Conceptual Indexing Using Thematic Representation of Texts

Boris V. Dobrov, Natalia V. Loukachevitch, Tatyana N. Yudina

Center for Information Research
339, Scientific Research Computer Center of Moscow State University
Vorobyevy Gory, Moscow, Russia
cir@online.ru

Abstract

We present the thesaurus-based indexing technology developed by the Center for Information Research under the Information System RUSSIA project. The technology is based on using basic properties of coherent text. Initially the technology was applied for automatic processing of Russian official (government) texts. Currently the instrument is adapted to process English texts for TREC-6 routing task.

1. Introduction

The indexing approach described here is the result of NLP-technology developed under the Information System RUSSIA project. The IS RUSSIA project pursues three main goals:

- to create and support a public domain computer-based library, designed and developed to also serve as a database for social studies and university education;
- to realize NLP technology for the Russian language; and
- to develop an adequate complex of searching tools and a user-friendly English interface in order to serve as a bilingual information resource available on-line for foreign users.

The technological approach realized under the IS RUSSIA project is based on research in linguistics. It is aimed at automatic Russian language text processing, understanding and information analysis (Yudina T., Dorsey P. 1995). The main approach is to analyze the content of a text. Currently a deep-structured search image is created for every text. In addition to traditional bibliographic fields, the search image also includes thesaurus-based components: subject headings, a list of topics described, main and specific thematic nodes, mentioned descriptors, and relations between topics. The thesaurus-based components provide for thematic representation of a text that is used for indexing, categorization and summarization of a text. Ranked query results presentation is based on this technique.

The technology is currently applied to the Russian official document electronic text corpora - one of the most complicated ones. The next text corpora it will be applied to is the news media. The system will provide automatic processing, indexing, and event categorization of messages and electronic editions of Russian leading informational agencies, newspapers and magazines. In the future we hope to realize a reference technique that will expand the analytical component of the system and enable it to keep track of a situation in its dynamics, as a next step - to compare reports coming from different sources.

The technology has made it possible to develop the IS RUSSIA as an integrated information warehouse that can be searched and retrieved across in its entirety. The search engine includes a system of subject headings and thesaurus-based retrieval as well as context search techniques. The most sophisticated instrument is the Thesaurus on Sociopolitical Life. Developed as part of the project, it incorporates more than 21,000 terms and 8,000 geographic names. It assists in navigation across the huge masses of textual data and enables query expansion based on the concept relationship encoded in the thesaurus.

The IS RUSSIA has been designed as part of the international information structure so as to serve not only Russian researchers but foreign specialists on Russia as well. This application required the creation of a set of bilingual tools including a user-friendly interface, help screens, reference databases, and search instruments. The search tools include the

"System of Subject Headings" (about 200 entries) and the bilingual version of the Thesaurus on Sociopolitical Life (currently with more than 10,000 equivalents), the work is underway to translate it in full and to compose the thesaurus in English (currently it is still a set of translations from Russian). The English translation is being done in concordance with the "System of Subject Headings" of the Library of Congress, the "Legislative Indexing Vocabulary" of the Congressional Research Service (LIV 1990), the United Nations thesaurus (UNBARS Thesaurus 1976), the LegiSlate thesaurus, the Westlaw thesaurus, and the EVROVOC (thesaurus of the European Economic Community).

The search tools include the optional use of subject headings systems that are mostly popular among foreign experts (those of the Congressional Research Service, US Library of Congress and the LegiSlate; the work underway is to include the system of subject headings of the European Economic Community).

The IS RUSSIA is being developed using Oracle Server (SCO UNIX). Linguistic software mainly run under MS-DOS. Four P/5-133 were used for the TREC-6 routing task.

The IS RUSSIA integrates a wide variety of official data and documents (laws, presidential edicts and directives, governmental enactments, acts and regulations), and exceeds 150 Mb of pure document text. The collection covers the period from 1994 till now. It is updated on a regular basis from official first-hand sources, and contains all open official documents. The system includes reference data on the Russian political system (brief history, prerogatives, structure and personnel of federal institutions, political parties, churches, etc.); extended reference information on the constituent members of the Russian Federation; economic indicators and election statistics.

The team of developers is a non-commercial organization - the Center for Information Research housed at the Research Computer Center of the Moscow State University. The team includes 20 specialists from academic institutions and universities of Moscow, and consists of system analysts, programmers, linguistic researchers and social scientists. Financial support of the project was provided by foreign charitable funds, the Russian government, and scientific funds.

The IS RUSSIA was initially designed to serve as an information warehouse for social investigations. This purpose requires a representative and regular updated complex of databases storing data and documents from a wide scope of resources. The Internet-based foreign resources may significantly enrich the information flow. Special part of the IS RUSSIA project are efforts aimed at applying the developed NLP technology on processing of large collections of English texts. TREC-6 is our first experience using English texts.

