Proceedings of the International Multiconference on Computer Science and Information Technology pp. 73–78
ISBN 978-83-60810-14-9
ISSN 1896-7094

Towards Word Sense Disambiguation of Polish

Dominik Baś, Bartosz Broda, Maciej Piasecki
Institute of Applied Informatics, Wrocław University of Technology, Poland
Email: {bartosz.broda, maciej.piasecki}@pwr.wroc.pl

***Abstract*—We compare three different methods of Word Sense Disambiguation applied to the disambiguation of a selected set of 13 Polish words. The selected words express different problems for sense disambiguation. As it is hard to find works for Polish in this area, our goal was to analyse applicability and limitations of known methods in relation to Polish and Polish language resources and tools. The obtained results are very positive, as using limited resources, we achieved the accuracy of sense disambiguation greatly exceeding the baseline of the most frequent sense. For the needs of experiments a small corpus of representative examples was manually collected and annotated with senses drawn from plWordNet. Different representations of context of word occurrences were also experimentally tested. Examples of limitations and advantages of the applied methods are discussed.**

I. Introduction

POLYSEMY of word forms seems to be an intrinsic feature of the natural language and can be observed in any natural language including Polish. It is a stumbling block for semantic text processing and complicates access to meanings in a semantic lexicon. One needs an algorithm choosing the most appropriate meaning of the given word form in relation to the given context. This need was noticed as early as in 50s, and continuous works on *Word Sense Disambiguation* (henceforth abbreviated to WSD) have been performed since 70s, e.g. the *Word Expert* system of Wilks [1].

Research on WSD has been conducted for English but also for many other languages. Despite its importance, it is very hard to find published works, working systems or practical results for Polish. Development of the corpus annotated by word senses and next development of a WSD algorithm is planned as a part of the project on the construction of the National Corpus of Polish [2].

Our main objective is to developed a robust WSD method for Polish which will be based on the Polish wordnet called *plWordNet* [3], as the source of lexical meanings. However, as such a far going enterprise requires a lot of time and workload, we decided to start with a much more limited experiment. The goal of the work presented here was to develop a WSD algorithm for selected Polish words going in the line of *lexical sample* task of Senseval Evaluation Exercises [4]. Also, lexical sample task is one approach to minimise the utilised resources, especially manual work.

As we wanted to achieve a relatively high accuracy, from the very beginning we assumed a supervised learning model and a construction of a small training corpus, in which selected words were manually annotated with plWordNet synsets. In the case of the supervised learning algorithms one can control which senses are learned and identified in text. Nevertheless, research on unsupervised sense discovery have been also performed in parallel, e.g. [5], [6].

II. Construction of the Training Corpus

As none of the existing Polish corpora has been semantically annotated, we decided to select subcorpus of the IPI PAN Corpus (IPIC) [7] and extend it with annotation by word senses for occurrences of chosen words. We selected 13 different base word forms corresponding to several polysemous lexemes and homonyms. Choosing the subset, we tried to represent the variety of different problems for WSD. We selected polysemous lexemes possessing from 2 to 7 senses, where some of the senses have homonymous character, i.e. those senses represent separate homonyms of the same morphological base form. We included homonyms, as they express very different meanings and it should be easy to differentiate between their senses. The important factor in selection was also the coverage of plWordNet, which still does not describe some rare senses. Finally, we also tried to compose the set of words as consisting of words that are intuitively less or more difficult for possible WSD algorithms. The selected set includes:

1) ***agent*** (4 senses, English translation: *agent*),
2) ***automat*** (3: *automaton*, *automatic machine* and *machine-gun*),
3) ***dziób*** (4: *beak*, *bow*, *mouth* (semantically marked) and *face* (semantically marked)),
4) ***język*** (2: *language* or *tongue*),
5) ***klasa*** (6: *class*, *classroom*, *form*, *grade*),
6) ***linia*** (6: *line*, *route*, *figure*, *contour*),
7) ***pole*** (3: *field*),
8) ***policja*** (2: *police*, *police station*),
9) ***powód*** (2: *reason* and *protection* (person)),
10) ***sztuka*** (6: *art*, *play*, *craft*, *item*),
11) ***zamek*** (4: *castle*, *door lock*, *zipper* and *breechblock*),
12) ***zbiór*** (7: *set*, *collection*, *harvest* and *file*),
13) ***zespół*** (5: *team*, *gruop*, *band*, *company*, *complex* and *unit*).

