Computer Science and Information Systems 11(1):133–155 DOI: 10.2298/CSIS130528021B

How to Combine Text-Mining Methods to Validate Induced Verb-Object Relations?

Nicolas Béchet[1], Jacques Chauché[2], Violaine Prince[2], and Mathieu Roche[2,3]

[1] GREYC – UMR 6072, CNRS – Univ. de Caen Basse-Normandie,
14032 Caen Cedex – France
[2] LIRMM – UMR 5506, CNRS – Univ. Montpellier 2,
34000 Montpellier – France
[3] TETIS – Cirad, Irstea, AgroParisTech,
34093 Montpellier Cedex 5 – France

Abstract. This paper describes methods using Natural Language Processing approaches to extract and validate induced syntactic relations (here restricted to the Verb-Object relation). These methods use a syntactic parser and a semantic closeness measure to extract such relations. Then, their validation is based on two different techniques: A Web Validation system on one part, then a Semantic-Vector-based approach, and finally different combinations of both techniques in order to rank induced Verb-Object relations. The Semantic Vector approach is a Roget-based method which computes a syntactic relation as a vector. Web Validation uses a search engine to determine the relevance of a syntactic relation according to its popularity. An experimental protocol is set up to judge automatically the relevance of the sorted induced relations. We finally apply our approach on a French corpus of news by using ROC Curves to evaluate the results.

Keywords: Text-Mining, Web-Mining, Syntactic Analysis

1. Introduction

Semantic knowledge acquisition is an important issue in Natural Language Processing (NLP), since such knowledge can be used in information retrieval and/or classification tasks. Semantic knowledge deals with the existing relationships between words (seen as concept names). Some of these relationships are classically known to help building ontologies, terminological sets and classifications, etc. Several other applications, such as machine aided translation or indexing, naturally rely on semantic knowledge.

Syntactic information, i.e., knowledge about sentence and phrase structures and therefore, knowledge about structural relations between words in discourse, is quite useful to build semantic knowledge [10]. It helps dealing with compound concept names (e.g. *country house*) and could be used to create conceptual classes (gatherings of words). For instance, the words *house*, *hangar*, and *farmhouse* can be gathered in a set representing the *building* concept. Since they also act as concept names, they can be hierarchically organized to build *conceptual hiearchies*. Note that the way these words are related in sentences (i.e. syntactic relations in the texts) can help to discover semantic relations.

Actually, syntactic relations are relevant features of fields associated to sentence constituents [16]. Two kinds of syntactic relations can be used to build concepts. First, the

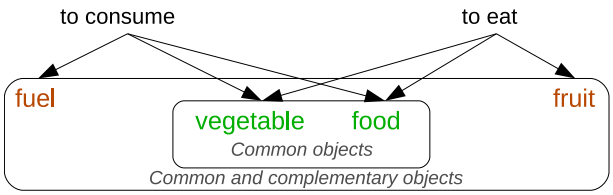

Fig. 1. Common and complementary objects of verbs 'to consume' and 'to eat'

so called **original** relations, directly extracted by a syntactic parser from a text [17], [28]. Second, relations can be considered as **induced** syntactic relations. The latter have been introduced in the Asium system [11]. The system underlying principle consists in gathering verbs object complements considered to be *close* according to a quality measure. For example, in Figure 1, if the verbs 'to consume' and 'to eat' are seen as *close verbs*, and if they are related to objects such as 'fuel, vegetable, meat, and fruit' in different sentences, then the latter are gathered in the same set, induced by the similarity of their syntactic position in sentences. Note that this method is not specific to verb-object relations. It also applies to verb-subject relations, and on might gather verb subjects considered *close*. The similarity is conveyed by verbs and delivered to their possible complements. Other approaches of the literature gather terms by using proximity measures such as cosine, Dice coefficient, or Jaccard [25,21]. These proximity measures are based on statistical information [3,27,25].

Our issue deals with the Asium system measure. The latter is the most adapted in building **induced syntactic relations** [2,13]. This type of relations is introduced in the following paragraph. In addition, in a syntactic relation proximity context, the Asium measure produces results very close to usual measures like cosine, Mutual Information, or Dice coefficient, as discussed in section 2.2.

The way closeness is defined is already related to the existence of syntactic relations: In fact, *consume* and *eat* having two common objects (*vegetable* and *meat* represent the seed of the 'food' concept), retrieved by a parser as original relations, they are assumed to be 'close'. As a feedback, closeness is assumed for other possible objects for which occurrences in corpora have appeared. As a summary, let V_1 and V_2 be two verbs that are said to be close if they have at least a common object (closeness will be measured as a function of the common objects number). Let $Obj_1^{V_1}...Obj_n^{V_1}$ and $Obj_1^{V_2}...Obj_m^{V_2}$ be the objects of the verbs V_1 and V_2, $Obj_i^{V_1}$ ($i \in [1,n]$) is called a common object if $\exists j \in [1,m]$ where $Obj_i^{V_1} = Obj_j^{V_2}$. If $Obj_k^{V_1}$ (resp. $Obj_k^{V_2}$) is not a common object then the V_2-$Obj_k^{V_1}$ relation (resp. V_1-$Obj_k^{V_2}$) is called an **induced syntactic relation** and $Obj_k^{V_1}$ is called a **complementary** object. For instance, induced relations in Figure 1 are *to eat fuel* and *to consume fruit*. They represent new knowledge since they were not present in the initial corpus. However, such knowledge is not always 'acceptable': For instance, *to eat fuel* is an odd combination of words (from a pragmatic point of view), not likely to occur unless in weird metaphors about heavy-consuming vehicles. Induced syntactic relations have to

be validated by a human expert who will assess their likelihood. A relation is said to be **likely** if it is accepted by the expert as a possible occurring sentence. If a relation is likely we can add the object to the seed concept (e.g. we can add *fruit* to the 'Food' concept).

This work aims at determining the quality of induced syntactic relations to alleviate the expert task in validating extracted knowledge (e.g. *to eat fuel*: Unlikely vs *to consume fruit*: Likely). We propose to rank induced relations by using different approaches. A first approach considers a syntactic relation as a combination of different concepts representations based on a thesaurus indexing words and their meanings with a small set of basic concepts. A second approach is deliberately Web-oriented, and uses statistical measures relying on the number of pages retrieved by a Web search engine. Both approaches are compared (with their pros and cons) and finally combined to get the best of both.

Next section describes the first validation techniques used to rank induced syntactic relations. Section 3 presents experimental results obtained by using a specific protocol. Protocol capacities and obtained results are then discussed. Conclusions are sketched in section 4.

2. Thesaurus and Web Based Methods Determining Induced Relations Likelihood

2.1. Contextual Semantic Vectors (SV)

We have chosen the Semantic Vectors Approach to represent words and sentences semantics. We first consider the representation of verbs and objects with *Contextual Semantic Vectors*. Then we compute a relevance measure between the verb and the object vectors, to finally rank all the induced syntactic relations by likelihood.

The vector base is defined with the space in which words and sentences are represented, i.e. the 873 concepts described in the Larousse French thesaurus [15], a French version of the Roget for English. The vector base is sort of a 'conceptual ontology', whose items index all dictionary words. Actually, more than 60,000 words are indexed with this ontology and thus represented as vectors. Each term, and ontology concepts are included, is indexed by one or many items. For example, 'to consume' is related to 'thin, nutrition, accomplishment, use, expense, and meal'. Thus, the resulting vector will be composed of active concepts as illustrated in Figure 2. The ontology concepts (vector base) are numbered (from 1 to 873). The word vector gets a 1 value if the ontology concept contributes to its meaning, 0 otherwise. In Figure 2, the following French concepts are positives: "Fin" (*i.e.* Thin), "Nutrition" (*i.e.* Nutrition), "Éducation" (*i.e.*Thin), "Accomplissement" (*i.e.* Accomplishment), "Usage" (*i.e.* Use), "Dépense" (*i.e.* Expense), "Repas" (*i.e.* Meal).

