Using Semantic Roles in Information Retrieval Systems

Paloma Moreda, Borja Navarro, and Manuel Palomar

Grupo de investigación del Procesamiento del Lenguaje y Sistemas de Información
Departamento de Lenguajes y Sistemas Informáticos. Universidad de Alicante
Alicante, Spain
{moreda,borja,mpalomar}@dlsi.ua.es

Abstract. It is well known that Information Retrieval Systems based entirely on syntactic contents have serious limitations. In order to achieve high precision Information Retrieval Systems the incorporation of Natural Language Processing techniques that provide semantic information is needed. For this reason, in this paper a method to determine the semantic role for the constituents of a sentence is presented. The goal of this is to integrate this method in an Information Retrieval System.

1 Introduction

It is well known that Information Retrieval (IR) Systems based entirely on syntactic contents have serious limitations. One of the challenges of these applications is to develop high quality or high precision systems. In order to do this, it is necessary to involve Natural Language Processing (NLP) techniques in this kind of systems. These techniques provide semantic information to IR systems. Among the different NLP techniques which would improve IR systems, Semantic Role Labelling (SRL) is found . In this paper an extension of a IR system making use of a Semantic Role Labelling method is presented. Such method improves retrieval performance by reducing the number of non-relevant documents retrieved. This research is integrated in the project R2D2[1].

A semantic role is the relationship between a syntactic constituent and a predicate. For instance, in the next sentence

(E0) The executives gave the chefs a standing ovation

The executives has the Agent role, *the chefs* the Recipient role and *a standing ovation* the Theme role.

To achieve high precision IR systems, recognizing and labelling semantic arguments is a key task for answering "Who", "When", "What", "Where", "Why",

[1] This paper has been supported by the Spanish Government under project "R2D2: Recuperación de Respuestas de Documentos Digitalizados" (TIC2003-07158-C04-01). Besides, it has been partially funded by the Valencia Government under project number GV04B-276

A. Montoyo et al. (Eds.): NLDB 2005, LNCS 3513, pp. 192–202, 2005.

etc. For instance, the following questions could be answered with the sentence (E0). The Agent role answers the question (E3) and the Theme role answers the question (E4).

(E1) Who gave the chefs a standing ovation?

(E2) What did the executives give the chefs?

These examples show the importance of semantic roles in applications such as Information Retrieval.

Currently, several works have tried using Semantic Role Labelling in IR systems, unsuccessfully. Mainly, it is due to two reasons:

1. The lower precision achieved in these tasks.
2. The lower portability of these methods.

It is easy to find methods of Semantic Role Labelling that work with high precision for a specific task or specific domain. Nevertheless, this precision drops when the domain or the task are changed. For these reasons, this paper is about the problem of Semantic Role Labelling integrated with a IR system.

The remaining paper is organized as follows: section 2 gives an idea about the state-of-art in IR systems using semantic information. Afterwards, the Semantic Role Labelling method is presented in section 3. Then, how this method improves the performance of a IR system is presented in section 4. Finally, 5 concludes.

2 Using Semantic Information in IR: Background

In several IR systems the meaning of documents resides solely in the words that are contained within them. So, these systems, based on mathematical models such as the Boolean model, the vector-space model, the probabilistic model and their variants [2], represent the meanings of documents and queries as bags of words. Even though they are well established, from the user's perspective, it is difficult to use these IR systems. Users frequently have problems expressing their information needs and translating those needs into queries.

For instances [21], consider the sentence *Harry loves Sally*. If it is considered as a query in a keyword matching system, the system would look for documents containing the terms *Harry*, *Sally* and *love*, and would not be able to distinguish among the following sentences (E3), (E4), (E5), (E6) and (E7).

(E3) Harry loves Sally

(E4) Sally loves Harry, but Harry hates Sally

(E5) Harry's best friend loves Sally's best friend

(E6) Harry and Sally loves pizza

(E7) Harry's love for Sally is beyond doubt

Several methods have been proposed to help users to choose searching terms and articulate queries making use of semantic information. Most of them work for a specific domain and use domain specific thesaurus. For instance, some systems use concepts[2]. So, the system presented in [19] first assigns a syntactic analysis to input from either a query or document about medical domain. The heart of the approach is a mapping of the phrases to concepts in UMLS domain model [1]. Then, the semantic interpretation specifies the relationship in a semantic case role form, which it is obtained between the concepts and the input phrases.

