An Ontology Enrichment Method for a Pragmatic Information Extraction System gathering Data on Genetic Interactions

Claude Roux[1], **Denys Proux**[2], **François Rechenmann**[3] and **Laurent Julliard**[4]

Abstract. We present in this paper a method to insert new concepts in an existing information extraction system based on a conceptual graph architecture. We use verb patterns as conceptual sub-graphs to characterize unknown terms in sentences. The objective is to help an information extraction system to cope with unknown expressions contained in a specific semantic context.

1 Introduction

Internet pushes today the electronic document processing in a new era, where efficient methods to extract information from large collections are more than ever needed [2]. Today, statistical methods for text processing are fast and reliable but yield too often absurd results as no intelligence really controls the final output. Therefore ontologies are used to enlarge queries to words that are similar along semantic relations. One of the problems raised by this kind of approach is the creation and maintenance of these bases of concepts.

Several projects have started to provide ready to use electronic ontologies for automatic information processing system. One of the most famous is WordNet[5]. But generally these bases are designed for generalist purposes and do not match closely enough the needs of specific domains such as semi-conductor micro-electronic or genomics. An efficient information extraction system needs a customized ontology to fit its needs. We also need some procedures to identify and categorize unknown pieces of information. Systems must be able to update their knowledge resources in an autonomous way.

In the context of a research project in genomics that involves biology and computer science laboratories, we are developing an information extraction system using a linguistic and a knowledge processing approach. The goal is to feed an object-oriented knowledge base on molecular interactions with data on several organisms. The problem encountered in genomics is that there is no formal nomenclature for gene naming. This specificity conducted us to adopt an open architecture enabling automatic knowledge update.

We propose in this paper a system to categorize these terms with the joint utilization of an existing ontology and verb patterns defined as small conceptual graphs. In this system, an ontology is a two-fold set that contain *passive* information, the lattice, combined with *active* information, the graph patterns.

2 Project Overview

We have decided to deal with the information extraction problem using a pragmatic approach based on a combination of robust technologies that have proved their efficiency. Our system has adopted a two-levels architecture: a linguistic analysis and a knowledge based processing.

The linguistic components are mainly based on the Finite State Transducer technology known to be very efficient in high speed text parsing [3]. It consists of a Part of Speech tagger that has been slightly customized for our domain specific corpus, and a Robust Parser [1]. Texts are processed in several step: tokenization, morphological analysis [5], disambiguation using an Hidden Markov Model technique [4], error corrections, and finally a contextual lookup to identify gene names. Syntactic dependencies are then extracted from that output.

On the knowledge processing side of the system we have chosen to adopt an architecture based on conceptual graphs [6]. The syntactic dependencies extracted at the linguistic level are used to build the semantic representation.

Nevertheless that architecture does not cover all our need. Hence, the adoption of a mechanism that provides an enrichment of our ontology using a set of learning conceptual graphs.

3 Architecture

The system has been designed as a multi-layered architecture where specialized modules may be inserted or removed according to the task. At the top of that architecture, we place a Robust Parser and a Conceptual Graph manager to use those data in a more elaborate way than just pattern matching in contrast with systems that provide a unified architecture such as the Loom system[6].

Most of the tools we use in this project have been developed in Xerox laboratories. These tools include a tokenizer and a morpho-analyzer with a high coverage of the English language. We use the Incremental Finite-State Parser developed by Ait-Mohktar and Chanod for English. We have built a complete architecture to use all those components as smoothly as possible. The conceptual graph manager is also a Xerox technology and has been custumized to handle a very large lattice of concepts (up to 30000).

[1] Xerox Research Centre Europe, 6 Chemin de Maupertuis, 38000 Meylan, France, email: Claude.Roux@xrce.xerox.com

[2] Xerox Research Centre Europe, 6 Chemin de Maupertuis, 38000 Meylan, France, email: Denys.Proux@xrce.xerox.com

[3] INRIA Rhne-Alpes, 655 avenue de l'Europe, 38330 Montbonnot, France, email: Francois.Rechenmann@inria.fr

[4] Xerox Research Centre Europe, 6 Chemin de Maupertuis, 38000 Meylan, France, email: Laurent.Julliard@xrce.xerox.com

[5] http://www.cogsci.princeton.edu/ wn/

[6] http://www.isi.edu/isd/LOOM/LOOM-HOME.html

3.1 The Robust Parser

The Robust Parser, we utilize in this project, yields for a given sentence the list of the most salient syntactic functions that link words in that sentence. For example, the sentence: "*Antp protein represses the BicD gene.*", yields the following list of syntactic functions:

SUBJECT	(repress, protein)
OBJECT	(repress, gene)

That set of syntactic functions is then mapped over a list of relations, in the guise of conceptual graphs.

