Control and Cybernetics

vol. **39** (2010) No. 2

Extraction of Polish noun senses from large corpora by means of clustering*

by

Bartosz Broda[1], Maciej Piasecki[1] and Stan Szpakowicz[2,3]

[1] Institute of Informatics, Wrocław University of Technology, Poland

[2] School of Information Technology and Engineering,
University of Ottawa, Canada

[3] Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland

Abstract: We investigate two methods of identifying noun senses, based on clustering of lemmas and of documents. We have adapted to Polish the well-known algorithm of Clustering by Committee, and tested it on very large Polish corpora. The evaluation by means of a WordNet-based synonymy test used Polish wordnet (plWordNet 1.0). Various clustering algorithms were analysed for the needs of extraction of document clusters as indicators of the senses of words which occur in them. The two approaches to word-sense identification have been compared, and conclusions drawn.

Keywords: corpus linguistics, semantic similarity, Polish nouns, word clustering, Clustering by Committee, co-occurrence retrieval models, rank weight function, Polish WordNet, WordNet-based synonymy test, document clustering, keywords extraction.

1. Introduction

The construction of a large-scale language resource describing lexical semantics, such as a large wordnet[1] requires a significant effort, takes much time and costs much money. There is a high pay-off: lexical-semantic resources and especially wordnets are essential in a fast-growing number of language processing applications. A possible way of lowering the required investment is to introduce automatic tools which extract lexical knowledge from text corpora and thereby support linguists. One of the most challenging problems in the construction of such tools is word polysemy.

Two paradigms of extracting semantic relation are distinguished (Pantel and Pennacchiotti, 2006): those based on patterns, and those based on clustering and

*Submitted: May 2009; Accepted: October 2009.

[1] A wordnet is any electronic thesaurus with a structure modelled on that of Princeton WordNet (PWN) (Fellbaum, 1998).

motivated by Distributional Semantics. The former paradigm relies on the construction of lexico-syntactic patterns which can be used to identify lemma pairs associated by a particular type of relation, such as hypernymy or hyponymy. Patterns seldom guarantee good coverage (recall), and precision of retrieval may be low. It is also not possible on a large scale to make sharp distinctions between near-synonyms and hypernyms, either direct or indirect. Relation instances – lemma pairs – originate from different senses of polysemous lemmas, but the patterns themselves do not enable the distinction between senses.

Distributional-semantic methods are founded on the *Distributional Hypothesis* (Harris, 1968): the similarity of the distribution of language expressions across different contexts of use is directly correlated with the similarity of the meaning of those expressions. Distribution can be described by a coincidence matrix – lemmas by context types – which stores the frequency of lemma occurrences in particular contexts. The matrix serves as the basis for the extraction of a *Measure of Semantic Relatedness* (MSR).

MSRs generate a continuum of relatedness values for lemma pairs. Even a casual look at a list of lemmas most related to a given lemma reveals many semantic relations. We can observe synonymy, various types of wordnet relations (semantic relations typically represented in wordnets) and relations based on some situation type, which involves entities that the two lemmas represent. MSR tend to deliver vague information, so the problem of identification of synonymous lemmas is even more difficult than for the pattern-based approaches. Moreover, the statistical nature of MSR extraction introduces a strong bias towards the dominance of the most frequent sense of a given lemma L in values produced for it. Often the lemmas most related to L reflect one or two senses most frequent in the corpus under consideration.

As a result, the nodes in an automatically extracted network of lexical semantic relations are lemmas, not lemma senses (often referred to as lexical units). This differs from the standard structure of wordnets and perhaps other lexical-semantics resources. That is why we sought a method which would allow us to identify different lemma senses based on data extracted directly from corpora. We also wanted to develop a solution for an inflectional language with weakly constrained word order and a limited number of available language tools and resources – the situation with Polish. The goal of the work presented here was to investigate two sense identification methods based on the clustering of lemmas and of documents.

By fuzzy clustering of lemmas using a MSR, we wanted to identify for a given lemma L all different lemma clusters containing L and interlinked by high MSR values. The clusters were assumed to express common meanings of the members and define a particular sense of the given lemma. It was not clear, however, how the problem of the dominance of frequent senses would influence the process.

Several clustering algorithms for the task of grouping lemmas have been discussed in the literature. Among them, Clustering by Committee (CBC) (Pantel, 2003; Pantel and Lin, 2002) has been reported to achieve good accuracy in evaluation based on Princeton WordNet (PWN). CBC is often referred to in the literature as one of the most interesting clustering algorithms (Pedersen, 2006). It relies only on a modestly advanced dependency parser and an MSR based on pointwise mutual information (PMI) extended with a discounting factor (Pantel and Lin, 2002). This MSR is a modification of Lin's measure (Li, 1998) analysed in Broda et al. (2008a) in application to Polish. Both measures are close to the RWF measure (Piasecki, Szpakowicz and Broda, 2007b) which achieves good accuracy in a comparison with Polish WordNet (Derwojedowa et al., 2008).

Our goal was also to analyse CBC's applicability to an inflected language, for which there is a limited set of language processing tools, and to extract lemma clusters. We expected to identify, for a polysemous lemma, several clusters of high internal similarity. We also wanted to improve CBC's accuracy and to analyse its dependence on several thresholds, which are introduced (explicitly or implicitly) in the description of CBC. We were looking for a more objective and straightforward evaluation of the algorithm results than originally proposed by Pantel and Lin (2002).

Applications of CBC to languages other than English are rarely reported in the literature. Tomuro et al. (2007) mentioned briefly some experiments with Japanese, but gave no results. And yet, differences between languages – especially in the availability of lexical resources – can affect the construction of the similarity function at the heart of CBC. The algorithm also crucially depends on several thresholds, whose values had been established experimentally. It is unclear to what extent those values can be reused or re-discovered for other languages and language resources.

Documents allow us to analyse lemma occurrences in a particular semantic context. Let us assume that the heuristics of 'one sense per discourse' (Agirre and Edmonds, 2006) is correct frequently enough with respect to a very large set of documents. We can then treat a single document as representing a specific, narrow semantic domain and pertaining to specific senses of lemmas, which occur in the given document. We can treat a cluster of similar documents as representing a larger domain, pertaining to more coarse-grained identification of lemma senses. Thus, given clusters of documents and their characteristic lemmas, one can expect the correspondence of lemmas and document clusters to correlate with distinctions between lemma senses. We have examined this hypothesis in the experiments presented in Section 6.

