How to Expand Dictionaries by Web-Mining Techniques

Nicolas Béchet
LIRMM, UMR 5506, CNRS,
Univ. Montpellier 2
France
bechet@lirmm.fr

Mathieu Roche
LIRMM, UMR 5506, CNRS,
Unv. Montpellier 2
France
mroche@lirmm.fr

Abstract

This paper deals with an approach of conceptual class enrichment based on the Web. In order to experiment our approach, we propose first to build conceptual classes. These ones are built using syntactic and semantic informations provided by a corpus. The concepts can be the input of a dictionary. Then, our web-mining approach dealing with a cognitive process simulates a human reasoning based on the enumeration principle. Experiments show the interest of our enrichment approach by adding new relevant terms into existing conceptual classes.

1 Introduction

Concepts have several definitions; one on the most general describes a concept 'as the mind representation of a thing or an item' (Desrosiers-Sabbath, 1984). Within a given domain such as ours, which deals with ontology building, semantic web, and computational linguistics, it seems quite appropriate to stick to the Aristotelian approach of a concept, and see it as a set of knowledge gathering of common semantic features. Features choice and gathering design are dependent upon criteria that we will try to explain hereafter.

This paper deals with the building of conceptual classes which can be defined as gathering of terms semantically close. First, we suggest building specific conceptual classes, by focusing on knowledge extracted from corpora. Conceptual classes are shaped through the study of syntactic dependencies between corpus terms (as described in section 2). Dependencies tackle relations such as Verb/Subject, Noun/Noun Phrase Complements, Verb/Object, Verb/Complements, and sometimes Sentence Head/Complements. In this paper, we focus on the Verb/Object dependency, because it is a good representative of a field. For instance, in computer science, the verb to load takes as objects, nouns of the conceptual class software (L'Homme, 1998). This feature also spreads to 'download' or 'upload' which have the same verbal root.

Corpora are rich ore in which mining for terminological information is fruitful. A terminology extraction of this kind is similar to a Harris-like distributional analysis (Harris, 1968) and literature displays an abundant set of works undergoing a distributional analysis to acquire terminological or ontological knowledge from textual data (e.g (Bourigault and Lame, 2002) for law, (Nazarenko et al., 2001; Weeds et al., 2005) for medicine).

After building conceptual classes (section 2), we describe an approach to expand concepts by using a Web search engine to discover new terms (section 3). In section 4, experiments conduced on real data enable to validate the proposed approach.

2 Conceptual Classes Building

2.1 Principle

A class can be defined in our approach as a gathering of terms having a common field. In this paper, we focus on objects of verbs judged to be semantically close regarding a measure. Thus, these objects are considered as instances of conceptual classes. The first step of building conceptual classes consists in extracting Verb/Object syntactic relations as explained in the following section.

2.2 Mining for Verb/Object relations

Our corpora are in French since our team is mostly devoted to French-based NLP applications. However, the following method is portable to any other language, provided that a quite reliable dependency parser is available. In our case, we use the SYGFRAN parser developed by (Chauché, 1984). As an example, in the French sentence "*Thierry Dusautoir brandissant le drapeau tricolore sur la pelouse de Cardiff après la victoire.*" (translation: 'Thierry Dusautoir brandishing the three colored flag on Cardiff lawn after the victory'), there is a syntactic relation verb-object: "*verb: brandir (to brandish), object: drapeau (flag)*", which is a good candidate for retrieval. The second step of the building process corresponds to the gathering of common objects related to semantically close verbs.

Semantic Closeness Assumption. The underlying linguistic hypothesis is the following: Verbs having a significant number of common objects are semantically close.
To measure closeness, the ASIUM score (Faure and Nedellec, 1999; Faure, 2000) is used as illustrated in the figure 1. This type of work is akin to distributional analysis approaches such as (Bourigault and Lame, 2002).

