Linguistically Informed Mining Lexical Semantic Relations from Wikipedia Structure

Maciej Piasecki, Agnieszka Indyka-Piasecka, and Roman Kurc

Institute of Informatics, Wrocław University of Technology, Poland
{maciej.piasecki,indyka,roman.kurc}@pwr.wroc.pl

Abstract. A method of the extraction of the wordnet lexico-semantic relations from the Polish Wikipedia articles was proposed. The method is based on a set of hand-written set of lexico-morphosyntactic extraction patterns that were developed in less than one man-week of workload. Two kinds of patterns were proposed: processing encyclopaedia articles as text documents, and utilising the information about the structure of the Wikipedia article (including links). Two types of evaluation were applied: manual assessment of the extracted data and on the basis of the application of the extracted data as an additional knowledge source in automatic plWordNet expansion.

1 Motivation, Related Work and Main Ideas

Electronic texts available in huge volumes include large amounts of valuable information and knowledge, but their use by the intelligent systems is problematic due to the limited development of the contemporary human language technology. However, *structured documents* in which text is extended by the additional layers of annotation expressing semantic information (e.g. text categorisation) or the structural organisation of a document, create new possibilities for automated, text-based knowledge extraction. Crowd-sourced encyclopaedias like Wikipedia are of special interest due to their substantial size and free availability for different languages. Moreover, Wikipedia, as being continuously edited by a number of users, reflects potentially contemporary language use.

The potential of Wikipedia as a lexical semantic knowledge base has been widely explored recently. It has been used in NLP tasks like text categorization [4], where article-category links were used for computing semantic relatedness between words in articles, information extraction [15], information retrieval [5], question answering [1], computing semantic relatedness [8,16], or named entity recognition and disambiguation [2]. All these tasks require well-constructed lexical semantic information, which could be acquired from linguistic knowledge bases like WordNet [3]. Ponzetto and Strube [14] combine the Wikipedia categories into a semantic network which served as basis for computing the semantic relatedness of words. The well-developed relatedness measures established for WordNet were used. In the further research they have developed the automatic method for assigning *isa* and *notisa* relations between the categories from Wikipedia [14].

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6591, pp. 297–306, 2011.

In our approach, we perceived Wikipedia as a set of structured documents in the natural language. Our objective is to extract knowledge concerning lexical semantics from them automatically. Similar works are mostly focused on the utilisation of the Wikipedia structure, e.g. calculation of similarity on the basis of the link structure, processing on the category names and their dependencies or mining meta-information. Our goal is to extract pairs of words and multi-word units that are linked by one of the wordnet lexico-semantic relations.

A wordnet is an electronic thesaurus whose construction follows the main lines of Princeton WordNet – the first and still the biggest wordnet. In a wordnet, words are grouped into sets of near synonyms – called *synsets* (basic building blocks). Synsets, but also individual words, are linked by lexico-semantic relations, that belong to a limited, linguistically motivated set, e.g. hypernymy, meronymy or antonymy (among words). Despite several known deficiencies of wordnets, cf e.g. [13], wordnets express the largest sizes among lexical semantics resources, are manually constructed and are useful in many applications. Wordnet development is cheaper than in the case of semantic lexicons of other types, but it is still a laborious process. A semi-automatic, supporting tool can be a substantial advantage, e.g. [10].

Analysis of the Wikipedia structured content gives an opportunity to combine pure text analysis with recognition of different kinds of annotation (e.g. links, categories etc.). In this paper we leave aside the linguistic analysis of the categories, studied in literature, e.g. [7] and instead we concentrate on processing the content of the Polish Wikipedia[1] articles together with the structural information encoded in them. Our goal is to extract from sequences: term – article, lemma[2] pairs that are *instances* of wordnet relations. Polish is an inflectional language and it seems to be more feasible to perform semantic processing on the level of lemmas, than on the level of word forms.

