PolNet - Polish WordNet: Data and Tools

Zygmunt Vetulani, Marek Kubis, Tomasz Obrębski
Adam Mickiewicz University
Faculty of Mathematics and Computer Science, 61-614 Poznań, ul. Umultowska 87
E-mail: {vetulani,mkubis,obrebski}@amu.edu.pl

Abstract

This paper presents the PolNet-Polish WordNet project which aims at building a linguistically oriented ontology for Polish compatible with other WordNet projects such as Princeton WordNet, EuroWordNet and other similarly organized ontologies. The main idea behind this kind of ontologies is to use words related by synonymy to construct formal representations of concepts. In the paper we sketch the PolNet project methodology and implementation. We present data obtained so far, as well as the WQuery tool for querying and maintaining PolNet. WQuery is a query language that make use of data types based on synsets, word senses and various semantic relations which occur in wordnet-like lexical databases. The tool is particularly useful to deal with complex querying tasks like searching for cycles in semantic relations, finding isolated synsets or computing overall statistics. Both data and tools presented in this paper have been applied within an advanced AI system POLINT-112-SMS with emulated natural language competence, where they are used in the understanding subsystem.

1. Introduction

The PolNet project aims at developing a linguistically motivated ontology useful for designing systems with emulated language competence. The project started in 2006 as a part of a larger text processing project (cf. Credits, below). Now it is being developed autonomously as a generic, application independent Polish Wordnet.

PolNet is a linguistically motivated ontology built on the basis of words, which reflect human conceptualization of the world (cf. Gruber, 1993). Ontologies are used to extend the classical logic by addition of a number of basic relations between concepts. In the past, ontologies were created independently of the way concepts are represented in language. Utility of ontologies for development of language technologies results in growth of the interest in ontologies directly connected with the lexical system of a natural language (like Princeton WordNet).

The main idea behind this kind of ontologies is to use words to construct formal representation of concepts. The words linked by the synonimy relations and forming the so called synsets are representations of concepts. Although the general ontologies are supposed to represent some universal conceptualization of the world common to all humans and therefore are language-independent, the WordNets are - by definition - language dependent. The idea of considering WordNets which are created for particular natural languages as ontologies corresponds to the intuition that the structure of a natural language reflects the structure of the world as it is perceived by the language users.

2. Methodology and implementation

We decided to develop PolNet from scratch on the basis of a monolingual lexicon with well-distinguished word senses and with a complete lexical and semantic coverage, rather than by applying the expand model consisting in a translation of some existing wordnet (usually Princeton WordNet serves as such reference resource). Producing a wordnet from scratch means that in order to generate synsets and relations the designer must apply the classifying criteria directly to the linguistic material (e.g. to apply synonymy tests to words in order to qualify them as belonging (or not) to the same synset). We were using criteria inspired by the EuroWordNet project (Vossen, 2003).

It is clear that the quality of the resources used has direct impact on the quality of the resulting wordnet. In case of Polish, we were in good position due to the existence of high quality multimedia lexicons (cf. Dubisz, 2006).

We distinguish several phases of PolNet development. After the first phase which consisted in creation of the project methodological foundations (selection of sources of linguistic data and elaboration of procedures (the algorithm) for creating synsets and relations), we proceeded to the second one resulting with the set of nominal synsets linked by the relations of hypernymy and meronymy. The third phase was that of formal validation of data obtained so far (formal quality check) and creation of the WQuery tool useful for accessing and validating the PolNet data (Kubis 2009a). At present (fourth phase), PolNet is being expanded in order to include verbs (more in (Vetulani and Obrębski, 2010)).

The algorithm (language independent) used for building the core part of PolNet is described in details in (Vetulani et al., 2009b).

The core PolNet composed of noun synsets was built during one year (2007) (phase 1 and 2). The lexicographical work (four PhD students with lexicographical training supervised by one senior lexicographer) consisting in application of the wordnet encoding algorithm (ibid.) amounts to 9.5 man-months

labour resulting with ca. 10,700 synsets. This result was obtained for ca. 10,000 words extracted from:

- IPI PAN Corpus (Przepiórkowski, 2004)

as well as from

- the Generative Lexicon of Polish Verbs (Polański, 1992). We considered the set of semantic descriptors (761 entry words) used for the description of semantic requirements of verbs
- corpora of text for the domain of homeland security terminology (1360 words), emergency situations dialogue corpus (630 words).