Words vectors, defined as such are 'inert vectors': All possible meanings are evoked, but only as a potential. Their active concepts are not differentiated in intensity. In corpora, words benefit from information conveyed by the sentences that use them, i.e., they are enriched by their neighbors. A contextual representation of a term enhances the original vector by modifying the components activity. Therefore, it is better to use a contextual representation of terms by relying on the sentence they come from. The SYGFRAN parser [5] can compute Semantic Vectors (SV) of sentences and produces SV for words. The parser

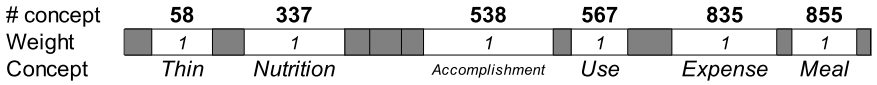

Fig. 2. The Semantic Vector representation of 'to consume'. Grey sections are filled with null values.

takes into account the syntactic role of a word or a phrase, in a sentence. SYGFRAN transforms a sentence in a constituents tree by identifying governing constituents from governed ones. Weights are defined as powers of 2 beginning with 2^0 for the leaf of the most dependent constituent to 2^p, p representing the rank of the highest governing component in the parsing tree (i.e. verb and subject are highest nodes). Then, the contextual Semantic Vector representation of a term extracted from a sentence includes contextual information by weighting these active concepts. An example of a contextual Semantic Vector computation for the verb 'to consume', in a sentence where 'to consume meat' occurs, is given in Figure 3. More details on the weight computation are given in [19]. It is obvious that the *Nutrition* and *Meal* components, semantically related to the word *meat* (they belong to its class) are quite enhanced when *meat* occurs with *to consume*.

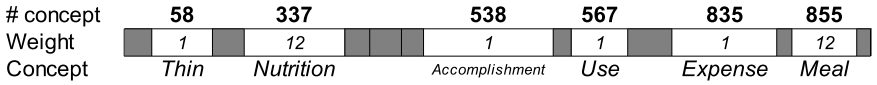

Fig. 3. The **contextual** Semantic Vector representation of 'to consume' in a sentence where 'to consume' and 'meat' co-occur

The purpose of this article is to determine if an induced Verb-Object relation is likely (i.e. **if the association of a complementary object and a verb is semantically relevant**). To do so, one has to rely on the largest contextualized representation of a verb. In other words, what are the most 'popular' concepts which are enhanced for a given verb, in the sentences where it is used? Therefore, we consider every sentence (of a given corpus) where a term appears and build all corresponding contextual Semantic Vectors. The global Contextual semantic representation of this term is defined as the centroid of all the contextual Semantic Vectors representing this term in the corpus sentences. An example of the global representation for the verb 'to consume' is given in Figure 4 (computed in the corpus used in the experiments described in next section).
Using Semantic Vectors actually leads to assessing the quality of an induced syntactic relation in the context of a given corpus. For example, according to the used corpus, the active vector concepts of the verb 'to consume' in Figure 4 have a homogeneous distribution of concepts (i.e. component values are close to each other), but concepts *thin* and *accomplishment* have a slightly higher weight over other concepts which are linked to the notion of eat (i.e., *nutrition*, or *meal*). The induced relation scope has to be defined,

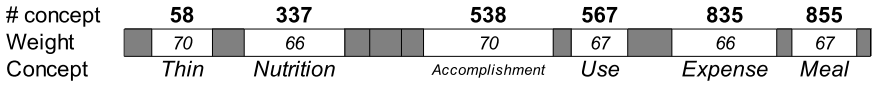

Fig. 4. The **contextual** global representation of 'to consume' in the corpus

and two options were available: Either the word itself or the whole sentence that contains it. Therefore two representations of an object have been used: The *Noun* itself and the *Noun-Phrase* or sentence which acts as an Object in the dependency relation determined by the parser, as well as two representations of the verb: The *Verb* itself and the *Verb-Phrase*. This paper presents the couple which performs the best in experiments: *Verb* and *Noun-Phrase*.

Measuring Semantic Relevance by Ranking Vectors Similarity Measuring the semantic relevance of an object to a verb in an induced syntactic relation, relies on a similarity assessment between an object vector, and the contextual global representation of the verb. The **cosine** measure has been the first to be used in Information Retrieval, and it is since regularly applied to determine vectors similarity. Its values rank between 0 and 1. When cosine values get close to 1, then vector angles tend to get very small. An angle between two vectors indicates their respective directions. The smaller the angle, the closer vectors are, thus emphasizing their similarity. Cosine is computed as the scalar product of both vectors divided by the norms product of both vectors.
Other similarity measures can be used such as the **Matching Measure** [19]. Let us consider two vectors $\boldsymbol{A}$ and $\boldsymbol{B}$. Their components values are sorted, from the most activated to the less activated (*i.e.* the most activated concepts are those having the most important weight, for instance concepts Nutrition and Accomplishment in Figure 4). Then, to have a reliable and discriminating comparison between two vectors, we considered that only the most activated concepts need to be kept. In this way, only the $873 \times 1/s$ first vectors components are used in order to compute the proximity between vectors, with s a scalar empirically chosen, and 873 the number of concepts in the Larousse thesaurus. Respective resulting vectors are noted $\boldsymbol{A_{tr}}$ and $\boldsymbol{B_{tr}}$. If these new vectors have no common components, the matching measure is 1. Otherwise, we need to compute the rank and the intensity differences.

The rank difference $E_{i,\rho(i)}$ is:

$$E_{i,\rho(i)} = \frac{(i - \rho(i))^2}{Nb^2 + (1 + \frac{i}{2})} \tag{1}$$

Where i is the rank of C_t a component of $\boldsymbol{A_{tr}}$ and $\rho(i)$ the rank of the same component in $\boldsymbol{B_{tr}}$, where Nb is the number of values kept.

The intensity difference $I_{i,\rho(i)}$, which compares the intensity of common strong components is:

$$I_{i,\rho(i)} = \frac{\left\| a_i - b_{\rho(i)} \right\|}{Nb^2 + (\frac{1+i}{2})} \tag{2}$$

Where a_i is the intensity of i rank component from $\boldsymbol{A_{tr}}$ and $b_{\rho(i)}$ the intensity of the same component in $\boldsymbol{B_{tr}}$ (its rank is $\rho(i)$).

Remark: Two remarks have to be made, in order to explain the rationale of these two formulas.
First both $E_{i,\rho(i)}$ and $I_{i,\rho(i)}$ are not symmetrical, thus leading to a matching measure that does not act as a distance *per se* (later on, for other needs, another formula, called 'concordance distance' has been defined by A. Labadié by transforming both intensity and rank differences into symmetrical elements). Second, in the rank difference, if the i rank component is leading (i.e close to 1) the value of the rank difference is emphasized whereas it is bit less important in the intensity difference. Intensity is an interesting measure only if ranks i and $\rho(i)$ are close, and if the rank of the leading component is a small figure (good ranks).This explains why, in the first formula, i is directly divided by 2 whereas it is $1+i$ in the second. When i is not leading (i.e. i much bigger than 1), then the difference between i and $1+i$ tends to fade, and the bigger i is, the more $1+\frac{i}{2}$ gets close to $\frac{1+i}{2}$. It means that the measure will emphasize the differences in rank and intensity mostly on the best ranks and will neglect the lesser ones.

