Weaving sloWNet using window-based co-occurence features

Darja Fišer,* Maciej Piasecki, † Bartosz Broda†

* Department of Translation, Facult of Arts, University of Ljubljana
Aškerčeva 2, SI-1000 Ljubljana, Slovenia
darja.fiser@ff.uni-lj.si
† Department of Artificial Intelligence, Institute of Informatics, Wroclaw University of Technology
ul. Wybrzeze Wyspianskiego 27, 50-370 Wroclaw, Poland
{bartosz.broda, maciej.piasecki}@pwr.wroc.pl

Abstract

This paper presents the first results of using statistical methods and linguistically annotated corpus data to extract lists of semantically similar words that are then incorporated into an existing wordnet for Slovene. The approach was originally developed for Polish but is attractive for other languages as well because, apart from a large corpus, it requires minimal NLP tools and resources, and can therefore be easily applied to any language that is still lacking an extensive wordnet or a similar semantic lexicon. Another important advantage of the adopted approach is that it relies on real linguistic evidence harvested from a corpus, yielding a linguistically sound organization of the vocabulary. As all the previous approaches used for the construction of Slovene wordnet were transfer-based and relied on the English Princeton WordNet, the encouraging results obtained in the presented experiment will be a welcome complement to the existing semantic network.

Spletanje sloWNeta na podlagi informacij o sopojavljanju besed v korpusu

V prispevku predstavljamo prve rezultate raziskave, v kateri smo z uporabo statističnih metod in jezikoslovno označenih korpusnih podatkov izluščili sezname semantično podobnih besed, ki smo jih nato vključili v wordnet za slovenščino. Pristop je bil prvotno razvit za poljščino, vendar je privlačen tudi za druge jezike, saj zanj razen obsežnega korpusa potrebujemo minimalna jezikovnotehnološka orodja in vire, zato ga je enostavno uporabiti tudi za jezike, za katere obsežen wordnet ali podoben semantični leksikon še ne obstaja. Druga pomembna prednost uporabljenega pristopa pa je, da temelji na izpričani jezikovni rabi, pridobljeni iz korpusa, ki se nato kaže v jezikovno utemeljeni organizaciji besedišča v izdelani semantični mreži. Glede na to, da so vsi naši dosedanji pristopi za izdelovo slovenskega wordneta celotno strukturo prevzeli iz Princeton WordNeta, ki je bil izdelan za angleščino, bodo spodbudni rezultati, dobljeni s pričujočo metodo, koristno dopolnjevali obstoječo semantično mrežo.

1. Introduction

sloWNet, a wordnet for Slovene, has been developed in a number of steps, taking advantage of several types of available bi- and multilingual language resources, such as bilingual dictionaries, parallel corpora and Wikipedia (Fišer and Sagot, 2008). All these approaches have in common that they take over the structure of Princeton WordNet (Fellbaum, 1998), the oldest and most extensive existing wordnet that was developed for English, and find Slovene equivalents for the same set of concepts.

The work presented in this paper tackles the problem from a completely different angle as it extracts all the relevant lexico-semantic information from a single resource, the largest Slovene reference corpus Gigafida (Logar Berginc and Šuster, 2009), yielding language-motivated lists of semantically related words and a linguistically sound organization of the vocabulary. The aim of this paper is to adapt the wordnet expansion algorithms, originally developed for Polish, to Slovene in order to test whether they work for another language as well. With the analysis of the first results we also wish to outline further refinements and enhancements of the approach for future work on fully automated methods of wordnet expansion for Slovene.

This paper is structured as follows: in the next section we present related work. Then, we focus on the resources and tools that were used in the experiment. In Section 4 we give an overview of the experimental setup, evaluate and discuss the results. We then conclude the paper with some final remarks and ideas for future work.

2. Related work

The task of extending a wordnet with additional literals or synsets consists of two parts: first, word pairs of sufficient semantic relatedness need to be extracted from a large corpus, and then they need to be attached to the most appropriate place in the existing semantic network.

