Corpus and Web: Two Allies in Building and Automatically Expanding Conceptual Classes

Nicolas Béchet, Jacques Chauché, Violaine Prince and Mathieu Roche
Équipe TAL, Laboratoire d'Informatique de Robotique et de Micro-électronique de Montpellier
UMR 5506, CNRS, Univ. Montpellier 2
34 392 Montpellier Cedex 5 - France
E-mail : {bechet, chauche, prince, mroche}@lirmm.fr

Keywords: terminology, conceptual classes, expansion, web validation

Received: March 30, 2010

In this paper, the approaches to building and expanding conceptual classes are presented. The classes are built with syntactic and semantic information provided by a corpus. Then, expansion is addressed by using the objects of syntactic relations found in the corpus. Relations between classes are thus designed. They are called induced relations. Then we use objects of induced syntactic relations (called complementary objects) to expand conceptual classes. We propose an automatic experimental protocol to measure the relevance of the provided concepts. The protocol helps alleviating the judgment effort of a human expert. The expansion method is evaluated and mixed in order to provide the most reliable technique in expanding conceptual classes.

Povzetek: V prispevku je opisan postopek izgradnje konceptualnih dreves s pomočjo spleta in korpusov.

1 Introduction

Several NLP (Natural Language Processing) applications use terminology. The latter can be defined as the study of technical terms of a field, as well as their signification. Two kinds of terminology studies can be proposed: one which is called ‘semasiologic’ and the other, ‘onomasiologic’. The first focuses on term signification to study sense. The second proposes to start from the conceptual level, and attaches terms as linguistic instantiations of concepts.

Concepts have born several definitions. One of the most general ones 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.

Starting from concepts needs to have, at start, an extensive representation of the terminology associated with each concept. Thus the onomasiologic approach better deals with restricted thematic fields (e.g. 'meteorology', 'tomato growth in agriculture', etc.). Concepts are first established and agreed upon, and terminology is associated with concepts. Afterwards, all types of processes could be undertaken with such a knowledge base. This approach outcomes are tied with the domain closure.

In an open, or yet incompletely browsed domain (such as Web pages might induce), onomasiology is less capable. Thus such cases are preferably investigated with semasiologic tools. The existing data are analysed and bring forth term which significations are otherwise arranged in order to create gatherings. Both concepts and terminology are incrementally enhanced, and shaping is a loop process with an important feedback. Very obviously, Semantic Web is better approached by the semasiologic method. However, such a method creates new problems as side effects. If onomasiology is better served by restricting the field, semasiology performs better when restricting the task. Tasks involve information retrieval (IR), text indexing, question answering, summarizing, translating, etc...Thus, the terminology built for a given task must not be used in other tasks without some care or partial rebuilding [Roche, 2005].

In this paper, we propose to build conceptual classes, expand them, and directly attach terminology under the framework of a semasiological process. The restrictions of semasiology are however alleviated by the fact that NLP techniques for classes building and term attachments are used on both domain corpora and cross-domains Web pages. Naturally, the most fitting task is IR, but to an extent, other tasks could be addressed by tuning the building and expansion process.

First, we uggest 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 the 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 or 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, we describe an approach to expanding the classes by using the corpus to discover new terms (in section 3). These terms are then ranked and proposed to an expert in a sorted list.

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. This type of work is akin to distributional analysis approaches such as [Bourigault et al. 2002].

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

3 Expanding conceptual classes

3.1 Principle

In order to expand conceptual classes, the main difficulty is to obtain new terms which can be instances of a conceptual class. The basic idea here is to use the corpus itself to acquire new instances with the same approach as in building classes (see 2.1). As it was said before, the process admits as instances of a class the common objects of close verbs. Thus expanding conceptual classes is a two steps procedure:

1) Retrieving **complementary objects (**to be explained hereafter)

2) Asserting the relevance of complementary object as a possible instance of a concept.

Both steps are introduced in the next sub-section.