With both differences, the **matching measure** P is computed as:

$$P(\boldsymbol{A_{tr}}, \boldsymbol{B_{tr}}) = (\frac{\sum_{i=0}^{Nb-1} \frac{1}{1+E_{i,\rho(i)} * I_{i,\rho(i)}}}{Nb})^2 \tag{3}$$

As P concentrates on components intensities and ranks, the overall components direction is introduced by mixing P with the angular distance noted $\delta(\boldsymbol{A}, \boldsymbol{B})$ for vectors $\boldsymbol{A}$ and $\boldsymbol{B}$. We note $\Delta(\boldsymbol{A_{tr}}, \boldsymbol{B_{tr}})$ the resulting measure:

$$\Delta(\boldsymbol{A_{tr}}, \boldsymbol{B_{tr}}) = \frac{P(\boldsymbol{A_{tr}}, \boldsymbol{B_{tr}}) * \delta(\boldsymbol{A}, \boldsymbol{B})}{w * P(\boldsymbol{A_{tr}}, \boldsymbol{B_{tr}}) + (1-w) * \delta(\boldsymbol{A}, \boldsymbol{B})} \tag{4}$$

Where w is a coefficient used to give more (or less) weight to P. To compute a distance, symmetry is needed. Therefore, the **matching measure** $D(\boldsymbol{A}, \boldsymbol{B})$ is define by the following formula:

$$D(\boldsymbol{A}, \boldsymbol{B}) = \frac{\Delta(\boldsymbol{A_{tr}}, \boldsymbol{B_{tr}}) + \Delta(\boldsymbol{B_{tr}}, \boldsymbol{A_{tr}})}{2} \tag{5}$$

In order to compute the scalar combination of the validation approaches, a final computation is applied to the matching measure. Thus, the resulting score between two close Semantic Vectors will be close to 1.

$$D(\boldsymbol{A}, \boldsymbol{B})_{Final} = 1 - D(\boldsymbol{A}, \boldsymbol{B}) \tag{6}$$

2.2. The Web Validation Approach (WV)

The previous subsection showed a method that relied on important NLP resources to determine the quality of an induced relation: A thesaurus, a vector representation of words

(in a lexical base of 60,000 vectors), a syntactic parser computing dependency relations, a Semantic Vector calculus procedure computing sentence vectors and, afterwards, contextualizing words vectors. Moreover, contextualized vectors are fed by sentences extracted from corpora. So the question was: Could we use NLP knowledge combined with statistical information in order to validate induced relations (i.e. constructed relations not present in texts)? Therefore, we tried to build a Web based method to measure the semantic relevance of an object to a verb in an induced syntactic relation. *Semantic Relevance is here assumed to be approached by Web Popularity*. This type of assumption already exists in several works that are briefly exposed hereafter.

Related Work. Web Validation has been more or less initiated by Turney [24], in a completely different context, but which has in common with ours that it tries to fathom a 'similarity' or 'compatibility' between items. The algorithm PMI-IR (Pointwise Mutual Information and Information Retrieval), described in [24], queries the Web via the AltaVista search engine to determine appropriate synonyms to a given query. For a given word, noted $word$, PMI-IR chooses a synonym among a given list. These selected terms, noted $choice_i$, $i \in [1, n]$, correspond to the TOEFL questions. The aim is to compute the $choice_i$ synonym that gives the best score. To obtain scores, PMI-IR uses several measures based on the proportion of documents where both terms i.e, '$word$' and '$choice_i$', are present. Turney's formula is given below (7): It is one of the basic measures used in [24]. It is inspired from Mutual Information[6].

$$score(\ choice_i\ ) = \frac{nb(\ word\ NEAR\ choice_i\ )}{nb(\ choice_i\ )} \tag{7}$$

– $nb(x)$ computes the number of documents containing the word x,
– $NEAR$ (used in the 'advanced research' field of AltaVista) is an operator indicating if two words are present in a 10 words wide window.

With formula (7), the proportion of documents containing both $word$ and $choice_i$ (within a 10 words window) is calculated, and compared with the number of documents containing the word $choice_i$. The higher this proportion is, the more $word$ and $choice_i$ are seen as synonyms. More sophisticated formulas have also been applied: They take into account the existence of negation in the 10 words windows. For instance, the words *big* and *small* are not synonyms if, in a given window, a negation associated to one of these two words has been detected. Let us note that antonymy, the lexical function tying *big* to *small*, is not considered here.
Other papers in the literature propose to use the Web. For instance, in [7], the authors detail a measure evaluating a similarity between words and phrases by using, among others, the Google search engine.

The approach of [14] is based on the Web to obtain frequencies of bigrams. They use Verb-Object, Noun-Noun, or Adjective-Noun patterns in order to extract the bigrams. They demonstrate the performance of Web frequencies in a pseudo-disambiguation task. Recently [4] use PMI-IR and Web frequencies in order to propose an evaluation methodology based on calculus of precision/recall.

Our Approach. We propose to query the Web with a syntactic relation represented by a string (for instance the French relation 'consommer un fruit', meaning 'to consume a fruit'). The underlying assumption is the following: *If this query happens to be present, and is frequent, then it is a measure of its consistency, and therefore, its likelihood.*

The query outcome is given by the $nb(x)$ function which is the number of pages provided by the search engine Yahoo by using an API [4]. In French,the language we use in our experiments, a verb and its object are usually separated by an article. So we consider five usual French articles *un, une, (i.e. 'a'), le, la, l' (i.e. 'the')* to compose the string representing our query. Then, $nb(v, o)$ is the number of pages provided by the search engine for the Verb-Object syntactic relation (v,o) where v and o are respectively the Verb and the Object. The following formula presents the $nb(v, o)$ computation:
$nb(v, o) = nb(v$ **un** $o) + nb(v$ **une** o$) + nb(v$ **le** o$) + nb(v$ **la** o$) + nb(v$ **l'** o$)$

$nb(v$ **un** $o)$ is the value returned by the search engine Yahoo for the string 'v **un** o'. Then, a statistic measure is applied to compute the semantic relevance of the object o and the verb v from the list of induced Verb-Object relations (among which we have 'to consume a fruit'). Four statistical measures have been examined to assess the quality of a given verb-Object pair. The first one is the **sum** defined by the $nb(v, o)$ computation. It is the most obvious one, and highlights the characters string occurrence frequency. The three others are presented next.

Mutual Information. One of the most commonly used measures to compute a sort of relationship between words composing what is called a **co-occurrence** of two words x and y, is Church's Mutual Information (MI) [6]. The formula is the following:

$$MI(x, y) = \log_2 \frac{P(x, y)}{P(x)P(y)} \tag{8}$$

Such a measure tends to extract rare and specific co-occurrences according to [9,23]. We suggest to apply this measure in a different context, to evaluate the quality of extracted Verb-Object relations. Let us notice that, in formula (8), the log_2 function is not mandatory, since it is strictly ascending. Thus, the order of Verb-Object relations likelihood provided by the measure is not impacted by the application of log_2. In this framework, the $P(v, o)$ measure is the probability of o being the object of v in the given language. For instance, in a sentence such as *The congressman talks to the consul'*, v is *to talk* and o is $consul$ (in French, "le député parle au consul")[5]. When simplified by making an empirical approximation based on maximum likelihood estimation, formula (8) could be written as follows, where nb designates the number of answers returned by the search engine:

$$MI(v, o) = \frac{nb(v, o)}{nb(v)nb(o)} \tag{9}$$

[4] http://api.search.yahoo.com

[5] Note the use in French of a preposition "au", which is not used in our work. Actually, we do not use prepositions because the number of queries necessary with web-based approach would be significantly increased in order to take into account all possibilities of relations between verbs and nouns.

Cubic Mutual Information. The Cubic Mutual Information is an empirical measure based on MI, that enhances the impact of frequent co-occurrences, a feature absent from the original MI [8]. This measure is defined by formula (10). Vivaldi *et al.* have estimated that the Cubic MI was the best behaving measure [26] in terminology extraction domain. We also suggest to apply this measure in the evaluation of extracted Verb-Object relations quality.

$$MI^3(v,o) = \frac{nb(v,o)^3}{nb(v)nb(o)} \tag{10}$$

Of course other formulas exist but we have preferred to focus on MI^3, because of its good behavior. We can cite for instance the MI correction factor of [18].

