

New Applications of “Ontology-to-Text Relation” Strategy for Bulgarian Language

Kamenka Staykova, Petya Osenova, Kiril Simov

Institute of Information and Communication Technologies, 1113 Sofia

Emails: staykova@iinf.bas.bg petya@bultreebank.org kivs@bultreebank.org

Abstract: *The paper presents new applications of the Ontology-to-Text Relation Strategy to Bulgarian Iconographic Domain. First the strategy itself is discussed within the triple ontology-terminological lexicon-annotation grammars, then – the related works. Also, the specifics of the semantic annotation and evaluation over iconographic data are presented. A family of domain ontologies over the iconographic domain are created and used. The evaluation against a gold standard shows that this strategy is good enough for more precise, but shallow results, and can be supported further by deep parsing techniques.*

Keywords: *Ontologies, semantic annotation, terminological lexicons, annotation grammars.*

1. Introduction

In recent years the semantic annotation has become an inevitable step in processing NLP chains, since it allows better identification and explication of knowledge information. The semantic annotation might range from named-entities to Minimal Recursion Semantics structures. It is also a very popular step in Semantic Web applications, which try to make the immense and still growing multilingual information in Web available to the international community. The most recent

works, such as [4] and [8], for example, concern the usage of semantic annotations for representation of multilingual information from cultural heritage texts, multilingual mapping, and language localization strategies for the Semantic Web. The problems of semantic annotation in connection with Natural Language Processing have been in the focus of many papers, among which [1]. Needless to say, there exist also approaches that do not benefit from language resources, like in [6].

In this paper we mean under “semantic annotation” the annotation of text chunks with concepts, structured in a domain ontology that fits the domain of the processed texts. Such a task is not trivial, since a domain ontology is needed (as a language-independent knowledge core), as well as terminological lexicons (as language dependent resources). But there is also a need of a mapping mechanism between the ontology and the lexicons.

Apart from the availability of these resources, another issue is the possibility of extending and adjusting both types of resources with respect to the coverage and precision. Thus, on one hand, the domain texts themselves provide material for the lexicons, as well as for the missing concepts in ontology. On the other hand, the annotation on in-domain texts evaluates the quality of the applied resources.

The paper is structured as follows: in the next Section the Ontology-to-Text Relation Strategy is presented with respect to the triple: domain ontology-terminological lexicon-annotation grammars. In Section 3 the focus is on the new specific applications of this strategy to Bulgarian Iconographic Domain. Section 4 concludes the paper.

2. Defining the Ontology-to-Text Relation Strategy

The approach evolved from the method to create, support and use Ontology-Based Lexicons [12]. The underlying idea of Ontology-to-Text Relation Strategy is to have some starting ontological notions, defined in a particular domain. Then, the adequate NLP technology is applied, so that the lexicalizations of the ontological notions to be recognized as verbalizations in natural language texts. The task is not trivial at all, because: (1) not all the ontological terms are (neither could be) lexicalized in a natural language form; (2) not all of the ontological notions occur in natural language texts in the way they are defined by experts of the ontology domain; and (3) the ontological terms could be presented in lexical elements or free phrases in various ways within the natural language texts. The strategy aims at handling the knowledge in domain ontologies, where the ontological notions and statements are agreed on by the experts of a particular domain of knowledge.

Ontology-to-Text Relation Strategy is defined as a connection between ontological notions of a given ontology and a terminological lexicon with lexicalizations of these notions in natural language. The terminological lexicon might play the role of a basis for the construction of annotation grammar of the given natural language, so that the grammar recognizes the ontological notions in natural language texts.

2.1. Resources

Ontology-to-Text Relation Strategy comprises two intermediate components: a terminological lexicon and an annotation grammar.

The terminological lexicon plays a twofold role. First, it inter-relates the concepts of the ontology to the lexical knowledge used by the grammar in order to recognize the concept role in the texts. Second, the lexicon represents the main interface between the user and the ontology. This interface allows the ontology to be navigated or represented in a natural way for the user. For example, the concepts and relations might be named with terms used by the stakeholders in their everyday activities and in their own natural language (e.g., Bulgarian). This could be considered as a first step to a contextualized usage of the ontology.

