Working on a botanic corpus

Eric de la Clergerie

January 30, 2001

1 Motivation

Extracting information from an encyclopedic corpus of botanic may be done by hand but it is a long and tedious work. More and more, it becomes interesting and possible to speed-up the process by automatizing it but still keeping an human expert for validation.

Among the different kind of information that may be extracted from a botanic corpus, we can cite terminology, conceptual information to model a specialized domain (for instance African tropical flora), and descriptions of a set of plants that follows the conceptual model.

The group ATOLL at INRIA Rocquencourt is interested in this area of research for different reasons. The group is primarily concerned with parsing for natural language and we believe that parsing is a key component for knowledge acquisition tasks from textual documents. Actually, we are already involved through participations to an ARC RLT (*Resources Linguistiques pour les TAGs* [1]) and to the working group A3CTE (*Applications, Apprentissage, Acquisition de Connaissances à partir de Textes Electroniques* [2]). For the European project TermIT [3], we also get some expertise in the domains of the conceptual structures such as thesaurus, ontologies, and semantic networks, and of the knowledge representation languages (conceptual graphs, descriptive logics). Finally, the group has worked on the issues of structuring and representing documents and linguistic resources (using for instance XML, and more anciently, using the systems Mentor and Centaur), issues that are relevant when processing encyclopedic corpus with a strong underlying structure.

The corpus of flora that we have examined seem to be promising for these tasks of acquisition. Indeed, they are strongly structured, following relatively precise patterns of presentation. Style is relatively formal in some parts (descriptive parts) and freer in others (explicative parts), allowing to test the dependence of acquisition w.r.t. style. More over, the corpus present a rich but very uncommon vocabulary to describe, for instance, colors, textures, and shapes. This vocabulary will not be found in available electronic dictionaries and will require a phase of extraction. Botanic conceptual models are not trivial but seems relatively well delimited and organized, having been developped over a long period of time. It means that a large knowledge can be instilled at the beginning of the acquisition process by an expert and that validation of extracted knowledge should be easier.

2 Program

We identify 4 main tasks.

2.1 Structuration of Corpus

The first step is the transformation of the original unstructured textual documents into a structured one, better adapted for further processing. The aim of the transformation is to delimit (with markup elements)

[1] `http://atoll.inria.fr/RLT/`
[2] `http://www-lipn.univ-paris13.fr/groupes-de-travail/A3CTE/`
[3] `http://www.mda.org.uk/term-it/`

the different entries and the different components of each entry (description part, explication part, bibliographic references, ...). The structured document may be encoded in XML, following a DTD (to be defined) characterizing this class of document.

The task may be achieved using scripts (Perl or Python) based on regular expressions.

Note that from the structured document may be derived an HTML version, adequate for on-line browsing. Such a browsable version would be helpful during the validation phases occurring in the following tasks.

2.2 Terminological Extraction

The first linguistic task consists in identifying the terminology associated to the corpus and building a list of terms with morphological variants.

This task may be achieved using scripts, morphological analyzing and terminological extractors (such as **FASTER** or **LEXTER**). Part-of-speech tagging and /or superficial parsing may also be needed to find the syntactic category (noun, adjective, ...) of the different terms (because we are dealing with a vocabulary not necessarily listed in available electronic dictionaries).

Note that the collected terminology may be used to efficiently index the corpus and speed-up searches in the on-line version.

2.3 Conceptual Acquisition

The next task would be the core of the research effort for ATOLL: we plan to organize the collected terminology into some kind of conceptual structure (such as a thesaurus, an ontology, or a semantic network). The result should model (more or less precisely) the concepts and relations between these concepts that emerge from the corpus. For instance, a flower is a kind of plant (*kind-of* relation) and has different parts (*part-of* relation); each part has different characteristics such as a color, a texture, a shape, a dimension, a number, ... A plant has also a geographic distribution (*location* relation).

The set of relations is very helpful to do inferences, in particular using heritage through the *kind-of* relation.

Of course, the modeling task may be helped a lot using a general model explicitly given by botanists. The acquisition process will complete and specialize this general model w.r.t. the corpus. In particular, it will assign semantic categories (such as color, shape, texture) to different terms (mostly adjectives).

