Evaluation of Thesaurus on Sociopolitical Life as Information-Retrieval Tool

Natalia V. Loukachevitch[*], Boris V. Dobrov[*]

[*] Research Computing Center of Moscow State University
339, Research Computing Center of Moscow State University,
Vorobyevy Gory, Moscow, 119899, Russia
{louk, dobroff}@mail.cir.ru

Abstract

In the paper we present description of Thesaurus on Sociopolitical life, which was constructed as a tool for automatic text processing of large text collections. Specific features of the thesaurus in comparison to conventional information-retrieval thesauri for manual indexing are described. Evaluation of thesaurus-based information retrieval for short queries showed considerable improvement of the model in comparison to vector model.

1. Introduction

In 60-90 years a lot of information-retrieval thesauri were created. They were intended to be used as a tool for a human subject to describe the contents of documents. But now in age of automatic systems these thesauri are hard to use in automatic text processing, that is, in automatic indexing and automatic query expansion. The problem is that conventional information-retrieval thesauri were a kind of artificial languages, and human specialists had to translate to and from these artificial languages. So it is important to understand if it is possible effectively to use thesauri in automatic text processing to enhance performance of information retrieval .

A technique of a thesaurus-based automatic indexing is a kind of conceptual indexing. The idea of conceptual indexing is often discussed (Woods, 1997; Voorhees 1999), but implementations of conceptual indexing in large domains and in real information systems are very rare. As an implementations of the conceptual indexing technique for filtration of financial news system SCISOR (Rau, 1997) can be considered. In the astronomy domain Ferret (Mauldin, 1991) system produced automatic conceptual indexing of stories. But these applications domain were narrow enough. Attempts to use WordNet (Miller et al., 1990), as a conceptual system for automatic conceptual indexing, described in (Voorhees, 1998; Richardson & Smeaton, 1995) did not show better performance of the search in comparison to vector model, based on word index.

However, in our opinion the unsuccessful attempts to enhance performance of information-retrieval systems with help of thesauri mean only that thesauri, used in automatic conceptual indexing, have to be specially constructed, have specific features, and for their effective use it is necessary to develop corresponding techniques of text processing.

Since 1994 we develop Thesaurus on Sociopolitical Life (below Sociopolitical thesaurus, the Thesaurus) for automatic conceptual indexing and since 1996 – use it (Loukachevitch et al., 1999) in such applications of automatic text processing as automatic text categorization, automatic text summarization and automatic conceptual indexing of Russian texts. The Sociopolitical thesaurus is a basic search tool in University Information System RUSSIA (Russian inter-University Social Sciences Information and Analytical consortium), which stores Russian official and legislative documents, newspaper articles.

Now Sociopolitical thesaurus includes more than 62 thousand terms, words and proper names, more than 27 thousand concepts and more than 102 thousand conceptual relations (about 700 thousand conceptual relations are inferred on a basis of relation properties).

In the paper we consider the structure of the Sociopolitical Thesaurus as a linguistic resource created specially for automatic conceptual indexing. We describe the technique of query expansion using the Thesaurus hierarchy and performance of thesaurus-based search in comparison to vector model search. At present we create Bilingual Russian-English Thesaurus on Sociopolitical Life, therefore we can illustrate the main features of the Thesaurus structure using corresponding English terms.

2. Specific Features of Thesaurus for Automatic Conceptual Indexing

The goal in developing a conventional information retrieval thesaurus is to describe terms necessary for representation of main topics of documents. More specific terms are not included. Ambiguous terms are provided with scope notes and comments convenient for human subjects (LIV, 1994). In fact a conventional information retrieval thesaurus describes an artificial language based on a real language of a domain. Human subjects have to use their domain, common sense, and grammatical knowledge not described in a thesaurus in order to index documents. Therefore conventional information-retrieval thesauri created for manual indexing are hard to utilize in an automatic indexing environment (Salton, 1989). 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.