Dice's Coefficient. An interesting quality measure is Dice's coefficient [22]. It is defined by the following formula (11).

$$D(x,y) = \frac{2 \times P(x,y)}{P(x) + P(y)} \tag{11}$$

Formula (11) leads directly to formula (12).[6]

$$Dice(v,o) = \frac{2 \times nb(v,o)}{nb(v) + nb(o)} \tag{12}$$

After computing the score of all Verbs and Objects of induced Verb-Object relations, the rank of all induced Verb-Object relations is obtained.

Note: The Web Validation approach was experimented on 50,000 induced Verb-Object relations (section 3). It needed 350,000 queries when using MI^3 (1 for the verb, 1 for the object and 5 for an induced Verb-Object relation: 50,000 $\times$ 7 = 350,000). Thus, the Web Validation approach is time-costly.

2.3. Combining Both Approaches

Both Contextual Semantic Vectors (SV) and Web Validation (WV) have pros and cons. SV are precise, linguistically grounded, but rely on important resources, and could be biased by the nature of the corpora at hand. On the other hand, WV is easy, available, browses a large amount of data, but could be criticized in its basic principles: Popularity is not necessarily a proof of quality, and Web pages are quite speckled with wrong sentences and poor style.

In order to take into account both types of knowledge, we searched for combinations that would give the best results. We tried two different combinations and selected the best fitting one.

[6] by writing $P(v) = \frac{nb(v)}{nb\_total}$, $P(o) = \frac{nb(o)}{nb\_total}$, $P(v,o) = \frac{nb(v,o)}{nb\_total}$

Combination 1: A Combined System with a Variant Scalar. In the first combination, a scalar $k \in [0, 1]$ has been introduced to reinforce one approach over the other when it behaves better. The results obtained with SV and WV methods are first normalized. Next, results of both SV and WV are combined according to the following formula for a Verb-Object relation c:

$$combine\_score_c = k \times SV + (1 - k) \times WV \tag{13}$$

Combination 2: An Adaptive Combined System. With the second combination, the first step consists in ranking Verb-Object relations with SV (and cosine as similarity measure). Then, the n first Verb-Object relations (obtained with SV) are ranked with the WV. This second process (WV applied on the ranked relations with SV) enables to accurately sort these n Verb-Object relations. Thus, with this adaptive combination, SV offers a global selection using semantic resources, and WV sorts out the most popular among the most semantically accurate relations. Let us note that we also experiment the opposite cascade, first WV and then SV but this kind of approach does not appear relevant.

3. Experiments

3.1. Experimental Protocol

Experiments used two corpora. A corpus from Yahoo's site [7]: 8,948 news items (16.5 MB), used as a test corpus, and called corpus T. The second one, called corpus V, comes from the French newspaper *Le Monde*, and plays the role of the validation corpus. It contains more than 60,000 news items (123 MB). It is needed to determine whether Induced Syntactic Relations (ISR) of corpus *T* are likely (We use here the general term 'syntactic relation', although experiments ran and were designed for Verb-Object, assuming that other relations such as Subject-verb will behave as such, as discussed in previous sections). Corpora T and V address the same field (newspaper with journalistic writing). The goal is to automatically evaluate the Induced Syntactic Relations quality from corpus T by checking if they occur in corpus V. If an ISR of corpus T appears in corpus *V*, we consider it as **positive**, else it is **negative**. The different approaches presented in section 2 (*Semantic Vectors*, *Web Validation*, and the *Combined Systems*) are used to rank induced syntactic relations. ROC curves measure the quality of the obtained ranking.

Initially ROC curves (Receiver Operating Characteristic), detailed in [12], come from the field of signal processing. ROC curves are often used in medicine to evaluate the validity of diagnosis tests. ROC curves show in X-coordinates the rate of false positives (in our case, rate of unlikely induced syntactic relations) and in Y-coordinates the rate of true positives (rate of likely induced syntactic relations). 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 surface under the curve is equivalent to the statistical test of Wilcoxon-Mann-Whitney (see [29]).

[7] http://fr.news.yahoo.com/

In the case of the ISR ranking with SV and WV measurements, a perfect ROC curve corresponds to obtaining all likely ISR at the beginning of the list and all unlikely syntactic relations at the end of the list. 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 $AUC = 0.5$. The Figure 5 is an instance of a ROC Curve with in

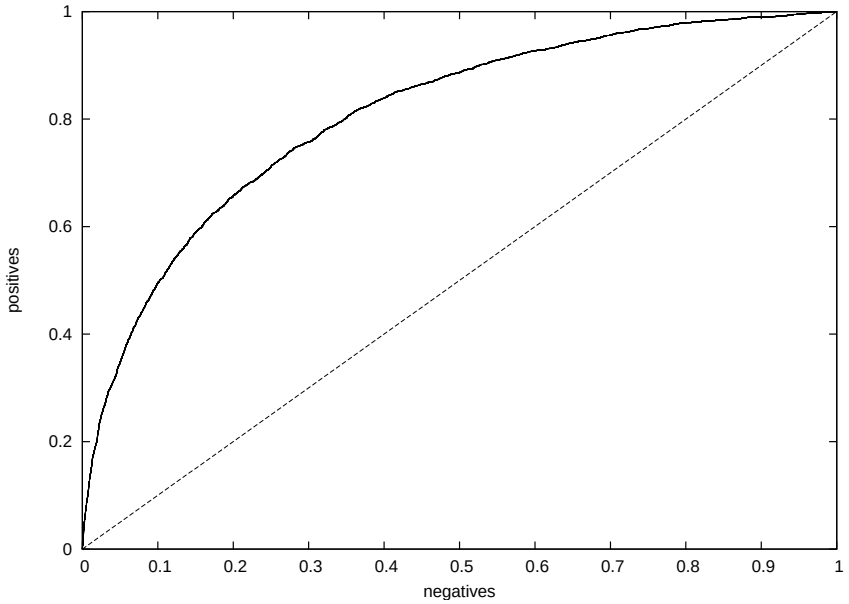

Fig. 5. Example of ROC Curve

diagonal a random system distribution. If the ISR are ranked by decreasing interest (*i.e.* all likely ISR are behind the unlikely ones in the ordered list) then $AUC = 0$. An effective measurement of interest to rank ISR consists in obtaining the highest possible AUC. This is strictly equivalent to minimizing the sum of the positive examples ranks.
The advantage of ROC curves comes from their resistance to imbalance (for example, an imbalance in number of positive and negative examples). The interest of this measure is developed in [20].

We should notice that unlikely relations (a negative case in the ROC curve application) can get a false negative value. Indeed, according to this evaluation protocol, an unfound ISR in the corpus V is an irrelevant one.

3.2. Results

Evaluation of Contextual Semantic Vector and Web Validation Approaches. The goal of our experiments is to have all positive relations (i.e. those likely and present in

CSV *Threshold*	***Cosine***	***Matching Distance***
5000	0.451	**0.520**
10000	0.502	**0.516**
15000	0.510	**0.532**
20000	0.501	**0.551**
25000	0.506	**0.573**
30000	0.512	**0.591**
35000	0.550	**0.606**
40000	0.578	**0.620**
45000	0.596	**0.634**
50000	0.603	**0.651**

Table 1. AUC obtained with the Contextual Semantic Vectors approach (CSV)

both corpora) at the top of the list. Results are presented with different thresholds. Table 1 presents AUC with different thresholds using the SV with the Verb and Noun Phrase pair. This threshold value is the number of induced syntactic relations we take into account to calculate the AUC. For example, for a threshold at $10,000$, we compute the AUC only on the $10,000$ first ranked relations. For all measures computed, results obtained with SV are poor, very close to a random distribution (AUC=0.5) for the first thresholds. This unsatisfactory results could be explained by the nature of the SV. Actually, Semantic Vectors are composed of 873 general concepts which are not always adapted to measure the quality of induced syntactic relations.

