Low-cost ontology development

Marek Grác and **Adam Rambousek**
NLP Center
Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno
Czech Republic
{xgrac,xrambous}@fi.muni.cz

Abstract

In this paper, we present the project building new lexical resource – shallow ontology derived from the corpora. The ontology should be used primarily for machine translation, syntactic parsing and word sense disambiguation. Currently, the ontology for Czech language is developed, but the methodology and tools are suitable for other languages with similar structure.

Ontology is based on BushBank corpus, which improves handling of ambiguity in natural language. BushBank data and tools are application-driven, thus reducing the time and costs needed to annotate the corpora and develop new lexical resources.

1 Introduction

Language resources for natural language processing are very important for development as well as improvement of existing natural language processing (NLP) tools. Situation for different European languages varies a lot. In the worst case there are almost no resources and we have to face the problem of creating them cheaply and quickly while maintaining high quality. We can attempt to build an ultimate corpus that will be useful for every application but we do not believe that such approach is successful often enough. We have decided to model our corpus using application-driven development. This approach should prevent major design flaws which might not be automatically recoverable later and could limit the usefulness of resulting work for ours needs.

This approach was used to build a multi-layered annotated corpus which is one of the resources used for creating our ontology. Application-driven development means that at begining corpus does not contain any data directly usable for creating ontology as we avoid creating data with no immediate application (even if it might be useful in future). In fact proposed ontology is independent of original corpus and we plan to use also other (larger) corpora for enriching ontology.

This paper focuses on a new type of annotated corpus named BushBank and an example study of building shallow ontology for Czech language on top of it. Semantic networks are among the most popular formalisms for knowledge representation. Like other networks, they consists of nodes and links. For English we have a number of possibilities from domain oriented to general ones like Princeton WordNet (Miller, 1990). It is very rich and complex network but unfortunately only few applications use its potential.

Creation of similar resources for other close languages such as Czech is very difficult and also time-consuming. Our goal is to create a simpler ontology which will be easy to create and use primarily in our existing applications. This application-driven approach should help us to avoid creating a perfect complex ontology by providing us with a simple one instead, which can be used in various projects right now. Simpler ontology should also help us to create similar resources for other languages and take advantage of it in machine translation. Such project can reuse many existing components that were created for different purposes and projects.

2 WordNet ontologies

A lexicon with information about how words are used and what they mean is a necessary component for any application working with natural language. Ontologies are one of the resources that can provide enough information for those. Ontology is a formal representation of a set of concepts within a domain and the relationship between those concepts. Ontologies can be based on different assumptions, for specific domains and

different purposes. Thus it is very difficult to compare them using objective metrics.

There are several ontologies built for the English language. For smaller European languages, one of the most important general ontologies is Princeton Wordnet (Miller, 1990). It contains many relations (e.g. hypo/hyperonym, is part of) connecting synsets (synonym set) which are equated with 'senses'. Specifically, according to WordNet's on-line glossary, a *sense* is a 'a meaning of a word in WordNet. Each sense of a word is in a different synset'. Princeton WordNet is available under free license also for commercial applications.

EurowordNet (Vossen, 1998) and Balkanet (Christodoulakis, 2004) were projects to localize (and improve) parts of the original version to Central and South East European languages. Thanks to ILI (inter lingual index), it is possible to connect ontologies and use the result as a multilingual dictionary. Unfortunately some of the problems of original WordNet still remain (Hanks and Pustejovsky, 2005), e.g. the assumption that membership in two or more synsets is equivalent to having more different senses. Some of the WordNet senses are indistinguishable from one another by any criterion. Attempt to build a WordNet-like ontology for new language was described in various papers (Pala and Smrž, 2004; Erjavec and Fišer, 2006). Creation of proper synsets and assigning the relations is a time-consuming process that needs expert in this field. One of the most serious problems of the EWN data is their very strict license.

