A Hybrid Approach to Validate Induced Syntactic Relations

Nicolas Béchet
bechet@lirmm.fr

Mathieu Roche
mroche@lirmm.fr

Jacques Chauché
chauche@lirmm.fr

LIRMM - UMR 5506, CNRS
Univ. Montpellier 2
34392 Montpellier Cedex 5 - France

Abstract

We propose in this paper to use NLP approaches to extract and validate induced syntactic relations (Verb-Object). We employ syntactic parser and a semantic proximity measure to extract them. Then, we focus on a Web Validation system, a Semantic-Vector-based approach, and finally we propose approaches to combine both in order to rank induced syntactic relations. The Semantic Vectors 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. The systems combine Web Validation approach and Semantic Vectors technique. We apply our approaches on corpus of news, using ROC curves to evaluate the results.

1 Introduction

The semantic knowledge acquisition is an important problem in Natural Language Processing (NLP). This knowledge can be used for information retrieval and/or classification tasks. Many other NLP applications employ semantic knowledge as automatic translation or indexing.

We can use syntactic informations to build semantic knowledge [10]. Then we can use them to make conceptual classes (gathering of words). For instance, the words *house*, *hangar*, and *farmhouse* can be gathered in a concept *construction*. Actually, these concepts can be hierarchically organized to build a conceptual classification. We focus in this paper on the building of semantic knowledge by using syntactic relations. Actually, syntactic relations are relevant features of fields [5]. Two kinds of syntactic relations can be used to build concepts. First, we have the **original** relations which can be extracted by a syntactic parser [16], [22]. The second kind of relations is based on the **induced** syntactic relations. They are introduced by the ASIUM system [11]. The ASIUM system consists in gathering the objects of verbs considering as close by a quality measure. For example, in figure 1, if the verbs "to consume" and "to eat" are close, we can gather objects "fuel, vegetable, food, and fruit", which are obtained by syntactic informations. Other approaches of the literature gather terms by using proximity measures as [3].

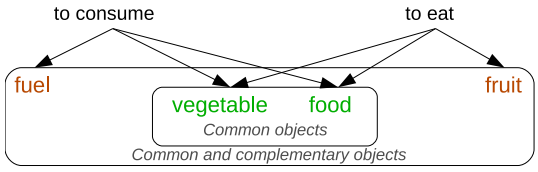

Figure 1. Common and complementary objects of verbs "to consume" and "to eat"

Globally, we consider two verbs V_1 and V_2 as close if they have a lot of common objects. Let $Obj_1^{V_1}...Obj_n^{V_1}$ and $Obj_1^{V_2}...Obj_m^{V_2}$ the objects of the verbs V_1 and V_2, $Obj_i^{V_1}$ ($i \in [0,n]$) 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, in the figure 1 the induced relations are *to eat fuel* and *to consume fruit*. Note that these induced syntactic relations represent new knowledge because they are not present in the initial corpus. The quality of the conceptual classes depends on the quality of the induced relations (i.e. quality of the complementary objects).

Our aim in this paper is to extract relevant induced syntactic relations to limit the task of the expert (e.g. *to eat fuel*: Irrelevant vs *to consume fruit*: Relevant). Thus we propose to rank these relations by using different approaches. We propose to compare two approaches and finally combine both. The first approach represents a syntactic relation as a combination of different concepts based on a thesaurus. The second approach consists in using statistical measures

based on the Web by considering the number of pages provided by a search engine.

We describe in the next section our validation approaches which are used to rank induced syntactic relations. The section 3 presents experimental results obtained by using a specific experimental protocol. The relevance of the automatic protocol are then discussed. We finally conclude in the section 4.

