Development of Bilingual Domain-Specific Ontology for Automatic Conceptual Indexing

Natalia V. Loukachevitch and Boris V. Dobrov

Research Computing Center of M.V.Lomonosov Moscow State University,
NCO Center for Information Research
339, Research Computing Center of M.V.Lomonosov Moscow State University,
Leninskie Gory, Moscow, 119992, Russia
{louk, dobroff}@mail.cir.ru

Abstract

In the paper we describe development, means of evaluation and applications of Russian–English Sociopolitical Thesaurus specially developed as a linguistic resource for automatic text processing applications. The Sociopolitical domain is not a domain of social research but a broad domain of social relations including economic, political, military, cultural, sports and other subdomains. The knowledge of this domain is necessary for automatic text processing of such important documents as official documents, legislative acts, newspaper articles.

1. Introduction

Any technique for cross-lingual information retrieval needs translation resources such as machine dictionaries, lexical knowledge bases, machine translation systems, aligned corpora (Gonzalo, 2001).

The first type of resources created for cross-lingual information retrieval were multilingual information retrieval thesauri. One example of such thesauri, thesaurus EuroVoc of European Community, is published on 9 languages of European Communities and nowadays used for retrieval of European documents (EUROVOC, 1995).

However such thesauri developed for manual indexing (monolingual and multilingual) have properties which make it impossible to use them in automatic text processing of contemporary large electronic collections (Salton, 1989). The goal in developing a conventional information retrieval thesaurus (for manual indexing) was to describe terms necessary for representation of main topics of documents. More specific terms were not included. Ambiguous terms were provided with scope notes and comments convenient for human subjects (LIV, 1994). Most relations were intended to serve for human navigation in such a thesaurus. Human subjects had to use their domain, common sense, and grammatical knowledge not described in a thesaurus in order to index documents. 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.

Since 1994 we develop Thesaurus on Sociopolitical Life as a special tool for automatic conceptual indexing and information retrieval. The domain of the thesaurus is a broad domain of social relations including economic, political, military, cultural, sports and other problems, which are discussed in governmental documents, legislative acts, newspaper articles. Now the Thesaurus includes more than 29 thousand concepts, 69 thousands terms, 112 thousand manually described relations. Since 1995 the Thesaurus is used in such information-retrieval tasks as automatic conceptual indexing, automatic text categorization and text summarization. The Thesaurus is a searching tool in University Information System RUSSIA (Russian Inter-University Social Sciences Information Consortium, UIS RUSSIA, www.cir.ru/eng/), containing more than 800 thousand documents (Loukachevitch & Dobrov, 2002).

In this paper we describe main stages of development of the bilingual Russian-English Thesaurus on Sociopolitical Life for automatic conceptual indexing of English and Russian documents and its current applications.

2. Specific Features of Sociopolitical domain

The domain of Sociopolitical Thesaurus is not domain of social research. It comprises situations and problems in social life of the contemporary society, which are discussed in official documents and newspapers. These problems are of great social significance, therefore there are corresponding words in the general lexicon and terms in professional terminologies of economy, law, defense, culture, sports and others. This domain can be considered as a transition area where concepts of the general lexicon and terminologies are intersected.

The domain inludes senses of general words practically coinciding with senses of terms of special terminologies (*arson, trolley-bus)* and senses of multiword terms understandable for native speakers (*internal migration, humanitarian aid).* Knowledge of the domain is necessary to organize a qualitative automatic processing of such important texts as governmental regulations, laws, international treaties, news reports and others.

At present the Sociopolitical Thesaurus includes also more and more terminologies of such non-manufacturing sectors as banking, accounting, taxes, customs. It does not include terms of manufacturing industries and specific sciences.

3. Structure of the Thesaurus

The Sociopolitical thesaurus (below the Thesaurus) is a hierarchical net of concepts. We consider it as a kind of a linguistic ontology. Concepts of the Thesaurus originate

from senses of language expressions, that is single words or multiword expressions.

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

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

A concept of Thesaurus has relations with other concepts. The main types of relations are taxonomic relations and specific set of conceptual relations based on ontological dependence relations. This set of relations was experimentally confirmed to be effective in information-retrieval applications (Loukachevitch & Dobrov, 2002).

