Using semantic word classes in text information retrieval systems

P. Gamallo[1], A. Agustini[1], P. Quaresma[2] and G. Lopes[1]

[1] CITI - Departamento de Informática
Universidade Nova de Lisboa
2825 Monte da Caparica, Portugal
gamallo|aagustini|gpl@di.fct.unl.pt

[2]Departamento de Informática,
Universidade de Évora,
7000 Évora, Portugal
pq@di.uevora.pt

July 14, 2002

Abstract

Intelligent text information retrieval systems need the capability to automatically extract and use knowledge from their text bases. In this paper an application of methodologies to automatically acquire semantic word classes and to use them in text information retrieval systems is described. Some examples from the Portuguese Attorney General's Office web information retrieval system are presented.

1 Introduction

Intelligent text information retrieval systems need the capability to automatically extract and use knowledge from their text bases.

In previous work Gamallo et al. [2, 3] showed how to automatically acquire word classes from text bases. In this paper the application of their results to improve the quality of text information retrieval systems is presented. The strategy is to use word classes to allow users to redefine their queries, selecting related concepts and specific subcategorization patterns. For instance, if an user asks for documents about "militaries", the IR system may give him/her the chance to select related concepts (employee, worker, ...) and to choose subcategorization patterns (military pension, active military, ...). Using this approach the IR system is able to cooperatively interact with users helping them to better specify their queries.

In section 2 the methodology used to acquire word classes from texts is briefly described. In section 3 the Portuguese Attorney General's Office web information retrieval system is presented. Finally in section 4, an example of the integration of the word classes with the IR system is shown.

2 Acquisition of Semantic Word Classes

We use an unsupervised method for acquiring word classes from partially parsed text corpora. Word classes are associated here to the words that can appear in specific contexts of subcategorisation. This method has been accurately described in [2, 3]. For the purpose of this paper, we will merely outline the basic assumptions the acquisition strategy is based on, and then we will sketch a general overview on its different steps and modules.

2.1 Basic Assumptions

Our method is based on two theoretical assumptions.

First, we assume a very general notion of linguistic subcategorisation. More precisely, we consider that in a Head-Modifier syntactic dependency, not only the Head imposes constraints on the Modifier, but also the Modifier imposes linguistic requirements on the Head [5]. So, for a particular word, we attempt to learn both what kind of modifiers and what kind of heads it subcategorises. For instance, consider the compositional behavior of the noun `republic` in a domain-specific corpus. On the one hand, this word appears in the Head position within dependencies such as `republic of Ireland`, `republic of Portugal`, and so on. On the other hand, it plays the role of Modifier in dependencies like `president of the republic`, `government of the republic`, etc. So, for a particular word, we attempt to learn both what kind of complements and what kind of heads it subcategorises. Given that there are interesting semantic regularities among the words cooccurring with `republic` in such linguistic contexts, our aim is to implement an algorithm letting us learn two different subcategorisation patterns:

- $< of; republic, [M] >$, where the M position must be filled by words referring to particular nations or states. Indeed, in semantic-pragmatic terms, only nations or states can be republics;

- $< of; [H], republic >$, where the H position must be filled by word heads denoting specific parts of the republic: e.g., institutions, organisations, functions, and so on.

The second assumption concerns the procedure for identifying and clustering similar subcategorisation patterns. We assume, in particular, that different subcategorisation patterns are considered to be semantically similar if they have similar word distribution [1]. Let's take, for instance, the following patterns: $< of; [H], republic >$, $< of; [H], state >$, $< of; delegate, [M] >$, $< on; be\_incumbent, [M] >$. All of them share the same semantic preferences. Since these patterns require words denoting the same semantic class, they tend to possess similar word distribution. Moreover, we also assume that the set of words required by similar subcategorisation patterns represents the extensional description of their semantic preferences.