2. Thematic Representation of Text

The core of the indexing technology is the thematic representation of a text. The thematic representation serves for description of contents of a document and is constructed using thesaurus knowledge about terms and property of text cohesion.

Text cohesion is achieved through semantically related terms, reference, ellipsis and conjunctions (Halliday and Hasan, 1976). Lexical cohesion is the most frequent type of cohesion. It can be achieved by repetitions, synonyms and hyponyms (reiteration) or by thematically related terms (collocation) for example: *aircompany*, *aircraft*.

Sequences of terms which the lexical cohesion relation holds can be incorporated into lexical chains. It is clear intuitively that lexical chains are connected with discourse and topical structure of the text, and so their recognition is very important for automatic text processing and representation of document content. To construct lexical chains, a linguistic resource describing relations between terms is needed. Both (Barzilay and Elhadad, 1997) and (Hirst and St-Onge, 1997) construct lexical chains based on WordNet (Miller et al. 1990). However, WordNet does not describe thematic relations between synsets (Climent et al. 1996) and therefore thematic relations are not used in the constructions of lexical chains.

Consideration of thematic relations changes a system of lexical cohesion relations in the text because a term can support some lexical chains simultaneously. For example, *minister* can support lexical chains of *government* and *ministry*, *astronaut* -- *cosmonautics* and *human*, *ratification* - lexical chains of *international treaty*, *the State Duma of Russia* and *the Congress of the USA* at the same time. It means that lexical cohesion is not based on a set of isolated lexical chains but on a complicated net of different relations between terms.

Semantically or thematically related terms of the text are not always connected with lexical cohesion relation. The existence of this relation is more likely for related terms in the same segment of the text than for terms in different segments and for domain specific terms than for words of common language. At the same time lexical cohesion can be the only means connecting text segments situated far from each other in the text. Thus it can be difficult to automatically decide if the relation of lexical cohesion holds between two related terms of the text.

In (Barzilay and Elhadad, 1997) the terms in text segments can be incorporated into lexical chains if they are members of synset of WordNet, if one is the child of the other in the hyperonym graph and in some cases if they are siblings in the hyperonym graph. Two lexical chains from different text segments are incorporated into a single chain if they contain a common word with the same sense. The lexical chains constructed in this manner can include terms that are not related to each other and have a bizarre form if they are represented as graphs of concepts.

We require that a lexical chain must represent a concept from the topical structure of text. Van Dejk (van Dejk, 1983) describes the topical structure of text **-** the macrostructure- as a hierarchical structure in the sense that the theme of a whole text can be identified and summed up to a single macroproposition. The theme of the text can usually be described in terms of less general themes which in turn can be characterized in terms of even more specific themes, and so on.

We approximate the highest macroproposition of the macrostructure with the set of macroconcepts that name the predicate of the macroproposition and its arguments. Each text is mainly devoted to description of the relations between these macroconcepts. This means (and our experiments confirm (Lukashevich, 1995)) that in most cases repetitions and synonyms of a macroconcept in the text are co-referent or are in relation of conceptual identity with the macroconcept. In most cases hyponyms, hyperonyms and thematically related terms of the macroconcept participating in subtopics of the text characterize different aspects of this macroconcept. Thus we can construct a lexical chain including a macroconcept and all text terms related to the macroconcept. We call such lexical chains 'thematic nodes'. The term that all terms of the thematic node are related to is called 'thematic center'.

Since we could construct thematic nodes for any term of the text as a thematic center, the question is how to distinguish thematic nodes of macroconcepts (main thematic nodes) from all possible thematic nodes of the text. Again we must remember that the text is devoted to description of relations between macroconcepts and so most sentences of the text must characterize these relations. This means that elements of different main thematic nodes occur together in sentences of the text more often than other terms. This distinguishes main thematic nodes from all other thematic nodes for texts of any size and different genres.

Thus the thematic representation of text is a hierarchical structure of terms where terms semantically or thematically 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. The thematic representation hierarchy characterizes the importance of terms in the text: the thematic center is more important than other terms of the thematic node, and terms of main thematic nodes are more important than terms of other thematic nodes.

Thematic representations are created on the basis of detailed description of the domain, represented as a thesaurus. Our Thesaurus was specially created as a tool for automatic processing of texts in the broad domain of sociopolitical life and has some essential distinctions from conventional thesauri created for manual indexing.

3. Thesaurus on Sociopolitical Life

We created our Thesaurus as a tool for automatic indexing -- the Thesaurus on Sociopolitical Life. It was constructed for indexing of different types of Russian texts in a broad domain of sociopolitics (such as official documents or news reports).

