INTEGRAL TECHNOLOGY OF HOMONYMY DISAMBIGUATION IN THE TEXT MINING SYSTEM "LOTA"

Olga Nevzorova, Vladimir Nevzorov, Julia Zin'kina, Nicolay Pjatkin

Abstract: The article describes the integral technology of homonymy disambiguation, realized in "LOTA" textual document analysis system. The technology contains a totality of homonymy disambiguation methods as well as a scheme of their interaction..

Keywords: natural language processing, functional homonymy, homonymy disambiguation..

ACM Classification Keywords: H.3.1.Information storage and retrieval: linguistic processing

1. Introduction

"LOTA" specialized system of textual documents processing is a system of Text Mining class. The system is intended for analyzing "The Logic of Work" specialized texts in Russian language, which describe the logic of work of a complex technical system in various functioning modes [Nevzorova et al., 2001]. The main task of the analysis is the extrication from the texts given of an informational model of the algorithms which solve a definite task in a particular problem situation$ and the control over structural and informational integrity of the chosen algorithm scheme.

The algorithm informational model includes:

- description of the input information flow (types of informational signals or a semantic description of the information flow with an indication of the information source - particular algorithm, particular measuring device);
- description of the processes of transforming input data into output data (acceptable way of problem solving);
- description of the output information flow (types of information signals or a semantic description of the information flow with an indication of the information receiving point).

Solution of the main task is provided by a complex of test processing technologies, which include:

- technologies of morphosyntactic analysis;
- technologies of semantic-syntactic analysis;
- technologies of interaction with applied ontology.

The indicated sum of technologies is formed on the basis of the central kernel, the applied ontology (further on, aviaontology), which supplies coordinated interaction of various program modules. Aviaontology conceptually describes the data domain of informationally maintaining different flight modes of anthropocentric systems [Dobrov et al., 2004]. Avaiontology is a notion net of the problem domain. The current ontology size is more than 1600 notions (about 5000 textual occurrences of the notions). Aviaontology belongs to the class of linguistic (lexical) ontologies and is intended for integration into various linguistic appendixes.

The program complex consists of three interacting subsystems: "Analyzer" subsystem of technical texts linguistic analysis, "OntoEditor+" subsystem and "Integrator" subsystem. The subsystem interaction is realized on the basis of "client-server" technology. Besides, subsystems act in different modes in various subtasks (server mode or client mode).

"OntoEditor+" software toolkit [Nevzorova et al., 2004] is a specialized database management system. The system is intended for manual editing of the ontologies kept in relational database in TPS format, as well as for user query service and outer programs. New system functionalities are provided by a "Linguistic toolbox"

functional set, by means of which the integration of the applied ontology into linguistic appendixes is realized. The most typical tasks solved with the help of "OntoEditor+" system toolbox are: studying structural features of the applied ontology with the help of "OntoEditor+" system research toolbox; constructing the applied ontology linguistic frame; the task of covering the text with ontological entries; forming conclusions on the applied ontology etc.

"Analyzer" subsystem realizes the main of linguistic text processing (graphematic, morphosyntactic and partly syntactic analysis). In this article, the integral technology of homonymy disambiguation will be investigated, which is aimed at functional, morphological and lexical homonymy disambiguation first of all.

"Integrator" subsystem solves external query extracting data from the text. External query structure contains algorithm informational model components. External query is interpreted in the course of interaction with "OntoEditor+" subsystem as a structure bound to the applied ontology. The extraction of informational model components is realized on the basis of identifying the elements of input text segment tree (interaction with "Analyzer" subsystem) with query structure elements (interaction with "OntoEditor+" subsystem).

2. Integral technology of homonymy disambiguation

"Logics" tests are real technical texts with complex syntactic structure. The texts contain many abbreviations, among them the author's ones, sentences with enumerations, homonymy of various types etc. The stage of linguistic analysis is supported by a range of standard linguistic resources, especially grammatical dictionary, formed on the basis of A.A. Zaliznyak's grammatical dictionary and substantially enlarged by adverbs and specialized vocabulary; standard abbreviations dictionary; collocation dictionaries etc. Under non-standard resources supporting linguistic text analysis are subsumed aviaontology linguistic frame and the indexed base of problem domain frequent collocations. Practically, these resources the main homonymy disambiguation technologies in the system, namely homonymy disambiguation on the basis of frequent collocations indexed base, as well as homonymy disambiguation on the basis of aviaontology linguistic frame.