4. Development of Bilingual Thesaurus for Automatic Text Processing

Development of a bilingual thesaurus intended for automatic text processing has the following specific features. It is necessary:

- to describe the most exact language variants of a concept in both languages. Such a bilingual resource has to be symmetric in distinction to conventional bilingual dictionaries, which can give a broader or narrower word as a translation variant. It often happens that a single-word term of one language corresponds to a multiword term in other language. Then it is necessary to search and describe such multiword terms and its synonymic variants. For example, Russian word 'mundir' is translated in bilingual dictionaries as "uniform", "coat" or "tunic", but more correct variants are "uniform coat", "uniform jacket" or "uniform tunic";
- to describe large synonymic sets for every concept in all languages;
- to describe as much multiword variants for a concept as possible as a basis for lexical disambiguation. Now Internet gives excellent possibilities to find such terms and check their real usage.

The development of the English part of the Thesaurus included four main stages. At the first stage we collected English translation for the Russian terms of the Thesaurus from bilingual dictionaries. Here we received 33 thousand English variants.

At the second stage we worked with American and English explanatory dictionaries, traditional information-retrieval thesauri, terminological dictionaries. We went through them entry by entry and searched for additional English variants for existing concepts and additional concepts that were missed in the Thesaurus or do not exist in Russian. During this stage more than 22 thousand words and expressions were added to the conceptual net of the Thesaurus.

The next stage was devoted to checking of translations in the English part of the Thesaurus and enrichment of English synonymic rows. We checked if found translations are really used in English texts, what senses of their contemporary usage are.

We found that a lot of words taken from well-known English and American dictionaries, from the best English-Russian dictionaries are not really used in contemporary texts. Usually we consider frequency 100-150 usages in Internet texts in English and American sites as a necessary minimum for inclusion or preserving of an expression in the Thesaurus.

For example, the following expression taken from English-Russian dictionaries (Multilex, 1996) were deleted because of their practical absence in real native English texts: *neo-mortality* (frequency – 19), *multiathlon* (frequency -100, means combined events), *narcologis*t (frequency – 500 in Russian sites and sites of Post-Soviet countries), *narcology (*4000 pages in Russian sites and sites of Post-Soviet countries) , *interindustrial balance* (frequency 2), *betterment of land* (frequency –53). Frequency is received from Google searches. The following expressions (and many others) taken from Unabridged Webster (1999) were deleted: *nunship (54), scattersite housing* (7), *mosquitoey* (127, also was in Merriam Webster), *monigamousness* (77), *mysticalness (125), ultramicrofiche* (60).

It often happened that English variants described in bilingual sources were not real, but we understood that translation equivalents had to exist, and we tried to find them. For example, for Russian word "motoblok" real translations are "*garden tractor*" or "*lawn tractor*" but not "motor block" as it was indicated in bilingual dictionaries.

To provide identification of a thesaurus concept in texts we try to collect various forms of its text expression, especially various multiword expressions. An English expression can have a mark indicating its origin. For example, a concept *EQUALITY BETWEEN MEN AND WOMEN* has the following synonymic expressions:

equal rights for women (WordNet's gloss)
equal rights of men and women (EuroVoc)
*equality between sexes (*Multilex*)*
*equality between women and men (*texts – documents of Council of Europe)
gender equality (texts)
sex equality (texts*).*

Now the English part of the Thesaurus includes more than 65,000 English terms.

5. Text categorization as an Evaluation Technique

An important step for evaluation of a created resource is its use in text processing applications. At the fourth stage of the development of the Thesaurus we test it in

automatic text categorization of English and Russian documents.

We have developed a thesaurus-based technique for automatic text categorization (Loukachevitch & Dobrov, 2003). The technique is based on the following principles:

- categories are connected with a relatively small number of 'supporting' concepts of the Thesaurus. Categories of other terms are established on the basis of properties of the Thesaurus relations. It became possible due to detailed presentation of various aspects of described concepts and careful testing of the Thesaurus relations;
- the possibility of processing texts of various types and sizes is based on thematic representation of text contents, where the terms of a text are divided to thematic nodes, simulating elements of the main theme and the subthemes of a text (Loukachevitch N., Dobrov B. 2000). Construction of the thematic representation is based on such a property of texts as lexical cohesion.

