OntosMiner Family: Multilingual IE Systems

Irina V. Efimenko, Vladimir F. Khoroshovsky, Victor P. Klintsov

AviComp AG, Moscow, Russia, 84/2, Vernadskogo Av., Russia, 119606
iefimenko@avicomp.ru, vkhoro@avicomp.ru, klintsov@avicomp.ru

Abstract

This article deals with the problem of Multilingual Information Extraction (MIE), which is one of the topics of the day since it is within the scope of interest for many international projects.

Presented in the article, OntosMiner Family Systems of IE are developed by Russian IT-company AviComp AG together with Ontos AG (Switzerland). AviComp AG is interested in various projects involving Natural Language Processing tasks such as Information Extraction, Semantic Documents Clustering, etc.

1. Introduction

Natural language processing is among the issues of the day both for research centers and IT industry. Special attention of researchers and developers of systems intended for practical use is paid to such tasks as Information Extraction (IE), Text Mining (TM), Semantic Documents Clustering (SDC) and Semantic Texts Classification (STC) [1].

Natural language text processing is the midpoint of knowledge management life cycle. In real practice, for all that, texts intended for processing come without any pre-editing and may contain errors or have some other characteristics, complicating their analysis. The NLP-system task is to get an adequate representation of the text meaning. Here we deal with the task of retrieving information important for a user, which is the IE-task. Among the systems of particular interest and importance one can mention IE-systems for multilingual document collections, including Internet documents, corporate and personal databases, news, etc.

The purpose of this article is to discuss OntosMiner systems, which are MIE systems. The development and implementation of OntosMiner family systems is carried out by Russian IT-company Avicomp AG\(^1\) together with Ontos AG\(^2\) (Switzerland).

2. State-of-Art in the NLP Area

2.1. NLP tasks and approaches

There are two basic groups of approaches to natural language processing: statistic and linguistic ones. Despite a number of difficulties, linguistic approaches seem to be more powerful. Moreover, they allow a possibility of statistic post-processing. The most promising approach is a hybrid one.

Main problems faced during natural language text processing can be divided into three groups:

- Formatting problems.
- Syntactic analysis problems.
- Semantic interpretation problems.

At first sight, the formatting problems seem to be solved since there are special filters transforming all most common formats to a plain text. However, it is not really so. Real texts (e.g. Internet-documents) often contain fragments that can not be correctly handled by existing filters. Therefore special filter development is required.

Syntactic analysis problems appear, first of all, from the complexity of syntactic constructions used in real texts, as well as from errors, misprints, etc. Nowadays linguistic science has no general models allowing construction of a syntactic analyzer for a text from an arbitrary domain. At the same time there is already a number of practically significant methods and algorithms for various syntactic patterns processing.

Anaphoric references processing can be mentioned as an example of semantic interpretation problem. An example of another type is a selection of correct interpretation for a polysemantic concept. The problems of this type are being solved nowadays through the use of domain models. One of the most up-to-date approaches is the domain ontology use.

All the existing NLP-systems can be classified according to the following characteristics:

- Domain restrictions

\(^1\) http://www.avicom.ru
\(^2\) http://www.ontos.ch
3. OntosMiner IE Systems

3.1. Basic steps of IE-task solution

The standard set of basic steps, providing an acceptable quality of IE-task solution, includes the following ones:

- **lexical normalization** (text formatting, basic text elements determination, text segmentation into sentences, paragraphs, etc.);
- **morphological analysis**;
- **syntactic analysis** (usually for separate sentences) with representation of results in a graphic form;
- **semantic analysis** of results from previous stages and transformation of a language-dependent “surface” representation to a language-independent one;
- **semantic interpretation** and comparison of the processing results with the knowledge about the domain.

3.2. Basic principles and features of OntosMiner IE Systems

In real practice, one often deals with multilingual document collection. Various language insertions in monolingual (in general) documents is also a typical case (e.g. names of organizations; medical product names - in Latin, etc.). For the above reasons, the OntosMiner project is Multilingual Information Extraction oriented. It is based on a number of principles. First of all, the OntosMiner family systems are NOT intended for solving a problem of FULL analysis of arbitrary natural texts. We formulate the problem in such a way to process those constructions, that can be processed correctly, and NOT to process those ones, that still can not be processed correctly. This model is based on the fact that people usually deal with a great number of informational materials describing the same event, situation or process, so one can get an adequate integral view even missing some of the materials. On the other hand, wrong results of some particular text analysis can distort the general picture and make further processing senseless.

The next principle, that the OntosMiner project is based on, is that of *reusable component* development for multi-platform implementation. Such components integration within the powerful environment allows fast development and implementation of IE systems for concrete domains.

The last but not least principle postulates the domain *ontology-driven analysis*.

So, the main requirements for OntosMiner family systems are the following:

- Work with multilingual document collection (at present OntosMiner systems deal with English, French, German and Russian texts).
- Work with monothematic document collection (at present it is, first of all, the so-called “Business Duties” domain. Such collections include informational materials about IT-companies, analytical materials about founding, investing, selling, buying, merging of companies, top-management CVs, etc.).
- An adequate processing of relevant objects and relations, according to the concrete ontology.
- Representation of processing results in a form of a cognitive map, that is a kind of semantic network.
- Multi-platform implementation of all systems of the family.