The concept annotation grammar is considered ideally as an extension of a general language deep grammar which is adapted to the concept annotation task. Minimally, the concept annotation grammar contains for each term in the lexicon at least one grammar rule aiming at the recognition of the term lexicalizations in natural language texts. Input texts annotation with grammatical features and lemmatization depends on the particular choice of grammatical resources used with the strategy. It is seen as a pre-processing step concerning the application of Ontology-to-Text Relation Strategy. The disambiguation rules exploit the local context in terms of grammatical features, semantic annotation and syntactic structure, and also the global context, such as topic of the text, discourse segmentation, etc. The following picture indicates the relation among the lexicon, the annotation grammar and the text:

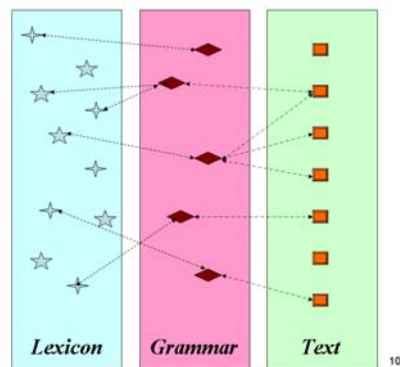


Fig. 1. Components of Ontology-to-Text Relation Strategy

2.2. The resource creation cycle

The realization of Ontology-to-Text Relation Strategy contains three steps, which could be repeated in a cycle to achieve better results for a particular natural language:

1. Create/enrich the terminological lexicon of ontological concept lexicalizations in a given natural language. In general, different ways, in which one concept could be represented in a natural language text are potentially infinite in

number, so only the most frequent lexicalizations could be represented or only the occurrences in a corpus of specialized texts could be considered.

2. Develop/tune some grammatical resources of a particular language to cover the lexicalizations of the terminological lexicon. It is important how deep the parsing should be with respect to the given task.

3. Evaluate the performance, for example, against a “gold standard” text corpus.

In Fig. 2 below the design of the applied Ontology-to-Text Relation Strategy is presented. The NLP technology is indicated in the box. It might consist of resources (morphological lexicons, annotation grammars) and other supporting tools, such as tokenizers, sentence splitting module, etc. The Input texts indicate the data that has to be annotated. The ontological concepts and the terms mapped to them from the terminological lexicons are connected to the text chunks via the annotation grammars.

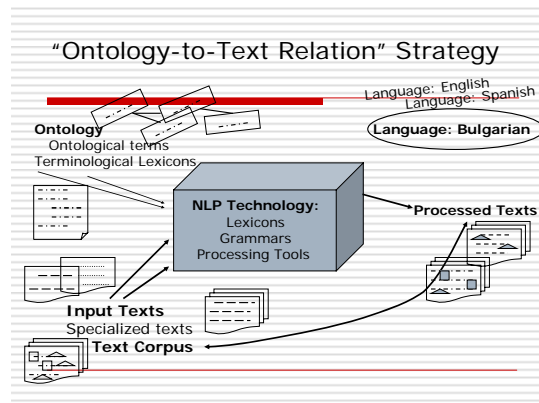


Fig. 2. The design of applied Ontology-to-Text Relation Strategy

2.3. Related works

The basic idea of Ontology-to-Text Relation Strategy is described in the works presented in [5], which focuses on the nature of Ontology-based Lexicons. The basic difference is the assumption behind the Ontology-to-Text Relation Strategy that the lexicon is based on the ontology, i.e., the word senses of all lexicon entries are represented by ontological concepts, relations or instances. The approach draws in many respects on the work, done on WordNet [3], EuroWordNet [14], SIMPLE [7]. With WordNet-like lexicons it shares the idea of grouping lexical units around a common meaning and in this respect the term-groups in Ontology-to-Text Relation correspond to synsets in the WordNet model. The difference is that the meaning is defined independently in the ontology.

An interesting application of such lexicons to semantic annotation task is demonstrated in [13]. In [10] the work of applying Ontology-to-Text Relation strategy to create an Ontology-Based Lexicon of Bulgarian is described.

Ontology-to-Text Relation Strategy is exploited with the EU project LT4eL “Language Technology for e-Learning” (<http://www.lt4el.eu>). The ontology used there is built up by notions in the domain of Computer Science (terms, such as

System, Computer, Program, Program Creator, etc.) and contains 1200 domain concepts with lexicalizations in eight natural languages. The input texts are lectures from the field of “Computer Science for Non-Computer Scientists” and could cover topics like operating systems; programs; document preparation – creation, formatting, saving, printing; Web, Internet, computer networks; HTML, websites, HTML documents; email, etc. The output is words or phrases, marked-up with ontological concepts.

