Development of Linguistic Ontology on Natural Sciences and Technology

Dobrov B., Loukachevitch N.

Research Computing Center of Moscow State University
Russia, Moscow, 119899, Vorobievy Gory, MGU NIVC
dobroff@mail.cir.ru, louk@mail.cir.ru

Abstract

The paper describes the main principles of development and current state of Linguistic Ontology on Natural Sciences and Technology intended for information-retrieval tasks. In the development of the ontology we combined three different methodologies: development of information-retrieval thesauri, development of wordnets, formal ontology research. Combination of these methodologies allows us to develop large ontologies for broad domains.

1. Introduction

Semantic Web activities are supposed to facilitate information sharing, which has to be based on large-scale ontologies in various domains including broad domains of science and technology.

Ontologies differ largely in the degree of formalization such as taxonomies, thesauri, axiomatized theories (fundamental ontologies) (Guarino, 1998).

It seems to be impossible to develop fundamental ontologies, axiomatized theories for such sciences as physics, chemistry or geology. At the same time formalization of conventional information-retrieval thesauri as tools of thematic search is not enough to use them in automatic text processing for information-retrieval tasks (Soergel et.al. 2004; Tudhope et.al 2001; Voorhees, 1999).

Therefore to develop an ontology for a natural science, on the one hand, it is important to describe concepts more formal than in information retrieval thesauri, on the other hand, to develop a large ontology (thousands of concepts) during a relatively short time (several years).

It means that it is necessary to understand what kind of the ontology structure is the most appropriate for information retrieval tasks, that is to develop special ontologies for information-retrieval applications.

Specific features of the information-retrieval tasks and the level of contemporary natural language processing systems require that relations in information-retrieval resources do not have to depend on the text context (ANSI/NISO, 2003; Will, 2004). Such context-independent relations include taxonomic relations, several subtypes of part-whole relations and relations of ontological dependence (Gangemi et.el 2001; Loukachevitch & Dobrov 2004).

The concepts of an ontology intended for information-retrieval purposes have to have direct relationships to language expressions – technical terms (single words and multiword expressions), therefore such an ontology has mainly to be based on senses of such expressions, to be a linguistic ontology (Gomez-Perez et.al., 2000).

In specific domains a linguistic ontology can be a formal ontology in the same time because senses of domain terms are very tightly linked to domain concepts (Sager, 1990).

On this basis we have developed Russian-English Socio-political thesaurus (32 thousand concepts, 79 thousand Russian terms, 85 thousand English terms) as a resource for automatic text processing for such information-retrieval tasks as automatic conceptual indexing, automatic text categorization and others (Loukachevitch & Dobrov, 2002) in a broad domain of the contemporary society life.

In the paper we describe the structure and steps of development of Linguistic Ontology in Natural Sciences and Technology (below SCI-Ontology), an integrated ontology for such sciences as mathematics, physics, chemistry, geography, geology and technological processes.

2. From Information-Retrieval Thesauri to Formal Ontologies

2.1. Information-Retrieval Thesauri

Since 1960s information-retrieval thesauri are considered as resources for thematic access to electronic collections. The thesauri comprise main terms of the domain, synonymous terms are linked to authorized terms (descriptors). Between descriptors relations such as Broader-Term – Narrower Term and Related Term are established (AGROVOC, 1999; LIV, 1994).

But conventional information-retrieval thesauri were developed for manual indexing by human indexer – now volumes of electronic collections surpass possibilities of manual indexing. In automatic regimes use of traditional information-retrieval thesauri for large full text collections usually leads to decrease of performance characteristics of information retrieval in comparison to “bag of words” methods (Voorhees, 1999).

The reason is that in fact an information-retrieval thesaurus describes an artificial language based on terms of natural language and intended for description of main topics of documents (LIV, 1984).

A human indexer had to translate a natural language of a document to an artificial language of a thesaurus. The whole procedure was based on domain, common sense, and grammatical knowledge of indexers. Many decisions in developing information retrieval thesauri were intended to make work of indexers more convenient and less subjective. For example, specific terms were not included, ambiguous terms were provided with scope notes and comments convenient for human subjects, relations were used mainly for convenient navigation.

