Learning term to concept mapping through verbs: a case study

Valentina Ceausu
CRIP 5 - Paris V University
45, Rue des Saints Pères
75006 Paris, France
ceausu@math-info.univ-paris5.fr

Sylvie Després
LIPN UMR CNRS 7030 - University of Paris 13
99 avenue Jean Baptiste Clément
93430 Villetaneuse, France
sylvie.despres@lipn.univ-paris13.fr

ABSTRACT

We propose in this paper an approach to learn term to concept mapping with the joint utilization of verb relations and an existing ontology. This is a non-supervised solution that can be applied to any field for which an ontology modeling verbs as relations holding between the concepts was already created. Conceptual graphs, representing a set of verb relations, are learned from a natural language corpus by using part-of-speech information and statistic measures. Labeling strategies are proposed to assign terms of the corpus to concepts of the ontology by taking into account the structure of the ontology and the extracted conceptual graphs. Results of this assignation could be used to automatically create semantic annotations of documents. A first experimentation in the field of accidentology was done and its results are also presented.

Categories and Subject Descriptors
D.3.3: term to concept mapping, semantic annotation.

General Terms
Experimentation.

Keywords
Ontology, verb relation, concept learning.

1. INTRODUCTION

The rapid evolution in the production of documents in natural language requires the definition of efficient automated approaches allowing finding relevant information in those documents. This paper presents an approach that uses verb relations and a domain ontology to assign terms of a given corpus to concepts of the field. Those assignations can be used thereafter for various exploitation scenarios, that is to say: semantic annotation of documents, estimating similarities between documents, etc.

This approach is based on an entirely automatic and non-supervised process, unless the use of a domain ontology to support the process.

The task to achieve could be described as follows: let O be a domain ontology and C a collection of domain-specific texts.

For this work, we assume that the ontology takes into account the linguistic level of entities. Thus, concepts and roles are labeled by terms, which are linguistic manifestation of ontology entities in a specific language (French, English, etc.). Therefore, ontology considered for this work has two levels: a conceptual level, describing domain specific entities (concepts and roles) and a linguistic level, providing expressions of those entities in a given language.

The goal of this approach is to identify within C terms t representing linguistic expression of concepts of O ontology. Thus, we can label terms identified in the corpus by concepts of ontology. We propose a three steps approach to carry out this labeling process:

(1) in a first stage, verb relations are extracted from the corpus. Each verb relation is composed of a verb, be that a general one or a field specific one, and a pair of terms connected by this verb.

(2) in a second phase, statistical processing is performed to structure verb relations as conceptual graphs. As the verb is considered to be the key element of a verb relation, it is placed at the top of the conceptual graph. Terms occurring as arguments of the verb are connected to this verb through links representing theirs syntactic function which could be subject or object.

(3) the last phase is based on the assumption that the domain ontology models verbs of the field as relations holding between the concepts. If this is the case, labeling strategies are using both the ontology and extracted conceptual graphs to assign field specific terms to field specific concepts.

We shall approach that topic by answering a number of questions: which method should be used to extract verb relations from corpus? How to learn conceptual graphs from the extracted verb relations? Those questions are analyzed in sections 2 and 3. Given a domain ontology and a set of conceptual graphs, which strategies will be used to assign terms to concepts? The solution is discussed in section 4. A first experimentation in the field of accidentology is described and its results are presented in section 5. Related work is presented in section 6. Conclusions and perspectives end this paper.

2. EXRACTING VERB RELATIONS FROM CORPUS

To extract verb relations from corpus, we adopted an approach based on pattern recognition. This approach is using part-of-speech information and consists in seeking within the corpus for particular associations of lexical categories. Such an association represents a lexical pattern. For example *Verb, Noun* or *Verb, Preposition, Noun* are lexical patterns.

We manually crafted a set of lexical patterns including a verb (among other categories). Associations of words matching patterns of this set are identified by a pattern recognition algorithm, described in [4]. The algorithm takes as input the corpus tagged by TreeTagger, see [19] and a set of lexical patterns including verbs. It is applied at sentence level and it automatically generates a set of word regroupings matching those patterns, such as (examples of this paper are translated in English, although they are extracted from a French corpus experimentation: *Verb, Preposition : diriger vers* (direct to); *Verb, Preposition, Noun: diriger vers place* (direct to square).