The very satisfactory speed of synset encoding was due to the use of the VisDic and DebVisDic tools developed at Brno University (Pala et al, 2007). A simplified example of a synsets formatted under VisDic editor is presented bellow.

```
<SYNSET>
<ID>PL_PK-518264818</ID>
<POS>n</POS>
<DEF>
instytucja zajmująca się
kształceniem</DEF>
<SYNONYM>
 <LITERAL lnote="U1"
 sense="1">szkoła</LITERAL>
 <LITERAL lnote="U4"
 sense="5">buda</LITERAL>
 <LITERAL lnote="U1a"
 sense="1">szkółka</LITERAL>
 ...
</SYNONYM>
<USAGE>Skończyć szkołę</USAGE>
<USAGE>Kierownik szkoły</USAGE>
...
<ILR type="hypernym" link="POL-
2141701467">instytucja
oświatowa:1</ILR>
<RILR type="hypernym" link="POL-
2141575802">uczelnia:1, szkoła
wyższa:1, wszechnica:1</RILR>
<RILR type="hypernym" link="POL-
2141603029">szkoła średnia:1</RILR>
...
<STAMP>Weronika 2007-07-15
12:07:38</STAMP>
<CREATED>Weronika 2007-07-15
12:07:38</CREATED>
</SYNSET>
```

(*szkoła, buda, szkółka...*above are Polish synonims of *school;* the definition *"instytucja zajmująca się kształceniem"* means *"teaching institution"*)

In the Figure 1 we show some hypernymy relations which link the synset PL_PK-518264818 ({szkoła 1, buda 5, szkółka 1,...} with some other PolNet synsets.

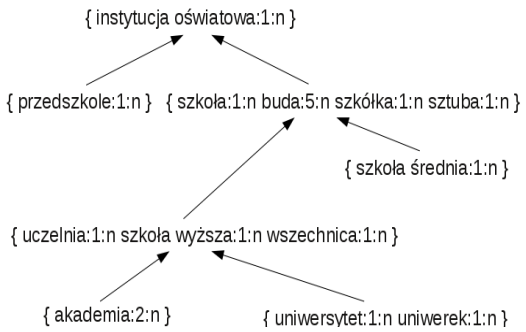

Figure 1. Sample PolNet synsets linked with hypernymy relation

In the following table we collected some numerical data about the PolNet resources obtained so far (computed using WQuery; cf. Section 4.3.).

Number of synsets	**10629**
Number of words	**10939**
Number of senses	**18851**
Monosemous words	7286
Polysemous words	3653
Number of hyponym-hypernym links	12156
Average polysemy including monosemous words	1.72
Average polysemy excluding monosemous words	3.16
Number of holo-part relations	114
Number of top synsets (no hypernym)	42
Number of synsets with more then one hypernym	1427
average depth of a synset in hypernymy relation hierarchy	5.1

Figure 2. PolNet statistics

In the future, we intend to link the PolNet data to the Global Wordnet Grid[1]. The alignment exercise for was completed successfully for some 1200 synsets (nouns) out of the list of 4689 Common Base Concepts proposed as basis for the first version of the Grid. Independently, we have aligned env. 2411 synsets directly to the PWN.

3. Tools

The two main software tools used for PolNet development and validation were: DebVisDic – at the synset encoding stage and WQuery for evaluation and correction.

DebVisDic (Pala et al., 2007) was the main tool used to

[1] www.globalwordnet.org/gwa/gwa_grid.htm (access date: 16.03.2010)

form synsets and to establish relations among them. It provides sufficient editing and browsing facilities, however its searching capacities appeared insufficient for evaluation and error detection purposes. They are limited only to queries by word (or its part), word and sense number or XML element and its content. In order to deal with more complex querying tasks like searching for cycles in semantic relations, finding isolated synsets[2] or computing overall statistics a new software tool was developed, namely the WQuery system.

WQuery is a system based on an artificial language (the WQuery language) designed to query wordnet-like databases. It is a mature system developed since 2007. WQuery proved its usefullness in the POLINT-112-SMS system (Vetulani et al., 2009a) where it is used in creation and composition of frames for SMS messages (Kubis, 2009).

3.1 WQuery system architecture

The architecture of the WQuery system is shown in Figure 3. Its central component is the WQuery language interpreter implemented in Java.