Henceforth the selected base word forms will be called *training words*.

For sense inventory we used plWordNet - a resource under development, but it has already reached the size of 14 677 lexical units and is publicly available for research [3]. plWordNet follows the general scheme of the Princeton WordNet but it was constructed from scratch in bottom-up way. The starting point was the list of about 10 000 most frequent word base forms from IPIC. As a result plWordNet includes the

most frequent and most general Polish words and multiword lexical units (some of them were added by lexicographers during construction). In a way typical for wordnets, plWordNet introduces a fine grained distinction of senses.

IPIC is the only available large, morphosyntactically annotated corpus of Polish and consists of about 254 millions of tokens. We started construction of the semantically annotated subcorpus with finding all IPIC documents including at least one of the training words. Next, the documents had to be manually inspected in order to balance the number of examples for different training words, and their occurrences representing different senses. As the frequencies of senses vary in large extent, it was necessary to analyse hundreds of training word occurrences in order to find an example of some rare sense. For some sense it was difficult to find enough examples, e.g. the training word *dziób* occurs 536 times in IPIC, but after the manual inspection of all cases we found that the emotionally marked sense of *dziób* meaning *mouth* can be spotted only 9 times. Our experiments confirm phenomenon observed for English [8]—even for general and frequent words some senses are underrepresented even in a large corpora.

While selecting documents for the training subcorpus we took into account their genres and origins. The subcorpus consists of literature works, press articles and news, scientific works and legal texts. We paid special attention to avoiding situations in which all examples for some sense would be taken from the same source text. It could negatively bias the disambiguation process, as some characteristic or even idiosyncratic properties of the given text (e.g. originating from the style of the given author) could be learned by the disambiguator.

Annotation of the subcorpus was done with the help of the modified version of the annotation editor called *Manufakturzysta* [9] constructed especially for the IPIC XML format [7] based on the XCES general format. Occurrences of training words were annotated with synset identifiers from plWordNet. Annotation was performed mainly by one of the co-authors. However, in many difficult cases in which the appropriate assignment of a sense to word occurrence was unclear, the other two co-authors were consulted. The final decision often was difficult as plWordNet is still under construction and there was always possibility that some sense of the given training word is not described in plWordNet yet.

Collocations, like *pole chwały* (*field of glory*) or *ugryźć się w język* (*to bite one's tongue*), appeared to be a problem. They should be treated as the multiword lexemes, but mostly there are not present in plWordNet and, moreover, the elimination of such occurrences of training words would decrease the limited number of examples for some sense. That is why we decided to annotate training words in collocation occurrences with their literal meaning in some cases. In the example above, *język* was annotated with the sense *tongue* (i.e. body part).

We had started with annotating all training word occurrences in the documents of the subcorpus, but shortly we realised that our time limitations are too tight and we would get a very imbalanced numbers of training examples for subsequent senses, as the sense frequencies vary a lot. So, in the second phase of manual annotation we tried to identify only examples including less frequent senses[1]. The whole subcorpus includes about 1 500 semantically annotated training word occurrences, however, the sense frequencies are not still balanced yet. For some senses we could find only few examples in the whole IPIC. The detailed numbers concerning annotated occurrences of training words and their senses are presented in Table I.

TABLE I
ANNOTATED TRAINING WORDS AND SENSES.