2 Pattern-Based Semantic Relation Extraction

There are two basic approaches to the extraction of lexical semantic knowledge from text corpora: *Distributional Semantics* and *pattern-based methods*. Distributional Semantics is based on measuring semantic relatedness among words on the basis of the similarity of their distributions in the corpus. High values of the relatedness can be correlated with a rich variety of relations. The measure can be a valuable knowledge source, but it is better to construct it on a corpus which is much larger than Polish Wikipedia.

Pattern-based methods originate from the seminal works of Hearst [6] on application of lexico-syntactic patterns to a corpus. Hearst's patterns have the expressive power of the regular expressions and recognise selected words and expressions, but require Noun Phrases boundaries to be identified by a shallow

[1] `http://pl.wikipedia.org/wiki/Strona_główna`

[2] A lemma is a pre-selected basic morphological form representing the whole set of words or multi-word expressions that differ only with respect to values of grammatical categories (like case or gender) but express the same lexical meaning.

parser. Each pattern is aimed at recognition of instances of a particular lexico-semantic relation. Hearst proposed patterns only for the nominal hypernymy, later attempts to apply patterns for the extraction of meronymy did not achieved accuracy on the practical level.

There is no robust shallow parser for Polish, however, patterns based on the language of morpho-syntactic constraints have been proposed and successfully applied to the extraction of data for the semi-automatic expansion of plWordNet, cf [13]. The patterns were expressed in the JOSKIPI language – a language of rules of the morpho-syntactic tagger [12]. JOSKIPI can be used to describe not only token sequences, but also morpho-syntactic relations between pairs of words, e.g. agreement on number, gender and case between a noun and an adjective modifying it. Here this approach will be extended to processing structurally annotated Wikipedia texts.

Wikipedia articles represent a rich variety of forms, but three main parts of a typical article can be distinguished: *term name*, *versions* and *description*. Versions occur in many articles, are enclosed in parentheses, and include information concerning translations, and term synonyms, but also etymology or language register, domain, etc. For instance:
Ciśnienie tętnicze (ang. "blood pressure" – BP) – ciśnienie wywierane przez
Gloss: Blood pressure (...) is the pressure exerted by

Term: *Ciśnienie tętnicze* — **Variants**: (ang. "blood pressure" – BP) — **Description**: ciśnienie ...

Description is generally a free text with links to other articles embedded in it, however, we can observe that a few first sentences are usually directly related to the term classification and, especially, links occurring there directly characterise the term (e.g. in terms of part/whole distinctions). When we go further from the description beginning interpretation of link semantics is becoming less and less predictable and author's intentions behind attachment of link to tokens[3] are much more difficult to be automatically discovered.

Following the article structure, two types of extraction rules were introduced:

- *article text rules* that can be applied to any piece of the description and do not take into account structural annotations,
- *heading rules* that are intended to be applied to the sequence consisting of the article term, versions and the first sentence of the description treated together as a one complex sentence.

2.1 Article Text Extraction Rules

Article text rules were developed on the basis of rules constructed for the general corpus, cf [13]. Information concerning the structure (links in first) is not used in them, as article text rules can be applied to any part of the description, even

[3] Links are not necessarily attached to proper expressions only – they very often encompass only selected words or symbols.

located far from the beginning and related in a remote way to the term or article categories. Each rule defines a scheme of textual context in which if two noun lemmas occur, it is very likely that they are associated by a particular lexico-semantic relation – the context is a marker. A rule describes lemma positions in the context, selected lexical elements and potential morpho-syntactic relations among lemma and context elements. The context is not limited to the token sequence between the two lemma occurrences but can be freely extended to the tokens preceding the first lemma and following the second lemma. Rules (schemes in fact) are next instantiated by a list of lemmas that represent lexical units for which we want to extract semantic information. Constraining the work of the rules to the preselected lemmas filters out information noise created by associations with infrequent Proper Names or their parts.

