Towards the Selection of Induced Syntactic Relations

Nicolas Béchet, Mathieu Roche, and Jacques Chauché

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

Abstract. We propose in this paper to use NLP approaches to validate induced syntactic relations. We focus on a Web Validation system, a Semantic Vector-based approach, and finally a Combined system. The Semantic Vector approach is a Roget-based approach which computes a syntactic relation as a vector. The Web Validation technique uses a search engine to determine the relevance of a syntactic relation. We experiment our approaches on real-world data set. The ROC curves are used to evaluate the results.

1 Introduction

This paper deals with the extraction of Verb-Object relations from textual data. But our approach is not based on "traditional" extraction in corpora because we discover induced Verb-Object relations (syntactic relations not present in the corpus). This knowledge can be used to enrich ontologies by adding relevant induced instances [6] or to expand contexts [1]. First, we introduce the "induced relations" term. The first step consists in extracting standard Verb-Object relation from a corpus with a syntactic parser [2]. We consider two verbs V_1 and V_2 as close if they have a lot of common objects [6]. 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**. For instance, with the relations *to consume vegetable*, *to consume food*, *to consume fuel*, *to eat vegetable*, *to eat food*, *to eat fruit*, 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. In order to determine which induced relations are relevant (i.e. *to eat fuel*: Irrelevant vs *to consume fruit*: Relevant), we propose to use ranking functions: Semantic Vectors approach (section 2.1), Web Validation approach (section 2.2), and Combined System (section 2.3).

2 Our approaches

The discovery of all the induced relations [1] based on the use of the Asium measure [6] is not the aim of this paper. Our approaches make possible the selection of **relevant relations** using ranking functions (i.e. the relevant relations have to be at the beginning of the lists).

2.1 The Semantic Vector approach

Many Roget-based usage are performed in different fields of NLP (e.g. Word-Sense Disambiguation, Information Retrieval, Text Cohesion, Text Classification, and so forth). For instance, the study of [7] uses the taxonomic structure of the Roget's Thesaurus to determine semantic similarity. Our approach proposes to use a Roget-based approach as a similarity measure. The first step of our approach is based on a vectorial representation of syntactic relations using SYGFRAN parser [2]. For the vector construction, each term is represented by a concept vector. These concepts come from a French thesaurus, the Larousse thesaurus (1992) which contains 873 concepts as Family, Evolution, Society, etc. A semantic vector of a syntactic relation Verb-Object is a linear combination of the concepts of the Verb and the concepts of the Object [3]. For instance, non-null components of the semantic vector based on the syntactic relation "to consume fruit" are relative to the Larousse concepts Thin, Nutrition, Education, Accomplishment, Use, Expense, Meal, and Bread. We compare induced relation based on the vector representation with existing relations. Then with the object *fruit* (example of the section 1), we compare the syntactic relations *to eat fruit* (real relation) and *to consume fruit* (induced relation) using their Semantic Vector representation. We compare the semantic vectors by the application of two different measures. The first one is the **cosine**. Cosine is the computation of the scalar product of both vectors divided by the norms product of both vectors. The second measure well adapted to the semantic vectors is the **matching distance** [3]. To compute the matching distance, first the difference between the most intense ranking components is calculated (ranking distance). Next, the intensity difference with the concepts is computed. The matching measure uses ranking distance, intensity difference, and the cosine measure (this measure is detailed in [3]).

2.2 The Web Validation approach

Our work using Web Validation are close to the Turney's approach [9]. The PMI-IR algorithm (Pointwise Mutual Information and Information Retrieval) described in [9] queries the Web via the AltaVista search engine in order to determine appropriate synonyms. This approach calculates the proportion of documents containing a word and its candidate synonym. In our case, we can apply an approach close to the Turney's method with other statistical measures described below. Then the dependence of the verbs (v) and the objects (o) is calculated for all the induced relations.

- One of the most commonly used measures to compute a sort of relationship between the words composing what is called a co-occurrence is Church's **Mutual Information** (MI) [4]: $MI(v, o) = \frac{nb(v,o)}{nb(v)nb(o)}$
- The **Cubic Mutual Information** is an empirical measure based on MI, that enhances the impact of frequent co-occurrences, something which is absent in the original MI [5]. Such as measure is defined by the following formula: $MI^3(v, o) = \frac{nb(v,o)^3}{nb(v)nb(o)}$

– The **Dice coefficient** is another interesting quality measure [8] which calculates the dependence of v and o: $Dice(v, o) = \frac{2 \times nb(v,o)}{nb(v)+nb(o)}$

In our work, we use the nb function which represents the number of pages provided by the search engine Yahoo. Our aim is to specify if a Verb-Object relation is relevant and popular in the Web. We consider five usual French articles *un, une, (i.e. a), le, la, l' (i.e. the)* to calculate the frequency nb. The nb value for a Verb-Object relation is: $nb(v, o) = nb(\text{"}v$ **un** $o\text{"}) + nb(\text{"}v$ **une** $\text{o"}) + nb(\text{"}v$ **le** $\text{o"}) +$ *etc.* Then we can evaluate nb, MI, MI^3, and *Dice* measures for all the syntactic relations in order to obtain ranked relations.

