Exploring and Extending Czech WordNet and VerbaLex

Zuzana Nevěřilová

Masaryk University, Faculty of Informatics
Botanická 68a, 602 00 Brno, Czech Republic
xpopelk@aurora.fi.muni.cz

Abstract. This paper presents usage of two major, linguist-made lexical resources of Czech language: WordNet and VerbaLex. First, a conversion to RDF was made. Afterwards, a Prolog program was used to analyse Czech language inputs.
In the second part of the article an extension to current VerbaLex is proposed. Possible pitfalls are discussed. In the conclusion, we emphasize the side-effect of this work: an important feedback for authors and administrators of both lexical resources.

Key words: VerbaLex; WordNet; semantic analysis; RDF; Prolog

1 Introduction

Since 2005 a database of verb valency frames is created. This database, VerbaLex [1], has form of frame-based lexical resource: it consist of verb valency frames with slots. Each slot contains two levels of semantic information:

- semantic role, such as *agent*, *patient*, *instrument*
- value restriction in form of bottommost hypernym, specified by literal and sense number in Princeton WordNet [2] (e.g. *person:1*)

Czech WordNet (CZWN) started as part of EuroWordnet [3] project in 1998 and it is still being actively developed.

VerbaLex and CZWN are two large linguist-made resources for Czech language. These resources can be and are expected to be used together thanks to the fact that in CZWN the IDs of synsets are linked to their translations in Princeton WordNet.

This article shows how these resources can be used for semantic analysis of sentences and proposes an extension that can add background knowledge to these sentences. This background knowledge is considered necessary for semantic discourse analysis [4].

For verb frame identification, semantic role assignment and subsequent inference SWI-Prolog and RDF were used.

In the experiments we deliberately omit syntactic analysis of the sentences and use only base form of nouns (singular nominative). We expect that syntactic analysis could improve the results notably. In practice intersection of our results and those of syntactic analysis will be used.

Petr Sojka, Aleš Horák (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2009, pp. 69–74, 2009.

2 Data Formats and the Program

Both CZWN and VerbaLex are stored in their own formats in the form of XML. For the purpose of their connection and inference, both data sources were converted to RDF [5] (in the form of XML). The conversion was done through XSL templates, since it is portable and easy to maintain (in case of slight changes in the structure of the XMLs).

The conversion does not cover all aspects of VerbaLex nor CZWN. For the reasons of reasonable size of the data, some features such as examples, human readable definitions etc. were omitted. In VerbaLex there is no ID for a frame, but during the conversion one is added for each verb frame. The ID consists of one of the lemmata (where czech accents were replaced by capitals), sense number and frame number (generated during the conversion). The ID is in form of URI according to RDF specification [6].

After experiments with RDF reasoners, Prolog with `rdf_db` module was chosen for inference. The advantages of this solution are:

- Prolog is able to work with large data. VerbaLex comes with more than 212 000 RDF triples, CZWN with nearly 100 000.
- It is possible to insert inference rules to the program and not to the data. The most resource-consuming relation is the hyperonymy, because it is a transitive relation. Since RDF is not able to handle transitivity, it would be necessary to use some kind OWL [7] guided with enormous increase of the number of RDF triples. Hyperonymy is handled in the Prolog program and thus the number of RDF triples is final.
- With an appropriate Prolog module, web interface can be made straightforwardly.

3 Finding Semantics through Verb Frames

Since this work does not concern syntactic analysis, almost no grammatical information is available for the analysis. The input is simple: the verb and a list of nouns in their base form (singular nominative).

In our analysis of a sentence, we can identify 3 kinds of bearers of the meaning:

- nouns occuring in the sentence identify *hypernyms* occuring in the verb frame
- *semantic roles* that the nouns play
- the verb frame *structure*, especially the number, semantic role and occupancy of other slots

The output contains the ID of a verb frame and nouns of the list with their semantic roles assigned:

```
?- find_roles('přicestovat',['ministr','zastávka'],FrameID,Roles).
FrameID = 'http://nlp.fi.muni.cz/verbalex#pRicestovat_1_2',
```

```
Roles = [ (ministr, 'AG', kdo1, obl), (zastávka, 'LOC', čeho2, opt)] ;
```

The input: verb *přicestovat* (arrive) and the nouns *ministr* (minister) and *zastávka* (station). The resulting role assignment: minister as AG(ent) and nominative animate (kdo1), obl(igatory) value of the slot, station as LOC(ation) inanimate genitive (čeho2), opt(ional).