To be effective in automatic text processing a thesaurus needs to include a lot of information that is usually missed in thesauri for manual indexing such as:

- description of much more descriptors of lower levels than in usual information-retrieval thesauri. Most information-retrieval thesauri for broad domains include 5-10 thousand descriptors. To be used in natural language processing it has to include tens of thousands of descriptors.
- Detailed description of synonyms of descriptors. A computer processing requires much more synonyms than it is necessary for a human, many of synonyms can be evident for a human indexer.
- Description of ambiguous terms.
- It is necessary to change the traditional system of thesaurus relations to more formal one.

2.2. Inclusion of Predicates and Axioms to Information-Retrieval Thesauri

Considering problems of formalization traditional information-retrieval thesauri to adapt them to contemporary level of electronic collections some authors propose to change the traditional system of thesaurus relations to a formalized set of predicates and describe axioms for such a set. For example, in (Soergel et.al. 2004) as an example of modification of AGROVOC thesaurus the following relations between thesaurus items are proposed:

milk
<*includesSpecific*> cow milk
<*containsSubstance*> milk fat
cow
<*hasComponent*> cow milk
Cheddar cheese
<*madeFrom*> cow milk

Examples of proposed axioms are as follows:

Rule1:
Part_X <*mayContainSubstance*> Substance_Y
IF Animal_W <*hasComponent*> Part_X
AND Animal_W <*ingests*> Substance_Y

Rule 2:
Food_Z <*containsSubstance*> Substance_Y**:**
IF Food_Z <*madeFrom*> Part_X
AND Part_X <*containsSubstance*> Substance_Y

Authors suppose that an automatic system having such rules can automatically receive that if Cheddar cheese contains (containsSubstance) milk fat and “if cows on a given farm are fed mercury-contaminated feed, that *Cheddar cheese* made from milk from these cows *<mayContainSubstance>mercury”*.

But to receive such an inference besides changes in thesaurus articles, it is necessary to have automatic tools for natural language processing in agricultural field. These tools have to allow full and exact extraction of facts from unrestricted coherent texts, to follow coreference, sequence of time periods and so on: there is mercury in some feed, these feeds belong to a farm, cows of this farm ate these feeds, cheese from milk of these cows is produced immediately after the contamination and so on.

But at present there is no such automatic systems processing large coherent texts with required precision and recall to provide reliable inference.

Therefore, in our opinion, huge labor costs on the described transformation of a traditional information retrieval thesaurus will not lead to real improvement of text processing for information-retrieval tasks, will not provide construction of resources better suitable for automatic regimes of text processing.

2.3. Relations in Ontologies Intended for Information Retrieval

Owing to current state of natural language processing linguistic resources intended for information-retrieval purposes will be utilized in indefinite contexts when for any concept mentioned in a text it is impossible to extract a full and precise set of facts about this concept or its examples discussed in the same text. So in such circumstances the only reliable conceptual relations are relations that do not depend (or weakly depend) on the text context, that is such relations that do not disappear or change or loss their significance in various text contexts.

The most known such a relation is a taxonomic relation. A birch in any situation will be a tree. This relation has such reliable properties as inheritance and transitivity. It is important to find and use other relations which are reliable in different text contexts and have the transitivity property.

In our opinion such reliable relations are relations of ontological dependence discussed in formal ontology research (Smith, 1998; Gangemi et.al., 2001).

The main question of the dependence theory is if an entity can exist by itself or it supposes the existence of something else. There are three main types of this relation:

- whether the existence of an entity supposes the existence of something else (rigid dependence), for instance, *boiling* is impossible without the existence of a certain volume of liquid which boils;
- whether existence of examples of a certain class (generic dependence) is supposed, like, the appearance of the concept *garage* is impossible without the existing concept *motor vehicle*, though a certain garage may appear without any reference to a certain motor vehicle;
- when the existence of an entity in moment *t* presumes the existence of another entity in moment *t1* before *t* (historical dependence), so, for instance, *motor vehicle* historically depends on *car plant*, as motor cars are produced in plants.