2.3 The Combined System

To exploit the performance of Semantic Vectors (SV) and Web Validation (WV) approaches, we propose to compute a combined system of both approaches. We use a parameter q to have the possibility to apply different weights to the normalized values obtained with SV and WV methods: $q \times SV + (1 - q) \times WV$

3 Experiments

In our experiments, we use two French corpora. The first one is a corpus extracted from Yahoo's site (http://fr.news.yahoo.com/). It contains 8,948 news (16.5 MB). It is used as a test corpus. We called it *corpus T*. The second one is used as a validation corpus. It is called *corpus V*. *V* comes from the French newspaper *Le Monde* (same field of corpus *T*). It contains more than 60,000 news (123 MB). We want to determine if induced relations of corpus T are relevant. Our aim is to evaluate the number of induced relations of corpus T that exist in corpus *V*. An induced relation of T that appears in V is considered as **positive**, else it is **negative**. We choose this method to have an automatic validation based on a large amount of data. We use the three approaches presented in section 2 (*Web Validation*, *Semantic Vectors*, and the *Combined system* approaches) to rank the induced relations (we consider the 12,000 first relations obtained with the Asium measure). To measure the quality of the obtained ranking, we use ROC curves. The ROC curves show in X-coordinate the rate of false positive (in our case, the rate of negative induced relations) and in Y-coordinate the rate of true positive. 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 the ranked syntactic relations, a perfect ROC curve provides all relevant relations at the beginning of the list and all irrelevant relations at the end. This situation corresponds to $AUC = 1$. We propose to evaluate the different thresholds (i.e. n first syntactic Verb-Object relations) of the ranking function. Table 1 presents AUC with different thresholds using the Semantic Vector approach (SV).[1] Table 1 shows that matching distance results are better than cosine. However, both

[1] $\beta = 0$ and $N = 0.5$ are applied for the matching distance described in [3]

results are poor, very close of a random distribution (i.e. AUC = 0.5). This unsatisfactory results could be explained by the nature of the Semantic Vectors. Actually, Semantic Vectors are composed of 873 concepts which could have an insufficient precision to rank our syntactic relations. The Web Validation (WV) approach gives better results than the Semantic Vector method (Table 1). For the first half of evaluated thresholds, the *Dice's measure* obtains the best results. On the other hand, MI^3 obtains best results in the second part. The three following ranking function $Frequency$, MI^3, and *Dice's measure* are very close with a small advantage for the MI^3 measure (by computing the average of different thresholds). The AUC obtained with the Combined System (section 2.3)

Threshold	**Semantic Vectors**		**Web Validation**			
	Angle	Match. Dist.	*Frequency*	*MI*	*MI³*	*Dice's*
1000	0,55	0,52	0,65	0,59	0,62	0,64
2000	0,55	0,53	0,68	0,62	0,68	**0,69**
3000	0,49	0,52	0,69	0,66	0,70	**0,70**
4000	0,51	0,53	0,70	0,67	0,71	**0,72**
5000	0,51	0,53	0,72	0,68	0,73	**0,73**
6000	0,52	0,52	0,74	0,69	**0,74**	0,74
7000	0,54	0,53	0,75	0,71	0,76	0,75
8000	0,53	0,53	0,76	0,72	**0,77**	0,76
9000	0,51	0,54	0,77	0,72	**0,78**	0,77
10000	0,51	0,54	0,79	0,73	**0,79**	0,78
11000	0,51	0,54	0,81	0,75	**0,81**	0,80
12000	0,52	0,55	0,82	0,77	**0,82**	0,81

Table 1. AUC obtained with the Semantic Vector and the Web Validation approaches

are given in Table 2. We propose to experiment the parameter $q \in [0, 1]$ with an increment of 0.1. $q = 0$ is equivalent to WV and $q = 1$ is equivalent to SV. When the Combined System favors the Semantic Vectors method (i.e. $q \in [0.8, 0.9]$) we obtain best results for few relations (small thresholds). The first thresholds based on the SV method (i.e. high value for q) return relevant global selections. The use of Web knowledge (applying the WV approach) on these global selections improves the quality of ranking.
In regard to a large amount of relations (high thresholds), the Combined System that favors Web Validation (i.e. $q \in [0, 0.2]$) is very efficient. Thus, following requests from experts (number of induced relation to take into account) we have to apply the relevant parameter q.
Different examples and experimental results (ROC curves) are presented on the web page: http://www.lirmm.fr/∼bechet/ECIR09.