Word	No. of senses	Annotated senses	No. of examples
agent	4	69/22/2/123	216
automat	3	28/31/46	105
dziób	4	31/17/24/9	81
język	2	22/54	76
klasa	6	9/51/6/13/29/7	115
linia	6	15/4/35/12/4/12	82
pole	3	69/25/2	96
policja	2	23/41	64
powód	2	137/122	259
sztuka	6	5/12/37/19/12/11	96
zamek	4	19/36/19/18	92
zbiór	7	16/16/1/15/3/32/3	86
zespół	5	60/1/3/28/29	121

III. APPLIED WSD ALGORITHMS

We based our current work on the previous experience from the creation of WSD systems, which were constructed mainly for English and were described in literature, e.g. [8]. We wanted to investigate behaviour of several known approaches adapted to the Polish language and Polish language tools and resources. We assumed the following scheme of processing:

1) Morphosyntactic processing of the training corpus (there is no shallow parser available for Polish).
2) Extraction of feature vectors describing occurrences of training words and storing them in the ARFF format (*Attribute-Relation File Format*).
3) Training and testing classifiers in the *Weka* system [11].

The key issue is the choice of types of features that will be used in training vectors. We surveyed types of features most frequently used in WSD systems, e.g. [8]. Lacking more advanced language tools, we assumed as the basic paradigm the *bag of words* model—Yarowsky and Florian [12] showed that omission of bag of words resulted in the decreased accuracy. Finally we have chosen five types of features:

1) *Parts of Speech* in the ± 2 text window around the training word occurrence (PoS)—information concerning parts of speech, or more precisely more fine grained division into 32 *grammatical classes*, comes from the TaKIPI morphosyntactic tagger [13] applied during preprocessing of the corpus. A training vector includes numerical identifiers of grammatical classes.

[1] A similar strategy was applied during the construction of the Basque semantically annotated corpus built for the Senseval-3 competition [10]. In this corpus, the minimal number of training examples per word w was set to $N_w = 75 + 15 \times |senses(w)|$, where $senses(w)$ is the set of senses of w.

2) *Collocation words* in the ± 2 text window (Coll)—word base forms of all grammatical classes (except punctuation signs) which occurred in the close context of the ± 2 text window. The lematisation was done by TaKIPI. No statistical filter was applied to the found occurrences. Found word base forms are represented by identifiers in the training vector.
3) The first noun to the left and to the right of the training word occurrence (Nouns)—nouns are important meaning bearers in the text. Information concerning nouns with which the training words is associated in text can be an important factor in determining the sense of the training word. Nouns are analysed on the level of their base forms. The appropriate elements of the training vector store the numerical identifiers.
4) *Wider context* of the training word (henceforth WCont)—described by base forms of words occurring in a larger text window. In a way typical for the bag of words model, the context is represented by the boolean vector, where 1 means that the corresponding base forms occurred in the given context. For the selection of base forms for the representation of contexts, we tested three methods:
   - selection by frequency—only most frequent base forms,
   - method of Ng and Lee [14],
   - selection by the Quantity of Information [15].

Concerning the description of the larger context, the selection of the most frequent base forms (4) is the simplest one. All base forms occurring in the context of training words more than the established threshold k are included in the set describing the context—the bag of words. The value of k is set experimentally.

According to the Ng and Lee method (4) we try to estimate the probability $p(s|b)$ of describing the given sense s of the training word w by the given base form b occurring in the context of w [14]:

$$p(s|b) = \frac{f(s,w,b)}{f(w,b)} \times \frac{1}{f(b)} \tag{1}$$

where:

- $f(s,w,b)$ is the frequency of b in those contexts in which w occurs in the sense s,
- $f(w,b)$—the frequency of b in the contexts of w in any sense,
- $f(b)$—the total frequency of b in the whole corpus.

In the measure of the Quantity of Information we attempt calculating how characteristic is the given base form b for the sense s of the training word w, i.e. how much information it delivers [15]:

$$Q(b,w) = -log\frac{1+N(b,w)}{1+|senses(w)|} \tag{2}$$

where $N(b,w)$ is the number of senses of w, such that they co-occur with b in the context.

We can also filter base forms used in the context description by a stop list or by the morphosyntactic properties of their occurrences, e.g. filtering out base forms of some grammatical classes.