Obtained word regroupings can be:

- a verb relation, highlighting domain relations, such as*: véhicule diriger vers bretelle* (vehicle direct to slip road);

- an incomplete verb relation such as *piéton traverser* (pedestrian crossing); or *diriger vers l'opéra* (direct to opera);

- or meaningless word regroupings, as we can see : *c, véhicule* (c, vehicle,) ; venir de i (come from i).

3. LEARNING CONCEPTUAL GRAPHS

The goal of this phase is to learn conceptual graphs from the results of pattern recognition algorithm.

A conceptual graph represents a hierarchy having as a top a verb and, on a second layer, arguments connected to the verb by their grammatical function, subject or object. We use the term *conceptual graph* as it was introduced by [18].

As many terms could be the subject or object of the same verb, a conceptual graph corresponds to a set of verb relations generated by the same verb. To learn conceptual graphs, the chain of treatments based on lexical similarity measures presented bellow is performed.

3.1 Lexical similarities and lexical distances

A lexical similarity measure associates a real number r to a pair of strings s,t. Important values of r indicate a significant similarity of strings s,t. In a similar way, a lexical distance measure associates a real number to a pair of strings but the interpretation is different: important values of r indicates minor similarity of strings s,t.

Many coefficients were proposed to calculate similarities or distances between strings. A number of them are presented in [5]. For this work, we have implemented the Jaccard, Jaro, Jaro-Winkler and Monge-Elkan coefficients–.

Jaccard coefficient calculates the similarity between two strings s and t by considering a string composed of several sub-strings. Jaccard coefficient is given by:

$$Jaccard(s,t) = \frac{\left| s \cap t \right|}{\left| s \cup t \right|}$$

This measure is takes into account the number of sub-strings common to s and t and the number of all sub-strings of s and t. If we consider characters as sub-strings, the coefficient expresses the similarity by taking into account the number of common characters of s and t only.

Jaro and Jaro-Winkler coefficients, introduced below, express the distance between two strings s,t by taking into account the number and the position of characters shared by s and t.

Let $s = s_1 \ldots s_k$ and $t = t_1 \ldots t_k$ be two strings. A character s_i in s is considered common to both strings if there exists t_j in t such as: $s_i = t_j$ and $i - h \le j \le i + h$, where $h = \frac{\min(s,t)}{2}$.

Let $s^1 = s_1^1 \ldots s_k^1$ be characters of s common to t and $t^1 = t_1^1 \ldots t_k^1$ characters of t common to s. We define a transposition between s and t as an index i such as: $s_i^1 \neq t_i^1$. If $T_{s,t}$ is the number of transpositions from s^i to t^i the Jaro coefficient calculates the lexical distance between s and t as follows:

$$Jaro(s,t) = \frac{1}{3}\left(\frac{\left|s^1\right|}{\left|s\right|} + \frac{\left|t^1\right|}{\left|t\right|} + \frac{\left|s^1\right| - T_{s,t}}{\left|s^1\right|} \right)$$

[13] proposes a variant of Jaro coefficient by using p, the length of the longer prefix common to both strings :

$$Jaro - Winkler(s,t) = Jaro(s,t) + \frac{p}{10}(1 - Jaro(s,t))$$

Presented coefficients calculate lexical similarity or distances iteratively and consider strings as blocks. There are also hybrid approaches calculating similarities recursively, by analyzing sub-strings of initial strings. Thus, Monge-Elkan coefficient calculates lexical similarity between $s^1 = s_1^1 \ldots s_k^1$ and $t^1 = t_1^1 \ldots t_l^1$ by performing two steps. First, the two strings are divided into sub-strings; then the similarity is given by:

$$Monge - Elkan(s,t) = \frac{1}{k}\sum_{i=1}^{k} \max_{j=1}^{l} sim(s_i, t_j)$$

where $sim(s_i, t_j)$ are given by some similarity function, for instance one of those previously presented. Such a function is called a *level 2 function*.