In contrast to a thesaurus for manual indexing a thesaurus for automatic indexing should have the following special features:

At the first place a thesaurus for automatic indexing (AI-thesaurus) has to include not only terms that "represent important concepts found in literature" (LIV, 1994), but also a wide range of specific terms, which can serve a basis for identification of such 'important concepts" in a text. For similar reasons in an AI-thesaurus sets of thematically close concepts had to be included -- in thesauri for manual indexing close concepts are usually

brought to a representative concept to avoid subjectivity of indexing.

Secondly synonymic rows of concepts had to be much richer to provide different ways of identification of thesaurus concepts in texts including not only nouns and noun groups but also adjectives and verbs. Moreover considerable efforts should be made to find new synonyms of thesaurus concepts in texts. In thesauri for manual indexing many types of synonyms are missed due to grammatical and commonsense knowledge of indexers.

Such extended rows of synonyms necessarily include ambiguous terms. Additional efforts should be made to find unambiguous multiword terms with an ambiguous term as a part and to include the multiword terms in a thesaurus as new concepts or as synonyms to existing concepts.

An extended conceptual basis of an AI-thesaurus and its use in automatic text processing significantly increase role of conceptual relations. Conceptual relations in AI-thesaurus must serve for solution of three different problems:

- for every concept in a AI-thesaurus determination of a set of concepts that can be used in automatic expansion of a query, containing a given concept;
- identification of semantically related concepts in a text as a basis for better recognition of the main theme and subthemes of a text;
- term disambiguation.

3. Thesaurus on Sociopolitical life

The Thesaurus on Sociopolitical Life is a hierarchical net of concepts constructed specially as a tool for different applications of automatic text processing. It contains a lot of terms from economical, financial, political, military, social, legislative, cultural and other spheres.

The Thesaurus has the following main features:

- the hierarchy of the thesaurus can be up to 10 levels;
- concepts in the Thesaurus have rich synonymic rows. Synonyms (especially multiword synonyms) are specially collected from documents of the text collections The number of synonyms of a concept can be up to 20 elements, as, for example, for the Russian concept *NATURE PROTECTION*;
- the Thesaurus has special means for description of term ambiguity. Ambiguous terms can be described in two ways in the Thesaurus. The first -- an ambiguous term can be a synonym of two or more concepts that represent different meanings of this term. For example, term *capital* is described as a synonym to two concepts *CAPITAL CITY* and *FINANCIAL CAPITAL*. The second -- if only one meaning of an ambiguous term is represented in the Thesaurus such term is marked with a special sign of ambiguity;
- additional efforts were made to find unambiguous multiword terms with an ambiguous term as a part and to include the multiword terms in the Thesaurus as new concepts or as synonyms to existing concepts. For example, the following terms include very ambiguous words, but when these terms are included in the thesaurus, they can diminish percentage of ambiguity in a document and help to disambiguate other terms: *construction area, travel field, to appeal case, produce market, stress disorder*;
- conceptual relations used in conventional information-retrieval thesauri (LIV 1984; EUROVOC, 1991; UNBIS, 1976) such as Broader Term (BT) -- Narrower Term (NT), Related Term (RT) were supplemented by additional relations to provide the navigation in the Thesaurus net for different goals. Conceptual relations in the Thesaurus are used for automatic query expansion, for recognition of the lexical cohesion in a text, as a basis for detecting the main theme and subthemes of a text, and for term disambiguation;
- the Thesaurus is constantly tested and corrected during automatic text processing.

Table 1 compares quantitative characteristics of Sociopolitical thesaurus as a linguistic resource for automatic text processing and Legislative Indexing Vocabulary (LIV), which is used for manual indexing of documents in Research Service of U.S. Congress since 1974. Domains of the thesauri are very similar, but one can see how large is difference in volumes of described knowledge. Number of concepts in our Sociopolitical thesaurus is more in almost 4 times, every concept has 60 percents more synonyms, number of manually described relations per a concept is twice more.

