Sociopolitical Domain As a Bridge from General Words to Terms of Specific Domains

Natalia Loukachevitch and Boris Dobrov

Research Computing Center of Moscow State University,
Leninskie Gory, Moscow, 119992, Russia
Email: louk@mail.cir.ru, dobroff@mail.cir.ru

Abstract. In the paper we argue that there exists a polythematic domain which is situated in an intermediate area between senses of a general language area and specific domains. The concepts of this domain can be naturally added to general wordnets together with publicly known technical terms. Such enhanced wordnets can provide much more considerable preliminary coverage of domain specific texts, improve efficiency of word sense disambiguation procedures.

1 Introduction

Majority of the texts in electronic collections contain as general words as terms from specific domains. To effectively organize automatic text processing, knowledge resources have to include descriptions of both types of language expressions. However for years general words and domain terms were studied by different research communities. Lexicology and lexicography studied meanings of general words, technical terms were considered by terminologists in the general theory of terminology. Wuster wrote that the main difference in consideration of general words by lexicologists and terms by terminologists was as follows: terminologists begin consideration from a concept, but lexicologists from a form of a linguistic expression [15]. He wrote that terminological research starts from the concept which has to be precisely delimited and that in terminology concepts are considered to be independent from their designations. This explains the fact that terminologists talk about 'concepts' while linguists talk about 'word meanings'.

But now when linguists began to develop wordnets for various languages, the situation is changing. Creating wordnets linguists construct hierarchical semantic networks, try to find similar "synsets" for different languages, build the top ontology of language-independent concepts [2]. These directions of lexical research are much closer to the study of concepts, therefore the distinction between approaches seems to be considerably less serious.

Recently researchers began development of wordnets for specific domains [1,14]. From this point of view it is very important to understand how a general wordnet and domain specific wordnets interact with each other, how development of domain specific wordnets correlates with terminology research, if it is possible to combine lexical and terminological knowledge in the same linguistic resource.

In this paper we argue that there exists a polythematic domain which is situated in an intermediate area between senses of general language and concepts of specific domains and partially intersects with both ones. The concepts of this domain can be naturally added to

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 163–168.

general wordnets together with the most known technical terms. Such enhanced wordnets can provide much more considerable preliminary coverage of domain-specific texts, to serve as a reliable source for development of domain-specific ontologies.

2 Features of Terms

There are a lot of definitions of a term given by terminologists. Most of them consider a term as a word or expression designating a concept in a special domain. A specific feature of a term is that its relations with other terms of the domain is described by a definition [11].

The whole set of terms of a domain is comprised by the terminology of the domain. This system of terms during the process of its development usually undergoes procedures of standardization and normalization to be understandable for all specialists in the domain.

For choice of appropriate terms in the standardization process terminologists consider the following features of an ideal term [12]:

- the term must relate directly to the concept. It must express the concept clearly;
- there should be no synonyms where absolute, relative or apparent;
- the contents of terms should be precise and not overlap in meaning with other terms;
- the meaning of the term should be independent of context.

According to [5] “the objective of term-concept assignment in a given special language is to ensure that a given term is assigned to only one concept is represented by only one term”.

This means that in ideal cases there must be a biunivocal relationship between concepts and terms in each special field of knowledge. For a terminology nothing could be better than that: no synonymy, no homonymy and no polysemy.

Though this ideal situation only happens in a few well structured fields and does not happen for the rest, this terminologists’ point of view stresses how considerable is difference between a term and a word of general language. However, in reality the gap a word – a term is not so broad.

3 Term Formation and Words of General Language

An impregnable barrier between words of a general language and terminologies does not exist. A lot of terms (for example, terms in technical domains) appeared in specific domains become elements of a general language. On the other hand a general language word can change its meaning and become an element of a terminology.

Among possible transitions from a general language to a terminology it is important to distinguish the following cases:

1. a sense of a general word and a sense of the same wordform as a technical term are really different. For example, a new sense of a term can result from metaphoric shift or domain specification of a general sense. So there is a general sense of word “function”, there is term “function” in biology, there is term “function” in mathematics and so on. A usage of word “function” can never have all or several of these meanings, that is an important rule of distinguishing of different senses of a word fulfills: “ senses of a lexical form are antagonistic to one another; that is to say, they can not be brought into play simultaneously without oddness” [3].