Automatic methods for the extraction of semantically related words from corpora fall into two main frameworks: pattern-based (Hearst, 1992) and those that follow the Distributional Hypothesis (Harris, 1968). The pattern-based approaches rely on a list of lexico-syntactic patterns in which two lexical units frequently occur in an identifiable lexical semantic relation, e.g., hypernymy (Pantel and Pennacchiotti, 2006). On the other hand, the distributional- based approaches assume that the similarity of distributions of some lexical units across different lexico-syntactic contexts is evidence of their close semantic relation. The stronger the similarity, the closer the meanings of the lexical units are. Unlike pattern-based approaches, which are limited only to the words that co-occur in a particular pattern, distributional-based techniques can be used for any word pair. Because high recall is an important desideratum in the work presented in this paper, we have opted for the latter.

Many measures of semantic relatedness have been proposed (cf. Ruge, 1992; Lin and Pantel, 2002; Weeds and Weir, 2005). They all share the starting point, which is the construction of a coincidence matrix of co-occurrences of lexical units (rows) and their lexico-syntactic contexts (columns) from a large corpus.

Their main differences between them are the following:

(1) how contexts are defined,
(2) how raw frequencies are normalized, and
(3) how the final value of the measure is calculated.

We have experimented with several different settings reported in literature in our previous work (cf. Piasecki and Broda, 2007; Broda and Piasecki, 2008), and are using the best-performing settings in this work (see Section 3.3).

Once lists of highly semantically related words have been generated, they need to be attached to the most appropriate positions in the existing semantic network. Most known taxonomy induction methods utilize only the existing hypernymy structure in incremental wordnet expansion. Several machine-learning methods have been used to induce a taxonomy from hypernym-hyponym pairs, such as decision trees (Witschel, 2005) or k-nearest neighbors (Widdows, 2003) for a limited set of domains of concrete and frequent nouns. In their seminal paper, Snow et al. (2006) propose a probabilistic wordnet-expansion method based on a probabilistic model of the taxonomy which reports promising results that however were not reproduced successfully in a reimplementation of their algorithm (see Piasecki et al., 2012a).

The approach used in this paper goes beyond the related work in three respects. First, in our previous work (Piasecki et al., 2012a), the wordnet hypernymy structure is perceived as intrinsically interlinked to other wordnet relations. Thus, we aim at utilizing all different types of links in the expansion of Slovene wordnet as well. Second, the algorithm is based on the assumption that the relation extraction method produces some noise in the results, so we cannot identify the exact place (synset) for a new lemma as such but an area (a wordnet subgraph). And last, contrary to a rich body of the related work, we do not assume any shape of the lexical semantic network, but we try to build it in a way that faithfully reflects the language data.

3. Resources and tools used

3.1. Gigafida

The Gigafida corpus is a 1.15 billion word reference corpus of Slovene and is as such currently the largest and most extensive text collection of Slovene (Arhar Holdt et al., 2012). It has been developed within the national project Communication in Slovene (2007-2013) and contains texts of various types and genres such as literary texts, newspaper articles and Internet contents that were published between 1995 and 2011. The corpus has been split into paragraphs and sentences, tokenized, part-of-speech tagged and lemmatized, so that is readily available for use via a concordancer as well as for NLP applications.

3.2. sloWNet

sloWNet is a concept-based semantic lexicon in which nouns, verbs, adjectives and adverbs are grouped into sets of synonyms (synsets) which are then organized into a hierarchical network with lexical and semantic relations, such as hyper- and hyponymy, antonymy, meronymy etc. The concepts that synsets represent are defined with a short gloss and usage examples while most synsets also have a domain label and a mapping to the SUMO/MILO ontology.

sloWNet is based on a Princeton WordNet that was originally developed for the English language (Fellbaum, 1998). Slovene equivalents for synsets were obtained automatically by leveraging existing bi- and multilingual resources, such as a bilingual dictionary, a multilingual parallel corpus and Wikipedia (see Fišer and Sagot, 2008). Recently, a large-scale extension of sloWNet has been achieved by training a maximum entropy classifier in order to determine appropriate senses of translation candidates extracted from heterogeneous bilingual resources (see Sagot and Fišer, 2012a). In addition, automatic detection of candidate outliers has been performed within the framework of distributional semantics by comparing the immediate neighborhood of literals in sloWNet and their contexts in a reference corpus (see Sagot and Fišer, 2012b) with the goal of eliminating noise from the automatically generated resource.

