Constructing of an Ontology-based Lexicon for Bulgarian

Kiril Simov, Petya Osenova
Linguistic Modelling Department, IPP, Bulgarian Academy of Sciences
Acad. G.Bonchev 25A, 1113 Sofia, Bulgaria
E-mail: kivs@bultreebank.org, petya@bultreebank.org

Abstract

In this paper we report on the progress in the creation of an Ontology-based lexicon for Bulgarian. We have started with the concept set from an upper ontology (DOLCE). Then it was extended with concepts selected from the OntoWordNet, which correspond to Core WordNet and EuroWordNet Basic concepts. The underlying idea behind the ontology-based lexicon is its organization via two semantic relations - *equivalence* and *subsumption*. These relations reflect the distribution of lexical unit senses with respect to the concepts in the ontology. The lexical unit candidates for concept mapping have been selected from two large and well-developed lexical resources for Bulgarian - a machine readable explanatory dictionary and a morphological lexicon. In the initial step, the lexical units were handled that have equivalent senses to the concepts in the ontology (2500 at the moment). Then, in the second stage, we are proceeding with lexical units selected on their frequency distribution in a large Bulgarian corpus. This step is the more challenging one, since it might require also additions of concepts to the ontology. The main applications of the lexicon are envisaged to be the semantic annotation and semantic IR for Bulgarian.

1. Introduction

The more demanding developments in natural language processing (such as, Machine Translation, Semantic Web applications, cross-lingual information retrieval, etc.) require richer types of knowledge in order to achieve a better coverage, a deeper processing and reliable results. Semantically annotated data, which are a mandatory preliminary step, must also provide context inference in order to resolve the linguistic ambiguities and extralinguistic usage constraints.

In this paper, we report on the creation process of a Bulgarian lexicon, based on ontology. It is a well-known fact, that despite its limitations (such as static state), the ontology provides the basic formalized knowledge about (some part of) the world. This knowledge supports the reasoning and is much richer relationally than the knowledge, represented within the other types of semantically-rich lexicons. Such an ontological lexicon is considered a part of the minimum semantic resources, necessary for the semantic annotation and applications of Bulgarian texts. We consider the following resources to be the minimal set of the semantic package:

- A lexicon for Bulgarian, mapped to an ontology. In its mapping to an upper ontology, it is viewed as a mechanism to cover the common lexica. In its mapping to domain ontologies, it is envisaged to cover the respective domain terminology;
- An annotation grammar for Bulgarian, based on the combination of the syntactic knowledge of the language, and the conceptual information from the ontology. The grammar itself comprises grammar rules for recognition of lexical units in the text as well as rules for selecting the correct interpretation in a given context;
- A corpus, manually annotated with ontological information in order to provide training and test environment for the machine learning components in the automatic word sense disambiguation modules. Thus, the appropriate concept for a lexical unit in context might be selected.

In this paper, we focus primarily on the process of the lexicon construction. In three recent European projects we have developed and have used the ontology-to-text relation, which facilitates the text annotation with domain concepts. Based on this experience, we started the development of an ontology-based lexicon for Bulgarian. It is constructed in an incremental manner. Our general view is as follows:

To support the semantic annotation for certain practical applications, we rely on an ontology-based lexicon. We also assume that there is a domain ontology which is used in the process of annotation. The domain ontology comprises three layers: its specific domain, middle and upper. The lexicon is mapped to the domain ontology. This mapping is based on relations between the meaning of the lexical units in the lexicon and concepts (relations and instances) in the ontology. Thus, we assume that the ontology contains the conceptual information necessary to model the word senses in the lexicon. The advantage of using an ontology is that the reflections of the conceptualization of the world become explicit.