IV. Experiments

For the experiments we concentrated our attention on Machine Learning algorithms which are implemented in Weka 3.4.14 [11] and which can be applied to small sets of training examples. Finally, we selected three algorithms representing different types of classifier used for WSD:

- *Naïve Bayes* (henceforth NB)— representing probabilistic methods in WSD,
- *k Nearest Neighbours* (kNN)—methods based on similarity to examples,
- *Decision Tables* (DT)—methods based on discriminating rules,

All experiments were done in *Weka* environment on the basis of previously prepared vectors describing training examples. Evaluation was performed in *Weka Experiment Environment*. Because of the small size of dataset, we used the scheme of *leave-one-out cross-validation* for all tests.

A. Thresholds selection

In the first set of experiments we tried to discover suboptimal values for the subsequent thresholds used in the methods of base form selection for the wider context. In the case of the Ng and Lee method and Quantity of Information we were looking for the values of both factors that result in higher WSD accuracy. For selection by frequency we were looking for the minimal frequency threshold above which the base forms have a positive influence on the accuracy.

In the case of all selection methods the higher the threshold is the smaller is the number of base forms used for the description of contexts. We remove less useful descriptors—base forms—and decrease the level of noise in data by increasing the thresholds values. As different Machine Learning algorithm (and constructed classifiers) express different possibilities in coping with noise in training data, we had to define separate sets of threshold value for the subsequent classifiers.

The used selection methods can be divided into two groups. The Ng and Lee method takes into account frequencies of base forms collected for the whole corpus. On this basis it eliminates base forms that do not have influence on the WSD process, like conjunctions and prepositions. This method has problems only with elimination of numerals, as lexemes not contributing to WSD.

The two other methods, i.e. Quantity of Information and selection by frequency do not include a similar mechanism. Thus, we had to extend them with a manually created filtering rules and a stop-list to eliminate such informationally vague base forms[2].

[2]Conjunctions, prepositions and numerals comprise about 18% of tokens in IPIC [16].

TABLE II
THRESHOLD VALUES ESTABLISHED EXPERIMENTALLY FOR THE CONTEXT DESCRIPTION SELECTION.

	Method of selection		
Classifier	Frequency	Ng and Lee	QI
NB	>2	>0.001	>0.5
kNN	>6	>0.001	>0.7
DT	>4	>0.001	>0.2

TABLE III
COMPARISON OF AVERAGE ACCURACY ACHIEVED USING DIFFERENT CLASSIFIERS AND SELECTION METHODS ON THE REDUCED SET OF TRAINING WORDS.

	Classifier		
Method	NB[%]	kNN [%]	DT[%]
Frequency	75.92	57.74	67.92
Ng and Lee	79.02	74.33	66.54
QI	78.54	69.10	69.50

The first experiments were performed on the data collected for four selected training words: *agent*, *sztuka*, *dziób* and *zamek*. The reason for this limitation was the high computation cost of the experiments. As the result we took the average from all experiments. During the experiments the training vectors included only the wider context features (WCont), as only these features are influenced by the selection methods.

The final sub-optimal values for thresholds were identified on the basis of analysis of values obtained for all base forms occurring in the contexts of the four training words. The obtained threshold values are presented in Table II—they were consequently applied in all following experiments.

In the case of the selection by frequency, results are dependent on the size of the text window and the number of training examples. The size of the window was set to ± 20 tokens. After changing the size we would have to define the threshold value again.

The highest threshold values were obtained for the KNN algorithm for all three methods. Table II shows that it is more sensitive to noise in data in comparison to the other two methods, which was expected. However, limiting the context description to the most informative base forms can increase its accuracy to a large extent.

The results obtained for the Ng and Lee method are similar to the results achieved for English [14].

Having the sub-optimal threshold values extracted, we compared all three methods of selection on the data set of the four training words. The results of the comparison presented in relation to all three classifier types are given in Table III.

In the case of both: Naïve Bayes and kNN the best results were achieved while using the Ng and Lee method. Contrary, Decision Table produced the best result in combination with the selection based on Quantity of Information.

B. Feature selection

In the next set of experiments, we wanted to identify a sub-optimal set of features for the description of training data. We performed these test on the full set of 13 training words. As all training words are nouns we could omit feature expressing grammatical class of the word being disambiguated (in the centre of the context) without loss of information.

