Knowledge Representation for Multilingual Text Categorization

Loukachevitch Natalia V.

Institute of the USA and Canada Studies
Khlebny per. 2/3
Moscow, Russia, 121814
dobrov@fly.triniti.troitsk.ru

Abstract

The described approach to text categorization is based on thematic representation of a text. Thematic representation includes nodes of thematically related terms simulating topics of the text and is provided with classes of their importance for the text. Thematic representation is created on the basis of detailed description of the domain and allows to process different types of texts, to use different systems of categories (in various languages) for text categorization, to adapt quickly the system to other formats and types of texts and/or other systems of categories, to categorize texts using several systems of categories simultaneously. The most part of the algorithm is not language-dependent.

Introduction

Text categorization is an important task in networks. There are a lot of everyday information from multilingual sources that ought to be thematically divided to satisfy various users needs.

Today there are two primary approaches to text categorization: knowledge engineering approach and machine learning approach.

A variety of existing machine learning approaches have been tested in text categorization (Goldberg 1996; Lewis & Rinquette 1992). They allow to construct text categorizers automatically by means of inductive learning, using texts pre-categorized by humans as examples. The highest known performance of these systems is close to 74 % of recall and precision (Goldberg 1996).

The knowledge engineering approach obtains more efficient results. The performance of CONSTRUE (Hayes 1992) is evaluated as around 90%. Riloff and Lehnert (1994) report about a high precision approach reaching 100% precision with over 60 % recall. Higher efficiency is based on manual creation of knowledge bases, rules and dictionaries describing the domains (Goodman 1991; Vledutz-Stokolov 1987). This requires a considerable amount of human labor and development time. Once a system was created changes in types and formats of texts, modification of categories or substitution of a whole system of categories result in to significant additional labor and time costs. Several tools were constructed to diminish the problem such as instrumental tools (Hayes 1992) or automatic generator of dictionary AUTOSLOG (Riloff 1993).

In our approach we describe knowledge about a very broad domain as a model of the world without fixation of any systems of pre-defined categories. The knowledge base is represented as the Thesaurus. The Thesaurus was specially created as a tool for automatic processing of texts in a broad domain of sociopolitical life of Russia and is developed now as bilingual Russian-English Thesaurus. Various systems of categories (in Russian or English) can be flexibly attached to Thesaurus units.

Our technique of text categorization is based on constructing thematic representation of a text including recognition of terms, incorporation of thematically related terms into thematic nodes, determination of importance of topics represented by thematic nodes in the text. After terms recognition is carried out the technology is not language-dependent.

This technology allows to process different types of documents (Russian or English) such as official documents or news reports by information agencies, to use different systems of categories for text categorization, to adapt quickly the system to new types of documents and/or systems of categories.

1. Thesaurus

Creators of conventional thesauri (LIV 1994; UNBIS Thesaurus 1976; Subject Headings 1991) take into account domain, commonsense, and grammatical knowledge of indexers, and therefore thesauri created for manual indexing are hard to utilize in automatic indexing environment (Salton 1989) -- important terms of texts are not found, less important terms are revealed, some terms are identified incorrectly because of their ambiguity. We created our Thesaurus as a tool for automatic indexing (AI-Thesaurus) -- the Thesaurus on Contemporary Life in Russia. The Thesaurus contains wide scope of terms from general to very specific ones, has means for representation of ambiguous terms, comprises developed system of relations between terms.

The Thesaurus has been created in semi-automatic mode using automatic processing of more than 70 Mb of Russian official texts (Lukashevich 1995). This procedure consists of two main stages. At the first stage it automatically processes new texts and reveals new terms-like language expressions. Such language expressions are determined on basis of their syntactical and lexical structure. The special dictionary containing more 30 thousand words directs this process. Lexical control helps to avoid consideration of such expressions as *large volume, new approach, new way, better results* to be terms.

At the second stage our specialists manually choose terms from gathered terms-like expressions.

