Cross-Language Information Retrieval Based on Multilingual Thesauri Specially Created for Automatic Text Processing

Natalia V. Loukachevitch, Boris V. Dobrov
{louk, dobroff}@mail.cir.ru
Research Computing Center of Moscow State University
Vorobyevy Gory, Moscow, 119899, Russia

Abstract.
In the paper we discuss necessity and possibility to develop special linguistic resources (monolingual, bilingual, multilingual), specially intended to be used in automatic processing of large text collections for information retrieval applications.

1. Introduction

Any techniques 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.

It is known that in monolingual information retrieval, the use of linguistic resources did not show improvements in retrieval performance that would be considerable enough to justify development costs in comparison to word-based models (Voorhees, 1999).

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

2. Development of Monolingual Thesaurus for Automatic Text Processing

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 are intended to serve for human navigation in such a thesaurus. 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 1994 we began development of Thesaurus on Sociopolitical life for automatic text processing .

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 27 thousand concepts, 64 thousands terms, 105 thousand manually described relations. To compare, conventional information-retrieval thesauri for the same domain has the following quantitative characteristics: Legislative Indexing Vocabulary – 6800 descriptors (concepts), 9800 terms, about 15 thousand relations between descriptors (LIV, 1994), English part of EuroVoc has 5933 descriptors, about 17 thousand relations between descriptors (EuroVoc, 1995).

Since 1996 Thesaurus on sociopolitical life is used in automatic processing applications. 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 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.

The Thesaurus is used as a basis for 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 hierarchical categorization systems:

- Subject Headings of Central Election Committee of the Russian Federation (450 categories, 3 levels of hierarchy),
- Subject Headings of Legislative Acts of the Russian Federation – more than 1000 categories, 4 levels of hierarchy).

In (Loukachevitch and Dobrov, 2002) we described an experiment that showed that performance of thesaurus-based information retrieval for short queries can achieve considerable improvement of the retrieval in comparison to a statistical retrieval model (Callan et.al., 1992).

Therefore we think that development of multilingual thesauri based on the same principles can lead to considerable improvement of performance in cross-lingual information retrieval.

3. Development of Multilingual Thesaurus for Automatic Text Processing

Development of a bilingual or multilingual thesaurus intended for automatic text processing also has its specific features. It is necessary:

- to describe the most exact language variants of a concept in different languages. Such a bilingual resource has to be symmetric in distinction tof 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, for example, Russian word ‘dissident’ corresponds to English “political dissident”. Then it is necessary to search and describe such multiword terms and its synonymic variants;
- to describe large synonymic sets for every concept in all languages;
- to describe as much multiword variants of a concept as possible as a basis for lexical disambiguation. Now Internet gives excellent possibilities to find such terms and check their real usage.

Fig.1.

We began development of bilingual Russian-English thesaurus on Sociopolitical life. Now English part of the thesaurus includes 55 thousand terms. 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,
- test collection of Council of Europe documents,

were automatically processed to be loaded to the system. We plan to begin evaluation of thesaurus-based bilingual retrieval next year.

Every (English or Russian) text can be searched (Fig.1) using formal charachteristics of a document or a word-based retrieval model. At the same time a text is automatically provided with a language independent conceptual index. Therefore thesaurus-based retrieval in our system is independent of language used in a query and in a text, and a retrieval set can contain texts in both languages.

4. Multilingual Thesauri and Visualization of Text Contents

In multilingual information retrieval there is a serious problem of how users can estimate the relevance of retrieved documents and how they can choose the most relevant documents for computer or human translation.

A multilingual thesaurus created for automatic text processing allows construction of a structural thematic summary of a text (Loukachevitch and Dobrov, 2000). The structural thematic summary describes the main theme and subthemes of a document, which are simulated by sets of semantically related terms – thematic nodes. It allows users to estimate the contents of a document at first sight. The thematic summary can be presented in any language of a thesaurus using corresponding language equivalents.