2 The ranking approaches

2.1 The Syntactic Verb-Object Relations

We present in this section how we get the induced syntactic relations from a corpus. We firstly use the SYGFRAN parser [6] to extract Verb-Object syntactic relations. Next, we evaluate the semantic proximity between verbs of extracted syntactic relations Verb-Object. To compute this proximity, we use the Asium measure [12]. This measure considers two verbs as close if they have a significant number of mutual features (objects). More globally, we consider the p and q verbs with their respective p_1,...,p_n and q_1,...,q_n objects. $NbOC_p(q_i)$ is the occurrence number of q_i object from q and p verbs (common objects). $NbO(q_i)$ is the occurrence number of q_i objects of q verb. Then, the Asium measure is:

$$Asium(p,q) =$$

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

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

We finally conserve the complementary objects of the closest verbs (determined by the Asium measure). Thus, we obtain a list of induced syntactic relations based on the complementary objects.

2.2 The Semantic Vectors approach

We present in this section our semantic vector approach, used to evaluate the semantic proximity between a verb and an object of a syntactic relation.
[23] argues that different features of a classic thesaurus like Roget can be relevant to NLP tasks. Many Roget-based usages are performed in some different areas of NLP (e.g. Word-Sense Disambiguation [24], Information Retrieval [4], Text Cohesion [18], Text Classification [7], or to determine Semantic Similarity [17], [14]).
[14] uses the taxonomic structure of the Roget's Thesaurus to determine semantic similarity. They consider two words as close if they occur in the same semicolon group in the thesaurus. Word pairs of the same paragraph are close. Word pairs of different paragraphs, which have the same part-of-speech tag and within the same category are quite close. They obtain better results than usual measures as LSA [15] or PMI-IR [20] by giving the correct answer of TOEFL, ESL, and Reader's Digest tests.
Our approach proposes to use a Roget-based approach as similarity measure in a different context.

Actually, we measure the semantic proximity between a verb and an object of an induced syntactic relation. This one is compared with the verb and the object from the original syntactic relation. With the figure 1, we measure the semantic proximity between syntactic relations "to consume – fruit" (induced relation) and "to eat – fruit" (real relation in the corpus). The goal is to obtain a score for each induced syntactic relation. This score enables to rank all the relations. We chose to represent a syntactic relation as a semantic vector. A vector is built by representing the terms in a space based on 873 dimensions. This space is organized as a conceptual ontology defined in [1]. Each term is indexed by one or many elements of this ontology. For example, "to consume " is relative to "nutrition, education, meal, and bread". The result of the semantic vector of the syntactic relation "to consume food" is given in figure 2.

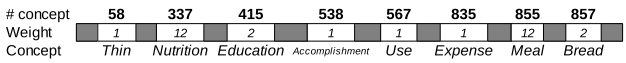

Figure 2. The semantic vector representation of "to consume food"

This vector is a linear combination of the vector representation of "to consume" and "food". The weights take into account the syntactic structure (in our case, a verb and the object) [6]. To measure the relevance of an induced syntactic relation, we evaluate if this relation shares the same concepts with the original syntactic relation. Thus, we compare both vectors using the cosine measure. Cosine is the computation of the scalar product of both vectors divided by the norms product of both vectors. Our aim is to obtain a ranking function based on this cosine measure to validate induced syntactic relations. An example of the usage of the semantic vectors approach is presented section 2.5.

2.3 The Web validation approach

We propose to use the Web to measure the dependence between a verb and an object of an induced syntactic relation. Finally we rank the syntactic relations by its "Web popularity". Thus, we query the Web with a syntactic relation represented by a string (for instance the French relation "consommer un fruit"). The query is given by the $nb(x)$ function which is the number of pages provided by the search engine Yahoo by using an API (http://api.search.yahoo.com). In French, language used in our work, a verb and its object is 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 our 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, we apply a statistic measure to compute the dependency between the verb v and the object o from an induced syntactic relation. Works of the literature using ranking functions ([21], [9]) have estimated that MI^3 (Cubic Mutual Information) was the best behaving measure. We propose to apply this measure to evaluate the quality of our extracted induced syntactic relations. The MI^3 is an empirical measure based on Mutual Information (MI [8]), that enhances the impact of frequent co-occurrences, something which is absent in the original MI [9]. We define the MI^3 as following:

$$MI^3(v, o) = \log \frac{nb(v, o)^3}{nb(v)nb(o)} \quad (1)$$

This measure enables to obtain a score to rank all the syntactic relations. An example of the Web validation approach is presented section 2.5.