The motivation for the construction of such a lexicon is the need for more precise semantic annotation. In order to ensure this, the lexicon has to provide more complex conceptual information than the one in computational lexicons like WordNet. The second requirement for the ontology-based lexicon is the coverage of the words in the text. The lexicon has to cover not only the domain terms, but also the non-specialized language. This is necessary for ensuring enough explicit knowledge for the application of word sense disambiguation methods. based on statistics. Since the development of a general ontology to support all the lexical units in a language is an intractable problem, we construct the ontology in an incremental way starting from the existing upper and middle layers, and contributing mostly to the domain specific parts. Then lexical units are mapped to this ontology via two relations – equality and subsumption.

The first is used when the appropriate concept for a meaning of some word is already represented in the ontology. The latter is used when such a concept is missing and only super-concepts are available.

The structure of the paper is as follows: the next section discusses related work; then the architecture of a domain ontology-based lexicon is discussed briefly; the fourth section presents the creation steps of the ontology-based lexicon of Bulgarian; the fifth section discusses the encoding of some special phenomena; and the last section concludes the paper.

2. Related Work on Ontology and Lexicon

Ontologies and lexicons are artifacts reflecting the human abilities for representing, processing and managing linguistic and conceptual knowledge. As such, they allow for the combination of many different approaches. A recent overview of the relation between ontologies and lexicons is presented in Hirst (2004). The paper discusses the structure of lexical entries, the knowledge recorded in them and mechanisms for interrelation of the lexicon elements. Special attention is given to the definition of *word sense*, its conceptual structure, relations between senses and problematic cases. The main topics under discussion are near-synonyms, gaps in the lexicon, and linguistic categorizations that are not ontological.

We assume that the lexicon is based on the ontology, i.e. the word senses are represented by concepts, relations or instances. Near-synonyms are words that share the same central conceptual information, but differ in the additional information they provide to the semantic interpretation module, such as small changes in the denotation, different implications, speaker attitude, etc. Our model does not solve this problem completely. We represent only the central part of the meaning of a word. The additional parts of the meaning (context related variations) can be encoded as additional information in the lexical entry or as an extension of the ontology where it is appropriate – similarly to the model used in (Edmonds and Hirst, 2002). The problem of lexical gaps is solved by allowing the storage of free phrases in the lexicon. Similarly, gaps in the ontology (a missing concept for a word sense, for example) are solved by appropriate extensions of the ontology. The non-ontological linguistic categorizations are not treated in our model.

As it was mentioned above, the construction of a Bulgarian ontology-based lexicon is motivated by the need to introduce more world knowledge into the semantic analysis of texts. In (Morris and Hirst, 2004) it is pointed out that most of the lexical relations necessary to determine the semantic content of lexical units are non-classical in contrast to the classical ones, i.e. *hyponymy*, *meronymy*, and *antonymy*. The non-classical relations are specific to some classes of meanings, i.e. *made-of*, *used-for*, etc. In our case, we assume that these relations are represented in the ontology. Thus, being formally defined, they can be used for the purposes of semantic inference and for the representation of some language phenomena like polysemy, metonymy, etc.

Our approach to the mapping between the lexicon and the ontology draws in many respects on the work done on WordNet (Fellbaum ,1998), EuroWordNet (Vossen, 1999), SIMPLE (Lenci et al., 2000). With WordNet-like lexicons we share the idea of grouping lexical units around a common meaning and in this respect the term groups in our model correspond to synsets in the WordNet model. The difference is that the meaning is defined independently in the ontology. With the SIMPLE model we share the idea to define the meaning of lexical units by means of an ontology, but we differ in the selection of the ontology which in our case represents the domain of interest, and in the case of SIMPLE reflects the lexicon model: Generative Lexicon – (Pustejovsky, 1995). Similar is the connection with EuroWordNet.

With the LingInfo model – (Buitelaar et. al., 2006a,b) – we share the idea that grammatical and context information also needs to be presented and linked to the ontology, but we differ in the implementation of the model and the degree of realization of the concrete language resources and tools.