3.2 An iterative approach to learn conceptual graphs

Conceptual graphs are learned from the set of lexical pattern instances extracted according to section 2. An iterative solution is proposed, performing a number of steps, each of them adding a new layer to the graphs.

(1) The first step identifies verb classes which represent the set of verb relations generated by the same verb, see Table 1.

Table 1.Extract from diriger (to direct) class

diriger vers (direct towards)
diriger vers lieu (direct towards place)
véhicule diriger vers (vehicle direct towards)
automobile diriger vers esplanade (car direct towards esplanade)

For each verb class, instances of patterns "*Verb"* and "*Verb, Preposition*, are added to the set of roots. We argue that for verbs accepting prepositions, each "*verb, preposition"* pattern accepts specific arguments and for this reason conceptual graphs are created for each instance of those patterns. This step creates a number of conceptual graphs having one level, which is to say the root (see Figure 1).

(2) For each root, its arguments are identified: terms that are subjects and objects. As each relation accepts many terms as subject or object, lists of arguments are obtained. This step is adding a second layer to each conceptual graph.

(3) We observe that, for a given verb, arguments can have different levels of granularity, as we can see in table 2:

Table 2.Granularity of arguments

partie (side)
partie gauche (left side)
partie droite (right side)
rétroviseur (rear view mirror)
rétroviseur extérieur (external rear view mirror)

Hence, a new layer can be added to each conceptual graph by clustering those arguments.

A cluster is a group of similar terms, having a central term c called centroid and its k nearest neighbors. Based on the heuristic that the greater number of words in a word regrouping there are, the more specific his meaning is, an algorithm is proposed to cluster arguments of verb relations. The clustering algorithm is written as follows:

(1) for each list of arguments, create the list L of centroids, composed of all one-word arguments;

(2) for each centroid c, calculate the lexical similarity with other terms of the list by using Monge-Elkan coefficient;

(3) add to cluster c terms having a similarity value greater than a given threshold. An expert intervention allows us to chose the value of this trehsolds.

At that stage, Monge-Elkan function is used because it carries out recursive comparisons between sub-strings. Consequently, it has the capacity to agglomerate around a word (as centroids of clusters are one-word terms, which is to say words), terms derived from this word.

We chose one-word terms as centroids as they have the most general meaning, and, by consequence, will be able to attract into a cluster terms that are similar from a lexical point of view and that have more specific meanings. Figures 1 and 2 show the iterative construction of conceptual graphs. We can see one-level conceptual graphs learned from *diriger* (to direct) class and two-level conceptual graphs learned from *circuler* (to circulate) class.

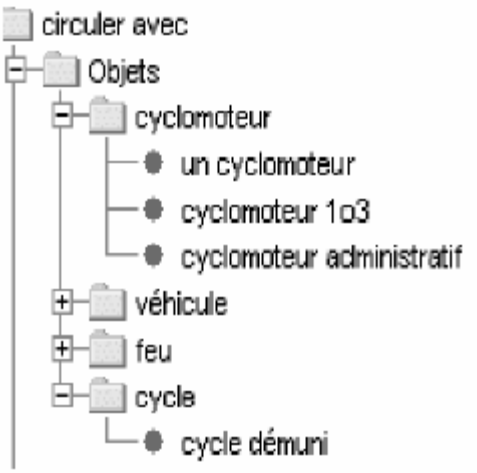

Figure 1. Conceptual graph modeling *circuler avec* (circulate with)

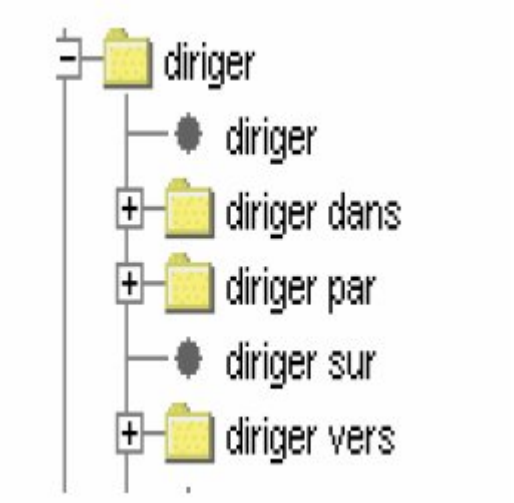

Figure 2. Conceptual graph modeling *diriger* (direct to)

