Generalizing the EM-based semantic category annotation of NP/PP heads to wordnet synsets

Elżbieta Hajnicz

Institute of Computer Science, Polish Academy of Sciences
ul. Ordona 21, 01-237 Warsaw, Poland
Elzbieta.Hajnicz@ipipan.waw.pl

Abstract

This paper contains an adaptation of EM-based NP/PP heads semantic categories disambiguation to entire wordnet senses. First, the preparation of a corpus to be semantically annotated and the wordnet on which the annotation is based are presented. Next, the process of semantic annotation is discussed. Finally, its results are evaluated.
Keywords: corpus linguistics, word sense disambiguation, wordnet, Polish

1. Introduction

The main goal of our work is to enrich the valence dictionary of Polish verbs by adding semantic information. This information may be represented by means of general wordnet semantic categories of nouns or by synsets from the entire net. The plain syntactic valence dictionary is a collection of predicates (here: verbs) provided with a set of verb frames. Verb frames consist of syntactic slots that represent phrases occurring in the corresponding position in a sentence. Thus, our goal is to provide syntactic slots (here: NPs/PPs) with a list of appropriate semantic categories for the corresponding nouns.

In order to automatically acquire semantic information for a syntactic valence dictionary, we need a large treebank where all NP/PP semantic heads are semantically annotated. In (Hajnicz, 2009) we presented an application of the EM selection algorithm to select the most probable semantic categories for each NP/PP head in a clause. In this paper a generalisation of the procedure for the entire wordnet hypernymy hierarchy is presented.

Our problem intersects with the Word Sense Disambiguation (WSD) task (Agirre and Edmonds, 2006). However, contrary to an typical WSD task, we are interested in the most general sense of a noun (i.e., including senses of its hypernyms) that is adequate for a particular context (a clause).

2. Data resources

Our main resource was the IPI PAN Corpus of Polish written texts (Przepiórkowski, 2004), referred to as Kipi. From this corpus, we selected a small subcorpus, referred to as SemKipi, containing 195 042 sentences. Selected sentences contain at least one verb from a preselected set of verbs. Details of the SemKipi creation can be found in (Hajnicz, 2009).

Sentences from SemKipi were parsed with the *Świgra* parser (Woliński, 2004, 2005) based on the metamorphosis grammar GFJP (Świdziński, 1992). The parser was provided with the valence dictionary prepared especially for our task, based on Świdziński's (1994) valence dictionary augmented with automatically created valence dictionary by Dębowski (2007). The dictionary entries of verbs preselected for the experiment were carefully elaborated.

Next, parsing trees of particular clauses were identified in parsing trees of each sentence and reduced to their flat forms representing only arguments of a verb (i.e., the subject and complements included in corresponding valence frames). As a result, we obtained *reduced parses* of a clause composed of a verb and a set of slots.

Świgra tends to produce large parse forests. The number of reduced parses of a sentence is much smaller than the number of entire parse trees, the more so as we have considered only actual arguments of verbs (without adjuncts). They were disambiguated by means of the EM selection algorithm proposed by Dębowski (2007) for the task of creating a syntactic valence dictionary.

The whole process is presented in (Hajnicz, 2009).

In order to prepare an initial sense annotation for NPs/PPs semantic heads (which would be later automatically disambiguated), we used the Polish WordNet (Derwojedowa et al., 2007, 2008a,b), called *Słowosieć* (English acronym PlWN). PlWN is a network of lexical-semantic relations, an electronic thesaurus with a structure modelled on that of the Princeton WordNet (Fellbaum, 1998) and those constructed in the EuroWordNet project (Vossen, 1998). Polish WordNet describes the meaning of a lexical unit (LU) of one or more words by placing this unit in a network which represents such relations as synonymy, hypernymy, meronymy, etc. For the construction of the semantic valence dictionary only the part of PlWN describing relations between nouns was needed. We have focused on the synonymy represented by synsets and on hypernymy represented by a directed acyclic graph.