The most recent version of sloWNet has 82,721 literals, which are organized into 42,919 synsets. Apart from single words sloWNet contains many multi-word expressions and proper names as well. Nouns are still by far the most frequent, representing more than 70% of all synsets. While 66% of all the literals in sloWNet are monosemous, their average polysemy level is 2.07.

The methodology of sloWNet construction has three important implications that we try to address in this work:

(1) The resource is based on a semantic network originally produced for a foreign language, so it might be biased towards the organization and distinction of senses typical of English and therefore inadequately reflects the semantic inventory of Slovene.

(2) Slovene equivalents for synsets were harvested from several already available language resources of limited coverage, which is why we were able to obtain equivalents only for some synsets while the rest are still empty, leaving gaps in the network.

(3) Due to automatic generation of synsets, word-sense disambiguation was not perfect, resulting in noisy synsets that have a negative impact on applications using sloWNet, and should therefore be eliminated as far as possible in the shortest possible time.

3.3. The SuperMatrix system for Distributional Semantics

SuperMatrix is a system for semantics analysis of text, especially aimed at supporting automatic acquisition of lexical semantic relations from large corpora (Broda and Piasecki, 2008). The main functionality of *SuperMatrix* is related to the automated construction of Measures of Semantic Relatedness (MSRs) on the basis of a corpus, and testing them on the basis of a wordnet. An MSR is a function that takes a pair of words and returns a value, which describes how closely semantically related the two words are. MSR construction follows a typical blueprint: corpus preprocessing, co-occurrence matrix construction, matrix filtering and transformation, and row similarity computation.

Corpus preprocessing depends on the available language tools. However, for morphologically rich languages lemmatization is a minimal requirement for obtaining a useful MSR. In addition, a corpus parsed by a shallow parser or a dependency parser is a good basis for the construction of a highly accurate MSR, i.e., an MSR which assigns higher values for pairs of lemmas linked by one of the lexico-semantic relations, e.g. synonymy, hyper-/hyponymy, holo-/meronymy and other relations described in wordnets.

Data collected from a corpus contain a lot of statistical noise, e.g. very low frequencies, accidental co-occurrences due to errors produced by language tools, thus the stored data must be filtered and transformed before they can be used for similarity calculations. In addition, raw frequencies produce skewed results, which is why several weighting algorithms have been implemented in *SuperMatrix*. Our previous experiments show that the Point-wise Mutual Information (PMI) measure (Lin and Pantel, 2002) produces the best results. *SuperMatrix* can also reduce dimensions of a matrix using, for example, singular value decomposition. Finally, a vector similarity measure is applied to the matrix in order to obtain a ranked list of similar lemmas. *SuperMatrix* offers most well-known similarity measures but it has been shown that the simple cosine measure produces best results in most cases.

The system also supports an automated evaluation of the selected MSR using synonymy tests that are automatically generated from wordnet, called *Wordnet-Based Synonymy Test* (WBST). The test is described in details in (Piasecki et al., 2009) but the procedure is quite straightforward. Each test item consists of a question word selected from the wordnet data, its synonym (the correct answer) taken from the same synset (or its direct hypernym in the case of singleton synsets including only the question word) and *k* distractors (words taken form other synsets). The task is to select the most related word to the question word among the presented candidates using only the MSR value. For example, for word *svet (council)* one has to choose between *gomolj* (tuber), *izvirnost* (originality), *odbor* (committee) – the correct answer – and *odobravanje* (approval).

3.4. WordnetWeaver

WordnetWeaver is a tool that extends the wordnet editing system called *WordnetLoom* (Piasecki et al., 2012b) with an automated wordnet expansion facility. It utilizes the results of the Activation Area Attachment Algorithm (AAAA) that generates suggested attachment positions for new lemmas, not yet present in wordnet. A suggested attachment is a synset to which a new lexical unit for the given new lemma can be added as a synonym or linked via a lexical or semantic relation, such as hypo- or hypernymy. Moreover, as all automated methods for the extraction of the lexico-semantic relations produce some errors, attachment points in *WordnetWeaver* are presented in the context of attachment areas – subgraphs of the wordnet hypernymy graph such that each synset of the selected subgraph express a strong enough semantic relation to the new lemma according to AAAA.