2. The CBC algorithm

The CBC algorithm has been well described by its authors (Pantel, 2003; Pantel and Lin, 2002). We will therefore only outline its general organisation, following Pantel and Lin (2002), and emphasise selected key points. We have reformulated some steps in order to name consistently all thresholds present in the algorithm. Otherwise, we keep the original names.

```
I Find most similar elements
    1. for each word $e$ in the input set $E$, select $k$ most similar words considering only $e$'s features[2] above the threshold $\theta_{MI}$ of mutual information
II Find committees
    1. extract a set of unique word clusters by average link clustering, one highest-scoring cluster per list
    2. sort clusters in descending order and for each cluster calculate a vector representation based on its elements
    3. going down the list of clusters in sorted order, extend an initially empty set $C$ of *committees* with clusters similar to any previously added committee below the threshold $\theta_1$
    4. for each $e \in E$, if the similarity of $e$ to any committee in $C$ is below the threshold $\theta_2$, add $e$ to the set of residues $R$
    5. if $R \neq \emptyset$, repeat Phase II with $C$ (possibly $\neq \emptyset$) and $E = R$
III Assign elements to clusters
    • for each $e$ in the initial input set $E$
        1. $S$ = identify $\theta_{T200} = 200$ committees most similar to $e$
        2. while $S \neq \emptyset$
            (a) find the cluster $c \in S$ most similar to $e$
            (b) exit the loop if the similarity of $e$ and $c$ is below the threshold $\sigma$
            (c) if $c$ "is not similar"[3] to any committee in $C$, assign $e$ to $c$ and *remove* from $e$ its features that *overlap* with $c$'s features
            (d) remove $c$ from $S$.
```

CBC has three main phases. Phase I prepares data representing the semantic similarity of words (in English; we work with lemmas rather than words in Polish, a richly inflected language). Here, CBC shows strong dependency on the quality of the applied MSR – the most important CBC parameter – and the MSR is transformed by taking into consideration only some features (the threshold θ_{MI}) and the k most similar words.

In Phases II and III, the set of possible senses is first extracted by means of committees; next, words are assigned to committees. A *committee* is a word cluster intended to express some sense by means of a cluster vector representation derived from features describing the words included in it. Committees are selected from the initial word clusters generated by processing the lists of the

[2] Pantel (2003) extracts features from dependency triplets produced by a parser. We use lexico-morphosyntactic relations as features; see Section 3 for further discussion.

[3] We interpret this as c's similarity being below an unmentioned threshold θ_{ElCom}.

k most similar words, see II.1 and II.2. Only the clusters dissimilar to other selected clusters, however, are added to the set of committees, because the committees should ideally describe all senses of the input words, see II.3. The set of committees is also iteratively extended in order to cover senses of all input words, see the condition in III.4.

Committees only define senses. They are not the final word clusters we are going to extract. Phase III uses committees to extract such word clusters – ideally the sets of near synonyms. Each word can be assigned to one or several clusters by the similarity to the corresponding committees. It is assumed that each sense of a polysemous word corresponds to some subset of features which describe the given word. In step III.2.c, whenever a word *e* is assigned to a committee *c* (meaning that the next sense of *e* has been identified), CBC attempts to identify the features which describe the sense *c* of *e* and to remove them before extracting the other senses of *e*. The idea behind this operation is to remove the sense *c* from the representation of *e*, in order to make other senses more prominent. The implementation of the *overlap* and *remove* operations is straightforward: the values of all features in the intersection are simply set to 0 (Pantel, 2003). This would be correct if the association of features and senses were strict, but it is very rarely the case. Mostly, one feature derived from lexico-syntactic dependency corresponds in different degree to several senses. A less radical solution for sense representation removal is proposed in Section 5.

3. CBC applied to Polish

Our initial intention was to re-implement CBC as published (Pantel, 2003; Pantel and Lin, 2002), in order to analyse and compare its performance for Polish. We faced two problems: there are significant typological differences between the two languages, and the availability of language tools differs. For one thing, unlike English (for which CBC was originally designed), Polish is generally a free word-order language; much syntactic information is encoded by rich inflection. This makes the construction of even a shallow parser for Polish more difficult than for English. For example, noun modification by another noun is marked by the genitive case, but genitive is also required by negated verbs, and the noun modifier can occur either in a pre-modifying or post-modifying position. On the other hand, there are possibilities of exploring morpho-syntactic relations between word forms (but not in the case of noun-noun modification). No verb subcategorisation dictionary is available for Polish, so the identification of verb arguments in text is almost impossible, and semantic description of nouns can only to a small extent be based on relations to verbs.

CBC for English begins by running a dependency parser on the corpus. No similar tool exists for Polish. In Piasecki, Szpakowicz and Broda (2007b), Broda et al. (2008a) a similar problem was successfully solved by applying several types of lexico-morphosyntactic constraints to identify a subset of structural dependencies mainly from morphological agreement among words in a sentence and a few positional features like noun-noun sequence of modification. A direct comparison of MSRs based on parsing and on constraints is not yet possible, but the constructed constraint-based MSRs have good accuracy when compared with *plWordNet* (Derwojedowa et al., 2008) by a modified version of *WordNet-Based Synonymy Test* (WBST) (Freitag et al., 2005). The constructed MSR gave results comparable with the results achieved by humans in the same task (Piasecki, Szpakowicz and Broda, 2007a). We therefore assumed that the constructed MSR is at least comparable in quality to that used in Pantel (2003), Pantel and Lin (2002), and we adopted the constraint-based approach here, applying the same constraints as in Piasecki et al. (2007b).

As in Broda et al. (2008a), Piasecki, Szpakowicz and Broda (2007b), the applied constraints are written in the JOSKIPI language and run by the engine of the TaKIPI morphosyntactic tagger (Piasecki, 2006). Each noun n is described by the frequency with which occurrences of n in the corpus meet two lexico-morphosyntactic constraints: modification by *a specific adjective* or *an adjectival participle*, and co-ordination with a *a specific noun*.

