Resources for Extending the PolNet - Polish WordNet with a Verbal Component

Zygmunt Vetulani
Adam Mickiewicz University
Faculty of Mathematics and Computer Science, 61-614 Poznań,
ul. Umultowska 87
vetulani@amu.edu.pl

Tomasz Obrębski
Adam Mickiewicz University
Faculty of Mathematics and Computer Science, 61-614 Poznań,
ul. Umultowska 87
obrebski@amu.edu.pl

Abstract
The paper presents the initial, basic step in extension of PolNet (Polish WordNet) with verbs. This step consists in formatting the source data necessary for final computer-assisted creation of verbal synsets including valency information. An algorithm for compiling verb descriptions contained in two human-oriented dictionaries into a computer tractable electronic resource is presented.

1 Credits
The research presented in this paper was partially covered by the on-going Polish Government research grant R00 028 02 "Text processing technologies for Polish in application for public security purposes" (2006-2009) within the framework of Polish Platform for Homeland Security.

2 PolNet – Polish WordNet Project
The long term PolNet project started in December 2006 and reached the maturity stage one year later with the core set of nominal synsets and the main hierarchical relation ISA corresponding to the hyponymy/hyperonymy between words (over 10600 synsets covering env. 18600 different word senses). This work was accomplished in 3 phases. The first phase consisted in elaboration of an algorithm for constructing synsets and relations, as well as in selection of the base concepts to be considered, cf. (Vetulani et al., 2009). The second and the third phases were those of creating synsets and relations. These phases appeared very efficient due to the tools provided by the Czech partners of the PolNet project: the VisDic system used during the second phase and the DEBVisDic used during the third one (Horák et al., 2007).

The algorithm, being highly language independent, allows for building a wordnet from scratch on the basis of a monolingual lexicon with well distinguished word senses and complete lexical and semantic coverage. It is clear that quality of the reference lexicon has direct impact on the quality of the output. In case of Polish, we were in good position due to the existence of high quality multimedia lexicons (cf. e.g. Dubisz, 2006).

In this paper we present the initial, basic step in extending PolNet with verbs. This step consists in proper (manual) formatting of source data necessary for final (software-assisted) synset creation. As initial lexical resource we have chosen two complementary human-user-oriented dictionaries (cf. Sections 3.1 and 3.2 below).

3 Initial Lexical Resources
The two source dictionaries selected as initial resources for PolNet are Uniwersalny Słownik Języka Polskiego (Universal Dictionary of Polish (UDP)) (Dubisz, 2006) and Słownik syntaktyczno-generatywny czasowników polskich (Generative Syntactic Lexicon of Polish Verbs (GSL)) (Polański, 2009).

3.1 UDP
UDP is a typical monolingual dictionary available for on-line consultation. For a given word it provides the description of its inflectional properties as well as the possible senses of this word. For each sense a definition and one or more examples are provided. The dictionary contains approximately 100,000 words. The structure of dictionary entries is as follows (Fig. 1.).
lemma register.
a) «definition1; synonyms»:
  example(s).
b) «definition2; synonyms»:
  example(s).
c) ...
inflectional_information (aspect, inflection_class, endings, derivates...)

Fig. 1. UDP entry structure

For the Polish verb mamić (to delude, to mislead) we find the following entry in the UDP (Fig.2.).
mamić książki.
a) «rozbudzać w kimś próżne nadzieje, zwodzić kogoś fałszywymi pozorami; łudzić, tumanić, manić»:
  Mamili ludzi obietnicami.
b) «działać przyciągająco; wabić, nęcić, manić»:
  Oczy mami mnogość towarów.
  Ameryka mamiła dobrobytem.

Fig. 2. The UDP entry for mamić (to delude) and its approximate English translation

to delude liter.
a) «to cause sb to believe sth that is false, to waken futile hopes in sb, to deceive»:
  They deluded people with promises.
b) «to entice, to lure, to tempt»:
  The abundance of goods deludes people’s eyes.
  America deluded people with welfare.