WordnetWeaver then presents top-scored suggestion in a visual, graph-based editor and enables their verification, correction as well as free manual editing of the wordnet structure. Contrary to other automated wordnet construction methods mentioned in Section 2, the aim of AAAA is to generate suggestions for lexicographers, who make the final wordnet expansion decisions, not to expand the wordnet fully automatically. Thus, AAAA is intentionally set up for slight sense over-generation in order to increase the coverage. The refinement of AAAA that would allow fully automated wordnet expansion is still an open research question.

The input to AAAA are sets of triples: *<l1, l2, w>,* where *l1* s a new lemma and *l2* a lemma already in wordnet. They are linked with a lexico-semantic relation according to a corpus-based relation extraction method, and *w* is the weight assigned to the pair by the given method. We refer to such a set of triples as a knowledge source (KS). AAAA does not assume a probabilistic interpretation of the weights and can work with any set of knowledge sources of any types, produced by any method. Each KS can also have an assigned weight, e.g. expressing the KS accuracy obtained from manual inspection of a sample.

Taking triples from the desired KSs, the AAAA algorithm is composed of two steps. First, the *semantic fit* between the input lemma *l1* and each synset *X* in a wordnet is calculated on the basis of the KSs and the neighborhood of *X*. And then, connected subgraphs *(activation areas)* of the lexical-semantic network are identified, (for details see Piasecki et al., 2012a, Broda et al., 2011).

AAAA has so far been successfully applied to the development of the Polish wordnet (plWordNet) on a practical scale (Piasecki et al., 2009). Also, an automated evaluation of the AAAA performance on Princeton WordNet (Fellbaum, 1998) has been performed (Broda et al., 2011).

4. Experimental setup

The application of the AAAA algorithm to a new language is limited only by the available language resources and corpus processing tools. The minimum requirements are: a large enough corpus and a means for constructing an MSR from it. For morphologically rich languages, Part-of-Speech tagging and lemmatization is also very useful.

In this initial experiment on Slovene wordnet extension with *WordnetWeaver,* we have limited our work to the most frequent single-word nouns, i.e. nouns that occurred at least 1,000 times in the Gigafida corpus. There were 36,026 such nouns, 8,981 of which are already in sloWNet. This was a pragmatic decision in order to examine the first results as quickly as possible and make any necessary changes for future large-scale experiments. But the selected setting is not a limiting factor of the algorithm as such as most of the methods developed for Polish were aimed at low- frequency data (see Piasecki et al., 2009). On the other hand, the results for very frequent words should be better due to the statistical nature of applied methods.

The corpus was PoS-tagged and lemmatized already. It was then converted to a simple plain-text format. In addition, sloWNet had to be converted to the plWordNet XML format for use in *WordnetWeaver*. Apart from that, no other changes were required, which is a great advantage of the tools that were initially developed for Polish because this means that they can be used with other resources and for other languages with relatively little effort.

4.1. Extracting semantically related words

The measure of semantic relatedness is the most fundamental knowledge source for AAAA as it has good coverage (i.e. it provides similarity values for every pair of lemmas that are frequent enough in the corpus), and facilitates the discovery of lexical-semantic relations between words. In comparison to a KS that contains pairs of semantically related lemmas extracted with manually constructed patterns, which has a much higher precision than MSR, the coverage of the pattern-based KS is much lower as only a limited number of pairs can be found in the corpus.

As work on dependency parsers for Slovene is still on-going and we wanted to avoid additional manual work required for pattern-based approaches in this preliminary work, the MSR was constructed with a simple window-based approach. That is, target lemmas are described by all the other content lemmas (nouns, adjectives, verbs, adverbs) co-occurring in a small text window (3 lemmas before and after the target lemma), stopping at paragraph boundaries.

Since there is no *a piori* best method for MSR development and several are implemented in SuperMatrix, we selected the best-performing one with WBSTs based on the existing part of sloWNet. We generated questions with three detractors and a correct answer. On the 20,308 generated questions we achieved the best results for PMI weighting extended with the discounting factor and cosine similarity function (Lin and Pantel, 2002). MSR chose the correct answer in 72.37% of all the questions in WBST.

4.2. Attaching the words to sloWNet

The most straightforward adaptation of AAAA to sloWNet requires importing sloWNet to the *WordnetWeaver* scheme and a preparation of knowledge sources.