TABLE IV
AVERAGE ACCURACY FOR ALL TRAINING WORDS IN RELATION DIFFERENT SETS OF TRAINING FEATURES.

	Classifier		
Features used	NB[%]	kNN [%]	DT [%]
WCont	89.80	72.94	75.61
WCont + PoS (± 2)	88.79	70.08	77.88
WCont + Coll(± 2)	88.88	71.24	73.52
WCont + Nouns	77.83	62.51	66.25

As the starting point we assumed the wider context set of features (one feature for each base form included to the context description). Next we extended WCont with combination of the other types of features. The results being average from all 13 training words are presented in Table IV.

On the basis of the average results for all words (Table IV) we can observe that only Decision Table classifiers produce higher result after adding PoS (± 2) features to the training vectors. The detailed analysis of the results obtained for subsequent words showed that the influence of PoS (± 2) features varies significantly in relation to particular words. For example, in the case of the training word *policja* and the kNN classifier after adding the PoS (± 2) features its accuracy increased by 7.8%, while in contrast, for the same word and Decision Tables, the combination of the WCont features with the Coll(± 2) features increased the accuracy by 34.4%. Such differences can be explained on the basis of the inspection of training examples. There are two senses distinguished for the word *policja* in the corpus:

- (English *police*) an institution protecting order and safety,
- (English *police station*) a place—a police station.

In the case of the second sense, the word *policja* is very often a part of the adverbial place describing some placement or destination, e.g. *pojechał na policję* (*went to the police station*). This kind of association is barely visible according to WCont features (bag of words), but after adding collocation words or grammatical classes from the nearest context as features becomes much more prominent.

We noticed that Naïve Bayes and kNN classifiers react in a similar way to the extension of training vectors by additional features. In contrast, the Decision Table classifier returns identical results for many words regardless of changes in the training features set. It can be caused by the low values achieved by this features and the lack of their influence on the final decisions.

C. Analysis of the results

The best results achieved for subsequent words using different classifiers are presented in Table V. The accuracy was calculated in all cases according to the one-leave cross-validation, and only the parameters of classifiers varied.

Base line has been calculated as the ratio of the number of examples for the dominating sense of the given word to the total number of examples for the given word, i.e. the base

TABLE V
BEST RESULTS IN ONE-LEAVE CROSS-VALIDATION FOR SUBSEQUENT TRAINING WORDS IN RELATION TO THE USED CLASSIFIERS.

	Classifier			
Word	NB [%]	kNN [%]	DT [%]	base line
agent	93.98	88.43	94.44	56.94
automat	90.95	64.76	57.14	43.81
dziób	76.54	64.2	59.26	38.27
język	78.95	71.05	71.05	71.05
klasa	82.5	56.67	63.04	44.35
linia	54.88	47.56	47.56	42.68
pole	95.83	87.5	91.67	71.88
policja	87.5	79.69	98.44	64.06
powód	84.17	74.9	91.12	52.90
sztuka	57.29	53.13	59.38	38.54
zamek	86.96	79.35	66.3	39.13
zbiór	79.07	50	60.47	37.21
zespół	74.38	67.77	67.77	49.59
average	80.23	68.08	71.36	50.03

line equals the accuracy of a simple majority classifier. For all words, base line calculated in this way is much higher than the base line of a random choice (compare Table I for the number of senses).

Almost all results are higher than the base line. Only for the word *język* the kNN and Decision Table classifiers produced results comparable to the base line.

One can notice the worst results, i.e. being only slightly above the base line, were achieved for words: *język*, *linia* and *sztuka*. In the case of *linia* and *sztuka* one could expect such resuts, as both words posses several polysemous senses, which are difficult to be differentiated. The low results for the word *język* is a little surprising, and it is probably an artefact of the training corpus.

The best results were achieved for the words: *zamek* and *agent*. All sense of the first one are exactly homonyms, so the meaning differences should be clear for classifiers. On contrary, the good result of *agent* is biased by the training corpus to some extent. The examples for two from the four senses of *agent* come from the same set of documents of the very similar type: the set of legal documents produced in the Parliament of Poland (*Dziennik Ustaw*). Thus the characteristic vocabulary occurring in these documents could simplify the differentiation of these two senses and could increase the average score.

