Some Methods of Texts Semantic Search by UNL Expressions Logical Analysis

Aram Avetisyan

Institute for Informatics and
Automation Problems of NAS RA
Yerevan, Armenia
e-mail:
a.avetisyan@undlfoundation.org

ABSTRACT

In this paper we present some research results and propose solutions for natural language texts semantic analysis.
A method is suggested for searching semantic structures corresponding to the given linguistic pattern. This method is based on natural language text analysis using UNL technologies [1]. As a result some experimental questions answering algorithm is created.

Keywords

UNL, natural language processing, indexing, search, pattern matching algorithm, semantic search.

1. INTRODUCTION

UNL (Universal Networking Language) is a meta-language representing semantics [1]. Its main purpose is to describe "the meaning" of natural language texts in a language independent format. Each sentence in UNL is a directed linked graph. Unlike natural languages, UNL expressions are not ambiguous. In the UNL semantic networks, nodes represent concepts, and arcs represent relations between concepts. These concepts are called "Universal words" (UWs). The UWs' connections are called "relations". They specify the role of each word in the sentence.

Many UNL centers and the UNDL Foundation itself have created a number of tools for working with UNL and Natural Language (NL) resources. In this paper we are highlighting some aspects of the development of algorithms for natural language text analysis and semantic search. The core tools of the project are the NL Analyzer and NL Generator [2] developed by the UNDL Foundation.

2. The problem

Currently available search engines return results of exact matches in the text, while the most advanced engines may perform some analysis and bring up results that contain words semantically or statistically (more common) related to the search phrase.
Since 1960s question answering search engines researches were started along with some experimental developments, such as ELIZA, DOCTOR, BASEBALL, LUNAR etc. [3] – [5]. All those engines were based on simple text pattern matching techniques, thus the result quality remained instable.
We implemented an experimental project aimed to create such engine, although the main aim of the project was to prove that UNL is a tool powerful enough to be used in the most sophisticated text analysis tasks.

3. UNL implementation

The algorithmic solutions developed by the UNDL Foundation can perform in-depth semantic and syntactic analysis of natural language texts. This gives us an opportunity for creation of language-independent question answering engine algorithms.
There are four main steps developed in the project.

Step 1. Natural Language Text Analysis

Before starting to search for a question answer we need to "understand" what information we have. For this reason the natural language text resources provided in advance must be analyzed. So in this step the NL Analyzer analyzes the text and creates UNL graphs for each sentence and stores them in database. Depending on the language, the corresponding dictionaries and grammar rules must be provided. The final result accuracy depends on these resources' quality.
After that, the indexing process starts. During the indexing process all the UWs found in the graphs are being indexed according to their appearances in the sentences. This data is also stored in the database.
When the first step is finished, as a result we have database containing indexed and completely analyzed language-independent text presented in a form of graph.

Step 2. Natural Language Question Analysis

The second step performs similar to the first step. The user provided question is considered and analyzed as a usual natural language sentence. As a result, we receiveve a semantic UNL graph. After that, the structure of the graph helps to identify the topic and the target of the question. Using that information a new, filtered from non-important information (some attributes or even relations) graph is created, containing a blank space for the target node. This new graph will be the semantic pattern to be matched.

Step 3. Pattern matching

In the third step the pattern matching process starts. First, using the indexed UWs, a list of UNL graphs are retrieved from the database as potential matches. After that the main matching process starts.
The highest priority is given to the direct matches that indicate a full pattern match. Then the algorithm retrieves graphs that contain similar to the pattern relations (the relation UWs are not ignored) including the relation(s) for the topic of the question and relation(s) containing the target UW.
If the matched graph's target UW is not a satisfying match (e.g. a pronoun "he"), the sibling sentence graphs are being analyzed to determine the most possible UW

(e.g. the algorithm searches for a noun that is indicated as a male person).

Examples
Q: "*Who took the green apple?"*

- Text contains a sentence "*Jack took the green apple from the basket.*"
  The algorithm will find a full match.
- Another sample when the text contains the following: "*There was a green apple on the table. Jack took it*".
  In this case the algorithm can't find a full match, but sees that "Jack took *(something)*". Now the algorithm has to find out what that "something" is and whether it is what we are looking for. When it analyzes the sibling sentence (the left sentence is always analyzed first) it finds two UWs that satisfy the main requirement: "*apple*" and "*table*", but since the second sentence refers to that object as a pronoun, then the target noun must be indicated as an entry node in the sibling sentence, thus the "*table*" cannot be considered.
- In the third case the text contains "*Jack saw a basket, full of fruits. He took a green apple and walked away.*"
  Here the algorithm will match the second sentence, but will continue the search to find out who "***he***" is. As "*Jack*" is the entry node of the sibling sentence, and is the only noun that can be indicated as "*he*", the algorithm considers this match satisfying.

Step 4. Answer generation
In the third step the algorithm found the satisfying answer as an original sentence UNL graph from the text or as a UNL sentence graph generated from two or more sibling UNL graphs from the text.
In the fourth step the result is passed to the NL Generation process [2], which generates a corresponding natural language sentence. This sentence is provided to the user as a most probable answer along with other less probable matches found during the process.

4. Conclusion

As it is presented in the examples above ("UNL implementation. Step 3."), the algorithm performs fairly well with simple questions targeted to a plain text. This can be the first step in development of an advanced text analyzing algorithm that may find its usage in question answering, text summary generation or retrieval of information from a large amount of text data. One of the key features of using UNL analyzing algorithms is the language-independent platform, which guarantees a great flexibility in terms of applications.

REFERENCES

[1] H. Uchida, M. Zhu, "The Universal Networking Language (UNL) specifications" version 7, UNDL Foundation, June, 2005.

[2] A. Avetisyan, "Development and Application of Interactive Algorithms for Natural Language to UNL and UNL to Natural Language Transformations" Mathematical Problems of Computer Science 34, IIAP NAS RA, 2011.

[3] F. Bert, Green Jr. et al., "Baseball: an automatic question-answerer" western joint IRE-AIEE-ACM computer conference. ACM, 1961.

[4] A. Woods, et al. "The Lunar Sciences Natural Language Information System." Cambridge, Mass., 1972.

[5] J. Weizenbaum, "ELIZA - A Computer Program For the Study of Natural Language Communication Between Man And Machine", Communications of the ACM, MIT, Cambridge, 1966.