The Thesaurus was created in semi-automatic mode using automatic processing of more than 150 Mb of Russian official texts (Lukashevich 1995). The thesaurus units represent real text expressions. In this sense Thesaurus is similar to such thesauri as WordNet (Miller et al. 1990) and Roget's thesaurus. Carefully gathered terms form rows of synonyms for concepts (descriptors of Thesaurus). Adjectives and verbs that are derivatives of a descriptor can also be its variants.

Ambiguous terms can be described in two ways in the Thesaurus. An ambiguous term can be a quasi-synonym of two or more descriptors that represent different meanings of this term. For example, (hereinafter we give fragments from the Thesaurus in English translation) term *capital* is described as a synonym to two descriptors *CAPITAL (City)* and

CAPITAL (Finance). If only one meaning of an ambiguous term is represented in the Thesaurus such term is marked with a special sign of ambiguity.

Existing relationships between descriptors in Thesaurus are: broader term (BT) -- narrower term (NT), related term (RT), whole-term (WT) -- part-term (PT). Latter relationship is used for description of physical parts, elements and objects of a concept.

Using these relations we developed our Thesaurus as a thesaurus inheritance system in which more specific concepts inherit information from more general concepts. In our system this means that relationship "related term" is inherited from a descriptor by its narrower descriptors and by its parts. Relationship "part-term" is inherited from a descriptor by its narrower descriptors. Relationships “broader term --narrower term” and “whole-term --part-term” are transitive relationships.

Thus every descriptor of Thesaurus is related to a wide scope of terms. For most descriptors the number of related descriptors is much larger than the number of direct indicated relationships. For example, descriptor *AGRICULTURE* has 26 direct relations with other descriptors, but through the properties of inheritance and transitivity it is related to more than 300 ones (branches of agriculture, agricultural enterprises, domestic animals and plants and so on).

This extended set of related terms in Thesaurus enables us to determine which terms of the text are semantically or thematically related to each other and can support a topic or a subtopic of the text. As an example, a description of the concept “fishing” is represented on Figure 1.

Currently the Thesaurus contains in Russian more than 21 thousand terms and 8 thousand geographic names (15,000 descriptors and about 40,000 relations between descriptors).

4. Construction of Thematic Representation

In this section we describe our technique of conceptual indexing initially used for processing of Russian texts. The technique was adapted to TREC-6 routing task with insignificant changes.

4.1. Identification of Terms in Texts

Text units are compared with the terms of the Thesaurus using morphological representation of the text and terms. If the same fragment of a text corresponds to different descriptors of the Thesaurus, ambiguity of the text unit is indicated.

Texts can include names that coincide with terms of the Thesaurus. A name that corresponds to a term of the Thesaurus but has different spelling (capital letters, quotes) is also marked as an ambiguous term.

After comparison with the Thesaurus the text is represented as a sequence of descriptors. All synonyms of any descriptor are represented by that descriptor and are not differentiated further. For every text descriptor related text descriptors are given. A set of text descriptors together with relations to related text descriptors is called a “thesaurus projection”.

