Large-Scale Linguistic Ontology as a Basis for Text Categorization of Legislative Documents

Natalia Loukachevitch [a,b], Boris Dobrov [a,b]
[a] *Research Computing Center of Moscow State University (NIVC MSU), Russia*
[b] *NCO Center for Information Research, Moscow, Russia*

Abstract. The paper describes the structure and properties of a large linguistic ontology – a new kind of information retrieval thesaurus - Thesaurus on Sociopolitical Life for Conceptual Indexing. The thesaurus is used in various real-scale information-retrieval applications in the legal domain. At present one of the main applications of the Thesaurus is knowledge-based text categorization. Categories are connected with the Thesaurus by flexible relationships. The categorization system can process text collections containing texts different in sizes and types.

Keywords. Linguistic Ontology, Information Retrieval Thesaurus, Text Categorization, Legal Documents.

Governmental and legislative bodies of various countries often need their documents to be classified according to large and complicated hierarchical systems of subject headings. As a rule, collections of legislative documents are very large, the documents differ greatly in sizes and styles. Manual classification of such collections is hard, time-consuming and tends to subjectivity, which can increase because of lack of qualified specialists. Therefore use of automatic methods of text categorization in the domain is very important.

Use of machine-learning techniques for text categorization of legal documents is hampered by such factors as serious inconsistency of training data, lack of enough training documents for each subject heading, poor explanation means of existing machine learning approaches.

On the other hand knowledge-based methods [1] require huge volumes of knowledge described, because normative documents regulate relations in various domains: state policy, governmental bodies, taxes, accounting, banking, industrial and agricultural production, housing and social welfare, nature protection, science, education, defense and many others. Terms from all these domains can be essential for the document processing.

In fact to process normative documents it is necessary to have a knowledge resource comprising “descriptions” of different situations and problems of the contemporary society life.

We call this polythematic domain “sociopolitical domain” and for more than ten years develop a linguistic resource called Thesaurus on Sociopolitical Life for Conceptual Indexing (or Sociopolitical thesaurus) [2].

Now the thesaurus is a hierarchical net of concepts comprising more than 32 thousand concepts, 79 thousand Russian terms, 80 thousand English terms. In construction of the thesaurus we combined three different methodologies:

- The methods of construction of **information-retrieval thesauri** (information-retrieval context, analysis of terminology, terminology-based concepts, a small set of relation types)
- The development of **wordnets** for various languages (word-based concepts, detailed sets of synonyms, description of ambiguous text expressions)
- Ontology and **formal ontology** research (strictness of relations description, necessity of many-step inference).

These features allow us consider our thesaurus as a **linguistic ontology**.

For development of a text-categorization system the categories are described as Boolean formulas using a relatively small number of 'supporting' concepts of the Sociopolitical thesaurus. The formulas can be expanded on the basis of properties of the thesaurus relations.

Possibility to process texts of various types and size is based on the thematic representation of the 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. Construction of the thematic representation is based on such a property of texts as lexical cohesion.

Concept-based automatic categorization gives opportunity to explain received results to legal experts.

In the legal domain we have developed two text-categorization systems: for governmental bodies and for a commercial organization.

The corporate subject headings system included more than 3000 subject headings. As training materials they provided more than 250 thousand legal documents (federal, regional, subordinate documents, court practice) with manually established subject headings.

Performance results of the text categorization system were as follows: 75% recall, 20% precision. During our work we revealed that these performance figures estimate not only the quality of the automatic categorization but also the quality of the manual categorization, its inconsistency and low recall. The concept-based approach to representation of categories allowed us to demonstrate these problems to experts of the company who began to seek ways to improve the situation.

References

[1] Ph. Hayes, Intelligent High-volume processing using shallow, domain-specific techniques, *In: Text-Based Intelligent Systems: Current Research and Practice in Information Extraction and Retrieval*. New Jersey, pp.227-242, 1992.

[2] N. Loukachevitch, B. Dobrov, Evaluation of thesaurus on sociopolitical life as information retrieval tool, *In: "Proceedings of Third International Conference on Language Resources and Evaluation (LREC2002)"*, M.Gonzalez Rodriguez, C. Paz Suarez Araujo, eds. – Vol.1 – Gran Canaria, Spain, pp.115—121, 2002.