3 VerbaLex

VerbaLex is the lexicon of verb valencies for Czech language, developed at the Faculty of Informatics, Masaryk University. VerbaLex (Hlaváčková and Horák, 2005) combines valency frames and formalism, used in previous projects (Balkanet and Vallex 1.0 (Hajic et al., 2003)), with other relevant information, such as verb aspect, verb synonymy and semantic verb classes based on VerbNet project (Schuler, 2005). VerbaLex contains 10 478 verbs, 21 123 verb senses and 19 360 valency frames. Information in VerbaLex is stored in the form of *complex valency frames* (CVF).

Complex valency frame is designed as a sequence of elements which form a list of necessary grammatical features (e.g. preposition and grammatical case).

opustit:4/leave office:1 (give up or retire from a position)
frame:
AG <person:1>$^{obl}_{who1}$ **VERB ACT**<job:1>$^{obl}_{what4}$
example:
Jarek opustil zaměstnání / Jarek left his job

Example sentence can show us that if "Jarek" has to be the agent (semantic role) then it has to be in nominative (numbered 1) case. Also it has to be a hyponym of person:1 in the WordNet ontology. Thanks to ILI we can have nodes named in English and use words from Czech EuroWordNet.

This notation exported to an XML format allows us to easily process both syntactic and semantic layer of the sentence.

4 Annotation process

Annotation of linguistic data is considered to be a task for experts. This is especially right for those corpora that attempt to cover more layers or structures of a language. Process of annotation is usually described in detail in an annotation manual. As an example, we can take annotation manual for the semantic layer of PDT2.0 which spans tens of pages (Hajič et al., 2005). In last years, we have witnessed several attempts to use crowdsourcing for small parts of linguistic annotation (Munro et al., 2010).

In order to use crowdsourcing we have to find a crowd that exceeds a critical mass. Thanks to services like Amazon Mechanical Turk, this is usually not a problem for widely used languages, such as English. Situation for languages like Czech (10 million speakers) is more complicated as no services of this type are available.

We have decided to involve students. Our annotators are mostly in their first year at the university and they have very limited amount of deeper linguistic knowledge. Our previous experience with student annotators gives us some hope that they can be trained to carry out simple linguistic tasks better than an average crowd-member, though.

We assume that an annotation standard is usually an attempt to approximate several mutually exclusive and contradictory constraints (Jakubíček et al., 2010):

1. **completeness**: the annotation should provide

complete linguistic insight into the particular area;

2. **consistency**: the annotation should be consistent, i. e. same or similar language phenomena should be handled in same or similar ways;

3. **usability**: the annotation should enable straightforward usage in the intended applications;

4. **simplicity**: the annotation should be as simple as possible to make high inter-annotator agreement achievable.

In our experience, most language resources try to find a trade-off among the constraints by prioritizing them in the order given above. They prefer completeness over consistency, and both of them over simplicity.

Following the so-called KISS[1] principle, we are strongly convinced that the reverse order of those constraints represents a much better priority list to be met when building a language resource. Thus, our priorities are:

- **simplicity**: so that annotators do not err too often;
- **usability**: so that the usage of the resource will be straightforward;
- **consistency**: following from simplicity;
- **completeness**: just in case everything is simple, usable and consistent.

Main objection against this new order of priorities can be that consistency is crucial to most NLP application. This applies to using the data both for testing/development and for machine learning. From our perspective, natural language and its semantics is too ambiguos and flexible to be easily and consistenly annotated. We have to face situations where even expert human annotator encounters a possibility of having more than one correct annotation. Inconsistencies between annotators are traditionally resolved by an expert who decides which annotation is correct. Qualified opinion of an expert can improve consistency of annotations but we do not prefered to use other than crowdsourcing methods. As we would like to know that these examples are clear and others are ambiguos for annotators. This can help us to better test applications as we can't expect machines to handle semantic ambiguity better than people and thus testing should be performed mostly on clear cases.

We had attampt to constrain the annotators as much as possible with a simple annotation scheme. Annotators can not add new noun phrases (nodes in ontology) and they have to work only with noun phrases found in source material. As can be seen on screenshot annotators can answer only yes/no. Limiting creativity and working with preprocessed data helps us to increase inter-annotator agreement and (therefore) also consistency.