Three productive article text rules were applied, all focused on the extraction of hypernymic pairs. The example of a rule is given below – by Noun1 and Noun2 we refer to the lemmas instantiating the rule:

```
and(
 not(Noun1 is in the genitive case and preceded by a noun in genitive),
 rlook(1,end,$C, in(base[$C],"i" and,"oraz" and) ),
 in(base[$+1C],"inny" other,"pozostały" the rest of),
 equal(nmb[$+1C],pl),
 only(1,$-1C,$X, adjectives, adverbs, nouns and commas ),
 not( conjunction or punctuation mark on the following position )
 rlook($+2C,end,$Y, in(flex[$Y],noun) ),
 equal(base[$Y],"Noun2"),
 equal(cas[$Y],cas[0]),
 not( Noun2 in genitive and precedes a noun in genitive )
 only($+3C,$-1Y,$Z,in(flex[$Z],adjectives and adverbs)) )
```

The above rule is expressed in JOSKIPI language. For the presentation clarity the exact code was simplified and summarized in some parts. The italic font marks abbreviations and glosses. In the rule, first we test if Noun1 is not an inner element of a sequence of nouns in the genitive case. In such a situation it is very likely that Noun1 is not a noun phrase head, and is not characterised by the relation identified by the rule. Next, we look for one of the particular conjunctions to the right (`rlook`) of Noun1 such that it is in a sequence with one of the particular adjectives in the plural number (`nmb`). Between Noun1 and the found concjunctions only adjectives, adverbs, nouns and commas can occur. Next, starting from the first position after the conjunction–adjective sequence we are looking for a Noun2 occurrence. The Noun2 and Noun1 occurrences must be in same case. Finally we check if only adjectives and adverbs occur between Noun2 and the conjunctive. Besides simple tests presented in the rule, JOSIPI offers also possibility of performing complex tests on morpho-syntactic agreement between pairs of words or across word sequences. Only Noun2 is explicitly referred to in the rule, as it is assumed that rule is run in contexts with Noun1 on the position 0. As the rules are instantiated with preselected lemmas, it is the task of the control mechanism to scan text for lemma occurrences and run

the appropriate rule instances. Henceforth, the above rule will be called R_*IInne*. The other two article text rules are based on two Polish copular constructions:

- R_*Jest* – built around the verb *być* 'to be'
- and R_*To* – built around a predicative word (a quasi-verb) *to* 'to be'.

R_*To* shows a bias towards the identification of synonymic pairs, while R_*Jest* describes a typical *is_a* context. All three rules were applied to the Polish Wikipedia articles extracted in textual form and preprocessed, cf Sec. 4.

2.2 Heading Extraction Rules

Heading rules are focused on the utilisation of the structural information and were designed to be applied for the initial parts of articles. In their construction, it is assumed that Noun1 is equal to the article term and is located on the position 0. As JOSKIPI constraints work within the limits of one sentence, the work of heading rules is limited to a sentence including: term, variants and the first sentence of the description. JOSKIPI works on the level of morphologically annotated text. In order to make the link positions in text visible to JOSKIPI they were marked by the additional symbols: "$LB" and "$LE".

Six heading rules were manually constructed. An example of the rule R_*ToRodzaj_Lnk* extracting pairs: hyponym – hypernym is presented below:

```
and( in(cas[0],nom,acc),
     rlook(1,10,$I,in(orth[$I],"-","to" is,"$LB")),
     not(equal(orth[$I],"$LB")),
     rlook($+1I,15,$R,in(flex[$R],subst,depr,ger)),
     inter(cas[$R],nom,acc),
     in(base[$R],"rodzaj" kind,"typ" type,"podtyp" subtype,
             "dziedzina" domain,"forma" form, "sposób" manner),
     rlook($+1R,$+10R,$N,or( in(flex[$N],subst,depr,ger),
                                  equal(orth[$R],"$LB") )),
     in(flex[$N],subst,depr,ger),
     equal(base[$N],"Noun2") )
```