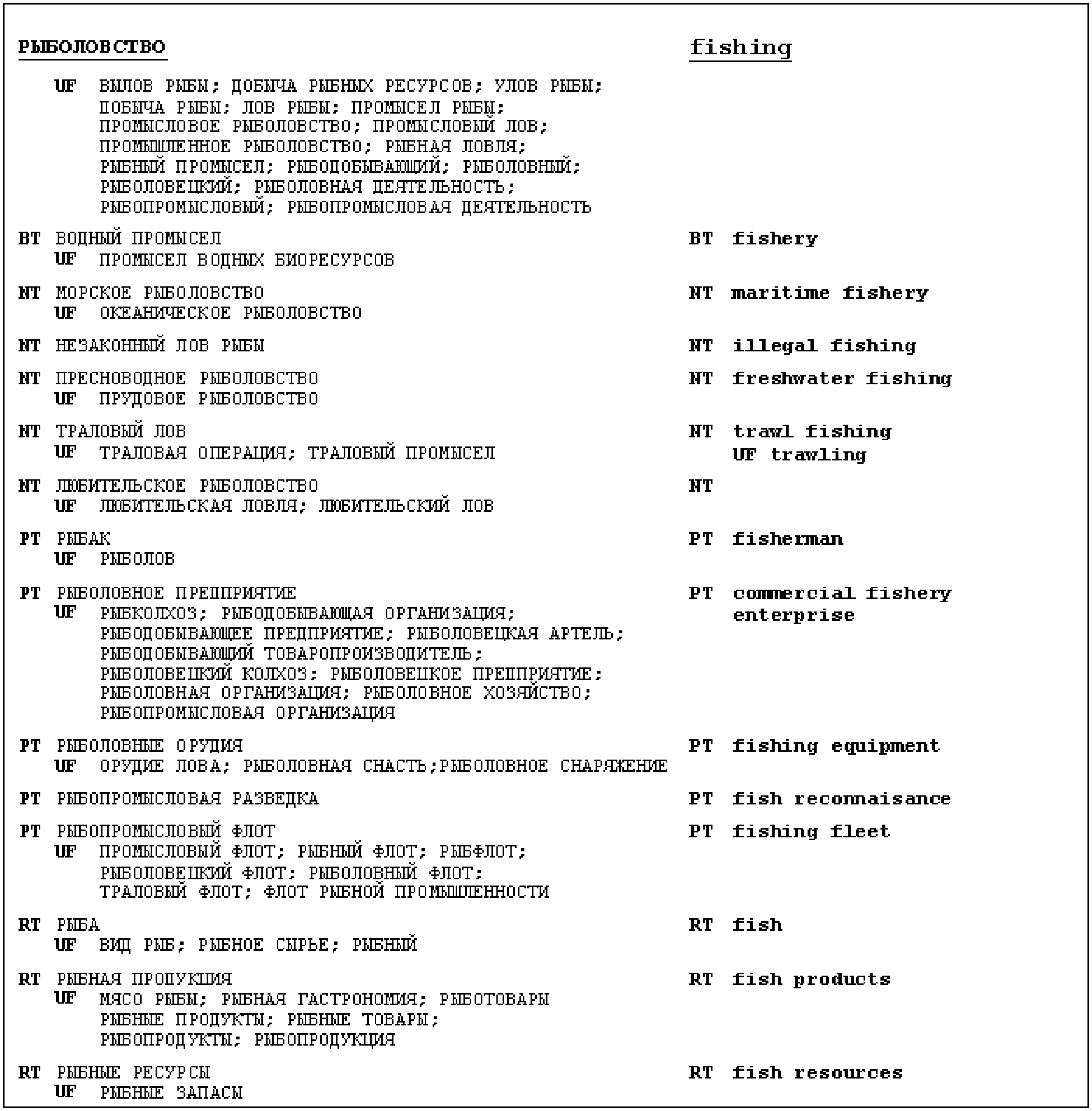
4.2. Disambiguation of Terms Using Thesaurus Projection

Descriptors corresponding to different meanings of ambiguous terms also participate in the construction of the thesaurus projection for a text. Using the thesaurus projection a proper meaning of an ambiguous term is chosen.

For every meaning of an ambiguous term the following conditions are checked. If one of the conditions is met, we consider the text to support this meaning of the ambiguous term.

1) A descriptor corresponding to a meaning of the ambiguous term is used in text in unambiguous form. For example, term *financial capital* is an unambiguous term for descriptor *CAPITAL(Finance)* and *capital* is an ambiguous term for this descriptor;