In the present experiment the version of PlWN containing 15870 LUs representing 11270 nouns and collected in 11780 synsets was used. Synsets were linked by 12550 hypernymy relation arcs. Each LU and each synset has its unique identifier (a natural number).

category	top synsets	category	top synsets	category	top synsets
act	action	communication	code	phenomenon	nature
	activity		communication		phenomenon
animal	animal		etnolect	plant	plant
artifact	concept		information		plant part
	medium		sound	possession	business
	net	event	event		economics
	thing	feeling	feeling		possession
attribute	detail	food	drink	process	process
	feature		food	quantity	degree
body	body	group	fauna		quantity
	corpse		flora		value
	gene		group	relation	connection
	meat		set		domain
	organ	location	container		institution
	tissue		line		relation
cognition	cognition		part		team
	destiny		place	shape	shape
	exercise		space	state	state
	idea	motive	duty	substance	substance
	intelligence		goal		sbst state
	psyche		position		sbst type
			solution	time	course
		object	object		period
		person	person		phase

Table 1: The set of tops in the hypernymy hierarchy correlated with the predefined set of semantic categories in Polish WordNet

Before the entire net was constructed, the set of LUs was divided according to the predefined set of 25 general *semantic categories*. The process of the manual creation of the net was performed to the large extent w.r.t. this division. Thus, all synsets contain LUs of the same category. To a large extent this concerns hypernymy as well (most exceptions concern synsets with several hypernyms). In Table 1 the list of LUs representing synsets positioned on the top of the hypernymy hierarchy together with the corresponding semantic categories.

In our experiments we use wordnet data transformed to simple tables assigning LUs lemmas, semantic categories and synsets they belong to to their identifiers and hypernymy hierarchy between synsets identifiers.

There exists another Polish wordnet (Vetulani et al., 2007). Since we do not use the internal structure of wordnet data, we can easily adapt it to our algorithms. The only problem would be caused by manually prepared data used for evaluation.

3. Semantic annotation

In (Hajnicz, 2009) the process of semantic annotation of verb arguments by means of semantic categories was presented. The EM selection algorithm (Dębowski, 2007) was adapted for the WSD task. Its three versions were proposed and compared: *EM-whole* treating each frame as a whole, *EM-indep* based on the assumption that senses of arguments occur in a clause independently and *EM-incr* adding slots incrementally to the most probable subframes of a particular length. In this paper the application of these algorithms to the entire PLWN structure is discussed.

Therefore, we start with assigning a list of synsets that contain an NP/PP head lemma. However, in contrast to the usual WSD task, we are interested in the most general sense of a noun adequate in a particular context. Thus, the list is extended with all the hypernyms of its elements. An example of a clause together with a valence schema selected for it, a corresponding reduced parse and lists of senses assigned for the arguments is presented in (1).

(1) `% 'Wspomniała pani, że mężczyzna widzi w kobiecie anioła.'`
(You have mentioned that a man sees an angel in a woman.)

```
<widzieć np:acc np:nom prepnp:w:loc>
4-9   widzieć   aff:fin:sg:_:ter
      [4-5:np:mężczyzna:sg:nom:m1:ter::
                5995 6047 6776,
      6-8:prepnp:w:kobieta:loc::
                6047 6129 6776,
      8-9:np:anioł:sg:acc:m12:ter::
                66 67 5908 6045 6047 6771 6778]
```

The clause *mężczyzna widzi w kobiecie anioła* contains two NPs *mężczyzna* (*a man*) and *anioł* (*an angel*) and one PP *w kobiecie* (*in a women*) predicated by the verb *widzieć* (*to see*). The syntactic frame chosen for the clause is presented (in <> brackets) together with the corresponding reduced parse. Syntactic information about arguments is augmented with a list of corresponding synsets' identifiers. Translations of LUs belonging to those synsets can be found in Ta-

ble 2. The clause is not ambiguous: a single semantic category person is assign to each noun. However, due to the hypernymy hierarchy, lists of synsets are multi-element even in this simple case.

syns. id.	list of LUs of a synset
66	angel, good person
67	angel, good spirit
5908	supernatural being
5995	man
6045	being (creature)
6047	person
6129	woman
6771	person positively judged
6776	person w.r.t. sex
6778	person w.r.t. his/her features

Table 2: Lists of lexical units belonging to synsets having identifiers presented in (1)

Next, we split the reduced parse into syntactic-semantic valence frames. Thus, we obtain tuples in which every NP/PP has only one category assigned. All pronouns obtain an artificial sense pron, represented by artificial synset 0 with no hyponyms and hypernyms. The disambiguation process consists in selecting (using the EM algorithm) the most probable frames. So, the reduced parse of the sentence (1) after splitting transforms into the $3 \times 3 \times 7 = 63$ frames, some of them listed in (2). The one selected by the *EM-indep* algorithm is marked by + symbol. Observe that this is the most general sense `person` in the case of every slot.