In the above rule, first, we check if the article term (Noun1) is in the nominative case (or accusative due to possible tagger errors) – a different case can signal the the article has not been written in a typical way. Next we are looking for an occurrence of a dash '–' or predicative quasi-verb *to* 'is' which should be found before the occurrence of the first link. The first link usually classifies the article term. The symbol *$LB* was added during preprocessing. Next, we look for the first noun that comes after the copular predicate (i.e. '–' or *to*). We expect it to be one of the nouns that signal a semantic relation: subordinate – superordinate. A list of such nouns is provided in the rule. The relation marker must be in the nominative (or accusative) case in this language construction. Finally we are looking for Noun2 which should follow the relation marker. Contexts in

which Noun2 occurs as a part of the link are excluded from R_*ToRodzaj*_*Lnk*, as being covered by one of the five other rules (`$LE` represents a link end):

R_Dash_*Lnk*: `Noun1`$_{case\in nom,acc}$ ... '-' ... `$LB` `Noun2`$_{case\in nom,acc}$ ... `$LE`
R_*ToElement*_*Lnk*: `Noun1`$_{case\in nom,acc}$ ... ('-' | `to`) ... (`element` *element* | `część` *part* | `fragment` *fragment*) ... `Noun2`
R_Dash_*Noun*: `Noun1`$_{case\in nom,acc}$ ... '-' ... `Noun2`$_{case\in nom,acc}$ – there is no beginning of a link before Noun2 and Noun2 is different that triggering words of the rules: R_*ToRodzaj*_*Lnk* and R_*ToElement*_*Lnk* (presented below).
R_After_Parantheses: `Noun1`$_{case\in nom,acc}$ ... '(' ... ')' ... `Noun2`$_{case\in nom,acc}$ – there is no link beginning before Noun2
R_In_Parantheses: `Noun1`$_{case\in nom,acc}$ '(' not(*verbs* and *punctuation marks*) `Noun2`$_{case\in nom,acc}$

R_Dash_*Lnk* reflects a relatively numerous article scheme in which the description starts with a dash after which a link to the superordinate term comes in a close distance. Rules: R_*ToRodzaj*_*Lnk* (presented earlier in details) and R_*ToElement*_*Lnk* describe less frequent, further specifications of this scheme. R_*ToRodzaj*_*Lnk* refers to lemmas directly signalling the hyponymy relation between the article term and the first link, while R_*ToElement*_*Lnk* identifies a narrow group of lemmas marking the meronymy relation (part of).

R_*ToElement*_*Lnk* extracts a limited number of instances, but their accuracy is relatively high. It is worth to be emphasised that the accuracy of rules similar to R_*ToElement*_*Lnk* is low when they are applied to a general corpus.

The other three rules explore information expressed by the article structure, not links, and the fact the first sentence is being processed. R_Dash_*Noun* and R_After_Parantheses are based on the similar assumption like R_Dash_*Lnk*, namely the first noun, if it is in the nominative case, represents a hypernym. All three rules were split, as their accuracy, and thus their reliability as knowledge sources, can be different, cf Sec, 4. R_In_Parantheses explores the fact, that synonyms of the term are often provided the article author in the variants.

3 Semi-automatic Wordnet Expansion

Pattern-based method extract relation instances with relatively high accuracy but limited coverage, cf Sec. 4, while methods of Distributional Semantics produce result for any pair of lemmas but the description is not focused on any single lexico-semantic relation. The idea of combing heterogeneous extraction methods in automated wordnet expansion became the basis for the Algorithm of Activation-area Attachment (henceforth AAA) and the WordnetWeaver system supporting semi-automatic plWordNet expansion. AAA is capable to utilise heterogeneous knowledge sources characterising lemma relations in suggesting senses for new lemmas. Knowledge sources extracted from Wikipedia are potentially valuable for AAA because of their expected high accuracy. AAA was presented in [10] for the description of its latest development see [11] in this volume. AAA is used in one of the two evaluation methods discussed in Sec. 4.