Within the EU project AsIsKnown “A Semantic-based Knowledge Flow System for the European Home Textiles Industry” (<http://www.asisknown.org>) the Ontology-to-Text Relation Strategy is used over domain ontology in textile and interior design. The ontology includes over 2000 concepts. The terminological lexicons cover around 1.4 lexical units per concept. The ontology was used for semantic annotation of texts from magazines with the aim to predict the new trends in the interior design. For example, the changes in the colours of the carpets and curtains, etc. In contrast to the LT4eL ontology, where the domain-specific relations were few, in AsIsKnown ontology they were added consistently (width and length of the carpet; the material it is made of; the elastic characteristics, etc.). The semantic annotation has been enabled also for the images in a special application. There the stakeholder can annotate the objects in a picture with concepts from the domain ontology.

Both projects, above mentioned – LT4eL and AsIsKnown used the domain-specific texts for extracting of domain terms and identifying new domain concepts. Additionally, in AsIsKnown home textile international standards have been used for ensuring better domain coverage. Both projects have been multilingual, and thus – had to take into account the language specific issues.

3. Application in the Domain of Iconography for SINUS Project

SINUS Project “Semantic Technologies for Web Services and Technology Enhanced Learning” (sinus.iinf.bas.bg) is a research project funded by the Bulgarian National Science Fund. Project’s main objective is to develop an environment that applies Semantic Web technologies as well as to create Technology-Enhanced Learning applications [2]. SINUS environment contains semantic repositories and some of the developed services perform semantic annotation of information objects [16]. Ontology-to-Text Relation Strategy is applied for SINUS project to support the semantic annotation within SINUS environment. Annotated information objects of the project use-case reflect real iconographical objects (icons, wall-paintings, etc.) which multimedia descriptions and pictures are stored in Multimedia Digital Library “Virtual Encyclopedia of East-Christian Art” [9]. The semantic annotations of iconographical objects are based on several ontologies created for purposes of SINUS project. Each one of the ontologies describes a different aspect of iconographical knowledge and art domain: descriptive (library style) information, technologies of iconography, religious aspects of iconographical images, etc. This is in contrast to LT4eL and AsIsKnown

projects, where one ontology per domain has been constructed, and it was mapped to an upper ontology. Here a layered approach has been taken.

The realization of Ontology-to-Text Relation Strategy is claimed to be useful for the performance of SINUS environment, because it suggests pragmatic use of available texts describing the iconographical objects. The semantic annotation model is based on features which might be mentioned in a natural language form within the particular descriptive texts. For example, if the current semantic annotation model of particular iconographical object does not contain a filler for the feature of *Primer* of type *Plaster*, but the following text describes the same iconographical object as having:

*Гинсов грунд, нанесен тънко и равномерно.
Plaster ground coat, applied thinly and evenly.*

Then the feature *Primer* of type *Plaster* could be added to the semantic annotation.

Two variants of Ontology-to-Text Relation Strategy are realized in SINUS project. The first starts from ontological concepts in SINUS_TechnologySpec Ontology (16 classes, 14 object properties, 43 ontological individuals). Occurrences of notions *Gilding*, *Lacquering*, *Primer*, *Type of Gilding*, *Condition of Lacquering*, etc., can be searched within descriptive texts of particular icons. The texts focus is on materials and techniques used in the icon creation, or on the current condition state of the icons. When some occurrences are found in the texts, ontological expressions are returned as possible parts of the semantic annotation model, which could be used further, namely confirmed or rejected by the human annotator. The second application of Ontology-to-Text Relation strategy for SINUS project recognizes the terms of ontology SINUS_IconographicalImage (7 classes, 5 object properties, 21 ontological individuals): *Iconographical Character*, *Type of Character Image*, *Festive Scenes*, etc. concepts are sought in texts describing religious images of real iconographical objects. In both cases the desired output is a list of ontological constructions corresponding to the recognized terms, and the details of linguistic realization are not of interest; the list of offered ontological constructions is connected to the processed text as a whole.