This task being more prospective, we only sketch here a few ideas that should be investigated. We plan to run several passes of partial parsing and of knowledge acquisition. Each pass of parsing will use the knowledge already present to be more complete and precise than the previous one, in particular in resolving ambiguities. From the partial parse trees may be collected a set of properties and relations on concepts that will be forwarded to the knowledge acquisition module. This module organizes this set, looking what properties and relations seems to be pertinent w.r.t. the whole corpus (based for instance on statistical or ponderation criteria). The emerging properties and relations may then be proposed to a human supervisor for validation. A new round of acquisition may then be started.

The efficiency of the parsing passes will be helped by first identifying the recurrent syntactic and stylistic patterns used in the corpus. This is most specially true for the entry parts that follow a quasi controlled pattern.

2.4 Text Mining

The last task of the proposition is also a research task but will be eased by the quality of the conceptual modeling given by the previous task. The objective is to process each entry in the corpus in order to fill a database. The processing of an entry will be guided by the conceptual structure in order to identify where each piece of information that is found should go in the database. By some aspects, this task is similar to the previous one, except that the work is done at the level on an entry rather than on the whole corpus and can no longer rely on statistical methods to correct errors. Parsing backed by knowledge and validation by a supervisor will again be used to achieve this task.

The adequation of language Delta to encode the information extracted form each entry will be examined.

3 Objectives

The objective for ATOLL is to explore several research issues during this work in the area of parsing and acquisition. We also wish to establish a methodology that may used to deal with other corpus of encyclopedic information and other domains than botanic. Beside the research effort, this project will imply the development of prototypes and, hopefully, the creation of resources (structured corpus, terminology, conceptual structure, database). It may be noted that, among of different tools that are needed, one should think to the validation tools for the human supervisors.

We consider that this project requires a long term commitment of, for instance, a PhD student to coordinate the different tasks and to dialog with all the involved communities (botanists and computational linguists).

4 Related works

4.1 WordNet and EuroWordNet

WordNet is a *computational lexicon* that classify English words into sets of synonyms (*synsets*) that denote concepts. Synsets are themselves organized in a *kind-of* taxonomy. Addtionnal relations between synsets are present, such as the *part-of* relations. WordNet has been done by hand, but experiences to automatize the building of brother wordnets for European languages (Project EuroWordNet) are being conducted.

The importance of WordNet as a source of conceptual information for all kinds of linguistic processing has been recognized with many different experiences and specialized workshops.

4.2 MindNet

MindNet is a massive semantic network built by a Microsoft Research Team by automatically extracting knowledge from the Machine Readable Dictionary LDOCE and more recently from the encyclopedia Encarta. Knowledge is extracted using a robust parser based on the grammar checker of Microsoft Office bundle and incorporated in the semantic network. The nodes of the network are either concepts (such as `car`) or relations (such as `drive`) and edges are labeled by several kinds of lexical, syntactic, thematic or semantic relations (listed in Figure 1).

Attribute	Goal	Possessor
Cause	Hypernym	Purpose
Co-Agent	Location	Size
Color	Manner	Source
Deep_Object	Material	Subclass
Deep_Subject	Means	Synonym
Domain	Modifier	Time
Equivalent	Part	User

Figure 1: Semantic relations in MindNet

5 Background knowledge at INRIA

Other INRIA research groups (at Rocquencourt or at the other sites) may provide useful knowledge or tools in the context of this project, in particular with representing knowledge and associating documents to knowledge.

Verso [4] Databases and XML. They are involved in XYLEME [5] "The Data Warehouse for the XML Documents of the WEB" and with C-Web [6] "Supporting Community-Webs". They already have collaboration with CSIRO.

Samie A very recent group led by Alain Michard, one of the main promoter of C-Web [7].

References

[1] Thomas Ahlswede and Martha Evens. Parsing vs, text processing in the analysis of dictionary definitions. In *Proc. of ACL'88*, pages 217–224, 1988.

[2] H Alshawi. Analysing the dictionary definitions. In *Computational lexicography for natural language processing*, pages 153–169. Longman, 1989.

[3] A. Analyti, N. Syratos, and P. Constantopoulos. On the definition of semantic network semantics. Technical Report FORTH-ICS-TR-187, FORTH, Hellas, 1997.

[4] Doug Beeferman. Lexical discovery with an enriched semantic network. In Harabagiu [10], pages 135–141.

[5] J. Chang and J. Chen. Acquisition of computational-semantic lexicons from machine readable resources. In *ACL'96 Workshop on the Breadth and Depth of Semantic Lexicons*, 1996.

