Chapter XIII
Automatic Acquisition of Semantics from Text for Semantic Work Environments

Maria Ruiz-Casado
Universidad Autonoma de Madrid, Spain

Enrique Alfonseca
Universidad Autonoma de Madrid, Spain

Pablo Castells
Universidad Autonoma de Madrid, Spain

ABSTRACT

This chapter presents an overview of techniques for semi-automatic extraction of semantics from text, to be integrated in a Semantic Work Environment. A focus is placed on the disambiguation of polysemous words, the automatic identification of entities of interest, and the annotation of relationships between them. The application of these topics falls into the intersection of three dynamic fields: the Semantic Web, Wiki and other work environments, and Information Extraction from text.

INTRODUCTION

Nonsemantic wikis and information desktops manage shared or personal information. The success of these platforms is mainly due to the interest they have arisen among potential contributors who are eager to participate because of their particular involvement in the domain under discussion. A stated by Pressuti and Missim (this volume), this success can be attributed to several features of these systems, such as the easiness to publish and review content on the Web through the Web browser, the few restrictions on what a wiki contributor can write, and the user friendly interface (both in wikis and personal information desktops). All these constitute a very flexible mechanism to create, publish, and annotate content.

Semantic Work Environments (SWEs) propose a blend of the Semantic Web technologies and individual or collaborative work environments.

The Semantic Web (SW) constitutes an initiative to extend the Web with machine readable content, primarily making the content explicit through semantic annotations. This allows an easy automated processing when searching and retrieving information. SWEs can also benefit from the tagging of relevant concepts and relations held among them, in such a way that the search and retrieval of information is greatly enhanced when semantic annotation is present. SWEs maintain explicit marks for these concepts and relations, so the tags can be easily detected and used by the computer to retrieve from the texts the information required by a user's query.

The addition of semantic annotations to documents can be achieved following the wiki philosophy of lowering the technical barriers, as simple labels attached to the hyperlinks. In fact, wiki communities have already proved to succeed in collaboratively producing at low cost vast information repositories. Next, internal tools may be provided to transform these annotations into a SW markup language. Through the semantic augmentation, wikis and other semantic work environments benefit from the automation of management, searching and retrieval pursued by the SW.

But the semantic tagging of work environments faces the well known bottleneck in the Semantic Web applications: placing tags in a large amount of existing content, sometimes also rapidly evolving, can be too costly if it has to be done manually. Moreover, the wiki contributors or the semantic desktop users may feel discouraged to review the full database (containing pre-existing information) in order to place semantic tags, or may hesitate about where and how to place them. This is the case of annotating large repositories of information like news databases, e-mails, or other personal and company-owned databases. In these cases, the number of users and information managers may be small and the amount of data very large. Even in a collaborative environment where the cost is shared among several contributors the tagging bottleneck is present. For instance, among the existing Wikipediae, as of March 2006, the English version has more than one million articles, and the German, Spanish, French, Italian, Japanese, Dutch, Polish, Portuguese, and Swedish Wikipediae are above one hundred thousand articles each. If all these entries were to be extended with semantic annotations manually in a reasonable amount of time, the cost of manually labeling them would be enourmous, if not unfeasible.

In answer to this need, the use of semi-automatic annotation procedures has been the object of extensive research for many purposes, including tackling the SW tagging bottleneck. In general, semantic tags provide the same information that a human reader can elicit (sometimes unconsciously) from an untagged natural-language text: names of locations, people, organisations and other entities mentioned in the text, and events and relationships involving them, for instance, that a particular person works in a specific company. Moreover, a human can usually discriminate the sense intended in a text for a polysemous word, for example, if the word *pipe* is mentioned referring to smoking, plumbing, or music. Finally, from the way in which the terms are used in a context, it is sometimes possible to imagine their meaning up to a certain degree. Motivated by the large amounts of textual information, different areas of Natural Language Processing (NLP) try to create algorithms to perform all these deductions automatically. The subfield of NLP that studies the automatic annotation of entities, events, and relationships in unrestricted texts is called Information Extraction (IE). Text Mining focuses on the discovery of information previously unknown (Hearst, 2003), for example, extracting patterns from text that express known relationships or events and using those same patterns to extract new knowledge. Finally, Word Sense Disambiguation (WSD) tries to identify the correct sense with which a word is being used in a given context.