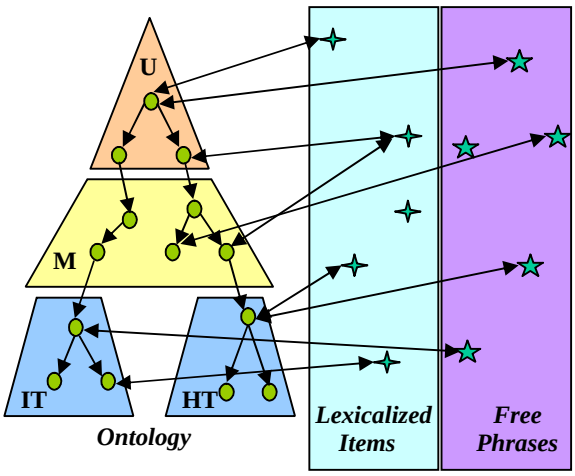
Finally, we would like to mention the work by (Nirenburg & Raskin. 2004) on the Ontology Semantics. It is very similar to our model except that we use existing ontologies like DOLCE, and we allow for an incremental construction of the lexicon.

3. Domain Lexicons

The model of the ontology-to-text relation used within the EU projects is described in (Simov & Osenova 2008; Osenova et al. 2008). It is based on the assumption that the ontology has a central role in the definition of the ontology-to-text relation and the language information reflects the available conceptual information in the ontology. The mapping is directed from the ontology to the lexicon, then from the lexicon to the grammar, and finally to the text. For each concept (relation, instance) in the ontology the lexicon contains at least one lexical unit. This requires the lexicon to contain non-lexicalized (fully compositional or free) phrases as well. Availability of different lexical units (lexicalized or not) for a given concept is used as a basis for the construction of the annotation grammar. This availability allows us to capture different wordings of the same meaning in a text. In general, a concept might have a few terms connected to it and a (potentially) unlimited number of free phrases expressing this concept in the language. Some of the free phrases receive their meaning compositionally regardless of their usage in a given text, other free phrases denote the corresponding concept only in a particular context.

The picture on the next page depicts the model. On the left side is the ontology, which is divided in three layers, which will be considered here: *Upper layer*. The alignment of the domain layer to an upper ontology is an obligatory step in each ontology creation methodology. This alignment ensures several properties of the domain ontology: (1) its consistency with the design of the upper ontology; (2) inheritance of the knowledge represented in the upper ontology. The inheritance requires the imposition of more specific constraints reflecting the structure of the domain. *Middle layer*. This layer contains concepts and relations which are neither part of the upper

layer, nor of the domain one, but play an important role for the alignment between them. For example, *carpet* is in the domain layer for the Home Textile ontology and *artifact* is in the upper layer, but the concept for *covering*, which is more specific than *artifact* and more general than *carpet* (defined as textile floor covering) is in the middle layer. *Domain layer*. At this layer we have the domain concepts and relations representing the main notions in the domain. These concepts and relations are used for solving different tasks, such as the representation of domain knowledge, the representation of common conceptualization for information exchange in the domain, the semantic annotation of domain texts, etc. On the right side of the picture, the mapping to the lexicon is given. The linguistic knowledge is encoded in addition as part of the lexicon and in the grammar. The assumption in the model is that the connection between the meaning of the words and the concepts in the ontology is equivalence.



Our experience in using this model showed some disadvantages of using only equivalence as a semantic relation between the meaning of words and the ontological concepts.

4. Constructing of an Ontology-based Lexicon of Bulgarian

The main problem with the above model of the *ontology-to-text* relation is the fact that the lexicon is mapped only in its domain part to the ontology. Thus, the annotation of domain texts with domain concepts is very sparse. For example, in the IT domain we have annotated 8 concepts within 100 tokens (with 14.8 tokens per sentence = 1.19 concepts per sentence at average). The sparse annotation blocks possibilities for using better methods for word sense disambiguation. This holds when the lexical units in the domain lexicon are ambiguous among themselves or with respect to the lexical units from the general lexicons. For example, the concepts *key-of-keyboard*, *key-of-database* and *key-for-door* have the same wording in English (*key*) and the last concept is not from the domain ontology. Therefore, we need a much better semantic annotation than the one which just uses the domain terms and grammar constructed on their basis. In order to overcome the problem, we have decided to extend the lexicon coverage also with the general (domain independent part of the ontology).