Characteristics	Sociopolitical Thesaurus	LIV
Number of concepts / descriptors	27,000	6,800
Number of terms	62,000	9,800
Terms described as ambiguous	4,500	No
Number of manually described relations between concepts	102,000	15,000
Number of relations inferred using logical properties of conceptual relations	686,000	Not defined

Table 1. Quantitative characteristics of the Sociopolitical Thesaurus for Automatic Conceptual Indexing and Legislative Indexing Vocabulary of the Research Service of the Congress of the USA

Specificity of the domain leads Sociopolitical thesaurus has to include a lot of “ordinary” words.

The Sociopolitical thesaurus differs from such linguistic resources as WordNet (Miller et al., 1990) and EuroWordNet (Climent et al., 1996):

- it describes terms of the specific domain and does not include words of general language that can be used in texts of any domains;
- ambiguity of the thesaurus terms is considered relative to the domain: ambiguous terms that almost always have one meaning in the domain are described as unambiguous;

- descriptions of terms include much thematic information: possible situations, reasons, results, participants, properties and so on;
- all conceptual relations in the Sociopolitical thesaurus are tested from the point of view of the information-retrieval task.

4. Establishing of Conceptual Relations in Sociopolitical Thesaurus

Queries in an information system can be very different: longer or shorter, simple or complicated (for example, with negation). But a linguistic resource intended to be used in an automatic regime has to effectively work at least for simple queries consisting of a single term. If it does not, it can not work better for complicated queries. If it does, it is an important step to study techniques for combining or cutting expansion trees of terms in long or complicated queries. Therefore every technique or a linguistic resource can be first of all tested on simple queries consisting of a single term.

Contemporary electronic collections are usually huge collections of electronic documents different in sizes, styles, structures. Therefore it is necessary to describe relations which do not depend or almost do not depend on the theme of specific texts. Such relations have not to disappear in specific situations described in texts.

To test changeability of a relation between concepts *C1* and *C2* it is necessary to answer the following questions:

1) if every example of a concept *C1* has the relation with an example of a concept *C2* (and vice versa), for example, not every tree is in a forest, but every forest has a tree as its part;

2) if an example of concept *C1* has the relation with *C2* (or its example) during all time of its existence, for example, concept *SHOES* can be considered as *CONSUMER GOODS* (as described in WordNet 1.6), but when a specific person wears shoes, they cease to be goods;

3) if all properties of a concept *C1* are properties of concept *C2*;

4) if existence of a concept *C1* is impossible without existence of concept *C2* or existence of an example of a concept *C1* is impossible without an example of another concept *C2* (dependency relations (Guarino, 1998)), for example, existence of concept *GARAGE* is impossible without existence of concept *AUTOMOBILE.*

Sociopolitical thesaurus has three basic relations: BT-NT relations (as in traditional information retrieval thesauri), WHOLE-PART relations for descriptions of conventional parts, properties and participants of situations, RT (related term) relations for description of all other relations, which can be symmetrical and nonsymmetrical.

Examples of WHOLE-PART relations are as follows:

JUVENILE DELINQUENT
WHOLE *JUVENILE CRIMES*

UNDERWOOD
WHOLE *FOREST*

But description of every type of relations depends of answers to questions 1-4.

So if we suppose that concept *C1* is a narrower term or a part of a concept *C2* and texts containing *C1* be relevant to a simple query about *C2* we have be sure that

1a) All examples of *C1* have the relations to C2 or its examples,

2a) All time of existence of an example of *C1* these examples have the relation to *C2* or its examples,

3a) A narrower term *C1* preserves all properties of a broader term.

If there are relatively small violations of restrictions 1a)-3a) we can preserve hierarchy of a relation but we specially mark them, so we inform a processing system that a relation is weaker.