2) A descriptor corresponding to a meaning of the ambiguous term is related to other descriptors in the thesaurus projection. For example, descriptor *PUBLIC ORGANIZATION* is connected by relationship NT with descriptor *POLITICAL PARTY* that corresponds to one of the meanings of ambiguous term *party*.



```
РЫБОЛОВСТВО                                            fishing

   UF  ВЫЛОВ РЫБЫ; ДОБЫЧА РЫБНЫХ РЕСУРСОВ; УЛОВ РЫБЫ;
       ДОБЫЧА РЫБЫ; ЛОВ РЫБЫ; ПРОМЫСЕЛ РЫБЫ;
       ПРОМЫСЛОВОЕ РЫБОЛОВСТВО; ПРОМЫСЛОВЫЙ ЛОВ;
       ПРОМЫШЛЕННОЕ РЫБОЛОВСТВО; РЫБНАЯ ЛОВЛЯ;
       РЫБНЫЙ ПРОМЫСЕЛ; РЫБОДОБЫВАЮЩИЙ; РЫБОЛОВНЫЙ;
       РЫБОЛОВЕЦКИЙ; РЫБОЛОВНАЯ ДЕЯТЕЛЬНОСТЬ;
       РЫБОПРОМЫСЛОВЫЙ; РЫБОПРОМЫСЛОВАЯ ДЕЯТЕЛЬНОСТЬ

BT ВОДНЫЙ ПРОМЫСЕЛ                                     BT fishery
   UF  ПРОМЫСЕЛ ВОДНЫХ БИОРЕСУРСОВ

NT МОРСКОЕ РЫБОЛОВСТВО                                 NT maritime fishery
   UF  ОКЕАНИЧЕСКОЕ РЫБОЛОВСТВО

NT НЕЗАКОННЫЙ ЛОВ РЫБЫ                                 NT illegal fishing

NT ПРЕСНОВОДНОЕ РЫБОЛОВСТВО                            NT freshwater fishing
   UF  ПРУДОВОЕ РЫБОЛОВСТВО

NT ТРАЛОВЫЙ ЛОВ                                        NT trawl fishing
   UF  ТРАЛОВАЯ ОПЕРАЦИЯ; ТРАЛОВЫЙ ПРОМЫСЕЛ               UF trawling

NT ЛЮБИТЕЛЬСКОЕ РЫБОЛОВСТВО                            NT
   UF  ЛЮБИТЕЛЬСКАЯ ЛОВЛЯ; ЛЮБИТЕЛЬСКИЙ ЛОВ

PT РЫБАК                                               PT fisherman
   UF  РЫБОЛОВ

PT РЫБОЛОВНОЕ ПРЕДПРИЯТИЕ                              PT commercial fishery
   UF  РЫБКОЛХОЗ; РЫБОДОБЫВАЮЩАЯ ОРГАНИЗАЦИЯ;              enterprise
       РЫБОДОБЫВАЮЩЕЕ ПРЕДПРИЯТИЕ; РЫБОЛОВЕЦКАЯ АРТЕЛЬ;
       РЫБОДОБЫВАЮЩИЙ ТОВАРОПРОИЗВОДИТЕЛЬ;
       РЫБОЛОВЕЦКИЙ КОЛХОЗ; РЫБОЛОВЕЦКОЕ ПРЕДПРИЯТИЕ;
       РЫБОЛОВНАЯ ОРГАНИЗАЦИЯ; РЫБОЛОВНОЕ ХОЗЯЙСТВО;
       РЫБОПРОМЫСЛОВАЯ ОРГАНИЗАЦИЯ

PT РЫБОЛОВНЫЕ ОРУДИЯ                                   PT fishing equipment
   UF  ОРУДИЕ ЛОВА; РЫБОЛОВНАЯ СНАСТЬ;РЫБОЛОВНОЕ СНАРЯЖЕНИЕ

PT РЫБОПРОМЫСЛОВАЯ РАЗВЕДКА                            PT fish reconnaisance

PT РЫБОПРОМЫСЛОВЫЙ ФЛОТ                                PT fishing fleet
   UF  ПРОМЫСЛОВЫЙ ФЛОТ; РЫБНЫЙ ФЛОТ; РЫБФЛОТ;
       РЫБОЛОВЕЦКИЙ ФЛОТ; РЫБОЛОВНЫЙ ФЛОТ;
       ТРАЛОВЫЙ ФЛОТ; ФЛОТ РЫБНОЙ ПРОМЫШЛЕННОСТИ

RT РЫБА                                                RT fish
   UF  ВИД РЫБ; РЫБНОЕ СЫРЬЕ; РЫБНЫЙ

RT РЫБНАЯ ПРОДУКЦИЯ                                    RT fish products
   UF  МЯСО РЫБЫ; РЫБНАЯ ГАСТРОНОМИЯ; РЫБОТОВАРЫ
       РЫБНЫЕ ПРОДУКТЫ; РЫБНЫЕ ТОВАРЫ;
       РЫБОПРОДУКТЫ; РЫБОПРОДУКЦИЯ

RT РЫБНЫЕ РЕСУРСЫ                                      RT fish resources
   UF  РЫБНЫЕ ЗАПАСЫ
```

Figure 1. Example of CIR Thesaurus concept description

If the text supports only one meaning of the ambiguous term the corresponding descriptor is chosen. If the text supports more than one meaning of the term we look through descriptors that are the nearest ones to every usage of the ambiguous term and choose the meaning of the descriptor supported by the nearest descriptors.

Only chosen descriptors participate in further processing of the text.

4.3. Construction of Thematic Nodes

We assume that the term that characterizes a topic of the text and therefore can become the thematic center of a thematic node is usually stressed in a text. It can be used in the title or in the beginning of the text or it can have the highest frequency among terms of the topic.

Any term of the Thesaurus (either general or specific one) can become the thematic center of a thematic node. For example, term *mathematics* can become the main term of a topic if the text is devoted to development of mathematics, or term *scientist* can become the main term of a topic if a text is about “brain drain” to foreign countries.

Creation of thematic nodes begins by choosing the thematic centers. First, descriptors mentioned in the title and first sentence of the text gather all related descriptors from the thesaurus projection and become the thematic centers of thematic nodes. Then the most frequent descriptors of the text can become thematic centers. A descriptor included into a thematic node cannot become the thematic center of a new thematic node.