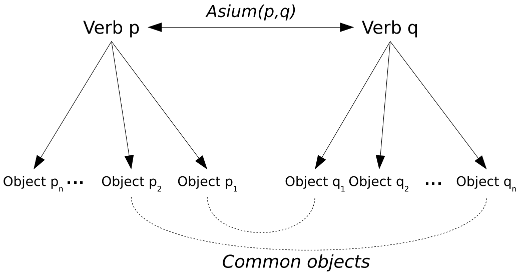

Figure 1: Common and complementary objects of verbs "to consume" and "to eat"

As explained in introduction, the measure considers two verbs as close if they have a significant number of mutual features (objects). Let p and q be verbs with their respective p_1,...,p_n and q_1,...,q_m objects. $NbOC_p(q_i)$ is the occurrence number of q_i objects from q also objects of p (common objects). $NbO(q_i)$ is the occurrence number of q_i objects of q verb. Then, the Asium measure is:

$$Asium(p,q) =$$

$$\frac{log_{Asium}(\sum NbOC_q(p_i))+log_{Asium}(\sum NbOC_p(q_i))}{log_{Asium}(\sum NbO(p_i))+log_{Asium}(\sum NbO(q_i))}$$

Where $log_{Asium}(x)$ equal to:

- for $x = 0$, $log_{Asium}(x) = 0$
- else $log_{Asium}(x) = log(x) + 1$

Therefore, conceptual classes instances are the common objects of close verbs, according to the ASIUM proximity measure.

The following section describes the acquisition of new terms starting with a list of terms/concepts obtained with the global process summarized in this section and detailed in (Béchet et al., 2008).

3 Expanding conceptual classes

3.1 Acquisition of candidate terms

The aim of this method is to provide new candidates of a given concept. Our approach is based on the enumeration on the Web of terms semantically close. For instance, with a query (string) "bicycle, car, and", we can find other vehicles. We propose to use the web to acquire new candidates. This kind of method uses information regarding the "popularity" on the web, it is independent of a corpus.

Our acquisition method is quite similar to the (Nakov and Hearst, 2008) approach. Authors propose to querying the Web by using the Google search engine in order to characterize the semantic relation between a pair of nouns. The Google's star operator is among others used in that aim. (Nakov and Hearst, 2008) refered the study of (Lin and Pantel, 2001) which used Web mining approach in order to discover inference rules missed by humans.

To apply our method, we initially consider the common objects of verbs semantically close. They are instances of reference concepts (e.g. vehicle). Let N concepts $C_{i \in \{1,N\}}$ and their respective instances $I_j(C_i)$. For each concept C_i, we submit to a search engine the following queries: "$I_{jA}(Ci)$, $I_{jB}(Ci)$, and" and "$I_{jA}(Ci)$, $I_{jB}(Ci)$, or" with j_A et $j_B \in \{1, ..., NbInstanceCi\}$ and $j_A \neq j_B$.

The search engine returns a set of results. We extract new candidate instances of a concept. For example, if we consider the query: "bicycle, car, and", one page returned by a search engine gives the following text:

Listen here for the Great Commuter Race (17/11/05) between bicycle, car and **bus***, as part of...*

Having identified the relevant features in the returned result (in bold on our example), we add the term "bus" in the initial concept "vehicle". We obtain new candidates for our concepts. This process can be repeated. In order to automatically determine the relevant candidates, these ones are filtered as shown in the following section.

3.2 Filtering of candidates

The quality of the extracted terms can be validated by an expert, or by using an automatic approach. This one can use the Web to determine if the extracted candidates (see section 3.1) are relevant. The principle is to consider a relevant term if this one is often present with the terms of the original conceptual class (kernel of words). Thus, we seek to validate a term "in context". From that point of view, our method is close to (Turney, 2001), which queries the Web via the AltaVista search engine to determine appropriate synonyms to a given term. As (Turney, 2001), we consider that the information about the number of pages returned by the queries can give a notion of relevance.

