TERMINOLOGICAL ANNOTATION OF THE DOCUMENT IN A RETRIEVAL CONTEXT ON THE BASIS OF TECHNOLOGIES OF SYSTEM "ONTOINTEGRATOR"

Olga Nevzorova, Vladimir Nevzorov

Abstract: In this article the method of terminological annotation of mathematical documents which is used in a context of text mining (in particular, for RDF network developing of a collection of mathematical documents) is considered. Terminological annotation of mathematical documents is carried out on the basis of universal design technology for applied problems solving developed in ontolinguistic system "OntoIntegrator" under control of system of ontological models.

Keywords: Natural language processing, ontological models, terminological annotation

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

Introduction

The idea of Semantic Web and the space of informational objects (a semantic net work of objects with an associate metadata) allows us deal with the information retrieval problem as the information retrieval problem in the space of informational objects. Such approach takes place in LinkedData project [Berners-Lee, 2011] where the space of informational objects represents as a PDF-network. Different semantic entities of the text documents such as structure elements of the document or terminological units or segments with different semantic might be represented as informational objects.

The information retrieval problem in the space of informational objects is essentially different from the searching methods and relevance assessment used nowadays. As distinct from keyword searching in the space of objects (in general keywords are represented by random symbolic sequences) implies searching in the space of names, relations and properties and could be realized on database meta-language. Therefore the important task is to generate the space of informational objects, to separate semantic objects of the text which are conceptual units of the semantic space out.

This paper is dedicated to the method of terminological annotation of mathematic documents, which is used in tasks of extraction semantic objects from the text (particularly in case of generating PDF-network of mathematic documents) and terminological indexation of collection of documents.

Terminological annotation of the document

It supposed that the quality of searching might be better if we use the terminological annotation of a source text and make up a terminological index.

The main problem of this approach is lack of accessible terminological sources in different areas of knowledge, on base of which we could realize the terminological annotation.

The alternative solution of the problem is an automatically extraction of terminology from the text. Nowadays there are 3 methods of terminology extraction: linguistic methods, statistic methods and combined methods. Linguistic methods are based on lexical-syntactic models of one-word and multi-words terminology and a system of filters when non-terminology is shifted out.

Statistic methods are based on idea of a terminological frequency. A term combination usually correspond with n-gramms (binomial, trinomial and tetratomic word combinations), which are characterized with the high level of steadiness. In case of evaluating the steadiness of word combinations in text we usually use *MI-score*, *t-score*, *Log-Likelihood*, *C-value*, χ^2 *criterion and other.*

Combined methods of terminology analysis suppose using both lexico-grammatical models, methods of terms word combination generation, the system of filters and statistic's instruments [Loukashevich et al., 2010].

The main problem of all methods is a filtration of generated terms word-combinations (a candidates for terms).

The other important task occurs in terminology searching is an evaluation of term relevance. In the information retrieval we use MI statistic measure (mutual information) and its modification (t-score measure).

The formula of *MI* measure is the following: $MI(ab) = \log \frac{N * freg(ab)}{freg(a) * freg(b)}$

where *freg ()* – the frequency of words and word-combinations, *N* – the number of words in collection. *MI* measure shows the difference between using the word in word- combination and the word's separate using.

The quality of extracted term word combinations might be realized on base of *AvP* measure (an average precision) [Ageev et al., 2004]. The precision *PrecTerm* of term word-combination extraction from a list of *n* word-combinations is defined as $\Pr ecTerm = T/n$, where *T* is a number of terms in a list.

In the arranged list we can get the precision on the level of n-terms $\Pr ecTerm(n)$, which is determine as the quantity of relevant terms among the first set of n in the distributed list divided by *n*.

In this case the average precision is counted according to the formula: $AvP = 1/k \sum_i \Pr ecTerm(i)$,

where *k* is the number of terms in the list of *n*-elements.

For example, if we have three elements in a list (N=3) and two terms (k=2) occur in the first and the third position of the list , we will get *AvP*=(1/2)(1+2/3)=5/6.

It should be made clear that when we figure out the *MI* measure the word order and their correlation (the relevance of syntactic structure of terms) doesn't take into account.