Even if the experiments have been conducted on journalistic documents, which deal with general topics (e.g. politics, sports, or sciences), some used terms are very specific. So the linguistic knowledge given by Semantic Vectors are not enough specialized. In our context, all available semantic resources have the same limit, they are too general. So we plan to investigate this point in order to enrich semantic resources such as Larousse French thesaurus to improve our SV approach.

However this approach performs better for important thresholds (0.60 for a threshold at $50,000$). In our experiments, we use two similarity measures: cosine and matching measure (see Section 2.1). We select matching measure that gives best results between both similarity measures (see Table 1).

Evaluation of the Web Validation Approach. The WV approach presented in Table 2 gives better results. For the first syntactic relations (small thresholds) the AUC are unsatisfactory. The WV approach is efficient for a large amount of syntactic relations. But according to the task in which induced syntactic relations are used, good results must be obtained with small thresholds. For example, induced syntactic relations could help to build conceptual classes. This process enables to extract a limited number of syntactic relations in order to evaluate syntactic relations by an expert.

Web Val.	*Statistic measure*			
Threshold	*Sum*	*MI*	*MI³*	*Dice*
5000	0.614	0.621	0.608	**0.628**
10000	0.650	0.641	**0.661**	0.658
15000	0.683	0.667	0.685	**0.687**
20000	0.708	0.681	**0.711**	0.708
25000	0.728	0.697	**0.732**	0.719
30000	0.744	0.708	**0.748**	0.737
35000	0.758	0.714	**0.760**	0.748
40000	0.772	0.722	**0.773**	0.759
45000	**0.786**	0.728	0.785	0.770
50000	**0.802**	0.747	0.800	0.786

Table 2. AUC obtained with the Web Validation approach

Comb. 1	*K value*										
Threshold	*0 (WV)*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10 (CSV)*
5000	0.608	0.620	0.643	0.666	0.654	0.674	**0.722**	0.696	0.660	0.615	0.520
10000	0.661	0.666	0.666	0.680	**0.691**	0.637	0.586	0.571	0.546	0.539	0.516
15000	0.685	0.691	**0.702**	0.697	0.668	0.630	0.596	0.574	0.558	0.549	0.532
20000	0.711	0.721	**0.722**	0.707	0.684	0.649	0.617	0.588	0.572	0.561	0.551
25000	0.732	**0.741**	0.738	0.724	0.698	0.667	0.634	0.611	0.598	0.585	0.573
30000	0.748	**0.756**	0.754	0.740	0.718	0.687	0.656	0.628	0.608	0.599	0.591
35000	0.760	**0.767**	0.764	0.752	0.732	0.707	0.677	0.649	0.625	0.611	0.606
40000	0.773	**0.776**	0.774	0.764	0.747	0.725	0.700	0.673	0.648	0.627	0.620
45000	0.785	**0.786**	0.782	0.773	0.759	0.739	0.717	0.694	0.672	0.647	0.634
50000	**0.800**	0.799	0.795	0.787	0.773	0.756	0.737	0.716	0.695	0.673	0.651

Table 3. AUC obtained with the first combined system

Evaluation of the Hybrid Approaches. In order to improve the first induced syntactic relations ranking (i.e. first thresholds), we computed combinations of both precedent approaches: SV and WV. The AUC obtained by applying the combination detailed in section 2.3 are given in Table 3. Since parameter k is variable, we tested different values of $k \in \{0.1...0.9\}$ with an increment of $1/10$. We also report the $k = 0$ results (WV) and $k = 1$ results (SV). Results are clearly better for small thresholds with a k value $\in \{2, 5\}$. For instance with a threshold at $5,000$, we improve the WV scores by 0.05 points with $k = 4$. However, these improvements are not sufficient since AUCs are still poor.

The second combination presented in section 2.3 was then applied. For example, for a threshold at $10,000$, all syntactic relations are first sorted with the SV approach and then the ordered $10,000$ first syntactic relations are sorted using the WV method. This approach improves all previous obtained AUC (Table 4). For instance with a threshold at $5,000$, the AUC of the second combination is 0.813 vs. 0.643 for the first combination and 0.608 for the WV. Thus, the second combination seems to be the better approach to rank induced syntactic relations.

Figures 6 and 7 show respectively the ROC Curves representation of the SV, the WV, and both combinations for a threshold at $5,000$ and $50,000$. With the previous results and

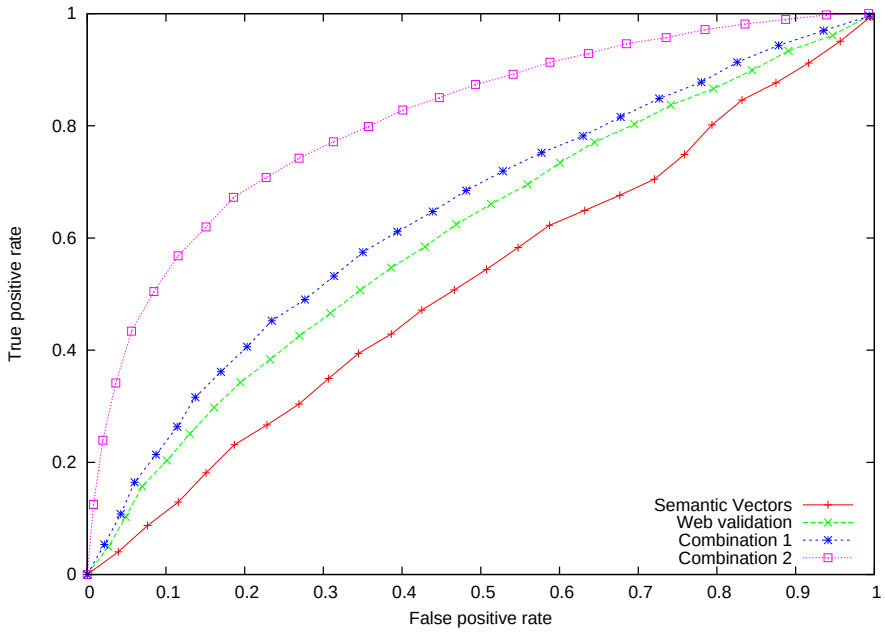

Fig. 6. ROC Curves for a threshold of 5,000

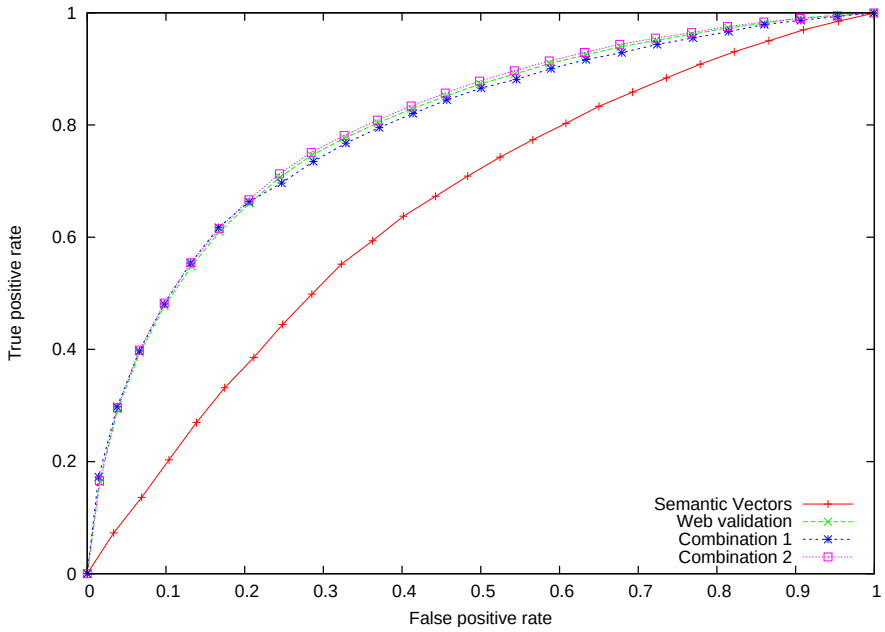

Fig. 7. ROC Curves for a threshold of 50,000

	Used approach			
Threshold	*CSV*	*WV*	*Comb. 1*	*Comb. 2*
5000	0.520	0.608	0.643	**0.813**
10000	0.516	0.661	0.666	**0.795**
15000	0.532	0.685	0.702	**0.788**
20000	0.551	0.711	0.722	**0.790**
25000	0.573	0.732	0.738	**0.789**
30000	0.591	0.748	0.754	**0.791**
35000	0.606	0.760	0.764	**0.793**
40000	0.620	0.773	0.774	**0.796**
45000	0.634	0.785	0.782	**0.799**
50000	0.651	0.800	0.795	**0.804**