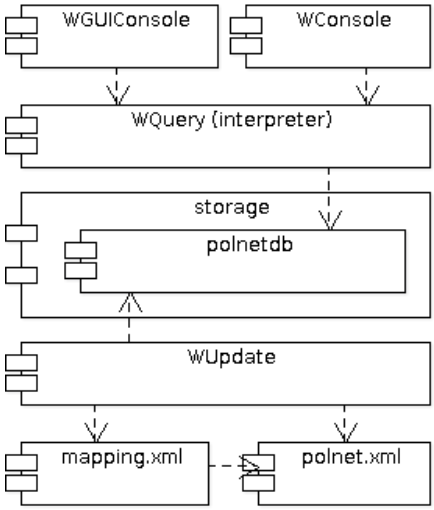

Figure 3. WQuery architecture

WUpdate (cf. Figure 3) is a module responsible for transforming PolNet stored in polnet.xml into a highly optimized internal representation suitable for WQuery interpreter. WUpdate is able to import any wordnet that is stored in a XML document in the Global WordNet Grid format. The transformation process may be optionally controlled by a mapping file (mapping.xml). This file is responsible for renaming specific relations found in the wordnet file (e.g from hypernym to hypernyms), registering relations that are inversions of the other ones (like hyponyms for hypernyms) and relations that are sums of the others (like meronyms relation which joins together several types of meronymy). WGUIConsole and its shell counterpart WConsole are two simple user interfaces. The first one is a window application and the second one is a console application suitable for batch processing.

3.2 The WQuery language

The WQuery system operates on wordnet related terms like synsets, word senses and words. Below, we describe the WQuery language syntax through simple examples. In the following examples *wq>* is the system prompt followed by the query. System answers are displayed in the line(s) below the query.

E.g. to check that PolNet contains the word *butelka* (eng. *a bottle or its content*) one may query:

```
wq> butelka
```

The system answer is:

```
butelka
```

To check that the second meaning of the word *butelka* exists one may ask the following query:

```
wq> butelka:2
{butelka:2n}
```
(eng. *bottle content*)

Finally, to find all synsets that contain *butelka* the following query may be submitted to the system. (Curly brackets in the query have function of the universal quantifier ("all")).

```
wq> {butelka}
{butelka:1:n, flaszka:1:n}
```
(first answer)
```
{butelka:2:n}
```
(second answer)

Datasets[3] may be transformed using dot operator followed by a relation name. In order to find hypernyms of synsets that contain the word *butelka* one may write

```
wq> {butelka}.hypernyms
{opakowanie:2:n}
```
(eng. *wrapping*)
```
{jednostka:4:n, jednostka miary:1:n,
 miara:1:n, miano:2:n}
```
(eng. *measure*)
```
{naczynie:1:n}
```
(eng. *dishware*)

The WQuery system processes, besides the semantic and lexical relations, also relations between different types of data. For example the pos relation associates the part of speech symbols to the synsets. Similarly the glosses relation transforms synsets into their descriptions (character strings).

```
wq> {flaszka}.pos
n
```
(*n* stands for *noun*)

For a relation between the same types of data the transitive closure may be computed by applying the "`!`" operator after the relation name.

```
wq> {butelka}.hypernyms!
{opakowanie:2:n}
{pojemnik:3:n}
```
(eng. *container*)
...

Datasets may be filtered by providing a conditional expression between the square brackets [and]. For example to find all noun synsets in PolNet the

[2] Synsets not related to any other.

[3] A dataset is a multiset of data sharing the same type (for example a multiset of synsets).

following query may be submitted.

```
wq> {}[pos = `n`]
    ...
```

A filter is applied separately to every element of the preceding dataset. The element which is passed to the filter may be referenced using # operator. The following query returns all synsets that contain the word *osoba* except the one that contains that word in its first noun sense (eng. *person*).

```
wq> {osoba}[# != {osoba:1:n}]
{ osoba:2:n } (eng. gram. person)
{ osoba:3:n } (eng. character /dram.)
```

Using # operator the query that finds all noun synsets in PolNet may be reformulated as

```
wq> {}[#.pos = `n`]
    ...
```

Datasets may also be passed as arguments to function calls. In order to check how many synsets that contain the word *butelka* exist in PolNet the following query may be executed.

```
wq> count({butelka})
2
```

Among other useful functions are: distinct, which removes duplicates from datasets; length, which returns the length of the character string and max which returns maximal element (if defined) of a dataset.

3.3 WQuery as a tool for PolNet evaluation

Several evaluation tasks justify usage of a versatile query tool.

Figure 4 presents examples of queries that fulfills such tasks as:

- searching for incorrect data
- searching for data to be improved
- computing the numbers of synsets, senses, words etc.