V1. If 1a) is violated, but a relation can be considered as a default relation or there are only two main alternatives, we mark the relation with modifier V (variability)

V2. If 2a) or 3a) are violated but a relation exists most time of existence of *C1* or of its example or most properties preserve we mark a relation with modifier A (aspect, point of view). For example,

SHOES
BT_A *CONSUMER GOODS*

STUDENT
$WHOLE_A$ *EDUCATION*

(last relation is not valid during all time when a person is a student, for example, when students eat, drink beer or visit a party, because a student means not only a role in education process but also social status and suppositional age).

But if violations of restrictions 1a-3a) exceed the described violations, then the relation between *C1* and *C2* can not be described, because we can not be sure that this relation can be stably useful in query expansion and in automatic text processing of various texts that mention *C1*. For example, we can not describe concept *TREE* as a part of *FOREST*, because a tree can grow in many other places.

Description of other relations is determined not by their semantic names, but restrictions 1-4.

Nonsymmetrical RT-relation (RT1 - RT2) requires fulfillment of restrictions 1, 2, 4 with possible violation V1,V2 for a lower concept. For example,

GARAGE
RT1 *AUTOMOBILE*

Symmetrical RT relation requires fulfillment of restrictions 1,2,3, with possible violations V1,V2 for both concepts.

5. Construction of the Thematic Representation of a Text

We use a specific technique for automatic text processing of documents based on thesaurus knowledge, constructing so called thematic representation of texts.

The technique of construction of the thematic representation of texts is based on the recognition of lexical cohesion in texts (Halliday & Hasan, 1976) and construction of lexical chains (Hirst & St-Onge, 1998) in the specific form of thematic nodes (detailed description

and argumentation see in (Loukachevitch, N. & Dobrov, B., 2000)).

The following main principles of lexical cohesion identification are used:

(1) One of consequences of global coherence of texts is that the main theme of a connected text can be formulated (van Dijk & Kintsch, 1983).

(2) In every connected text chains of semantically related terms passing through the whole text can be found. Such lexical chains usually contain terms from the main theme of the text and detection of such lexical chains helps to identify the main theme of the text.

(3) Existence of lexical chains in a text is greatly based on preliminary knowledge needed for understanding of the text. To detect lexical chains automatically it is necessary to restore conceptual net standing behind the text;

(4) Lexical chains are textual manifestation of specific conceptual structures - "thematic nodes" consisting of the central concept ("thematic center") and semantically related concepts. The thematic center is more important for text content then other concepts of the thematic node and had to be somehow stressed in the text. It can be used in the title or in the beginning of the text or it can have the highest frequency among related concepts;

(5) Thematic nodes with thematic centers belonging to the main theme of a text (main thematic nodes)are basis for lexical chains passing through the whole text;

(6) The main principle for automatic detection of main thematic nodes in texts is that terms corresponding main thematic nodes occur together in sentences of the text more often than other terms. This feature of main thematic nodes allows to recognize them among thematic nodes of other terms for texts of any size and different genres (Loukachevitch et al. 1999). It allows us not to follow chains of related terms from sentence to sentence to find main topics of a text. Instead we can construct various thematic nodes for the text terms using knowledge about relatedness of terms from the Thesaurus and choose such a subset of these thematic nodes whose terms are situated near each other more often than other ones. Terms of all main thematic nodes must be neighbored to each other in some context and so they form a triangle, a tetrahedron and so on of mutual coocurrence.

The thematic representation of text comprises different thematic nodes detected in the text. The thematic representation of text is a hierarchical structure of concepts where concepts semantically related to thematic centers are gathered in thematic nodes. Thematic nodes whose thematic centers can characterize contents of the text are called main thematic nodes. Hierarchy of thematic representation characterizes importance of concepts in the text: a thematic center is more important than other concepts of a thematic node, concepts of the main thematic nodes are more important than concepts of other thematic nodes.