MSRs and clustering algorithms constructed for Polish can undergo evaluation based on *plWordNet* (Derwojedowa et al., 2008), but *plWordNet* is still quite small in comparison to PWN. It includes mostly general lexical units and lacks many senses for the lemmas described. This complicates the analysis of the evaluation.

All experiments were run on the IPI PAN Corpus (Przepiórkowski, 2004), the largest annotated corpus of Polish, extended with a corpus of the on-line edition of the Polish daily *Rzeczpospolita* in 1993-2001 (Rz) (Weiss, 2008) and a corpus of large electronic text documents in Polish collected from the Internet, ≈ 214 million tokens. The *joint corpus* includes about 581 million tokens, around 3.5 times more than the corpus used in Pantel and Lin (2002). The joint corpus, however, is not well balanced: legal and scientific texts are over-represented, so intuitively rare words may have inflated frequencies, but many "popular" words have low frequencies. TaKIPI does not distinguish proper names. Lemmatization is less accurate than it is the case for English.

Several thresholds used in the CBC algorithm, plus a few more in the evaluation, pose a major difficulty for exact re-implementation. Moreover, no method of optimising CBC in relation to thresholds was proposed in Pantel (2003), Pantel and Lin (2002) and the values of all thresholds were established experimentally in Pantel (2003). There also was no discussion of their dependence on the applied tools, corpus and language characteristics. We now discuss the values of most of these thresholds:

- k – the tested value range: $[10, 20]$ (Pantel, 2003, p. 53), but the final choice is not given.
- θ_{MI} – the exact value is not presented, but it is claimed that θ_{MI} "had no visible impact on cluster quality" (Pantel, 2003, p. 53).
- $\theta_1 = 0.35$ (Pantel, 2003, p. 55).

- $\theta_2 = 0.25$ (Pantel, 2003, p. 55).
- $\theta_{T200} = 200$ (Pantel, 2003, p. 58).
- σ – different values tested (Pantel, 2003, pp. 95-96), while the best score was reported with $\sigma = 0.18$, but in the chart on p. 96 of Pantel (2003) the best result is presented for $\sigma = 0.1$, which we assumed as the default value.

A crucial threshold, θ_{ElCom}, is not overtly named in the algorithm[4] (Pantel and Lin, 2002, Pantel, 2003); the values assigned to θ_{ElCom} are unknown. The possibility that θ_{ElCom} is identical with σ is excluded by the order of steps: 2b comes before 2c. For θ_{T200} no other values were tested but it is reasonably high: it is unlikely that there ever are more than 200 senses of a word. Besides the unknown value of θ_{ElCom}, other thresholds seem to depend on the corpus and, especially, on the properties of the MSR.

To extract clusters in Phase II, we applied the CLUTO package (Karypis, 2002), which allowed us to analyse the influence of several clustering strategies – *i1*, *i2*, *h1*, *slink* and *wclink* – besides the average-link clustering originally applied in CBC. During the first experiment, we used an MSR based on PMI, constructed according to the equations presented by Pantel and Lin (2002). The results of this experiment appeared in Broda et al. (2008b).

The experiments presented in Piasecki, Szpakowicz and Broda (2007a), Broda et al. (2008a), used MSR based on Rank Weight Function (RWF) to transform feature frequencies. RWF generally surpassed several other MSRs known from the literature, some of them similar to the PMI measure applied in CBC. A recent direct performance comparison of RWF and PMI in a synonymy test (Broda, Piasecki and Szpakowicz, 2009) again showed an RWF-based measure to perform significantly better.

The RWF measure is based on the assumption that no corpus is perfectly balanced, so some feature frequencies are accidental. In a comparison of two lemmas, the *exact* feature values are not important. Instead, we compare *ranks* of features. This allows a generalization away from corpus frequencies.

After gathering corpus data, the first step of RWF calculates a measure of association between every feature and every word; any measure can be used, including t-score and even PMI. This closely resembles the standard approach to MSR construction. The RWF method adds another step: features are sorted in ascending order according to the association measure used. The value of each feature is then replaced with its rank – the most highly associated feature get the highest value. One more step removes all but k highest ranked features for every lemma ($k = 1000$ is a possible cut-off value). After this transformation to a ranking space, lemmas can be compared by any standard similarity calculation, for example the cosine measure.

We have discussed the highlights of the RWF method. See, Szpakowicz and Broda (2007a), Broda et al. (2008a), Broda, Piasecki and Szpakowicz (2009) for

[4] The threshold influences the process of assigning elements to word clusters in Phase III.

details. In the subsequent experiments with CBC, we used only MSR based on RWF.

4. Evaluating CBC on Polish

All experiments were run on the joint corpus – see Section 3. We wanted to evaluate the ability of the algorithm to reconstruct *plWordNet* synsets. That would confirm the applicability of the algorithm in the semi-automatic construction of wordnets. We put nouns from *plWordNet* on the input list of nouns (E in the algorithm). Because *plWordNet* is constructed bottom-up, the list consisted of 13,298 most frequent nouns in the IPI PAN Corpus, plus some most general nouns (Derwojedowa et al., 2008). The constraints were parameterised by 41,599 adjectives and participles, and by 54,543 nouns – 271,563 features in total.

4.1. Evaluating extracted word senses

Pantel and Lin (2002) and Pantel (2003) proposed an evaluation of the extracted word senses; it is expressed by word clusters – lemma clusters in our experiments. The evaluation is based on comparing the extracted senses with those defined for the same words in PWN. It is assumed that for a word w a correct sense is described by a word cluster c such that $w \in c$ if a synset s in PWN such that $w \in s$ is sufficiently similar to c. The latter condition is represented by another threshold θ.

The notion central to the evaluation in Pantel and Lin (2002), Pantel (2003) is similarity between wordnet synsets. The definition of similarity was based on probabilities assigned to synsets and derived from a corpus annotated with synsets. This kind of synset similarity is very difficult to estimate for languages, for which there is no such corpus, as is the case of Polish. In order to avoid any kind of unsupervised estimation of synset probabilities, we used a slightly modified version of Leacock's similarity measure (Agirre and Edmonds, 2006):

$$sim(s_1, s_2) = -log(\frac{Path(s_1, s_2)}{\max_{s_a, s_b} Path(s_a, s_b)}), \tag{1}$$

$Path(s_a, s_b)$ is the length of a path between two synsets in *plWordNet*.

