Actualising lexico-semantic annotation of *Składnica* Treebank to modified versions of source resources

Elżbieta Hajnicz

Institute of Computer Science, Polish Academy of Sciences

Abstract

In this paper a method of automatic update of the lexico-semantic annotation of *Składnica* treebank by means of PlWN wordnet senses is described. Both resources are under intensive development. The method is based on information which is considered invariant across subsequent versions of the resource.

Keywords: corpus linguistics, lexico-semantic annotation, treebanks, wordnet, Polish

1. Introduction

It is widely acknowledged that linguistically annotated corpora play a crucial role in NLP. There is even a tendency towards their ever-deeper annotation. In particular, semantically annotated corpora become more and more popular, because they have several applications in word sense disambiguation (Agirre and Edmonds, 2006) or automatic construction of lexical resources (McCarthy, 2001; Schulte im Walde, 2006; Sirkayon and Kawtrakul, 2007). An important part of semantically annotated corpora are semantically annotated treebanks.

Some electronic resources are completed and they do not undergo any changes. However, most of them are under more or less intensive development. Therefore, the consistency between various levels of annotation of a corpus and between its annotation and the resources the annotation is based on become an important problem, the more serious, the more complicated is the corpus itself.

Certainly, this problem concerns only corpora annotated manually; new versions of automatically annotated ones can be simply reannotated. However, the manual annotation should be transferred from the "old" version to the "new" one with minimal human effort.

In this paper, we present the automatic transfer of the lexico-semantic annotation of *Składnica* Treebank from its former to later version and from the "old" to the "new" version of Polish WordNet the annotation is based on. The first task consists in finding a terminal in a parse tree to which the lexico-semantic interpretation of a token should be transferred, whereas the the second task consists in updating the lexico-semantic interpretation itself.

Section 2. contains the description of resources being used. In section 3. the method of transferring annotations to new versions of *Składnica* is presented, whereas in section 4. the way how it is updated to new versions of PlWNis discussed.

2. Data resources

2.1. Polish WordNet—*Słowosieć*

Autosemantic tokens in *Składnica* are annotated with very fine-grained semantic interpretations represented by wordnet lexical units. For this sake we used the Polish WordNet (Piasecki et al., 2009), known as *Słowosieć* (English acronym PlWN).

```
124     aparycja        1
136     apteka          1
139     arbiter         2
198     atrybut         3
199     atrybut         1
18382   atrybut         2
19474   arbiter         1
```

Figure 1: The fragment of the table of triples ⟨identifier, lemma, meaning⟩ of PlWN 1.6

A lexical unit (LU) is a string which has its morphosyntactic characteristics and a meaning as a whole. Therefore, it may be an idiom or even a collocation, but not a productive syntactic structure (Derwojedowa et al., 2008). An LU is represented as a pair ⟨lemma, meaning⟩, the last being a natural number. Technically, any LU has also its unique numeric identifier. Each lexical unit belongs to a synset, which is a set of synonyms. Synsets have their unique numeric identifiers as well. A fragment of the table of triples ⟨identifier, lemma, meaning⟩ is presented in Fig. 1.

2.1.1. Named entities in PlWN

Polish WordNet contains some number of named entities, selected rather randomly. They are represented in the same way as common words, by means of lexical units. LUs representing NEs are grouped in synsets as well, since the same object can be identified by means of several NEs (e.g., a full name and its acronym). The only difference is that they are connected by 'type' and 'instance' relations instead of 'hypernym' and 'hyponym'.

The representation of NEs in PlWN is far from satisfactory. Therefore, a table of names (a sort of a gazetteer) was created, in which a list of semantic types represented by PlWN synset identifiers is assigned to every NE lemma. The order of synsets in a list reflects their preference.

The version 2.0 of PlWN is used for semantic annotation of tokens. By contrast, the annotation of named entities was performed using PlWN 1.6.

```
<node nid="48" from="7" to="8" chosen="true">
    <nonterminal>
        <category>formarzecz</category>
    </nonterminal>
    <children rule="n_rz1" chosen="true">
        <child nid="49" from="7" to="8"/>
    </children>
</node>
<node nid="49" from="7" to="8" chosen="true">
    <terminal token_id="morph_6.75-seg">
        <orth>pokoleń</orth>
        <base>pokolenie</base>
        <f type="tag">subst:pl:gen:n</f>
    </terminal>
<plwn_interpretation .../>
</node>
```

