Learning term to concept mapping: an ontology based approach

Valentina Ceausu[1] and Sylvie Desprès[1]

Paris 5 University,
Paris, 75006,
France,
ceausu@math-info.univ-paris5.fr
sd@math-info.univ-paris5.fr

Abstract. We propose in this paper an approach to learn term to concept mapping with the joint utilization of an existing ontology and verb relations. 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. This paper presents the approach proposed to learn the conceptual graphs from the corpus and the labeling strategies. A first experimentation in the field of accidentology was done and its results are also presented.

Keywords: concept learning, ontology, verb relation.

1 Introduction

The rapid evolution in the production of documents in natural language requires to define efficient automated approaches allowing to find 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: indexing collection of documents, estimating similarities between documents, annotating 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 fallows : let O be a domain ontology and C a collection of domain-specific texts. With this approach, our goal is to identify within C terms W representing linguistic expression of concepts of O ontology. Thus, we can label terms identified in the corpus by concepts of ontology.

Three steps have been proposed to carry out this labelling process: (1) in a first stage, verb relations are extracted from the corpus. Each verb relation represents a triple 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 those 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, that could be subject or object. (3) The last phase is based on the assumption that the field ontology models verbs of the field as relations holding between the concepts. If this is the case, labelling strategies are using 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 on 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 approach are presented in section 6. Conclusions and perspectives of this work will end the paper.

2 Extracting verb relations from corpus

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

A set of lexical patterns including, among other categories,a verb, is defined. A pattern recognition algorithm, described in [1], is using part-of-speech information to identify associations of words matching the patterns of this set. The algorithm takes as input the corpus tagged by TreeTagger(see [2]) 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, as we can see bellow :

[1] *Verb, Preposition : diriger vers (direct to);*

Verb, Preposition, Noun : diriger vers place (direct to square);

Obtained word regroupings can be: a verb relation, highlighting relations of the field, 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).*

[1] Examples of this paper are translated in English, although they are extracted from a French corpus experimentation

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 theirs grammatical function, subject or object. We use the term *conceptual graph* as it was introduced by [3]. 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, a chain of treatments are performed that are based on lexical similarity measures presented bellow.

3.1 Lexical similarities

A similarity measure associates a real number R to a pair of strings $S1, S2$. Important values of R indicate a significant similarity of strings $S1, S2$. Many approaches are proposed to calculate similarities between strings. A number of them are presented in [4]. For this work, several similarity measures - Jaccard, Jaro, Jaro-Winkler, Monge and Elkan are - implemented. Jaccard coefficient considers a string composed of several sub-strings and calculates the similarity between two strings S and T as :
$Jaccard(S,T) = \frac{|S \bigcap T|}{|S \bigcup T|}$
This measure is given by the number of sub-strings common to S and T compared to the number of all sub-strings of T and S. 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 similarity by taking into account the number and the position of characters shared by S and T. Let $a = a_i..a_k$ and $b = b_1..b_l$ be two strings.A character $a_i \in s$ is considered common to both strings if there is a $b_j \in t$ such as: $a_i = b_j$ and $i - H \leq j \leq i + H$, where $H = \frac{min(|S|,|T|)}{2}$. Let $s^1 = a_1^1..a_k^1$ characters of s common to t and $t^1 = b_1^1..b_l^1$ characters of t common to s. We define a transposition between s^1 and b^1 as an index i such as: $a_i^1 \neq b_i^1$. If T_{s^1,t^1} is the number of transpositions from s^1 to t^1 the Jaro coefficient calculates the similarity between s and t as follows:

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

[5] proposes a version of this coefficient by using P , the length of the longer prefix common to both strings. Let $P^1 = max(P, 4)$, then Jaro-Winkler is written:

$$Jaro - Winckler(s,t) = Jaro(s,t) + \frac{P^1}{10}(1 - Jaro(s,t))$$

Presented coefficients calculate similarity between strings iteratively and consider strings as blocks. There are also hybrid approaches calculating similarities recursively, by analyzing sub-strings of initial strings. Thus, Monge-Elkan uses two steps to calculate similarity between $s^1 = a^1_1..a^1_k$ and $t^1 = b^1_1..b^1_l$: the two strings are divided into sub-strings then the similarity is given by:

$$sim(s,t) = \frac{1}{k}\sum_{i=1}^{k} max_{j=1}^{L}(sim^1(a_j,b_j))$$

where values of $sim^1(a_j,b_j)$ are given by some similarity function, called basic function, for example one of those previously presented. Such a function is called a *level 2 function.* For this work, Monge-Elkan is implemented by using the coefficients of Jaccard, Jaro and Jaro-Winkler as a basic function.
Statistic measures will be used in different phases of our approach.