Let us note that we experiment the Web validation approach on 60,000 induced syntactic relations (section 3). Then we need 420,000 queries by using MI^3 (1 for the verb, 1 for the object and 5 for an induced syntactic relation: 60,000 $\times$ 7 = 420,000. Thus, the Web validation approach is time-costly.

2.4 The combinations

To exploit the performance of Semantic Vectors (SV) and Web Validation (WV) approaches, we propose to compute a combined system of both methods.

2.4.1 Combination 1: A combined system with a variant scalar

We propose in the first combined approach to introduce a scalar $k \in [0, 1]$ to reinforce one or another approaches. The results obtained with SV and WV methods are first normalized. Next, we combine both results of SV and WV with the following formula:

For a syntactic relation c,
$combine\_score_c = k \times SV + (1 - k) \times WV$

2.4.2 Combination 2: An adaptive combined system

We present a second combined system between SV and WV approaches: The **adaptive combined system**. First we rank syntactic relations with Semantic Vectors (SV). Finally we rank the n first syntactic relations (obtained with SV) with the Web Validation technique. This second process (WV applied on the ranked relations with SV) enables to accurately sort these n syntactic relations. Thus, with this adaptive combination, SV proposes a global selection using semantic resources, and WV manages this first selection.

2.5 An example with five syntactic relations

French relations		English translation	
Induced	Regular	Induced	Regular
lancer recherche	mener recherche	launch research	conduct research
poursuivre réforme	demander réforme	pursue reform	ask reform
réussir évaluation	faire évaluation	succeed evaluation	make evaluation
dépasser recherche	faire recherche	exceed research	do research
dire croisade	poursuivre croisade	say crusade	continue crusade

Table 1. The five syntactic relations used

We present in this section an example of our different approaches previously described, the Semantic Vectors, the Web Validation, and the combined approaches. We use five French syntactic relations translated in table 1. This table shows induced syntactic relations and the regular syntactic relations (Verb with the object found in a corpus).

Verb-Object relations		Cosine
Induced	Regular	
pursue reform	ask reform	0.60
exceed research	do research	0.52
succeed evaluation	make evaluation	0.41
say crusade	continue crusade	0.37
launch research	conduct research	0.27

Table 2. The Semantic Vector approach

First we compute the Semantic Vector approach. We represent the induced syntactic relation and the regular syntactic

relation by semantic vector with SYGFRAN. Then we compute the cosine. The results are given in Table 2. Next, we

Verb-Object	nb(Verb)	nb(Object)	nb(Verb, Object)	MI³
launch research	82,700,000	863,000,000	2,299,288	0.71
pursue reform	46,200,000	39,000,000	45,914	0.49
say crusade	370,000,000	4,120,000	72	0.41
succeed evaluation	27,600,000	57,900,000	1,366	0.35
exceed research	15,900,000	863,000	363	0.28

Table 3. The Web Validation approach

query the Yahoo search engine to apply the Web Validation approach. We are querying the Web for the verbs, the objects, and the syntactic relations. Then we can compute the MI^3. The results are given Table 3.

Verb-Object relations	SV	WV	Combination 1	Combination 2
launch research	0,60	0,49	0,55	1,49
pursue reform	0,41	0,35	0,38	1,35
succeed evaluation	0,52	0,33	0,43	1,33
exceed research	0,37	0,13	0,25	0,37
say crusade	0,27	0,71	0,49	0,27

Table 4. Scores of combined approaches

We normalize results as shown in Tables 2 and 3 to compute the first combined system. We chose to compute the first combined system with $k = 0.5$, and we define a threshold to 3 for the second combined system.

SV	WV	Combination 1	Combination 2
pursue reform	launch research	pursue reform	pursue reform
exceed research	pursue reform	launch research	succeed evaluation
succeed evaluation	succeed evaluation	exceed research	exceed research
say crusade	exceed research	succeed evaluation	say crusade
launch research	say crusade	say crusade	launch research

Table 5. The ranked syntactic relations

The results for both combined approaches are given in Table 4. Table 5 presents the ranking of syntactic relations for all computed measures in this example.