4. TERM TO CONCEPT MAPPING USING THE ONTOLOGY

At this stage, arguments of verb relations can be assigned to concepts of the domain by using the previous conceptual graphs and a domain-ontology. We make the assumption that, for a given conceptual graph, the verb R representing its root node is already modeled by the ontology. If this is the case, let r be the corresponding relation and $Range_r, Domain_r$ concepts of the ontology connected by r. Those concepts and theirs descendants will be used to label arguments of the verbs. As arguments are connected to verb by links corresponding to the syntactic function, $Domain_r$ will be used to label subject arguments, while $Range_r$ will be used to label object arguments. Assignation of terms to concepts is performed by one of labelling strategies described bellow.

A first strategy ignores the hierarchical organization of arguments. Thus, similarities between each argument and terms naming concepts of the ontology are calculated using one of presented similarity measures. The argument is assigned to the concept maximizing this similarity, if the value of this similarity is greater than a pre-defined threshold. If similarity values are below the threshold, the term will be labelled as inconnu (*unknown)*. This is a non-oriented strategy because all arguments are considered at the same level.

Further on, we present two strategies (the second and the third) which take into account the hierarchical structure of arguments. Therefore, each cluster of arguments is considered as a hierarchy having on its first level the centroid and on its second level terms that are specializations of centroid.

The second strategy we propose is a top-down strategy. In the first phase, it identifies concepts of ontology which label the centroid of the cluster. If the centroid of a cluster is labeled as unknown, the same label is assigned to each term of the cluster. If the centroid of a cluster is labeled by a concept c of ontology, labels for other terms of the cluster are searched only in the set of sub-concepts of c. In this way, the top-down labelling strategy reduces the search space.

A third strategy is based on a bottom-up approach. For each cluster, the similarities between its terms and the concepts of ontology are calculated by using one of presented coefficients. If values of similarities are higher than a threshold, the concept labels the term. If this is not the case, the term will be labeled as inconnu (*unknown)*. Based on the assignments of each term of cluster to ontology concepts, the similarity between the centroid and a concept of ontology is given by:

$sim(Centroid, c) = \frac{1}{k}\sum_{i=1}^{k} sim(t_i, c)$, where t_i is a term of the cluster, c is a concept of ontology, $sim(t_i, c)$ is the similarity between t_i and c, and k is the number of terms labeled by c.

Those three labelling strategies are used in a first experimentation in the field of accidentology which is described in the next section.

5. EXPERIMENTATION IN ACCIDENTOLOGY AND FIRST RESULTS

Results of our approach can be affected by different parameters: the corpus we use, that is to say its size and its nature (a domain specific corpus or a general one) and the ontology. Our first experimentation was performed in accidentology and aimed to points out how different ontologies affect the outcome.

For this experimentation, we used a corpus and we aimed to assign terms extracted from this corpus to concepts of two different ontologies. Here after we describe those resources.

The corpus we used is composed of about 250 accident reports of accidents which occurred in and around Lille region (130 KO, 205 000 words). Accident reports are documents created by the police describing road accidents. They are written by policemen, according to declarations of people involved in accident and testimonies of witnesses.

A first case study was done by using an ontology created from accident reports O_1, see [4]. The ontology was created with Terminae, see [3], and it is expressed in OWL, see [6] and [21]. It models the domain of accidentology as it appears through documents created by the police. In this case study, the ontology and the corpus are created by the same community.

Our second case study was done by using an ontology O_2, created from accident scenarios, see [7]. Accident scenarios are documents created by researchers in road safety which describe prototypes of road accidents. The ontology was created with

Protégé see [16] and it is expressed in OWL. This ontology models the domain of accidentology as it appears through documents created by road safety researchers. In this second case study, the corpus and the ontology are created from two different communities.

Each ontology models concepts (see figure 3) and roles.