Figure 2: A fragment of the representation of a sentence in *Składnica*

```
<plwn_interpretation sem_id="sem_5">
    <plwn_units case_agreement="true"
     polysemy="true">
        <unit luid="sem_5-sv1" chosen="true">
            <lubase>pokolenie</lubase>
            <lusense>1</lusense>
            <luident>20791</luident>
            <synset>2418</synset>
        </unit>
        <unit luid="sem_5-sv2">
            <lubase>pokolenie</lubase>
            <lusense>2</lusense>
            <luident>5921</luident>
            <synset>7789</synset>
        </unit>
    </plwn_units>
</plwn_interpretation>
```

Figure 3: XML representation of a polysemic common word

2.2. *Składnica*

2.2.1. Representation of the syntactic structure

Składnica (Świdziński and Woliński, 2010; Woliński et al., 2011) is a bank of constituency parse trees for Polish sentences taken from the balanced manually annotated subcorpus of NKJP. The whole paragraphs from NKJP were selected. To attain consistency of the treebank, a semi-automatic method was applied: trees were generated by an automatic parser[1] and then selected and validated by humans. The resulting version 0.5 of *Składnica* contains 8227 manually validated trees for 19998 sentences.

The consequence of the applied method is that some sentences do not have any correct parse tree assigned, if *Świgra* has not generated any tree for a particular sentence or no generated tree was accepted as the correct one.

Parse trees are encoded in XML, each parse being stored in a separate file. Each tree node, terminal or nonterminal, is represented by means of an XML `node` element, having the `from` and `to` attributes which determine the boundaries of the corresponding phrase. Terminals additionally contain the `token_id` attribute linking them with corresponding NKJP tokens.

A fragment of the representation of sentence *Taki był u nas zwyczaj od pokoleń.* (*There was such a habit among us for generations.*) in *Składnica* is shown in Fig. 2. `plwn_interpretation` node was added during the semantic annotation of *Składnica* (cf. next section for details).

2.2.2. Representation of lexico-semantic information

PLWN contains lexical units representing three open parts of speech: adjectives, nouns and verbs. Therefore, only tokens belonging to these POS are annotated. On the other hand, only sentences having correct parse trees are annotated.

Semantic annotation is introduced into the XML structure of a parse tree as a new type child element of the element `node`: a terminal node (element `plwn_interpretation`) for common words and a nonterminal node[2] (element `named`) for named entities (Hajnicz, 2013). All corresponding LUs (synsets for named entities) are included, the correct ones having the attribute `chosen="true"` (see Fig. 3 for the noun *pokolenie—generation*).

Apart from LUs having the same lemma as a tagged token, multi-word units, synonyms and hypernyms are used for annotation. The former are used for a more precise annotation, the latter are applied in the case, when the appropriate LUs are absent in PLWN (see Hajnicz, 2014, for details).

Additionally, the root element is augmented with three attributes, `name-plwn_version, sense-plwn_version, final-plwn_version` indicating out which version of PLWN was used for a particular phase of semantic annotation.

3. Transferring annotations to new versions of *Składnica*

Składnica is a resource under development. However, its development is not limited to adding new sentences which have a correct parse tree chosen by linguists. First, wrong decisions are corrected when detected. What is more, the grammar underlying the parser is modified in order to cover a larger set of linguistic phenomena and consequently a larger set of sentences having a correct parse. If the parse tree chosen by the linguist is present in the set of parses generated by the new version of the grammar, it is automatically accepted. Otherwise the manual selection procedure has to be repeated.

[1] *Świgra* parser (Woliński, 2005) based on the revised version (Świdziński and Woliński, 2009) of metamorphosis grammar GFJP (Świdziński, 1992).

[2] The reason for doing this is that named entities are very often multi-word units.

Regardless of the reason of the change of the actual parse tree of a sentence, the procedure of transfer of the semantic annotation linked to a particular node was based on two pieces of information:

- `from` and `to` attributes of a node,
- the lemma connected with any terminal node.

This procedure would be simple, if the segmentation of sentences was always preserved. Unfortunately, in *Składnica* there exist orthographic words that can be represented both by single tokens and by sequences of tokens. The most important reason for that are so-called agglutinates (Przepiórkowski, 2004). For instance, the orthographic words *gdzieś, coś* can be represented as single tokens or as pairs of tokens *gdzie+ś, co+ś*. The other reason are punctuation marks (like hyphen '-' or apostrophe '") that can be included in a token or can constitute a separate token (e.g., *SLD-PSL* in *naprawiając błędy przyjętej przez koalicję SLD-PSL ustawy* (*correcting errors in the act passed by the coalition of SLD and PSL*)[3] erroneously treated as a single token).