This everyday procedure adds constantly new terms to the Thesaurus. If the first megabytes of texts could give up to 1000 new terms per megabyte, now every megabyte of texts gives about 10 terms in average.

Carefully gathered terms form rows of quasi-synonyms (UF references) -- sometimes up to 20 elements. Adjectives and verbs that are derivatives of a descriptor can also be its quasi-synonyms.

Ambiguous terms can be described in two ways in the Thesaurus. The first -- 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 the Thesaurus are: broader term (BT) -- narrower term (NT), associative term (RT), whole-term (WT) -- part-term (PT).

Latter relationship is used for description of physical parts, elements or actants of a concept. For example, this relationship connects such descriptors as *AVIATION* and *AIRCRAFT, AGRICULTURE* and *FARMER*, and others.

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 "associative 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 AI-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 *FINANCE* has 13 direct relations with other descriptors, but in fact according to the properties of inheritance and transitivity it is related to more than 400 ones.

This extended set of related terms in AI-Thesaurus allows to determine which terms of a document are related to each other and to provide the disambiguation of terms during automatic indexing.

For example, description of concept *INSURANCE* is as follows:

INSURANCE

BT	*FINANCIAL ACTIVITY*
NT	*COINSURANCE*
NT	*PERSONAL INSURANCE*
NT	*PROPERTY INSURANCE*
NT	*RE-INSURANCE*
PT	*INSURANCE CONTRACT*
PT	*INSURANCE COVERAGE*
PT	*INSURANCE ORGANIZATION*
PT	*INSURANCE PREMIUM*
PT	*INSURANCE RISK*
PT	*INSURANCE TARIFF*
PT	*INSURANT*
PT	*INSURED RISK*
PT	*FRANCHISE*
RT	*INSURANCE LEGISLATION*
RT	*INSURANCE SUPERVISION*
RT	*INSURANCE MARKET*

Currently the Thesaurus contains more than 18 thousand terms and 7 thousand geographic names.

Russian descriptors and their quasi-synonyms were translated into English and formed English sub-system of the Thesaurus consisting of English descriptors and synonyms. Synonymic rows were supplemented with terms from the thesauri (LIV 1994; UNBIS THESAURUS 1976, Miller et al.. 1990). We plan to organize the procedure of processing English texts for enriching the synonymic rows.

English ambiguous terms are described in the Thesaurus in the same way as Russian ambiguity.

2 Relations between the Thesaurus and Categories

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

We consider any category as user defined query that has to be represented by descriptors of the Thesaurus. 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 some descriptors. Now we use two types of category representation by a set of supporting descriptors.

The first type of representation is disjunction of supporting descriptors

$$D_1 \cup D_2 \cup .... \cup D_n.$$

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

Other type of representation is a conjunction of disjunctions of supporting descriptors

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

For example, category "Taxes and Budget of the Russian Federation" is represented with the following sequence of supporting descriptors: *(TAX* $\cup$ *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 following algorithm:

Step 1. Verify if a given descriptor is a supporting descriptor. If it is then a corresponding category is found, else go to Step 2.

Step 2. Look through descriptors related to the given descriptor with relationships BT, WT, RT. If some of these descriptors are supporting ones then add corresponding categories to a list of categories of the given descriptor. If some descriptors are not supporting ones and related to the initial descriptor with relationships BT and WT they are added to a buffer for further search of categories.

Step 3. If the buffer is not empty every descriptor of the buffer is processed as at Step 2.

As a result most descriptors of the Thesaurus are connected with some categories indicating disjunction it belongs to. A descriptor can have no category.

The establishing of such flexible relationships between categories and descriptors of the Thesaurus allows to take into consideration specific features of documents and categories without changing thesaurus relationships. For example, in the Thesaurus descriptor *GOVERNMENT COMMISSION* is related to descriptor *GOVERNMENT.* But if it is known that all documents of a collection are decrees by government of the Russian Federation then descriptor *GOVERNMENT COMMISSION* has to correspond to category "Government of the Russian Federation". To obtain it we can do descriptor *GOVERNMENT COMMISSION* supporting descriptor of this category.