We have prepared two KSs based on MSR. The first one is based on the similarity lists for lemmas. That is, for each lemma l_x we compute 20 most similar lemmas l_y using the above described MSR. This KS then takes the form of pairs $<l_x, l_y, msr(x,y)>$, where $msr(x,y)$ is a value of MSR between the two lemmas.

The other KS uses *bi-directional* similarity lists. It is a subset of the above knowledge source with additional filtering. For l_x the pair $<l_x, l_y, msr(x,y)>$ is included only if there is also a pair $<l_y, l_x, msr(y,x)>$ among the 20 most similar items for l_y.

4.3. Evaluation of the results

WordnetWeaver and AAAA were designed to help a linguist in expanding an existing wordnet structure with new lemmas. Thus, the evaluation of the algorithm's performance should focus on this practical aspect. In order to gain a comprehensive insight into the performance of the adopted approach, we perform the results both automatically and manually.

4.3.1. Automatic evaluation

For automatic evaluation of the results, we follow the evaluation methodology proposed by (Broda et al., 2011). The idea of the evaluation is simple: first, we remove some literals from the existing sloWNet structure; then we run AAAA for those literals and see how close to the original place in sloWNet (along hyper-/hyponymy paths) the removed literals were re- attached by the AAAA. Ideally, we would like to remove all occurrences of one lemma in sloWNet at a time and then reattach it, in order to alter sloWNet structure as little as possible, but this is computationally very expensive. Thus, we remove a package of 50 lemmas at a time. For evaluation purposes, we randomly selected a sample of the 1,000 nouns meeting the frequency threshold that was also set to 1,000 (see Section 3).

Several evaluation strategies are possible, each giving a different perspective on the algorithm performance (Broda et al., 2011). From the lexicographers' point of view, the algorithm performs well if there is at least one correct suggestion that is relatively close to the proper place in a wordnet structure, i.e., the *closest path* strategy. For a given lemma, this method only checks the attachment of the closest path to one of the lemma's original position in the wordnet. On the other hand, the *strongest* supported strategy evaluates only the highest-ranked suggestion provided by the wordnet expansion algorithm. The last strategy we use evaluates *all* the propositions returned by the algorithm.

Table 1 presents the results of the described evaluation methodology for all three strategies. The *acceptable distance* to the original place was set to 6 by the lexicographers during the construction of plWordNet (Piasecki et al., 2009). The distance is measured on the hypo-/hypernymy and mero-/holonymy graphs with the exception that we can only traverse one edge of mero-/holonymy (as this relation can take us to completely unrelated parts of the wordnet very quickly).

Dist.	Closest[%]	Best[%]	All[%]
0	15.0	5.9	3.7
1	19.7	13.9	4.6
2	19.0	13.9	6.0
3	11.7	8.2	4.9
4	8.1	9.0	5.3
5	5.5	6.4	6.8
6	0.2	0.7	0.8
Σ	79.2	57.9	32.2

Table 1: Results of the automatic evaluation procedure for sloWNet expansion.

The achieved results are significantly lower than for Polish (Broda et al., 2011), which was expected as we have employed much simpler and less precise, window-based MSR, and we did not used additional, pattern-based KSs. On the other hand, the results are encouraging as for almost 80% of the words the algorithm suggested at least one correct place for attachment. Also, the correct attachment places are mostly close to the original place in

the wordnet structure (i.e., the results are shifted towards closer distances than 6). AAAA provided a suggestion for 94% of words from the random sample and found 29.6% of word senses for each word on average.

4.3.2. Manual evaluation

For a more qualitative insight into the results, we also performed a manual evaluation on 100 random lemmas included in the automatic evaluation. In manual evaluation, 5 highest-ranking attachment suggestions were checked for each lemma, amounting to 500 candidate-attachment pairs.

The evaluated lemmas were first categorized into monosemous or polysemous. Based on the attachment suggestions for polysemous lemmas, we checked whether our algorithm was able to detect only one of its senses or more. Next, we tried to label each attachment suggestion with one of the 10 lexico-semantic relations: *synonymy, hypernymy, hyponymy, holonymy, meronymy, co-hyponymy, co-meronymy, antonymy, close, vague*, or *no relation*. The *no relation* label is intended for clear errors of the algorithm. The *close* label is used for cases where the candidate-attachment pair is clearly semantically related but the relation type is not found in the current version of sloWNet (e.g. *Occupation-Place* such as *pošta-poštar* [post-postman], *Activity-Occupation* such as *učenje-učitelj* [teaching-teacher]). The *vague* label, on the other hand, is used for cases where the candidate-attachment pair is in a more loose associative relation that will probably not be encoded in wordnet (e.g. same semantic field such as *politika-debata* [politics-debate]).

