Proceedings of the International Multiconference on Computer Science and Information Technology pp. 83–89
ISBN 978-83-60810-14-9
ISSN 1896-7094

Sense-Based Clustering of Polish Nouns in the Extraction of Semantic Relatedness

Bartosz Broda*, Maciej Piasecki*, Stanisław Szpakowicz†‡
*Institute of Applied Informatics, Wrocław University of Technology, Poland
{bartosz.broda,maciej.piasecki}@pwr.wroc.pl
†School of Information Technology and Engineering, University of Ottawa, Canada
szpak@site.uottawa.ca
‡Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland

***Abstract*—The construction of a wordnet from scratch requires intelligent software support. An accurate measure of semantic relatedness can be used to extract groups of semantically close words from a corpus. Such groups help a lexicographer make decisions about synset membership and synset placement in the network. We have adapted to Polish the well-known algorithm of Clustering by Committee, and tested it on the largest Polish corpus available. The evaluation by way of a *plWordNet*-based synonymy test used Polish WordNet, a resource still under development. The results are consistent with a few benchmarks, but not encouraging enough yet to make a wordnet writer's support tool immediately useful.**

I. Introduction

THE construction of a wordnet for yet another language, especially from scratch, requires a significant effort. There is a high pay-off: wordnets are essential in a fast growing number of applications. One way of reducing the cost is to facilitate wordnet development by automatic tools that suggest missing synsets and relations among them. Two paradigms of extracting semantic relation are cited [1]: based on patterns and based on clustering. The existing methods, however, extract relations between *lexical units* (LUs), while the problem of synset construction has been left open.

Measures of Semantic Relatedness (MRSs) based on *distributional semantics* generate a continuum of relatedness values for pairs of LUs. Even a casual look at a list of LUs most related to a given unit u reveals numerous semantic relations: synonymy and antonymy (with similar distributional patterns), hypernymy, meronymy, metonymy, LUs semantically linked to u by some situation type, and so on. Precise annotation of a list of pairs of related LUs with different types of semantic relations is a difficult manual task, and low inter-annotator agreement is likely if the task is performed without context. Any definition of hypernymy, meronymy and in particular synonymy is stated as a textual description that relies on the annotator's language competence and, at best, is supported by a diagnostic substitution test. Semantic and pragmatic constraints make many LUs semantically related to many other LUs. This view of wordnet relations, especially including near-synonymy – the basis for determining synsets – suggests that these relations are just weakly identifiable characteristic subspaces in the continuum of semantic relatedness.

Carefully selected lexico-syntactic patterns can identify pairs of LUs by a hypernymy-type relation, but their coverage (recall) is limited, precision imperfect, and a sharp distinction among remote, indirect hypernyms and more direct close hypernyms, or even near-synonyms, is not possible on a large scale.

MSRs and lexico-syntactic patterns deliver only some clues. The notion of a synset is not exactly constructively defined. To approximate synset structure automatically, several clues may have to be combined. A densely interlinked group of LUs strongly related via a MSR seems to be a natural first approximation.

The main objective of our research is the identification of tightly interlinked groups of words as representing near-synonymy and close hypernymy. Next, on the basis of the extracted groups, we identify LUs represented by lemmas belonging to groups and their sense as described by groups. Finally, we propose to extend a wordnet semi-automatically with the collected synsets and LUs.

Several clustering algorithms have been discussed in literature for the task of grouping LUs. Among them, Clustering by Committee (CBC) [2], [3] has been reported to achieve good accuracy in comparison to *plWordNet*. It is often referred to in the literature as one of the most interesting clustering algorithms, e.g. [4].

CBC relies only on a modestly advanced dependency parser and a MSR based on pointwise mutual information (PMI) extended with a discounting factor [3]. This MSR is a modification of Lin's measure [5] analysed in [6] in application to Polish. Both measures are close to the RWF measure [7] that achieves good accuracy in comparison to Polish WordNet [8].

Our goal was to analyse CBC's applicability to an inflected language for which there is a limited set of language processing tools, and to extract LU groups for the purpose of extending Polish WordNet. We expected to identify several groups of high internal similarity for a polysemous word. Moreover, we wanted to improve CBC's accuracy and to analyse its dependence on several thresholds which are explicitly, but also implicitly, introduced in CBC. We were looking for a more objective and straightforward evaluation of the algorithm results than originaly proposed in [3].

Applications of CBC to languages other than English are rarely reported in the literature. Tomuro et al. [9] mentioned briefly some experiments with Japanese, but gave no results. However, differences between languages, and especially differences in resource availability for different languages, can affect the construction of the similarity function at the heart of CBC. Moreover, CBC crucially depends on several thresholds whose values were established experimentally. It is quite unclear to what extent they can be reused or re-discovered for different languages and language resources.