Main stages of the automatic construction of the thematic representation of a text are as follows:

- identification of concepts in a text using morphological representation of the text and the terms of the Thesaurus;
- identification of semantic relations between the text concepts using thesaurus relations and their properties;
- disambiguation of terms based on conceptual neighborhoods of concepts corresponding to different meanings of terms (for example three meanings of a term *congress*);
- for every text concept fixation of textual context. Experimentally we received that 3 concepts to left and right sides is the best context. Such contextual concepts represent 'textual relations' of a text concept;
- creation of thematic nodes. It begins from choosing concepts to be thematic centers. The thematic centers are usually more frequent or used in the title of a text. The chosen concept gathers all semantically related concepts of the text into the thematic nodes and becomes its thematic center. A concept included into a thematic node can not become the thematic center of a new thematic node. After construction of thematic nodes frequencies of textual relations of concepts in thematic nodes are summed up, and textual relations between them are received;
- identification of main thematic nodes that
- have textual relations with all other main thematic nodes and
- have a sum of frequencies of textual relations between themselves greater than the sum of frequencies for the same number of other thematic nodes in the text. Such topological evaluation of main thematic nodes allows us to identify them in texts of any size and of great variety of genres;
- identification of specific thematic nodes and mentioned concepts. Specific thematic nodes represent subthemes discussed in a text. Concepts that are not elements of main or specific thematic nodes are called 'mentioned concepts'.

Thus all concepts of the text are divided into five classes of different importance for the text and every class has its own basic weight $n(c, D)$:

- main concepts of main thematic nodes – 0.95,
- other concepts of main thematic nodes – 0.70,
- main concepts of specific thematic nodes – 0.75,
- other concepts of specific thematic nodes – 0.60,
- mentioned concepts – 0.20.

The basic weight of a concept in the thematic representation is received as a result of aggregate distribution of thesaurus-related terms in a text. The basic weight is additionally modified to account for the frequency of a given concept.

The output weight $V(c, D)$ of concept c in document D is:

$$V(c; D) = l \cdot n^*(c, D) + (1 - l) \cdot \frac{freq(c, D)}{freq^*(D)}$$

where

$$n^*(c, D) = \max_{Thems(c, D)} n(c, D)$$

maximum of thematic weights of concept c in thematic nodes; the optimal value of l is 0.7;

$freq(c, D)$ is the frequency of concept c in document D, $freq^*(D) = \max freq(d, D)$ is the maximum frequency among concepts in a document.

6. Properties of Conceptual Relations in the Thesaurus and Automatic Query Expansion

When a linguistic resource is used for automatic query expansion, we have to determine what thesaurus relations and paths of which length can be applied for query expansion. In our thesaurus relations NT, NT_A, NT_V, PART, $PART_A$, $PART_V$ can be used in query expansion. Use of synonymic relations in query expansion follows from construction of a conceptual index. All text variants are represented as corresponding concepts in this index. Therefore terms search (unlike word search) presupposes retrieval of documents using all synonyms. Length of paths for query expansion is determined by properties of thesaurus relations.

Properties of conceptual relations in the Thesaurus are as follows:

1) Transitivity of relations NT and PART

NT	+	NT	=	NT
PART	+	PART	=	PART

2) Partial transitivity of relations NT and PART with modifiers A and V. Relations with modifiers are transitive from the point of view the name of a relation, but it is not transitive for the full relation set (name of relation, modifier):

NT	+	$NT_{A(V)}$	=	$NT_{A(V)}$
$NT_{A(V)}$	+	NT	=	$NT_{A(V)}$
PART	+	$PART_{A(V)}$	=	$PART_{A(V)}$
$PART_{A(V)}$	+	PART	=	$PART_{A(V)}$

3) Inheritance of relations WHOLE, PART, RT, RT_1 to lower concepts. For relations used in query expansion it means:

RT	+	NT	=	RT
PART	+	NT	=	PART
RT2	+	NT	=	RT2
RT2	+	PART	=	RT2

Thus, every concept of the thesaurus has a set of lower concepts obtained on basis of properties of conceptual relations and which can be used for expansion of a query containing this concept. Such a set is called 'expansion tree'.