3.2 Step 1: Extraction of object features

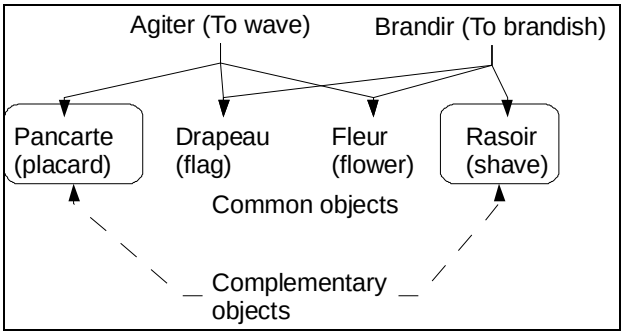

Figure 1: Complementary and Common objects of verbs to *wave* and to *brandish*.

Two types of objects appear as an output of the preceding action: Objects that are **common** to two given verbs, and objects that are called **complementary** since they appear in association with one but not with the other. In Figure 1, the considered pair of verbs is ('to *brandish'*, 'to *wave*'). Their common objects are in the pair ('*flag'*, *'flower'*) (either given from start, or already retrieved from a corpus by a previous step of the process). "*Flag"* and *flower*" are instances of a concept "*symbol*", the gathering class of '*brandish'* and '*wave'* objects. Their complementary objects, i.e., objects that appear either with one or the other, are ('*placard*', '*shave*') where '*placard*' is a retrieved object of '*wave*', and '*shave*' is a retrieved object of ' *brandish*'.

To measure the quality of our expansion approach, we propose to answer the question: Are complementary objects relevant instances of the conceptual class defined by common objects? To answer this, we have several ways to provide an evaluation protocol, and this paper will show different methods. But first, a human evaluation determines what is likely and what is not.

3.3 Step 2: Human evaluation of the quality of complementary objects

The procedure is the following.

A few concepts are selected, since they are addressed by a given corpus. For instance, in Figure 1, concept '*Symbol'* is chosen.

Conceptual classes are built with verb common objects as explained in 3.1. Here, '*Symbol'* is populated with '*flag'* and '*flower'*.

Then complementary objects are considered, and human judges have to evaluate to which extent there terms are relevant instances of the concerned concept. In the example, '*placard'* and '*shave'* are judged as possible instances of '*Symbol'*.

Evaluation consists in selecting a figure associated to one of the following propositions:

- 2: Completely relevant
- 1: Possibly relevant
- 0: Not relevant
- N: No opinion

The principle underlying this method is the following: We assume that complementary objects retrieval is a good way to discover new terms of conceptual classes because some of complementary objects are possible instance of concepts.

Human evaluation was undertaken (see experiment in section 5.1) to assert the likelihood of such an assumption. Complementary objects could very possibly be of no use for conceptual classes expansion.

Once the benefit of such an assumption acknowledged, however accurate, human evaluation might prove to be tedious, time consuming and difficult to undertake (as reported in the experiment). Thus, we have designed a **filtering procedure** that automatically sorts complementary objects by decreasing relevance. This procedure introduces a ranking function, and relies on a pre-formal data structuring called **Induced Syntactic Relation** (ISR), presented in the next section.

4 Induced syntactic relations (ISR): definition, and relevance

4.1 Defining induced relations

According to the *semantic closeness assumption*, 'to *wave'* and 'to *brandish'* (in Figure 1) are supposed to be rather close (and closeness is measured) since they have common objects. An important add-on of our approach is to **assess the status of complementary objects**. More precisely, we call **induced syntactic relation (ISR)** the following relation:

Definition:
Let v_1 and v_2 be two semantically close verbs. Let V/O be a Verb-Object relation.
Let CO_1 be the complementary object of v_1. V/O (v_1,CO_1) is true (there is a Verb/Object relation between them).
Let CO_2 be the complementary object of v_2. V/O (v_2,CO_2) is true.
V/O(v_1, CO_2) and V/O(v_2, CO_1) are the syntactic relations induced by the semantic closeness of v_1 and v_2. They are proposed as new knowledge, and their validity is evaluated.