We follow Pantel and Lin (2002) and Pantel (2003) in most aspects of word sense evaluation (though we work with lemmas instead of words). There is a difference in how we handle synset similarity. It is used to define the similarity between a word w and a synset s. Let $S(w)$ be a set of wordnet synsets including w (its senses). The similarity between s and w is defined as follows:

$$simW(s, w) = max_{t \in S(w)} sim(s, t). \tag{2}$$

The similarity of a synset s (a sense recorded in a wordnet) and a cluster of words c (extracted sense) is defined as the average similarity of words belonging to c. Word clusters extracted by CBC have no strict limits – their members are of different similarity to the corresponding committee (sense pattern). The core of the word cluster is defined in Pantel and Lin (2002), Pantel (2003) via a threshold κ on the number of words belonging to the core[5]. Let also c_κ be the core of c – a subset of κ most similar members of c's committee. The similarity of c and s is defined as follows:

$$simC(s, c) = \frac{\sum_{w \in c_\kappa} simW(s, w)}{\kappa}. \tag{3}$$

We assume that a cluster c corresponds to a correct sense of w if

$$max_{s \in S(w)} simC(s, c) \geq \theta. \tag{4}$$

The wordnet sense of word w, corresponding to w's sense expressed by a word cluster c, is defined as a synset which maximizes the value in formula (4):

$$arg\, max_{s \in S(w)} simC(s, c). \tag{5}$$

The question arises why this evaluation procedure is so indirect. Why do we not compare the cores of the words (or lemma) clusters with wordnet synsets? The answer appears simple. Both in Polish and in English, certain matches are hard to obtain. Word clusters are indirectly based on the MSR used. They do not have clear limits, and still express some closeness to a sense, but not to a strictly defined sense. On the other hand, wordnet synsets also express a substantial level of subjectivity in their definitions, especially when they are intended to describe *concepts*, which are not directly observable in language data. The proposed indirect evaluation will measure the level of resemblance between the division into senses made by wordnet writers and that extracted via clustering.

As stated previously, the selection of committees is critical, because it affects the remainder of the algorithm. Obviously, the criterion function for agglomerative clustering used in step of Phase II is important in this process. We therefore measure the precision of assigning lemmas to correct sense using different criterion functions. The results appear in Table 1. We used default values for thresholds: $\theta_1 = 0.35$, $\theta_2 = 0.25$, $\sigma = 0.1$, $\theta_{MI} = 250$ and $k = 20$. We assumed that default value for θ_{ElCom} is 0.2. Previous investigation of the properties of Rank Weight Function (RWF) (Broda et al., 2008a) revealed that it behaves differently than MSRs based on (pointwise) mutual information. We chose different default values for RWF: $\theta_1 = 0.2$, $\theta_2 = 0.12$. Also, θ_{MI} does not apply to RWF, so for fair comparison we used another threshold – on the minimal frequency, with which a lemma appears in any relation, $min_{tf} = 200$, and on the minimal number of different relations, in which the lemma appeared with $min_{nr} = 10$.

[5] We changed the original symbol k to κ so as not to confuse it with k in the algorithm.

The selection of threshold values was based on experiments. Automating this process is a very difficult problem, as the whole process is computationally very expensive – one full iteration takes 5-7 hours on a PC 2.13 GHz and 6GB RAM. That makes, for example, the application of Genetic Algorithms barely possible.

The differences between *slink*, *UPGMA* and *i2* (see Table 1) are very small. We have chosen the *i2* criterion for further experiments because of its efficiency.

Table 1. Precision for different criterion functions of the agglomerative clustering algorithm. The last column shows how many lemmas were assigned to clusters.

	RWF	
	Precision	No. of lemmas
UPGMA	53.90	1436
i1	53.62	1589
i2	55.49	1593
h1	37.43	926
slink	54.28	1415
wclink	53.11	1574

In Table 1 we can see that the differences in the algorithm of agglomerative clustering used in generating committees influence the final precision. The best, *i2*, leads to visibly better committees and lemma clusters.

Because the value of σ is so important for the result, we tested its several values with the other parameters fixed (RWF MSR, *i2* clustering, $\theta_{ElCom} = 0.2$):

- $\langle \sigma = 0.1$, precision = 55.49, number of lemmas assigned = 1593$\rangle$,
- $\langle \sigma = 0.12, P = 55.84, N = 1582 \rangle$,
- $\langle \sigma = 0.15, P = 56.74, N = 1558 \rangle$,
- $\langle \sigma = 0.18, P = 56.77, N = 1522 \rangle$.

With the increasing value of σ the precision increases, but the number of lemmas clustered drops significantly. The tendency persists for higher values of both thresholds, for example

$\langle \theta_{ElCom} = 0.3, \sigma = 0.25, P = 56.08, N = 1405 \rangle$.

When we set σ small and θ_{ElCom} we get relatively good precision but more lemmas clustered, for example

$\langle \theta_{ElCom} = 0.3, \sigma = 0.1, P = 55.5, N = 1593 \rangle$.

It means that, contrary to the statement and chart in Pantel (2003), tuning of both thresholds was important in our case.

In order to illustrate the working of the algorithm, we selected two examples of correct lemma senses extracted for two polysemous lemmas. The lemma senses are represented by committees described by numeric identifiers. We thus emphasise that committee members define only some lemma sense and are not necessarily near synonyms of the given lemma.

lemma: **bessa** *economic slump*
id=95 committee:{ niezdolność *inability*, paraliż *paralysis*, rozkład *decomposition*, rozpad *decay*, zablokowanie *blockage*, zapaść *collapse*, zastój *stagnation* }
id=153 committee:{ tendencja *tendency*, trend *trend* }

lemma: **chirurgia** *surgery*
109 committee:{ biologia *biology*, fizjologia *physiology*, genetyka *genetics*, medycyna *medicine* }
196 committee:{ ambulatorium *outpatient unit*, gabinet *cabinet*, klinika *clinic*, lecznictwo *medical care*, poradnia *clinic*, przychodnia *dispensary* }

Now, the same but with the proposed *heuristic of minimal value activated*, see Section 5.
lemma: **bessa**
64 committee: {pobyt *stay*, podróż *travel*} – a spurious sense
95 committee: **as above**
153 committee: **as above**

lemma: **chirurgia**
109 committee: **as above**
171 committee: {karanie *punishing*, leczenie *treatment*, prewencja *prevention*, profilaktyka *prophylaxis*, rozpoznawanie *diagnosing*, ujawnianie *revealing*, wykrywanie *discovering*, zapobieganie *preventing*, zwalczanie *fight*, ściganie *pursuing, prosecuting*} – a correct additional sense found
196 committee: **as above**

Next, two examples of committees and the generated lemma clusters.