(2)
```
% 'Wspomniała pani, że mężczyzna widzi
    w kobiecie anioła.'
<widzieć np:acc np:nom prepnp:w:loc>
acc:  66,    nom:  5995,  w_loc:  6047
acc:  66,    nom:  5995,  w_loc:  6129
acc:  66,    nom:  5995,  w_loc:  6776
acc:  66,    nom:  6047,  w_loc:  6047
acc:  66,    nom:  6047,  w_loc:  6129
acc:  66,    nom:  6047,  w_loc:  6776
acc:  66,    nom:  6776,  w_loc:  6047
acc:  66,    nom:  6776,  w_loc:  6129
acc:  66,    nom:  6776,  w_loc:  6776
acc:  67,    nom:  5995,  w_loc:  6047
acc:  67,    nom:  5995,  w_loc:  6129
acc:  5908,  nom:  5995,  w_loc:  6047
acc:  5908,  nom:  5995,  w_loc:  6776
acc:  6045,  nom:  5995,  w_loc:  6047
acc:  6045,  nom:  6047,  w_loc:  6047
acc:  6047,  nom:  5995,  w_loc:  6129
acc:  6047,  nom:  6047,  w_loc:  6047  +
acc:  6771,  nom:  5995,  w_loc:  6047
acc:  6771,  nom:  6776,  w_loc:  6129
acc:  6778,  nom:  5995,  w_loc:  6047
acc:  6778,  nom:  6776,  w_loc:  6776
```

All hypernyms of each synset always co-occur with it. Thus, the frequency of synsets increases according to hypernymy relation. Hence, if a synset S and its hypernym S_H have the same highest probability for a particular slot of a particular valence frame of a verb, than no hyponym of S_H being a hypernym of S (including S_H itself) appears in this slot in the same context. Therefore, the tuples of the less general synsets $\langle \check{S}_1, \ldots, \check{S}_n \rangle$ are finally selected from the results of the entire EM algorithm.

The method of using the relationship between a predicate and its argument in order to disambiguate the sense of the latter (by means of maximal relative entropy) was proposed by Resnik (1993, 1997). However, he considered only one argument at once, whereas we disambiguate the whole predicate-argument structure of a clause. The secondary difference is that Resnik assigns only the actual senses of words, whereas we accept their hypernyms. Nevertheless, this is easy to change in both methods.

4. The experiment

4.1. Manually annotated data for an evaluation of the algorithm

In order to evaluate the algorithms, a small subcorpus of SemKipi was syntactically and semantically annotated by a group of linguists. 240 sentences for each of 32 preselected verbs were selected randomly from SemKipi.[1] The linguists performed three different tasks:

1. a correction of morphosyntactic tagging (tagger errors),
2. a division of sentences into phrases, i.e., pointing out their boundaries and syntactic and semantic heads,
3. an assignment of a single PlWN semantic category to each noun in a sentence.[2]

We have selected sentences for manual annotation before SemKipi was parsed. Unfortunately, only 43% of manually annotated sentences were accepted by *Świgra* (the problem of coverage of *Świgra* on sentences from SemKipi is discussed in Hajnicz and Woliński (2009)). In order to enlarge the test set and to minimise the influence of errors resulting from the preprocessing phase we extended the set of automatically preprocessed sentences with manually annotated ones. If a sentence belongs to both sets, the manually annotated version was chosen. Manually annotated semantic categories of nouns were certainly deleted.

The results of the manual annotation were transformed to the format of *Świgra* post-processing (presented in (1)).

4.2. Efficiency of the algorithm

In order to disambiguate semantic categories, the algorithms have to reduce the number of categories

[1] More precisely, only single-verb sentences were chosen in this sampling.