3.1 Features, Problems and Solutions

The result of the analysis brings following advantages:

- appropriate verb meaning recognition
- frame identification
- semantic roles assignment
- grammatical information (cases)

It is necessary to keep in mind that the result is a set. In the case above, this set has only one element.

Problems occuring during the analysis can be following:

- verb not found in VerbaLex. This is not expected to occur often, since VerbaLex contains 19 360 valency frames from more than 10 000 verbs [8]. But if this case occurs, the analysis brings no result.
- word from the list not found in CZWN. This occurs almost in every sentence, since CZWN is much smaller than Princeton WordNet. Moreover it does not contain proper names at all. The instant solution is to take subsets of the input set and try to assign as much nouns as possible. A long-term solution consists of improving CZWN and using other resources for proper names.
- no suitable frame for the list of words. In VerbaLex, only *common use* is encoded. In some cases, language users do not follow the common use. This occurs rarely. Most often there are words not related to the verb (e.g. parts of noun phrases) or nouns contained in adverbial phrases. Solution is again to take subsets of input set.
- no suitable hyperonym for a word. This came in sight as the most difficult problem. It seems that there is not much consensus about the bottommost hypernyms in frame slots. For example the verb *koupit* (to buy) has the OBJ(ect) slot value *goods:1*. But the object of buying can be almost every object or even animal. Thus it seems that the value of the OBJ slot should be *object:1*. In this case verb frame will not offer much information.

4 Proposed Extension of VerbaLex

VerbaLex is a frame-based lexical resource. Like other resources, such as FrameNet [9], it contains slots describing typical situations (in this case noun phrases related to the verb), with restriction on their values (in this case WordNet hypernyms).

Contrary to FrameNet, VerbaLex frames are not related together, there is no hierarchy among the frames.

According to [10] it makes sense that frame information should be inherited through type hierarchy. Frame-based representation can be also used to encode additional information not mentioned in the sentences. This underlying knowledge is believed to be very useful in interpreting language. In particular knowledge about causality is very important. Frame-based knowledge representations consist at least of:

- preconditions
- effects
- decompositions

FrameNet, as a representant of large frame-based resources, contains even more types of relations. Proposed extension rests in introducing these three relations to the frames. Prolog program was extended that it supports inference rules.

These inference rules are in form of another RDF (encoded in XML) and related to VerbaLex through RDF IDs. Only information is: type of relation (precondition, effect, decomposition), relation to another frame and mapping between the slots:

```
<proposition rdf:about="#pRicestovat_1_1_effect_1">
  <action rdf:resource="#pRicestovat_1_1"/>
  <effect rdf:resource="#nachAzet_se_1_1"/>
  <mapping>
    <map>
      <from rdf:resource="#AG"/>
      <into rdf:resource="#ENT"/>
    </map>
  </mapping>
  <mapping>
    <map>
      <from rdf:resource="#LOC"/>
      <into rdf:resource="#LOC"/>
    </map>
  </mapping>
</proposition>
```

In this piece of XML the *effect* of *přicestovat* (arrive) is to *nacházet se* (inhere). Mapping is done from AG(ent) to ENT(ity) and from LOC(ation) to another LOC(ation). Note that in the example above the grammatical change occurs on the basis of VerbaLex information. No other information is needed in the inference rule.

With these data program is able to output:

```
?- find_effect('přicestovat',['ministr','zastávka'],FrameID,Roles).
FrameID = 'http://nlp.fi.muni.cz/verbalex#nachAzet_se_1_1',
Roles = [ (ministr, 'ENT', kdo1, obl), (zastávka, 'LOC', čem6, opt)] .
```

The input: verb *přicestovat* (arrive) and the nouns *ministr* (minister) and *zastávka* (station). With the inference rule that *přicestovat* (arrive) *has the effect* of *nacházet se* (inhere): minister as ENT(ity) and nominative animate (kdo1), obl(igatory) value of the slot, station as LOC(ation) inanimate locative (čem6), opt(ional).