Table 4. AUC obtained with the second combined system compared to others

their ROC Curves, we can conclude that the WV is more adapted for high or medium thresholds (for instance in a corpus expansion task) and the second combination is more adapted with small thresholds (for instance in a conceptual classes building). Results obtained with the second combination seem to be the most promising. They are based on a first selection determined by the SV approach, which, as previously said, relies on linguistic information to ground semantic relevance of an object to a verb. However, AUC obtained with the SV is very close to a random distribution. From a sole statistical point of view, linguistic information does not introduce a noticeable change. We wanted to investigate this 'negative result' which, epistemologically, is quite problematic. We produced a random rank of the induced syntactic relations. Then we sorted the first relation with WV. We called this approach the 'random' second combination. Table 5 compares the AUCs obtained with the 'classical' second combination and the 'random' second combination. AUCs for both approaches are similar.

	SV-WV		*Random-WV*	
Threshold	*AUC*	*+*	*AUC*	*+*
5000	0.813	1362	0.801	750
10000	0.795	2675	0.808	1542
15000	0.788	3809	0.808	2323
20000	0.790	4790	0.810	3078
25000	0.789	5575	0.809	3863
30000	0.791	6248	0.809	4648
35000	0.793	6838	0.808	5438
40000	0.796	7332	0.808	6229
45000	0.799	7758	0.807	6982
50000	0.804	8070	0.806	7758

Table 5. Comparing the combination 2 with SV and random distribution

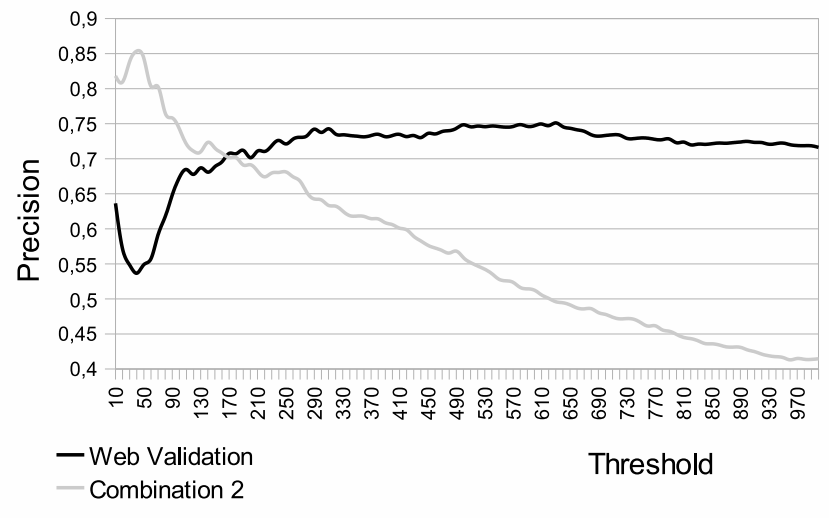

Fig. 8. Lift curve comparing Web Validation and Combination 2 approaches

However, the difference is in the number of positive relations (i.e. the number of induced syntactic relations found in the Validation Corpus) obtained with both approaches. The 'classical' second combination approach gets almost twice the number of relations than obtained by the 'random' one, while maintaining the same ratio of true positives over false positives. This means that the linguistic information is truly helpful since it noticeably gathers more 'good' relations. This effect tends to diminish when augmenting the threshold.

3.3. Discussion

Quality. The results given by the ROC curves are a relevant indication measuring the quality of the presented approaches. But this evaluation criterion does not give indications about precision. The likelihood of the first induced syntactic relations is an important feature for building of conceptual classes. Actually we can considerably reduce the expert task by giving an expert the first correctly ranked induced syntactic relations. Thus we propose to compute precision of both WV and second combination approaches scores for the first $1,000$ induced syntactic relations. We define precision as the number of positive syntactic relations divided by the total number of syntactic relations.

Figure 8 shows lift curves (precision as a function of the number of syntactic relations) obtained (threshold at $1,000$). The lift curve provides a global view of the precision. This figure shows the regularity of the precision, whatever the threshold, with the WV approach (scores about 0.70 and 0.75). For a threshold lower than 200, precision is quite interesting in the second combination because best syntactic relations are placed at the top of the sorted list. These results confirm our previous conclusion: WV is adapted to medium and large thresholds and second combination is adapted to small thresholds.

French Induced Syntactic Relations	*Rank with Comb. 2*	*Rank with VW*	*Positive / Negative*
mettre_note	0	7	+
éviter_récession	50	642	+
établir_sorte	100	1132	+
connaître_décision	150	1645	+
nommer_entraîneur	200	2142	-
réaliser_retour	250	2639	+
suspendre_traitement	300	3314	+
maintenir_ouverture	350	3854	+
accepter_déclaration	400	4452	+
provoquer_séparation	450	4878	-
gérer_ouverture	500	5424	-

Table 6. Comparing rank of WV and Combination 2 position

When focusing on the first sorted induced syntactic relations with the second combination, the next step was to determine if the first syntactic relations sorted by it are the same relations found by WV. Table 6 shows an extract of the first 500 relations sorted by the second combination, and the respective rank with the WV. Results show that the first ranked induced syntactic relations with the second combination are not Web 'popular' relations. Actually, the syntactic relations found with the second combination are not frequently used in the Web (i.e. because they are not at the top of the list returned by the WV approach). Two possible explanations: Either these relations are too 'literary', and the Web mostly deals with everyday language, or they are effectively new information, and corpora may contain knowledge nuggets.

The Minimum Size of the Validation Corpus in the Experimental Protocol. Quality has been discussed in the previous paragraph. Here, the experimental protocol needs to be considered. A second corpus (named V) is used to validate relevant induced syntactic relations. A question arises: *Does the Validation Corpus size impact quality*? In other words, we needed to find the required minimum size of corpus V to produce a correct AUC. **A Correct AUC** is defined as the AUC performed with the entire corpus V. Thus, the Validation corpus is split in n sub-sections. Then a validation of the induced syntactic relations of the Test corpus is performed with all sub-sections of the Validation corpus, by computing a cross validation. In other terms, for $n = 1000$, we make 1000 experiments to compute the average AUC. $n = 1000$ is similar to a corpus size of $50,000/1000$, that is, 50 average articles in the Validation corpus. When $n = 1$, then the entire Validation corpus is considered.

Figure 9 presents the AUC obtained for $n \in \{1; 10,000\}$. Here, only results obtained with the WV approach (for the 4 statistical measures) are presented, with a threshold of $50,000$. These experiments show the possibility to reduce the size of the Validation corpus by $4,000$. Actually, for a threshold included in $n \in \{1; 4,000\}$ the resulting AUC are similar, more or less 2%. After $n = 5,000$ scores decrease. This is caused by the small number of covered relations (the number of syntactic relations found in the Validation corpus) which is not enough to compute a correct AUC.