3 Experiments

3.1 Exerimental protocol

To measure the quality of induced syntactic relations extracted from a first corpus, we use a second corpus. The first one is a corpus from Yahoo's site (http://fr.news.yahoo.com/): 8,948 news (16.5 MB). It is used as test corpus. We called it *corpus T*. The second one (corpus V) is used as a validation corpus. V comes from the French newspaper *Le Monde*. It contains more than 60,000 news (123 MB). We need it to determine if Induced Verbal Syntactic Relations (IVSR) of corpus T are relevant. The corpora T and V come from the same field. Thus, our aim is to automatically recover IVSR of corpus *T* that exist in corpus *V*. If an IVSR of corpus T appears in corpus *V*, we consider it as **positive**, else it is **negative**. We use different approaches presented in section 2 (*Semantic Vectors*, *Web Validation*, and the *Combined Systems*) to rank induced syntactic relations. To measure the quality of the obtained ranking, we use the ROC curves.

Initially the ROC curves (Receiver Operating Characteristic), detailed in [13], come from the field of signal treatment. ROC curves are often used in the field of medicine to evaluate the validity of diagnostic tests. The ROC curves show in X-coordinate the rate of false positive (in our case, rate of irrelevant induced syntactic relations) and in Y-coordinate the rate of true positive (rate of relevant 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. In the case of IVSR ranking, a perfect ROC curve corresponds to obtain all relevant IVSR at the beginning of the list. This situation corresponds to $AUC = 1$. The interest of this measure is developed in [19]. Note that an IVSR could be away in the validation corpus but it does not mean that this verbal syntactic relation is really irrelevant. We choose this validation to have an automatic validation based on a large amount of data.

3.2 Results

We present results for different thresholds. The goal of our experiments is to have all positive relations at the top of the list.

Threshold	*WV*	*SV*	*Threshold*	*WV*	*SV*
5000	0,61	0,51	*35000*	0,75	0,55
10000	0,65	0,52	*40000*	0,76	0,56
15000	0,68	0,54	*45000*	0,78	0,55
20000	0,70	0,54	*50000*	0,79	0,54
25000	0,72	0,55	*55000*	0,80	0,54
30000	0,74	0,55	*60000*	0,81	0,54

Table 6. AUC obtained with the Semantic Vectors and the Web Validation approaches

Table 6 presents AUC with different thresholds using the Semantic Vectors and the Web Validation approaches. Results obtained with Semantic Vectors are poor, very close of a random distribution (AUC=0.5). This unsatisfactory results could be explained by the nature of the Semantic Vectors. Actually, Semantic Vectors are composed of 873 general concepts which are not always adapted to measure the quality of IVSR. The Web Validation approach gives better

Threshold	*k* = 0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
5000	0,61	0,61	0,62	0,62	0,63	0,63	**0,66**	0,55	0,56
10000	0,66	0,66	0,66	0,67	0,67	**0,67**	0,67	0,64	0,57
15000	0,68	0,68	0,68	0,69	0,70	0,70	**0,71**	0,65	0,56
20000	0,70	0,71	0,72	0,73	**0,74**	0,73	0,71	0,66	0,58
25000	0,73	0,75	0,75	0,75	**0,76**	0,75	0,73	0,69	0,62
30000	0,76	0,77	0,78	0,78	**0,78**	0,77	0,75	0,71	0,63
35000	0,78	0,79	**0,79**	0,79	0,78	0,77	0,75	0,72	0,64
40000	0,79	**0,79**	0,79	0,79	0,78	0,77	0,75	0,71	0,65
45000	**0,79**	0,79	0,79	0,79	0,79	0,78	0,76	0,73	0,67
50000	**0,80**	0,80	0,79	0,78	0,75	0,74	0,72	0,69	0,64
55000	**0,80**	0,79	0,78	0,75	0,73	0,71	0,69	0,66	0,62
60000	**0,79**	0,78	0,76	0,74	0,72	0,70	0,68	0,65	0,61