In Figure 1, '*Placard'* and '*Shave'*, complementary objects need to be validated as possible instances of '*Symbol'*. Presently, object '*Shave'* is not a valid instance of the concept "*Symbol'*. As a filtering procedure, the automatic procedure will examine the two following induced syntactic relations:
To brandish a placard
To wave a shave

If these utterances are to be considered, by a way or another, as likely, then this is a good clue to consider '*placard'* and '*shave'* as possible instances of '*Symbol'*. So, induced syntactic relations (called ISR from now on) relevance needs to be be defined and assessed.

ISR Relevance Assumption
Let v_1 and v_2 be two semantically close verbs.
Let (KO_1, KO_2, …, KO_m) be their common objects. By definition, V/O(v_1, KO_j) is true, and V/O(v_2, KO_j) is true, for j={1,..,m}. Let K_a be their common concept (the KO_j are instances of K_a). K_a is assumed to be the conceptual class of v_1 and v_2.
CO_1 and CO_2 are possible instances of K_a if V/O(v_1, CO_2) and V/O(v_2, CO_1) are relevant.

In other words, we suppose that the complementary object is a *valid instance of the concept* defined by the common objects of the two verbs if an IRS is *relevant.* By the result presented in section 5, we have proved that our hypothesis is relevant.

Relevance is the first step before assessing complete validity. Next section shows how it is dealt with.

4.2 Ranking functions

ISR can be submitted to human approval, as objects could be submitted (see 3.3), but this is not the point: ISR has been introduced in order to pre-filter possible objects, and not to add complexity. So the best method was to examine functions that might rank ISR according to their assumed relevance [Béchet et al., 2009a].
Therefore, we need to describe the three following items:
1-How do we define the semantic relevance of a complementary object to a conceptual class
2-How this semantic relevance is computed: The methods and measures that have been chosen to achieve computation
3-Last, how IRS has been ranked according to each measure.

4.2.1 Semantic relevance definition

Definition
Let v_k be a verb.
An item I_n is assumed to belong to the conceptual class of v_k objects, if:
It has appeared as such in a corpus and has been retrieved, i.e. V/O (v_k, I_n) is satisfied.
I_n has not been retrieved but is a semantically relevant object of v_k.

4.2.2 Semantic relevance measuring process

Semantic relevance is measured as such:
1-A semantic representation of the original Verb/Object relation is computed for complementary objects. This representation is based either on a vector model, or is a digital output representing a statistical information. Both measures are detailed hereafter.
2-The same semantic representation is produced for the IRS.
3-A distance measure (or more precisely a closeness measure) is then used to assess the possible similarity between the IRS and the original relation.
3-The expected result is: **The closest the IRS and the original relation are, the more relevant to the verb, is the CO**.
For instance, in Figure 1, we measure the proximity between both syntactic relations "*to wave a placard*" (original relation) and "*to brandish a placard*" (induced relation).

4.2.3 Semantic measures

Two semantic measures belonging to two semantic modelling paradigms have been determined: *Semantic Vectors*, and Corpus co-occurrence also called *Web Validation*. Both are briefly described hereafter.

Semantic Vectors (SV): Contribution of a Vector Model to the Verb Object Relation Representation

A Semantic vector is built by projecting one or many terms on a close space vector of 873 concepts. Concepts are taken out of an ontology defined in the French Larousse Thesaurus [Larousse, 1992], a Roget-based dictionary indexing all language entries with one or several items taken from the 873 concepts ontology. For instance, the French verb "*brandir*" (to *brandish*) is associated with the concept of "*agitation*" and the noun "*drapeau*" (*flag*) is indexed by the concepts of "*paix (peace), armée (army), funérailles (funerals), signe (sign)",* and *"cirque (circus)*". The ISR vector is the result of a linear combination between verb and object vectors. Coefficients take into account the syntactic structure of the relation [Chauché, 1990]. The vector closeness is finally evaluated by a **cosine** computation between both semantic vectors (vector of the original and the induced relations).