4. Final remarks

Both data and tools presented in this paper have been applied within an advanced AI system POLINT-112-SMS (Vetulani et al., 2009a) where they are used by the natural language understanding subsystem (NLP Module and Situation Analysis Module). PolNet is particularly useful in construction of the knowledge representation frames. Its extension with a verbal component (operated now, cf. Vetulani and Obrębski, 2009) will permit further, substantial reinforcement of the reasoning competence of the system. Demos of both PolNet and WQuery will be ready for presentation at the LREC 2010. The site www.wquery.org is now under construction It will contain a demo version by the LREC 2010. We intend to distribute free of charge the PolNet for the LREC participants for their non-commercial purposes.

5. Credits

This work has been partly supported by the Polish Ministry of Science and Higher Education, grant R00 028 02 (project „Polish Text Processing Technologies for the Public Security Oriented Applications" within the Polish Platform for Homeland Security).

Query	**Result**
`{}[# in hypernyms!]`	Cycles in hypernymy relation
`{}[pos != `n` and pos != `v`]`	Synsets that have incorrect part of speech symbol assigned
`{}[count(glosses)!=1]`	Synsets with incorrect number of glosses
`{}[max( length(glosses))<20]`	Synsets with too short glosses
`{}[ count(hypernyms)=0 and count(hyponyms)=0]`	Synsets that do not take part in hypernymy relation
`C ount({})`	Number of synstets
`count({}[ count(hypernyms)=0])`	Number of top synsets
`count( ''[count(senses)>1])`	Number of polysemous words
`count(::)/count('')`	Average polysemy

Figure 4. Evaluation related queries

6. References

Stanisław Dubisz (Ed.). 2006. *Uniwersalny słownik języka polskiego PWN*, (*Universal dictionary of Polish*, in Polish)(2nd edition) Warszawa: Wydawnictwo Naukowe PWN.

Karel Pala, Aleš Horák, Adam Rambousek, Zygmunt Vetulani, Paweł Konieczka, Jacek Marciniak, Tomasz Obrębski, Przemysław Rzepecki, Justyna Walkowska. 2007. DEB Platform tools for effective development of WordNets in application to PolNet. In: Z. Vetulani (ed.*). Proceedings of the 3rd Language and Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, October 5-7, 2007, Poznań, Poland.* Wyd. Poznańskie, Poznań : 514-518.

Marek Kubis. 2009. An access layer to a lexical database in POLINT-112-SMS. In: Z. Vetulani (ed.*). Proceedings of the 4th Language and Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, November 6-8, 2009, Poznań, Poland.* Wyd. Poznańskie, Poznań : 437-441.

Kazimierz Polański. (Ed.) 1992. *Słownik syntaktyczno - generatywny czasowników polskich (Generative Syntactic Lexicon of Polish Verbs*, in Polish*)*, vol. I-IV, Ossolineum, Wrocław,1980-1990, vol. V, Kraków: Instytut Języka Polskiego PAN.

Adam Przepiórkowski. 2004. *The IPI PAN Corpus*, IPIPAN, Warszawa.

Zygmunt Vetulani, Jacek Marciniak, Paweł Konieczka, Justyna Walkowska. 2009a. *An SMS-based System Architecture (Logical Model) to Support Management of Information Exchange in Emergency Stuations. POLINT-112-SMS.* Intelligent Information Processing IV (*Book Series:* IFIP International Federation for Information Processing, *Subject collection:* Computer Science), Volume 288/2009, Springer-Boston: 240-253.

Zygmunt Vetulani, Justyna Walkowska, Tomasz Obrębski, Jacek Marciniak, Paweł Konieczka, Przemyslaw Rzepecki. 2009b. An Algorithm for Building Lexical Semantic Network and Its Application to PolNet - Polish WordNet Project. In: Z. Vetulani, H. Uszkoreit. *Human Language Technology. Challenges of the Information Society*, Third Language and Technology Conference, LTC 2007, Poznan, Poland, October 5-7, 2007, Revised Selected Papers. LNAI 5603. Springer Verlag, Heidelberg: 369-381.

Zygmunt Vetulani and Tomasz Obrębski. 2010. Resources for Extending the PolNet-Polish WordNet with a verbal Component, in: Pushpak Bhattacharyya, Christiane Fellbaum, Piek Vossen, *Principles, Construction and Application of Multilingual Wordnets. Proceedings of the 5th* Global Wordnet Conference, Narosa Publishing House: New Delhi, Chennai, Mumbai, Kolkata: 325-330.

Piek Vossen. 2003. Euro WordNet. General Document, Version 3. University of Amsterdam.