The main problem in this case is the fact that there is no appropriate ontology to which the general lexical units from the Bulgarian lexicon to be mapped via equivalence relation. Thus, we have decided to add a second semantic relation – subsumption – which to be used for mapping between lexical units with more specific senses than the concepts available in the ontology. The aim was to have a smaller ontology which covers the basic conceptual level of Bulgarian and also specifies the appropriate level of granularity of senses. This ontology has to provide a wide range of concepts on upper and middle level in order to provide appropriate matching concepts for the Bulgarian lexica and enough specific information for the semantic annotation tasks. We decided to construct such an ontology by using several resources: (1) DOLCE foundational ontology as an upper ontology; (2) OntoWordNet as a source of more specific concepts (OntoWordNet is WordNet 1.6 aligned to DOLCE); (3) Core WordNet [1] (CWN contains 5000 synsets from WordNet on the basis of analysis of British National Corpus) and EuroWordNet Base Concepts (Vossen, 1999) as a source of basic level conceptualization to be mapped to the Bulgarian lexicon. The concepts selected from Core WordNet and EuroWordNet are considered as the middle layer of the ontology.

The procedure for the construction of the Bulgarian Ontology-based lexicon is as follows:

- From OntoWordNet the concepts are selected, which correspond to the synsets in Core WordNet and EuroWordNet Base Concepts. We used the synsets from OntoWordNet because they have been already mapped to DOLCE ontology;
- Together with DOLCE, they formed a starting point for the necessary ontology. One important feature of the ontology created in this way is that the concepts formed on the basis of OntoWordNet inherit properties from the concepts in DOLCE. Although these properties are very general, they provide some constraints on the definition of the more specific concepts. They also can be further made more specific for each more concrete concept;
- Using an English-Bulgarian dictionary and an Explanatory Dictionary of Bulgarian, the candidate Bulgarian lexical units were selected. The English-Bulgarian lexicon is used in order to select potential lexical entries from the Bulgarian Explanatory Dictionary for each English wordform in the corresponding synset;
- Manual selection of the appropriate Bulgarian sense from the candidates has been performed. The lexical units corresponding to the senses are linked to the concept.

Currently we have processed more than 2500 concepts including the concepts from DOLCE, the intersection of Core WordNet and EuroWordNet Base Concepts as well as some more specific concepts from our previous work on domain ontologies. Our next tasks are to cover all the synsets selected in Core WordNet as much as they are selected on the basis of the sense distribution. This guaranties the central role of corresponding concepts in the semantic annotation. The next step will be the extension of the lexicon with lexical units for which there

[1] http://wordnet.princeton.edu/wordnet/download/standoff/

is no concept in the ontology, equivalent to the sense of the lexical unit. We have selected the candidate lexical units from a large corpus of Bulgarian texts (currently more than 130 million running words). The corpus was lemmatized and the lemmas were ranked on the basis of the frequency of their word forms in the corpus as well as the number of documents in which they appear. These lexical units are already present in our morphological lexicon and they will be gradually mapped to the ontology via the subsumption relation.

5. Encoding of Special Phenomena

Including of a formal ontology in the lexicon construction provides many possibilities for using the knowledge, represented in the ontology and the services related to it, such as the inference mechanism. In this section, we present the encoding of some important phenomena for the task of word sense disambiguation: metonymy and verb frames representation. The metonymy covers also a substantial part of the cases of the regular polysemy. For an overview on regular polysemy, its representation and importance of this representation see (Barque and Chaumartin 2008). We assume that the patterns described by the authors can be represented as inference patterns in our model of lexicon to ontology mapping.

A general assumption in the treatment of the above mentioned phenomena is that the related word senses are already represented in the ontology. In this way, the lexical representation of the corresponding patterns (metonymical or frame) is done via appropriate mappings to the corresponding concepts in the ontology. The application of such patterns for creation of new senses is not explored in our work.