3.2 An iterative approach to learn conceptual graphs

Conceptual graphs are learned from the set of word regroupings 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, that represent the set of verb relations generated by the same verb (see Table 1).

Table 1. Extracts from *diriger (direct to)* class

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

For each verb class, instances of *Verb* and *Verb, Preposition* patterns are added to the set of roots. We argue that for verbs accepting prepositions, each *verb, preposition* structure is specific and for this reason we create conceptual graphs for any of those structures. This steps create a number of conceptual graphs having one level, which is to say the root (as the Figure 1 shows).
(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 :

Partie *(side);*
partie *gauche (left side)*
partie *droite (right side)*
rétroviseur *(rear view mirror);*
rétroviseur *extrieur (external rear view mirror)*

***rétroviseur** intérieur (internal 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 observation 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 the single-word arguments;
(2) For each centroid C, calculate the similarity with other terms of the list by using Monge-Elkan function.
(3) Add to cluster C terms having a similarity value greater than a given threshold.

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 terms derived from this word. We chose single words 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. Figure (1) shows 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.

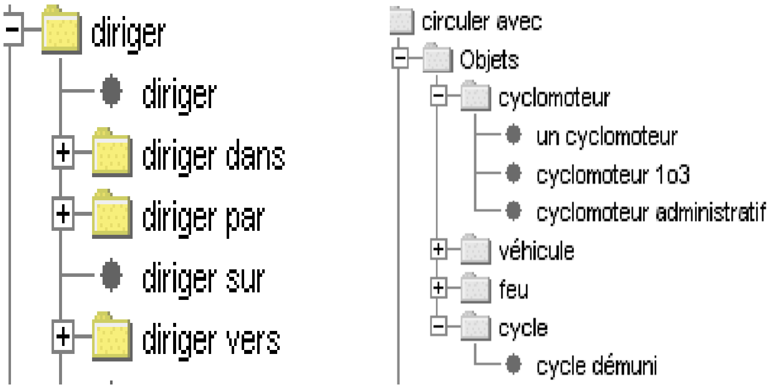

Fig. 1. Iterative learning of conceptual graphs

4 Term to concept mapping using the ontology

At this stage, arguments of verb relations can be assigned to concepts of the field by using those conceptual graphs and a domain-ontology. We make the assumption that, for a given conceptual graph, the verb R representing its top node is already modelled 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 the 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 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 similarity is greater than a pre-defined threshold. If the similarities between the term and the concepts of ontology are below the threshold, the term will be labelled as *unknown.* This is a non-oriented strategy because all the arguments are considered at the same level.

Further on, we present two strategies which take into account the hierarchical structure of arguments. Therefore, each argument cluster 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 *unknown.* Based on the assignments of each term of cluster to ontology concepts, similarity between the centroid and a concept of ontology is given by:

$$sim(Cen, C) = \tfrac{1}{k} \sum_{i=1}^{k} sim(t_i, C)$$

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

A first experimentation of this approach was done in the field of road accidents. A corpus composed of 250 accident reports of road accidents which occurred in and around the Lille region is used for this experimentation.
We used an ontology of road accidents which was also created from accident reports. The ontology was created with Terminae (see [6]), and it is expressed in OWL (see [7]). This ontology is composed of about 450 concepts describing road accidents. Concepts are connected by roles. The ontology contains about 300 roles expressed by verbs.
Figure 2 shows the concept *Véhicule (vehicle)* as it is implemented in this ontology and relations connecting *Véhicule* with other concepts of the ontology.

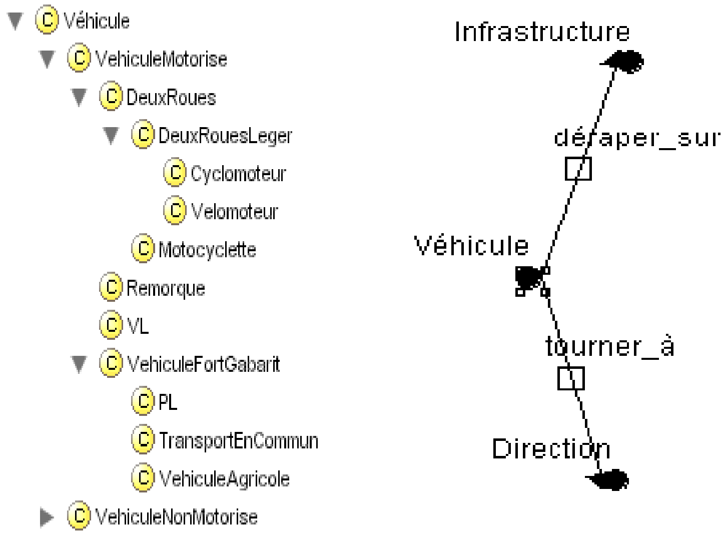

Fig. 2. Concept *Véhicule (vehicle)* and its roles in the ontology of accidentology