Moreover, one should remember, that all the words being disambiguated are nouns, and for nouns the results achieved in WSD are mostly higher than for other Parts of Speech.

Large part of the results supports our initial assumptions that words with many homonymous senses are easier for WSD. However, we could also observe some exceptions to this scheme, which can be explained on the basis of detailed analysis of the results and the training corpus.

V. CONCLUSIONS

During the performed experiments, WSD algorithm based on the Naïve Bayes achieved the accuracy 30% above the baseline on average, the Decision Table classifier 21% above the base line on average, and the kNN classifier 18% above the base line. It is worth to emphasise that the worst results was produced by the classifiers based on kNN algorithm, which is claimed in literature, e.g. [8], to be one of the best for WSD. But according to Escuedro [13], we need to introduce weighting of examples and features and sophisticated similarity metrics in order to achieve better accuracy with kNN than with Naïve Bayes. As we wanted only to compare different approaches in relation to Polish, we did not apply such extensions. For the kNN algorithm, the introduction of the different values for the k parameters could be helpful, as well.

We got also quite low results while extending training vectors with additional features, that is often performed in WSD systems. But a closer look into the detailed results for subsequent words shows that in the case of at least some words (*policja*, *język*, *powód*, *automat*) the accuracy increased with additional features. It seems that the optimal solution is selection of different sets of training features for subsequent words and applying them in relation to a word being disambiguated. It shows that the general schemes worked out for English should not be directly transferred to Polish.

The most informative type of features appeared to be the wider context, i.e. the bag of words approach. Its positive influence can be even increased by the application of a thesaurus and grouping synonyms or clustering based on automatically extracted Measures of Semantic Relatedness [17], [18], [19].

The main disadvantage of the wider context based representation is its strong dependence on the type of text for which it is applied. It is especially visible in the case of specific texts like: legal texts or scientific works, in which a specialised vocabulary is over-represented[3]. Words from this specific vocabulary rarely occurs in the rest of a large corpus, so are highly ranked by the automatic methods of selection. Classifiers trained on the basis of the wider context representation are difficult to transfer from one domain to the other.

Moreover, we noticed that the wider context representation is especially sensitive to some errors in the corpus annotation. For example, in the first phase of corpus annotation all occurrences of training words were annotated. In the case of some occurrences, training words were located very close to each other even in the range of the text window of the wider context. So the same base form occurrences were taken to the representation of more than one training word, e.g. we present below a snippet from the training corpus which includes two occurrences of the word *automat*:

"*Nikt natomiast nie ubezpieczył się od zabicia przez automat z zimnymi napojami. A takie automaty zatłukły już 15 osób, które usiłowały siłą wyciągnąć puszkę po wrzuceniu pieniędzy i bezskutecznym naciskaniu guzika [...]*"

(*Nobody has insured himself from being killed by a drinks machine. In the meantime such machines have already beaten to death 15 persons who were trying to pull out a can using*

[3]However, training classifier based on the wider context representation on a representative subcorpus is its advantage.

force after they had thrown money into it and had been pressing a button without result.)

In the example above, the distance between the occurrences of *automat* is 6 tokens, i.e. much less than the text window of the wider context. The descriptions of both occurrences are collect from the same tokens in large extent. Thus two very similar training examples are constructed. During training they can introduce some bias, when the number of training examples is small, during testing they can increase the result while separated into the training and testing part of the corpus. In order to avoid such cases, we need to carefully select locations of training examples or to filter an existing corpus.

The results of our investigations are very positive for processing the Polish language. The performed experiments showed the construction of WSD system for Polish on the basis of limited language resources and tools is possible. Obviously with the larger number of disambiguated words one can expect decrease in the average result, but still methods developed for English seem to work for a typologically different language like Polish.

ACKNOWLEDGMENT

Work financed by the Polish Ministry of Education and Science, project No. 3 T11C 018 29.

REFERENCES

[1] Y. Wilks, "Preference semantics," in *Formal Semantics of Natural Language*, E. L. Keenan, Ed. Cambridge, UK: Cambridge University Press., 1975, vol. III, pp. 329–348.