Let us consider the case of metonymy in more detail. In general, metonymy is defined as a trope in which one entity is used to stand for another associated entity[2]

Our treatment follows the ideas of (Hobbs 2003) who interpreted it by introduction of a function which relates the mentioned object with the intended one. The function is different for different cases of metonymy and it can be context dependent. In order to implement the same idea, we assume that the function is determined by an inference over the ontology and the context. This function is a composition of relations from the ontology. We consider the representation of such compositions in the lexicon as an important device for facilitation of text annotation. Our view of these compositions is that they are very specific inference rules. In future, we will investigate the possibility to encode the metonymy relations reported in the literature (like the ones presented in (Barque and Chaumartin 2008)) as such special inference rules. Here we present the interpretation of two cases of metonymy.

Let us suppose that we have to annotate the sentence "She was wearing stripe." First we annotate 'stripe' as a kind of *property* and as such it is connected to 'cloth' via the *property-of* relation and 'cloth' is annotated as *material* and it is connected to 'clothing' via the *made-of* relation. The concept 'clothing' is of the relevant type for the object of the verb 'to wear'. Thus, the understanding of the sentence is something like: "She was wearing a clothing made from a textile with a stripe design."' The composition of the corresponding relations is stored in the lexical entries for the corresponding lexical units. In the case of metonymy this is a better option, because the possible patterns are (potentially) infinite in number. Representing each metonymy usage as a separate meaning will result in many strange meanings for the lexical units. In this way, we represent the most frequent metonymy uses as inference patterns and the actual inference is done during the analysis of the discourse where the lexical unit is used metonymically.

When the regular polysemy is an example of metonymy, we represent it in the same way. The different meanings are represented in the ontology as different concepts and these concepts are connected via appropriate relations. The main difference here is that for each of the meanings we construct a separate lexical entry. This means that during the analysis of the text we have to disambiguate between these senses. In some cases, more that one of the senses is visible via one usage of the lexical unit. For example, in the sentence "This large book is very interesting." the word 'book' is used simultaneously as a *physical object* selected by 'large' and as an *information object* selected by 'interesting'.

Currently, we do not represent in the lexicon the relation between the literal meaning of a given word and its metaphorical meaning. In contrast to metonymy, metaphorical meanings are not always closely related in the ontology. They require a special kind of inference by analogy, which differs in many respects from the inference necessary to deal with metonymy.

The encoding of verbs is also a very important for the task of semantic annotation. We assume that the appropriate information is represented in two ways: (1) in the ontology each verb is connected to an event concept related to the meaning of the verb. In the ontology all the participants (irrespectivly of whether they are considered to be arguments, adjuncts, etc.) are represented as such via appropriate relations; (2) the linguistic behavior is encoded in the lexicon as a set of frames. These frames determine the role of each participant in the given event. During the annotation, the verbs are annotated with the frames from the lexicon and the corresponding relations are connected with appropriate phrases from the text. Some of them remain unconnected when the corresponding participant is not explicitly mentioned in the text.

The encoding of relational adjectives is done on the basis of ontology statements instead of connection to a concept in ontology. These statements in logical terms are formulas with one free variable, which corresponds to the object to which the meaning of the adjective applies. For example *golden* in one of its senses will be connected to statement: *x made-of gold*. The restriction on the variable x will be inferred from the ontology, or it can be stated explicitly in the associated statement.

6. Conclusion

In the paper we presented the construction of an ontology-based lexicon for Bulgarian on the basis of an ontology and machine readable dictionaries. This lexicon originates from the practical task of semantic annotation of domain texts, and semantic retrieval over them. Our initial efforts went to the mapping from domain ontologies to terminological lexicons. However, due to

[2] http://www.sil.org/linguistics/GlossaryOfLinguisticTerms/WhatIsMetonymy.htm

the sparseness of the resulting concept annotation, the coverage was extended to the general lexicon. Because the size of the ontology is much smaller than the target size of the lexicon, we rely on two relations between lexical units and the concepts in the ontology: equality and subsumption. The first is used primarily for the domain ontology and the second for the middle and upper part of the ontology.