5 BushBank corpus

BushBank is a concept that extends TreeBanks, which are sets of annotated syntactic trees, by reducing the requirements for unambiguity and making them closer to real language. Like other modern corpora, bushbank usually covers several layers of lingusitic annotation. For this reason, we have decided to use NXT NITE (Carletta et al., 2005), which was developed for multimodal corpora. We do not plan to have a multimodal corpus, but using existing libraries for complex search queries and the XML format persuaded us. On top of this toolkit, we have built our own library which maps elements in the corpus to objects, so that programmers do not need to care about the internal NXT NITE structures or about XML elements.

One of our main objections against existing annotated corpora is the fact that they treat language as an unambigious structure and possible ambiguities are solved by the annotation manual or by expert decision. This leads to a situation when corpora users are not able to determine whether they are handling cases that were easy to determine or cases where even human annotators were not really sure. For various NLP applications, it is crucial to know whether the application can handle correctly at least the clear cases and only later focus on areas which are hard even for humans.

Ambiguity in a BushBank is one of its main advantages. In fact, only the first layer has to be disambiguated. This layer contains marks for sentences and a token for every word in the corpus. We are aware that even on this layer, it is possible to have ambiguities but both simplicity and usability will be corrupted if we introduce ambiguity at

[1] Keep It Simple and Stupid

this level. Currently for the Czech BushBank (as first case-study of bushbank concept) we have the following layers:

1. **tokens**: contains tokens and marks for begin/end of sentence.
2. **morphology**: defines lemma and morphologic tag for tokens.
3. **syntactic structures**: defines short noun phrases, verb phrases, coordination and clauses. This structures uses the token layer.
4. **relations between syntactic structures**: for every short noun phrase we define its dependency parent.

We believe that corpus users should be able to select proper resolution model for their needs and thus they should have access to the existing annotation also in the form of raw data. All our results are easily reproducible and can be reproduced by anyone interested in doing so.

6 Building the Sholva ontology

The Sholva ontology, currently being developed at FI MU, attempts to create a new lexical resource for Czech language which will be free to use for any purpose. Proposed methods and implementation of the tools are also suitable for other languages with similar structure (e.g. Slovak) and we believe that those languages will follow soon. Our ontology should be used primarily for machine translation, syntactic parsing and word sense disambiguation. Application driven extensions should be possible.

We do not intend to create an ontology with dozens of relations and complicated. For our usage, only hypo/hyperonymical relations can be used directly, and basic ontology will not contain any other relations. In EWN, senses of words are numbered, but splitting word senses is also a very subjective task. More importantly, this has no direct relevance to our primary goals. For word sense disambiguation, we need to be able to distinguish various senses of a given word. We believe that for this purpose, the knowledge of path from the root node to any given word created by hyponymical relations defines *word sense*. It is possible that a word will have several hyperonymical relations, but we do not know if they refer to the same or a different thing.

The process of creating an ontology in this style is very similar to corpora annotation. Annotation tool provides users with very simple interface. Annotation tool is web-based and optimized for mobile devices, see screenshots 1 and 2. Each annotator is given a set of around 500 tokens per annotation round. This set can be annotated on iPad device in 5–6 minutes as annotator can only answer by clicking/touching *yes* or *no* buttons.

Each token consists of semantic class (e.g. person:1) and noun phrase for which annotator have to select its validity. Annotator do not have possibility to look at whole original sentence or found examples in corpus. This may look like a step backwards as language resource build on top of corpus are used for quite long time. But it is not. Our noun phrases are from corpus but we want ontology that do not contains figurative language. For figurative language context of noun phrase is very important so we have decided to preventively ignore context at all.

Candidate tokens for annotation are prepared by combination two layers from bushbank: syntactic structures (noun and verb phrases) and relations of syntactic structures and verb valency lexicon. Using bushbank helps us to find only those noun phrases which are mapped to verb as valencies. These noun phrases are then mapped to valency frames of given verb. As each of the slot of VerbaLex valency lexicon points to node in WordNet we are able to suggest potential semantic class for noun phrase.