It is easy to see that in case of the rigid dependence the existence of a dependent concept is very tightly connected with the existence of a main concept. It is difficult to imagine a situation (and a text) where a dependent concept participates and this situation has no relation to a main concept.

In case of the generic dependence examples of a dependent concept usually participate in situations related to a main concept, however sometimes situations, not relevant to a main concept, can arise (for example, a crime in a garage can have no relation to automobiles).

At last the historical type of dependence is the weakest type among existential situations. A main concept is necessary for appearance of a dependent concept, but then a dependent concept can exist for a long time and participate in various situations not relevant to the main concept.

There differences in subtypes of conceptual dependence relations lead to differences in behaviour of these relations in information-retrieval context.

In (Gangemi et.al. 2001) authors postulate transitivity of relations of ontological dependence. In (Loukachevitch & Dobrov 2004 a,b) it was shown that relations of strict and generic ontological dependence are effective for development of linguistic resources for information-retrieval purposes.

3. Main Principles for Development of SCI-Ontology

In construction of the SCI-Ontology we combine three different methodologies:

- the methods of construction of **information-retrieval thesauri;**
- the development of **wordnets** for various languages;
- ontology and **formal ontology** research.

From the methodology of information-retrieval thesauri development the following principles are important:

- information-retrieval context,
- work with domain-specific terminology (including multiword expressions),
- descriptor-synonym organization,
- a small set of relation types.

From the methodology of wordnets development the following principles are significant (Miller et.al., 1990):

- Concepts are mainly based on senses of real language expressions,
- Description of ambiguous terms,
- Many-level organization of terminological system.

From formal ontologies development the following principles are used:

- concept-based organization. This does not contradict to the principle of linguistically-motivated units, because senses of domain terms are tightly connected to domain concepts;
- use of taxonomic relations and relations of ontological dependence,
- many-step inference.

4. Thesaurus on Socio-political Life as a Model for Development of SCI-Ontology

We have already used principles discussed in previous section for development of Socio-political thesaurus as a tool for automatic text processing for information-retrieval tasks.

The Socio-Political Thesaurus is a hierarchical net of concepts. We consider it as a kind of a linguistic ontology. The concepts of the Thesaurus originate from senses of language expressions, that is single words or multiword expressions.

The main unit of the Socio-Political Thesaurus is a concept. When a new concept is introduced into the Thesaurus, it is necessary to assign its name. The name of a concept has to be clear and unambiguous for native speakers. In the Russian-English thesaurus a concept has to have a name in Russian and a name in English. These names are used in different representations of text processing results.

A concept has a set of linguistic expressions that can be used for reference to the concept in texts. A set of linguistic expressions of a concept is called ‘text entries of a concept’ and can be considered as a synonymic row. In the Russian-English thesaurus a concept has a set of Russian text entries and set of English text entries. These text entries are used to recognize a concept in texts.

Concepts often have more than 10 text entries including single nouns, verbs, adjectives and noun or verb groups.

A concept within the Thesaurus has relations with other concepts. The main types of relations are taxonomic relations and a specific set of conceptual relations based on ontological dependence relations.

So the types of conceptual relations in the Thesaurus are:

- taxonomic relations,
- generalized part-whole relations describing internal characteristics of entities (physical parts, properties, participants for situations). In establishing of part-whole relations we use an important rule: concept-parts have to be ontologically dependent from concept-wholes. Therefore in the Thesaurus a tree is not a part of a forest (in fact, only the concept FOREST TREE can be described as a part of the concept FOREST). This rule provides transitivity of part-whole relations of the Thesaurus,
- external relations of ontological dependence. So in the Thesaurus the concept FOREST is described as a dependent concept from the concept TREE, because forests can not exist without trees, but trees can grow in many others places, not only in forests,
- related term (RT) relation is used for description of relations between very similar concepts not merged to the same concept.