Let us analyze document FBIS-F001-0015 (Figure 2). Some thematic nodes that were constructed during automatic processing of the example text (the right column represents descriptor frequency in the text) are as follows:

Russia (Russian)	10
Far East	1
Curile	1
President of Russia	1
state (country)	6
territorial waters	*9*
ocean	3
ship	4
island	1
state (country)	6
President of Russia	1
fish	11
fishing	5
fisherman	2
illegal fishing	1
pouching (pouch)	5
illegal activity	1
illegal fishing	1
fisherman	2

Border Troops `Putina' Exercise to Control Poaching

[Text] The border troops "are not saber rattling" in Russian territorial waters in the Far East as the mass media, especially the Japanese mass media, are attempting to portray it. Servicemen have been legally granted the right to utilize all of the tools at their disposal, including weapons, to put a stop to poaching. Russian Border Troops Commander-in-Chief Colonel-General Andrey Nikolayev stated that to an ITAR-TASS correspondent while stressing that his subordinates are conducting a strict policy to put a stop to the illegal activities of foreign boats. He noted that the President of Russia supports the position of the border troops for the full observance of the law in the country's territorial waters.

Recently, we have become accustomed to reports on the entry of Japanese fishing boats into Russian territorial waters to poach fish. According to official data, the number of such violations has increased by a factor of 3.5-4 in 1993, in contrast to 1990. And although the Russian border guards, who are experiencing great difficulties in logistics-technical support due to the well-known economic situation in the country, have been able to observe approximately 140 foreign fishing boats and to fine poachers a sum of more than 21 million rubles and over 100,000 U.S. dollars in 199, so far, their efforts are a drop in the sea. These fines have hardly made up for the damage from more than 7,500 pirate entries into Russia's territorial waters.

....

(full text size is about 7 Kb).

Figure 2. Fragment of FBIS-F001-0015 document
(terms of four thematic nodes are underlined)

4.4. Determination of Status of a Thematic Node

In the previous stage thematic nodes were gathered. Each thematic node includes descriptors of the thesaurus projection that are related to its thematic center. Thematic nodes correspond to topics or subtopics discussed in a text. At this stage it is necessary to evaluate the importance of topics and thematic nodes representing these topics in the text. The first step is to determine main topics of the text, that is to choose main thematic nodes.

In our approach we assume that in normal, conventional texts main topics pass through the whole text and are discussed in combination with each other. This means that descriptors of different main thematic nodes are usually located together all over the text. To find out how descriptors of thematic nodes are distributed in the text we use the notion "textual relation": a given descriptor has textual relations with those descriptors of the text that are located not further than N descriptors from the given descriptor (location order is not important). Currently N=2, so every usage of a descriptor in the text is considered in a sequence of descriptors by length 7. Thus we assume that in a text descriptors of thematic nodes are usually repeated over seven descriptors. This approach originates on the basis of experiments in psychology and linguistics.

As a result we obtain a set of textual relations for every descriptor of a text.

Textual relations between descriptors are determined at the stage of comparison of text with Thesaurus. After thematic nodes are constructed, textual relations frequencies of descriptors in each thematic node are summed to compute the textual relations between thematic nodes.

In our approach we assume that first of all main thematic nodes are those ones that

- have textual relations with all other main thematic nodes and
- have a sum of frequencies of textual relations between these nodes greater than the sum of frequencies for the same number of other thematic nodes of this text.

The thematic nodes for the example in Figure 1 are thematic nodes with main descriptors *territorial waters, fish, Russia, Japan, border troops, poaching, boat, ...*
Thus we can produce a "thematic summarization" of text (right column represents total frequency of thematic node descriptors):

territorial waters; state (country);ship; ocean; island; President of Russia	24
fish; fishing; fisherman; illegal fishing;	20
Russia (Russian); state (country); Far East; Curile; President of Russia	19
Japan; continental shelf; state(country)	18
border troops; border guard, state(country)	17
pouching (pouch); fisherman; illegal activity; illegal fishing	9
boat	7

These requirements for main thematic nodes determine a threshold that distinguishes main thematic nodes form all other thematic nodes of a text. This threshold is an average frequency of descriptors in determined main thematic nodes. The initial set of main thematic nodes is supplemented with those thematic nodes whose frequency is more than the threshold.

In addition to main thematic nodes there are specific thematic nodes and mentioned descriptors. Specific thematic nodes represent primary characteristics of main topics discussed in the text. Specific nodes are those thematic nodes that have textual relations with at least two different main thematic nodes. Descriptors that are not elements of main or specific thematic nodes are called mentioned descriptors.

In our example specific thematic nodes are:

logistics
    equipment
    monitor
    computer

mass media
    correspondent

The first one is represented in the following paragraphs of example (Figure 3). Mentioned descriptor are *weapon, expert, ice situation ....*

.... (3rd paragraph)

But then again, we can explain the definite impunity of violators not only through the problems in logistics-technical support, due to which border troops maritime units and aircraft have been compelled to reduce their activities (for example, last year the United States had 3.2 ships per 100,000 square kilometers of economic zone, Japan had 8.2, and Russia had 2.1), but also through the obvious delay in the adoption of the laws "On the Russian Federation's Exclusive Economic Zone" and "On the Russian Federation's Continental Shelf",..