The Web Validation (WV)

The second approach method uses the Web to measure the dependency between a verb and an object of an IRS. It is based on Turney's method [Turney, 2001] summarized as follows: A string is submitted as a query to a search engine. The number of returned results defines the dependency measure. In addition, different statistical measures such as Mutual Information [Church and Hanks, 1990] or Dice's coefficient [Smadja and al., 1996] are employed. With these measures, one can weight the IRS relevance, depending on the verb and the object composing the relation. Here, only Mutual Information is run on experiments, since this measure performed the best in previous works. The Mutual Information measure, adapted for this task, is defined as:

$$MI(v,o) = \frac{nb(v,o)}{nb(v)nb(o)}$$

where nb(v), nb(o), and nb(v,o) are respectively the number of returned results by the search engine with the submission of the verb *v*, the object *o*, and the syntactic relation *vo*. The Web validation process uses external knowledge to measure the relevance of a candidate to a concept. Thus, this validation allows for a more global evaluation of relevance for the final concepts.

Combining Measures

Combining measures has been contemplated in order to improve accuracy. Two different procedures have been defined.
- Combination 1: The first combination introduces a scalar k ∈ [0, 1] to reinforce one approach or the other. The results obtained with SV (Semantic Vectors) and WV (Web Validation) methods are first normalized. Next, both results (the figures are named SV and WV after their methods) are combined with the following formula for a syntactic relation *c*:

$$combined\_score_c = k * SV + (1 - k) * WV$$

- Combination 2: The second combined system between *SV* and *WV* has been computed. First, syntactic relations are ranked with *SV.* Then, the *n* first syntactic relations obtained with *SV* are ranked with *WV*. This second process (*WV* applied on the ranked relations with *SV*) enables to accurately sort these syntactic relations. Thus, with this adaptive combination, *SV* provides a global selection using semantic resources, and *WV* handles this first selection.

The next section presents experiments we made to measure the quality of these validation methods.

5 Experiments

5.1 Experimental setting and goals

We use a French corpus from Yahoo's site (*http://fr.news.yahoo.com/)* composed of 8,948 news

(16.5 MB) from newspapers (called corpus *T)*. Experiments are performed on 60,000 induced syntactic relations [Béchet and al. 2009b]. We have selected *manually* five concepts. Instances of these concepts are the common objects of verbs defining the concept[1]. The French selected concepts are presented in Table 1.

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)

Table 1. The five selected concepts and their instances.

The goal of these experiments are the following:
1- Evaluating the consistency of a procedure manning conceptual classes with complementary objects: A human evaluation of the complementary objects quality has been conducted as a feasibility study
2- Evaluating the quality of the filtering procedure based on semantic measures. The aim is to select the best complementary objects before giving them to a human expert. Thus, CO are ranked according to SV, WV and combined measures, as presented in section 4.2. Then the quality of the resulting lists of ranked objects is measured with experimental protocols presented in following sub-sections.

5.2 Human evaluation of the quality of complementary objects

Eight human judges have undergone the following protocol: An evaluation form was available on a specific Web page. This form allowed them to judge whether given terms can be instances of a given concept as explained in section 3.

Figure 2 gives a screen-copy of the submitted form. Resulting scores can be computed by submitting the different results to a voting system. So a term is positive if a percentage of *p* judges estimate the term to be relevant. A relevant term for a judge gets the value 1 or 2. We fix *p* at 75%.