Taxonomic relations and part-whole relations (with the above mentioned restrictions) are considered as transitive. Taxonomic relations, part-whole relations and external relations are hierarchical relations. Therefore a concept of the Thesaurus can have a set of hierarchically lower concepts – a tree of the concept.

Socio-political thesaurus is provided with a program suite of the full technological cycle including term extraction for new collections, morphological analysis, thesaurus matching, term disambiguation, construction of thematic representation of a text, an interface for thesaurus testing in applications.

Now the Socio-political thesaurus is used in such information-retrieval tasks as automatic conceptual indexing, automatic text categorization, automatic text summarization, visualization of retrieval results (Loukachevitch & Dobrov 2002).

The thesaurus contains a lot of concepts related to science classifications, scientific organizations, research process and so on. So it was important to utilize the described knowledge in development of a new ontology.

5. Main Stages of Initial Ontology Development

The initial stages of development of SCI-Ontology were as follows.

At the first stage for every considered science text collections were gathered. Sizes of collections were from 50 till 90 Mb. Sources of the collections included specialized Internet sites, school lessons and university lectures materials.

At the next stage term-extraction procedures were applied to the text collections. Term extraction procedures were as follows:

- extraction of one to three words noun compounds including adjectives and/or nouns in genitive (adjective+noun, noun+noun_in_genitive and so on). These expressions are the most frequent structures of technical terminology in Russian.
- extraction of repeated sequences of nouns, adjectives and propositions, which in a given text are mentioned more often together than separately.
- extraction of terms found in Socio-political thesaurus.

So we received lists of candidate terms for each science. The lists were ordered by frequency. The 25 thousand most frequent terms from every list were given to knowledge experts for fast check. The knowledge experts removed evident mistakes, general expressions, complex expressions consisting of several terms and marked domains of terms. Several marks for a term were allowed. The integrated list of 32 thousand terms with domain marks was received.

This terminological list became a basis for extraction of relevant concepts, terms and relations from Socio-political Thesaurus, which contains general scientific and technological concepts. If a term from the list matched the thesaurus term, the corresponding concept together with upper-level concepts were loaded to the SCI-Ontology.

At last terms from the terminological list were loaded to the SCI-Ontology as names of candidate concepts with only one relation to its science (sciences) concept. So we received the preliminary version of our ontology.

After that our knowledge experts began to work with the SCI-Ontology project.

6. Methodology of Knowledge Experts Work

Knowledge experts can introduce a new concept or approve a candidate concept. They try to find a definition (definitions) of a corresponding term or analyse document contexts clarifying a sense (senses) of a term. Found definitions are compared to real term usages because there are a lot of outdated definitions or definitions narrowed in a specific subdomain.

The experts try to reveal and analyse ambiguity of terms and describe this ambiguity introducing several concepts.

A new concept is supplied with a list of terms (text entries) whose senses correspond to the concept. This list of text entries has to be as full as possible to provide stable recognition of the concept in texts.

Term definitions and text contexts became a basis for description of concept relations.

The experts work with various sources for terminological knowledge.

The initial source of terminological knowledge from real texts are candidate concepts:

- Candidate concepts allowed us to extract relevant knowledge from Socio-political thesaurus.
- Candidate concepts can be real concepts of the domain. An expert can approve a candidate concept (=to delete candidacy mark) and add text entries and relations to the concept.
- A candidate concept can represent a variant of another concept. An expert can join two candidate synonymous concepts.
- Candidate terms can provide useful examples of terminological usage in real texts, which can confirm or serve as counterexamples to found definitions.
- Every item from candidate concepts list has to be approved, or joined, or deleted as not necessary. So candidate concepts urge to process all found terminological phenomena.

The experts study entries of glossaries and items of academic programs trying to encompass maximal volumes of terminology.

Texts of term definitions can be a basis not only for description of relations but also can provide additional useful concepts or text entries for a known concept. For example, reading definitions we can understand that it is necessary to introduce such a concept as COMPUTATION OF INTEGRALS because this is an important mathematical task and there exists a set of mathematical methods to solve it. There is not such an entry in glossaries.