Our future goals are to implement a system for automatic word sense disambiguation and for detection of metonymical uses in the text. The extension of the lexicon coverage is also one of our tasks. In addition, the general lexicon together with the ontology could be used for the creation of domain ontologies and lexicons. We also plan an annotation of a corpus with concepts from the middle and upper part of the ontology.

Another envisaged task of ours is to enrich the ontology with more information from different sources, such as dictionary definitions, wikipedia, and other ontologies. The final goal of this work is to gather together as much knowledge as possible at availability for various practical tasks..

7. Acknowledgements

The work reported here is partially done within the context of the EU project – Language Technology for Lifelong Learning (LTfLL). We would also like to thank the three anonymous reviewers for their valuable remarks.

8. References

Barque L. and Chaumartin Fr. R. (2008). *Regular polysemy in WordNet*. In: Proceedings of Lexical-Semantic and Ontological Resources Maintenance, Representation, and Standards, KONVENS 2008. Berlin, Germany.

Buitelaar P., M. Sintek, M. Kiesel. (2006). *A Lexicon Model for Multilingual/Multimedia Ontologies*. In: Proceedings of the 3rd European Semantic Web Conference (ESWC06), Budva, Montenegro, June 2006.

Buitelaar P., Th. Declerck, A. Frank, S. Racioppa, M. Kiesel, M. Sintek, R. Engel, M. Romanelli, D. Sonntag, B. Loos, V. Micelli, R. Porzel, Ph. Cimiano. (2006). *LingInfo: Design and Applications of a Model for the Integration of Linguistic Information in Ontologies*. In: Proc. of OntoLex06, a Workshop at LREC, Genoa, Italy, May 2006.

Edmonds Ph. and Hirst Gr. (2002). *Near-synonymy and lexical choice*. Computational Linguistics, Vol. 28:2, pp. 105-144.

Fellbaum Chr. (1998). Editor. *WORDNET: an electronic lexical database*. MIT Press.

Hirst Gr. (2004). *Ontology and the lexicon*. In: Steffen Staab and Rudi Studer (editors), Handbook on Ontologies. Springer Verlag, Berlin, Germany. pp 209-229.

Hobbs J. R. (2003). *Discourse and Inference*. University of Southern California, Marina del Rey, California. Unpublished manuscript. http://www.isi.edu/~hobbs/disinf-tc.html

Lenci A., F. Busa, N. Ruimy, E. Gola, M. Monachini, N. Calzolari, A. Zampolli, E. Guimier, G. Recourcé, L. Humphreys, U. von Rekovsky, A. Ogonowski, C. McCauley, W. Peters, I. Peters, R. Gaizauskas, M. Villegas. (2000). *SIMPLE Work Package 2 - Linguistic Specifications*, Deliverable D2.1. ILC-CNR, Pisa, Italy.

Morris J. and Hirst Gr. (2004). *Non-Classical Lexical Semantic Relations*. In: Proceedings of the HLT Workshop on Computational Lexical Semantics. Boston, USA. pp 46-51.

Nirenburg S. and V. Raskin. (2004). *Ontological Semantics*. MIT Press.

Osenova P., K. Simov, E. Mossel. 2008. *Language Resources for Semantic Document Annotation and Crosslingual Retrieval*. In: Proc. of LREC 2008, ELRA.

Pustejovsky J. (1995). *The Generative Lexicon*. Cambridge, MA: MIT Press.

Simov K. and P. Osenova. *Language Resources and Tools for Ontology-Based Semantic Annotation*. OntoLex 2008 Workshop at LREC 2008, pp. 9-13. 2008.

Vossen P. 1999. Editor. *EuroWordNet General Document*. Version 3, Final, July 19, 1999. http://www.hum.uva.nl/~ewn