The syntactic structure of multiword term determines types of syntactic relations between components of the word-combination.

Let introduce the notion of precision of term extraction with the specified syntactic structure *R* - ($\Pr ecTerm\_R$),which defines as a quantity of relevant terms with the specified syntactic structure *R* taken from the list of an *n* word-combinations, in other words $\Pr ecTerm\_R = T - R/n$, where *T_R* is a number of terms with the specified *R* syntactic structure in the list.

It supposed that the relevant documents should contain word-combinations with the same syntactic structure as the structure of the word combinations in request. For example, if there is a *rank of an Abelian group* word-combination in the request, the documents which contain a *torsion-free Abelian groups of rank one* word-combination should be recognized as irrelevant, because the main word of the *rank of an Abelian group* word-combination is *rank* since in the *torsion-free Abelian groups of rank one* the main word *rank* has a dependent position (attributive relation).

Thus a new approach of the terminology searching improvement grounds on the idea, that the searching based on terminological annotation should rely on semantic-syntactic equivalence of the document models with the searching request, what is different to a traditional keyword searching.

Let us overview the syntactical relations of Russian language might be used for the term word combinations generation. The general syntactical model for term word combination is a nominal group that includes the main word (a noun) and the modifier (dependent word). The structure of nominal group determines by syntactical relations between the main and dependent words. The syntactical relations found on correlation between lexical meaning of the words and their grammatical forms.

There are five main types of syntactical relations of word-combinations in Russian language: an attributive, an objective, a subjective, an adverbial and a completive.

Attributive relations. We can talk about attribute relations when a noun (with the general lexical meaning of subject) correlates with the word of attributive meaning that might be coordinated and uncoordinated with the noun. A formal model of the attribute relations could be represented as:

1) $Atr \cap_N N$ – is a model with a coordinated attribute *Atr* (it's coordinated by the whole set of grammatical categories), for example, in Russian *конечная группа (a finite group).*

2) $N + Atr$ – is a model with an uncoordinated attribute (a parataxis), for example, in Russian *группа без кручения (a* torsion-free group), *in Rissian* группа с обычной арифметической операцией *умножения* (a group with standard arithmetic multiplication operator).

Objective relations. We can talk about objective relations when a verb (a participle or an adverb) correlates with a noun or more rare infinitive. These word combinations are semantically bounded, because the main word has lexical meaning of action, sense, perception, since the dependent word means an object of this action, sense, perception (it's a direct object mainly). The noun of action takes an acting structure after the verb (in Russian решить уравнение – решение уравнения).

A formal model of the objective relations could be represented as:

$N_V + N_{p2}$, where N_V is a verbal noun, N_{p2} is a genitive case noun. Particular model of objective relations rely on a model of government of the verb.

Subjective relations characterize the word-combinations with a verb or participle in a passive voice. The dependent word in this case shows an actor (an instrumental case). For example, in Russian предложенный автором *(*метод*).* A formal model of the subjective relations could be represented as: $V^2 / A_V^2 + N_{p5}$, where V^2 / A_V^2 *is a verb or participle in a passive voice and,* N_{p5} is a noun in instrumental case.

Adverbial relations characterize a verb word-combination and rely on the lexical meaning of the process. The adverbial relations are specificated as the adverbial relations of attribute, time, place, cause and purpose.

The example of an adverbial relation of place is in Russian группа параллельных переносов в линейном пространстве (a group of parallel transfers in the linear space), compare with строка в таблице (in Russian) (a row in the table). The latter is an attributive model.

A formal model of adverbial relations could be represented as: $N_V + PPNP$, where N_V is a verbal noun, *PPNP* is prepositional phrase (more often in an adverbial meaning).

Completive relations appear in idiomatic word-combinations. A formal model of the completive relations might be represented as a list of the corresponding word combinations.

In general, a multi-word term might be generated as a superposition of the aforenamed relations. As an example, in Russian *абелева группа без кручения первого ранга* (*torsion-free Abelian groups of rank one*) word-combination is a superposition of the word-combinations which are *((абелева группа) без кручения) первого ранга)*, (группа (параллельных [переносов в линейном пространстве])), where round brackets mark out the attributive relations and square brackets indicate the adverbial relations.