Every concept in an expansion tree has its weight, which depend on the type of the relation to the initial concept and does not depend on length of path from the initial concept of the tree. The weights were received experimentally and now they are as follows for concept *c* from tree of concept *t*:

$c \in Tree(t)$

Q(*t* NT *c*)	=	0.9
Q(*t* PART *c*)	=	0.8
Q(*t* RT2 *c*)	=	0.6
Q(*t* RT *c*)	=	0.5

A document can contain several different concepts from an expansion tree. We sum up the weights of these concepts to do more relevant the documents containing different concepts from the expansion tree:

$$W(t,D) = 0.7 \cdot \max_{c \in Tree(t) \cap D} \{ V(c,D) \cdot Q(t,c) \} + 0.3 \cdot \max \left\{ V(t,D), \frac{R(t,D)}{1+R(t,D)} \right\}$$

where $$R(t,D) = \sum_{d \in Tree(t) \cap D} V(d,D) \cdot Q(t,d)$$

If a document contains a concept which has a relation with a modifier to the initial concept of a expansion tree, then a special technique of estimation of document relevance is used. A modifier of a conceptual relation means that a relation can be not relevant to a situation described in some texts, therefore such a relation requires additional confirmation.

We suppose that such a relation is confirmed if in a document there is another concept from the expansion tree that do not require confirmation (for example, the initial concept). In this case such a relation with a modifier is considered as a relation without any modifier.

If such a relation is not confirmed, then the modifier diminishes the weight of the relation in 2 times.

$Q(NT_{A(V)})$	=	0.45	=	(0.9/2)
$Q(PART_{A(V)})$	=	0.40	=	(0.8/2)

7. Evaluation of Thesaurus in Information Retrieval Applications

Thesaurus on sociopolitical life is used in automatic processing applications since 1996. The Thesaurus is a searching tool in University Information System RUSSIA (UIS RUSSIA, www.cir.ru/eng/), containing more 600 thousand documents. The text collection of this information system includes such various types of documents as official documents of Russian Federation, legislative acts, international treaties, newspaper articles and statistical reports.

Our Thesaurus and technique of construction of the thematic representation of a text allowed us to develop technology of flexible knowledge-based text categorization. Our text categorization system can be easily adapted to new systems of categories or other text collections from the domain. Seven text categorization systems were created. Two of these systems categorize texts using very large and hierachical categorization systems: Subject Headings of Central Election Committee of the Russian Federation (450 categories, 3levels of hierarchy, Subject Headings of Legislative Acts of the Russian Federation – more than 1000 categories, 4 levels of hierarchy).

In 1998, using our thesaurus-based technology of text summarization, we participated in the SUMMAC conference in a text categorization task. Our summaries of 'best length' had a maximal F-score (SUMMAC Final Report, 1998) . The F-score of our 10% summaries was more than medium.

The system was fully based on the Thesaurus knowledge and could not include any processing of manifold proper names, which were very important in the texts. Notwithstanding we received good results.

Therefore we consider our results in this competition as confirmation of the quality of our representation of text contents and representation of knowledge in the Thesaurus.

In this paper we present evaluation of the Thesaurus in retrieval of documents. To evaluate thesaurus-based information retrieval in University Information System RUSSIA we took 20 topics from list of "Subject Headings for Legislative Acts" adopted as an official system of subject headings in the Russian Federation. The system has 1168 subject headings and 20 main thematic subdivisions. The topics for evaluation were extracted from every subdivision of the system.

Topics usually have very short form and consist of 1-4 words. Examples of chosen subject headings are as follows: "Water supply", "Use of nuclear energy", "Migration of population". Documents were searched in the subcollection of the whole collection consisting of 50 thousand legislative acts of the Russian Federation to provide correspondence between queries and documents.