In [6] y [7] a method implemented on Digital Library using a hierarchical perspective based on a concept system is presented. The EDR [8] electronic dictionary was used as a thesaurus. The meanings of words extracted from queries and text are represented by concepts and are used for retrieval. The "concept of a word" indicates the concept which represents the shared synonyms for the meaning of a word.

Complex nominal sequences must undergo a specific semantic treatment in order to increase the performance of IR systems [5]. This work defines three objectives using semantic information on English compounds: determination of the conditions under which the concept expressed by a compound is presented in a text, recognition of equivalent reformulations of the compounds and a weighting of the words of the compounds proportional to their importance. An extension of this work is proposed in [4] adding an objective of disambiguation of polysemous words. The work shows, by using concrete examples from an experimentation conducted on a French system of telematic services, how a rich semantic model for binomial sequences can be used in order to increase both the recall and precision rates of an IR system. This system, named CNET, is composed of three modules: the linguistic analyzer, which generates a structured representation of a text in which each word is replaced by the list of its meanings; the indexing module, which generates the list of the indexes, and its weights according to their frequencies in the text, that represent the contents of a text; and the matching module, which valuates the relevance of a text for a given question.

On the other hand, several methods have investigated IR systems making use of relationships for specific domains or specific tasks. For instance, the use of relation matching in IR is discussed in [21]. Terms and relationships between terms expressed in the query are matching with terms and relationships found in the documents. In this method, non-domain specific knowledge was used. Nevertheless, the cause-effect relation was the only one studied.

The work of [23] is based on an algorithmic approach of concept discovery and association. Concepts are discovered using an algorithm based on an automated thesaurus generation process. Subsequently, similarities among terms are computed using the cosine measure, and the relationships among terms are established using a method known as *max-min* distance clustering.

NLP techniques with the structured domain knowledge provided by the UMLS, were applied to texts concerning to the coronary arteries in order to ex-

[2] "Concept" and "term" words are used according to the terminology used by respective authors

tract arterial branching relationships from cardiac catheterization reports [20]. First, the coronary artery terminology occurring in the sentence is identified. Next, the processing constructs a complete branching predication where a correspondence between a syntactic entity and a semantic predicate is established.

Besides the semantic relations are explored and evaluated in cross-language IR in the medical domain making use UMLS as the primary semantic resource and a corpus of English and German medical abstracts, in [22]. A method for selecting relevant relations from those proposed by UMLS and a method for extracting new instances of relations based on statistical and NLP techniques are described. First the specialist lexicon provides lexical information. Second, the metathesaurus is the core vocabulary component used for assigning a identifier for each term. Third, the semantic network provides a grouping of concepts according to their meaning into a semantic type and specifies potential relations between those semantic types.

Other researchers have studied general methods for extracting semantic relations for IR. In response, Lu [12] investigated the use of case relation matching using a small test database of abstracts. Using a tree-matching method for matching relations, he obtained worse results than from vector-based keyword matching. The tree-matching method used is probably not optimal for IR and the results may not reflect the potential of relation matching [21].

Liu [9] tried to match individual concepts together with the semantic role that the concept has in the sentence. Instead of trying to find matches for *term1-relation-term2*, his system sought to find matches for *term1-relation* and *relation-term2* separately. Liu used case roles and the vector-space retrieval model, and was able to obtain positive results only for long queries (abstracts that are use as queries).

The DR-LINK project attempted to use general methods for extracting semantic relations for IR. Non-domain specific resources were used. However, preliminary results found few relation matches between queries and documents.

3 The SemRol Method

In this section, the Semantic Role method, named SemRol, is presented.

The problem of the Semantic Role Labelling is not trivial. In order to identify the semantic role of the arguments of a verb, two phases have to be solved, previously. Firstly, the sense of the verb is disambiguated. Secondly, the argument boundaries of the disambiguated verb are identified.

First, the sense of the verb has to be obtained. Why is it necessary to disambiguate the verb? Following, an example shows the reason for doing so.