Fig. 3. GSL entry structure

As a simple example, let us take the entry MAMIĆ. It has several meanings, one of them is represented by the lexicon entry given in Fig. 4.

Syntactic frame is an expression describing arguments of the verb on syntactic level. It may contain operators of alternative (disjunction) and optionality (round brackets in Fig. 4.), so several sentence patterns may be encoded by one such expression. This notation is very compact but at the same time computationally very inconvenient. For the purpose of including syntactic information in the WordNet without making the processing too hard, we decided to expand these expressions into a list of simple patterns with no additional operators.

For each syntactic frame the syntactic_frame_slots are associated with feature-based/descriptor-based semantic classes characterizing the semantic requirement of the verb with respect its arguments.

1 More detailed description may be found in (Vetulani, 2004)
MAŁYCH
I. 'to deceive, to mislead (on purpose), to lead astray'
1. NP⁻¹ → [+Hum]
2. NP⁻¹ → [+Hum]
NP⁻¹ → [+Abstr, -Anim][+Abstr]
NP⁻¹ → [-Abstr, -Anim][+Abstr]
NP⁺¹ → [oczy][wzrok][+Hum]
examples of use omitted

Fig. 4. The GSL entry for mamić (to delude) and its approximate English translation

For each syntactic frame the syntactic frame slots are associated with feature-based/descriptor-based semantic classes characterizing the semantic requirement of the verb with respect its arguments.

In order to express semantic requirements, Polański uses semantic classes of two types:

- feature-based classes e.g. [+Hum], [-Abstr, -Anim]
- descriptor-based classes e.g. [oczy],[wzrok]

Feature-based classes are defined using the following basic semantic features:

<table>
<thead>
<tr>
<th>Feature Based Class</th>
<th>Descriptor Based Class</th>
</tr>
</thead>
</table>

These features and a number of their combinations are enough to express semantic requirements for the major part of verbs, nevertheless, the necessity to use more detailed specifications is quite common. This was the reason for Polański to complete the short list of semantic features with 1600 concepts expressed by common nouns (simple or compound). These nouns will be henceforth called semantic descriptors (for more details cf. (Vetulani 2003)). Semantic descriptors are used to define descriptor-based classes.

It is important to notice that the semantic descriptors proposed by Polański refer either to:

- "concrete entities perceivable by senses and located at any point in time in a three-dimensional space", e.g. roślina (plant), zwierzę (animal), and therefore belong to the category of 1stOrderEntities in the terminology of EuroWordNet (Vossen 2003), or to
- "unobservable propositions that exist independently of time and space", e.g. myśl (thought), pamięć (memory), and therefore belong to the category of 3rdOrderEntities in the terminology of EuroWordNet (Vossen 2003).

All the semantic descriptors used in GSL were included into PolNet. This means that PolNet is sufficiently reach to serve as reference ontology for semantic description of verbs contained in the GSL.

### 4 Algorithm 1: the algorithm for verb encoding

Access to the initial project resources described in Section 3 permits to benefit from a huge amount of manual work done by lexicographers and language engineers to gather the essential syntactic and semantic knowledge about words. The next step involving important investment of manual work was bringing the linguistic data to the format appropriate for further automatic processing and precise enough to obtain a high quality final product (extension of PolNet with verbs).

The Algorithm 1 compiles information from the two source dictionaries into a data structures formally simplified with respect to GSL entries (no optionality, no alternatives) but extended by addition of semantic roles and exhaustive/systematic usage examples. In order to ease further processing, the categorial symbols used in the output forms are simplified with respect to the human-addressed notation of the input data (GSL in particular): no indices, no stratified notation.