Table 7. AUC obtained with the first combined system

Threshold	*WV*	*Comb. 1*	*Comb. 2*
5000	0,61	**0,66**	**0,83**
10000	0,65	**0,67**	**0,82**
15000	0,68	**0,71**	**0,83**
20000	0,70	**0,71**	**0,83**
25000	0,72	**0,73**	**0,83**
30000	0,74	**0,75**	**0,83**
35000	0,75	0,75	**0,83**
40000	0,76	0,75	**0,83**
45000	0,78	0,76	**0,83**
50000	0,79	0,72	**0,82**
55000	0,80	0,69	**0,82**
60000	0,81	0,68	**0,81**

Table 8. AUC obtained with the combinations 1 and 2

results. For the first syntactic relations (small thresholds) the AUC are unsatisfactory. The Web Validation approach is efficient for a large amount of syntactic relations. But our aim is to obtain good results with small thresholds. This process enables to extract a limited number of syntactic relations in order to evaluate syntactic relations by an expert.

Thus, we propose to apply the first combined system. The AUC obtained are given in Table 7. We propose to experiment the parameter $k \in [0.1...0.9]$ with an increment of $1/10$. We do not report the $k = 0$ results (Web Validation) and $k = 1$ results (Semantic Vectors). The results are particularly interesting for small values of the threshold with $k = 0.6$ or $k = 0.7$.

Now we apply the second combination presented in section 2.4.2. For example, for a threshold at $10,000$, we first sort all syntactic relations with the Semantic Vectors approach and then, we sort these $10,000$ first syntactic relations using the Web Validation method. This approach improves all previous obtained AUC (Table 8). For instance with a threshold at $5,000$, the AUC of the second combination is 0.83 vs. 0.66 for the the first combination and 0.61 for the web validation (Table 8).

3.3 Discussion

The results given by the ROC curves are a well indication to measure the quality of our presented approaches. But this evaluation criterion does not give indications about the precision of our approaches. We define precision as the number of positive syntactic relations divided by the number of syntactic relations.

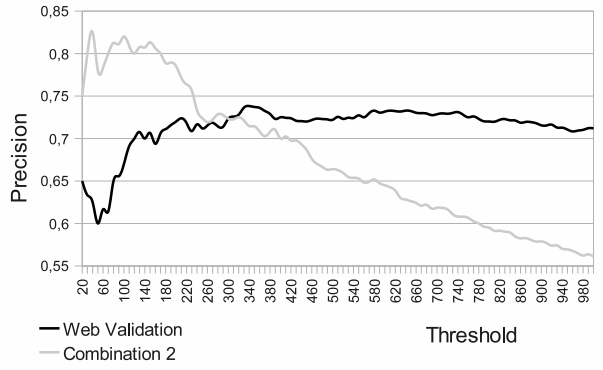

Figure 3. Lift curve comparing Web Validation and Combination 2 approaches

We present in figure 3, lift curves (precision in function of the number of syntactic relations) for the first $1,000$ syntactic relations using different approaches (threshold at $1,000$). The lift curve provides a global view of the precision. For a threshold lower than 300 (between 270 and 320 in figure 3), the precision is very interesting with the second combination. It means that the best syntactic relations are placed at the top of the ranked list. These global results confirm the AUC results (Table 8). We want to know if the first syntactic relations ranked by the second combination are the same relations found by the Web Validation approach. For instance, for the *300* first relations found by the second combination, *37 (12 %)* are ranked between $[0; 300[$ with WS, *39 (13 %)* between $[300, 600[$, *39 (13 %)* between $[600, 900[$, and *186 (62 %)* are beyond. Thus, the syntactic relations found with the second approach are not frequently used in the Web (i.e. because they are not at the top of the list returned by the Web Validation approach). Thus, we discover new knowledge nuggets from a corpus.

4 Conclusion

An induced syntactic relation is a syntactic relation which are not initially present in a corpus. To extract them

we measure the verb proximity. These syntactic relations can be used for example to improve the ontology acquisition.

We present in this paper few approaches to validate and to rank candidate syntactic relations. The first one consists in representing syntactic relations by semantic vectors which represent terms as combination of concepts of the Larousse French thesaurus. Then, we measure the vector proximity with the cosine measure.
The second one is a Web Validation method-based. It consists in querying the Web with induced syntactic relations. We use the *Cubic Mutual Information* to sort results given by a search engine (Yahoo API). We finally combine both previous approaches with two methods.