In later stages of building ontology, we plan to use existing ontology to improve precision of our mapping by using only those valency frames where at least one of the slot will be filled by noun phrase which is already known and can be in expected semantic class.

7 Conclusion

Application driven development that lead us to creation of new ontology means that ontology was created to solve our specific problems. Our main idea is to use this information to better disambiguate mapping of verb valency frames and to test current recall and precision. We expect that succesfull mapping should helps us to improve machine translation dictionary as we will be able to translate valency frames instead of verbs. Additionationaly this mapping can be used to improve anaphora resolutions.

In the future, the project will continue with the annotation of additional resources and we plan to develop methods to use also large corpora that are not annotated in as detailed way as our bushbank. We plan to release our corpus to the research community. Along with that, our linguistic tools and resources will be improved by fixing problems discovered in the process of annotation. We will gladly help to create a similar resource for other languages. We believe that this can be a way for even smaller languages to obtain valuable linguistic resources, using this very low-cost approach.

Acknowledgements

This work has been partly supported by the Ministry of Education of CR within the Center of basic research LC536 and LINDAT-Clarin project LM2010013, by the Czech Science Foundation under the projects P401/10/0792 and 102/09/1842, and by the Ministry of the Interior of CR within the project VF20102014003.

References

J. Carletta, S. Evert, U. Heid, and J. Kilgour. 2005. The NITE XML toolkit: data model and query language. *Language resources and evaluation*, 39(4):313–334.

D. Christodoulakis. 2004. *Balkanet Final Report*. University of Patras, DBLAB. No. IST-2000-29388.

T. Erjavec and D. Fišer. 2006. Building slovene wordnet. In *Proceedings of the 5th International Conference on Language Resources and Evaluation LREC*, volume 6, page 24.

J. Hajic, J. Panevová, Z. Urešová, A. Bémová, V. Kolárová, and P. Pajas. 2003. Pdt-vallex: Creating a large-coverage valency lexicon for treebank annotation. In *Proceedings of The Second Workshop on Treebanks and Linguistic Theories*, volume 9, pages 57–68.

J. Hajič, J. Panevová, E. Buráňová, Z. Urešová, A. Bémová, J. Kárník, J. štěpánek, and P. Pajas. 2005. Anotace na analytické rovině. *Návod pro anotátory. Praha: ÚFaL MFF UK*.

J. Hajic. 1998. Building a syntactically annotated corpus: The prague dependency treebank. *Issues of valency and meaning*, pages 106–132.

P. Hanks and J. Pustejovsky. 2005. A pattern dictionary for natural language processing. *Revue Française de linguistique appliquée*, 10(2):63–82.

Dana Hlaváčková and Aleš Horák. 2005. Verbalex – new comprehensive lexicon of verb valencies for czech. In *Proceedings of the Slovko Conference*, Bratislava, Slovakia.

M. Jakubíček, V. Kovář, and M. Grác. 2010. Through low-cost annotation to reliable parsing evaluation. In *PACLIC 24 Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation*, pages 555–562.

G. Miller. 1990. Five Papers on WordNet. *International Journal of Lexicography*, 3(4). Special Issue.

R. Munro, S. Bethard, V. Kuperman, V.T. Lai, R. Melnick, C. Potts, T. Schnoebelen, and H. Tily. 2010. Crowdsourcing and language studies: the new generation of linguistic data. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 122–130. Association for Computational Linguistics.

Karel Pala and Pavel Smrž. 2004. Building Czech Wordnet. *Romanian Journal of Information Science and Technology*, 7(1-2):79–88.

K.K. Schuler. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon*. Ph.D. thesis, University of Pennsylvania.

P. Vossen, editor. 1998. *EuroWordNet: a multilingual database with lexical semantic networks for European Languages*. Kluwer.