A semantic structure of multi-word terms word-combinations might be represented in a structure of the terminological annotation, in which we can distinguish a type of relation, a main word and a dependent word. The correlation of the elements are similar to the subordination relations of compound sentences.

The terminological annotation might be organized as an XML notation with the following characteristics:

- the type of relationship is represented by the value of a LINK attribute;

- the Holder attribute indicates the main element of a word-combination;

-the Dependent attribute defines the dependent element of a word-combination according to a syntactical relation.

The announcement of attributes in XML format are designated as:

- <!ATTLIST TERM Link (attributive | objective | subjective | adverbial | completive)
- <!ATTLIST TERM Holder CDATA >
- <!ATTLIST TERM Dependent CDATA >

Thus, the terminological annotation of the 'Abelian group' word-combination in XML format could be represented as: <TERM Link="attributive" Holder="группа" Dependent="абелева">

абелева группа

</TERM>

Design technology for applied problems solving in "OntoIntegrator" system

The "OntoIntegrator" system is an ontolinguistic research software development kit for the solution of applied problems, connected with an automatic text processing. The "OntoIntegrator" system contains the following functional subsystems, which are [Nevzorova, 2007]:

- an "Integrator" subsystem;
- an "OntoEditor+" subsystem of ontology modeling;

- a "Text analyzer" subsystem;
- a subsystem of dealing with external linguistic recourses;
- a subsystem of ontological models.

The "OntoEditor+" [Nevzorova, 2006] subsystem of ontology modeling provides the main table functions for dealing with an ontology (addition, modification, deletion, automatic correction; keeping of more than one or compound ontologies, in other words with the general lists of relations, classes, text equivalents and others; an import of the ontologies with the different formats of data; a filtration of ontology; keeping of statistics automatically, searching for chains of relations and others). The functions of the visualization unit support different graphic modes of system, including the graphic mode of the ontology modeling.

The “Text analyzer” subsystem includes linguistic tools are useful for solution the problem of the morphology analysis, of the ontological markup, of the polysemy resolution, of segmentation and the applied linguistics modeling. The subsystem of dealing with the external linguistic recourses supports keeping the basic linguistic recourses that contain a grammatical dictionary and a set of specialized linguistic data bases. The Integrator subsystem provides integrated base of the applied linguistic problem solution and control under the applied problem solution generation.

Let us overview the process of generation the solution of applied problems in the "OntoIntegrator" system by the example of the terminological annotation task. The process of the solution is realized under control of the ontological models system with a reflective kernel (the system is represented by the relational data bases) . The system of the ontological models includes different types of ontologies; there are applied ontologies (domain ontology and the relations for inference in it), the ontology of models and the ontology of problems [Nevzorova et. al., 2011].

In terms of structure the system of ontological models represents a ternary associative system. The components of the system are the semantic networks (the ontological subsystems); there are the ontology of problems, the ontology of models and the applied ontology. The system allows us the interpreting of the applied ontology as a set of ontologies of different domains, which can be external, attached by a user, and internal, integrated into "Ontointegrator" system (with the possibility to spread-out, to edit and the support of calculation) with the purpose of the applied problems solution.

In order to build a solution for the linguistic problem it is necessary to decompose it to structural elements represented in problem ontology. Then, structural elements of solution for the linguistic problem are mapped to the set of model ontology structures. For easier interpretation all model-concepts are split to conditional groups supported by complex visualization mechanisms:

- Basic models providing the minimal functionality of system of ontological models;
- Syntactical models reflecting T-models (text models);
- Structure-semantic models creating adequate to applied problem solution structure and interpretation of its results;
- User models which are dynamically created by user.

Meaningfully, models are used for implementation of solution to problem-concept and allow to set (to interpret) solution components as a problem of setting the property, or as a problem of identification of the relation or the

problem with known evaluation algorithm. The solution for applied task is composed on the basis of text fragments (T-models) which are derived from the models from model ontology (S-models) by model identification procedure. Syntactic models, which are the basis for the extraction of terminological word-combinations, are based on the rules of syntactic analysis of NP. Basic syntactic models of NP structure are distinct by type of syntactic relations between the noun and modifiers. Also, arbitrary number of modifiers and combinations of any types of syntactic relations (attributive, objective, subjective and adverbial) are allowed.