[2] B. L.-T. Adam Przepiórkowski, Rafał L. Górski and M. Łaziński, "Towards the national corpus of polish," in *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, E. L. R. A. (ELRA), Ed., Marrakech, Morocco, may 2008.

[3] M. Derwojedowa, M. Piasecki, S. Szpakowicz, M. Zawisławska, and B. Broda, "Words, concepts and relations in the construction of Polish WordNet," in *Proc. Global WordNet Conference, Seged, Hungary January 22-25 2008*, A. Tanács, D. Csendes, V. Vincze, C. Fellbaum, and P. Vossen, Eds. University of Szeged, 2008, pp. 162–177.

[4] P. Edmonds, "Introduction to senseval," *ELRANewsletter*, vol. October 2002, 2002.

[5] B. Broda and M. Piasecki, "Experiments in documents clustering for the automatic acquisition of lexical semantic networks for polish," in *Proceedings of the 16th International Conference Intelligent Information Systems*, 2008, to appear.

[6] B. Broda, M. Piasecki, and S. Szpakowicz, "Sense-based clustering of polish nouns in extracting semantic relatedness," June 2008, to appear in the AAIA'08 Conference Proceedings.

[7] A. Przepiórkowski, *The IPI PAN Corpus: Preliminary version*. Institute of Computer Science PAS, 2004.

[8] E. Agirre and P. Edmonds, Eds., *Word Sense Disambiguation: Algorithms and Applications*. Springer, 2006.

[9] M. Piasecki, G. Godlewski, and J. Pejcz, "Corpus of medical texts and tools," in *Proceedings of Medical Informatics and Technologies 2006*. Silesian University of Technology, 2006, pp. 281–286.

[10] T. Pedersen, "The duluth lexical sample systems in Senseval-3," in *Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, Barcelona*, 2004, pp. 203–208.

[11] Weka, "Weka 3: Data Mining Software in Java," 2008, http://www.cs.waikato.ac.nz/ml/weka/.

[12] D. Yarkowsky and F. R., "Evaluating sense disambiguation across diverse parameter spaces," *Journal of Natural Language Engineering*, vol. 8, no. 4, pp. 293–310, 2002.

[13] M. Piasecki, "Polish tagger TaKIPI: Rule based construction and optimisation," *Task Quarterly*, vol. 11, no. 1–2, pp. 151–167, 2007.

[14] T. Ng and H. Lee, "Integrating multiple knowledge sources to disambiguate word senses: An examplar-based approach," in *Proceedings of the Thirty-Fourth Annual Meeting of the Association for Computational Linguistics*, 1996, pp. 40–47.

[15] C. Loupy, M. El-Béze, and P. Marteau, "WSD based on three short context methods." in *SENSEVAL Workshop, Herstmonceux Castle, England*, 1998.

[16] A. Przepiórkowski, "The potential of the IPI PAN Corpus," *Poznań Studies in Contemporary Linguistics*, vol. 41, pp. 31–48, 2006. [Online]. Available: http://nlp.ipipan.waw.pl/~adamp/Papers/2005-psicl-numbers/

[17] M. Piasecki, S. Szpakowicz, and B. Broda, "Automatic selection of heterogeneous syntactic features in semantic similarity of Polish nouns," in *Proc. Text, Speech and Dialog 2007 Conference*, ser. LNAI, vol. 4629. Springer, 2007.

[18] M. Piasecki, S. Szpakowicz, and B. B., "Extended similarity test for the evaluation of semantic similarity functions," in *Proc. 3rd Language and Technology Conference, October 5-7, 2007, Poznań, Poland*, Z. Vetulani, Ed. Poznań: Wydawnictwo Poznańskie Sp. z o.o., 2007, pp. 104–108.

[19] B. Broda, M. Derwojedowa, M. Piasecki, and S. Szpakowicz, "Corpus-based semantic relatedness for the construction of polish wordnet," in *Proceedings of the 6th Language Resources and Evaluation Conference (LREC'08)*, 2008, to appear.