(E8) John gives out lots of candy on Halloween to the kids on his block

(E9) The radiator gives off a lot of heat

Depending on the sense of the verb a different set of roles must be considered. For instance, Figure 1 shows three senses of verb *give* (give.01, give.04,

and give.06)) and the set of roles of each sense. So, sentence (E0) matches with sense give.01. Therefore, roles *giver*, *thing given* and *entity given to* are considered. Nevertheless, sentence (E1) matches with sense give.06 and sentence (E2) matches with sense give.04. Then, the sets of roles are (*distributor*, *thing distributed*, *distributed*) and (*emitter*, *thing emitted*), respectively. In sentence (E1), *John* has the distributor role, *lots of candy* the thing distributed role, *the kids on his block* the distributed role and *on Halloween* the temporal role. In sentence (E2), *the radiator* has the emitter role and *a lot of heat* the thing emitted role. These examples show the relevance of WSD in the process of assignment of semantic roles.

```
<roleset id="give.01" name="transfer"> <roles>
  <role n="0"  descr="giver"  vntheta="Agent"/>
  <role n="1"  descr="thing given"  vntheta="Theme"/>
  <role n="2"  descr="entity given  vntheta="Recipient""/>
</roles>

<roleset id="give.04" name="emit"> <roles>
  <role n="0"  descr="emitter"/>
  <role n="1"  descr="thing emitted"/>
</roles>

<roleset id="give.06" name="transfer"> <roles>
  <role n="0"  descr="distributor"/>
  <role n="1"  descr="thing distributed"/>
  <role n="2"  descr="distributed"/>
</roles>
```

Fig. 1. Some senses and roles of the frame *give* in PropBank [17].

In the second phase, the argument boundaries are determined. For instance, in the sentence (E0), the argument boundaries recognized are

[The executives] gave [the chefs] [a standing ovation]

Once these two phases are applied, the assignment of semantic roles can be carried out.

So, our method, named SemRol, presented in this section consists of three phases:

1. Verb Sense Disambiguation phase (VSD)
2. Argument Boundaries Disambiguation phase (ABD)
3. Semantic Role Disambiguation phase (SRD)

These phases are related since the output of VSD phase is the input of ABD phase, and the output of ABD phase is the input of SRD phase. First, the process to obtain the semantic role needs the sense of the target verb. After that, several heuristics are applied in order to obtain the argument boundaries of the sentence. And finally, the semantic roles that fill these arguments are obtained. So, the success of the method depends on the success of the three phases.

Both, Verb Sense Disambiguation phase and Semantic Role Disambiguation phase are based on conditional Maximum Entropy (ME) Probability Models

Table 1. Results of the SRD phase.

	Precision	Recall	$F_{\beta=1}$		Precision	Recall	$F_{\beta=1}$
A0[a]	92.17%	90.50%	91.33	AM-MNR	99.70%	97.90%	98.79
A1	83.17%	96.31%	89.26	AM-MOD	100.00%	100.00%	100.00
A2	98.14%	88.26%	92.94	AM-NEG	100.00%	98.47%	99.23
A3	99.08%	72.48%	83.72	AM-PNC	100.00%	99.00%	99.50
A4	92.86%	35.37%	51.23	AM-PRD	100.00%	100.00%	100.00
A5	100.00%	50.00%	66.67	AM-PRP	0.00%	0.00%	0.00
AM-ADV	99.71%	98.58%	99.14	AM-REC	0.00%	0.00%	0.00
AM-CAU	98.11%	98.11%	98.110	AM-TMP	99.04%	94.99%	96.99
AM-DIR	96.30%	86.67%	91.23	R-A0	0.00%	0.00%	0.00
AM-DIS	97.50%	76.47%	85.71	R-A1	0.00%	0.00%	0.00
AM-EXT	96.00%	48.98%	64.86	R-A2	0.00%	0.00%	0.00
AM-LOC	100.00%	98.70%	99.34	R-AM-LOC	100.00%	75.00%	85.71
R-AM-TMP	85.71%	100.00%	92.31	V	97.44%	97.44%	97.44
				all	**92.46%**	**92.38%**	**92.42**
				all–{V}	**90.53%**	**90.41%**	**90.47**

[a] The semantic roles considered in PropBank are the following [3]: Numbered arguments (A0-A5, AA):, arguments defining verb-specific roles; adjuncts (AM-), general arguments that any verb may take optionally, for instance, AM-LOC is location or AM-CAU: cause; references (R-), arguments representing arguments realized in other parts of the sentence; and verbs (V), participant realizing the verb of the proposition.