Overall, the results of manual evaluation are very encouraging as no cases were found where all the attachment suggestions for a lemma would be completely unrelated. What is more, only 1 out of 100 lemma received no better attachment suggestion than a vague association, and an additional 1 got at best a closely related one. On the other hand, as many as 38 lemmas had no erroneous attachment suggestions, which means that the lexicographers who are responsible for selecting the best attachment candidates will be presented with very little noise that would slow down their work.

Category	Freq.	%
synonym	22	4.40%
hypernym	74	14.80%
hyponym	9	1.80%
holonym	9	1.80%
meronym	12	2.40%
antonym	1	0.20%
co-hyponym	40	8.00%
co-meronym	2	0.40%
closely related	171	34.20%
vaguely related	50	10.00%
unrelated	110	22.00%
total	500	100.00%

Table 2: Frequency counts of association candidates per relation type.

As Table 2 shows, almost 34% of the suggested association candidates can easily be labeled with one of the standard lexico-semantic relation types from wordnet. By far the most frequent one is the hypernymy relation that was selected in almost 15% of the cases. There were quite a lot of co-hyponymy (8%) and synonymy (4%) attachments as well while the rest of the relations were much more rare. A further 34% of the suggestions were very closely related to the lemmas, 10% were loosely associated to them while 22% of the association candidates were not related at all to the lemmas they were assigned to.

When analyzing the semantic nature of the randomly selected lemmas in the evaluation sample, we observe that 62% of them are monosemous and 38% polysemous. This is very similar to the polysemy level of nouns in the latest version of sloWNet, where 66% of the literals are monosemous. A single sense prevailed for 58% of the otherwise polysemous lemmas in the evaluation sample, while association candidates refer to different senses in 42% of the cases. This is a well-known phenomenon of distributional semantics where a Zipfian distribution of senses in the corpus causes skewed context vectors of polysemous words, which are thus heavily biased towards the most frequent sense in the corpus.

Table 3 shows frequency counts of semantic categories that appeared at least once among the association suggestions per lemma. Because we counted all the relation types that were suggested for each lemma, and a single lemma could have suggestions belonging to a single category or up to five different categories, the total count is more than 100. Hypernymy and co-hyponymy are still the most frequent in this setting, suggested for 60% and 28% of the lemmas, respectively. Both are more frequently suggested for monosemous nouns, while polysemous ones have more suggestions for synonyms, hyponyms, holonyms, meronyms and co-meronyms. Polysemous nouns contain a slightly higher number of erroneous attachment candidates and a much higher number of vaguely and closely related suggestions than polysemous ones. Interestingly, the polysemous nouns for which only one sense was detected by the algorithm, contain the least noise and vague association candidates.

Cat.	Mono.	Poly.			Σ
		1 sense detected	>1 sense detected	Σ poly	
	62	22	16	38	100
syn	11	5	3	8	19
hyper	40	13	7	20	60
hypo	3	1	2	3	6
holo	4	2	3	5	9
mero	4	5	2	7	11
anto	1	0	0	0	1
co-hypo	19	6	3	9	28
co-mero	1	1	0	1	2
close	51	16	2	18	69
vague	22	4	5	9	31
error	37	11	10	21	58

Table 3: Frequency counts of lemmas with at least 1 association suggestion per category.

5. Conclusions

In this paper we presented the first results of applying *WordnetWeaver* to Slovene data in order to extend Slovene wordnet. The approach uses statistical methods to extract lists of semantically similar words from a large reference corpus of Slovene, and then identifies the part of the wordnet hierarchy these words should be attached to. Automatic and manual evaluations of the results show that the algorithm was successfully ported to a new language and is already useful in its most basic setting. However, the state-of-the-art results for Polish suggest that further improvements of measures of semantic relatedness are still possible, for example by using a constraint-based approach, a dependency parser, and testing more measures with more parameters. Similarly, the attachment algorithm could further be improved by optimizing parameters of the algorithms, for example by using meta-heuristics like in (Kłyk et al., 2012), and providing additional knowledge sources, such as pattern-based lists of semantically related word pairs.