Every search was implemented twice. The first search was implemented using vector model. In the second search we manually represented a topic as boolean expression of words which are absent in the Sociopolitical thesaurus and terms from the Thesaurus. Such translation was literal without any additions or deletions. For example, subject heading "Use of nuclear energy" was represented as:

"use (word)" and "nuclear energy (term)".

During search every term was automatically expanded using its full thesaurus tree:

Word = '*USE*'
AND Concept(with Tree) = '*NUCLEAR ENERGY*'

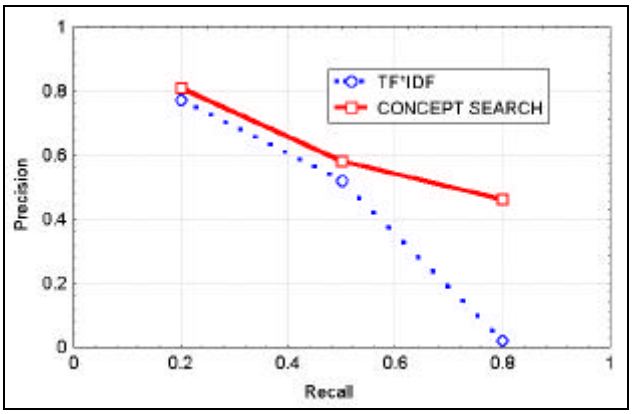

Figure 1. 3-point Recall-Precision Data for Word-Based Search (tf*idf) and Thesaurus-Based Search (Concept Search)

Most of the queries resulted in several hundreds documents from the full subcollection. To economize time of evaluation without losses in quality of evaluation we reduced time interval to receive 30-40 documents.

We used "3-point" evaluation to evaluate average precision for 0.2, 0.5, 0.8 recall values (Voorhees, 1999), as a measure of retrieval effectiveness.

Results of evaluation are presented on Figure 1 and in Table 2.

Type of search	0.2	0.5	0.8	Average
Vector Search	0.77	0.52	0.02	0.44
Thesaurus Search	0.81	0.58	0.46	0.62

Table 2.

8. Example of Search for Query *"PROTECTION OF LABOUR"*

The search was in Russian, here we present English translation of Russian terms.

Word-based Query:
Word='*PROTECTION*' AND Word='*LABOUR*'

Thesaurus-based Query:
Concept(with Tree) = '*PROTECTION OF LABOUR*'

Dates of documents were from 01.10.2000 till 01.01.2001.

The used tree of concept *PROTECTION OF LABOUR* was in Russian, here we present English translation of Russian terms:

PROTECTION OF LABOUR (labor protection, labour protection, protection of labour)
1 LABOUR SAFETY (employee safety, industrial safety, job safety, labor safety, occupational safety, safety of workers, work safety)
1 INDUSTRIAL SAFETY MAINTENANCE (safety engineering, safety maintenance in industry, safety measures in industry)
2 ACCIDENT-FREE USAGE
1 HYGIENE OF LABOUR (hygiene of labor, factory hygiene, industrial hygiene)
1 SPECIAL WORKING CONDITIONS
2 BAD LABOUR CONDITIONS (bad labor conditions, harmful labor conditions)
2 WORKS ON THE NIGHT SHIFT (night shift, night work, graveyard work, graveyard shift, lobster shift)
2 OVERTIME WORK (overwork, overtime(Amb.), work overtime, work extra)
1 OCCUPATIONAL TRAUMA (employment injury, industrial injury, work-related injury, work-related personal injury)
1 OCCUPATIONAL DISEASE (occupational health, work-related disease, work-related illness, health at work, health of workers, employee health)
2 ANTRACOSIS (black lungs, coal miner's lungs)
2 ASBESTOSIS
2 BERYLLIOSIS
2 BISSINOSIS (brown lungs)
2 SILICOSIS
2 OCCUPATIONAL MORBIDITY
...

There were 31 concepts in the tree and 70 text entries.
Results of word-based search - 33 documents .

Results of thesaurus-based search - 26 documents (only 17 without tree).

Number of relevant texts - 28 documents.