Because of that, the procedure of actual semantic annotation transfer is preceded by node alignment (precisely, their `from/to` attributes). For this reason, a boundary `shift` variable is used. Initially, its value is set to 0 and it does not change unless lemmas of corresponding terminals are equal.

Detecting inequality of lemmas starts the alignment procedure. Since the only reason for such inequality is the change of segmentation, two symmetric cases are considered:

1. Splitting one "old" segment into two or three; we seek for identity of the segment next to the "old" one and the segment following the "new" one by two or three. The `shift` variable is increased by 2 or 3, respectively.
2. Joining two or three "old" segments into one; the procedure is symmetric to the previous one.

The above procedure could fail in two cases:

- if more than 3 segments were split/joined into one;
- if two adjacent segments were split or joined.

Both of these possibilities are highly unlikely and were not met in practice.

The remaining problem is what to do if a split segment or one of joined segments was semantically annotated. The following heuristics are used in order to automatically choose the correct semantic interpretation for the maximal number of tokens:

- If more than one of "new" split segments or "old" joined segments are autosemantic, the sentence is sent to a human annotator for reannotation.
- If annotation contains an anaphoric link, it is copied to the corresponding node.
- If an LU with the "new" lemma belongs to the same synset as the previously chosen LU, it is accepted as a correct semantic interpretation of the "new" token.
- If exactly one LU with the "new" lemma belongs to a direct hyponym/hypernym synset of the previously chosen LU or they have a common hypernym, the procedure is the same.
- Otherwise, the sentence is sent to a human annotator for reannotation.

Across *Składnica* 0.5 and *Składnica* 0.6, 130 sentences lost validated trees, 254 sentences acquired validated trees. Validated trees were changed for 1083 sentences, whereas the structure of shared forests was changed for 5966 trees. The segmentation has changed in 9 sentences, whereas the lemmas of tokens were changed in 45 sentences. Therefore, the segmentation is the stable part of *Składnica*.

4. Updating annotations to new versions of PlWN

PlWN has undergone substantial changes during its development, which poses a much more formidable challenge to the task of updating lexico-semantic annotations in *Składnica* than changes in *Składnica* itself do. The changes in PlWN can be classified in the following way:

1. adding a new lemma,
2. adding a new lexical unit for an existing lemma,
3. moving an LU to another synset,
4. changing the sense number of an LU,
5. changing the identifier of an LU,
6. changing the identifier of a synset,
7. deleting an LU,
8. deleting a synset,
9. changing some relations linking lexical units or synsets.

We divide the procedure of updating the lexico-semantic annotation of *Składnica* to a new version of PlWN into two phases: identification of changes in PlWN and introducing them into *Składnica*.

4.1. Identification of differences between two PlWN versions

For two subsequent versions of PlWN, there exists a file coding changes in sense numbering (if there were any). New and old unit representations are separated by comma, each of them consists of a lemma, a POS identifier (1—verb, 2—noun, 4—adjective) and a sense number separated by dots (see Fig. 4).

Units absent from this file were deleted. Using this information, we can also compare old and new version of PlWN in order to find LUs with a modified identifier, as well as new LUs (and whole lemmas) with no counterpart in the old version. For all deleted LUs actually used for annotation, a counterpart that should

[3] *SLD* and *PSL* are acronyms of Polish parties.

```
aparycja.2.1;aparycja.2.1
aparycja.2.2;aparycja.2.2
aranżować.1.2;aranżować.1.1
aranżować.1.1;aranżować.1.2
atrybut.2.1;atrybut.2.1
atrybut.2.3;atrybut.2.2
atrybut.2.2;atrybut.2.3
```

Figure 4: A fragment of a file coding the change of sense numbering between PLWN 1.4 and 1.5

Table 1: LU changes between PLWN versions

		adjectives	nouns	verbs
PLWN 1.8	changed	3 (0.06)	59 (0.06)	18 (0.06)
	moved	399 (7.65)	775 (0.80)	265 (0.84)
	deleted	51 (1.02)	473 (0.49)	86 (0.27)
PLWN 2.0	changed	9 (0.18)	100 (0.10)	18 (0.06)
	moved	579 (11.53)	1316 (1.36)	275 (0.87)
	deleted	99 (1.97)	777 (0.81)	141 (0.45)

replace it in annotation must be determined manually. With this limitation, sense numbers are the stable part of PLWN.