Another developed method allows the extraction of NP from constructions with conjunctive reduction. The inverse problem of extraction of construction's potential components for their recognition as independent terminological entities is solved through terminological analysis of conjunctive syntactic constructions of certain types. For example, the components "*natural numbers addition operator*" and "*natural numbers multiplication operator*", which are recognized as independent terminological entities, are extracted from the syntactic construction "*natural numbers addition and multiplication operators*". The extraction of components from conjunctive constructions is made on the basis of special rules, which take the phenomenon of "semantic homogeneity" into consideration. Semantic homogeneity assumes the composition of syntactic construction with semantically homogeneous parts, in other words, all members of homogeneous constructions must belong to the same semantic class. During the phase of construction of rules, two main semantic classes - the concrete and abstract entities - are separated. The semantic homogeneity principle allows the composition of conjunctive construction for either concrete or abstract entities. For example, the constructions of type "*Eigen vectors and values of matrix*" (concrete homogeneity) or "*number addition and multiplication*" (abstract homogeneity) are acceptable. Likewise, the semantic homogeneity of classes of attributes is required for composition of conjunctive attributive set (conjunction by attributes). For example, the conjunction by attribute "number sign" ("*positive numbers and negative numbers"*) is presented in the "*positive and negative numbers*" construction; but, attributes from different semantic classes are defined in the "*nondegenerate and symmetric matrices*" construction and, therefore, this construction is treated as attributive. The rules of extraction of components from conjunctive constructions are well-defined in [Nevzorova et. al., 2009].

Recognition of terminology in text on the basis of main syntactic NP and conjunctive reduction models is performed in "OntoIntegrator" system on the basis of applied ontological resource - the ontology of mathematical knowledge (theory of groups section). The basic list of mathematical terms (prop terms) from this section is used for experiments. Syntactic link type of the prop term and its syntactic role (main or dependent word) is determined for prop term containing terminological word-combination selected in text. The corresponding data is represented via XML notation.

In general case, the technology of problem preparation and solving in "OntoIntegrator" system consists of following phases:

1. Preparation of linguistic resources. Basically, this phase involves the supplement of system dictionaries with new terminology of given problem field. The selection mechanisms of candidates to terms are provided in the system.

2. System configuration. Besides the ability of creating new ontology system or connection one developed earlier, the possibility of combining the ontology of models and ontology of problems from different systems into applied ontology system is provided in the "OntoIntegrator" system. Adding new concepts to ontology of models and the redefinition of them in the reflexive kernel as corresponding relations in ontology of problems and applied

ontology is sufficient in order to accomplish this. Also, dynamic change of applied concepts (instances) classification on the basis of model-concepts (classes) is possible through connection of different classification, which was developed earlier.

