Use of natural language processing methods for scientific papers

I.E. Nikolaev
Institute of Information Technology
Chelyabinsk State University
Chelyabinsk, Russia
e-mail: criticjim@gmail.com

Abstract[1]

In the article consider two approaches of processing natural language text, as well as the possibility of their use for processing of scientific publications.

1. Introduction

High-tech development, technology and software advance which facilitate production of information products has led to a dramatic increase in quantity of information and knowledge that a man has to handle in course of his professional activity.

Science theorist believe that it took 1750 years since the beginning of our era to double the stock of knowledge, then it doubled for the second time in 1900 and for the third time in 1950, i.e. already in 50 years while the quantity of information grew 8-10 times within these 50 years. And this tendency is strengthening since the quantity of knowledge grew 2 times and the quantity of information 30 times by the end of the XX century.

An analogy can be drawn between industry and intelligent information production. Overproduction takes place both in industry and information sphere. There appears too much information. One explanation to it is that any person capable of brainwork can produce information. Nowadays anyone has an opportunity to publish their thoughts and conclusions in any available way – in mass media or on the Internet.

Incredible it may seem, but such avalanche of information has led to the information deprivation when information quantity many times exceeds physiological capability to receive, process and extract necessary information from the overall information stream [2].

The problem of information deprivation was one of the factors to induce development of cognitive methods of **natural language processing (NLP)**. This field of science aims at extracting qualitative information (information possessing such qualitative characteristic as accuracy, adequacy, relevance, completeness) from natural language input.

Recently the problem of information deprivation has become urgent in the sphere of scientific papers. It may be explained by reduction of knowledge production cycle. Nowadays big information repositories contain a huge number of scientific papers in various fields and their number is constantly increasing. A person has to spend immense amount of time to find and examine in detail papers necessary for his professional activity.

The problem of extraction of a qualitative information product based on cognitive methods of NLP is one of the most relevant top-priority tasks in scientific research. The solution of this problem will allow finding necessary texts in the shortest time, carrying out an expertise of scientific papers, revealing contradictions in them and thus revealing further scientific problems demanding solution.

Now 2 approaches to NLP can be singled out:

- Semi-automatic processing (with human participation in case of questionable and controversial situations);
- Automatic processing (with no or minimal human participation in case of questionable and controversial situations).

Let us consider in detail main blocks and sequence of semi-automatic NLP stages. The first block which the input meets is the block of lexical analysis. The processing in this block is carried out at several stages [3]:

- The stage of partitioning of the input (i.e. text) into separate sentences, word combinations and words (fig. 1. (1.1));
- The stage of morphological (analysis of each word and determination its morphological characteristics – part of speech, case, declension, conjugation, etc.) and morphemic analysis (prefix, root, suffix, ending). (fig.1. (1.2));

[1] **Proceedings of the 14th international workshop on computer science and information technologies CSIT'2012, Ufa – Hamburg – Norwegian Fjords, 2012**

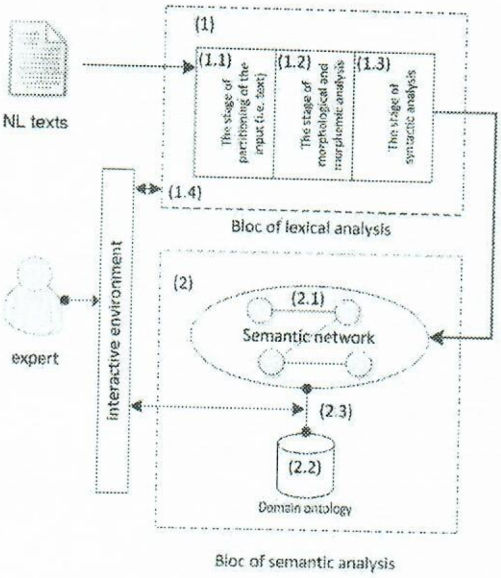
(1)
(1.1)
The stage of partitioning of the input (i.e. text)
(1.2)
The stage of morphological and morphemic analysis
(1.3)
The stage of syntactic analysis
NL texts
(1.4)
Bloc of lexical analysis
interactive environment
expert
(2)
(2.1)
Semantic network
(2.3)
(2.2)
Domain ontology
Bloc of semantic analysis

Fig. 1. Diagram of semi-automatic language processing

- The stage of syntactic analysis (punctuation marks, etc.) (fig.1.(1.3)).

At this stage separate lexical units, words, word combinations, speech patterns are determined and preliminary joined according to the rules of a certain language.

The next stage is semantic analysis. At this stage a semantic network, which reflects interaction between objects and object properties of a domain is formed. It should be mentioned that semantic networks nowadays are de facto standard of representation of domain semantics.

The process of semantic network formation is accompanied by a check of forming relations (object-object, object-property) to evaluate conformity with the domain rules which usually are stored in form of a database of all possible combinations of object-object and object-property relations. Such databases are called *domain ontology*.

A domain ontology is usually formed for a definite domain by a human expert or a group of experts. Experts describe *all possible* combinations of relations object-object, object-property or in other words they describe semantic (conceptual) rules of domain structure.

Unfortunately, in practice some information sources, namely NL texts – scientific literature, mass media, etc. – may contain actual contradictions and mistakes.

2 kinds of mistakes can be singled out: textual mistakes and semantic contradictions (semantic collisions, knowledge collision).

The following can be referred to textual mistakes:

- Misprints;
- Orthographic mistakes – incorrect spelling of a word;
- Grammatical mistake – using an incorrect form of a word;
- Stylistic mistake – inappropriate using of a form of a word.