II. The CBC Algorithm

The CBC algorithm has been well described by its authors [2], [3]. We will therefore only outline its general organisation, following [3] and emphasising 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 above the threshold θ_{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 on the basis of its elements
3) going down the list clusters in sorted order, extend an initially empty set C of *committees* with clusters similar to any previously added committee below the threshold θ_1
4) for each $e \in E$, if the similarity of e to any committee in C is below the threshold θ_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 a cluster $c \in S$ most similar to e
     b) exit the loop if the similarity of e and c is below the threshold σ
     c) if c "is not similar" to any committee in C[1], 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, marked by Roman numerals above. In the initial Phase I, data expressing semantic similarity of LUs are prepared. 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 LUs.

In the next two phases, the set of possible senses is first extracted by means of committees; next, LUs are assigned to committees. A *committee* is an LU cluster intended to express some sense by means of a cluster vector representation derived from features describing the LUs included in it. Committees are selected from the initial LU clusters generated by processing the lists of the k most similar LUs, see II.1 and II.2. However, only the groups dissimilar to other selected groups are added to the set of committees, because the committees should ideally describe all senses of the input LUs, see II.3. The set of committees is also iteratively extended in order to cover senses of all input LUs, see the condition in III.4.

Committees only define senses. They are not the final LU groups we are going to extract. The final LU groups – ideally sets of near synonyms – are extracted on the basis of committees in Phase III. Each LU can be assigned to one of several groups on the basis of the similarity to the corresponding committees. It is assumed that each sense of a polysemous LU corresponds to some subset of features which describe the given LU. In step III.2.c, each time a LU e is assigned to some committee c (i.e. the next sense of e has been identified) CBC attempts to identify the features describing the sense c of e and remove them before the extraction of 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. However, the implementation of the *overlap* and *remove* operations is straightforward: values of all features in the intersection are simply set to 0 [2]. It 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 amount to several senses. A less radical solution for sense representation removal is proposed in Section V.

III. CBC Applied to Polish

Our initial intention was to re-implement CBC as published in [2], [3], in order to analyse and compare its performance for Polish. However, we face two problems – there are significant typological differences between the two languages and the availability of language tools differs. For example, 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, e.g. 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 the noun-noun modification). As no verb subcategorisation dictionary is

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

available for Polish, the identification of verb arguments in text is almost impossible, and semantic description of nouns can be based on relations to verbs only to a small extent.

CBC begins by running a dependency parser on the corpus. No similar tool exists for Polish. In [7], [6] a similar problem was successfully solved by applying several types of lexico-morphosyntactic constraints to identify a subset of structural dependencies mainly on the basis of 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* [8] by a modified version of *WordNet-Based Synonymy Test* (WBST) [10]. By applying the constructed MSR we got results comparable with the results achieved by humans in the same task [11]. We therefore assumed that the constructed MSR is at least comparable in quality to the one used in [2], [3], and we adopted the constraint-based approach here, applying the same constraints as in [7].

As in [6], [7], the applied constraints are written in the JOSKIPI language and run by the engine of the TaKIPI morphosyntactic tagger [12]. 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*.

MRSs and clustering algorithms constructed for Polish can be evaluated on the basis of *plWordNet*, but *plWordNet* is still quite small in comparison to Princeton *plWordNet* (henceforth PWN). It includes mostly general words and lacks many senses for the words described. This complicates the analysis of the evaluation.

All experiments were run on the IPI PAN Corpus [13] (IPIC), the largest annotated corpus of Polish, extended with a corpus of the on-line edition of a Polish daily, 1993-2001) [14] (Rz). The joint corpus (IPIC+Rz) includes about 368 million tokens, around 2.56 times more than the corpus used in [3]. IPIC+Rz, 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 makes more errors than it is the case for English.

Several thresholds used in the CBC algorithm (plus a few more in the evaluation) are the major difficulty in its exact re-implementation. Moreover, any method of the CBC optimisation in relation to thresholds was not proposed in [2], [3] and the values of all thresholds were established experimentally in [2]. There also was no discussion of their dependence on the applied tools, corpus and characteristics of the given language. We will discuss the values of most of these thresholds:

- k – the tested value range: $[10, 20]$ [2, 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" [2, p. 53],
- $\theta_1 = 0.35$ [2, p. 55],
- $\theta_2 = 0.25$ [2, p. 55],
- $\theta_{T200} = 200$ [2, p. 58],
- σ – different values tested [2, pp. 95-96], while the best score was reported with $\sigma = 0.18$, however, in the chart on pp. 96 of [2] the best result is presented for $\sigma = 0.1$, which we assumed as the default value.