2. a sense of a term in a domain-specific terminology is only slightly refined in comparison to a sense of the same word as a general language expression. Let us consider several terms from criminal law that also exist as words of general language. In this cases dictionaries often use terminological definitions as glosses such as ***arson**-Law. the malicious burning of another's house or property, or in some statutes, the burning of one's own house or property, as to collect insurance [10].*

If one supposes that there exist two senses of such legislative terms as *arson, murder* or *bail*, then one have to agree that for too many usages it is impossible to distinguish the general usage of a word from the terminological use, especially in media texts. So news reports can be understood by ordinary people and at the same time such texts can contain a lot of domain-specific terms.

In such situations we should not distinguish two senses of such words. In fact, the same sense "works" in a general language and in domain-specific language.

One can argue that terminological definitions delimit domain concepts stricter than definitions of explanatory dictionaries. Indeed, the borders of a general language sense can be very vague. Using a general language word we distinguish typical cases and can mistake or doubt in complicated cases (as previously one could think that a whale is a fish). A terminology tries to provide a concept with more definite boundaries, for example, legislators use a page long definition to distinguish "new construction" from "repair" for taxation needs. However we think that if there is an agreement in typical cases the problem vague vs. strict boundaries of a sense is not a reason to separate senses. We suppose that people do not think about concept boundaries because of lack of necessity. If necessary they readily use domain definitions as a support. So for the most known legislative terms general dictionaries use law definitions.

4 Notion of Sociopolitical Domain

It is important to understand how many senses of general language words practically coincide with senses in specific domains. A scope of such senses is not restricted with the legal domain. Let us take word "Building" as a noun in sense 1: *a relatively permanent enclosed construction over a plot of land, having a roof and usually windows and often more than one level, used for any of a wide variety of activities, as living, entertaining, or manufacturing* [10].

Terms with similar senses are necessary at least in two fields of public activity such as the construction trade and the field of public utilities. It means that majority of artifact senses of general language words coexist as terms in two fields of business activity: a field of industrial production of the artifact and a field using the artifact.

Main classes of such "dual" concepts include transportation means, job positions, technical devices, food, agricultural plants and animals, other natural objects, social, political and economic processes, art work and so on. These concepts are very important in everyday life, therefore people need language expressions to speak about them. At the same time fields of social activities, social sciences include them in their special languages. We estimate that almost 40 percents of general language word senses are used in various social subdomains. (For all estimations the lexical and terminological resource of Russian language RuThes containing more than 105 thousand words, collocations and terms [7] is used).

Thus we can distinguish a large specific domain, incorporating all these concepts – a domain of political, economic and social life, a domain that comprises general language senses coinciding with concepts of various domains of social activities. We call this polythematic domain "sociopolitical domain".

The sociopolitical domain has very interesting properties. These properties do it very useful to distinguish a sociopolitical zone in wordnets conceptual systems for automatic text processing goals.

5 Properties of Sociopolitical Domain

The sociopolitical domain has the following properties**.**

Property 1. Location of senses of the sociopolitical zone in general wordnets. Synsets belonging to the sociopolitical zone are situated mainly in the lower levels of the wordnet's conceptual system. Therefore the senses are the most thematically definite. The consequences of the fact are as follows: if such a general word as "creation" is used in a text, it can relate to different entities, different elements of the text structure. If such a "sociopolitical" word as "transportation" is mentioned several times in a text it is possible to suppose that all usages of the word are elements of the same topic structure and use this fact, for example, for construction of lexical chains and identification of the thematic structure of a text [9].

Property 2. Lexical ambiguity within the sociopolitical zone of the general language conceptual structure is much lower. For instance, in the current version of RuThes the ratio, denoting amount of second, third and other senses of expressions,

N = (number of relations "word-concept" – number of different words)
/ number of different words

is more than 4 times lower in the sociopolitical zone than in the whole resource.

Property 3. Lexical disambiguation for synsets within the sociopolitical zone is much easier, because different senses are often situated in different social subdomains and have rather different contexts of their usage in texts. For information-retrieval purposes synsets of the sociopolitical zone are much more important. Therefore it is possible to divide word sense disambiguation into three parts:

- disambiguation within the sociopolitical zone;
- disambiguation of term senses belonging the sociopolitical zone and general levels of the language conceptual system, to decide if a sociopolitical sense was applied;
- work with undisambiguated words out of the sociopolitical domain.

This combined approach to lexical disambiguation can diminish problems of incorrect disambiguation in automatic text processing in wordnet-based information-retrieval systems.