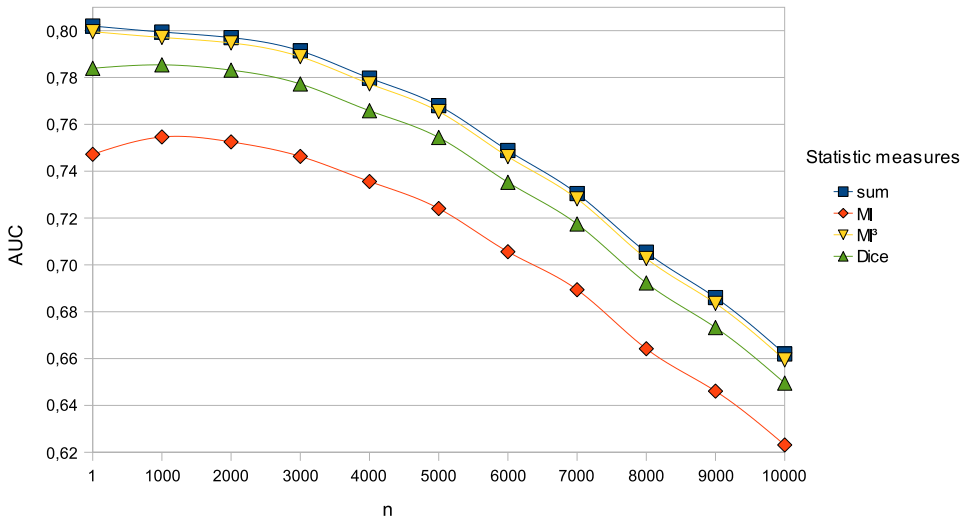

Fig. 9. AUC obtained for different statistical measures depending of the size of the corpus V $(50,000/n)$

Table 7 presents the average of the number of covered relations according to the size of the Validation corpus (size of the corpus is the size of the original Validation corpus divided by the n parameter). With a n value lower than $4,000$, AUC is very similar to scores obtained with the original Validation corpus. Thus, in the Table 7, we see that only 5 syntactic relations covered are required to apply our automatic evaluation protocol. The Figure 10 presents the AUC obtained with the statistical measures used in WV approach. We confirm that for $n \in \{1; 4,000\}$, AUC are similar, and for n score upper than $4,000$ AUC is decreased. We also show in this table that whatever the n value, the rank defined by the statistical measure is respected.

4. Conclusion

We present in this paper different approaches in order to reduce the human involvement in producing and validating new knowledge, either to build conceptual classes (domain ontologies, terminology networks) or to produce indexes, and discover knowledge in texts. We focus on the automatic validation of syntactic relations called induced relations, which are new combinations of text segments. These induced relations are not originally present in a corpus. An induced relation is built by assigning the object of a verb, and both object and verb are extracted from a corpus, to another verb, which is determined as semantically close. *This assignment is seen as an assumption, and the likelihood of such a combination is questioned*. Since several of these relations might appear as unlikely for many possible reasons, this paper describes different computational approaches to rank them.
The first considered approach is the Contextual Semantic Vectors (SV) approach. It claims

n	*Found relations*
1	*8268*
1000	*20*
2000	*10*
3000	*6,6*
4000	*5*
5000	*4*
6000	*3,33*
7000	*2,8*
8000	*2,5*
9000	*2,2*
10000	*2*

Table 7. Number of syntactic relations found according to the size of corpus V $(50,000/n)$

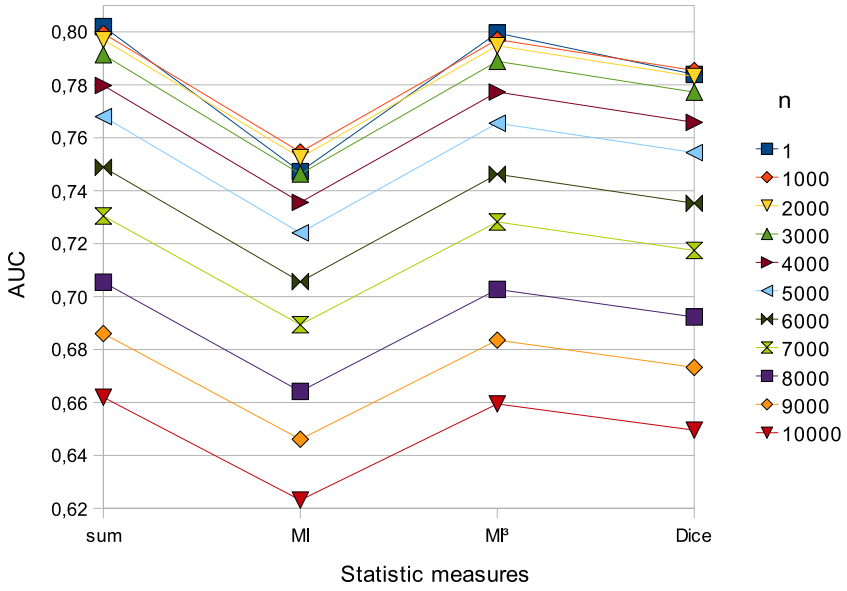

Fig. 10. AUC obtained depending of different statistical measures used with the Web Validation approach

to rely on linguistic information (lexical semantics, syntactic dependency determination in sentences, sentence semantics seen as a function of word semantics modified by syntactic roles). It represents words, and more generally, terms, as combinations of concepts defined in Roget-like thesaurus. It proposes to compute the semantics of any portion of texts with vector operations, and provides a contextual word vector, representing the word 'semantics' when the latter is waded in a given corpus acting as its context. The semantic closeness between a verb and a possible object (in an induced relation) is calculated with measures such as cosine.

The second approach is a WV approach which consists in querying the Web with induced syntactic relations. Statistical measures such as *Sum*, the *Mutual Information*, the *Cubic Mutual Information*, and the *Dice's Coefficient* are used to sort results given by a search engine (Yahoo API).
Both approaches rely on different claims: SV try to stick to linguistic consistency, WV relies on popularity, and frequency in a large amount of produced discourse as a clue for validity. Each one has its weaknesses and strengths. Therefore, we tried to combine both approaches within two possible procedures: One adding both results with a parameter affecting each method, and the other, using the first (SV) as a prime filter, and the second (WV) as a ranking function on a filtered list from which linguistically aberrant associations have been rejected.

To evaluate such methods, we chose to define an automatic protocol by opposition to a human evaluation. Actually the important number of induced syntactic relations results in too many potential tests to be manually performed. The proposed protocol considers an induced syntactic relation as **likely** if this relation exists in another corpus of the same field. We show in section 3.3 that a small corpus of an average of only five syntactic relations found can be used. A ROC Curves-based score, with the AUC, is computed to evaluate the different ranking methods quality. Various thresholds have been experimented in order to propose our validation procedure for different tasks. Actually for the ontology acquisition improvement task by using new knowledge (i.e. the induced syntactic relations), only a quality ranking of the first relations is required.

With small thresholds (i.e. a small number of induced syntactic relations) the second combined approach is the most accurate one. Otherwise, the best AUC are obtained with the WV method with the MI^3 and Sum statistical measures.
One of the biases introduced by our interpretation of the ROC setting within this frame is that it defines likelihood as positively assessed by the relation appearance in a corpus chosen as a reference (the V corpus). A likely relation might not appear in another corpus, if this corpus is randomly chosen. We tried to study the impact of a decreasing size, but not of an increasing one, and not with another corpus. The underlying idea according to which any randomly chosen corpus might act as a reference corpus tends to transform some possibly linguistically grounded relations into unlikely ones, which restricts the scope of discovered knowledge. This type of measure anyway favors large data sets, and therefore is biased in favor of WV, which shares with it the same rationale.

We consider evaluating the quality of this experimental protocol by comparing the human evaluation of induced syntactic relations quality results with the protocol results, as a future work. One of the possible future directions would be to let humans focus on those relations suggested by the combined approach (SV + WV in second combination) but rejected by the ROC AUC.