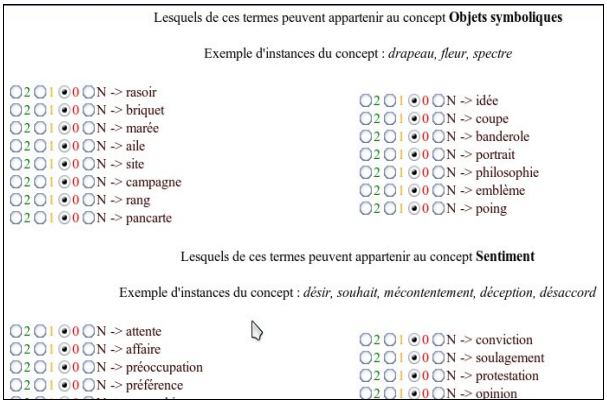

Figure 2. Screen-copy of the French form.

We obtain an accuracy score definition as the number of complementary objects divided by the number of relevant term according by the judges. The obtained score obtained is **0,14** (75 relevant terms divided by 553 complementary objects). This result shows the interest of complementary objects as instances of conceptual classes. It also shows that the number of potential candidates is high, and that an automatic procedure needs to be performed, as an aid to experts.

5.3 Evaluation of induced syntactic relations

We focus in this section on the quality of the ranking function presented in section 4.2. Here, the asset is to assert the reliability of the process, as a 'good' filter for sorting complementary objects.

Thus, we present two different evaluation protocols: A human and an automatic.

5.3.1 Evaluating relevance

Automatic Evaluation (AUTO)

The method we used to automatically measure the quality of IRS focuses on the use of a second French corpus, bigger than the first one, created from the French newspaper Le Monde (called corpus V). It contains more than 60,000 news (123 MB). It helps to determine if those ISR found in corpus T are relevant. Corpora T and V come from the same field. Thus, the first step is to automatically recover the ISR of corpus T existing in corpus V. If an ISR of corpus T appears in corpus V, it is marked down as positive (existing as a real object for the other verb), else it is negative.

Let us note that negative relations can be false negatives. Actually, a syntactic relation not found in the corpus V is not inevitably a negative relation. In addition, a relevant complementary object from an induced syntactic relation can be an irrelevant instance for a concept which has been defined 'on the spot', after the features of existing common objects. Therefore, a manual evaluation protocol, relying on human approval, is needed.

Human Evaluation (VOTING)

The human evaluation is the same as presented in subsection 5.2, except that we measure here the quality

1 From those concepts which have obtained an Asium score higher than 0.7 [Faure, 2000]

of validation approaches and not the quality of complementary objects. Thus, a relevant term for a judge gets the value 1 or 2.

The notion of 'relevant term' being defined for both AUTO and VOTING protocols, the quality of the ranked relations list is evaluated with ROC curves.

5.3.2 Evaluating ranking functions

ROC curves (Receiver Operating Characteristic), detailed in [Ferri02] are often used in medicine to evaluate the validity of diagnosis tests. ROC curves show in X-coordinate the rate of false positives (in our case, the rate of irrelevant IRS) and in Y-coordinate the rate of true positives (rate of relevant IRS). The surface under the ROC curve (AUC - Area Under the Curve), can be seen as the effectiveness of a measurement of interest. The criterion related to the AUC is equivalent to the statistical test of Wilcoxon-Mann-Whitney [yan03].

In the case of the ISR ranking using SV and WV measurements, a perfect ROC curve corresponds to a configuration where all relevant ISR are at the beginning of the list and all irrelevant syntactic relations at the end. This situation corresponds to AUC=1. The diagonal corresponds to the performance of a random system, progress of the rate of true positives being accompanied by an equivalent degradation of the rate of false positives. This situation corresponds to AUC=0.5. Figure 3 is an instance of a ROC Curve with in diagonal a random system distribution.

If the ISR are ranked by decreasing interest (i.e. all relevant ISR are after the irrelevant ones) then AUC=0.