3.3. OntosMiner Systems architecture

According to the above mentioned principles, the NLP-systems development technology was developed and patented within the OntosMiner project. It is based on the powerful multi-platform GATE (General Architecture for Text Engineering) environment, that was created in Sheffield University, Great Britain [2]. In the context of this technology, some GATE components were modified. In addition to that, some new components were developed and implemented. They allow:

- Knowledge representation in ontology form and convenient handling of such knowledge;
- Information retrieval from Internet with the help of special agents addressing to the most common search engines with the posterior use of Oracle DataWarehouse;
- Linearization of various format texts with Oracle 9i filters.

It is also worth mentioning that in the last GATE version the Debugger developed by Avicomp AG appears as one of components.
In addition to the above components, a number of special components for NLP problem solution were developed and implemented. They allow:
- lexical text formatting which provides elementary units of the text for its further processing;
- text segmentation into sentences;
- morphological analysis;
- processing of predicate constructions (verbs, various analytic forms);
- processing of domain relevant Named Entities (NE);
- ontology-driven semantic analysis and such representation of the results from previous stages which makes it possible to generate cognitive maps, describing text meaning.

The common architecture of OntosMiner family systems is presented in figure 1.

There are three groups of components in this diagram: a) language-independent components; b) components that are partially language-dependent; c) components unique for each language.

For the block of morphological analysis two components were developed and implemented. One of them deals with inflection in English, providing processing of 56000 dictionary entries (Müller dictionary). Another one provides processing of 98000 Russian dictionary entries, as well as analysis of unknown words, based on heuristics connected to Russian word-formation models.

Named Entities, which are processed by the OntosMiner systems, include the following:
- People;
- Organizations;
- Titles and Job titles;
- Scientific degrees;
- Various kinds of Addresses;
- Money;
- Locations;
- Cars;
- Dates and Periods of Time.

In the context of the OntosMiner project, a linguistic model of Time was developed and implemented for English, German, French and Russian languages. Classification of time entities is presented in table 1.

As for semantic analysis and interpretation, it was already mentioned that OntosMiner systems work within “Business Duties” domain. They also deal with a forensic domain and work with collections of police reports. Relations between Named Entities, which are processed by the OntosMiner systems, include the following:
- Affiliate;
- Buy-Sell;
- Employ;
- Found;
- Graduate;
- Invest;
- Joint Venture;
- Own;
- Rival;
- Locate In;
- Reside;
- Belong;
- Hijack;
- Investigate

Processing results are represented in the form of XML documents that specify cognitive maps. Such maps can be integrated into knowledge bases and visualized according to user demands and preferences.

3.4. Future trends

Our plans for future include the following:
- language spectrum enhancement;
- domain spectrum enhancement;
- use of more sophisticated domain models and increase of the number of entities and relations intended for processing;
- use of more sophisticated linguistic models and processing of more sophisticated constructions;
- use of IE results in Information Retrieval, Summarization and Question&Answering systems.

4. Conclusion

Presented in the article, OntosMiner Family Systems are MIE systems of practical significance. Enhancement of system facilities, as well as collaboration with other IT-companies, are among the plans for future

5. References


“Figure 1. OntosMiner Systems architecture”
<table>
<thead>
<tr>
<th>Time NE major type</th>
<th>Time NE minor type</th>
<th>Time NE minor subtype</th>
<th>Attributes</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>abstract</td>
<td>date class</td>
<td>on [Easter]date class:Easter day (En)</td>
</tr>
<tr>
<td></td>
<td>historical point</td>
<td>concrete</td>
<td>event</td>
<td>в день [избрания В.В.Путина]event: Putin election (Ru)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>abstract</td>
<td>event class</td>
<td>в день [выборов]event class:elections (Ru)</td>
</tr>
<tr>
<td></td>
<td>moment-of-speech oriented point</td>
<td>concrete</td>
<td>distance (number + unit, direction, unit kind)</td>
<td>[il y a]direction:past [quatre années]distance: number:4, unit:year (Fr); das [letzte]direction:past, distance:last before now [Jahr]unit kind:year (Ge); [yesterday]direction:past, distance:last before now, unit kind:day (En)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>abstract</td>
<td>direction, remoteness</td>
<td>недавно]direction:past, remotness:small (Ru)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>start-point&amp; length period</td>
<td>start point, length (number + unit)</td>
<td>[four years]length: number:4, unit:year</td>
</tr>
<tr>
<td></td>
<td></td>
<td>end-point&amp; length period</td>
<td>end point, length (number + unit)</td>
<td>в течение [нескольких лет]length: number:several, unit: year, преждествовавших [объединении компаний]end point:joint venture (Ru)</td>
</tr>
<tr>
<td></td>
<td>period with length</td>
<td>length (number + unit)</td>
<td>für [zwei]number:2 [Monaten]unit:month]length: number:2, unit: month (Ge)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>open period</td>
<td>period with start point</td>
<td>start point</td>
<td>à compter du [1er]day:1 [janvier]month:1 [2003]year:2003 start point:1.1.2003 (Fr)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>period with end point</td>
<td>end point</td>
<td>before [now]end point:now (En)</td>
</tr>
</tbody>
</table>