Result of inference brings in addition to features mentioned above following:

- new frame identification
- change of roles assignment (AG → ENT)
- change of grammatical information (čeho2 → čem6)

4.1 Problems and Solutions

Main problem of this extension is how to build effectively set of inference rules.

Proposition is to group verbs according to structure of their frames and assign rules depending on which group each verb joins.

For example: LOC(ation) slot with genitive indicates that one of role representants (either AG(ent) or PAT(ient) changes LOC(ation)). In most cases, s/he either starts or stops to be placed in that LOC(ation). Verbs fulfilling this structure are the verbs of motion [1] such as *dorazit*, *přicestovat* (arrive), *dojízdět* (commute), or the verbs of sending and carrying such as *cpát* (crowd), verbs of spatial configuration such as *klesat*, *svažovat se* (slope down).

This grouping can lead to a semi-automatically created inference rules set.

4.2 Introducing New Entities and New Roles to the Discourse

According to [10], knowledge about usual situations in which actions occur is useful for language interpretation. Moreover if these situations are defined, the knowledge reveals new objects that do not have to be mentioned, but exist in the discourse.

For example buying something involves four objects: the buyer, the seller, the object and an amount of money. Even if the money is not mentinoned in discourse, it is contained in it.

Decomposition of buying is:

- buyer gives money to seller
- seller gives object to buyer

Moreover agents in the discourse can play new roles. Every living person can be buyer or seller, but during the act of buying, AG(ent) has the role of buyer (buyer is not a new entity in the discourse, but it is a new role of the entity previously mentioned).

In future work we will concentrate on encoding these new entities and roles to inference rules so they can be used in the discourse semantic analysis.

5 Conclusion

We have introduced Prolog program that is able to analyse verb and nouns occuring in a sentence. The analysis acquire following information:

- valency frame identification
- semantic role assignment
- grammatical information

We have proposed an extension to VerbaLex that can imply new propositions. Main problem is how to build an appropriate set of rules. With this extension we can even introduce new object to the discourse or to assign new roles to the agents previously mentioned. This background knowledge is believed to be useful for language interpretation.

Side-effect of this analysis is that on corpus sentences it offers an important feedback to the authors and administrators of VerbaLex and CZWN. Namely choice of bottommost hypernym in VerbaLex slots can be checked.

Acknowledgements This work has been partly supported by the Ministry of Education of CR within the Center of basic research LC536.

References

1. Hlaváčková, D.: Databáze slovesných valenčních rámců VerbaLex. Master's thesis, Masarykova univerzita, Filozofická fakulta, Ústav českého jazyka (2007).
2. Fellbaum, C.: WordNet: An Electronic Lexical Database (Language, Speech, and Communication). The MIT Press (1998).
3. Vossen, P.: EuroWordNet – a multilingual database with lexical semantic networks (1998).
4. van Dijk, T.A.: Semantic discourse analysis. In: Handbook of Discourse Analysis: Dimensions of Discourse. Volume 2., London, Academic Press (1985).
5. Beckett, D., McBride, B.: RDF/XML syntax specification (2004).
6. Lassila, O., Swick, R.R.: Resource Description Framework, (RDF) model and syntax specification (1999).
7. McGuinness, D.L., van Harmelen, F.: OWL Web Ontology Language Overview (2004).
8. Hlaváčková, D.: Počet lemmat v synsetech VerbaLexu. In: After Half a Century of Slavonic Natural Language Processing, Brno, Czech Republic, Tribun EU (2009).
9. Baker, C.F., Fillmore, C.J.: FrameNet (2009) [Online; accessed 30-July-2009].
10. Allen, J.: Natural Language Understanding (2^{nd} ed.). Benjamin-Cummings Publishing Co., Inc., Redwood City, CA, USA (1995).