HOW TO SEARCH THE FUTURE WEB

Marek Schmidt
Master Degree Programme (2), FIT BUT
E-mail: xschmi01@stud.fit.vutbr.cz

Supervised by: Pavel Smrž
E-mail: smrz@fit.vutbr.cz

ABSTRACT

The research described in this paper deals with the Semantic web – the vision of a future web accessible not only for humans but also for automatic processing by machines. The paper also discusses that this ultimate goal cannot be accomplished without a semi-automatic transformation of plain text into a structured form. It employs XML technologies such as XSLT, XML database and XQuery to manipulate, store and query semantically enriched dependency trees obtained by syntactic analysis and demonstrates an advantage of this approach.

1 INTRODUCTION

The emerging technologies around the Semantic Web concept try to bring the power of structured information to the World Wide Web. Such technologies, if widely used, would allow computer agents and search engines to actually understand the data on the web and to perform reasoning to infer useful information out of different data sources.

There are currently not many web sites providing structured information, while there are billions of pages on the WWW which contain unstructured information. People should not be forced to change their ways of publishing information. Instead, new technologies for automatic information extraction should be developed to bring the power of Semantic Web technologies into the 'old' WWWW.

This work explores a way of using available natural language processing tools to transform unstructured information into a form processable by standard XML technologies.

2 SYSTEM ARCHITECTURE

The processing of the input text data has three steps:

1. *Syntactic analysis*, which parses the text with a dependency parser.
2. *Semantic annotation*, which adds semantic information to pieces of text.
3. *Storage, indexing and query*, which stores the data in an XML database.

2.1 SYNTACTIC ANALYSIS

Syntactic analysis is the process of creating syntactic trees out of plain text. The MiniPar parser for the English language [4] has been used in this project. This parser transforms plain text sentences into dependency trees, in which nodes contain a lemma of the original word and a part of speech category assigned by the parser, while edges describe syntactic relationships between two words in the sentence.

The output of the MiniPar parser is then transformed to an XML based format. A pipeline of XSLT transformations is then applied to the tree with the purpose of simplifying the syntactic structure of sentences.

2.2 SEMANTIC ANNOTATIONS OF WORDS

The meaning can be given to words if they can be related to by other words and by structured data. We define these basic requirements for a mapping between words and their meaning:

- Each sense shall have a unique identifier
- When referring to a word, we shall implicitly cover all its synonyms (ie. *car* → *automobile*) and all its hyponyms (ie. *car* → *jeep*)

The WordNet lexical database of English [2] has been used as the mapping between words and their meaning. The basic element of the WordNet is a synonym set (synset). Each synset has an unique ID. Each word in WordNet may be a part of several synsets, which brings ambiguity to the system.

All synonyms have the same synset identifier in WordNet. A transitive closure of a hypernymy relation (which is inverse to the hyponymy relation) is computed and stored in each node.

Proper nouns, such as names identifying a person, location or a company, are generally not represented in the WordNet. Such words need to be annotated separately by a named entity recognizer, which recognizes named entities in a text and classifies them into ‘person’, ‘location’ or ‘company’ category.

2.3 XML DATABASE

The Oracle Berkeley DB XML database [1] have been chosen as the storage and query engine. This database supports indexing element and attribute values which have significant effect on search performance for some queries.

The standard XQuery language can be used to perform queries. While there are languages arguably more suitable for linguistic queries [3], the use of XML allows for interoperability with other data sources.

3 DATA

The system has been developed and tested on samples from English Gigaword corpus, which comprises English newswire articles. The data have been split into separate packages of approx. 100.000 sentences to allow parallel processing.

The system has been used to perform queries for gathering statistics of linguistic features of WordNet concepts. Table 1 shows the result of a simple query to extract persons in a role of a head of a state, based on a trivial syntactic pattern (a person node and noun-noun modifier dependent on a head-of-state concept).

Occurrences	Country	Title	Name
102	US	president	Bill Clinton
26	Ukrainian	president	Leonid Kravchuk
24	French	president	Francois Mitterrand
16	South African	president	Nelson Mandela
5	Chinese	premier	Li Peng

Table 1: Subset of results for the query based on search for head-of-state WordNet concept. (The news articles were from the year 1994)

The problem in searching for semantic relations between concepts lies in the fact, that there are many possible syntactic ways to express a particular semantic relation. A proof-of-concept tool to expand the query by similar syntactic patterns has been developed as part of this project [5]

4 CONCLUSION

A system which transforms plain text documents into a structured form based on syntax analysis has been implemented. It is already used for extracting statistics of linguistic features of words and concepts. Further methods for semantic relations extraction will be developed and implemented before the system will be truly useful for making structured information out of unstructured text sources. The future work will focus on automatic expansion of syntactic queries.

REFERENCES

[1] *Oracle Berkeley DB XML.*
URL <http://www.oracle.com/database/berkeley-db/xml/>

[2] *WordNet: An Electronic Lexical Database.*
URL <http://wordnet.princeton.edu/>

[3] Lai, C.: A Formal Framework for Linguistic Tree Query. 2006.
URL <http://eprints.unimelb.edu.au/archive/00001594/>

[4] Lin, D.: *MiniPar: broad-coverage parser for the English language.*
URL <http://www.cs.ualberta.ca/~lindek/minipar.htm>

[5] Lin, D.; Pantel, P.: DIRT discovery of inference rules from text. In *Knowledge Discovery and Data Mining*, 2001, s. 323–328.
URL <citeseer.ist.psu.edu/lin01dirt.html>