[6] Martin Chodorow, Roy J. Byrd, and George E. Heidorn. Extracting semantic hierarchies from a large on-line dictionary. In *Proc. of ACL'85*, pages 299–304, 1985.

[7] Giuseppe De Giacomo and Maurizio Lenzerini. A uniform framework for concept definitions in description logics. *Journal of Artificial Intelligence Research*, 6:87–110, 1997.

[8] Xavier Farreres, German Rigau, and Horacio Rodríguez. Using WordNets for building WordNets. In Harabagiu [10], pages 65–72.

[9] Thomas R. Gruber. Toward principles for the design of ontologies used for knowledge sharing. In *Formal Ontology in Conceptual Analysis and knowledge representation*. Kluwer Academic Press, 1993. also Technical Report KSL 93-04, Stanford University.

[10] Sanda Harabagiu, editor. *COLING-ACL'98 Workshop on "Usage of WordNet in Natural Language Processing Systems"*. Université de Montréal, 1998.

[11] M. Doerr I. Dionysiadou. Mapping of material culture to a semantic network. In *Proc. 1994 JOINT ANNUAL MEETING, International Council of Museums Documentation Committee and Computer Network*, 1994.

[12] N. Ide and J. Veronis. Extracting knowledge bases from machine-readable dictionaries. In *Proc. of KB&KS'93*, pages 257–266, 1993.

[13] Daniel Kayser. *La représentation des connaissances*. Hermes, 1997.

[14] J. Klavans, M. Chodorow, and N. Wacholder. From dictionary to knowledge base via taxinomy. In *Proc. of the sixth conf. of the University of Waterloo, Canada*, 1990.

[15] Oi Yee Kwong. Bridging the gap between dictionary and thesaurus. In *COLING-ACL'98* [26], pages 1487–1489.

[16] Claudia Leacok, Martin Chodorow, and George A. Miller. Using corpus statistics and WordNet relations for sense identification. *Computational Linguistics*, 24(1):147–166, Mars 1998.

[5] `http://www-rocq.inria.fr/verso/LEVEL1/Xyleme.html`
[6] `http://cweb.inria.fr/`
[7] `http://cweb.inria.fr/`

[17] J. Markowitz, T. Ahlswede, and M. Evens. Semantically significant patterns in dictionary definitions. In *Proc. of ACL'86*, pages 112–119, 1986.

[18] G. Miller. Five papers on WordNet. *Special issue of Int. Journal of Lexicography*, 3(4), 1990.

[19] S. Montemagni and L. Vanderwende. Structural pattern vs. string pattern for extracting semantic information from dictionaries. In *Proc. of ACL'92*, pages 546–552, 1992.

[20] Tom O'Hara, Kavi Mahes, and Sergei Nirenburg. Lexical acquisition with WordNet and the Mikrokosmos ontology. In Harabagiu [10], pages 94–101.

[21] Éric Villemonte de la Clergerie. Multilingual terminology production through an intermediate kn owledge level: Knowledge acquisition methods and techniques. Tâche 3.3.2 du Projet LE4-8356 Term–IT, devant être inclus da ns le document D3.1, June 1999.

[22] Stephen D. Richardson, William B. Dolan, and Lucy Vanderwende. MindNet: Acquiring and structuring semantinc information from text. In *COLING-ACL'98* [26], pages 1098–1102.

[23] C. Rigau, J. Atserias, and E. Agirre. Building acurate semantic taxinomies from MRDs. In *COLING-ACL'98* [26], pages 1103–1109.

[24] John Sowa. *Lexical Structures and Conceptual Structures*. Kluwer, 1989.

[25] John Sowa, editor. *Principles of Semantic Network*. Morgan Kaufman, 1991.

[26] Université de Montréal. *COLING-ACL'98*. Morgan Kaufmann Publishers, August 1998.

[27] P. Vossen, P. Diez-Orzas, and W. Peters. The multilingual design of EuroWordNet. In *Proc. of the ACL/EACL-97 workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP application*, 1997.

[28] Piek Vossen, editor. *EuroWordNet: A multilingual database with lexical semantic networks*. Kluwer Academic Publishers, 1998.

[29] Pierre Zweingenbaum and Jacques Bouaud. Construction d'une représentation sémantique en graphe conceptuels à partir d'une analyse LFG. In *Proc. of TALN'97*, 1997.