Every included entry is checked through its usage in Internet. It is possible to find a lot of additional information in real texts:

- ambiguity of a term that was not described in a glossary,
- non-usage of a term in real texts,
- additional text entries to a concept,
- additional useful concepts and so on.

We recheck information received from Internet texts.

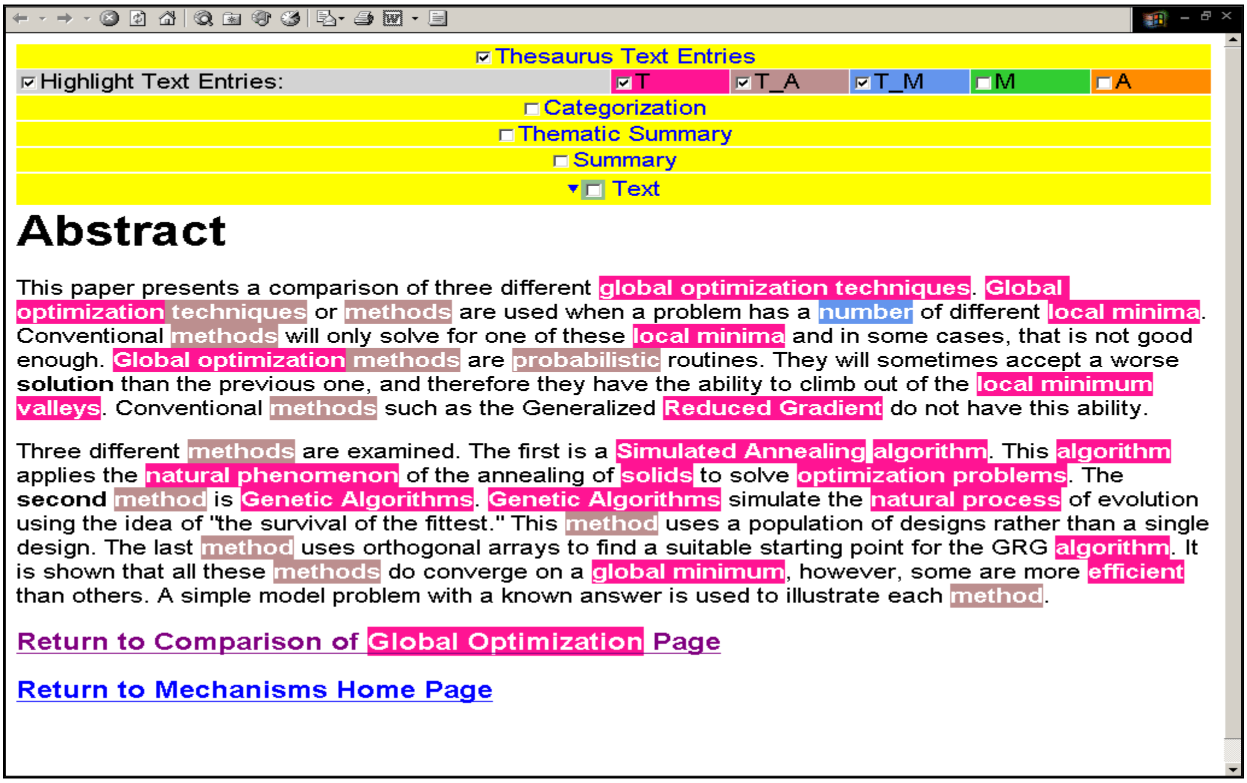
Our experts are non-professionals in the domain of natural sciences and technologies. Usually they are professional linguists. They can not apply formulae, decide a task, or do experimental work. But they can formally work with texts, compare text fragments, analyse definitions.

From our previous work with Socio-political thesaurus we know that domain experts can easily find mistakes in ready descriptions, but it is difficult for them to create descriptions from the beginning.