In the last few years the group of authors was developing a universal technology of functional homonymy disambiguation on the basis of contextual rules method [Nevzorova et al., 2006]. The technology given is based on linguistic scrutiny of the homonyms' syntactic behavior and homonyms' grammatical characteristics specification. This scrutiny revealed a range of actual problems of Russian functional homonyms' lexicographic description. New statistical basis for homonymy classification and subtypes selection was proposed. The work upon developing a new Russian functional homonyms dictionary on the basis of corpus research was started. The contextual method of homonymy disambiguation is the basic one in the integral technology of homonymy disambiguation in "LOTA" system. However, practical tasks of the system revealed some important aspects of linguistic analysis, which stimulated the development of new homonymy disambiguation methods. Initially, the work related to obtaining quantitative and qualitative assessments of specialized textual base of system documents. The analysis revealed quantitative assessments and the distribution of homonyms into types. Another important assessment was obtained in the course of typical homonym contexts. This research revealed the degree of technical texts homonymy (on average, 15-20 % of homonyms); the frequent homonyms were listed, as well as their typical contexts. These results became a basis for new applied technologies of homonymy disambiguation.

Thus, the integral homonymy disambiguation technology developed in "LOTA" system includes the following methods:

- contextual method of functional homonymy disambiguation;
- method of functional, grammatical and lexical homonymy disambiguation based on the indexed collocation base;
- method of functional, grammatical and lexical homonymy disambiguation based on the ontology linguistic frame.

2.1. Contextual method of functional homonymy disambiguation

Contextual method of functional homonymy disambiguation comes to developing for every functional homonymy type a group of rules defining the syntactic context of the homonym disambiguation and forming the group control structure which defines the rule application order. In work [Nevzorova et al., 2006] the main benefits and drawbacks of this method were described, concrete structures of generalized rules for some functional homonymy types disambiguation were given. Contextual disambiguation method is applied at the stage of morphosyntactic

text analysis, as rather frequently a syntactic method of building homogeneous groups is used in homonymy disambiguation. Disambiguating a homonym outside the group allows considering not only the local homonym context, but also the more remote one. Assuredly, it is one of the main benefits of the method. In the integral technology, the contextual method is a basic one and is applied last among the homonymy disambiguation methods.

2.2. Method of functional homonymy disambiguation based on the indexed collocation base

In order to realize this method, an integrated program technology of building up the homonym contexts base index was developed. The program technology developed includes the modules of creating and leading the homonym index, the module of coordinating the index base with the main linguistic resource (grammatical dictionary) and the mechanism of solving input queries on disambiguation (search) of typical homonymy contexts in the input text on the basis of homonyms index. The technology developed was realized on the basis of main "OntoEditor +" and "Analyzer" subsystems interaction.

"OntoEditor+" system, in order to provide effective integration into linguistic appendixes, supports a group of interconnect protocols of informational exchange with outer system program modules and outer dictionary databases, supplying client-server mode work. The functional, morphological and lexical homonymy disambiguation in the input texts is realized on the basis of a mechanism of recognizing the homonym contexts fixed in the indexed context base. Three main mechanisms of enlarging the indexed functional homonyms context base were developed:

- manual input and editing data on typical homonym contexts;
- import of typical homonym contexts from a textual file prepared in a special format of data representation;
- import of typical homonym contexts discovered by special search mechanism of the "Analyzer" subsystem.