A crucial threshold θ_{ElCom} – it influences the process of assigning elements to word groups in Phase III – is not overtly named in the algorithm [3], [2]; the values applied 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 ever to have 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 [15], which allowed us to analyse the influence of several clustering strategies, namely: *UPGMA*, *i1*, *i2*, *h1*, *slink* and *wclink*, besides the average-link clustering originally applied in CBC. During the first experiment, we used a MSR based on PMI, constructed according to the equations presented in [3]. The results of this experiment appear in Table I.

In the experiments presented in [11], [6], MSR based on Rank Weight Function used for the transformation of feature frequencies generally surpassed several other types of MRS known from the literature, some of them similar to the PMI measure applied in CBC, e.g. see [11], [6]. In the second experiment we replace PMI MSR with RWF MSR.

IV. Evaluating CBC on Polish

As we wrote in section III, all experiments were run on the IPIC+Rz corpus. We wanted to evaluate the algorithm's ability 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 13298 most frequent nouns in IPIC plus some most general nouns, see [8]. The constraints were parameterised by 41599 adjectives and participles, and 54543 nouns – 96142 features in total.

A. Evaluating Extracted Word Senses

Evaluation of the extracted word senses proposed in [3], [2] 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 group 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 proposed in [3], [2] 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[16]:

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

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

Except for synset similarity, we follow [3], [2] strictly in other aspects of word sense evaluation. Synset similarity 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:

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

Similarity of a synset s (a sense recorded in a wordnet) and a group of LUs c (extracted sense) is defined as the average similarity of LUs belonging to c. However, LU groups extracted by CBC have no strict limits – their members are of different similarity to the corresponding committee (sense pattern). The core of the LU group is defined in [3], [2] via a threshold κ[2] on the number of LUs belonging to the core. 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:

$$simC(s, c) = \frac{\sum w \in c_\kappa simW(s, u)}{\kappa} \quad (3)$$

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

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

The wordnet sense of LU w, corresponding to the sense of w expressed by a LU group c, is defined as a synset which maximizes the value in formula 4:

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

The question arises why this evaluation procedure is so indirect. Why do we not compare the cores of the LU groups with wordnet synsets? The answer is seemingly simple. Both in Polish and in English, certain matches are hard to obtain. LU groups 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 words to correct sense using different criterion functions. The results appear in Table I. 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 RWF [6] revealed that it behaves differently than MSRs based on 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 word appears in any relation $min_{tf} = 200$ and on the minimal number of different relation in which the word appeared with $min_{nz} = 10$.

The selection of threshold values was done on the basis of experiments. Automatising 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 6 GB RAM, that makes e.g. application of Genetic Algorithms barely possible.

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

TABLE I
PRECISION FOR DIFFERENT CRITERION FUNCTION OF THE AGGLOMERATIVE CLUSTERING ALGORITHM.

	PMI		RWF	
	Precision	No. of words	Precision	No. of words
UPGMA	22.59	2993	38.42	682
i1	23.45	2980	35.72	744
i2	22.37	2995	38.81	742
h1	—	—	31.88	345
slink	22.70	2982	37.59	665
wclink	22.98	2981	34.14	703

The comparison – presented in Table I – of the influence on CBC of both MSRs used, PMI and RWF, is a little misleading; in these cases the number of clustered words is very different. This was caused by keeping the same value of the threshold $\sigma = 0.1$ for both versions. It seems that the value of σ must be carefully selected for each type of MSR separately.

In Table I 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 word groups.

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 = 38.81, number of words assigned = 742$\rangle$,
- $\langle \sigma = 0.12, P = 40.33, N = 719 \rangle$,
- $\langle \sigma = 0.15, P = 40.99, N = 688 \rangle$,
- $\langle \sigma = 0.18, P = 42.14, N = 655 \rangle$.

With the increasing value od σ the precision increases, but the number of words clustered drops significantly. The tendency persists for higher values of both thresholds, e.g. $\langle \theta_{ElCom} = 0.3, \sigma = 0.25, P = 45.4, N = 522 \rangle$. When we set σ small and θ_{ElCom} we get relatively good precision but more words

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

clustered, e.g. $\langle \theta_{ElCom} = 0.3, \sigma = 0.1, P = 38.81, N = 742 \rangle$. It means that, contrary to the statement and chart in [2], tuning of both thresholds was important in our case.