Clarity and readability of the structural thematic summary is provided by:

- the automatic knowledge-based clustering of semantically related terms and knowledge-based evaluation of their significance for the text contents;
- translation of multiword terms, which is easier than translation of a single word or a sentence of a text;
- automatic disambiguation of ambiguous terms.

For example, let us suppose that a user's query was "*Border troops*" for a Russian collection and a text with the following structural thematic summary was received:

```
                ============================================================
****    |  TERRITORIAL WATERS; WATER TRANSPORT; STATE; OCEAN; ISLAND
        |
        |        ================================================
****    |  X     |  BORDER TROOPS; STATE; FRONTIER GUARD
        |        |
        |        |        ========================================
****    |  X     |  X     |  JAPAN; STATE; FOREIGN COUNTRY;
        |        |        |  MINISTRY OF FOREIGN AFFAIRS
        |        |        |        ================================
****    |  X     |  X     |  X     |  FISH; FISHING; NATURAL RESOURCES; FISHERMAN;
        |        |        |        |  FISHING VESSEL; FISH RESOURCES; ILLEGAL FISHING
        |        |        |        |        ================================
****    |  X     |  X     |  z     |  z     |  RUSSIAN FEDERATION; STATE; FAR EAST;
        |        |        |        |        |  CURILE; PRESIDENT OF RUSSIA
        |        |        |        |        |        ============================
****    |  X     |  z     |  z     |  z     |  z     |  POACHING; OFFENCE; ILLEGAL FISHING;
```

A structural thematic summary contains the following parts:

- the terms of the main thematic nodes ordered by frequency and situated horizontally;
- the marks of relative frequencies of main thematic nodes denoted by different numbers of '*' :
  - '****' - node frequency is more than 75% of maximum node frequency,
  - '***' - node frequency is more than 50 % of maximum;
  - '**' - node frequency is more than 25 % of maximum,
  - '*' - in other cases.
- marks of strength of mutual coocurrrence between different thematic nodes:
  - -- 'X' - very frequent cooccurrence (frequency of cooccurrence between nodes is more than 75% of maximum frequency of cooccurrence);
  - -- 'z' - frequent cooccurrence (frequency is more than 25% of maximum cooccurrence);
  - -- '.' - rare cooccurrence.

The structural summary allows us to know that the text is devoted to illegal fishing in Russian territorial waters in the Far East.

5. Proposal for CLIR Research

We think that it is important to develop domain-specific monolingual and multilingual linguistic resources specially constructed for automatic text processing of large text collections.

The most important domain is a broad sociopolitical domain – its knowledge allows processing of texts of large social significance: governmental acts, international treaties, legislation, newspaper articles, news reports. A sociopolitical thesaurus can serve as an initial source of general terminology for developing thesauri in more specific domains.

From our experience we know that development of a sociopolitical thesaurus of 20 thousand terms plus geographical subthesaurus of 5-7 thousand geographic names is enough as a basis for the first knowledge-based experiments in various information retrieval applications. Having large electronic collections of documents and using contemporary techniques of term extraction it is possible to develop such sociopolitical thesauri in multilingual environment.

Bibliography

(Callan et.al., 1992) 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.

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

(Gonzalo, 2001) 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) LIV, 1994, Legislative Indexing Vocabulary. Congressional Research Service. The Library of Congress. Twenty-first Edition.

(Loukachevitch and Dobrov, 2002) Loukachevitch N., Dobrov B., 2000, Thesaurus-Based Structural Thematic Summary in Multilingual Information Systems. – Machine Translation Review, N 11, December 2000, p. 10-20.

(Loukachevitch and Dobrov, 2002) Loukachevitch, N., Dobrov., B., 2002, Evaluation of Thesaurus on Sociopolitical Life as Information-Retrieval Tool // Proceedings of Third International Conference on Language Resources and Evaluation (LREC2002) / M. Gonzales Rodriguez, C. Paz Saurez Araujo (Eds.) – Vol.1 – Gran Canaria, Spain – p.115-121.

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

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