This mechanism is organized as a query to the "Analyzer" subsystem, with "OntoEditor+" subsystem transferring to it a textual corpus where the search is being carried out. While processing the "Analyzer" subsystem transfers into "OntoEditor+" subsystem the information about the homonym contexts discovered. This is written either into the homonym index, or in the automatic mode, or in the mode of a dialogue with the operator. The special feature of the dialogue mode is the self-study mode, which is realized using the event diary mechanism. Depending on the settings, the diary records some important events in the system, for example, information change in the homonym index or the interaction with the "Analyzer" subsystem. In the self-study mode the sequence of earlier generated dialogues is saved and controlled, which provides the generation of unique dialogues only and homonymy disambiguation without repetition.

Based on the experimental text collection, as well as a range of linguistic resources, among them the most substantial being the National Corpus of the Russian Language (www.ruscorpora.ru), and Russian Associative Dictionary in 2 volumes (Karaulov Yu.N., Cherkasova N.V. etc. - M.: OOO "Izd.-vo Astrel'", 2002) a base of functional homonymy disambiguating collocation was built (about 30000 collocations currently). A program module was developed, which supplies the disambiguating collocation items generation according to their models in the course of forming a functional homonyms base index. The collocation model defines the functional or lexical homonym disambiguating context. Currently, about 2000 collocation models were formed on the basis of the abovementioned resources. The collocation model consists of two parts. In the first part, the collocation

component word forms are represented (as a rule, binary or ternary), the second part contains code parameters of the word form inner description according to the grammatical dictionary, along with the position and the distance of the disambiguating word form in relation to the homonym. Such model allows generating all disambiguating contexts, which are differed by the disambiguating word form. For example, the Russian collocation model 'relatively short' *'otnositel'no korotkij'* (Russian functional homonym *otnositel'no* disambiguated as an adverb) is expanded by the whole of the Russian adjective '*korotkij'* paradigm. Statistical analysis of the collocation model types allowed revealing the most frequent types where the following models belong:

- homonym (adverb/short form of adjective) + verb (disambiguating word form), for example, 'act effectively' (*'effectivno dejstvova') ;*
- homonym (noun/adjective) + noun, for example, 'close combat' (*'blizhnij boj')*.

The homonymy disambiguation method based on the collocation models is effectively used in disambiguating complex homonymy cases, for example for 'this/it' 'eto','all' 'vse' homonyms there were compiled more than 200 disambiguating collocations.

2.3. Method of homonymy disambiguation based on the ontology linguistic frame

"OntoEditor+" subsystem linguistic toolbox provides integrating the ontology into various appendixes related to text processing. The linguistic toolbox realizes the functions of text corpus download; automatical statistics leading on different corpus objects; functions of pre-syntactic text processing (sentence segmentation, abbreviation recognition, homonymy disambiguation based on the special protocols of interaction with outward lexicographical resources); forming the ontology linguistic frame; recognizing the applied ontology terms in the input text (cover task). The coupling of the ontological and linguistic (grammatical) resources is realized through the mechanisms of the ontology linguistic frame. The ontology linguistic frame is created with the help of the developed program toolbox, through which grammatical information about ontological concepts and their textual forms is fixed. Each ontological entry (как a composite term as a rule) is supplied by the corresponding grammatical information; in this process the corresponding homonymy (functional, morphological, lexical) is disambiguated. Grammatical information is transferred into the "OntoEditor+" subsystem from the "Analyzer" subsystem on the basis of special protocols of interaction. Functional, morphological, lexical homonymy disambiguation is completed on the basis of special dialogues with an expert linguist. Special procedures check the word forms in a term entry on the consistency of their grammatical characteristics. Lexical information reliability control is also carried out. Reliability control traces the changes both in the grammatical dictionary and the ontology. Allowing for the complexity and numerous stages of the abovementioned procedures, a master of linguistic frame construction was developed in the "Ontoeditor+" subsystem; the master is called out by a command from the main menu.

The mechanism of homonymy disambiguation on the basis of the ontology linguistic frame is related to solving the task of recognizing ontological entries in the text (text covering task). For each recognized ontological entry containing a homonym, the information about the grammatical characteristics of this homonym in the context of the ontological entry is transferred. The method allows disambiguating functional, morphological and lexical homonymy.