In order to illustrate the work of the algorithm, we selected two examples of correct word senses extracted for two polysemous LUs. The word senses are represented by committees described by numeric identifiers. In this way it is emphasised that committee members define only some word sense and are not necessarily near synonyms of the given LU.

LU: **bessa** *economic slump*

id=95 committee:{ niezdolność *inability*, paraliż *paralysis*, rozkład *decomposition*, rozpad *decay*, zablokowanie *blockage*, zapaść *collapse*, zastój *stagnation* }

id=153committee:{ tendencja *tendency*, trend *trend* }

LU: **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 V.

LU: **bessa**

64 committee: {pobyt *stay*, podróż *travel*} – a spurious sense

95 committee: **as above**

153 committee: **as above**

LU: **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 word groups.

- **committee 57**: {ciemność *darkness*, cisza *silence*, milczenie *silence = not speaking*}
- **LU group**: {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*}
- **LU group**: {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 LU group generated on this basis. The group clearly consists of two separate parts: animals and zodiac signs.

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

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

B. Evaluating 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 [3], [2]. 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 words belong to the same word group, it is a strong evidence of their being near-synonyms or at least being closely related in the hypernymy structure.

In WordNet-Based Synonymy Test (WBST) [10], [11], for each LU q we create a set of four possible 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 on the basis of MSR: we automatically select $max_{a \in A} MSR(q, a)$. In the evaluation of clustering on the basis of WBST we use sequentially two criteria in answering a single WBST question. The results of clustering is the primary criterion, and the MSR is secondary. Here is the algorithm of selecting the answer for a pair $\langle q, A \rangle$:

1) if there is only one a such that a belongs to a LU group of q, return a
2) if there is a subset $W_A \subseteq A$ whose every element is in one word group with q (not necessarily the same one), for each $a \in W_A$:
   a) calculate the rank position of $rank(a, q)$ in a LU group of q on the basis of 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 possible answers belong to one of the LU groups of a, we need to compare them. Each element of a LU group has some similarity to this group's 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 word group are lower. Instead of the exact similarity values, we arrange all LU group elements in the linear order of their similarity. The resulting ranks are next used in step 2a to compare different possible answers.

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

We generated 2726 WBST questions from *plWordNet*. The RWF MSR applied alone to solving the test gave 90.97% accuracy (2480 correct and 246 incorrect answers).

TABLE II
ACCURACY IN WBST TEST. SIZE$_{CBC}$ EXPRESSES % OF RESPONSES GIVEN BY CBC.

	Acc. [%]	Acc., CBC only[%]	CBC q.	Size$_{CBC}$[%]
UPGMA	90.61	93.94	495	18
i1	90.68	94.30	491	18
i2	90.54	94.32	493	18
h1	89.62	85.89	319	12
slink	90.35	93.13	466	17
wclink	90.28	93.50	523	19

The application of the combined algorithm based on CBC and RWF MSR achieved the accuracy of 90.68% (see Table II). The result of CBC-based algorithm is only slightly worse, but the conclusion is that CBC clustering did not bring any improvement to RWF MSR in its ability to distinguish between a near-synonym and non-related LUs.

In the next experiment we applied RWF MSR and the CBC-based algorithm to solving a (much more difficult) Enhanced WBST (EWBST) proposed in [11]. In EWBST wrong answers are randomly selected from LUs which are *similar* to the proper answer. The similarity is defined on the basis of a wordnet, *plWordNet* in our case. RWF MSR scores 55.52% in EWBST. The result of CBC-based algorithm is significantly lower in EWBST than the result of RWF MSR alone. LU groups generated by CBC include too many loosely related LUs. Assigning a LU to a LU group 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 LUs. This complicates setting the value of θ_{ElCom} and generating more consistent word groups.

The results of the evaluation by synonymy test are consistent with the results in IV-A and reveal the source of low precision: loosely related LUs are too often grouped in the same groups. The achieved results of CBC evaluation are in contrast with the better score of RWF MSR alone.

V. IDENTIFYING SUBSEQUENT SENSES

CBC can assign a LU w to several LU groups, 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 the features overlapping with the committee centroid v_c are removed from the vector of w in step 2c. 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. Moreover, some features have low, accidental values, while some 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, e.g. σ in step 2b of CBC.

TABLE III
ACCURACY IN EWBST TEST. SIZE$_{CBC}$ EXPRESSES % OF RESPONSES GIVEN BY CBC.