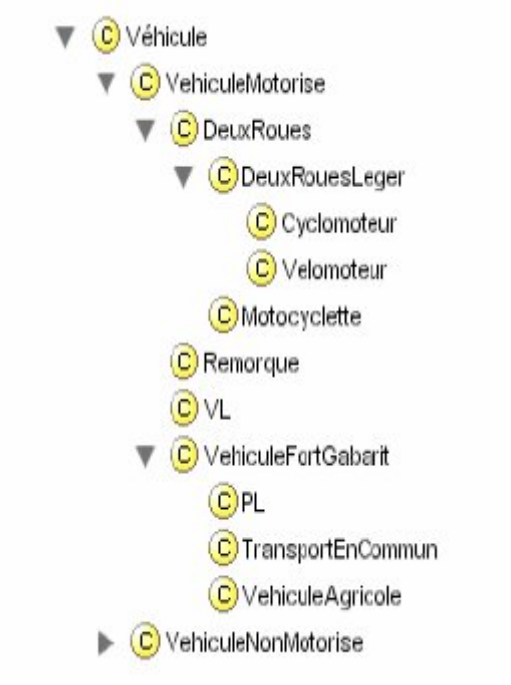

Figure 3. The concept Véhicule (Vehicle)

Roles are designated by domain specific verbs, see fig. 4.

As the community of road safety researchers is smaller, the number of entities of O_2 is less important, see table 3.

Table 3. O_1 and O_2 : number of entities

	Concepts	Roles
O_1	450	320
O_2	130	70

The analysis of results is done by using a new measure which we defined. This measure is called assignation degree, and it is given by:

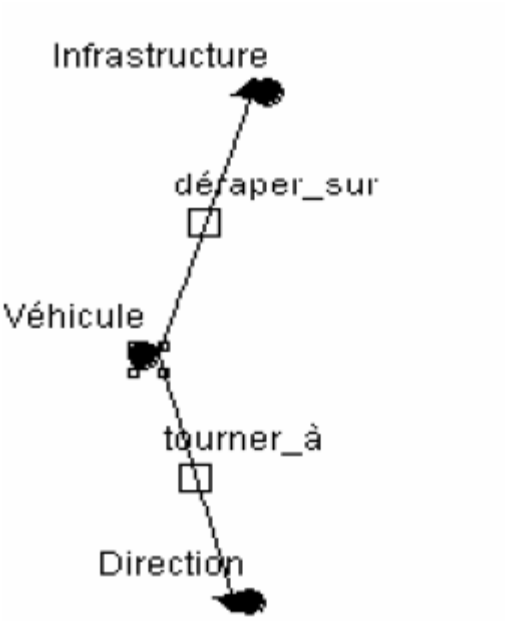

Figure 4. Roles of concept Véhicule (Vehicule)

$D(Corpus, O) = \frac{T_a}{T_{total}} * \frac{C_a}{C_{total}}$, where T_a is the number of arguments of verb relations assigned to concepts of the ontology; T_{total} is the number of terms extracted from corpus (arguments of verb relations); C_a is the number of concepts to whom arguments of verb relation are assigned, and C_{total} is the number of concepts of the ontology.

This definition is based on relative measures which enables us to compare results obtained by using different corpus and ontologies.

Values of assignation degree rank from 0 (all terms extracted from corpus are labelled as unknown) to 1 (each term extracted is assigned to a concept and each concept of the ontology labels at least one term).

For each case study, terms are assigned to concepts by using a labelling strategy. Results obtained are presented below:

Table 4. Assignation degree : non oriented strategy

Case study	Corpus	Ontology	Assignation degree
1	accident reports	O_1	70,5
2	accident reports	O_2	30,5

Table 5. Assignation degree : top-down strategy

Case study	Corpus	Ontology	Assignation degree
1	accident reports	O_1	68,5
2	accident reports	O_2	25,5

Table 6. Assignation degree : bottom -up strategy

Case study	Corpus	Ontology	Assignation degree
1	accident reports	O_1	68,5
2	accident reports	O_2	25,5

Result analysis is two-fold: for each case study, we compare the results provided by each labelling strategy; for the same case study (which is to say the same ontology), we compare the results provided by each labelling strategy.