AAA introduces a notion of a *semantic fit* between two lemmas and also between a lemma and a synset (defining a sense). In the later phase synsets that fit the input lemma are grouped into *activation areas* describing the input lemma senses. Fit for a lemma pair depends on how well the given pair is supported by the knowledge sources, e.g. the pair was extracted by several reliable knowledge sources. The function *score* assigns a value to every lemma pair on the basis of weighted voting across all knowledge sources. Reliability of different knowledge sources is estimated by manual evaluation of the accuracy of the extracted pairs. The reliability values are a basis for weighted voting present in fit and *score*.

AAA works in two phases. During the first phase semantic fit between an input lemma x and each synset Y is computed on the basis of: the semantic fit between x and lemmas belonging to the synset Y and and, additionally, the semantic fit between x and synsets linked to Y by the hypernymy or hyponymy relation (up to several links). During the second phase, on the basis of the semantic fit between x and synsets, connected subgraphs of the hypernymic wordnet structure, called *activation areas*, are identified – an activation area includes only synsets for which semantic fit to x is above some threshold; each activation area is assigned its semantic fit values to x which is equal to the maximum of the semantic fit values between x and synsets of the area.

Proposed automatic evaluation checks how well can AAA reconstruct plWordNet in the lower parts of the hypernymy structure, see [11]. Three strategies for evaluating AAA's proposals were proposed: *All* , *One* (the highest-scoring attachment site), $Best_{P\geq 1}$ (one closest attachment site). In case of all suggestions based on strong fit located not further than 2 hypernymic links from the appropriate synset (the range of acceptable errors) the accuracy was 42.87%. For the same range, the accuracy of *One* was 67.99% and $Best_{P\geq 1}$ was 75.02%. Full results can be found in [10]. Almost half of the suggestions based on strong fit were near the correct place. The result for $Best_{P\geq 1}$ strategy shows support for a linguist: the number of words with at least one useful suggestion.

4 Evaluation

Evaluation was based on the Polish Wikipedia dumps from the 29th September 2009. Articles were extracted with the help of the *Wikipedia Extractor* tool[4]. A text corpus of the size 172 millions tokens consisting only of articles in textual form (without additional elements like info-boxes, categories etc.) was created. The corpus was pre-processed by TaKIPI [9] (the Polish morpho-syntactic tagger) and next link limits were marked by the "$LB" and "$LE" symbols. All test were performed on the set of noun lemmas (one word and multi-word) extracted from plWordNet version 1.1 (from July 2010). The list consisted of 39 039 noun lemmas including 6 957 multi-word lemmas. As there is no robust shallow parser for Polish, only multi-word lemmas that were covered by the list could be recognised in the corpus.

[4] `http://medialab.di.unipi.it/wiki/Wikipedia_Extractor#Related_Work`

Two types of evaluation were performed: *direct* during which professional linguists verified manually the accuracy of relation instances extracted by the rules; and *indirect* — instance sets extracted by different rules were used as separate knowledge sources in automatic plWordNet 1.1 reconstruction – improved precision was expected.

During direct evaluation samples of the size 300 instances were randomly drawn from sets extracted by the rules. An instance was placed by linguists in one of the predefined classes:

1. *P*, *M* – proper linguistic hypernymy or meronymy, respectively, as in dictionaries or wordnets;
2. *PT*, *MT* – a form of conceptual hypernymy or meronymy supported by local context;[5]
3. *PG*, *MG* – correct, but given smart linguistic processing tools;[6]
4. *hypo*, *mero* – already added to plWordNet as hyponyms (meronyms);
5. *F* – not slotted into any other class – here treated as errors.

We write *allHypo* to denote the combination PT+P+PG+hypo (and *allMero* for MT+M+MG+mero) and *lingHypo* to denote P+PG+hypo (and *lingMero* for M+MG+mero). Instances extracted by the rules but of the opposite direction relation, i.e. were counted under the label *Hyper*. The results of the manual evaluation are given in Table 1.