Now there have been developed and are widely used software tools (e.g. Microsoft Word, search engines Google, Yandex, etc.) which allow to reveal the most of textual mistakes (fig.1 (1.4)).

The second kind of mistakes is semantic contradictions. Generally any non-conformity with domain semantics rules or in other words discrepancy with the relations object-object, object-property in the domain ontology (problem of ambiguity) can be referred to this kind.

Engaging experts for domain ontology formation and solving semantic contradictions at this stage of development is necessary, reasonable and at times non-negotiable and is explained by the fact that in some cases usage of extra-linguistic information which is not extracted directly from the text(s) under analysis is essential to solve some of semantic contradictions. In number of tasks sensitive to semantic accuracy (analysis of documents, meaning extraction, in deep semantic analysis) this rule is not debatable. However, are these "unsolvable" contradictions so significant in more practical tasks (abstracting, text clustering, annotating) of NLP systems? This question is a subject for further studies.

Of no little interest is the fact that semantic analysis is based on the results of previous stage of text processing, namely on the information extracted at the stage of linguistic and grammatical analysis. Recently a number of new methods of information extraction from NL texts have been devised in the field of computational linguistics. These methods can give additional meta-information from NL texts which in its turn can be effective for contradictions solving at the stage of semantic analysis. That is why analysis of latest achievements of computational linguistics should be carried out.

Basing on the above mentioned facts, we suggest a different structure of NLP (see fig.2).

As can be seen in fig.2 the structure of language processing is changed in comparison with the semi-automatic model. Participation of a human here is minimal, he accounts for NL texts selection and controls processing at the stage of linguistic analysis, which is virtually censoring of the input and correction of textual mistakes.

Semantic analysis in this model is supposed to be fully automatic, with no human participation, based on frequency analysis and information acquired from the block of linguistic analysis. In this connection the structure and the way of ontology formation is being completely reviewed, i.e. it doesn't include any extra-linguistic information but facts and knowledge in form of object-object, object-property relations extracted from other NL texts.

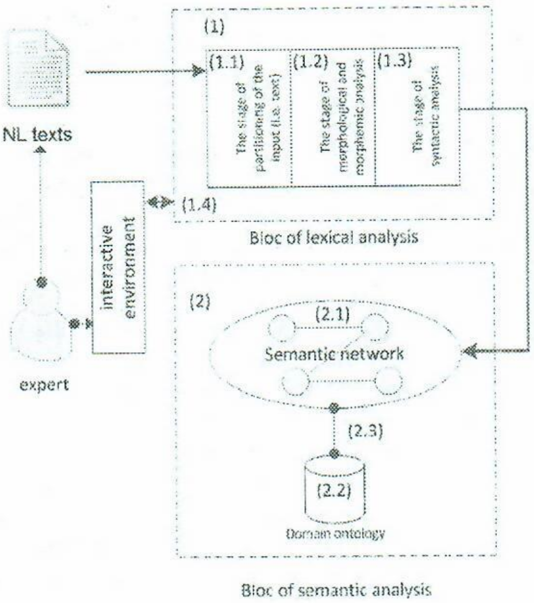
(1)
(1.1)
The stage of partitioning of the input (i.e. text)
(1.2)
The stage of morphological and morphemic analysis
(1.3)
The stage of syntactic analysis
NL texts
(1.4)
interactive environment
Bloc of lexical analysis
(2)
(2.1)
Semantic network
expert
(2.3)
(2.2)
Domain ontology
Bloc of semantic analysis

Fig. 2. Diagram of automatic language processing

Semantic analysis in this model is supposed to be fully automatic, with no human participation, based on frequency analysis and information acquired from the block of linguistic analysis. In this connection the structure and the way of ontology formation is being completely reviewed, i.e. it doesn't include any extra-linguistic information but facts and knowledge in form of object-object, object-property relations extracted from other NL texts.

Automatic model of language processing involves a number of assumptions and cannot act as a tool of deep semantic analysis but it is quite viable and useful in a number of applied tasks as an element of DataMinig. For instance, it can be such tasks as text classification and clusterization according to facts which they include, checking of facts in the text to find out whether they coincide with ones in previously processed texts, revealing of implicit connection between domain objects. Besides, principles of domain ontologies formation can be applicable to unify and join different domain ontologies according to key concepts.

In the course of further research we plan the following:

- To carry out an analysis of latest achievements in the field of computational linguistics and linguistics analysis methods.
- To carry out a deeper analysis of existing systems of semi-automatic and automatic NLP.
- To carry out an analysis of application of semi-automatic and automatic approaches to analyze published scientific texts of NL.

References

1. Khlopotov M.V. Thesis. Methods and Algorithms of Finding Semantic Mistakes in Hypertextual Information Resources Basing on Ontological Models. – SUSU. 2004.
2. Alvin Toffler. The Third Wave. Translation from English: - Moscow: AST, 2002, 776 pages.
3. Seleznev K. Natural Language Processing, 2003 http://www.osp.ru/os/2003/12/183694/.
4. Iomdin L.L., Boguslavsky I.M., Lazursky A.V., Mityushin L.G., Sizov V.G., Kreidlin L.G., Berdichevsky A.S. Interactive Solving of Ambiguity in the Process of Automatic Language Processing: Problems and Prospects. Conference Dialogue-21. 2005.
5. V.D. Gusev, A.V. Zavertaylov, N.G. Zagoruyko, S.P. Kovalev, A.M. Naletov, N.V. Salomatina. System "Ontogrid" for Ontologies Formation. Conference Dialogue-21. 2005.
6. Voskresensky A.L., Khakhalin G.K. Means of Semantic Search. Conference Dialogue-21. 2005.