As tables above show, the assignation degree has more important values if the corpus and the ontology share the same community. This could be explained by the similarity between the linguistic level of the ontology (terms designing its entities) and the corpus. By using a corpus and an ontology belonging to different communities, the assignation degree decrease drastically. In order to overcome this problem, we can use lexical resources such as WordNet, see [14], allowing us to take into account synonymy between terms when estimating theirs similarity.

Among the labelling strategies, the bottom-up one shows lows values of assignation degrees in both case studies. This is because the most clusters we have obtained have less than 10 words, and this strategy fails in case of small sized clusters.

The non oriented strategy and the top-down strategy provide similar values of assignation degree. Nevertheless, the top-down strategy performs faster, as it reduces the search space.

6. RELATED WORK

Approaches proposed in different application fields, such as ontology learning or word-sense disambiguation are at the origin of this work.

Among them, [10] propose Asium, a machine learning system which acquires subcategorization frames of verbs based on syntactic input. Asium hierarchically clusters nouns based on the verbs that they are syntactically related with and vice versa.

The work of [24] concerns the identification meaning of unknown verbs using the context of occurrence of the verb. The system Camille uses WordNet, see [14] as background knowledge and generates assumptions concerning the meaning of verbs. The assumptions are formulated according to linguistic criteria's.

[13] use a principle from information theory to model selectional preferences for verbs. Several classes may be appropriate for modeling selectional preferences.

[20] propose RelExt, a system which is capable of automatically identifying highly relevant triples (pairs of concepts connected by a relation). RelExt extracts relevant terms and verbs from a given text collection and it estimates relations between them through a combination of linguistic and statistical processing. Extracted triples can be integrated in an already existing ontology.

[18] propose a system having a multi-layered architecture aiming to extract information from genetic interaction data. The system uses verb patterns modelled as conceptual sub-graphs to characterize unknown terms in sentences. The goal is to enrich an existing ontology by integrating discovered concepts.

Our approach is based on the previous work presented in [18], whose major drawback is the impossibility to assign terms composed of many words (multi-words terms) to concepts of ontology. In order to overcome this limitation, our approach takes into account arguments of verb relations which have different levels of granularity. Therefore, we represent verb relations by conceptual graphs having three levels: the verb (first level), one-word arguments (second level) and multi-words arguments (the third level).

7. CONCLUSION AND FUTURE WORK

We have presented a non supervised approach developed to automatically assign terms of a corpus to concepts of ontology. This approach is using jointly verb relations and a domain ontology. Results provided could be used to semantically annotate or index documents.

A first experimentation in the accidentology domain was done in order to point out how different ontologies affect the outcome. In order to evaluate the results of this evaluation we have defined a new measure, called assignation degree. This evaluation shows that the approach provide better results if the corpus and the ontology belong to the same community.

If they belong to different communities, values of assignation degree decrease. This experimentation shows that our approach is sensitive to lexical level (changing the vocabulary, by passing from a community to another, affects values of assignation degree).

As a future work, new evaluation scenarios have to be proposed in order to study how other factors (namely the corpus: its size and its nature) affect the results.

Another perspective concerns the exploitation of lexical resources such as WordNet, in order to take into account the synonymy between terms. Namely, this will allow us to overcome the problem of lexical variation between different communities.

As a continuation of this work, a feedback could be added in order to enrich the domain ontology by integrating new concepts **.**

8. REFERENCES

[1] Alfonseca, E., Manandhar, S.: Improving an ontology refinement method with hyponymy patterns. *In proceedings of the Third International Conference on Language Resources and Evaluation*, 2001.

[2] Aussenac-Gilles, N., Seguela, P.: *Les relations sémantiques: du linguistique au formel*. Cahiers de grammaire 25 (175), 2000.

[3] Biébow, B., Szulman, S.: A linguistic-based tool for the building of a domain ontology. *In proceedings of the International Conference on Knowledge Engineering and Knowledge Management*, 1999.

[4] Ceausu, V., Desprè, S.: Towards a text mining driven approach for terminology construction. *In proceedings of the 7th International conference on Terminology and Knowledge Engineering*, TKE 2005, 2005.

[5] Cohen, W., Ravikumar, P., Fienberg, S.: A comparison of string distance metrics for name-matching tasks. *In proceedings of IJCAI-2003, Workshop on Information Integration on the Web pages*, 2003.