The analysis of the results is two-fold: for the same similarity coefficient, we compare the results provided by each labelling strategy; for the same labelling strategy, results provided by each coefficient are then compared.
For instance, arguments that are objects of the *circuler avec (circulate with)* relation are labelled. Results of the not-directed labelling strategy, using the Jaro coefficient and having a value of 0.85, as a threshold, are presented in Table 2. The top-down approach assigns the following labels to centroids (see Table 3): Table 4 presents the results in the second phase of the top-down strategy.

We observe that *véhicule blanc* is labelled as *inconnu (unknown)*. Consequently, the term can be ignored. The labelling of *véhicule lourd* is less appropriate. This strategy is faster than the previous one, because the concepts which can label the terms are sub-concepts of concept assigned to centroid.

Table 2. Non-oriented labelling, Jaro coefficient.

Term	Concept
cyclomoteur	cyclomoteur
véhicule blanc	Véhicule
véhicule lourd	véhicule lourd
véhicule	véhicule

Table 3. Top-down labelling, labels assigned to centroids

Centroid	Concept
véhicule	Véhicule Léger
cyclomoteur	cyclomoteur
cycle	Cycle
feu	Inconnu

The bottom-up approach initially assigns labels to terms of the cluster; then it uses those labels to assign a concept of the ontology to the centroid of the cluster.

Labels assigned to centroids are shown in Table 5. This strategy allows us to eliminate the centroid *feu (fire)*, which is labelled as *inconnu (unknown)*. On the downside, clusters having a small number of terms are penalized with this strategy. Centroids of clusters containing a small number of terms are assigned to concepts of an ontology with a low coefficient, or are labelled as *unknown*. For Jaro-Winkler coefficient, results of the three strategies are similar to results obtained with Jaro coefficient. This similarity results from the fact that Jaro-Winkler represents just a variation of Jaro measure. For Jaccard coefficient, the bottom-up strategy shows a failure as it assigns the term *véhicule* to the concept *véhicule de service*. As Jaccard coefficient is a measure based on the number of common characters of the two strings only, it assigns an important number of the terms of *véhicule* cluster to the concept *véhicule de service*. As a result, the centroid is labelled by the same concept. Independent of the coefficient that in used, the top-down strategy performs faster.
For the same couple *term, concept*, values of the Jaccard coefficient are slightly lower than values of Jaro and Jaro-Winkler. To assign labels by using the Jaccard coefficient, the selected threshold will therefore need to be lower than the thresholds used for Jaro and Jaro-Winkler coefficients.

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, [8] propose Asium, a machine learning system which acquires subcategorization frames of verbs based on syntactic input. Asium system hier-

Table 4. Top-down labelling, labels assigned to terms of clusters

Term	Concept
cyclomoteur	cyclomoteur
véhicule blanc	Inconnu
véhicule lourd	Véhicule Léger

Table 5. Bottom-up approach, labels assigned to centroids of clusters

Centroid	Concept
cyclomoteur	cyclomoteur
véhicule	Véhicule Léger
feu	Inconnu
Cycle	Cycle

archically clusters nouns based on the verbs that they are syntactically related with and vice versa.

The work of [9] concerns the identification of significance of the unknown verbs using the context of occurrence of the verb. The system Camille uses WordNet, (see [10]) as background knowledge and generates assumptions concerning the meaning of verbs. The assumptions are formulated according to linguistic criteria's.

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

[12] 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.

[3] 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.

7 Conclusion and future work

We have presented an approach allowing us to assign terms of a corpus to concepts of an ontology. This approach is using jointly verb relations and a domain ontology. Different measures which estimate similarity between strings have been implemented and used in the various phases of our approach.

A first experimentation in the field of road accidents shows that Jaro and Jaro-Winkler coefficients provide better similarities estimation than the Jaccard coefficient. Among the labelling strategies, the top-down strategy performs faster and generates better assignments of the terms to concepts of ontology. Those

are only preliminaries conclusions and further case studies are needed.
The evaluation of our approach is another important issue to address. An evaluation strategy must be defined, and experimentations should be performed in order to see how the size of the corpus, different domain-specific text collections or different kinds of ontologies affect the outcome.
As a further direction, a feedback could be added in order to enrich the domain ontology by integrating some of the arguments of verb relations.