Using this technique we have implemented more than ten automatic categorization systems of Russian and English documents with the number of categories up to 3000. A special visualization tool of the categorization systems allows us:

- to look through all terms recognized in a text and search for missed text entries for thesaurus concepts. For example, analyzing results of text processing of "Declaration on the Rights of Persons Belonging to National or Ethnic, Religious and Linguistic Minorities" adopted by UN General Assembly we could add the following new text entries to the existing concepts: *collaboration among nations* for concept *INTERNATIONAL COLLABORATION*, *development of society* for *SOCIAL DEVELOPMENT*, *right of persons* for *HUMAN RIGHTS*.
- to see all ambiguous terms of a text and results of their disambiguation,
- to see all categories assigned to a text automatically and a terminological basis for every category.

For example, for category "Legal system" a terminological basis of concepts in the UN Declaration is as follows: *LAW, LEGISLATION, CRIME, LEGAL NORM, RULE OF LAW, LAWMAKING*.

An expert analyzes results of text categorization and especially texts that were categorized unsuccessfully and can easily to find problems in description of terms in the Thesaurus, to find new useful expressions to include to the Thesaurus.

Evaluation of the Thesaurus is carried out in real applications. Now bilingual economic terminology is tested in text categorization on JEL (Journal of Economic Literature) subject headings – we develop a tool for assistance to authors to categorize their papers in economics in the SocioNet project.

Legal terminology verification is based on the hierarchical system of 1168 subject headings, adopted by the Russian president's decree. So we prepare the Thesaurus to serve as a tool for automatic processing of English documents and Russian queries to provide better access in Russian to materials of European Court for Human Rights.

6. Bilingual Thesaurus as an Information Retrieval Tool

In University Information System RUSSIA the first version of thesaurus-based bilingual retrieval is implemented. Several collections of English documents:

- RePEc (Research Papers in Economics, www.repec.org) abstracts and full papers,
- test collection of Council of Europe documents,

were automatically processed to be loaded to the system.

Every (English or Russian) text can be searched using formal characteristics of a document or a word-based retrieval model. At the same time a text is automatically provided with a language independent conceptual index. This process includes the following stages:

- the matching of text substrings with the thesaurus terms on a morphological basis and identification of a corresponding thesaurus concept,
- in case of ambiguity of a term, which can be included in synonymic rows of several different concepts, automatic term disambiguation is initialized. The analyzer compares the context of a term in a text and thesaurus neighborhoods of the concepts-senses (Loukachevitch and Dobrov, 2000). A concept, that has the coincidence between a text and the thesaurus neighborhood in the most minimal text distance, is chosen.

After these stages for Russian or English texts the language-independent index of the Thesaurus concepts is generated. Therefore thesaurus-based retrieval in our system is independent of a language used in a query and in a text, and a retrieval set can contain texts in both languages (Fig.1).

The right column of the screen shows concepts specific for the retrieval set. A user can modify the query, add or delete the concepts of the right column from the query using only one key. Names of these concepts can be also formulated in both languages. Therefore a user can refine a query using his/her native language, and only after this refinement stage a user has to begin reading or translation of texts in another language.

Conclusion

In the paper we have described specific features of Russian-English Sociopolitical thesaurus intended for automatic text processing of Russian and English texts. To develop the Thesaurus we studied well-known American and British dictionaries and thesauri, bilingual dictionaries, checked usages of words and multiword expressing through Internet search. Now the Thesaurus is used and evaluated in several applications of bilingual information retrieval.