#### Algorithm 1

1. Find the entry for V in UDP dictionary
2. On the basis of the UDP make a list [V:1, V:2, ...] of all meanings for V
The algorithm applied to the GSL entry from Fig. 3. (MAMIĆ) will result with the structure presented in Fig. 6. The new structure for MAMIĆ is:

```
MAMIĆ:1
ref Ua
def "książk. rozbudzać w kimś próżne
nadzieje, zwodzić kogoś fałszywymi
pozorami; łudzić, tumanić, manić"
framel
  fref F11
  pat 1 n a "Oszukiwał i mamił nas, żeby
  osiągnąć swój cel"
  pat 2 n a i "Hamili ludzi
  obietnicami(UDP)"
  sem n -> Agent [+Hum]
  sem a -> Experiencer [+Hum]
  sem i -> Instrument [-Abstr,-Anim]+[Abstr]
frame2
  fref F12
  pat 1 n a "Mamili ludzi jego obietnice"
  sem n -> Cause [-Abstr,-Anim]+[Abstr]
  sem a -> Experiencer [oczy][wzrok][+Hum]
```

TO DELUDE:1
ref Ua
def "liter. to cause sb to sb to believe sth
that is false, to waken futile hopes in sb,
to deceive"
framel
  fref F11
  pat 1 n a "He cheated and deluded us in
  order to reach his aim."
  pat 2 n a i "They deluded people with
  promises.(UDP)"
  sem n -> Agent [+Hum]
  sem a -> Experiencer [+Hum]
  sem i -> Instrument [-Abstr,-Anim]+[Abstr]
frame2
  fref F12
  pat 1 n a "His promises deluded us."
  sem n -> Cause [-Abstr,-Anim]+[Abstr]
  sem a -> Experiencer [eyes][glance][+Hum]

Fig. 6. Algorithm 1 output for MAMIĆ (with English translation)

5 Application of the Algorithm 1

The Algorithm 1 have been applied by an experienced lexicographer².

At the present stage, the algorithm was applied to 350 verbs. In this number are in particular the most frequent verbs of general Polish (Przepiórkowski, 2004) as well as verbs selected among the most frequent in the field of public security (Walkowska, 2009) within the project "Text processing technologies for Polish in application for public security purposes" (cf. Section 1). PolNet is used in this project as ontology supporting reasoning in the POLINT-112-SMS system with emulated language competence (Ve-

² Beata Nadzieja, Faculty of Modern Languages, UAM.
tulani et al. 2008). From this list of 421 verbs we obtained ca. 1572 entries (verb senses) in the format shown in Fig. 5. The average number of frames per entry was 1,13 and the average number of patterns per frame was 3,12.

6 Algorithm 2: from set of Structures to a Verb Network

The Algorithm 1 described above resulted in an electronic, fully computer processable lexicon. The Algorithm 2 transforms the lexicon into a network:

- a verbal synset (initially containing a single item) is created for each verb meaning
- each verbal synset is connected with nominal synsets corresponding to semantic restrictions on arguments with relations labelled with roles
- all syntactic patterns and semantic constraints constitute a part of the synset description

It is a step towards full integration of verbs into PolNet. The initial requirement is that all descriptors are included in PolNet. This is already done. Next, correspondence between feature-based classes and nominal synsets has to be established.

A version of the algorithm for building verbal synsets is presented below.

Algorithm 2

For each verb sense V:i
1. Create a synset S with V:i as its unique lexical element
2. Copy the definition of V:i to S
3. Copy all patterns of V:i to S
4. For each semantic constraint of the form
   “sem e → role classes” consider classes;
   for each class in classes create an Intralingual Relation whose type is role and the target is set as follows (depending on class):
   - if class is a feature-based class then make the corresponding synset the target of the relation
   - if class is a descriptor based class then choose appropriate sense for the descriptor and make the synset containing this sense the target synset of the relation

An example of a verbal synset created in accordance with Algorithm 2 for the verb MAMIĆ (to delude) is shown in Fig 7.

7 Conclusion

Creation of a real size application requires always an important effort. Language processing applications involving the modeling of human language competence are an example of practical problems where the final success of a computer system depends on mainly manual work invested in preparation of language data for being computer tractable. The work reported in this paper serves this objective. The data tools we have obtained so far will be used shortly in the POLINT-112-SMS application to be applied in the PolNet-based ontology, and first of all as main tool in the process of extending PolNet with the verbal component.
References