2.4. The interaction of homonymy disambiguation methods

The integral technology of homonymy disambiguation includes three abovementioned methods of homonymy disambiguation. The interaction of methods in solving the task of homonymy disambiguation is provided by the interaction of the main subsystems of the "LOTA" system. "OntoEditor+" subsystem provides the realization of homonymy disambiguation method based on collocations and the one based on the ontology linguistic frame. In the development of these methods, the engineering approach is used, which allows selecting the typical frequent language cases, which are actively used in technical language. Initially, in the course of homonymy disambiguation based on these methods, general and special system knowledge is used, which is stored in

various databases. "Analyzer" subsystem provides the realization of homonymy disambiguation method based on contextual rules, so the linguistic system knowledge is used. This method is a universal one, not depending on the specific problem domain. In the current version it supplies disambiguation accuracy of not less than 95 %. However, there exist some types of functional homonymy which are too complex for this method, for example, "conjunction/particle" type. The disambiguation of such homonymy is often possible only after the completion of a full syntactic analysis.

"OntoEditor+" and "Analyzer" subsystems interaction is realized on the basis of special interconnect protocols of interaction. In the course of the integral technology application, the homonymy disambiguation is carried out in two stages. On the first stage, the "Analyzer" subsystem (client) transfers a query on input text homonymy disambiguation to the "OntoEditor+" subsystem (server). The "OntoEditor+" subsystem returns the information about the disambiguated homonyms based on its own methods to the "Analyzer" subsystem. On the second stage, the "Analyzer" subsystem disambiguates the rest of the homonyms on the basis of contextual rules.

3. Conclusion

The integral technology of homonymy disambiguation is effectively used at the stage of pre-syntactic analysis in "LOTA" system. Essentially, the integral technology is a combination of engineering and linguistic approach to the solution of the given task. The integral technology projection is based upon the processes of coordinated interaction of different language levels, first of all the ontological level (providing the system model of knowledge about the world) with different language levels (morphological and syntactic). In the system, there was realized an effective mechanism of various subsystems interaction, which supply the realization of different methods in the integral technology.

4. Acknowledgements

The work has been completed with partial support of Russian Foundation Basic Research (grant № 05-07-90257).

5. Bibliography

[Nevzorova et al., 2001]. Nevzorova O.A., Fedunov B.E. Sistema analiza tehnicheskih tekstov "LoTA": osnovnye koncepcii i proektnye reshenija. // Izv. RAN. Teorija i sistemy upravlenija.– 2001. № 3. S. 138-149. In Russian.

[Dobrov et al., 2004] Dobrov B.V., Lukashevich N.V., Nevzorova O.A., Fedunov B.E. Metody i sredstva avtomatizirovannogo proektirovanija prikladnoj ontologii // Izvestija RAN. Teorija i sistemy upravlenija. M.: 2004. № 2. S. 58-68. In Russian.

[Nevzorova et al., 2006]. Nevzorova O..A., Zin'kina JU.V., Pjatkin N.V. Metod kontekstnogo razreshenija funkcional'noj omonimii: analiz primenimosti // Trudy mezhd. konf. Dialog'2006. M., Nauka, 2006. S. 399 – 402. In Russian.

[Nevzorova et al., 2004]. Nevzorova O.A., Nevzorov V.N. Sistema vizual'nogo proektirovanija ontologij "OntoEditor": funkcional'nye vozmozhnosti i primenenie //IX nacional'naja konferencija po iskusstvennomu intellektu s mezhdunarodnym uchastiem KII-2004. M.: Fizmatlit, 2004. T. 3. S.937-945. In Russian.

Authors' Information

Olga Nevzorova – Research Institute of Mathematics and Mechanics, Tatar State al University of Humanities and Pedagogiks, Kazan, Russia; e-mail: olga.Nevzorova@ksu.ru

Vladimir Nevzorov – Kazan State Technical University, Russia; e-mail: nevzorov@mi.ru

Julia Zin'kina – Kazan State University, Russia; e-mail: zjuliv@mail.ru

Nicolaj Pjatkin – Research Institute of Mathematics and Mechanics, Kazan, Russia; e-mail: nikolaip@mail.ru