3-point precision for word-based search (0.60; 0.66; 0.00).

3-point precision for thesaurus-based search (0.86; 0.93; 0.99).

9. Conclusion

In the paper we present description of Thesaurus on Sociopolitical life, which was constructed as a tool for automatic text processing of large text collections. Specific features of the thesaurus in comparison to conventional information-retrieval thesauri for manual indexing are described. A thesaurus constructed for automatic text processing has to include considerably more concepts, synonyms and relations between concepts, has specific means for description of term ambiguity. Evaluation of thesaurus-based information retrieval for short queries showed considerable improvement of the model in comparison to vector model.

In future we plan to study how to combine and cut expansion trees for long queries and how use thesaurus expansion in automatic processing of natural language queries. At present we develop Bilingual Russian-English Thesaurus on Sociopolitical Life to use it in bilingual information retrieval. Now English part of the Thesaurus contains about 47 thousand English terms.

10. Acknowledgements

Partial support for this research is provided by the Russian Foundation for Humanities through grant # 00-04-00272.

11. References

Callan, J.P., Croft, W.B. and Harding, S.M., 1992. The INQUERY Retrieval System. In A.M. Tjoa and I. Ramos (eds.), *Database and Expert System Applications*. Springer Verlag, New York.

Climent, S., Rodriguez, H. and Gonzalo, J., 1996. Definitions of the links and subsets for nouns of the EuroWordNet project. - Deliverable D005, WP3.1, EuroWordNet, LE2-4003.

van Dijk, T.A. and Kintsch, W., 1983. *Strategies of Discourse Comprehension*. New York. Academic Press.

Halliday, M. and Hasan, R. 1976. *Cohesion in English*. Longman, London.

Hirst, G. and St-Onge, D., 1998. Lexical Chains as representation of context for the detection and correction malapropisms. In C. Felbaum (ed.), *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, Massachusetts, London, England,305-332.

Guarino, N., 1998. Some Ontological Principles for Designing Upper Level Lexical Resources. In *Proceedings of First International Conference on Language Resources and Evaluation*.

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

Loukachevitch, N.V., Salii, A.D. and Dobrov, B.V., 1999. Thesaurus for Automatic Indexing: Structure, Developement, Use. In P. Sandrini (ed.), *Proceedings Fifth International Congress on Terminology and Knowledge Engineering*. TermNet, Vienna. 343-355.

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

Mauldin, M.L., 1991. Retrieval Performance in FERRET: A Conceptual Information Retrieval System *The 14th International Conference on Research and Development in Information Retrieval*.

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

Rau, L.F., 1997. Conceptual Information Extraction and Retrieval from Natural Language Input. In K. Sparck Jones and P. Wilett (eds.), *Readings in Information Retrieval*. Morgan Kaufmann, San Francisco, 527-533.

Richardson, R. and Smeaton, A. 1995. Using WordNet in a Knowledge-Based Approach to Information Retrieval. *School of Computer Applications. Working Paper*, CA-0395.

Salton, G., 1989. *Automatic Text Processing - The Analysis, Transformation and Retrieval of Information by Computer*. Addison-Wesley, Reading, MA.

SUMMAC Final Report, 1998. The TIPSTER SUMMAC Text Summarization Evaluation. Final Report. *MITRE Technical Report*, MTR 98W0000138.

EUROVOC, 1995. *Thesaurus EUROVOC*. Vol.1-3, European Communities. – Luxemburg: Office for Official Publications of the European Communuties - Ed.3. – English version.

UNBIS, 1976. *UNBIS Thesaurus*. English Edition, Dag Hammarskjold Library of United Nations, New York.

Voorhees, E.M, 1998. Using WordNet for Text Retrieval. In C. Felbaum (ed.), *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, Massachusetts, London, England, 285-303.

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

Woods, W.A., 1997. Conceptual indexing: a better way to organize knowledge. *SunMicrosystems Laboratories Technical Report*, SMLI TR-97-61.