- **committee 57**: {ciemność *darkness*, cisza *silence*, milczenie *silence = not speaking*}
- **lemma cluster**: {cisza, milczenie, ciemność, spokój *quiet*, bezruch *immobility*, samotność *solitude*, pustka *emptiness*, mrok *dimness*, cichość *silence (literary)*, zaduma *reverie*, zapomnienie *forgetting*, nuda *ennui*, tajemnica *secret*, otchłań *abyss*, furkot *whirr*, skupienie *concentration*, cyngiel *trigger*, głusza *wilderness*, jasność *brilliance*}
- **committee 69**: {grota *grotto*, góra *mountain*, jaskinia *cave*, lodowiec *glacier*, masyw *massif*, rafa *reef*, skała *rock*, wzgórze *hill*}
- **lemma cluster**: {góra, skała, wzgórze, jaskinia, masyw, pagórek *hillock*, grota, wzniesienie *elevation*, skałka *small rock*, wydma *dune*, górka *small mountain*, płaskowyż *plateau*, podnóże *foothill*, lodowiec, wyspa *island*, wulkan *volcano*, pieczara *cave*, zbocze *slope*, ławica *shoal*}

Finally, an example of a polysemous committee and the lemma cluster generated from it. The cluster clearly consists of two separate parts: animals and zodiac signs.

- **committee 11**: bestia *beast*, byk *bull*, lew *lion*, tygrys *tiger*
- **lemma cluster**: {lew, byk, tygrys, bestia, wodnik *aquarius*, koziorożec *capricorn*, niedźwiedź *bear*, smok *dragon*, skorpion *scorpio*, bliźnię *twin*, nosorożec *rhinoceros*, lampart *leopard*, bawół *buffalo*}

The last examples clearly show the role of the committee in defining the main semantic axis of the lemma cluster. Two general but semantically different lemmas occurring in the same committee make it ambiguous between at least two senses. Such a committee results in inconsistent lemma clusters created from it. Thus, the initial selection of committees is crucial for the quality of the whole algorithm, and the quality of CBC depends directly on the MSR applied.

4.2. Evaluation by a synonymy test

The estimation of synset similarity is not reliable without synset probabilities, at least as the basis of a reimplementation of the evaluation proposed in Pantel and Lin (2002), Pantel (2003). We have therefore constructed an additional measure of the accuracy of clustering. We assumed that proper clustering should be able to clear the MSR from accidental or remote associations. That is to say, if two lemmas belong to the same cluster, it is a strong evidence of their being near-synonyms or at least being closely related in the hypernymy structure.

We have applied the WordNet-Based Synonymy Test (WBST) (Freitag et al., 2005; Piasecki, Szpakowicz and Broda, 2007a). For each lemma q we create a set of four answers A in such a way that only one $p \in A$ belongs to the same synset as q. The three detractors are selected randomly but do not belong to any synset either of q or p. Next, we evaluate the accuracy of choosing p among A using MSR: we automatically select $max_{a \in A} MSR(q, a)$. In the evaluation of clustering based on WBST we use sequentially two criteria in answering a single WBST question. The result of clustering is the primary criterion, and the MSR is secondary.

We now present the algorithm of selecting the answer for a pair $\langle q, A \rangle$.

1. If only one a belongs to a lemma cluster of q, return a;
2. If there is a subset $W_A \subseteq A$ whose every element is in one (not necessarily the same) lemma cluster with q, for each $a \in W_A$:
    (a) calculate the rank position of $rank(a, q)$ in a lemma cluster of q based on the similarity to the committee;
    (b) select subset $W_H R \subseteq W_A$ of elements with the highest rank;
    (c) if $|W_H R| > 1$, return $max_{a \in W_H R} MSR(a, q)$.
3. Return $max_{a \in A} MSR(a, q)$.

If more answers belong to one of the lemma clusters of q, we need to compare them. Each element of a lemma cluster has some similarity to this cluster committee, but the similarity values depend on the size of the committee. Committees are represented by centroids calculated from feature vectors of the members. With more members the number of non-zero features increases, and the average values for most features are smaller, so the resulting values of the similarity to the elements of the lemma cluster are lower. Instead of the exact similarity values, we arrange all lemma cluster elements in the linear order of their similarity. The resulting ranks are next used in step 2a to compare different answers.

If the results of clustering do not give enough evidence to select the answer, we select the answer using the MSR alone.

We generated 10,428 WBST questions from *plWordNet*. The RWF MSR applied alone to solving the test gave 81.08% accuracy (8,455 correct and 1,973 incorrect answers).

Table 2. WBST test accuracy. CBC$_{OK}$ shows how many answers were correct. Size$_{CBC}$ shows % of CBC responses.

	acc. [%]	acc. CBC only [%]	CBC$_{OK}$	Size$_{CBC}$[%]
UPGMA	81.04	96.46	762	7.6
i1	81.09	96.72	823	7.9
i2	81.03	96.08	816	7.8
h1	81.05	95.03	624	6.0
slink	81.06	96.25	693	6.9
wclink	81.03	96.32	785	7.8

The application of the combined algorithm, based on CBC and RWF MSR achieved the accuracy of 81.09% (Table 2). The result of the CBC-based algorithm is only slightly and insignificantly better, but the conclusion is that CBC clustering did not bring any improvement to RWF MSR in its ability to distinguish between a near-synonymic and unrelated lemmas.

In the next experiment we applied RWF MSR and the CBC-based algorithm to solving a (much more difficult) Enhanced WBST (EWBST) proposed by Piasecki, Szpakowicz and Broda (2007a). In EWBST wrong answers are randomly selected among lemmas which are *similar* to the correct answer. The similarity is defined via a wordnet, *plWordNet* in our case. RWF MSR scores 64.29% in 8031 EWBST questions. The result of CBC-based algorithm is significantly lower in EWBST than the result of RWF MSR alone. Lemma clusters generated by CBC include too many loosely related lemmas. Assigning a lemma to a lemma cluster depends on the similarity to the committee vector and the implicit threshold θ_{ElCom}. Both MSRs generated on our corpus using morphosyntactic constraints can have different levels of values for different lists of the most semantically related lemmas. This complicates setting the value of θ_{ElCom} and generating more consistent lemma clusters.