[2] During manual annotation, the entire PlWN net was not available.

algorithm	semcats		synsets		selected		transformed	
source	1.715	1.840	6.472	6.543	——	——	1.838	1.962
EM-indep	1.002	1.002	1.037	1.042	1.005	1.005	1.003	1.003
EM-whole	1.036	1.042	1.417	1.478	1.067	1.076	1.044	1.051
EM-incr	1.040	1.046	1.343	1.393	1.058	1.066	1.038	1.045

Table 3: Efficiency of the algorithms

algorithm	data set	c-corr	n-corr	acc	prec	rec	F
EM-indep	semcat	60.40	77.63	71.98	77.63	78.01	77.82
	synset	57.49	75.55	71.12	75.44	76.04	75.74
	selected	57.52	75.59	71.23	75.59	76.03	75.81
EM-whole	semcat	61.36	76.53	70.86	75.44	79.70	77.51
	synset	58.42	74.05	69.60	72.56	78.35	75.34
	selected	58.63	74.37	70.17	73.22	78.32	75.69
EM-incr	semcat	61.74	76.57	71.06	75.56	79.04	77.68
	synset	58.30	74.37	69.91	73.09	77.95	75.44
	selected	58.46	74.64	70.42	73.69	77.92	75.75

Table 4: Results of evaluation of the algorithms

assigned to a noun. We call this feature the efficiency of the algorithm.

In Table 3 we present the mean of the number of semantic categories assigned to occurrences of nouns by all algorithms, including source data. The column *semcats* contains data for the algorithms run on semantic categories, whereas the column *synsets* contains data for the algorithms run on entire wordnet synsets. The results of selection of less general synsets are presented in column *selected*, and the results of transformation of selected synsets to their semantic categories are shown in column *transformed*. Observe that the mean of semantic categories obtained after transformation of source data is a bit larger than the mean of categories directly assigned. The reason is that a synset and its hypernym could be differently categorised.

Each column is divided in two, the first concerning all senses and the second calculated without pronouns, which always have the single sense, namely pron. This leads to the decrease of the mean. The experiment presented in (Hajnicz, 2009) was performed with pronouns distributed among all 25 semantic categories, hence the mean calculated for all nouns is larger than calculated without pronouns.

The median is always 1, even for source data. This does not concern source data for synsets, for which the median is 5.

Efficiency of all the algorithms is high, especially when they are performed on synsets. The best one is *EM-indep*, about one order of magnitude better than the other two algorithms.[3]. Observe that *EM-whole* is more efficient for semantic categories whereas *EM-incr* is more efficient for synsets. Note also that selection of less general synsets decreases the mean about 1 order of magnitude.

4.3. Evaluation of the algorithm

For the sake of evaluation a small subcorpus HandKipi of SemKipi containing 5634 simple (single verb) sentences manually annotated with verb arguments boundaries and syntactic and semantic heads and semantic categories of nouns (cf. Hajnicz, 2009). As a consequence of limited manual annotation in HandKipi, we cannot evaluate actual results of tuples of synsets selection, since we do not have data to compare with. Instead, we reduce synset annotation to corresponding semantic categories annotation. Thus, we can only appraise whether we gained or lost some knowledge.

In Table 4 the results of evaluation of all the algorithms are presented. Again, *semcat* means running algorithm on semantic categories, *synsets* means running the algorithm on wordnet synsets, whereas *selected* means evaluating results after less general synsets selection. *c-corr* means correctness calculated for whole clauses (i.e., all slots should have properly assigned senses) and *n-corr* means usual correctness calculated for single nouns.

The results of evaluation are very similar for all the algorithms. The best results shows *EM-indep*, except clause correctness and recall, which is coherent with Table 3. *EM-incr* is a bit better than *EM-whole*. All the algorithms show the best results, when they are performed on semantic categories (where the task is easier); selecting less general synsets helps a bit. This is not really important result; the reason is that the linguists used only the semantic categories of actual nouns in annotated sentences, so semantic categories of their hypernyms worsen the evaluation results.

[3]Since the mean is always less than 1.5, the comparison is made only for its fraction part, the more so as the best possible efficiency is 1.

5. Conclusions

In the paper we presented an adaptation of EM-based NP/PP heads semantic categories disambiguation to entire wordnet senses. The results were evaluated after transformation to semantic categories of selected synsets. The results obtained for synsets were about 2 percentage points worse. This is probably the cost of more precise information.