Thus, we submit to a search engine different strings (using quotes). Queries consist of the new candidate and both terms of the concept. Formally, our approach can be defined as follows. Let N concepts $C_{i \in \{1,N\}}$, their respective instances $I_j(C_i)$ and the new candidates for a concept C_i, $N_{i_k \in \{1, NbNI(C_i)\}}$. For each C_i, each new candidate N_{i_k} are queried to a Web search engine. In practice the three terms are alternately separated by a comma and the words "or" or "and".[1] For each query the search engine returns the number of results (i.e. number of web pages). Then, the sum of these results is calculated using all possible combinations between "or", "and", and the three words (words of the kernel plus candidate word to enrich). Below, an example with the kernel words "car", "bicycle" and the candidate "bus" to test is given (with Yahoo):

- "car, bicycle, and bus" : 71 pages returned
- "car, bicycle, or bus" : 268 pages returned
- "bicycle, bus, and car" : 208 pages returned
- and so forth

Global result: $71 + 268 + 208...$

The filtering of candidates consists in selecting the ***k*** **first candidates by class** (i.e. having the highest sum), they are added as new instances of the initial concept. We can reiterate the acquisition approach by including these new terms. Then, the acquisition/filtering process can be repeated several times.

We present in the next section conducted experiments in order to measure the quality of our approach.

[1] Note that the commas are automatically removed by the search engines.

4 Experiments

4.1 Evaluation protocol

We use a French corpus from Yahoo's site (http://fr.news.yahoo.com/) composed of 8,948 news (16.5 MB) from newspapers. Experiments are performed on 60,000 syntactic relations (Béchet et al., 2008; Béchet et al., 2009) to build original conceptual class. We have selected manually five concepts (see Figure 2). Instances of these concepts are the common objects of verbs defining the concept (see section 2.2).

Concepts	**Organisme /Administration** *(Civil Service)*	**Fonction** *(work)*	**Objets symboliques** *(symbols)*	**Sentiment** *(feeling)*	**Manifestation de protestation** *(protest)*
Instances	parquet *(prosecution)*	négociateur *(negotiator)*	drapeau *(flag)*	mécontentement *(discontent)*	protestation *(remontrance)*
	mairie *(city hall)*	cinéaste *(filmmaker)*	fleur (flower)	souhait *(wish)*	grincement *(grind)*
	gendarme *(policeman)*	écrivain *(writer)*	spectre *(specter)*	déception *(disappointment)*	indignation *(indignation)*
	préfecture *(prefecture)*	orateur *(public speaker)*		désaccord *(disagreement)*	émotion *(emotion)*
	pompier *(fireman)*			désir *(desire)*	remous *(swirl)*
	O.N.U. *(U.N.)*				tollé *(collective protest)*
					émoi *(commotion)*
					panique *(panic)*

Figure 2: The five selected concepts and their instances.

For our experiments, we use an API of the search engine *Yahoo!* in order to get new terms. We apply the following post-traitements for each new candidate term. They are initially lemmatized. Therefore, we only keep the *nouns*, after applying a PoS (Part of Speech) tagger, the TreeTagger (Schmid, 1995).

After these various post-treatments, we manually validate the new terms by three experts. We compute the precision of our approach to each expert. The average is calculated to define the quality of the terms. Precision is defined as follows.

$$\text{Precision} = \frac{\text{Number of relevant terms given by our system}}{\text{Number of terms given by our system}}$$

The next section presents the evaluation of our method.

4.2 Experimental results

Table 1 presents the results of term acquisition method (i.e for each acquisition step, we apply our approach to filter candidate terms). The table presents for each step, the obtained precision after expertise:

- **All candidates.** We calculate the precision before the filtering step.
- **Filtered candidates.** After applying the automatic filtering by selecting k terms per class, we calculate the obtained precision. Note that the automatic filtering (see section 3.2) reduces the number of proposed terms, and thus reduce the recall. [2]

We finally show in Table 1 the number of terms generated by the acquisition system.

	Precision		*Term number*
Steps #	*All terms*	*Filtered terms*	*(without filter)*
1	0,69	0,83	29
2	0,69	0,77	47
3	0,56	0,65	103

Table 1: Results obtained with k=4 (i.e. automatic selection of the k first ranked terms by the filtering approach).