The results of the evaluation by the synonymy test, consistent with the results in Section 4.1, reveal the source of low precision: loosely related lemmas are too often placed in the same clusters. The achieved results of CBC evaluation are in contrast with the better score of RWF MSR alone.

Table 3. EWBST test accuracy. CBC_{OK} shows how many answers were correct. Size_{CBC} shows % of CBC responses.

	acc. [%]	acc. CBC only [%]	CBC_{OK}	Size_{CBC}[%]
UPGMA	63.94	82.46	781	9.7
i1	63.85	78.89	810	10.1
i2	64.01	80.82	850	10.6
h1	63.73	73.41	598	7.4
slink	63.87	79.40	738	9.2
wclink	63.92	78.75	805	10.0

5. Identifying subsequent senses

CBC can assign a lemma w to several lemma clusters, because w can be similar to several committee centroids. It is assumed that the representation of different senses can depend on different features. In order to emphasise the representation of subsequent senses in the vector of w, some of the features overlapping with the committee centroid v_c are removed from the vector of w in step III2c of the CBC algorithm (Section 2). We found this technique too radical. We performed a manual inspection of data collected in a co-occurrence matrix. We concluded that it is hard to expect any group of features to encode some sense unambiguously. Some features also have low, accidental values, while others are very high. Finally, vector similarity is influenced by the whole vector, especially when we analyse the absolute values of similarity by comparing it to a threshold, for example σ in step III2b of the CBC algorithm.

Assuming that a group of features and some part of their "strength" are associated with a sense just recorded, we want to look for an estimation of the extent to which feature values should be reduced. The best option seems to be the extraction of some association of features with senses, but for that we need an independent source of knowledge for grouping features, as it was done by Tomuro et al. (2007). Unfortunately, it is not possible in the case of a language with limited resources like Polish. Instead, we tested two simple heuristics; $w(f_i)$ is the value of feature f_i, $v_c(f_i)$ – the value of f_i in the committee centroid:

- minimal value – $w(f_i) = w(f_i) - min(w(f_i), v_c(f_i))$,
- the ratio of committee importance – $w(f_i) = w(f_i) - w(f_i)\frac{v_c(f_i)}{\sum v_c(\bullet)}$.

In the minimal-value heuristics we make quite a strong assumption that a feature is associated only with one sense on one of the sides: lemma and committee. The lower value identifies the right side. The ratio heuristics is based on a weaker assumption: the feature corresponds to the committee description only to some extent.

The application of both heuristics was tested experimentally. We used the settings that resulted in the best precision in Table 1, namely RWF MSR, *i2* used for initial clustering and the original technique of removing features. The minimal-value heuristics increased the precision from 55.49% to 57.1% and the number of lemmas clustered from 1,593 to 1,608. The ratio heuristic gave a slightly worse result – the precision rose to 56% with 1,605 lemmas clustered. A manual inspection of the results shows that the algorithm tends to produce too many overlapping senses while using the ratio heuristic.

Modifications in the original CBC algorithm have resulted in increased accuracy, up to 57.1%. The achieved level of accuracy can be useful for many applications, but the low recall is a problem: only 1,608 lemmas were placed in lemma clusters which had senses identified. The limited possibility of MSR-based extraction of unambiguous committees limits the increase of recall. The MSR applied has a high accuracy in performed tests (Piasecki, Szpakowicz and Broda, 2007b), so we have to look for a different way of tackling the problem of low recall.

6. Estimation of senses via document clustering

There exists another approach to the identification of word senses, which does not rely on direct clustering of words (or lemmas). When documents are clustered instead, extraction of word senses is based on the resulting clusters. There is a plethora of clustering algorithms that differ in terms of the criterion function, efficiency, feasibility for textual data, and so on. Following a review of existing algorithms (Broda, 2007) we chose two algorithms for further analysis:

- ROCK (*RObust Clustering using linKs*) (Guha, Rastogi and Shim, 2000) – a hierarchical clustering algorithm designed for handling large sets of non-numerical data using concepts of *neighbours* and *links*,
- The *Growing Hierarchical Self-Organizing Map* (GHSOM) (Rauber, Merkl and Dittenbach, 2002) – a natural extension of Kohonen's idea idea of Self Organizing Maps (SOM) (Kohonen et al., 2000); GHSOM does not require *a priori* definition of the map structure.

For the evaluation we used a news collection which is a part of the IPI PAN Corpus (Przepiórkowski, 2004). Two test sub-corpora were isolated: DZP_{98} – 25486 short articles from the daily „Dziennik Polski" from 1998 and DZP_{04} – 7776 short articles from „Dziennik Polski", January-April 2004. DZP_{98} was divided into a few general categories: *Cracow*, *Economy*, *Sports*, *Magazine*, *Home News*, *World News*. In DZP_{04} regional categories were added.

For the domain of textual document clustering, evaluation methods should refer to some *external criteria*, such as comparing the results with some pre-existing categories created manually. We applied several evaluation measures to capture different aspects of created clusters, for example cluster *purity* (Forster, 2006) to measure cluster homogeneity, and the *Rand Index* (Forster, 2006) to measure the accuracy by decisions performed for the subsequent document pairs.

We evaluated only the first layer clusters (GHSOM) and clusters at the top level of the hierarchy (ROCK) to make the results comparable. The ROCK measure of similarity was set to the cosine between document vectors weighted by two functions: $tf.idf$ (Salton and McGill, 1983) and *logent*. Landauer and Dumais (1997) used *logent* as a weighting scheme in LSA; it combines the logarithmic scaling with entropy normalization.

Clustering gave good results. The purity values for DZP_{98} were in the range of 0.86-0.96, the Rand Index in 0.68-0.75.[6] Clustering on DZP_{04} has a very low precision and recall. Careful manual inspection of the clusters showed that many documents are ambiguously categorized (for example, articles about sporting events assigned to regional categories instead of sport). We found no mixing of major topics in clusters (for example, no document on *Sport* and *Economy* together). The algorithm also found more categories than actually present in the corpus (for example, different sport disciplines were extracted into separate clusters). An important drawback of ROCK is that it sometimes produces a very deep and unbalanced hierarchy.