[18]. It has been implemented using a supervised learning method that consists of building classifiers using a tagged corpus [15]. Argument Boundaries Disambiguation and Semantic Role Disambiguation phases take care of recognition and labelling of arguments, respectively. VSD module means a new phase in the task. It disambiguates the sense of the target verbs. So, the task turns more straightforward because semantic roles are assigned to sense level. A more detalied approximation of this method is presented in [14].

Results about SRD phase are shown in table 1. In order to evaluate it, right senses of the verbs and right argument boundaries have been presumed. The results have been obtained using the PropBank corpus [17]. There are 26 different kinds of roles in this corpus. One of theme, the *V* role, refers verbs. In this case, the precision is about 97%. In most of cases, the precision is over 83%, being over 96% in sixteen of them (only four cases are below 96%) and 100% in seven of them.

4 IR System Extended with SemRol

The goal of this paper is to integrate the method presented in the previous section (section 3) in an IR system. In this particular case, in the IR-n system[11], [10].

The architecture of an IR system extended with the SemRol method is shown in the figure 2. The architecture presented consist of four modules: IR system, selection module, annotation module and module of heuristics.

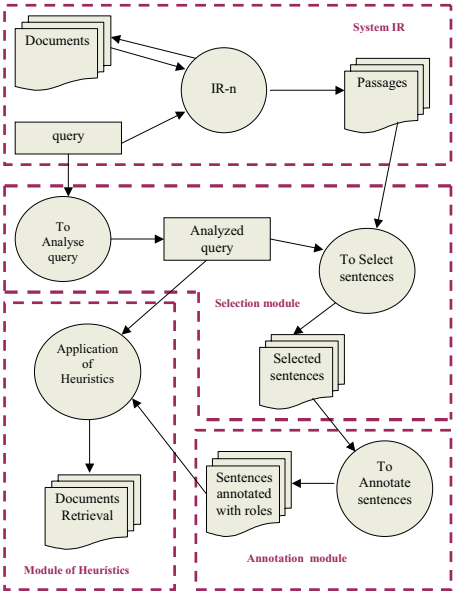

Fig. 2. The architecture of an IR system extended with the SemRol method.

When a query is done, the IR system IR-n retrieves a set of passages. It is supposed that these passages contain the answer of the query. Then the verbs of the sentences of these passages are compared with the verb of the query, and a list of verbs related with it, in order to select only the sentences containing a verb of this list. Next, the selected sentences are annotated with semantic roles making use of SemRol method. Finally, a set of heuristics are applied. These heuristics establish a relation between queries and semantic roles. So, only the sentences contained the right semantic roles are selected and the number of passages retrieved is reduced.

4.1 IR System IR-n

Passage Retrieval is an alternative to traditional document-oriented Information Retrieval. IR-n system is a passage retrieval system. These systems use contiguous text fragments (or passages), instead of full documents, as basic unit of information. So, IR-n system uses the sentences as atoms with the aim to define the passages. Thus each passage is composed by a specific number of sentences. This number depends in a great measure of the collection used. For this reason, the system requires a training phase to improve its results. IR-n system uses overlapping passages in order to avoid that some documents can be considered not relevant if words of the question appear in adjacent passages.

First, the system calculates the similarity between the passages and the user query. Next, the system determines the similarity of the documents that contain these passages making use of the best passage similarity measure. This approach

is based on the fact that if a passage is relevant then the document is also relevant.

As most of IR systems, IR-n system uses also techniques of query expansion. In current version, the most frequent terms in the documents are added.

4.2 The Extension

The new modules needed to extend the IR system are analyzed below.