4 Conclusion

In this paper we have established few approaches to order induced relations. The first one consists in representing syntactic relations by semantic vectors

Threshold	q coefficient										
	0 = WV	1/10	2/10	3/10	4/10	5/10	6/10	7/10	8/10	9/10	1 = SV
1000	0,62	0,62	0,62	0,64	0,64	0,66	0,68	0,69	**0,70**	0,70	0,52
2000	0,68	0,68	0,68	0,69	0,68	0,68	0,69	0,71	**0,71**	0,71	0,53
3000	0,70	0,70	0,69	0,69	0,69	0,69	0,71	0,74	0,76	**0,77**	0,52
4000	0,71	0,72	0,72	0,72	0,74	0,74	0,76	0,78	0,79	**0,80**	0,53
5000	0,73	0,75	0,76	0,77	0,78	0,79	0,80	0,81	0,82	**0,82**	0,53
6000	0,74	0,78	0,79	0,80	0,81	0,81	**0,81**	0,81	0,80	0,79	0,52
7000	0,76	0,81	**0,81**	0,81	0,79	0,75	0,70	0,64	0,62	0,62	0,53
8000	0,77	**0,79**	0,78	0,77	0,75	0,73	0,68	0,65	0,63	0,63	0,53
9000	**0,78**	0,78	0,77	0,76	0,75	0,72	0,68	0,65	0,64	0,64	0,54
10000	**0,79**	0,78	0,77	0,76	0,75	0,73	0,70	0,67	0,65	0,65	0,54
11000	**0,81**	0,78	0,78	0,77	0,76	0,74	0,71	0,68	0,67	0,66	0,54
12000	**0,82**	0,79	0,78	0,78	0,76	0,75	0,72	0,70	0,69	0,68	0,55

Table 2. AUC obtained with Combined system

as the combination of concepts of the French Larousse thesaurus. We measure the vector proximity with the cosine measure and the matching distance. The second one is a Web Validation method-based. It consists in querying the Web with induced syntactic relations. We use four ranking functions (i.e. *Frequency*, *Mutual Information*, *Cubic Mutual Information*, *Dice's measure*) to order results given by a search engine. In addition we propose to combine the systems. We evaluate our results with the ROC curves. We obtain good results with the Web Validation and the Combined System approaches. Now, we plan to perform more complex combinations in order to improve the quality of the results. Finally, we will apply our approach on other domains and languages.

References

1. N. Béchet, M. Roche, and J. Chauché. How the ExpLSA approach impacts the document classification tasks. In *Proc. of IEEE International Conference on Digital Information Management*, 2008.
2. J. Chauché. Un outil multidimensionnel de l'analyse du discours. In *Proceedings of Coling, Standford University, California*, pages 11–15, 1984.
3. J. Chauché and V. Prince. Classifying texts through natural language parsing and semantic filtering. In *International Language and Technology Conference*, 2007.
4. K. W. Church and P. Hanks. Word association norms, mutual information, and lexicography. In *Computational Linguistics*, volume 16, pages 22–29, 1990.
5. B. Daille. *Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques*. PhD thesis, Université Paris 7, 1994.
6. D. Faure and C. Nedellec. Knowledge acquisition of predicate argument structures from technical texts using machine learning: The system ASIUM. In *Proc. of Knowledge Acquisition, Modelling and Manag. Workshop*, pages 329–334, 1999.
7. M. Jarmasz and S. Szpakowicz. Roget's thesaurus and semantic similarity. In *Proc. of Conference on Recent Advances in NLP*, pages 212–219, 2003.
8. F. Smadja, K. R. McKeown, and V. Hatzivassiloglou. Translating collocations for bilingual lexicons: A statistical approach. *Comp. Linguistics*, 22(1):1–38, 1996.
9. P.D. Turney. Mining the Web for synonyms: PMI–IR versus LSA on TOEFL. *Proc. of ECML, LNCS*, 2167:491–502, 2001.