Property 4. Besides linguistic expressions having dual functions as general language means and terminological means there are a lot of terms (usually multiword terms) in domains of public affairs which can be understood by majority of the native speakers such as *aircraft industry, crime prevention, military assistance, internal migration*. The existence of such a polythematic terminological level, its importance for various information needs was recognized by developers of information-retrieval thesauri. Several general sociopolitical

thesauri [6,13] have been created and are used for indexing and retrieval of such important types of documents as governmental, parliamentary, international documents.

This set of terms can be naturally added to the sociopolitical zone of a wordnet. The inclusion of multiword expressions gives additional information for disambiguation. Such an enhanced wordnet becomes a valuable initial source for development of domain-specific ontologies. So for development of Avia-Ontology, describing interaction of an operator (air crew) and board equipment in various flight situations (1200 concepts, 3400 terms), almost a third part of the ontology was taken from thesaurus RuThes [4], comprising a lot of Russian sociopolitical terminology.

6 Related Work

Broadly speaking, the sociopolitical domain can be compared with an aggregate of all subject fields proposed in [8], except the Factotum field. The main differences are as follows:

- Systems of subject fields can be quite different. We propose not to work with any given system but analyze if a synset belongs a set of possible domains of social activity.
- The main point here is not to find such domains for maximal number of synsets but provide real analysis of domains otherwise multiple overgeneration of subject field codes can arise.
- It is important not only to mark "sociopolitical" synsets but recognize the existence of a broad layer of synsets belonging to as the general language system as to upper levels of various specific domains' hierarchies.

7 Conclusion

A border between a general language lexicon and terminologies of specific domains is not sharp and abrupt. It looks more as a broad strip and contains general language senses practically coinciding with concepts of social subdomains and concepts of specific domains understandable for native speakers.

Detailed description of concepts, terms, words from this "transition area", called "sociopolitical domain", can be naturally added to a wordnet' semantic network and facilitate solution of such problems as lexical disambiguation and identification of the text structure, enhance coverage of domain-specific texts by wordnets' synsets, improve effectiveness of the wordnets use in various automatic text processing applications.

Acknowledgements

Partial support for this work is provided by the Russian Foundation for Basic Research through grant # 030100472.

References

1. Buitellar, P., Sacalenau, B.: Extending Synsets with Medical Terms In proceedings of the NAACL workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations, Pittsburg, USA (2001).
2. Climent, S., Rodriguez, H., Gonzalo, J.: Definitions of the links and subsets for nouns of the EuroWordNet project. – Deliverable D005, WP3.1, EuroWordNet, LE24003 (1996).
3. Cruse: Lexical Semantics. – Cambrige (1986).
4. Dobrov, B., Loukachevitch, N., Nevzorova, O.: The Technology of New Domains' Ontologies Development. Proceedings of X-th Intern. Conf. KDS 2003 "Knowledge-Dialogue-Solution". June 16–26, Varna, Bulgaria. pp. 283–290. (2003).
5. ISO/DIS 704: Terminology work – Principles and methods. Geneva, ISO (Revision of second edition 704:1987) (1999).
6. LIV: *Legislative Indexing Vocabulary*. Congressional Research Service. The Library of Congress. Twenty first Edition (1994).
7. Loukachevitch, N. V., Dobrov, B. V.: Development and Use of Thesaurus of Russian Language RuThes. Proceedings of workshop on WordNet Structures and Standartisation, and How These Affect WordNet Applications and Evaluation. (LREC 2002) / Dimitris N. Christodoulakis, Gran Canaria, Spain – p. 65–70 (2002).
8. Magnini, B., Cavaglia, G.: Integrating Subject Field Codes into WordNet. – Proceedings of LREC 2000, Athens (2000).
9. Morris, J., Hirst G.: Lexical cohesion computed by thesaural relations as an indicator of the structure of a text. Computational linguistics, 17(1), pp. 21–48 (1991).
10. Random House Unabriged dictionary: Random house, Inc. (1999).
11. Rondeau, G.: Introduction a a terminologie. Quebec (1980).
12. Sager, J.C.: A Practical Course in Terminology Processing. Amsterdam: J. Benjamins (1990).
13. Thesaurus EUROVOC: Vol. 1–3 / European Communities. – Luxembourg: Office for Official Publications of the European Communities, Ed. 3. – English Language (1995).
14. Vossen, P.: Extending, Trimming and Fusing WordNet for Technical Documents. In: Proceedings of WordNet and Other Lexical Resources: Applications, Extensions and Customizations, Pittsburg, USA (2001).
15. Wuster, E.: Einfurung in die Allgemeine Terminologielehre and terminologishe Lexicographie. – Wien; N.Y., 1979/Bd 1–2.