3.2 The Conceptual Graph Manager

The next module is a conceptual graph manager. Its goal is to store information as semantic graphs, where concepts represent words or terms and are connected to other concepts through relations. The concepts are described in a lattice where each node is related semantically to the other nodes along an "isA" relation. Traditionnaly, the verb in a sentence is considered to be the key element as it generally indicates the kind of action involved. Therefore it is placed at the top of the conceptual graph structure symbolizing the sentence. Nouns occuring in subject or object groups, are connected to this verb through links representing their syntactic relation.

For example, we could represent the fact: "*Antp protein represses the BicD gene.*", with the following graph.

— (agent) — protein — (Related-To) — Antp

represses |

— (target) — gene — (Related-To) — BicD

3.3 New Concepts

This work aims at building automatic ontologies where concepts are inserted in a lattice according to the verb those concepts are connected to in a text [7]. A concept is not simply a word, it can be an expression or a proper noun. For instance, in the system we present here, one of the modules is designed to recognize a certain sequence of symbols as a gene name. This gene name is then handled as one single concept and not as a list of its components.

We assume that a text contains two main sorts of data:

- First, data that are easily identified by their immediate context. Small-specialized modules that are based on simple pattern matching can then be utilized to handle those data.
- Second, data that are more remotely coordinated and which can only be extracted with more refined techniques that take into account syntactic dependencies. The results can then be constrained by the data extracted by the first modules.

When a new word appears in the text that has not been yet referenced as a concept in the lattice, the problem consists of adding this word to the lattice. As the lattice comprises concepts that are connected to each other along semantic paths, this imposes to categorize this new concept in order to find its correct slot in the lattice.

The system we propose here manages new concepts with a specific device to detect certain configurations of verbs that will assign their position in the lattice. Those configurations are specific graphs in which one of the concept is a verb that expects certain semantic attributes.

For example the verb "to repress" associated with a "protein" type concept usually expects its target to be a gene type concept.

If more than one graph fires, then all the possible meanings that these graphs cover will be attached to the new concept. In order to exploit the mechanism of projection, the "empty slot" concept is the "Universal" concept. The "Universal" concept is the most general concept in the hierarchy (located at the top of the lattice) so that it can match with any other. Types associated to the "empty slot" concept will serve to connect the unknown term that matches in a specific semantic context, under corresponding nodes inside the ontology.

Example:
Learning Graph (1)

— (agent) — protein

represses |

— (target) — Universal: [gene]

Input sentence : *Antp protein represses BicD.*

Sentence Graph (2)

— (agent) — protein — (Related-To) — Antp

represses |

— (target) — BicD

The new concept is "BicD".

We project the graph (1) on the graph (2). "repress" matches "repress", "protein" matches "protein", "Universal" matches everything and in this case "BicD".

The special value "gene" that is associated to the node "Universal" in graph (1) is then used to categorize the new concept "BicD" which will be appended in the lattice under the node "gene".

In this simple example the target of the learning sub-graph is gene names, but it can of course be applied to more complex expressions.

4 Conclusions

We present an information extraction system that relies on a linguistic and a knowledge processing approach. The linguistic tools we use are based on the Finite State Transducer technology, combining a Part Of Speech tagger and a Robust Parser. The extraction mechanism is built upon a conceptual graph architecture, in conjunction with a domain specific ontology to improve its efficiency. To solve the problem of the lack of coverage of ontologies for domain specific applications we have decided to implement a learning mechanism to automatically update our hierarchy of concepts. This method is based on the projection of customized conceptual sub-graphs with typed concepts on specific nodes. The objective is to cope with new or unknown domain specific expressions. The mechanism is currently under development and tests are planned to validate this approach.

REFERENCES

[1] S. Ait-Mokhtar and J.P. Chanod, 'Incremental finite-state parsing.', in *Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP'97)*, pp. 72–79, Washington, USA, (March 31st to April 3rd 1997).

[2] D. Appelt, 'Introduction to information extraction', *Artificial Intelligence Communications*, **12**, 161–172, (1999).

[3] J. Hobbs, D. Appelt, J. Bear, D. Israel, and M. Tyson, *FASTUS: A Cascaded Finite-State Transducer for Extracting Information from Natural-Language Text*, MIT Press, Cambridge MA., 1996.

[4] J. Kupiec, ‘Robust part-of-speech tagging using a hidden markov model’, *In Journal of Computer Speech and Language*, **6**, (1992).

[5] A. Schiller, ‘Multilingual part-of-speech tagging and noun phrase markup’, in *Proceedings of the 15th European Conference on Grammar and Lexicon of Romance Languages*, Munich, Germany, (September 19th to 21st 1996).

[6] J.F. Sowa, *Conceptual Structures. Information Processing in Mind and Machine*, Reading, Mass.: Addison-Wesley, 1984.

[7] P. Wiemer-Hastings. Automatic acquisition of word meaning from context. University of Michigan, Dissertation, 1994.