Achieving good document clusters is the first step in word sense extraction. Next, we seek *representative lemmas* to label document clustered in the hierarchy organised as a tree. Lemmas, which describe clusters of documents closer to the root of the tree should be more general than those describing documents in the leaves. Ideally, the labels would add up to a hierarchical thesaurus.

Keyword extraction can be supervised or unsupervised. The former requires costly, manually constructed resources. We therefore only worked with unsupervised methods, which try to capture statistical properties of lemma occurrences to identify lemmas best describing the given document. The statistics can be computed from data in a single document, or estimated from a large body of text. To benefit both from such local and global computation, we took the method proposed by Indyka-Piasecka (2004) – its simple yet effective heuristics combine *tf.idf* weighting with *cue validity*[7] – and extended it with the algorithm of Matsuo and Ishizuka (2004). Both components of the resulting hybrid keyword extraction method rely only on lemma frequencies in the corpus, document and cluster. Matsuo and Ishizuka (2004) propose a local approach – data come from one document. It is also more complex. Morphological analysis precedes lemma clustering in the document. We simultaneously used two clustering strategies: based on Jensen-Shannon divergence and mutual information. The former measures the similarity between the distribution of lemmas; the latter is used to find similar co-occurrence patterns between lemmas. Finally, the χ^2 test checks if there is a bias in lemma occurrences in the cluster.

We evaluated this hybrid method on plWordNet (Derwojedowa et al., 2007). We compared plWordNet hyponymy hierarchy with the automatically created thesaurus (Broda, 2007; Broda and Piasecki, 2008). This was unsuccessful: only

[6] This performance depends on a few parameters; see (Broda and Piasecki, 2008) for details.

[7] The frequency of a lemma in a cluster divided by its frequency in the whole corpus.

86 hyponymy instances (less then 1% of all relations) appear in the thesaurus. Clustering whole documents might be a factor in low accuracy, but experiments with segmenting documents into smaller parts (Broda, 2007) decreased the quality of clustering. On the other hand, keyword extraction methods, developed primarily for Information Retrieval, might be not suitable for the extraction of relations between lemmas which describe different clusters of documents.

Nevertheless, the extracted cluster labels are very descriptive. For example, a cluster of documents about "interventionist purchase of grain and harvest in the area of Małopolska" are labelled with: *zboże* (*grain*), *pszenica* (*wheat*), *tona* (*tonne*), *rolnik* (*farmer*) and *agencja* (*agency*). Another possible use of extracted lemmas is to measure the degree of polysemy, because different lemma meanings occur on different branches of the hierarchy. Labels also help choose which cluster to use for training domain MSR.

7. Conclusions and further research

We have compared two methods of extracting word senses of Polish nouns: by clustering lemmas using a high-accuracy MSR, and by clustering documents considered as defining narrow semantic domains for lemmas they contain. The latter did give mixed results, so only the former approach should be open to further research. The main problem is to extend the coverage of the method based on lemma clustering.

Several explicit and implicit thresholds defined in the CBC algorithm make its re-implementation difficult. Moreover, most of the thresholds seem to depend on the MSR used and, somewhat infelicitously, on the corpus. Any optimisation method would be difficult to apply because of the complexity of the whole CBC process. One full iteration takes 5-7 hours on a PC 2.13 GHz and 6GB RAM (excluding the initial collection of feature frequencies from the corpus). A method that associates the thresholds with some properties of the corpus or MSR would be necessary. We plan to investigate the ways in which at least a subset of thresholds could be derived from the properties of the used MSR, and statistical properties of corpora used for the construction of the MSR.

Our experiments on the application of various clustering algorithms to committee extraction show the dependency of the whole CBC on this initial step. Moreover, committees often represent more than one sense. This results in inconsistent lemma clusters. Once created, a committee is not verified or amended later in the algorithm. It would be hard, but some method of committee splitting or verifying could improve the consistency of clusters.

The achieved precision is much lower than reported in Pantel (2003), Pantel and Lin (2002), but quite comparable to that reported by Tomuro et al. (2007) for a re-implementation of CBC for English. Thus, despite limited resources for Polish (such as the lack of a dependency parser) and typological differences between Polish and English, we successfully transferred the method. The achieved accuracy shows the limitations of CBC.

The selection of committees in Phase II is restricted to one committee per a list of related lemmas. However, such a list can represent more than one sense in the case of a polysemous lemma, for which the list was generated.

Infrequent lemmas in the corpus are a serious problem, because they generate high values of MSR with other infrequent lemmas. Committees generated for such lemmas negatively bias the whole CBC algorithm. We achieved better results when we constructed committees only from lemmas that are frequent in the corpus, for example ≥ 1000 occurrences.

The original solution of feature removal when assigning lemmas to lemma clusters seemed simplistic. We considered two simple heuristics of decreasing feature value in relation to the extent, in which the feature potentially corresponds to the sense represented by the given committee. Both heuristics resulted in the improvement of the precision of lemma sense extraction. We will investigate this issue further.

Most senses and lemma clusters generated by CBC are helpful, but may be of too low accuracy to be a tool willingly used by a fastidious linguist who works on extending a wordnet. Nonetheless, even a few correctly discovered relations might help increase coverage during manual construction of a wordnet.

We have identified several key elements in CBC that determine its accuracy: MSR applied, the clustering algorithm used for the identification of committees, the identification of feature-sense association, together with the algorithm of extraction of subsequent senses from lemma description, and finally the problem of optimising the numerous threshold values. Except the last point, we proposed some solutions to all elements. Still, while we achieved improvement in all of them, they all appear to be open research questions.

References

Agirre, E. and Edmonds, P., eds. (2006) *Word Sense Disambiguation: Algorithms and Applications (Text, Speech and Language Technology)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Broda, B. (2007) Mechanizmy grupowania dokumentòw w automatycznej ekstrakcji sieci semantycznych dla języka polskiego.

Broda, B. and Piasecki, M. (2008) Experiments in Documents Clustering for the Automatic Acquisition of Lexical Semantic Networks for Polish. In: M.A. Kłopotek, A. Przepiórkowski, S.T. Wierzchoń and K. Trojanowski, eds., *Proceedings of the 16th International Conference Intelligent Information Systems.* EXIT, Warsaw, 203–212.