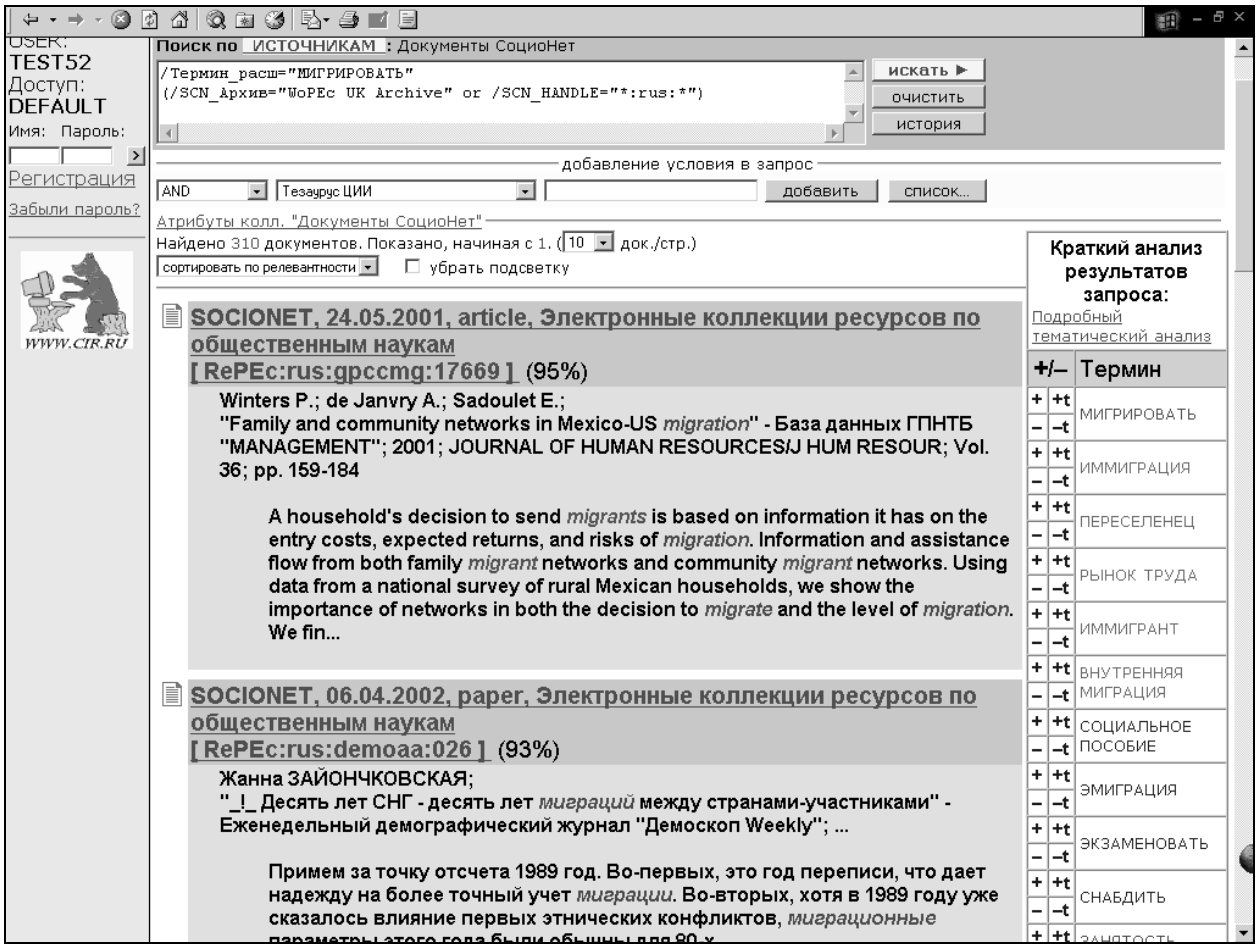

Figure 1

Acknowledgements

Partial support for this work is provided by the Russian Foundation for Basic Research through grant # 03-01-00472.

References

Thesaurus EUROVOC: Vol 1-3 / European Communities. – Luxembourg: Office for Official Publications of the European Communities, 1995. – Ed.3. – English Language.

Gonzalo, J. (2001). Language Resources in Cross-Language Information Retrieval: a CLEF perspective. - Cross-Language Information Retrieval and Evaluation: Proceedings of the First Cross-Language Evaluation Forum, LNCS, Springer-Verlag.

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

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

Loukachevitch N. & Dobrov, B. (2002). Evaluation of Thesaurus on Sociopolitical Life as Information-Retrieval Tool. In M.Gonzalez Rodriguez, C. Paz Suarez Araujo (Eds.) The Third International conference on Linguistic Resources and Evaluation (LREC-2002). – Vol.1 – 2002, Gran Canaria, Spain – p.115-121.

Loukachevitch, N.V. & Dobrov, B.V. (2003). Knowledge-Based Text Categorization of Legislative Documents. In F.Kiefer, J.Pajzs (Eds.) Proceedings of 7th Conference on Computational Lexicography and Text Research (COMPLEX 2003) – Research Institute for Linguistics, Hungarian Academy of Sciences, Budapest, 2003. – pp.57-66.

Multilex. (1996). *Multilex 1.0a. Anglo-russkiy elektronniy slovar*. Medialingua Ltd.

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

Webster. (1999). *Random House Webster's Unabridged Dictionary*. Version 3.0. Random House, Inc.