.... (8th paragraph)

It is noteworthy that the poachers' schooners have been well adapted for "wolf-like" swoops into our territorial waters. They have excellent navigation equipment, they are equipped with computers and they are maneuverable. Maneuverability also helps them to feel quite confident in themselves even under conditions of a complex ice situation (up to 4-5 balls)....

Figure 3. Fragments of FBIS-F001-0015 document

Thus all descriptors of the text are divided into five classes of decreasing importance for the text:

- main descriptors of main thematic nodes,
- other descriptors of main thematic nodes,
- main descriptors of specific thematic nodes,
- other descriptors of specific thematic nodes,
- mentioned descriptors.

5. Text Categorization Using Thematic Representation of Text

The thematic representations of texts can serve as a basis for text categorization. It was used for processing of TREC-6 routing task when TREC-6 topics were described as categories for text categorization.

5.1. Relations between the Thesaurus and Categories

Our technique allows us to carry out text categorization using different systems of categories.

We consider a category to be a user defined query that has to be represented by descriptors of the Thesaurus. The hierarchical structure of the Thesaurus allows to choose a subtree of the Thesaurus corresponding to the category and connect the category with upper descriptor of this subtree. We call such a descriptor “supporting descriptor” of the category.

A category can be represented by a set of descriptors. We define two types of category representation over a set of supporting descriptors.

The first type of representation is a disjunction of supporting descriptors

$$D_1 \bigcup D_2 \bigcup .... \bigcup D_n. \qquad (1)$$

For example, the category “Taxes and Budget” can be represented with expression *TAX* $\bigcup$ *BUDGET SYSTEM*.

The other type of representation is a conjunction of disjunctions of supporting descriptors

$$(D_{11} \bigcup D_{12} \bigcup ... \bigcup D_{1n}) \ \& \ ... \ \& \ (D_{21} \bigcup D_{22} \bigcup ... \bigcup D_{2m}) \ \& \ ... \ \& \ (D_{k1} \bigcup D_{k2} \bigcup ... \bigcup D_{kr}). \qquad (2)$$

For example, category “Taxes and Budget of the Russian Federation” is represented with the following sequence of supporting descriptors: *(TAX* $\sqcup$ *BUDGET SYSTEM)* & *RUSSIAN FEDERATION.*

After relations between categories and supporting descriptors are fixed, categories corresponding to other descriptors of the Thesaurus are established automatically using the hierarchy of Thesaurus. As a result most descriptors of the Thesaurus are connected with some categories indicating the disjuncts it belongs to. A descriptor can have no category.

5.2. Text Categorization Using Different Systems of Categories

Text categorization of official documents of the Russian Federation is fulfilled for Information System RUSSIA (Yudina & Dorsey 1995). The system of categories consists of 180 categories that are connected with 210 supporting descriptors of the Thesaurus. Categories are represented as disjunctions of supporting descriptors. (Loukachvitch, 1997).

Text categorization for news reports uses 35 categories that are connected with 145 supporting descriptors of the Thesaurus. Most categories are represented as conjunctions of two disjunctions of supporting descriptors.

To provide convenient access to Russian official documents via the Internet for users accustomed to one of well-known thesauri (LIV 1990; UNBIS THESAURUS 1976), we took top categories (top terms, subject headings) from these thesauri and created relations between the categories and our Thesaurus. Every such thesaurus has a systematic part describing correspondence between its descriptors and top categories. Thus these systematic parts determine interpretation of each top category. For example, Legislative Indexing Vocabulary (LIV 1990) has 89 top terms that were connected with 250 supporting descriptors of our Thesaurus. In particular, top term “Medicine” containing 400 descriptors in LIV was connected with 7 supporting descriptors and currently 460 descriptors of our Thesaurus correspond to this top term.

6. TREC-6 Routing Task

We assumed that after matching a text with thesaurus units the remainder of our technique is language-independent. Thus to process TREC text collections we had to perform the following tasks:

- supplement the English translations of Thesaurus terms with synonymic expressions (size of Russian synonymic rows reach 20 and more elements);
- create morphological analyzer of English words;
- describe ambiguity of English terms by means of our Thesaurus;
- represent topics of TREC-6 as logical expressions of supporting descriptors.

6.1. Description of TREC-6 Topics

TREC-6 routing task carried out by Center for Information Research was close to the general strategy of CIR for automated text processing.

We used manually query construction where TREC-6 topics were represented as categories for text categorization.