Table 1. The accuracy [%] of rules according to the manual evaluation

Rule	No	allHypo	lingHypo	Hyper	allMero	lingMero
R_IInne	11984	**58,00%**	23,00%	0,33%	–	–
R_*To*	10564	33,33%	13,33%	8,33%	–	–
R_*Jest*	15309	39,33%	20,33%	3,33%	–	–
R_*ToRodzaj*_Lnk	678	64,67%	48,67%	2,33%	–	–
R_Dash_Lnk	3220	**65,33%**	33,67%	2,00%	–	–
R_*ToElement*_Lnk	401	11,00%	7,33%	1,67%	57,33%	49,00%
R_Dash_Noun	8617	48,33%	23,33%	0,67%	–	–
R_After_Para	4960	54,67%	17,00%	0,33%	–	–
R_In_Para	924	**59,46%**	48,31%	2,03%	–	–

The number of instances extracted by the heading rules seems to be small, but the rules' applicability was limited to these article terms that were covered by the lemmas extracted from plWordNet 1.1, i.e. less than 10 000 terms. Heading rules explore highly specific patterns and express relatively high accuracy. As construction and testing of the heading rules took only less than two man-work

[5] For instance, examples include a relation linking a named entity with its hypernym signalled by the head noun; a single-word lemma as a remote hypernym in place of the proper multi-word lemma; or hypernymy supported by a role played by some object in the particular local context.

[6] For example: wrong number (Carpathian Mountains versus mountain) or wrong – but semantically related – lemma (tournament versus compete).

days and they are applied to the subsequent versions of Wikipedia, the effort is profitable. The best result among article text rules was achieved by R_*IInne* which is a logical conjunction of a few patterns identifying language constructions representing a kind of enumeration. Among the heading rules the best precision was achieved by R_Dash_Lnk referring to the text under link.

The indirect evaluation was based on the application of the extracted instance sets as additional knowledge sources in automatic AAA-based plWordNet 1.1 reconstruction, cf Sec. 3. Test was performed on lexical units located in the hypernymic structure on the depth equal or greater 4. As a baseline we took the results obtained without the use of the manually written extraction rules. Next AAA results obtained with the data extracted from Wikipedia were compared to the baseline. For the acceptable error range of 2 hypernymic links and the evaluation strategy *All*, the accuracy was increased by 3.7% for the suggestions based on strong fit, 7.7% for weak fit, and 6.2% for the combined result. Concerning the limited size of the instance sets extracted from the Wikipedia, the achieved improvement is valuable. Moreover, we can extract from Wikipedia information about lemmas that are not frequent in the general corpus and for which it is difficult to acquire reliable knowledge sources.

5 Conclusions and Further Research

Our objective is the development of a semi-automatic tool supporting Polish wordnet expansion. As the plWordNet development gradually moves into the domains of infrequent (even in very large corpora) and mostly specific lemmas, the use of publicly available, semi-structured text corpora like Wikipedia is becoming more and more important. We proposed a set of rules extracting instances of the wordnet lexico-semantic relations from the Wikipedia articles. The rules were constructed manually, but for the cost of only less than one man-week of workload. Two groups of rules were developed. Rules of the first group work on articles treated as text documents. They can be applied to a general text corpus too, but they achieve much better results on the Wikipedia due to its informative content. Rules of the second group utilise structural information present in the Wikipedia articles, extract less relation instances, but mostly with better accuracy. The extracted knowledge sources improved the accuracy of the semi-automatic plWordNet expansion.

The most significant problem is the automatic recognition of new multi-word lemmas. Rules are now applied under the assumption that the list of multi-word lemmas is predefined and they all have been syntactically described that facilitates their recognition. We need to develop a method of the automatic acquisition of the linguistically described multi-word lemmas from Wikipedia in combination with a very large corpus. It should increase the coverage of the rules a lot. Moreover, the problem of the Proper Name recognition must be solved.

Acknowledgments. Work co-financed by the European Union within European Economy Programme project POIG.01.01.02-14-013/09.