In construction of the ontology we try to provide maximal coverage of educational levels for every science.

Now we achieved description of higher school levels. Our next point – undergraduate levels of knowledge, and then the graduate levels. For this goal we match terminology described in the ontology and corresponding academic programs.

At present the Ontology includes 20 thousand concepts, 45 thousand terms and term variants from natural sciences and technological processes.



☑ Thesaurus Text Entries

☑ Highlight Text Entries: ☑ T ☑ T_A ☑ T_M □ M □ A

□ Categorization

□ Thematic Summary

□ Summary

▼□ Text

Abstract

This paper presents a comparison of three different global optimization techniques. Global optimization techniques or methods are used when a problem has a number of different local minima. Conventional methods will only solve for one of these local minima and in some cases, that is not good enough. Global optimization methods are probabilistic routines. They will sometimes accept a worse **solution** than the previous one, and therefore they have the ability to climb out of the local minimum valleys. Conventional methods such as the Generalized Reduced Gradient do not have this ability.

Three different methods are examined. The first is a Simulated Annealing algorithm. This algorithm applies the natural phenomenon of the annealing of solids to solve optimization problems. The **second** method is Genetic Algorithms. Genetic Algorithms simulate the natural process of evolution using the idea of "the survival of the fittest." This method uses a population of designs rather than a single design. The last method uses orthogonal arrays to find a suitable starting point for the GRG algorithm. It is shown that all these methods do converge on a global minimum, however, some are more efficient than others. A simple model problem with a known answer is used to illustrate each method.

Return to Comparison of Global Optimization Page

Return to Mechanisms Home Page

Figure 1. Ontology text entries found in a text

7. Testing of SCI-Ontology

Any ontology intended for natural language processing has to be tested on domain-specific texts.

We have a special interface program to look through results of ontology-based natural language processing. The program allow us to see:

- all terms found in text and their corresponding concepts,
- the ambiguity status of every term: an unambiguous term, an ambiguous term and alternative concepts, results of the disambiguation procedure
- relations of any concept found in a text to other concepts of the same text.
- the thematic summary of a text consisting of main thematic nodes constructed for the text – that is sets of semantically-related concepts which are considered as the most important for the text content (Loukachevitch & Dobrov 2000).
- categories automatically assigned to a text in automatic text categorization procedure. Several systems of categories can be a basis for processing at the same time.

Fig 1. shows coverage of a mathematical abstract with SCI-Ontology terms. Highlighted terms are terms included in current version of the ontology. An expert can easily see loss of important terms, for example, *orthogonal array*.

Fig 2. shows main sets of semantically related concepts of SCI-Ontology found in the text – so-called thematic summary. Concepts from the highlighted string in the thematic summary are also highlighted in the text. terms described in the ontology are presented in the bold font. An expert can easily reveal that genetic algorithms are not described in the ontology as optimisation methods and correct the mistake.

Another string of the thematic summary "local minimum, global minimum, number" reveals incorrect results of the disambiguation procedure for word *number*.