In order to reflect properly specific features of document collection and categories we add special "empty category" to any system of categories. We use it when thesaurus description of concepts is not appropriate for a given document collection. For example, USA is described as a foreign country from the point of view of our Thesaurus. But if we process such documents as treaties between the Russian Federation and the US then USA is a participant of any document. For this document collection category "Foreign country" has to correspond to any country except the Russian Federation and USA. In this case descriptor *FOREIGN COUNTRY* is a supporting descriptor for a category "Foreign country" but descriptors *USA* and *RUSSIAN FEDERATION* are supporting descriptors for "empty category".

To provide convenient access to Russian official documents (Yudina & Dorsey 1995) via 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 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 1994) 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 now 460 descriptors of our Thesaurus correspond to this top term.

3. Text Categorization Using Thematic Representation of Text

Text units are compared with terms of a 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.

After comparison with the Thesaurus the text is represented as a sequence of descriptors and the following steps of the algorithm are not language dependent. All quasi-synonyms of any descriptor are represented by the that descriptor and are not differentiated further.

Now it is necessary to determine what descriptors of the text are related to each other. We can do it using thesaurus relationships and properties of inheritance and transitivity. A set of text descriptors and relationships between them that are obtained using properties of thesaurus relationships is called "thesaurus projection".

Descriptors corresponding to different meanings of ambiguous terms also participate in construction of thesaurus projection for a text. Using thesaurus projection a proper meaning of an ambiguous term is chosen. Efficiency of term disambiguation is more than 75 percents of chosen correctly descriptors.

At the next stage it is necessary to identify topics of the text and describe them constructing thematic nodes.

Every topic discussed in a text is usually expressed with a set of related terms. For example, discussion of scientific problems can be expressed in a text by means of the following terms: *mathematics, physics, fundamental research, applied research, academic institute,* and so on. The term that characterizes the topic 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 main term of a topic. 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.

Thematic relations between terms in a text are represented by relationships between corresponding descriptors in the thesaurus projection. The thesaurus projection usually consists of some separate fragments. A fragment of the thesaurus projection can have a complex structure and contain descriptors that are not really thesaurus related to each other. Thus it is necessary to subdivide these fragments further in the thesaurus projection.

Our experiments show that for the most effective division of the thesaurus projection it is necessary to use the notion of "thematic node". A set of descriptors from a text that have thesaurus relationships with one and the same descriptor D0 in the thesaurus projection of the text is called "thematic node". Descriptor D0 is called "main descriptor" of this thematic node.

Let us see fragments of a thematic nodes with main descriptor *CUSTOMS FORMALITY* that were constructed during automatic processing of Customs Code of the Russian Federation (the right column represents descriptor frequency in the text).

CUSTOMS FORMALITY	520
CUSTOMS DUTY	165
CUSTOMS CONTROL	153
CUSTOMS DECLARATION	47
CUSTOMS BENEFITS	21
IMPORT TAX	12
EXPORT TAX	8

During automatic processing of Customs Code more than 140 thematic nodes were constructed. (Size of the document is more than 500 Kb).

At the next stage it is necessary to evaluate the importance of topics and thematic nodes representing these topics in the text. At first we have 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. It 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).

As a result we obtain a set of textual relations for every descriptor of a text. For example, here are fragments of a set of textual relations of descriptor CUSTOMS BORDER received during processing of Customs Code (on the right side frequency of textual relations is indicated):

CUSTOMS BORDER	
GOODS	8
MEANS OF TRANSPORT	5
CUSTOMS TERRITORY	3
FREE CUSTOMS ZONE	1
CUSTOMS DUTY	1

Textual relations between descriptors are determined at the stage of comparison of text with Thesaurus. After construction of thematic nodes textual relations frequencies of descriptors in each thematic node are summed up, and we receive textual relations between thematic nodes.