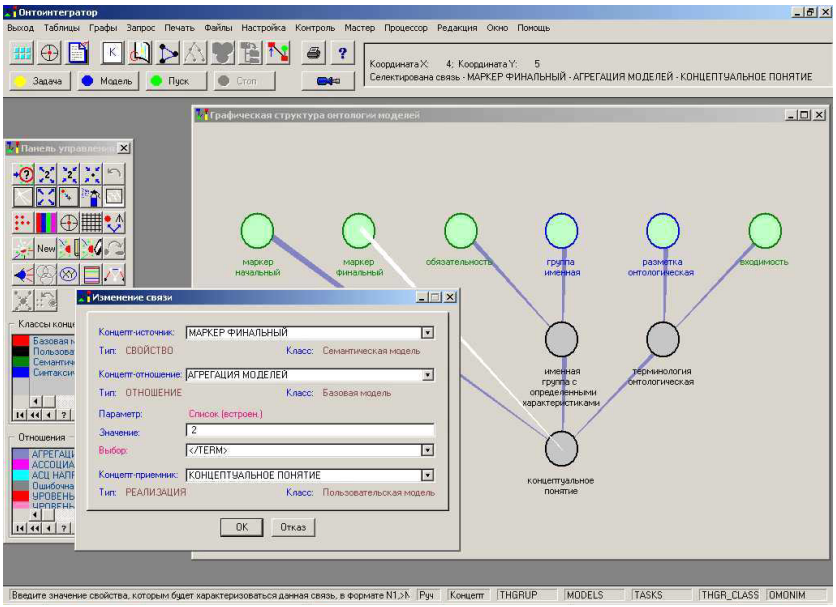

Fig.1. Construction of model-concepts of "Implementation" type

3. Construction of problem-concepts of "Implementation" type. In the context of this article, this means the construction of solution process of abstract problem "Document tagging on the basis of problem model" structure in the ontology of problems basis using inclusion relationship. Problem-concept of type "Implementation" is meaningfully a link (node) to semantic subnet, which describes the sequence (structure) of solution process.

4. Construction of model-concepts of "Implementation" type. Several such models could be used simultaneously during the problem solving, but they have to be combined in the semantic subnet with the node (link) as model-concept of type "Implementation" that would be treated as problem model, using model aggregation relationship. The problem model "Abstract concept", which is used for terminological tagging of document, is presented on fig.1. This concept has the "Final marker" property which is updated through parameter of link with </TERM> value, chosen from integrated ontology of text markers. Integration of chosen models is performed on the base of aggregation relationship.

5. Supplement of integrated ontologies. In problem under discussion, markers <TERM Link="link relationship" Holder="main element" Dependent="dependent element"> and </TERM> should be added to integrated ontology of text markers.

6. Preparation of applied ontology. During problem solving, it is sufficient to supplement applied ontology with basic list of terms (prop terms) without identifying the relationships between them and to configure its linguistic shell, which would provide the ontological tagging of document being processed.

7. Classification of applied concepts. For problem under discussion this phase is performed by default, because the model "Conceptual idea" does not contain model-concepts of type "Implementation" that identified in text through their instances (concepts of applied ontology).

8. Choice of problem-concept and model-concept. Text selection for processing and/or building-up the natural language query is viewed as part of the solution to problem of tagging.

9. Startup of solution process (problem-concept).

The result of terminological tagging of the document is presented in OMDOC format in the Internet-browser window on fig.2. For example, the tagging of terminological group *finite p-groups* includes indication of syntactic relationship type (attributive word-combination), selection of head (*p-group*) and the dependent word (*finite*). Prefix ("$$$") denotes the form (*p-*).

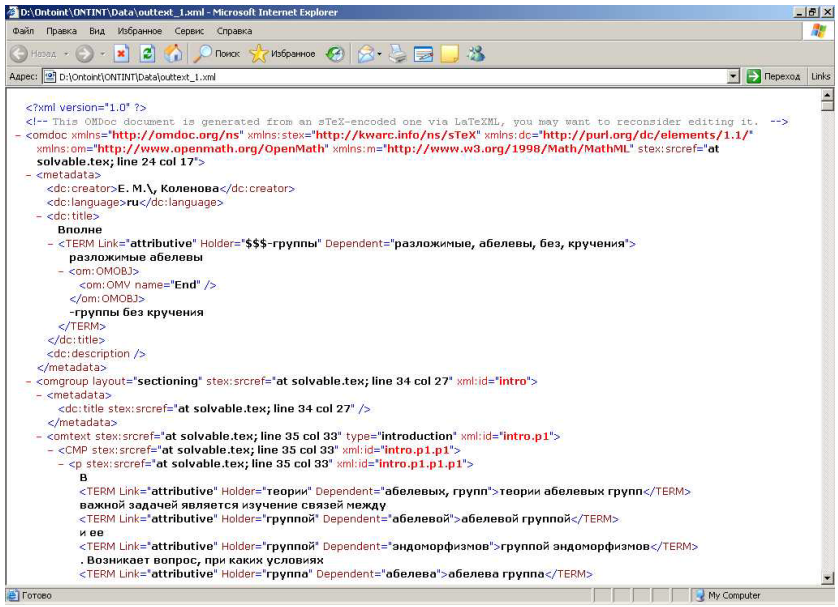

Fig.2. The result of terminological tagging in OMDOC format

Experiments

Experiments for terminological annotation of mathematical articles are done on the experimental collection of group theory articles and on the original terminological list of group theory concepts from the corresponding section of DBPedia.

