) the sentence with the use of F-EXT-WS, we are able to extract necessary lexical-semantic information from this sentence and complete the tuple (company, A0, share, A1).

[2] http://www.cis.upenn.edu/~treebank/

(A0 (DT The) (NNS financial-services) (NN company)) (MD will) (V (VB pay)) (A1 (CD 0.82) (NN share) (IN for) (DT each) (NNP Williams) (NN share)).

Fig. 2. Sentence annotated with F-EXT-WS

The second stage aims to produce the object-attribute pairs that will give origin to the FCA formal context. From each tuple (np_1,sr_1,np_2,sr_2) extracted from the texts, two object-attribute pairs are created: $(np_1,sr_1\_\text{of}\_np_2)$ and $(np_2,sr_2\_\text{of}\_np_1)$. So, from the tuple (company, A0, share, A1) the pairs (company, A0_of_share) and (share, A1_of_company) are created. Frequently, A0 corresponds to Agent and A1 to Patient. With the use of semantic roles, we can better determine the relationship between the nouns: company is an agent of share, and share is a patient of company. As many pairs can be generated, in order to avoid an excessively sparse formal context, we group concepts, as described in [9]. The pairs created, the formal context can be built. SemLat's last stage consists of the generation of the conceptual structure (Fig. 3). In order to accomplish this task, FCA algorithms, such as Bordat [11], can be used. Another alternative is to use a specific tool to generate lattices such as Concept Expert[3] 1.3 .

5 Studies concerning SEMLAT

We compare structures built with the SemLat method (Fig. 3b) to those built with FCA exclusively based on the syntactic relations between verbs and their arguments, as proposed by Cimiano in [12] (Fig. 3a). In order to accomplish this task, we use Wikicorpus[4] 1.0 comprising Wikipedia texts. We randomly took from Wikicorpus 322 texts of the Finances domain and 284 texts of the Tourism domain. These subsets were named correspondingly, WikiFinance and WikiTourism. Both corpora were annotated with lexical-semantic information using F-EXT-WS. We lemmatized nouns present in the identified noun phrases with TreeTagger[5]. To analyse the contribution brought with the semantic roles in the formal concepts formation, we outlined two case studies:

- case (np, v): describes syntactical relations of the type verb-argument.
- case (np, sr_of_np): describes semantic relations obtained with SemLat.

With these two studies and using WikiFinance and WikiTourism corpora, we produced four conceptual structures to be examined (only relations with a minimum frequency of 2 were considered):

- TourismFCA (np, v): from case (np, v) for the WikiTourism corpus.
- TourismFCA (np, sr_of_np): from case (np, sr_of_np) for the WikiTourism corpus.

[3] http://sourceforge.net/projects/conexp

[4] http://nlp.lsi.upc.edu/wikicorpus/

[5] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

- FinanceFCA (np, v): from case (np, v) for the WikiFinance corpus. A subset of this structure is shown in Fig. 3a.
- FinanceFCA (np, sr_of_np): from case (np, sr_of_np) for the WikiFinance corpus (subset in Fig. 3b).

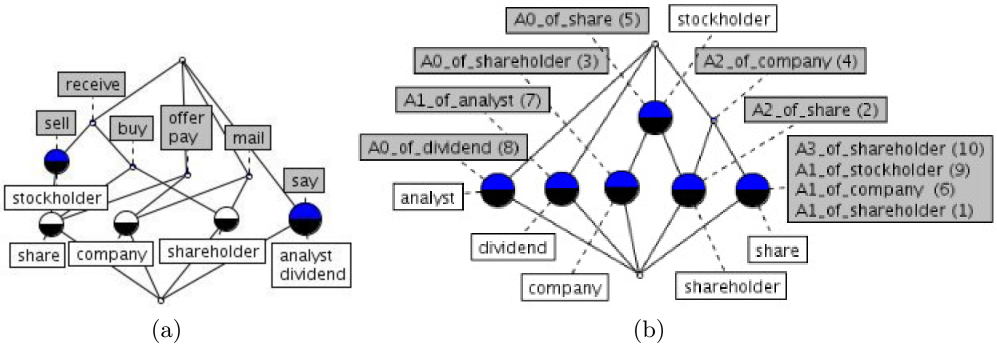

Fig. 3. Syntactic and semantic lattices

Table 1 contains information on these structures, including number of objects, attributes, concepts and edges. We notice that case (np, sr_of_np) has in average 4 times more attributes than case (np, v). This number of attributes was already expected, since in case (np, sr_of_np) the attributes are much more specific. This specificity increases in around 7% of the concepts in the Finances domain and in approximately 30% in the Tourism domain. This fact may be related to the scope of the texts in each domain.