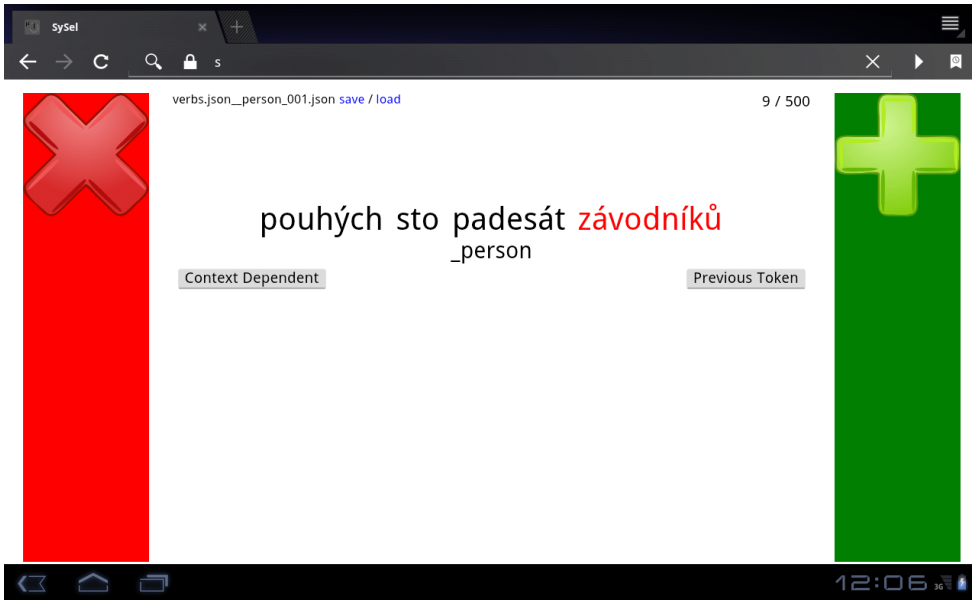

Figure 1: Annotation tool on Android tablet

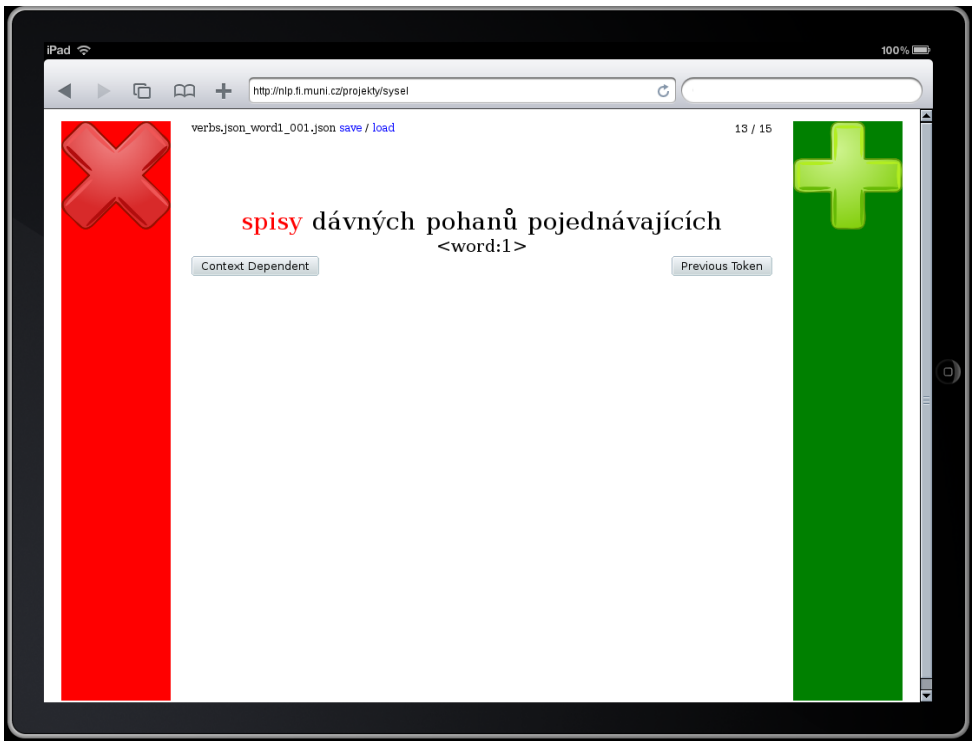

Figure 2: Annotation tool on iPad