2.2 Method Overview

Our learning method consists in the following steps. Raw Portuguese text is automatically tagged and partially analysed in sequences of basic chunks. Then, binary syntactic dependencies are identified on the basis of some symbolic heuristics of attachment. Then, we extract subcategorisation patterns from the binary dependencies, by following the first assumption outlined above. Finally, subcategorisation patterns with similar word distributions are clustered into more general classes taking (second assumption). Similarity between patterns is calculated by using a particular weighted Jaccard coefficient [4]. For instance, if $< of; republic, [M] >$ and $< of; [H], state >$ are considered as similar patterns, they can be merged into a general class:

$$< of; [H], republic > \cup < of; [H], state >$$

which is constituted by those words occurring in both patterns: e.g., `president, assembly, minister, government, administration` etc. This allows us to build clusters of words representing the semantic preferences of similar subcategorisation restrictions.

3 PGR: Portuguese Attorney General's Office web IR system

In previous work Quaresma and Rodrigues [7, 6] have described the web legal information retrieval system of the Portuguese Attorney General's Office.

In this system, the text base has around 7,000 documents since 1940 with near 10,000,000 words. The documents have a specific structure and they were saved in an XML format. The document structure was analysed and an ontology was defined in the daml+oil language [8].

The documents were also changed by adding to them linguistic information obtained through off-line processing. Namely, it was added to each word:

- Its canonical form
- Its morpho-syntactic tag (verb, noun, ...)

The canonical form of each word is obtained through the access to a lexical database, POLARIS[1], which has more than 900,000 words with information about their canonical forms and possible morpho-syntactic tags.

4 Example

Suppose the user asks for documents about militaries[2].

- militares - militaries

The PGR system answers:

- There are 1559 documents.
- You may cluster them in concepts, or select related concepts or choose subcategorization patterns.

If the user chooses to select subcategorization patterns he/she will obtain the following list:

- pensão de militar - military pension
- militar por categoria - military by category
- ...

Then the user may select "pensão de militar" and obtain the following answer:

- There are 2 documents.
- Document 150/1972/P000341972.html
- Document 150/1951/P000731951.html

Finally, it will be possible to retrieve the intended documents (2 out of 1559!).

[1] POrtuguese Lexicon Acquisition and Retrieval Interaction System

[2] Due to space constraints the system's graphical output is not shown here.

References

[1] David Faure and Claire Nédellec. Asium: Learning subcategorization frames and restrictions of selection. In *ECML98, Workshop on Text Mining*, 1998.

[2] Pablo Gamallo, Alexandre Agustini, and Gabriel P. Lopes. Selection restrictions acquisition from corpora. In *10th Portuguese Conference on Artificial Intelligence (EPIA'01)*, pages 30–43, Porto, Portugal, 2001. LNAI, Springer-Verlag.

[3] Pablo Gamallo, Alexandre Agustini, and Gabriel P. Lopes. Using co-composition for acquiring syntactic and semantic subcategorisation. In *ACL-SIGLEX'02*, Philadelphy, USA, 2002.

[4] Gregory Grefenstette. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers, USA, 1994.

[5] James Pustejovsky. *The Generative Lexicon*. MIT Press, Cambridge, 1995.

[6] Paulo Quaresma and Irene Rodrigues. Using logic programming to model multi-agent web legal systems – an application report. In *Proceedings of the ICAIL'01 - International Conference on Artificial Intelligence and Law*, St. Louis, USA, May 2001. ACM.

[7] Paulo Quaresma and Irene Pimenta Rodrigues. Pgr: Portuguese atorney general's office decisions on the web. In Osamu Yoshie, editor, *Proceedings of the 14th International Conference on Applications of Prolog*, University of Tokyo, Tokyo, Japan, October 2001. REN Associates, Inc. ISSN 1345-0980. To be published by Springer Verlag's LNAI.

[8] www.daml.org. *DAML+OIL – DARPA Agent Markup Language*, 2000.