Concerning the evaluation of performance, the decision taken for the SINUS applications is a “gold standard” text corpus to be used. The “gold standard” text corpus is prepared using a manual mark-up procedure supported by the CLARK system [11] (<http://www.bultreebank.org/clark/>). Three types of ontological terms lexicalizations are distinguished:

- lexical variations of the term: added to the terminological lexicon; for example, the term *Lacquering* could be realized in Bulgarian as *лак* (*lacquering*) or *лаково покритие* (*a layer of lacquering*);
- syntactic variations of term lexicalizations: object of target recognition by the grammar; for example, an alternative lexicalization *изображението е традиционно* (*the image is traditional*) for the term *традиционно иконографско изображение* (*traditional iconographical image*);

- lexicalizations of the term on semantic level, which are not a part of our task to tune grammars; for example, the human reader finds in the following text occurrence of *OilPrimer* notion: *Темперна живопис с повишено съдържание на масло в свързвателя. (A distemper painting with an increased content of oil in the vehicle).*

The application of Ontology-to-Text Relation Strategy which works with the notions of SINUS_TechnologySpec Ontology uses a terminological lexicon with 87 basic lexicalizations, mapped to 62 ontological concepts. The target lexicalizations, marked up in the “gold standard” text corpus, are 899. The second application for SINUS project with input by the SINUS_IconographicalImage ontology works with 96 ontological terms. The target occurrences in the “gold standard” text corpus are 563.

Similarly to LT4eL and AsIsKnown projects, SINUS project applications rely on the CLaRK system and its technology to apply Finite State Automata and to build grammars that recognize the lexicalizations, mapped to the ontological terms. In both cases, with SINUS_TechnologySpec Ontology and with SINUS_IconographicalImage Ontology, all possible variations of the elements within the chunks are lemmatized by the Bulgarian Morphological Lexicon [15]. Some non-local syntactic constructions are also modelled in the input expressions of the grammar by means of regular expressions of the system CLaRK, for example: `<RE>(изображение, #*, традиционно)</RE>`, for recognizing phrases like *изображението е традиционно (the image is traditional)*.

Interesting coordination is also modelled without additional syntactic analyses of input texts: the coordination of mentioned religious **Character** name and occurrence of **TypeOfCharacterImage** notion (*допоясно изображение, цял ръст, на кон и т.н./ half length image, full length image, on horse, etc.*); for example, *Христос е изписан в цял ръст (Jesus Christ is portrayed in full length)*.

The evaluation of recognition applied on “gold standard” text corpuses is again supported by the CLaRK system. The initial evaluation of performance of Ontology-to-Text Relation Strategy applied for SINUS_TechnologySpec Ontology shows Precision = 0.984 and Recall = 0.842. These metrics are much better after two cycles of tuning the grammars: Precision = 0.989 and Recall = 0.910.

Target occurrences in “gold standard” text corpus for the application with SINUS_IconographicalImage Ontology are 563, and 449 of them are recognized correctly. Examples of omissions are phrases as *Изображението се различава в значителна степен от традиционното...* (*The representation differs to a considerable degree from the traditional one...*), which should be recognized as the ontological individual **Untraditional Iconographical Image**. Such semantic level lexicalizations need much more mark-up and full parsing of input text on the top of the partial grammars that have been used so far. The current annotation grammars rely basically on the occurrences of phrases like *традиционно изображение (traditional image)* for recognition of the term **Traditional Iconographical Image** and on phrases like *нетрадиционно изображение (untraditional image)*, *специфично изображение (specific image)* for recognition of the term **Untraditional Iconographical Image**. There are only three unachieved

occurrences against 150 well recognized lexicalizations of the two terms in the text corpus. It is worth mentioning the results with unrecognized coordination Character – TypeOfCharacterImage, where nearly 2/3 of the target expressions are recognized, but the number of wrong and missed recognitions is relatively high. The unsatisfactory results here reach the line of trade-of between the cost of deeper parsing and the gain of right recognitions.

4. Conclusion

In general, there have been several challenges behind the task to apply Ontology-to-Text Relation Strategy. First of all, the type of the ontology that is being mapped to the text lexicalizations. Our survey showed that the more domain-level is ontology, the better is the recognition of the corresponding lexicalizations. Another issue is the recognition of relations. At the moment we have identified only the concepts (ontological classes and individuals), but not the relations between them, similarly to LT4eL project and in contrast to AsIsKnown Project.

For the task of the semantic annotation, some prerequisites are required in the area of Natural Language Processing. These are the following resources: terminological lexicons in the specific language of interest for the given domain; grammatical resources, parsers for this language, available specialized texts rich in domain terms, etc. Thus, the usual difficulties are: the change of the natural language and the change of the domain. The solution to the first problem depends on the availability of the necessary resources and technology, while the solution to the second problem depends on the availability of ontology in the specific domain and on the terminological lexicons corresponding to it. In any case, cross-linguality is a challenging, but interesting problem with promising capacity when applying Ontology-to-Text Relation Strategy.

