How to Expand Dictionaries with Web-Mining Techniques

Nicolas Béchet
LIRMM, UMR 5506, CNRS,
Univ. Montpellier 2

bechet@lirmm.fr

Mathieu Roche
LIRMM, UMR 5506, CNRS,
Univ. Montpellier 2

mroche@lirmm.fr

Abstract

This paper presents an approach to enrich conceptual classes based on the Web. To test our approach, we first build conceptual classes using syntactic and semantic information provided by a corpus. The concepts can be the input of a dictionary. Our web-mining approach deals with a cognitive process which simulates human reasoning based on the enumeration principle. The experiments reveal the interest of our approach by adding new relevant terms to existing conceptual classes.

1 Introduction

Concepts have several definitions; one of the most general describes a concept 'as the mind's representation of a thing or an item' (Desrosiers-Sabbath, 1984). In a domain such as ours, i.e. ontology building, semantic webs, and computational linguistics, it seems appropriate to stick to the Aristotelian approach to a concept, and consider it as a set of knowledge (gathered information) on common semantic features. The choice of the features and how the knowledge is gathered depend on criteria we will explain below.

In this paper, we deal with the building of conceptual classes, which can be defined as gathering semantically close terms. First, we suggest building specific conceptual classes by focusing on knowledge extracted from corpora.

Conceptual classes are shaped by 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 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 extends to 'download' or 'upload', which have the same verbal root.

Corpora are rich sources of terminological information that can be mined. A terminology extraction of this kind is similar to a Harris-like distributional analysis (Harris, 1968) and many works in the literature have been the subject of 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 conducted on real data enable us to validate our approach.

2 Building Conceptual Classes

2.1 Principle

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

Proceedings of the 2nd Workshop on Cognitive Aspects of the Lexicon (CogALex 2010), pages 33–37,
Beijing, August 2010

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 can be used for any other language, provided a 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 verb-object syntactic relation: *"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.

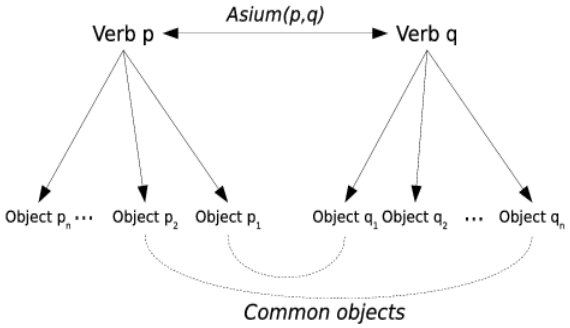

Figure 1: Common and complementary objects of the verbs "to consume" and "to eat"

Assumption of Semantic Closeness. The underlying linguistic hypothesis is the following: Verbs with a significant number of common objects are semantically close.

To measure closeness, the ASIUM score (Faure and Nedellec, 1999; Faure, 2000) is used (see figure 1). This type of work is similar to distributional analysis approaches such as that of (Bourigault and Lame, 2002).

As explained in the introduction, the measure considers two verbs to be close if they have a significant number of common 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 number of occurrences of q_i objects from q that are also objects of p (common objects). $NbO(q_i)$ is the number of occurrences of q_i objects of q verb. The Asium measure is then:

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

Where $log_{Asium}(x)$ is 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 approach is to provide new candidates for a given concept. It is based on enumeration on the Web of terms that are 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" of the web and is independent of a particular corpus.

Our method of acquisition is quite similar to that of (Nakov and Hearst, 2008). These authors propose to query the Web using the Google search engine to characterize the semantic relation between a pair of nouns. The Google star operator among others, is used to that end. (Nakov and Hearst, 2008) refer to the study of (Lin and Pantel, 2001) who used a Web mining approach to discover inference rules missed by humans.

To apply our method, we first consider the common objects of semantically close verbs, which 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: "$I_{jA}(C_i)$, $I_{jB}(C_i)$, and" and "$I_{jA}(C_i)$, $I_{jB}(C_i)$, or" with jA and $jB \in \{1, ..., NbInstanceC_i\}$ and $jA \neq jB$.

The search engine returns a set of results from which 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 result returned (in bold in our example), we add the term "bus" to the initial concept "vehicle". In this way, we obtain new candidates for our concepts. The process can be repeated. In order to automatically determine which candidates are relevant, the candidates 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 automatically by using the Web to check if the extracted candidates (see section 3.1) are relevant. The principle is to consider a relevant term if it is often present with the terms of the original conceptual class (kernel of words). Thus, our aim is to validate a term "in the context". From that point of view, our method is close to that of (Turney, 2001), which queries the Web via the AltaVista search engine to determine appropriate synonyms for a given term. Like (Turney, 2001), we consider that information concerning the number of pages returned by the queries can give an indication of the relevance of a term.

Thus, we submit to a search engine different strings (using citation marks). A query consists 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, $N_{ik \in \{1, NbNI(Ci)\}}$. For each C_i, each new candidate N_{ik} is sent as a query to a Web search engine. In practice the three terms are separated either by a comma or the word "or" or "and"[1]. For each query, the search engine returns a number of results (i.e. number of web pages). Then, the sum of these results is calculated using all possible combinations of "or", "and", or of the three words (words of the kernel plus candidate word to enrich it). Below is an example with the kernel words "car", "bicycle" and the candidate "bus" to test (using 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. with the highest sum), they are added as new instances of the initial concept. We can reiterate the acquisition approach by including these new terms. The acquisition/filtering process can be repeated several times.

In the next section, we present experiments conducted to evaluate the quality of our approach.

4 Experiments

4.1 Evaluation protocol

We used a French corpus from the Yahoo site (http://fr.news.yahoo.com/) composed of 8,948 news items (16.5 MB) from newspapers. Experiments were performed on 60,000 syntactic relations (Béchet *et al.*, 2008; Béchet *et al.*, 2009) to build original conceptual classes. We manually selected 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	Fonction	Objets symboliques	Sentiment	Manifestation de protestation
	(Civil Service)	*(work)*	*(symbols)*	*(feeling)*	*(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.

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

For our experiments, we use an API of the search engine Yahoo! to obtain new terms. We apply the following post-treatments 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 post-treatments, we manually validate the new terms using 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.

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

In the next section, we present the evaluation of our method.

4.2 Experimental results

Table 1 gives the results of the term acquisition method (i.e. for each acquisition step, we apply our approach to filter candidate terms). For each step, the table lists the degree of precision obtained 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 precision obtained. Note that the automatic filtering (see section 3.2) reduces the number of terms proposed, and thus reduces the recall[2].

	Precision		*Terms 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).

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

Finally Table 1 shows the number of terms generated by the acquisition system.

These results show that a significant number of terms can be generated (i.e. 103 words). For example, for the concept 'feeling', using the initial terms given in figure 1, we obtained the following eight French terms (in two steps): "horreur (horror), satisfaction (satisfaction), déprime (depression), faiblesse (weakness), tristesse (sadness), désenchantement (disenchantment), folie (madness), fatalisme (fatalism)".
This approach is appropriate 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 precision obtained in the first two steps was high (i.e. 0.69 to 0.83). The third step returned lower scores; noise was introduced because we were too "far" from the initial kernel words.

5 Conclusion and Future Work

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

References

Béchet, N., M. Roche, and J. Chauch´e. 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 *Proceed ngs 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 *Proceedings of Terminology*, volume 11, pages 107–141.