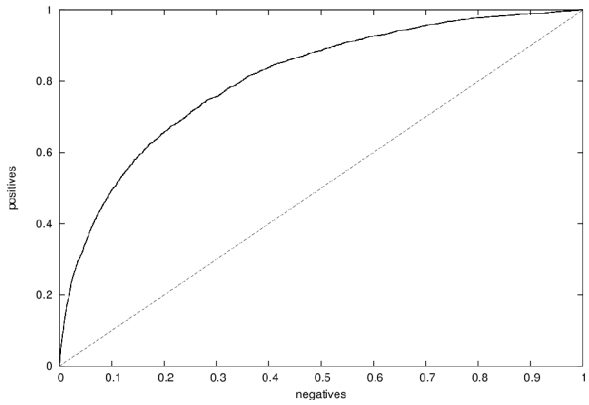

Figure 3: Example of a ROC Curve with a random distribution in diagonal.

An effective measurement of interest to order ISR consists in obtaining an AUC the highest possible value. This is strictly equivalent to minimizing the sum of the rank of the positive examples.

The advantage of the ROC curves comes from its resistance to imbalance (for example, an imbalance in number of positive and negative examples). The interest of this measure is developed in [Roche and Kodratoff, 2006].

Term	***Manual validation***
Conviction (conviction)	+
Opinion (opinion)	+
Préférence (preference)	-
Attente (waiting)	-
Colère (anger)	+

Table 2: Example of evaluation of terms for French concept "sentiment" (*feeling*).

Table 2 presents an example of ranked terms by the second combination approach. Terms are rated by a manual evaluation for the French concept "sentiment" (*feeling*). The resulted ROC Curve is given in figure 4. We finally get an AUC of 2/3 with this example (in blue in figure 4).

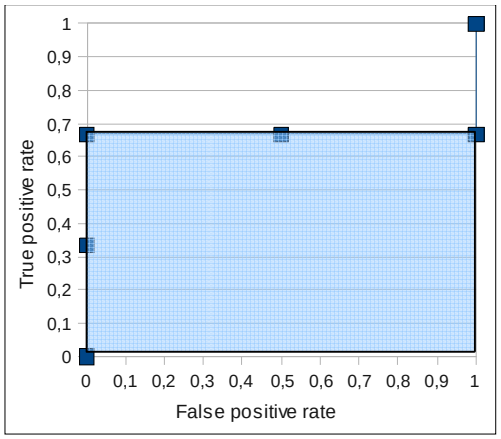

Figure 4: ROC Curve resulting of the example in Table 2

5.3.3 Experimental results

SV, WV and combined approaches propose to validate induced semantic relations by offering sorted lists of relations. The number of induced syntactic relations is taken into account by introducing a threshold of considered relations.

A fixed threshold at 100 means that AUC is computed for the only 100 first ranked (with our validation approaches) induced syntactic relations. Table 3 presents the obtained AUC for both approaches "Web validation" and "combination 2". This table compares manual and automatic evaluations. We present AUC obtained for the automatic validation by using a second validation corpus. We also present results obtained with the manual evaluation by using the voting system. A positive relation is validated if 75% of the judges give the score of 2. Better results in the Table 3 are given by "combination 2".

The manual evaluation gives good results for the second combination (AUC up to 0.83) with the first induced syntactic relations (i.e. small thresholds). Results are fair up to a threshold of 350 to finally decrease with all the induced syntactic relations (AUC close to a random distribution 0.5). Thus, we cannot provide an expert with a complete sorted list of relations but only with a selected part. So we favor the precision and the quality of the sorted list by reducing the number of possible instances to a concept.

	Web Validation		***Combination 2***	
	Vote	**Auto**	**Vote**	**Auto**
Threshold	***AUC***		***AUC***	
50	0,64	0,59	**0,81**	**0,90**
100	0,50	0,60	**0,83**	**0,87**
150	0,62	0,66	**0,80**	**0,84**
200	0,61	0,65	**0,76**	**0,79**
250	0,56	0,66	**0,71**	**0,75**
300	0,51	0,65	**0,70**	**0,74**
350	0,57	0,67	**0,69**	**0,75**
400	0,59	0,67	**0,67**	**0,74**
450	0,61	0,67	**0,65**	**0,71**
500	0,56	0,68	**0,57**	**0,70**
550	**0,52**	**0,69**	**0,52**	**0,69**

Table 3: AUC scores for the Web validation and the combination 2, with the manual (Vote) and the automatic evaluation.