Each topic was described as logical expression:

$$\bigsqcup X_i \; = \bigsqcup \; (\; \& \; x_{ij}) \;\; .$$

For each operand x_{ij} some supporting descriptors from the Thesaurus were chosen. After that the query was expanded by narrowed descriptors from Thesaurus.

Finally

$$x_{ij} = \bigsqcup w_{ijk} \quad ,$$

where w_{ijk} descriptors from Thesaurus.

For example the query for Topic 012 "Water Pollution - document is about the pollution of a body of water" was defined as:

$$X_1 \bigcup X_2$$

$$X_1 = x_{11};\ \ x_{11} = \mathrm{A};\quad X_2 = (x_{21}\ \&\ x_{22});\ \ x_{21} = \mathrm{B};\ \ x_{22} = \mathrm{C}$$

Figure 4 gives the detailed description of TREC-6 topic 012.

	x_{ij}	w_{ijk}
012	A “water pollution”	federal water pollution control act; federal water pollution control administration; hot water pollution; sewage disposal pollution of sea environment; sewage water pollution; water purification water supply and pollution control division
012	B “pollution”	ground pollution; oil distribution supertanker shipwreck oil pollution; oil spill
012	C “body of water”	body of water; animalis aquaticus; basin; fresh water; freshwater fishing; freshwater aquaculture; inland waterways freshwater reservoir; maritime fishery; lake ocean; ocean resources; reservoir; river; salt water; sea; sea animal; sea fish sea flora; sea mammal; sea-water; water basin sources of water; surface waters; water biological resources; water plant water resources; water scoop; water supply water-way; watershed

Figure 4. Topic 012 description

6.2. Processing Documents

We created an English morphological analyzer using standard morphological rules and WordNet exception lists. A morphological representation was built for every English entry of the Thesaurus.

During processing of a document we calculated the weights of any topics that were found.

The general rule was

$$\mu_D = \max_i(\ \mu_X(X_i)\ )\ \ ,$$

where weight of operand group is:

$$\mu_X(X_i) = \Pi_j\ \mu_x(x_{ij}) = \mu_x(x_{i1})\cdot\ \mu_x(x_{i2})\cdot....\ \cdot\mu_x(x_{im})\quad ,$$

weight of operand calculated as:

$$\mu_a(x_{ij}) = max\{\mu_0 , \nu_T(w_{ijk})\} ,$$

here $\mu_0 = 0.001$,

$$\nu_T(a_{ij}) = \begin{cases} 1.00, & \text{if } a_{ij} \text{ represents the main descriptor of main thematic node,} \\ 0.60, & \text{if } a_{ij} \text{ represents a descriptor of main thematic node,} \\ 0.30, & \text{if } a_{ij} \text{ represents the main descriptor of specific thematic node,} \\ 0.10, & \text{if } a_{ij} \text{ represents a descriptor of specific thematic node,} \\ 0.05, & \text{if } a_{ij} \text{ represents a mentioned descriptor,} \\ 0.00, & \text{otherwise.} \end{cases}$$

7. Analysis of Results

Our TREC6- routing results are close to median of the Category A routing results thus confirming the basic principles of our technology.

During our TREC-6 processing we encountered the following problems:

- ambiguity of English terms considerably differs from ambiguity in Russian and its description requires additional information;
- some subunits of TREC-6 topics could not be expressed by means of our Thesaurus.

The Thesaurus is to be further developed and carefully horned and tested in order to obtain better results using our technology of conceptual indexing for English texts.

Bibliography

Brazilay R., Elhadad M. 1997. Using Lexical Chains for Text Summarization. - ACL/EACL Workshop Intelligent Scalable Text Summarization.- Madrid.

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

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

Halliday M., Hasan R. 1976. Cohesion in English. Logman, London.

Hirst G., St-Onge D. 1997. Lexical Chains as representation of context for the detection and correction malapropisms. In C. Fellbaum, editor, WordNet: An electronic lexical database and some of its applications. Cambrige, MA: The MIT Press.

LIV 1990. Legislative Indexing Vocabulary 19th Edition. - Washington: The Library of Congress.

Loukachevitch N. 1997. Knowledge Representation for Multilingual Text Categorization . AAAI Symposium on Cross-Language Text and Speech Retrieval, AAAI Technical Report, 1997, pp. 133–142.

Lukashevich N. 1995. Automated Formation of an Information-Retrieval Thesaurus on the Contemporary Sociopolitical Life of Russia. *Automatic documentation and mathematical linguistics*. 29(2): 29-35.

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

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

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

Subject Headings 1991. Subject Headings. 14th Edition. - Cataloging Distribution Service, Library of Congress, Washington, D.C.

Yudina T., Dorsey P. 1995. IS RUSSIA: An Artificial Intelligence-Based Document Retrieval System. *Oracle Select*. 2(2), 12-17.