In our approach we assume that 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.

Evaluated in such a way main thematic nodes determine a threshold that distinguishes main thematic nodes among all other thematic nodes of a text. 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 our example of the thematic representation for the Customs Code main thematic nodes were thematic nodes with main descriptors *GOODS, CUSTOMS FORMALITY, CUSTOMS COMMITTEE, LAW.*

Besides 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.

A set of thematic nodes constructed for the text with evaluated status of these thematic nodes is called "thematic representation" of the text.

Thus all descriptors of the text are divided into five classes of different 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.

Division of descriptors into classes of importance is used for text categorization. A category represented as a disjunction of supporting descriptors became a category of the text if one of main descriptors of main thematic nodes correspond to this category. If a category is a conjunction of two disjunctions, a special function f(k1,k2,r) evaluates if this category is a category of the text, where k1 is the highest class of descriptors corresponding to the first disjunction of the category, k2 is the highest class of descriptors corresponding to the second disjunction and r is a frequency of textual relations between descriptors corresponding to different disjunctions of the category.

Conclusion

Now we have evaluation of performance of our system only for Russian texts.

Our evaluation of performance was obtained as a result of the following procedure. The system processed texts. We looked through categories obtained for each text and determined

- how many categories were obtained -- S_i;
- how many categories correspond to the contents of the text -- P_i;
- how many categories this text has in our opinion -- W_i .

Precision of the whole process was estimated as $\Sigma(P_i)/\Sigma(S_i)$, recall -- $\Sigma(P_i)/\Sigma(W_i)$.

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. Efficiency of text categorization -- 91.2% of precision and 94.2% of recall --was tested on 700 documents that were not used for construction of the Thesaurus

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. Evaluation of text categorization received as a result of analysing 1200 reports of IMA-PRESS information agency is as follows: precision 91.1%, recall - 93.8%.

We have shown that it is possible to provide effective text categorization of various text collections using description of domain in the Thesaurus created as a special tool for automatic text processing, and constructing thematic representations of texts. Categories are connected with the Thesaurus by flexible relationships. The system can be quickly adapted to other types of texts and other systems of categories in various languages; it can process texts using different systems of categories simultaneously.

We plan to develop text categorization of English texts using Russian categories and to provide access to Russian official documents using top categories of well-known thesauri (LIV 1994; UNBIS THESAURUS 1976).

Acknowledgments

This research was funded in part by John and Catherine MacArtur Foundation N. 95-30332A-FSU.

Bibliography

Goldberg J.L. 1996. A machine learning method for text categorization. Ph.D. diss., Texas A&M University.

Goodman M. 1991. Prism: A Case-Based Telex Classifier. In Proceedings of the Second Annual Conference on Innovative Applications of Artificial Intelligence. AAAI Press. 25-37.

Hayes Ph. 1992. Intelligent High-Volume Processing Using Shallow, Domain-Specific Techniques. Text-Based Intelligent Systems: Current Research and Practice in Information Extraction and Retrieval. New Jersey, P.227-242.

Lewis D.D. and Ringuette M. 1992. Text Categorization by Inductive Learning. Proceedings of AAAI-92.

LIV 1994. Legislative Indexing Vocabulary 21th Edition. - Washington: The Library of Congress.

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.

Riloff E.1993. Automatically Constructing a Dictionary for Information Extraction Tasks. Proceedings of the Eleventh National Conference on Artificial Intelligence. AAAI Press/ The MIT Press. 811-816.

Riloff E. and Lehnert W. 1994. Information Extraction as a Basis for High Precision Text Classification. *ACM Transactions on Information Systems,* 12(3):296-333.

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.

Vledutz-Stokolov N. 1987. Concept Recognition in an Automatic Text-Processing System for the Life Sciences. *J. of the American Society for Information sciences* 38, P.269-287.

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