The experiences with the presented herein new applications of Ontology-to-Text Relation Strategy for the purposes of SINUS project show that this variant of semantic annotating, although not-too-big with respect to the number of concepts, has its pragmatic value for the SINUS environment. The lack of terminological lexicons and a domain-tuned deep parser does not seem to be crucial for our task. The reason is that the language in iconographic domain can be considered as a “controlled language” with predictable syntactic and lexical variation. Thus, another conclusion can be drawn here, namely: it is necessary to estimate in advance the efforts and resources, required for each particular task.

The “gold standard” text corpus prepared for SINUS project applications is available for public use together with the paralleled English translations on the Bulgarian CLARIN website.

Acknowledgements: The research presented in this paper was partially funded by the Bulgarian NSF project D-002-189 SINUS “Semantic Technologies for Web Services and Technology Enhanced Learning”.

References

1. Bontcheva, K., H. Cunningham, A. Kiryakov, V. Tablan. Semantic Annotation and Human Language Technology. Semantic Web Technology: Trends and Research. John Wiley and Sons, Ltd., 2006.
2. Dochev, D., G. Agre. Towards Semantic Web Enhanced Learning. – In: Proceedings of the International Conference on Knowledge Management and Information Sharing, Madeira, 2009, 212-217.
3. C. Fellbaum, Ed. WORDNET: An Electronic Lexical Database. MIT Press, 1998.
4. Jones, D., A. O'Connor, Y. M. Abgaz, D. Lewis. A Semantic Model for Integrated Content Management, Localization and Language Technology Processing. – In: Workshop “Multilingual Semantic Web” at ISWC, 23-25 October 2011, Bonn, Germany. <http://iswc2011.semanticweb.org/fileadmin/iswc/Papers/Workshops/MSW/Dominic.pdf>
5. Hirst, G. Ontology and the Lexicon. – In: Steffen Staab and Rudi Studer, Eds. Handbook on Ontologies. Berlin, Springer Verlag, 2004, 209-229.
6. Cui, H., D. Boufford, P. Selden. Semantic Annotation of Biosystematics Literature without Training Examples. – In: Journal of American Society for Information Science and Technology, Vol. 61, 2010, Issue 3, 522-542.
7. 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. SIMPLE Work Package 2 – Linguistic Specifications. Deliverable D2.1. ILC-CNR, Pisa, Italy, 2000.
8. Moert, K., T. Declerck, P. Lendvai, T. Varadi. Accessing and Creating Multilingual Data on the Web for the Semantic Annotation of Cultural Heritage Texts. Workshop “Multilingual Semantic Web” at ISWC, 23-25 October 2011, Bonn, Germany. <http://iswc2011.semanticweb.org/fileadmin/iswc/Papers/Workshops/MSW/Karlheinz.pdf>
9. Pavlova-Draganova, L., V. Georgiev, L. Draganov. Virtual Encyclopaedia of Bulgarian Iconography. – Information Technologies and Knowledge, Vol. 1, 2007, No 3, 267-271.
10. Simov, K. Ontology-Based Lexicon of Bulgarian. – Journal for Language Technology and Computational Linguistics, Vol. 24, 2009, No 2, 40-55.
11. Simov, K., Z. Peev, M. Kouylekov, A. Simov, M. Dimitrov, A. Kiryakov. CLaRK – An XML-Based System for Corpora Development. – In: Proc. of the Corpus Linguistics Conference, 2001, 558-560.
12. Simov, K., P. Osenova. Applying Ontology-Based Lexicons to the Semantic Annotation of Learning Objects. – In: Proc. of the Workshop on NLP and Knowledge Representation for eLearning Environments, RANLP-2007, 49-55.
13. Simov, K., P. Osenova. Language Resources and Tools for Ontology-Based Semantic Annotation. – In: Al. Oltramari, L. Prévot, Chu-Ren Huang, P. Buitelaar, P. Vossen, Eds. Proc. of the OntoLex Workshop at LREC'2008, 2008, 9-13.
14. P. Vossen, Ed. EuroWordNet General Document. Version 3. Final. 19 July 1999. <http://www.hum.uva.nl/ewn>
15. Popov, D., K. Simov, S. Vidinska. Vocabulary of Orthoepy, Spelling and Punctuation. Sofia, Atlantic, 1998 (in Bulgarian).
16. Agre, G. SINUS – A Semantic Technology Enhanced Environment for Learning in Humanities. – Cybernetics and Information Technologies, Vol. 12, 2012, No 4, 5-24.