In the future, we wish to investigate methods that would enable us to extend the current functionality of the attachment algorithm to expand sloWNet fully automatically, requiring no human intervention for reaching the final decision where to add a new word in wordnet. A somewhat different but very interesting area of research would be to adapt the attachment algorithm to be able to use corpus data in order to analyze the semantic network in sloWNet that is based on Princeton WordNet and find suspicious areas in the network that does not correspond to the linguistic evidence harvested from the corpus and should therefore be improved.

Acknowledgments

We would like to thank Tomaž Erjavec for performing conversions of sloWNet into the plWordNet XML format and of the Gigafida corpus into plain text.

6. References

Š. Arhar Holdt, I. Kosem and N. Logar Berginc. 2012. Izdelava korpusa Gigafida in njegovega spletnega vmesnika. In *Proceedings of 8th Eighth Language Technologies Conference IS-LTC-12*. Ljubljana, Slovenia.

Z. S. Harris. 1968. *Mathematical Structures of Language*. Interscience Publishers, New York.

M. A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceeedings of COLING-92*. Nantes, France, pp. 539–545.

Ł. Kłyk, P. B. Myszkowski, B. Broda, M. Piasecki and D. Urbansky. 2012. Metaheuristics for Tuning Model Parameters in Two Natural Language Processing Applications. In *Proceedings of the 15th AIMSA conference*, Varna, Bulgaria.

C. Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

P. Pantel and M. Pennacchiotti. 2006. Espresso: Leveraging generic patterns for automat- ically harvesting semantic relations. In *Proceedings of COLING-ACL-06*, Sydney, Australia, pp. 113–120.

G. Ruge. 1992. Experiments on linguistically-based term associations. *Information Processing and Management* 28(3), pp. 317–332.

Lin, D. and Pantel, P. 2002. Concept discovery from text. In Proceedings of *COLING-02*, Taipei, Taiwan, pp. 577–583.

M. Piasecki and B. Broda. 2007. Semantic similarity measure of Polish nouns based on linguistic features. *Business Information Systems 10th International Conference, Volume 4439 of Lecture Notes in Computer Science*, Springer.

B. Broda and M. Piasecki. 2008. SuperMatrix: a General Tool for Lexical Semantic Knowledge Acquisition. *Speech and Language Technology conference, Volume 11 of Lecture Notes in Computer Science*, Springer, pp. 239–254.

M. Piasecki, S. Szpakowicz and B. Broda. 2009. *A Wordnet from the Ground Up*. Oficyna Wydawnicza Politechniki Wrocławskiej, Wrocław.

R. Snow, D. Jurafsky and A. Y. Ng. 2006. Semantic taxonomy induction from heterogenous evidence. In *Proceedings of ACL-06*, pp. 801–808.

M. Piasecki, R. Kurc, R. Ramocki and B. Broda. Lexical Activation Area Attachment Algorithm for Wordnet Expansion. 2012a. In *Proceedings of the 15th AIMSA conference*, Varna, Bulgaria.

M. Piasecki, M. Marcińczuk, R. Ramocki, M. Maziarz. 2012b. WordnetLoom: a Wordnet Development System Integrating Form-based and Graph-based Perspectives. *International Journal on Data Mining, Modelling and Management*.

B. Broda, R. Kurc, M. Piasecki, R. Ramocki. 2011. Evaluation Method for Automated Wordnet Expansion. In *Security and Intelligent Information Systems*. Springer, 2011.

D. Widdows. 2003. Unsupervised methods for developing taxonomies by combining syntactic and statistical information. In *Proceedings of NACL-03*.

H. F. Witschel. 2005. Using decision trees and text mining techniques for extending taxonomies. In *Proceedings of Learning and Extending Lexical Ontologies by using Machine Learning Methods, Workshop at ICML-05*.

J. Weeds, D. Weir. 2005. Co-occurrence retrieval: A flexible framework for lexical distributional similarity. *Computational Linguistics* 31(4), pp. 439–475.