References

1. Ceausu, V., Desprès, S.: Towards a text mining driven approach for terminology construction. In: 7th International conference on Terminology and Knowledge Engineering. (2005)
2. Schmid, H.: Probabilistic part-of-speech tagging using decision trees. In: International Conference on New Methods in Language Processing. (1994)
3. Roux, C., Prouxet, D., Rechenmann, F., Julliard, L.: An ontology enrichment method for a pragmatic information extraction system gathering data on genetic interactions. In: Ontology Learning Workshop at ECAI. (2000)
4. Cohen, W., Ravikumar, P., Fienberg, S.: A comparison of string distance metrics for name-matching tasks. In: IJCAI-2003,Workshop on Information Integration on the Web pages. (2003)
5. Monge, A., Elkan, C.: The field-matching problem: algorithm and applications. In: Second International Conference on Knowledge Discovery and Data Mining. (1996)
6. Biébow, B., Szulman, S.: A linguistic-based tool for the building of a domain ontology. In: International Conference on Knowledge Engineering and Knowledge Management. (1999)
7. Szulman, S., Biébow, B.: Owl et terminae. In: 14-me Journe Francophone d' Ingnierie des Connaissances. (2004)
8. Faure, D., Nedellec, C.: Asium, learning subcategorization frames and restrictions of selection. In: 10th European Conference On Machine Learning, Workshop on text mining, Chemnitz, Germany (1998)
9. Wiemer-Hastings, P., Graesser, A., Wiemer-Hastings, K.: Inferring the meaning of verbs from context. In: Twentieth Annual Conference of the Cognitive Science Society. (1998)
10. Miller, G.: Wordnet: A lexical database for english. CACM **38** (1995) 39–41
11. Li, H., Abe, N.: Generalizing case frames using a thesaurus and the MDL principle. Computational Linguistics **24** (1998) 217–244
12. Schutz, A., Buitelaar, P.: Relext: A tool for relation extraction from text in ontology extension. In: International Semantic Web Conference. (2005) 593–606
13. Alfonseca, E., Manandhar, S.: Improving an ontology refinement method with hyponymy patterns. In: Third International Conference on Language Resources and Evaluation. (2001)
14. Aussenac-Gilles, N., Seguela, P.: Les relations smantiques: du linguistique au formel. Cahiers de grammaire **25** (2000) 175
15. Euzenat, J., Valtchev, P.: An integrative proximity measure for ontology alignment. In: ISWC-2003 Workshop on Semantic Information Integration. (2003)
16. Faatz, A., Steinmetz, R.: Ontology enrichment with texts from the www. In: SemanticWeb Mining 2nd Workshop at ECML/PKDD. (2002)

17. Gagliardi, H., Haemmerl, O., Pernelle, N., Sas, F.: An automatic ontology-based approach to enrich tables semantically. In: The first International Workshop on Context and Ontologies : Theory, Practice and Applications. (2005)
18. Hearst, M.: Automatic acquisition of hyponyms from large text corpora. In: 14th International Conference on Computational Linguistics. (1992)
19. Monge, A., Elkan, C.: An efficient domain-independent algorithm for detecting approximately duplicate database records. In: Workshop on data mining and knowledge discovery, SIGMOD. (1997)
20. Parekh, V., Jack, P.G., Finin., T.: Mining domain specific texts and glossries to evaluate and enrich domain ontologies. In: International Conference on Information and Knowledge Engineering. (2004)
21. Valarakos, A., Paliouras, G., Karkaletsis, V., Vouros, G.: A name matching algorithm for supporting ontology enrichment. In: 3rd Hellenic Conference on Artificial Intelligence. (2004)
22. Ville-Ometz, F., Royaut, J., Zasadzinski, A.: Filtrage semi-automatique des variantes de termes dans un processus d'indexation contrle. In: Colloque International sur la Fouille de Textes. (2004)
23. Xiaomeng, S.: Semantic Enrichment for Ontology Mapping. PhD thesis, Norwegian University of Science and Technology (2004)
24. Warin, M., Oxhammer, H., Volk, M.: Enriching an ontology with wordnet based on similarity measures. In: MEANING-2005 Workshop. (2005)
25. Widdows, D.: Unsupervised methods for developing taxonomies by combining syntactic and statistical information. In: Human Language Technology Conference, HTL-NAACL. (2003)