As an improving process, we might also use the methods presented in this paper to enhance the ExpLSA approach [1]. ExpLSA enables context expansion to improve document classification tasks. This approach consists in making expansion by using syntactic Verb-Object relations extracted from a corpus. The goal should be to use induced relations in expansion process.

References

1. Béchet, N., Roche, M., Chauché, J.: How the ExpLSA approach impacts the document classification tasks. In: Proceedings of IEEE International Conference on Digital Information Management (ICDIM). pp. 241–246 (2008)
2. Béchet, N., Roche, M., Chauché, J.: Towards the selection of induced syntactic relations. In: Proceedings of ECIR, LNCS, Springer. pp. 786–790 (2009)
3. Bourigault, D.: UPERY : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus. In: Actes de TALN, Nancy. pp. 75–84 (2002)
4. Bronzi, M., Guo, Z., Mesquita, F., Barbosa, D., Merialdo, P.: Automatic evaluation of relation extraction systems on large-scale. In: Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction. pp. 19–24. AKBC-WEKEX '12 (2012)
5. Chauché, J.: Un outil multidimensionnel de l'analyse du discours. In: Proceedings of Coling, Standford University, California. pp. 11–15 (1984)
6. Church, K.W., Hanks, P.: Word association norms, mutual information, and lexicography. In: Proceedings of the 27th Annual Conference of the Association of Computational Linguistic. vol. 16, pp. 76–83 (1989)
7. Cilibrasi, R., Vitanyi, P.M.B.: The google similarity distance. IEEE Transactions on Knowledge and Data Engineering 19, 370 (2007), http://www.citebase.org/abstract?id=oai:arXiv.org:cs/0412098
8. Daille, B.: Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques. Ph.D. thesis, Univ. Paris 7 (1994)
9. Daille, B.: Study and Implementation of Combined Techniques for Automatic Extraction of Terminology. In: P. Resnik and J. Klavans (eds). The Balancing Act: Combining Symbolic and Statistical Approaches to Language, MIT Press. pp. 49–66 (1996)
10. Fabre, C., Bourigault, D.: Extraction de relations sémantiques entre noms et verbes au-delà des liens morphologiques. In: TALN'06, 10-13 avril 2006. pp. 121–129 (2006)
11. Faure, D.: 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 2000)
12. Ferri, C., Flach, P., Hernandez-Orallo, J.: Learning decision trees using the area under the ROC curve. In: Proceedings of ICML'02. pp. 139–146 (2002)
13. Furstenau, H., Rambow, O.: Unsupervised induction of a syntax-semantics lexicon using iterative refinement. In: Proceedings of Joint Conference on Lexical and Computational Semantics. pp. 180–188 (2012)

14. Keller, F., Lapata, M.: Using the web to obtain frequencies for unseen bigrams. Comput. Linguist. 29(3), 459–484 (Sep 2003), http://dx.doi.org/10.1162/089120103322711604
15. Larousse, T.: Thésaurus Larousse - des idées aux mots, des mots aux idées. Ed.Larousse, Paris (1992)
16. L'Homme, M.C.: Le statut du verbe en langue de spécialité et sa description lexicographique. In: Cahiers de Lexicologie 73. pp. 61–84 (1998)
17. Lin, D.: Extracting collocations from text corpora. In: In First Workshop on Computational Terminology. pp. 57–63 (1998)
18. Pantel, P., Ravichandran, D.: Automatically labeling semantic classes. In: HLT-NAACL. pp. 321–328 (2004)
19. Prince, V., Chauché, J.: Building a Bilingual Representation of the Roget Thesaurus for French to English Machine Translation. In: LREC'08: Sixth International Language Resources and Evaluation Conference. pp. 438–446 (2008)
20. Roche, M., Kodratoff, Y.: Pruning terminology extracted from a specialized corpus for cv ontology acquisition. In: OTM Workshops (2). pp. 1107–1116 (2006)
21. Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Commun. ACM 18(11), 613–620 (1975)
22. Smadja, F., McKeown, K.R., Hatzivassiloglou, V.: Translating collocations for bilingual lexicons: A statistical approach. Computational Linguistics 22(1), 1–38 (1996)
23. Thanopoulos, A., Fakotakis, N., Kokkianakis, G.: Comparative Evaluation of Collocation Extraction Metrics. In: Proceedings of LREC'02. vol. 2, pp. 620–625 (2002)
24. Turney, P.: Mining the Web for synonyms: PMI–IR versus LSA on TOEFL. In: Proceedings of ECML'01, Lecture Notes in Computer Science. pp. 491–502 (2001)
25. Turney, P.D., Pantel, P.: From frequency to meaning: Vector space models of semantics. J. Artif. Intell. Res. (JAIR) 37, 141–188 (2010)
26. Vivaldi, J., Màrquez, L., Rodríguez, H.: Improving term extraction by system combination using boosting. Proc of ECML, LNCS, 2167, 515–526 (2001)
27. Weeds, J., Weir, D., McCarthy, D.: Characterising measures of lexical distributional similarity. In: IN COLING-04. pp. 1015–1021 (2004)
28. Wermter, J., Hahn, U.: Collocation extraction based on modifiability statistics. In: COLING '04. p. 980. Association for Computational Linguistics, Morristown, NJ, USA (2004)
29. Yan, L., Dodier, R., Mozer, M., Wolniewicz, R.: Optimizing classifier performance via an approximation to the Wilcoxon-Mann-Whitney statistic. In: Proceedings of ICML'03. pp. 848–855 (2003)

Nicolas Béchet is Associate Professor (Maître de Conférences) at Université de Bretagne-Sud, France. He obtained his PhD in Computer Sciences from Université Montpellier 2 in 2009. His main research domains are Text Mining and NLP, and he leads the Text topic of the EXPRESSION team at IRISA laboratory. He is involved in many academic and industrial projects and has supervised many students in master degree. He published more than 20 papers in international conferences and journals.

Jacques Chauché is pensioner full Professor. His main research domain is Natural Language Processing. He is member of the GETA team (Groupe d'étude pour la traduction automatique Grenoble) and obtained his "Doctorat d'état" in 1975 from the University Joseph Fourier (Grenoble). He worked in the CELTA team (Centre d'étude pour la traduction automatique, University of Nancy 1) and moved from university of Le Havre to

University of Montpellier 2 in 1992. He developed the Sygmart system used for the french syntaxico-semantic analyzer Sygfran. He published many papers about this in numerous international conferences of which Coling.

Violaine Prince is full Professor at Montpellier 2 University in Computer Science since 2000. Her main research domain is Natural Language Processing. She had led the NLP research team at LIRMM, the Computer Science Lab in Montpellier for 10 years, as well as the CS Reasearch Department (2010-2012), and the CS Teaching Department (2003-2006). She obtained her PhD in 1986 from University Denis Diderot (Paris 7), then her HDR in 1992 from University Paris 11 (Orsay). She moved from Paris 11 to Ecole Normale de Cachan as an associate professor, then to Paris 8 as a full professor where she co-founded a CS lab and Distant Learning curricula in CS, and was elected president of the University National Council for Computer Science. She had supervised 24 PhD Students, participated to European and international research projects, and led a few national academic and industrial projects. Violaine Prince published and edited several books, counts more than 100 papers in major international conferences and journals, and also chaired several conferences.

Mathieu Roche is CIRAD Research Scientist (PhD, HDR) at TETIS lab. Between 2005 and 2013, he has been Associate Professor (Maître de Conférences) at the University Montpellier 2, France. Mathieu Roche obtained a PhD in Computer Science from University Paris XI (Orsay) in 2004. He defended his HDR (Habilitation à Diriger des Recherches - Accreditation to supervise research) in 2011. Mathieu Roche led several academic and industrial projects in text-mining. He has supervised 7 PhD students. Mathieu Roche published more than 120 papers in major international conferences and journals.

Received: April 22, 2013; Accepted: November 22, 2013.