[6] Dean, G. Schreiber, P. Patel-Schneider, P. Hayes, and I. Horrocks. *Owl web ontology language reference.* Technical report, W3C Proposed Recommendation, 2004.

[7] Després, S. *Contribution à la conception de méthodes et d'outils pour la gestion des connaissances*. Habilitation à diriger des recherches, Université René Descartes, 2002.

[8] Euzenat, J., Valtchev, P.: An integrative proximity measure for ontology alignment. *In proceedings of ISWC-2003, Workshop on Semantic Information Integration*, 2003.

[9] Faatz, A., Steinmetz, R.: Ontology enrichment with texts from the www. *In proceedings of the SemanticWeb Mining 2nd Workshop at ECML/PKDD*, 2002.

[10] Faure, D., Nedellec, C.: Asium, learning subcategorization frames and restrictions of selection. *In proceedings of the 10th European Conference On Machine Learning, Workshop on text mining*, Chemnitz, Germany, 1998.

[11] Gagliardi, H., Haemmerl, O., Pernelle, N., Sas, F.: An automatic ontology-based approach to enrich tables semantically. *In proceedings of the first International Workshop on Context and Ontologies: Theory, Practice and Applications*, 2005.

[12] Hearst, M.: Automatic acquisition of hyponyms from large text corpora. *In proceedings of the 14th International Conference on Computational Linguistics*, 1992.

[13] Li, H., Abe, N.: G*eneralizing case frames using a thesaurus and the MDL principle*. Computational Linguistics 24, 217–244, 1998.

[14] Miller, G.: Wordnet: A lexical database for english. CACM 38, 39–41, 1995.

[15] Monge, A., Elkan, C.: The field-matching problem: algorithm and applications. *In proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 1996.

[16] Noy, N., Fergerson, R. W. and Musen, M. A.. *The knowledge model of Protégé-2000 : Combining interoperability and _exibility*. In Proceedings of the International Conference on Knowledge Engineering and Knowledge Management, 2000.

[17] Parekh, V., Jack, P.G., Finin., T.: Mining domain specific texts and glossaries to evaluate and enrich domain ontologies. *In Proceedings of the International Conference on Information and Knowledge Engineering*, 2004.

[18] Roux, C., Prouxet, D., Rechenmann, F., Julliard, L.: An ontology enrichment method for a pragmatic information extraction system gathering data on genetic interactions. In *Proceedings of Ontology Learning Workshop at ECAI*, 2000.

[19] Schmid, H.: Probabilistic part-of-speech tagging using decision trees. *In Proceedings of the International Conference on New Methods in Language Processing*, 1994.

[20] Schutz, A., Buitelaar, P.: Relext: A tool for relation extraction from text in ontology extension. *In Proceedings of the International Semantic Web Conference*, 593–606. 2005.

[21] Szulman, S., Biébow, B.: Owl et Terminae. *In actes de la 14ème Journée Francophone d'Ingénierie des Connaissances*, 2004.

[22] Valarakos, A., Paliouras, G., Karkaletsis, V., Vouros, G.: A name matching algorithm for supporting ontology enrichment. *In Proceedings of the 3rd Hellenic Conference on Artificial Intelligence*, 2004.

[23] Ville-Ometz, F., Royaut, J., Zasadzinski, A.: Filtrage semi-automatique des variantes de termes dans un processus d'indexation contrle. *In actes du Colloque International sur la Fouille de Textes*, 2004.

[24] Wiemer-Hastings, P., Graesser, A., Wiemer-Hastings, K.: Inferring the meaning of verbs from context. *In Proceedings of the Twentieth Annual Conference of the Cognitive Science Society*, 1998.

[25] Xiaomeng, S.: *Semantic Enrichment for Ontology Mapping*. PhD thesis, Norwegian University of Science and Technology, 2004.

[26] Warin, M., Oxhammer, H., Volk, M.: Enriching an ontology with wordnet based on similarity measures. *In: MEANING-2005 Workshop, 2005*.

[27] Widdows, D.: Unsupervised methods for developing taxonomies by combining syntactic and statistical information. *In proceedings of Human Language Technology Conference, HTL-NAACL, 2003.*