From Text to Knowledge*

M. Fernández[1], E. de la Clergerie[2], M. Vilares[1]

[1] Department of Computer Science, University of Vigo
Campus As Lagoas s/n, 32004 Ourense, Spain
{mfgavilanes,vilares}@uvigo.es
[2] Institut National de Recherche en Informatique et en Automatique
Domaine de Voluceau, Rocquencourt, B.P. 105, 78153 Le Chesnay Cedex, France
Eric.Clergerie@inria.fr

Nowadays, in spite of the increasing amount of information available in electronic format, most of human knowledge is still only available in textual format, from which it is not possible to directly consider automatic management tasks. This makes practicable knowledge acquisition a highly interesting topic, in particular in the case of technical and/or scientific documents with a highly structured wording that could simplify their computational treatment. In this context, we focus on semantical data extraction from text. The goal is to generate a knowledge structure to develop question-answering facilities on textual documents.

In order to favour understanding, we introduce the proposal from a botanic corpus describing the West African flora. It is composed of about forty volumes in French, organized as a sequence of sections, each one dedicated to one species and following a systematic structural schema. So, for example, sections include a descriptive part enumerating morphological aspects such as color, texture, size or form. This implies the presence of nominal phrases, adjectives; and also adverbs to express frequency and intensity, and named entities to denote dimensions.

A first phase consisting of performing such a translation has been applied using an OCR platform and a complementary error correction technique [5], although the description of this initial task is not of interest for the purposes of this paper. The next step consists of capturing the structure of the text using a combination of mark-up language, such as XML, and chunking tasks. The goal is to establish the linguistic context the analyzer will work with in order to serve as a guideline for the later knowledge acquisition process. Also, as a result, we can browse the document.

We are now ready to introduce knowledge acquisition, by extracting and later connecting terms in order to detect pertinent relations and eliminate non-deterministic interpretations. To deal with this, two principles are considered: the *distributional semantics model* [4] establishing that words whose meaning is close often appear in similar syntactic contexts; and the assumption that terms shared by these contexts are usually nouns and adjectives [1]. [2] As a starting point, we parse the text on the basis of the meta-grammar concept [2], providing both

* Research partially supported by the Spanish Government under project TIN2004-07246-C03-01, and the Autonomous Government of Galicia under projects PGIDIT03SIN30501PR and PGIDIT05PXIC30501PN.

total and partial recognition including XML information. The parse takes the form of a graph whose arcs represent relations of the type *governor/governed*, which allows the vocabulary to be concentrated around *pivot terms* and even makes it possible to establish similarity measures between these [3].

Term extraction is organized around the recognition of generic lexical and/or syntactic patterns from these pivot terms. We profit from this topological information to apply automatic learning techniques in order to locate those dependencies that are more frequent and less ambiguous, focusing the meaning of the text on what we baptize as *strong dependencies.* These dependencies constitute the semantical skeleton of the text, from which we look for more concrete properties involving pivot terms. In this sense, linguistic marking information allows primary conceptual adquisition from text. So, we can consider coordination schema such as *“X et X”* (‘‘X and Y’’), interval definitions of the type *“de X à Y”* (‘‘from X to Y’’); or more explicit physical information such as *“en forme de X”* (‘‘in form of X’’) or *“de couleur X”* (‘‘of color X’’). The result serves to take out simple concepts such as the value for `color`, `form` or `domain` attributes; or detect enumerations that can propagate some of these values.

We now infer a number of semantic tags that we use for text indexing, as in the case of *“des sépales* [**organe**] *ovales-aigus* [**forme**]*, glabres* [**texture**] *ou éparsement hérissés* [**texture**]*”* (‘‘Sepals [**organ**] oval-pointed [**form**], smooth [**texture**] or scattered bristly [**texture**]’’). Also, once basic syntagms and properties have emerged from the text, we focus on more sophisticated patterns connecting them in order to derive more complex semantical relations. Such is the case of *“SN à SN”* (‘‘NS with NS’’) in *“Fleurs à pétales ovales”* (‘‘Flowers with oval petals’’); from which we can derive an *hypernymy* relation.

At this point, we have at our disposal a preliminary tool-kit to deal with the automatic generation of ontologies. The identification of semantic classes is approached from the detection of similar syntactic contexts around pivot terms. From here, existing relations between those semantic classes are approached from the lexical and/or syntactic patterns connecting them.

References

1. J. Bouaud, B. Bachimont, J. Charlet, and P. Zweigenbaum. Methodological principles for structuring an ontology, 1995.
2. E. de la Clergerie. From metagrammars to factorized TAG/TIG parsers. In *Proc. of IWPT’05*, pages 190–191, Vancouver, Canada, October 2005.
3. L. Denoue and L. Vignollet. L’importance des annotations, application à la classification des documents web. *Special Issue of the Journal of Pragmatics*, 1:1–22, 1996.
4. Z.S. Harris. *Mathematical Structures of Languages.* John Wiley & Sons, New York, U.S.A., 1968.
5. B. Sagot and E. de la Clergerie. Error mining in parsing results. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 329–336, Sydney, Australia, July 2006. Association for Computational Linguistics.