Table 1. Information on the conceptual structures generated

FCA	**case**	**#objects**	**#attributes**	**#concepts**	**#edges**
Finance	(np,v)	631	237	760	2055
	(np,sr_of_np)	631	1018	819	1919
Tourism	(np,v)	383	121	239	529
	(np,sr_of_np)	383	535	343	633

In a subjective and shallow analysis, we perceive the Tourism domain texts are more restricted than the Finances ones. While Tourism texts mostly approach subjects related to attractions, texts from Finances include descriptions on the key terms in the domain. In the following sections we study the contribution of semantic roles in the formation of concepts.

5.1 Lexical cohesion

Although extensively studied, the evaluation of conceptual structures is still an issue to be further investigated. When we evaluate FCA-based structures, difficulties increase due to the fact that this investigation is recent. We found two measures ideally appliable to this evaluation [13,14] both comparing FCA structures regarding the objects and the formal attributes of their concepts. As the formal concepts generated from the case studies were not equivalently configured (they had different attributes), we could not apply these measures satisfactorily. So we focused our analysis on formal objects. The evaluation of these lattices was based on the structural Semantic Similarity Measure (SSM) [15]. SSM indicates how close are the concepts that match (exactly or partially) the search terms in an ontology. In the present study, SSM became a sort of lexical cohesion measure, as it was applied to the objects of each formal concept from the FCA structure. Typically, synonymy, hypernymy and meronymy are considered, when calculating cohesion. In order to obtain such cohesion value, as recommended in [15], we used in the SSM estimation the measure defined by Wu and Palmer [16] which takes semantic relations from an ontological structure to calculate the semantic distance between words. Equation (1) indicates the average lexical cohesion among the N concepts in a FCA structure, regarding a conceptual structure E.

$$SSM_E = \frac{1}{N} \textstyle\sum_{i=1}^{N} ssm_i \tag{1}$$

As detailed in Equation (2), ssm_i computes the similarity in the set of objects G of a concept i in a FCA structure, using Wu e Palmer (wup) measure. In case the cardinality of G is 1, ssm_i is zero.

$$ssm_i = \begin{cases} \dfrac{1}{|G_i|} \sum\limits_{j=1}^{|G_i|-1} \sum\limits_{k=j+1}^{|G_i|} wup_E(o_j, o_k) \;\; for \; |G_i| > 1 \; and \; o_j, o_k \in G_i \\ \qquad\qquad 0 \qquad\qquad\qquad\qquad o/w \end{cases} \tag{2}$$

Besides WordNet[6], we applied SSM over domain ontologies: LSDIS_Finance[7] and Finance[8] for the Finances domain, and Travel[9] and TGPROTON[10] for the Tourism domain. Although the extension and richness in WordNet relations, these relations are mostly general and do not refer to a specific domain. We believe that the measure proposed by Wu and Palmer [16], applied to the WordNet structure, might not fully capture the expected semantic relations so producing less expressive values. Besides, even if domain ontologies have a more concise concepts set (regarding its domain), it is more frequent to find n-gram labelled concepts ($n > 1$) as for the present studies. So, it is possible to assert that the relations among concepts are domain relations. These points may conduct to more significant results, from a semantic point of view, when concerning the quality of the clusters of concepts. Table 2 shows the results obtained from the application of SSM. In this table, W, F, L, TG and T correspond to the lexical resources used: WordNet, LSDIS_Finance, Finance, TGPROTON and Travel, respectively. As we imagined, SSM showed a low cohesion for both domains when using WordNet. As we expected, the domain ontologies have a cohesion

[6] http://wordnet.princeton.edu/

[7] http://lsdis.cs.uga.edu/projects/meteor-s/wsdl-s/ontologies/LSDIS_Finance.owl

[8] http://www.fadyart.com/ontologies/data/Finance.owl

[9] http://protege.cim3.net/file/pub/ontologies/travel/travel.owl

[10] http://goodoldai.org/ns/tgproton.owl

distinct from that found in the texts we used. In this case, cohesion values were low because less than 10% of the objects in concepts from a lattice were present in the ontologies. For the Tourism domain, we believe the variety of the texts was the main reason for the low matching. The absence of non-hierarchical relations in the selected ontologies caused some difficulties to the evaluation as well. As the semantic roles should express non-hierarchical relations, the cohesion of these relations were not computed in the evaluation results. As a next step we performed a qualitative analysis.