Table 1 shows the number of LUs' changes between PLWN 1.6 and PLWN 1.8 (the upper part) and PLWN 2.0 (the lower part). The row *changed* concerns the change of an identifier, the row *moved* concerns the change of a synset. The percentage of the LUs that have undergone changes is given in brackets. However, the absolute numbers are more important, because they determine the human effort (fortunately, only for deleted LUs).

Next, each LU belonging to both versions of PLWN is checked for moving to another synset. Finally, this information can be used in order to detect changes in the composition of synsets. The following changes are detected:

- deletion of an LU removed from the database,
- deletion of an LU shifted to another synset,
- addition of an LU shifted from another synset.

This information is used to recognise synsets with a new identifier but having the same content, deleted synsets and completely new synsets. Changes which add new LUs only are ignored. Table 2 shows the number of LUs' changes between PLWN 1.6 and PLWN 1.8 (the upper part) and PLWN 2.0 (the lower part). The row *modified* concerns synsets having at least one LU added or deleted.

4.2. Updating the table of names

The table of names used for named entities annotation is a resource handled independently from the actual PLWN. Therefore, it should be updated to a new version of PLWN before updating the semantic annotation of *Składnica*.

For this reason, a list of synsets used in the table is established. The above lists of synsets being changed in PLWN are intersected with this list. This information is used to manually find synsets which composition changes to the extent forcing replacement by other synsets. The resulting list of deleted synsets (between version 1.6 and 2.0) has decreased to 6 elements (36 modified synsets) for nouns (0 and 5, respectively for adjective NE derivatives).

Table 2: Synset changes between PLWN versions

		adjectives	nouns	verbs
PLWN 1.8	changed	31 (1.00)	57 (0.08)	7 (0.03)
	modified	542 (17.44)	1451 (2.07)	396 (1.85)
	deleted	74 (2.38)	659 (0.94)	116 (0.54)
PLWN 2.0	changed	55 (1.77)	103 (0.15)	8 (0.04)
	modified	824 (26.52)	2351 (3.36)	455 (2.13)
	deleted	195 (6.28)	998 (1.43)	141 (0.66)

4.3. Introducing changes detected in PlWN into *Składnica*

The changes of the type 3–6 detected in PLWN are introduced into the existing structure of the XML file. The modification is marked by the addition of a new attribute `update` having value `sense, unit` or `synset`.[4]

The method of maintaining changes of the type 7 depends on whether the deleted unit was chosen in the particular context and whether the unit supposed to replace the deleted one was present in the previous version of PLWN. First, the deleted unit gets the attribute `update="deleted"`.[5] Next, if the deleted unit was the chosen one, the attribute `chosen` changes value to `old`, whereas the suggested unit gets the attribute `chosen="true"`. Finally, if the last one was already present, it is added with the attribute `update="close"`, whereas a new element gets the attribute `update="added"`. Figure 5 contains the modified interpretation of an NE *Ministerstwo Spraw Wewnętrznych* (*Ministry of Home Affairs*).

The changes of the type 8 are only concerned with named entities interpreted by synsets and are introduced analogously.

New lexical units are used for updating the semantic interpretation of tokens annotated by synonyms or hypernyms due to the absence of adequate units in the source version of PLWN. Therefore, for each token annotated in that way, new LUs with a corresponding lemma are checked whether they are synonyms or hyponyms of the current rough annotation (for synonymy, only direct hyponyms are considered). The LU closest to the rough one in hypernymy hierarchy is selected (if there exists more than one appropriate unit, the one with the lowest sense number is chosen). The procedure of assigning attributes is analogous to the standard case.

[4]The attribute `update` consists of a list of values.

[5]No element is deleted from the file.

```
<named name_id="named_105.27-s_n6">
   <nambase>Ministerstwo Spraw Wewnętrznych
   </nambase>
   <nkjp_type type="orgName"/>
   <plwn_units part="head" case_agreement="Full"
               polysemy="true">
      <unit luid="n6-sv1" status="auto"
            chosen="old" update="deleted">
         <lubase>ministerstwo</lubase>
         ..........................
      </unit>
      <unit luid="n6-sv3" chosen="true"
            update="close">
         <lubase>ministerstwo</lubase>
         ..........................
      </unit>
   </plwn_units>
</named>
```

Figure 5: XML representation of an updated NE

5. Conclusions

In this paper a method of automatic update of the lexico-semantic annotation of a treebank by means of wordnet senses was presented. Both the treebank and the wordnet are under intensive development, which is by no means additive. The method utilises all information that is invariant across two versions of relevant resources in order to match corresponding nodes in the two versions of *Składnica* and corresponding LUs and synsets in the two versions of PLWN.