8. Conclusion

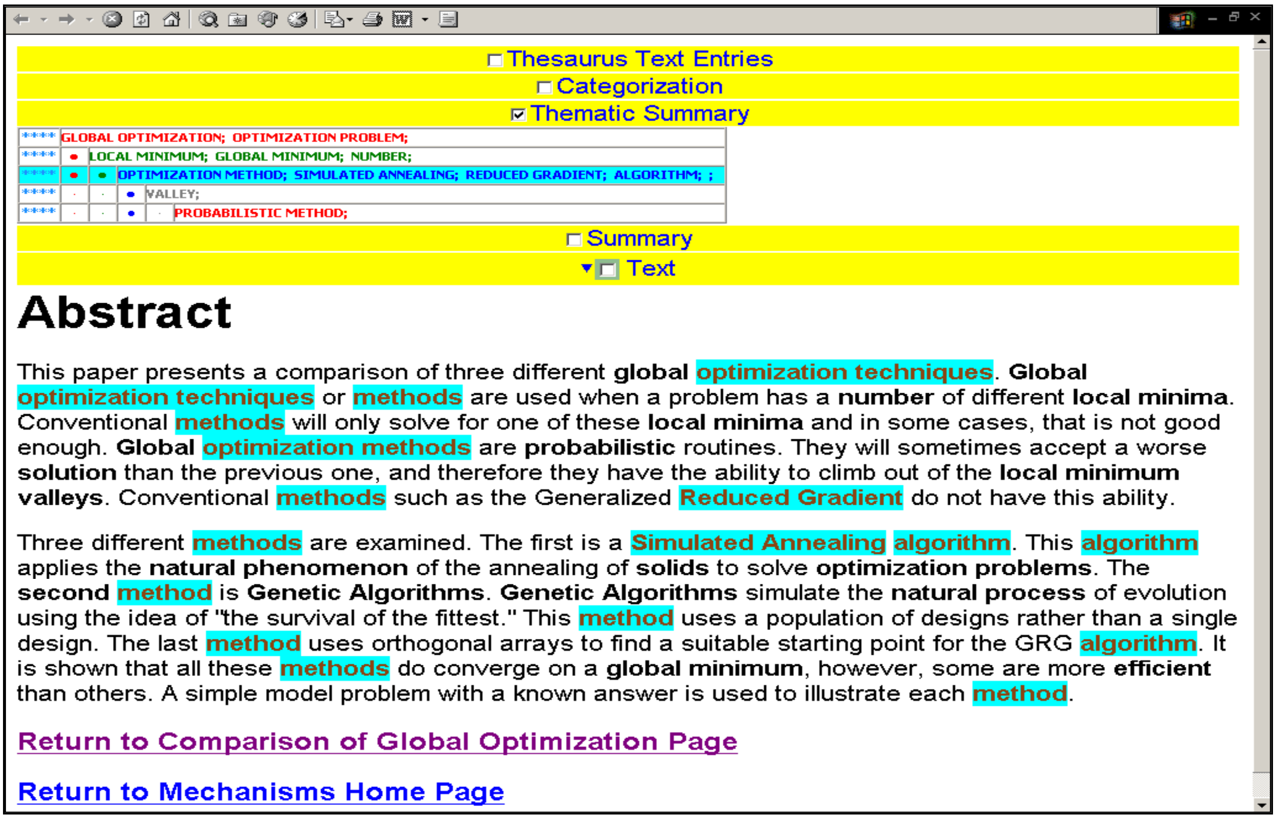
We described the main principles of development and current state of Linguistic Ontology on Natural Sciences and Technology intended for information-retrieval tasks.

In the development of the ontology we combined three different methodologies: development of information-retrieval thesauri, development of wordnets, formal ontology research. Combination of these methodologies allows us to develop large ontologies for broad domains.

Now the ontology is mainly in Russian, but in future we would like to develop the multilingual ontology of free access to facilitate the semantic search in the field of science and technology.

9. Acknowledgments

Partial support for this research is provided by the Russian Foundation for Basic Research through grant #05-07-90391.



□ Thesaurus Text Entries

□ Categorization

☑ Thematic Summary

*****	GLOBAL OPTIMIZATION; OPTIMIZATION PROBLEM;				
*****	●	LOCAL MINIMUM; GLOBAL MINIMUM; NUMBER;			
*****	●	●	OPTIMIZATION METHOD; SIMULATED ANNEALING; REDUCED GRADIENT; ALGORITHM; ;		
*****	·	·	●	VALLEY;	
*****	·	·	●	·	PROBABILISTIC METHOD;

□ Summary

▼□ Text

Abstract

This paper presents a comparison of three different global optimization techniques. Global optimization techniques or methods are used when a problem has a number of different local minima. Conventional methods will only solve for one of these local minima and in some cases, that is not good enough. Global optimization methods are probabilistic routines. They will sometimes accept a worse solution than the previous one, and therefore they have the ability to climb out of the local minimum valleys. Conventional methods such as the Generalized Reduced Gradient do not have this ability.