	Acc. [%]	Acc., CBC only[%]	CBC q.	Size$_{CBC}$[%]
UPGMA	54.47	59.81	642	24
i1	54.55	60.82	684	25
i2	54.80	62.42	660	24
h1	54.43	49.08	379	14
slink	54.47	60.60	637	23
wclink	54.40	56.07	601	22

Assuming that a group of features and some part of their 'strength' are associated with a sense just recorded, we wanted 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 in [9]. 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 the f_i feature, $v_c(f_i)$ – 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: LU 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 I, namely RWF MSR, i2 used for initial clustering and the original technique of removing features. The minimal-value heuristics increased the precision from 38.8% to 41% on 695 words clustered. The usage of the ratio heuristic improves the result even further – the precision rises to 42.5% on 701 words clustered. A manual inspection of the results showed that the algorithm tends to produce too many overlapping senses while using the ratio heuristic.

VI. Conclusions and Further Research

Several explicit and implicit thresholds defined in the algorithm make the re-implementation of CBC difficult. Moreover, most of the thresholds seem to depend on the MSR used and, unfortunately, 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 6 GB 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 shows the dependency of the whole CBC on this initial step. Moreover, committees often express more than one sense. That results in inconsistent LU groups. Once created, a committee is not verified or amended during the subsequent steps of the algorithm. It would be hard, but some method of committee splitting or verifying could improve the consistency of groups.

The achieved precision is much lower than reported in [2], [3] but quite comparable to that reported for a re-implementation of CBC for English done in [9]. Thus, instead of limited resources for Polish, e.g. lack of a dependency parser and typological differences of Polish in relation to English, we were successful in transferring 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 LUs. However, such a list can represent more than one sense in the case of a polysemous LU for which the list was generated.

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

The original solution of feature removal when assigning LUs to LU groups seemed to be too simplistic. We considered two simple heuristics of decreasing feature value in extent related to the potential feature correspondence to the sense represented by the committee. Both heuristics resulted in the improvement of the precision of word sense extraction. We will investigate this issue further.

Most senses and LU groups generated by CBC are helpful but of too low accuracy to be a tool willingly used by a fastidious linguist who works on extending a wordnet.

We have identified several key elements in CBC that decide about its accuracy: applied MRS, clustering algorithm used for the identification of committees, identification of feature-sense association together with the algorithm of extraction of subsequent senses from LU description and finally the problem of optimisation of the numerous threshold values. Except the last point, we proposed some solutions to all elements but, while we achieved improvement in all of them, all of them seem to be still open research questions.

Acknowledgment

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

References

[1] P. Pantel and M. Pennacchiotti, "Espresso: Leveraging generic patterns for automatically harvesting semantic relations," ACL 2006, Ed. ACL, 2006, pp. 113–120. [Online]. Available: http://www.aclweb.org/anthology/P/P06/P06-1015

[2] P. Pantel, "Clustering by committee," Ph.D. dissertation, Edmonton, Alta., Canada, Canada, 2003, adviser-Dekang Lin.

[3] P. Pantel and D. Lin, "Discovering word senses from text," in *Proc. ACM Conference on Knowledge Discovery and Data Mining (KDD-02)*, Edmonton, Canada, 2002, pp. 613–619.

[4] T. Pedersen, "Unsupervised corpus based methods for wsd," E. Agirre and P. Edmonds, Eds. Springer, 2006, pp. 133–166.

[5] H. Li, "A probabilistic approach to lexical semantic knowledge acquisition and structural disambiguation," Ph.D. dissertation, Graduate School of Science of the University of Tokyo, 1998.

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

[7] 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.

[8] 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.

[9] N. Tomuro, S. L. Lytinen, K. Kanzaki, and H. Isahara, "Clustering using feature domain similarity to discover word senses for adjectives," in *Proc. 1st IEEE International Conference on Semantic Computing (ICSC-2007)*. IEEE, 2007, pp. 370–377.

[10] D. Freitag, M. Blume, J. Byrnes, E. Chow, S. Kapadia, R. Rohwer, and Z. Wang, "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, June 2005, pp. 25–32.

[11] M. Piasecki, S. Szpakowicz, and B. Broda, "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.

[12] M.Piasecki, "Handmade and automatic rules for Polish tagger," ser. Lecture Notes in Artificial Intelligence, P. Sojka, I. Kopeček, and K. Pala, Eds. Springer, 2006.

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

[14] D. Weiss, "Korpus Rzeczpospolitej," [on-line] http://www.cs.put.poznan.pl/dweiss/rzeczpospolita, 2008, corpus of text from the online edtion of Rzeczypospolita.

[15] G. Karypis, "CLUTO a clustering toolkit," Department of Computer Science, University of Minnesota, Technical Report 02-017, 2002. [Online]. Available: http://www.cs.umn.edu/~cluto

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