- **Selection module.** First, a list of verbs related to the verb in the query is obtained. In order to do this, an electronic lexical database has been used, WordNet [13]. In it, nouns, verbs, and adjectives are organized into synonym sets, each representing underlying lexical concepts. To create WordNet several kinds of semantic relations were used, such as synonymy, hyponymy, meronymy and antonymy. In the case of verbs, this semantic relations have been adapted to fit the semantics of them (for instance, troponymy is the adaptation of hyponymy).
  In our system, the list of related verbs is extracted making use of synonymy and troponymy relations.
  Secondly, the verbs of the sentences of the passages retrieval by IR-n are compared with this list of verbs. So, only passages containing sentences with one of these verbs are selected and those sentences are marked.
- **Annotation module.** The sentences marked in the previous module are annotated with semantic information by using the SemRol method. So, the argument boundaries of the sentences are recognized and the semantic roles that fill this arguments are identified.
  As a result, a set of annotated sentences with the roles of the arguments of the verbs is obtained.

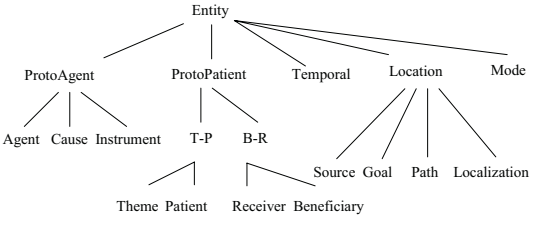

Fig. 3. Set of roles.

The set of roles [16] used for the annotation process is shown in the figure 3. This set of roles is different to the set of roles used for testing the SemRol method (See section 3). It is not a problem because a mapping between both set of roles can be done easily.

– **Module of heuristics.** Depending on the kind of question a different set of roles could be considered. So, it is possible to define a set of heuristics in order to establish a relationship between questions and semantic roles. For instance, questions such as "When", "What + time expression" or "In what + time expression" must be answered with the Temporal semantic role and must not be answered with the Agent, Patient, Location, Cause or Mode semantic role; and "Where", "In where + location expression" or "In what + location expression" must be answered with the Location semantic role and must not be answered with the Agent, Patient, Temporal, Cause or Mode semantic role. A summary of these heuristics is shown in figure 4.

Question	Role		No role
Where In where In what + exp At what + exp	Location		ProtoAgent Mode Temporal Cause ProtoPatient
When In what + exp What + exp	Temporal		ProtoAgent Mode Location Cause ProtoPatient
How	Mode Theme (if it is a diction verb)		ProtoAgent Location Temporal Cause Patient Beneficiary
Who	Agent - ProtoAgent Patient - ProtoPatient		Mode Temporal Location Theme Beneficiary
What	Cause Theme		
Whose	Receiver Beneficiary Patient	ProtoPatient	Agent Location Mode Temporal Theme Cause

Fig. 4. Set of heuristics.

Then, making use of these rules only the sentences containing the right semantic roles are selected and the number of passages retrieval is reduced.

5 Conclusions and Working in Progress

In this paper, an extension of a IR system using semantic role information is presented. When a query is done, the IR system IR-n retrieves a set of passages. Then the verbs of the sentences of these passages are compared with a list of verbs related with the verb of the query. Next, the sentences containing a verb of this list are annotated with semantic roles by using the SemRol method. Finally, several heuristics are applied in order to establish a relation between queries and semantic roles. So, only the sentences containing with the right semantic roles are selected and the number of passages retrieved is reduced.

The SemRol method is used to annotate selected sentences with semantic information. This method, based on conditional Maximum Entropy (ME) Probability Models, identifies and labels the constituents of a sentence with semantic roles. It consists of three phases. First, the process to obtain the semantic role needs the sense of the target verb. After that, several heuristics are applied in order to obtain the argument boundaries of the sentence. And finally, the semantic roles that fill these arguments are obtained.

Currently, we are developing the extension modules. Shortly, we will show results about this IR system extended with semantic role information and will evaluate them in appropriate forum.

On the other hand, it is important to say that the current version of the SemRol method only works with English corpus. In order to overcome this limitations, some kind of adaption must be done.