The best results were obtained by the *EM-indep* algorithm. However, it selects top synsets in 97.9 % cases, whereas *EM-whole*— in 84.8 % cases and *EM-incr* in 87.4 % cases. Thus, evaluation on sentences manually annotated with entire synsets can give different and more reliable results.

For the experiment presented in (Hajnicz, 2009) the correctness (calculated only for semantic categories) was substantially better. However, the experiment was performed with pronouns distributed among all semantic categories. Moreover, the set of categories assigned to each noun has changed during the entire net construction, which could influence the manual annotation process as well. Thus, these results are incomparable.

Acknowledgements

This paper is a scientific work supported within the Ministry of Science and Education project No N N516 0165 33

References

Agirre, E. and Edmonds, Ph. (Eds.) (2006). *Word Sense Disambiguation. Algorithms and Applications*, volume 33 of *Text, Speech and Language Technology*. Dordrecht, The Netherlands: Springer-Verlag.

Derwojedowa, M., Piasecki, M., Szpakowicz, St. and Zawisławska, M. (2007). Polish WordNet on a shoestring. In: *Data Structures for Linguistic Resources and Applications: Proceedings of the GLDV 2007 Biannual Conference of the Society for Computational Linguistics and Language Technology*. Universität Tübingen, Tübingen, Germany.

Derwojedowa, M., Piasecki, M., Szpakowicz, St., Zawisławska, M. and Broda, B. (2008a). Words, concepts and relations in the construction of Polish WordNet. In: Tanacs, A., Csendes, D., Vincze, V., Fellbaum, Ch., Vossen, P. (Eds.) *Proceedings of the Global WordNet Conference*. Seged, Hungary.

Derwojedowa, M., Szpakowicz, St., Zawisławska, M. and Piasecki, M. (2008b). Lexical units as the centrepiece of a wordnet. In: Kłopotek, M.A., Przepiórkowski, A., Wierzchoń, S.T., (Eds.) *Proceedings of the Intelligent Information Systems XVI (IIS'08)*, Challenging Problems in Science: Computer Science. Zakopane, Poland: Academic Publishing House Exit.

Dębowski, Ł. (2007). Valence extraction using the EM selection and co-occurrence matrices. In: arXiv.

Fellbaum, Ch. (Ed.) (1998). *WordNet — An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Hajnicz, E. (2009). Semantic annotation of verb arguments in shallow parsed Polish sentences by means of EM selection algorithm. In: Marciniak M., Mykowiecka, A. (Eds.) *Aspects of Natural Language Processing*, volume 5070 of *LNCS*. Springer-Verlag, pp. 211–240.

Hajnicz, E. and Woliński, M. (2009). How valence information influences parsing Polish with *Świgra*. In: Kłopotek, M.A., Przepiórkowski, A., Wierzchoń, S.T., Trojanowski K. (Eds.) *Recent Advances in Intelligent Information Systems*, Challenging Problems in Science: Computer Science. Warsaw, Poland: Academic Publishing House Exit.

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

Resnik, Ph. (1993). *Selection and Information: A Class-Based Approach to Lexical Relationships*. Ph.D. thesis, University of Pennsylvania, Philadelphia, PA.

Resnik, Ph. (1997). Selectional preference and sense disambiguation. In: *Proceedings of the ACL Workshop on Tagging Text with Lexical Semantics, Why, What and How?*. Washington, DC.

Vetulani, Z., Walkowska, J., Obrębski, T., Konieczka, P., Rzepecki, P. and Marciniak, J. (2007). PolNet — Polish WordNet project algorithm. In: Vetulani, Z. (Ed.) *Proceedings of the 3rd Language & Technology Conference*. Poznań, Poland.

Vossen, P. (Ed.) (1998). *EuroWordNet: a multilingual database with lexical semantic network*. Dordrecht, Holland: Kluwer Academic Publishers.

Woliński, M. (2004). *Komputerowa weryfikacja gramatyki Świdzińskiego*. PhD thesis, Institute of Computer Science, Polish Academy of Sciences, Warsaw.

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

Świdziński, M. (1992). *Gramatyka formalna języka polskiego*. Rozprawy Uniwersytetu Warszawskiego. Warsaw: Wydawnictwa Uniwersytetu Warszawskiego.

Świdziński, M. (1994). *Syntactic Dictionary of Polish Verbs*. Uniwersytet Warszawski / Universiteit van Amsterdam.