We also compare the manual evaluation and the automatic scores given in Table 3. Results are similar for both evaluations. Actually, results of combination 2 for both evaluation protocols are relevant for small thresholds and decrease with all relations. Web Validation gives regular results close to 0.60 with the manual evaluation and 0.65 with the automatic.

As a conclusion about IRS relevance measure, we can say that:

Combination 2 has the best scores for all threshold values, thus is the best semantic measure among the studied ones

The first 150 ranked IRS have an AUC of and over 0,80, whatever the evaluation method is, so this means that if we retrieve the first 150 IRS with combination 2, these are a valuable material for retrieving complementary objects being possible instances of our conceptual classes, as termed in the IRS relevance assumption.

However the obtained scores are too highly rated with the automatic evaluation. These differences can be explained by the fact that two aspects are addressed by the protocols. The manual protocol addresses the relevance of a given term as an instance of a concept. The automatic protocol tries to measure the relevance of a syntactic relation built with a verb and a complementary object. These close tasks have not the same goals. Actually, automatically measuring the quality of a terms belonging to a concept is a more difficult task than measuring the quality of a syntactic relation.

6 Conclusion

This paper aims at showing and evaluating procedures that help building and expanding conceptual classes. Those tasks are quite common in terminology and ontology design. As several others, this research mines textual knowledge to do so. However, unlike others, NLP knowledge is not restricted to lexical relations but engulfs syntactic knowledge, focusing here on the verb-object dependency as a valuable relation for building and expanding conceptual classes.

One of the original features is to build classes by using common objects of semantically close verbs in a given corpus. Semantic closeness is measured with the ASIUM measure. Then, classes are expanded with complementary objects, being those ‘left over’ data, since they are not common objects.

This information source has proved to be interesting through a feasibility study conducted with a human evaluation protocol (see section 5.2). However, since it is a very abundant set of knowledge, ploughing it manually must not be considered as a possible task, since it is tedious, and time and effort consuming.

This consideration has led us to contemplating an automatic filtering procedure that would rank objects according to their relevance to the conceptual classes. Several methods could have been performed, however, we wanted here to pursue further in the light of the verb object relation, by studying the consistency of what we called the Induced Syntactic Relation, i.e., when the ‘local’ (complementary) object of a verb is exported to close verb. We made the assumption that if that Induced Syntactic Relation was to be relevant then this complementary object should play the same role as a common object, and thus should be a possible instance of the conceptual class (IRS relevance assumption).

So the problem shifted from populating a conceptual class towards measuring and asserting semantic relevance of a verb-object relation.

The second original feature of this paper was to unite both Web-based and Corpus-based techniques in order to fetch as many possible occurrences of an IRS, or to assume its inconsistency when not finding any clue about it. Among the several possible semantic models for corpora data, semantic vectors were chosen since they mix syntactic and semantic representations in a same numeric structure. Also among Web queries measures, it is Turney’s approach that has been chosen. Experiments have shown that a particular combination of measures (combination 2) proved to be the most efficient. Measures and evaluation protocols have shown that the first 150 relations, ranked with combination 2, have the best AUC scores (over 0,80), which means that they are utterly reliable.

Although very concluding, these filtering methods could be improved, at least by introducing contextual information. For Web Validation, context could be introduced in the search engine queries. With the semantic vectors approach, contextual vectors can be used. These vectors take into account the morpho-syntactic structure of the sentence containing the terms to be validated. Thus, combination 2 results might hopefully increase, either by increasing the number of acceptable IRS (AUC over 0,8) or by improving the AUC value for a fixed number of IRS.