The updates to new versions of *Składnica* and PLWN are performed independently and separately.

What is important, the human intervention is limited to the rare cases of segmentation change in *Składnica* and LUs deletion in PLWN. It is indispensable in order to make the update reliable and error-free. The only heuristic element is the update of annotation of tokens originally tagged using synonyms or hypernyms, but the risk of an error is minimised by the fact that the original semantic annotation is always preserved.

Acknowledgements This research is supported by the POIG.01.01.02-14-013/09 project which is cofinanced by the European Union under the European Regional Development Fund.

References

Agirre, Eneko and Philip Edmonds (eds.), 2006. *Word Sense Disambiguation. Algorithms and Applications*, volume 33 of *Text, Speech and Language Technology*. Dordrecht, the Netherlands: Springer-Verlag.

Derwojedowa, Magdalena, Maciej Piasecki, Stanisław Szpakowicz, Magdalena Zawisławska, and Bartosz Broda, 2008. Words, concepts and relations in the construction of Polish WordNet. In A. Tanacs, D. Csendes, V. Vincze, C. Fellbaum, and P. Vossen (eds.), *Proceedings of the Global WordNet Conference*. Seged, Hungary.

Hajnicz, Elżbieta, 2013. Mapping named entities from NKJP corpus to *Składnica* treebank and Polish WordNet. In M. A. Kłopotek, J. Koronacki, M. Marciniak, A. Mykowiecka, and S. T. Wierzchoń (eds.), *Proceedings of the 20th International Conference on Language Processing and Intelligent Information Systems*, volume 7912 of *LNCS*. Springer-Verlag.

Hajnicz, Elżbieta, 2014. Lexico-semantic annotation of *Składnica* treebank by means of PLWN lexical units. Accepted for publication in GWC 2014 Proceedings.

McCarthy, Diana, 2001. *Lexical Acquisition at the Syntax-Semantics Interface: Diathesis Alternations, Subcategorization Frames and Selectional Preferences*. PhD thesis, University of Sussex.

Piasecki, Maciej, Stanisław Szpakowicz, and Bartosz Broda, 2009. *A Wordnet from the Ground Up*. Wrocław, Poland: Oficyna Wydawnicza Politechniki Wrocławskiej.

Przepiórkowski, Adam, 2004. *The IPI PAN corpus. Preliminary version*. Warsaw, Poland: Institute of Computer Science, Polish Academy of Sciences.

Schulte im Walde, Sabine, 2006. Experiments on the automatic induction of German semantic verb classes. *Computational Linguistics*, 32(2):159–194.

Sirkayon, Chaloemphon and Asanee Kawtrakul, 2007. Automatic lexico-semantic acquisition from syntactic parsed tree by using clustering and combining techniques. In *Proceedings of the International Workshop on Intelligent Systems and Smart Home (WISH 2007)*, volume 4743 of *LNCS*. Springer-Verlag.

Świdziński, Marek, 1992. *Gramatyka formalna języka polskiego*. Rozprawy Uniwersytetu Warszawskiego. Warsaw, Poland: Wydawnictwa Uniwersytetu Warszawskiego.

Świdziński, Marek and Marcin Woliński, 2009. A new formal definition of Polish nominal phrases. In Małgorzata Marciniak and Agnieszka Mykowiecka (eds.), *Aspects of Natural Language Processing*, volume 5070 of *LNCS*. Springer-Verlag, pages 143–162.

Świdziński, Marek and Marcin Woliński, 2010. Towards a bank of constituent parse trees for Polish. In P. Sojka, A. Horák, I. Kopeček, and K. Pala (eds.), *Proceedings of the International Conference on Text, Speech and Dialogue TSD 2010*, volume 6231 of *LNAI*. Springer-Verlag.

Woliński, Marcin, 2005. An efficient implementation of a large grammar of Polish. In Z. Vetulani (ed.), *Proceedings of the 2nd Language & Technology Conference*. Poznań, Poland.

Woliński, Marcin, Katarzyna Głowińska, and Marek Świdziński, 2011. A preliminary version of Składnica — a treebank of Polish. In Z. Vetulani (ed.), *Proceedings of the 5th Language & Technology Conference*. Poznań, Poland.