Table 2. SSM application results

	Finance			**Tourism**		
case	**SSM**$_w$	**SSM**$_F$	**SSM**$_L$	**SSM**$_w$	**SSM**$_{TG}$	**SSM**$_T$
(np,v)	0.33	0.33	0.27	0.18	0.05	0.02
(np,sr_of_np)	0.20	0.21	0.16	0.09	0.01	0.01

5.2 Qualitative Analysis

In this section we address, from a qualitative perspective, the importance of semantic roles in the formation of the formal concepts. Features inherent to semantic roles may help distinguish, classify and, essentially, better associate the elements extracted from texts. To illustrate this analysis, we used a subset from FinanceFCA (*np*, *sr*_of_*np*) and FinanceFCA (*np*, *v*) lattices. These subsets are those presented in Fig. 3. We perceived that the semantic roles caused the generation of an extra concept. The nouns *analyst* and *dividend* were not clustered in a same concept. However, the relation between them was not lost. In the structure obtained from case (*np*, *sr*_of_*np*) from Fig. 3b, transversal relations appear as attributes. The object *analyst* is defined as A0_of_dividend, meaning that it is the Agent of *dividend*. And the object *dividend* is A1_of_analyst, its patient. In both cases, the structures produce a concept for "share". In case (*np*, *sr*_of_*np*) we get to more clearly interpret the relation between *share* and the other elements of the domain. We can notice that share is usually patient (A1) of *stockholder*, *company* e *shareholder*. The *stockholder* concept showed to be a superconcept in both structures but, in case (*np*, *sr*_of_*np*), *share* was not its subconcept. This relation was expressed in the attributes. In case (*np* ,*sr*_of_*np*), *stockholder* as well as *company* and *shareholder*, its subconcepts, are agents (A0) of *share*. From this analysis we noticed that, even if the semantic roles make the concepts more specific, they are much more informative than the verbs. The concepts generated from case (*np*, *sr*_of_*np*) are semantically richer, from an intentional point of view, than those from case (*np*,*v*).

6 Conclusions

In this paper we depicted the SemLat method, which allows to build concepts based on semantic roles, using FCA as a conceptual clustering method. We then investigated the contribution of SemLat in the formation of concepts. From a structural and lexical point of view, it is still difficult to objectively evaluate the contribution of semantic roles in the building of formal concepts. The cohesion computed by SSM for the Tourism and Finances domains was inconclusive. From a qualitative point of view, we perceived semantically richer formal concepts. The inclusion of semantic roles in the formal attributes improved the intentional description of concepts. We are interested in the extrinsic evaluation of the concept lattices generated by SemLat. Presently, we are analysing the contribution of these structures in the text categorization task. Work on more appropriate methods for the evaluation of ontological structures is also important for future directions of the present study.

References

1. Kamphuis, V., Sarbo, J.: Natural language concept analysis. In: NeMLaP3/CoNLL98, ACL, Yellow Book (1998) 205–214
2. Wille, R.: Conceptual graphs and formal concept analysis. In: ICCS '97, London, UK, Springer-Verlag (1997) 290–303
3. Valverde-Albacete, F.J., Peláez-Moreno, C.: Galois connections between semimodules and applications in data mining. In: ICFCA. (2007) 181–196
4. Yule, G.: The Study of Language. Cambridge University Press, Cambridge (1996)
5. Jurafsky, D., Martin, J.H.: Speech and Language Processing: An Introduction to Natural Language Processing, Computational Lingustics, and Speech Recognition. Artificial Intelligence. Prentice Hall (2008)
6. Fernandes, E.R., Milidiu, R.L., Santos, C.N.: Portuguese language processing service. In: 18th International World Wide Web Conference. (2009)
7. Stumme, G., Darmstadt, T.H., Mathematik, F.: Exploration tools in formal concept analysis. In: Proc. OSDA 95. Studies in Classification, Data Analysis, and Knowledge Organization 8, Springer (1995) 31–44
8. Priss, U.: Relational Concept Analysis: Semantic Structures in Dictionaries and Lexical Databases. PhD thesis, Darmstadt, Aachen (1998)
9. Moraes, S.M.W., Lima, V.L.S.: Combining formal concept analysis and semantic information for building ontological structures from texts : an exploratory study. In: LREC'12, Istanbul, Turkey (2012)
10. Moraes, S.M.W.: Construção de Estruturas Ontológicas a partir de textos: um estudo baseado no método Formal Concept Analysis e em papéis semânticos. PhD thesis, Faculdade de Informática, PUCRS, Brazil (2012)
11. Fu, H., Nguifo, E.M.: Lattice algorithms for data mining. In: Revue ISI (Ingénierie des Systèmes d'Information). Volume 9., Edition Hermes (2004) 109–132
12. Cimiano, P., Hotho, A., Staab, S.: Learning concept hierarchies from text corpora using Formal Concept Analysis. JAIR **24** (2005) 305–339
13. Alqadash, F., Bhatnagar, R.: "Similarity Measures in Formal Concept Analysis". Annals of Mathematics and Artificial Intelligence **61** (2011) 245–256
14. Formica, A.: Concept similarity in formal concept analysis: An information content approach. Knowledge-Based Systems **21** (2008) 80–87
15. Alani, H., Brewster, C.: Metrics for ranking ontologies. In: Proceedings of the 4th EON2006 at the 15th WWW 2006, Edinburgh, Scotland (2006) 24–30
16. Wu, Z., Palmer, M.: Verbs semantics and lexical selection. In: ACL '94, Stroudsburg, PA, USA, Association for Computational Linguistics (1994) 133–138