References

1. UMLS Unified Medical Language System (2005AA release). http://www.nlm.nih.gov/research/umls/umlsdoc.html, January 2005.
2. R.A. Baeza-Yates and B.A. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press / Addison-Wesley, 1999.
3. X. Carreras and L. Màrquez. Introduction to the CoNLL-2004 Shared Task: Semantic Role Labeling. In *Proceedings of the Eighth Conference on Natural Language Learning (CoNLL-2004)*, Boston, MA, USA, Mayo 2004.
4. C. Fabre and P. Sebillot. Semantic interpretation of binomial sequences and information retrieval. In International ICSC Congress on Computational Intelligence: Methods and Applications (CIMA). Symposium on Advances in Intelligent Data Analisys (AID), editors, *Proceedings of the Symposium on Advances in Intelligent Data Analisys*, Rochester, USA, 1999.
5. L.S. Gay and W.B. Croft. Interpreting nominal compounds for information retrieval. *Information Processing and Management*, 26(1):21–38, 1990.
6. C. Horii, M. Imai, and K. Chihara. An information retrieval using conceptual index term for technical paper on digital library. In *Proceedings of International Symposium on Research, Development and Practice in Digital Libreries*, volume 97, pages 205–208, Tsukuba, November 1997. International Symposium on Research, Development and Practice in Digital Libreries (ISDL).
7. C. Horii, M. Imai, and K. Chihara. Information retrieval using conceptual index terms for technical papers in a digital library. *Systems and Computer in Japan*, 32(8), 2001.
8. Japan Electronic Dictionary Laboratory. EDR electronic dictionary technology guide (2nd edition, revised). http://www.iijnet.or.jp/edr, 1995.
9. G.Z. Liu. Semantic vector space model: Implementation and evaluation. *Journal of American Society for Information Science*, 48(5):395–417, 1997.
10. F. Llopis and R. Muñoz. Cross Language experiments with IR-n system. In *Proceedings of Workshop of Cross-Language Evaluation Forum (CLEF 2003)*, Trondheim, Norway, 2003.
11. F. Llopis and J.L. Vicedo. IR-n system, a passage retrieval system at CLEF 2001. In *Proceedings of Workshop of Cross-Language Evaluation Forum (CLEF 2001)*, pages 244–252, Darmstadt, Germany, 2001.

12. X. Lu. *An application of case relations to document retrieval*. PhD thesis, University of Western, Ontario, 1990.
13. G. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller. Five papers on wordnet. csl report 43. Technical report, Cognitive Science Laboratory, Princeton University, 1990.
14. P. Moreda, M. Palomar, and A. Suárez. Assignment of semantic roles based on word sense disambiguation. In *Proceedings of the 9TH Ibero-American Conference on AI*, Puebla, Mexico, Noviembre 2004.
15. P. Moreda, M. Palomar, and A. Suárez. Identifying semantic roles using maximum entropy models. In *Proceedings of the International Conference Text Speech and Dialogue*, Lecture Notes in Artificial Intelligence, Brno, Czech Republic, 2004. Springer-Verlag.
16. B. Navarro, P. Moreda, B. Fernández, R. Marcos, and M. Palomar. Anotación de roles semánticos en el corpus 3lb. In *Proceedings of the Workshop Herramientas y Recursos Linguísticos para el Español y el Portugués*, Tonantzintla, México, November 2004. Workshop Herramientas y Recursos Linguísticos para el Español y el Portugués. The 9TH Ibero-American Conference on Artificial Intelligence (IBERAMIA 2004).
17. M. Palmer, D. Gildea, and P. Kingsbury. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 2004. Submitted.
18. A. Ratnaparkhi. *Maximum Entropy Models for Natural Language Ambiguity Resolution*. PhD thesis, University of Pennsylvania, 1998.
19. T.C. Rindflesch and A.R. Aronson. Semantic processing in information retrieval. In *Proceedings of the 17th Annual Symposium on Computer Applications in Medical Care*, pages 611–615, 1993.
20. T.C. Rindflesch, C.A. Bean, and C.A. Sneiderman. Argument identification for arterial branching predications asserted in cardiac catheterization reports. In *Proceedings of the AMIA*, 2000.
21. C. Soo and G. Kho. The use of relation matching in information retrieval. *LIBRES: Library and Information Science Research*, 7(2), September 1997.
22. S. Vintar, P. Buitelaar, and M. Volk. Semantic relations in concept-based cross-language medical information retrieval. In *Inproceedings of the Workshop on Adaptive Text Extraction and Mining*. Workshop on Adaptive Text Extraction and Mining (ECML/PKDD), 2003.
23. J. Zhang, J. Mostafa, and Himansu Tripathy. Information Retrieval by Semantic Analysis and Visualization of the Concept Space of D-lib Magazine. *D-lib Magazine*, 8(10), October 2002.