We evaluate our results with the ROC curves measure and AUC. We obtain better AUC with a combination of Web Validation and Semantic Vectors (second combination). We discuss results with the precision measure to confirm that first syntactic relations given by the second combination are relevant. With this approach, we can for example improve the ontology acquisition tasks using new knowledge, given by induced syntactic relations. We consider as future work to use the methods presented in this paper to improve the ExpLSA approach [2]. ExpLSA enables to expand the context to improve document classification tasks.

References

[1] *Thésaurus Larousse - des idées aux mots, des mots aux idées*. Ed.Larousse, Paris, 1992.

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

[3] D. Bourigault. UPERY : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus. In *Actes de TALN, Nancy*, pages 75–84, 2002.

[4] R. Boyd, J. R. Driscoll, and I. Syu. Incorporating semantics within a connectionist model and a vector processing model. In *TREC*, pages 291–302, 1993.

[5] L. M. C. Le statut du verbe en langue de spcialit et sa description lexicographique. In *Cahiers de Lexicologie 73*, pages 61–84, 1998.

[6] J. Chauché. Un outil multidimensionnel de l'analyse du discours. In *Proceedings of Coling, Standford University, Califonnia*, pages 11–15, 1984.

[7] J. Chauché and V. Prince. Classifying texts through natural language parsing and semantic filtering. In *3rd International Language and Technology Conference, Poznan, Pologne*, 2007.

[8] K. W. Church and P. Hanks. Word association norms, mutual information, and lexicography. In *Computational Linguistics*, volume 16, pages 22–29, 1990.

[9] B. Daille. *Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques*. PhD thesis, Université Paris 7, 1994.

[10] C. Fabre and D. Bourigault. Extraction de relations smantiques entre noms et verbes au-del des liens morphologiques. In *TALN'06, 10-13 avril 2006*, pages 121–129, 2006.

[11] D. Faure. *Conception de mthode d'apprentissage symbolique et automatique pour l'acquisition de cadres de sous-catgorisation de verbes et de connaissances smantiques partir de textes : le systme ASIUM*. PhD thesis, Universit Paris-Sud, 20 Dcembre 2000.

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

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

[14] M. Jarmasz and S. Szpakowicz. Roget's thesaurus and semantic similarity. In *Conference on Recent Advances in Natural Language Processing*, pages 212–219, 2003.

[15] T. Landauer and S. Dumais. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104(2):211–240, 1997.

[16] D. Lin. Extracting collocations from text corpora. In *In First Workshop on Computational Terminology*, pages 57–63, 1998.

[17] M. L. McHale. A comparison of wordnet and roget's taxonomy for measuring semantic similarity. In *Proc COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems.Montreal, Canada*, pages 115–120, 1998.

[18] J. Morris and G. Hirst. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Comput. Linguist.*, 17(1):21–48, 1991.

[19] M. Roche and Y. Kodratoff. Pruning Terminology Extracted from a Specialized Corpus for CV Ontology Acquisition. In *Proceedings of onToContent'06 workshop - OTM'06, Springer Verlag, LNCS*, pages 1107–1116, 2006.

[20] P. Turney. Mining the Web for synonyms: PMI–IR versus LSA on TOEFL. *Lecture Notes in Computer Science*, 2167:491–502, 2001.

[21] J. Vivaldi, L. Màrquez, and H. Rodríguez. Improving term extraction by system combination using boosting. *Lecture Notes in Computer Science*, 2167:515–526, 2001.

[22] J. Wermter and U. Hahn. Collocation extraction based on modifiability statistics. In *COLING '04*, page 980, Morristown, NJ, USA, 2004. Association for Computational Linguistics.

[23] Y. Wilks. Language processing and the thesaurus. In *National Language Research Institute*, 1998.

[24] D. Yarowsky. Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. In *Proceedings of COLING-92*, pages 454–460, Nantes, France, July 1992.