Content processing of elements of mathematical documents includes text segmentation to sentences, text objects (extraction of formulas, number sequences, words, punctuation marks, abbreviations etc.) recognition, recognition of NP containing terms from applied ontology, recognition of complex syntactic constructions (conjunctive reduction groups) and other procedures (for example, homonyms extraction and classification).

In articles being considered, single- and multiword NP are selected on the basis of corresponding syntactic models (*finite group, Sylow 2nd subgroup of group, Klein group, dihedral group, nilpotent non-Abelian group* etc.). It is necessary to note that mathematical texts contain a large amount of words with prefix-formulas (*p-subgroup, 2-subgroup*) and postfix-formulas (*group G, subgroup K*). Arbitrary formulas and expressions could be used as prefixes. These objects are not contained in system dictionary and are processes with special methods, which are separating left prefix-formulas and are working on the basis of right word-part syntactic model. Postfix-formula containing words are processed on the basis of NP with abbreviation syntactic model (*group G, subgroup K*).

Conclusion

New ideas of consideration of syntactic structure of terminological word-combination in search problem context and the "OntoIntegrator" system solution technology of problem of terminological annotating are considered in this article. Main technological phases of problem solution preparation are described on the basis of concrete applied problem.

Proposed technology allows the unification of process of solution composition of wide range of applied problems that are oriented on ontology usage and methods of automated text processing. All phases of applied problem solution composition are supported by convenient graphical interfaces and specialized graphical editors. Detailed description of conceptual and technological solutions made in "OntoIntegrator" system could be found in bibliographic links.At the present time, the developed technology of applied problems solution are probed in concrete applications linked to terminological and structural annotation of mathematical documents, basic linguistic problems, such as ontological tagging of text, context rule-based homonymy disambiguation etc.

Acknowledgements

The work has been completed with partial support of Russian Foundation of Basic Research (grant № 11-07-00507).

Bibliography

[Berners-Lee, 2011] T. Berners-Lee. Linked Data - Design Issues. At http: //www.w3.org/DesignIssues/LinkedData.html/. (Accessed on January 18, 2011).

[Ageev et al., 2004] Ageev M., Kuraleonok I. Official metrics of ROMIP'2004. In Proceedings of the second Russian seminar on the evaluation of methods of information retrieval, Saint-Petersburg, pp. 142-150 (2004). In Russian.

[Loukashevich et al., 2010] Loukashevich N.V., Logachev A.M. Attribute combination for automated term extraction. Computational methods and programming (11), 108-116 (2010). Available at: http://num-meth.srcc.msu.su/. In Russian.

[Nevzorova, 2006] Nevzorova O. Instrumental system "OntoEditor+" for visual design of ontologies in linguistic applications. KSTU named after A.N.Tupolev newsletter, (3), 56-60, (2006). In Russian.

[Nevzorova et. al., 2011] Nevzorova O., Nevzorov V. Multi-layer ontological system for applied problems solution planning. In Proceedings of international conference "Open Semantic Techologies for Intelligent Systems" (OSTIS'2011), pp. 323-330, (2011). In Russian.

[Nevzorova, 2007] Nevzorova O. Ontolingvistic systems: Technologies of applied ontology interaction. Kazan State University scientific notes, Physic-mathematical sciences, 149 (2), pp.105-115 (2007). In Russian.

[Nevzorova et al., 2009] Nevzorova O., Nevzorov V. Ontological analysis of the domain: Automated methods of term extraction in "OntoIntegrator" system. In "Modelling methods" symposium, Kazan, pp.196-208 (2009). In Russian.

Authors' Information

Olga Nevzorova *– Chebotarev Research Institute of Mathematics and Mechanics, Institute of Applied Semiotics of Tatarstan Academy of Sciences, Kazan, Russia; e-mail: olga.nevzorova@ksu.ru*

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