These results show that a significant number of terms is generated (i.e. 103 words). For example, we obtain for the concept *feeling* by using initial terms given in figure 1 the following eight French terms (with two steps): "horreur (horror), satisfaction (satisfaction), déprime (depress), faiblesse (weakness), tristesse (sadness), désenchantement (disenchantment), folie (madness), fatalisme (fatalism)".

This approach is relevant to produce new relevant terms to enrich conceptual classes, in particular when we select the first terms ($k = 4$) returned by the filtering system. In a future work, we plan to test other values of the automatic filtering. The obtained precision in the first two steps is high (i.e. 0.69 to 0.83). The third step returns lower scores; noise is introduced because we are "far" of the initial kernel words.

[2] The recall is not calculated because in an unsupervised context this one is difficult to estimate.

5 Conclusion and Future Work

This paper has focused on an approach for conceptual enrichment classes based on the Web. We apply the ”enumeration” principle in order to find new terms using Web search engine. This approach has the advantage of being less dependent on the corpus. Note that the use of the Web requires validation of candidates. Thus, we propose an automatic filtering method to propose relevant term to add in the concept. In a future work, we plan to use other statistical web measures (e.g. Mutual Information, Dice measure, and so forth) to validate terms in an automatic way.

References

Béchet, N., M. Roche, and J. Chauché. 2008. How the ExpLSA approach impacts the document classification tasks. In *Proceedings of the International Conference on Digital Information Management, ICDIM'08*, pages 241–246, University of East London, London, United Kingdom.

Béchet, N., M. Roche, and J. Chauché. 2009. Towards the selection of induced syntactic relations. In *European Conference on Information Retrieval (ECIR), Poster*, pages 786–790.

Bourigault, D. and G. Lame. 2002. Analyse distributionnelle et structuration de terminologie. application à la construction d'une ontologie documentaire du droit. In *TAL*, pages 43–51.

Chauché, J. 1984. Un outil multidimensionnel de l'analyse du discours. In *Proceedings of COLING, Standford University, California*, pages 11–15.

Desrosiers-Sabbath, R. 1984. *Comment enseigner les concepts*. Presses de l'Université du Québec.

Faure, D. and C. Nedellec. 1999. Knowledge acquisition of predicate argument structures from technical texts using machine learning: The system ASIUM. In *Proceedings of the 11th European Workshop, Knowledge Acquisition, Modelling and Management, number 1937 in LNAI*, pages 329–334.

Faure, D. 2000. *Conception de méthode d'apprentissage symbolique et automatique pour l'acquisition de cadres de sous-catégorisation de verbes et de connaissances sémantiques à partir de textes : le système ASIUM*. Ph.D. thesis, Université Paris-Sud, 20 Décembre.

Harris, Z. 1968. *Mathematical Structures of Language*. John Wiley & Sons, New-York.

L'Homme, M. C. 1998. Le statut du verbe en langue de spécialité et sa description lexicographique. In *Cahiers de Lexicologie 73*, pages 61–84.

Lin, Dekang and Patrick Pantel. 2001. Discovery of inference rules for question answering. *Natural Language Engineering*, 7:343–360.

Nakov, Preslav and Marti A. Hearst. 2008. Solving relational similarity problems using the web as a corpus. In *ACL*, pages 452–460.

Nazarenko, A., P. Zweigenbaum, B. Habert, and J. Bouaud. 2001. Corpus-based extension of a terminological semantic lexicon. In *Recent Advances in Computational Terminology*, pages 327–351.

Schmid, H. 1995. Improvements in part-of-speech tagging with an application to german. In *Proceedings of the ACL SIGDAT-Workshop, Dublin*.

Turney, P.D. 2001. Mining the Web for synonyms: PMI–IR versus LSA on TOEFL. In *Proceedings of ECML'01, Lecture Notes in Computer Science*, pages 491–502.

Weeds, J., J. Dowdall, G. Schneider, B. Keller, and D. Weir. 2005. Weir using distributional similarity to organise biomedical terminology. In *In Proceedings of Terminology*, volume 11, pages 107–141.