Anyway, the IRS relevance assumption not being invalidated by experiments, we think that other dependency relations could also be contemplated: Why not the Verb-Subject relation, or the Verb-other Complements one, depending on the type of terminology or ontology one needs to retrieve. Other works have provided results in ontology populating by using the Subject-Verb-Object relation pattern in a specialized domain (e.g., Makki et al. 2009). Here, we go further by assuming ‘non retrieved’ but likely relations and ranking

them. This tends to show that NLP techniques have still a lot to offer to Web Semantics, and Ontology Design and Population.

7 References

[1] Béchet, N., Roche M., Chauché J. (2009a) A Hybrid Approach to Validate Induced Syntactic Relations. In *AINA Workshops 2009*, pp. 727-732.

[2] Béchet, N., Roche M., Chauché J. (2009b) Towards the Selection of Induced Syntactic Relations. In *ECIR'09* (poster), pp. 786-790.

[3] Bourigault D., Lame G. (2002) Analyse distributionnelle et structuration de terminologie. Application à la construction d'une ontologie documentaire du Droit, in *TAL*, pp. 43-51.

[4] Chauché, J. (1984) Un outil multidimensionnel de l'analyse du discours. In *Proceedings of COLING*, Standford University, California, pp. 11–15.

[5] Chauché, J. (1990) Détermination sémantique en analyse structurelle : une expérience basée sur une définition de distance. In *TA Information*, pp. 17–24.

[6] Church, K. W. and Hanks P. (1990) Word association norms, mutual information, and lexicography. In *Computational Linguistics*, Vol. 16, pp. 22–29.

[7] Desrosiers-Sabbath (1984) *Comment enseigner les concepts* - Sillery: Presses de l'Université du Québec.

[8] Faure D., Nedellec C. (1999) Knowledge Acquisition of Predicate Argument Structures from Technical Texts Using Machine Learning: The System ASIUM. In *EKAW 1999*, pp. 329-334.

[9] Ferri, C., Flach P., and Hernandez-Orallo J. (2002) Learning decision trees using the area under the ROC curve. In *Proceedings of ICML'02*, pp. 139–146.

[10] Harris, Z. (1968) *Mathematical Structures of Language*, New-York, John Wiley & Sons.

[11] Larousse, T. (1992) *Thésaurus Larousse - des idées aux mots, des mots aux idées*. Ed.Larousse, Paris.

[12] L'Homme M. -C. (1998) Le statut du verbe en langue de spécialité et sa description lexicographique. *Cahiers de Lexicologie,* 73, pp. 61-84.

[13] Makki, J., Alquier A-M., Prince V. (2009) Ontology Population via NLP Techniques in Risk Management. *International Journal of Humanities and Social Sciences* **3**, 3, pp. 212-217.

[14] Nazarenko A., Zweigenbaum P., Habert B, Bouaud J. (2001) Corpus-based Extension of a Terminological Semantic Lexicon. In *Recent Advances in Computational Terminology*, pp. 327-351.

[15] Smadja, F., McKeown K. R., and V. Hatzivassiloglou (1996) Translating collocations for bilingual lexicons : A statistical approach. *Computational Linguistics*, Vol. 22, No. 1, pp.1–38.

[16] Roche C. (2005) Terminologie et ontologie. *Revue Langages,* No. 157, Éditions Larousse, mars 2005, pp. 48-62.

[17] Roche M., Kodratoff Y. (2006) Pruning Terminology Extracted from a Specialized Corpus for CV Ontology Acquisition. In *Proceedings of onToContent'06 workshop (Ontology content and evaluation in Enterprise) - OTM'06*, Springer-Verlag, LNCS, october 2006, Montpellier, France, pp.1107-1116.

[18] Turney, P. (2001) Mining the Web for synonyms : PMI– R versus LSA on TOEFL. In *Proc. of ECML*, LNCS, 2167, pp. 491–502.

[19] Weeds J, Dowdall J., Schneider G., Keller DaB., and D.J. (2005) Weir Using distributional similarity to organise biomedical terminology. *Terminology*, Vol 11, No. 1, pp. 107-141.