Broda, B., Derwojedowa, M., Piasecki, M. and Szpakowicz, S. (2008a) Corpus-based Semantic Relatedness for the Construction of Polish WordNet. In: *Proc. 6th Language Resources and Evaluation Conference (LREC'08), 1800–1807.*

Broda, B., Piasecki, M. and Szpakowicz, S. (2008b) Sense-Based Clustering of Polish Nouns in the Extraction of Semantic Relatedness. In:

Proceedings of the International Multiconference on Computer Science and Information Technology – 2nd International Symposium Advances in Artificial Intelligence and Applications (AAIA'08), 83–89.

Broda, B., Piasecki, M. and Szpakowicz, S. (2009) Rank-based transformation in measuring semantic relatedness. In: Y. Gao and N. Japkowicz, eds., *Canadian Conference on AI.* **LNCS 5549**, Springer, 187–190.

Derwojedowa, M., Piasecki, M., Szpakowicz, S. and Zawisławska, M. (2007) Polish WordNet on a Shoestring. In: *Proceedings of Biannual Conference of the Society for Computational Linguistics and Language Technology, Tübingen, April 11-13 2007*, Universität Tübingen, 169–178.

Derwojedowa, M., Piasecki, M., Szpakowicz, S., Zawisławska, M. and Broda, B. (2008) Words, Concepts and Relations in the Construction of Polish WordNet. In: A. Tanács, D. Csendes, V. Vincze, C. Fellbaum and P. Vossen, eds., *Proc. Global WordNet Conference, Szeged, Hungary January 22-25 2008*, University of Szeged, 162–177.

Fellbaum, C., ed. (1998) *WordNet – An Electronic Lexical Database.* The MIT Press.

Forster, R. (2006) Document Clustering in Large German Corpora Using Natural Language Processing. PhD thesis, University of Zurich.

Freitag, D., Blume, M., Byrnes, J., Chow, E., Kapadia, S., Rohwer, R. and Wang, Z. (2005)New Experiments in Distributional Representations of Synonymy. In: *Proc. Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, Ann Arbor, Michigan, Association for Computational Linguistics, 25–32.

Guha, S., Rastogi, R. and Shim, K. (2000) ROCK: A Robust Clustering Algorithm for Categorical Attributes. *Information Systems* **25** (5), 345–366.

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

Indyka-Piasecka, A. (2004) Modele użytkownika w internetowych systemach wyszukiwania informacji (User models in the web information retrieval systems; in Polish). PhD thesis, Politechnika Wrocławska.

Karypis, G. (2002) CLUTO a clustering toolkit. Technical Report 02-017, Department of Computer Science, University of Minnesota. URL `http://www.cs.umn.edu/~cluto`.

Kohonen, T., Kaski, S., Lagus, K., Salojrvi, J., Honkela, J., Paatero, V. and Saarela, A. (2000) Self-organization of a massive document collection. *IEEE Transactions on Neural Networks*, **11**, 574–585.

Landauer, T.K. and Dumais, S.T. (1997) A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition. *Psychological Review* **104** (2), 211–240.

Li, H. (1998) A Probabilistic Approach to Lexical Semantic Knowledge Acquisition and Structural Disambiguation. PhD thesis, Graduate School of Science of the University of Tokyo.

Matsuo, Y. and Ishizuka, M. (2004) Keyword extraction from a single document using word co-occurrence statistical information. *International Journal on Artificial Intelligence Tools* **13** (1), 157–169.

Pantel, P. (2003) Clustering by committee. PhD thesis, University of Alberta, Computing Science, Edmonton, Alta., Canada.

Pantel, P. and Lin, D. (2002) Discovering Word Senses from Text. In: *Proc. ACM Conference on Knowledge Discovery and Data Mining (KDD-02)*, Edmonton, Canada, 613–619.

Pantel, P. and Pennacchiotti, M. (2006) Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations. In: *Proc. 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, 113–120. URL http://www.aclweb.org/anthology/P/P06/P06-1015.

Pedersen, T. (2006) Unsupervised Corpus Based Methods for WSD. In: E. Agirre and P. Edmonds, eds., *Word Sense Disambiguation: Algorithms and Applications (Text, Speech and Language Technology).* Springer-Verlag New York, Inc., Secaucus, NJ, USA, 133–166.

Piasecki, M. (2006) Handmade and Automatic Rules for Polish Tagger. In: P. Sojka, I. Kopeček and K. Pala, eds., *Proc. Text, Speech and Dialog 2006 Conference*, **LNAI 4188**, Springer, 205–212.

Piasecki, M., Szpakowicz, S. and Broda, B. (2007a) Extended Similarity Test for the Evaluation of Semantic Similarity Functions. In: Z. Vetulani, ed., *Proc. 3rd Language and Technology Conference, October 5-7, 2007, Poznań, Poland*, Wydawnictwo Poznańskie Sp. z o.o., Poznań, 104–108.

Piasecki, M., Szpakowicz, S. and Broda, B. (2007b) Automatic Selection of Heterogeneous Syntactic Features in Semantic Similarity of Polish Nouns. In: *Proc. Text, Speech and Dialog 2007 Conference.* **LNAI 4629**, Springer.

Przepiórkowski, A. (2004) *The IPI PAN Corpus: Preliminary version.* Institute of Computer Science PAS.

Rauber, A., Merkl, D. and Dittenbach, M. (2002) The growing hierarchical self-organizing maps: exploratory analysis of high-dimensional data. *IEEE Transactions on Neural Newtorks* **13** (6), 1331–1341.

Salton, G. and McGill, M.J. (1983) *Introduction to Modern Information Retrieval.* McGraw-Hill Inc.

Tomuro, N., Lytinen, S.L., Kanzaki, K. and Isahara, H. (2007) Clustering Using Feature Domain Similarity to Discover Word Senses for Adjectives. In: *Proc. 1st IEEE International Conference on Semantic Computing (ICSC-2007)*, IEEE, 370–377.

Weiss, D. (2008) Korpus Rzeczpospolitej.
[on-line] http://www.cs.put.poznan.pl/dweiss/rzeczpospolita. Corpus of text from the online edition of *Rzeczpospolita.*