Three different methods are examined. The first is a Simulated Annealing algorithm. This algorithm applies the natural phenomenon of the annealing of solids to solve optimization problems. The second method is Genetic Algorithms. Genetic Algorithms simulate the natural process of evolution using the idea of "the survival of the fittest." This method uses a population of designs rather than a single design. The last method uses orthogonal arrays to find a suitable starting point for the GRG algorithm. It is shown that all these methods do converge on a global minimum, however, some are more efficient than others. A simple model problem with a known answer is used to illustrate each method.

Return to Comparison of Global Optimization Page

Return to Mechanisms Home Page

Figure 2.Semantically related concepts found in a text

10. References

AGROVOC (1999) Multilingual Agricultural Thesaurus. Fourth Edition.

ANSI/NISO (2003) Z39.19. Guidelines for the Construction, Format, and Management Monolingual Thesauri. – 2003.

Gangemi, A., Guarino, N., Masolo, C., Oltramari, A. (2001). Understanding Top-Level Ontological Distinctions // Proceedings of IJCAI 2001 workshop on Ontologies and Information Sharing.

Gomez-Perez, A., Fernandez-Lopez, M., Corcho, O. (2000). OntoWeb. Technical Roadmap. D.1.1.2. - IST project IST-2000-29243.

Guarino, N. (1998). Formal ontology and information systems. In Nicola Guarino, editor, *Formal Ontology and Information Systems, (FOIS'98)*. IOS Press.

LIV (1994). Legislative Indexing Vocabulary. Congressional Research Service. The Library of Congress. Twenty-first Edition.

Loukachevitch, N., Dobrov, B. (2000). Thesaurus-Based Structural Thematic Summary in Multilingual Information Systems. *Machine Translation Review*, 11:10-20

Loukachevitch, N.V., Dobrov, B.V. (2002). Evaluation of Thesaurus on Sociopolitical Life as Information Retrieval Tool // Proceedings of 3-rd International Conference on Language Resiurces and Evaluation (LREC2002) – Vol.1 – 2002, Gran Canaria, Spain – p.115-121.

Loukachevitch, N.V., Dobrov, B.V. (2004a). Development of Ontologies with Minimal Set of Conceptual Relations // Proceedings of 4-th International Conference on Language Resources and Evaluation / Eds: M.T.Lino et.al. – vol. VI. – 2004. – pp.1889-1892.

Loukachevitch, N.V., Dobrov, B.V. (2004b). Ontological Types of Association Relations in Information Retrieval Thesauri and Automatic Query Expansion // Proceedings of OntoLex 2004: Ontologies and Lexical Resources in Distributed Environments / Eds: A.Oltramari et.al. – 2004. – pp. 24-29.

Miller, G., Beckwith, R., Fellbaum, C., Gross, D., Miller, K (1990). Five papers on WordNet. - CSL Report 43. Cognitive Science Laboratory, Princeton University.

Sager, J.C. (1990). A Practical Course in Terminology Processing. Amsterdam: J. Benjamins.

Smith, B. (1998). *Basic tools of formal ontology*. In N. Guarino, (Ed.) Formal Ontology in Information Systems.

Soergel, D., Lauser, B., Liang, A., Fisseha, F., Keizer, J., Katz, S. (2004). Reengineering Thesauri for New Applications: the AGROVOC Example. - Article No. 257, 2004-03-17.

Tudhope, D., Alani, H., Jones, Cr. (2001). Augmenting Thesaurus Relationships: Possibilities for Retrieval. – Journal of Digital Libraries. Volume 1, Issue 8.

Voorhees, E. (1999). Natural Language Processing and Information Retrieval. In M.T.Pazienza (ed.). - Information Extraction: Towards Scalable, Adaptable Systems, New York: Springer, pp. 32-48.

Will, L. (2004). Thesaurus consultancy. – The thesaurus: review, renaissance and revision / Sandra K. Roe and Alan R. Thomas, editors. - New York ; London : Haworth, 2004. - 209p.