References

1. Ahn, D., Jijkoun, V., Mishne, G., Müller, K., de Rijke, M., Schlobach, S.: Using Wikipedia at the TREC QA Track. In: Proceedings of TREC (2004)
2. Bunescu, R., Pasca, M.: Using Encyclopedic Knowledge for Named Entity Disambiguation. In: Proc. of the 11th Conf. of the European Chapter of ACL, pp. 9–16. ACL, Trento (2007)
3. Fellbaum, C. (ed.): WordNet – An Electronic Lexical Database. The MIT Press, Cambridge (1998)
4. Gabrilovich, E., Markovitch, S.: Overcoming the Brittleness Bottleneck using Wikipedia: Enhancing Text Categorization with Encyclopedic Knowledge. In: Proc. of the 21st National Conference on AI and the 18th Innovative Applications of AI Conference. AAAI Press, Boston (2006)
5. Gurevych, I., Müller, C., Zesch, T.: What to be? – Electronic Career Guidance Based on Semantic Relatedness. In: Proc. of the 45th Annual Meeting of ACL, Prague, Czech Republic, June 2007, pp. 1032–1039. ACL (2007)
6. Hearst, M.A.: Automatic acquisition of hyponyms from large text corpora. In: Proceedings of the Conference of the International Committee on Computational Linguistics, pp. 539–545. ACL, Nantes (1992)
7. Nastase, V., Strube, M.: Decoding Wikipedia Categories for Knowledge Acquisition. In: Proc. of the 23rd AAAI Conf., Chicago, pp. 1219–1224 (2008)
8. Nastase, V., Strube, M., Boerschinger, B., Zirn, C., Elghafari, A.: WikiNet: A Very Large Scale Multi-Lingual Concept Network. In: Proc. of LREC 2010, pp. 1015–1022 (2010)
9. Piasecki, M.: Polish tagger TaKIPI: Rule based construction and optimisation. Task Quarterly 11(1-2), 151–167 (2007)
10. Piasecki, M., Broda, B., Głąbska, M., Marcińczuk, M., Szpakowicz, S.: Semi-automatic expansion of polish wordnet based on activation-area attachment. In: Recent Advances in Intelligent Information Systems, pp. 247–260. EXIT (2009)
11. Piasecki, M., Kurc, R., Broda, B.: Heterogeneous knowledge sources in graph-based expansion of the polish wordnet. In: ACIIDS 2011. LNCS (LNAI), vol. 6591, pp. 307–317. Springer, Heidelberg (2011)
12. Piasecki, M., Radziszewski, A.: Morphosyntactic constraints in acquisition of linguistic knowledge for polish. In: Mykowiecka, A., Marciniak, M. (eds.) Aspects of Natural Language Processing (a festschrift for Prof. Leonard Bolc). LNCS, vol. 5070, pp. 163–190. Springer, Heidelberg (2009)
13. Piasecki, M., Szpakowicz, S., Broda, B.: A Wordnet from the Ground Up. Oficyna Wydawnicza Politechniki Wrocławskiej, Wrocław (2009)
14. Ponzetto, S.P., Strube, M.: Deriving a large scale taxonomy from Wikipedia. In: Proc. of the 22nd Conference of the Advacement of Artificial Intelligence, Vancouver B.C., Canada, July 22-26, pp. 1440–1445 (2007)
15. Ruiz-Casado, M., Alfonseca, E., Castells, P.: Automatic assignment of Wikipedia encyclopedic entries to WordNet synsets. In: Szczepaniak, P.S., Kacprzyk, J., Niewiadomski, A. (eds.) AWIC 2005. LNCS (LNAI), vol. 3528, pp. 380–386. Springer, Heidelberg (2005)
16. Zesch, T., Gurevych, I., Mühlhäuser, M.: Comparing Wikipedia and German Wordnet by Evaluating Semantic Relatedness on Multiple Datasets. In: Proc. of NAACL-HLT 2007, pp. 205–208. ACL (2007)