

Experiments in Clustering Documents for Automatic Acquisition of Lexical Semantic Networks for Polish

Bartosz Broda and Maciej Piasecki

Institute of Applied Informatics, Wrocław University of Technology, Poland

Abstract

The aim of this work is to explore document clustering techniques for the needs of semi-automatic construction of a lexical semantic network for Polish. Although the majority of research in this area is based on measures of distributional similarity calculated from co-occurrences of words in large collections of documents, we wanted to approach a difficult problem of meaning ambiguity resolution from a different direction. During the research we have faced and addressed several problems of keyword extraction from document groups and document segmentation.

Keywords: document clustering, document segmentation, keywords extraction

1 Introduction

The majority of research in automatic acquisition of lexical semantics knowledge from corpora has been done on the basis of *measures of semantic relatedness* (henceforth *MRS*) calculated from word co-occurrences, e.g. Curran (2004); Lin (1998); Pantel (2003); Piasecki *et al.* (2007). All of them follow a similar scheme: first the co-occurrences of words and some features are counted, then some weighting scheme like *Pointwise Mutual Information* is used to reveal more significant associations. Last step involves usage of clustering algorithms to group similar words together and to create associations between groups.

One of the problems of this approach is word ambiguity. The previous research showed that MSR tends to group different senses of one word together or, in the case one sense is predominant in the text, MSR does not distinguish between different senses at all (Piasecki *et al.*, 2007). It can be problematic, because the sense distribution is not known a priori. Following the *one sense per discourse* heuristics (Agirre and Edmonds, 2006) one can hope to solve this problem by incorporating knowledge concerning the domain of the used documents as the heuristic states that a given word is used only in one of its senses in one discourse.

Although documents in properly constructed corpora are categorized we do not assume existence of such information. Pre-existing categories are often created with a different purpose in mind or are too coarse-grained. One category can cover different domains or one domain can encompass more than one category. Finally, to solve data sparseness problem one needs to use more than one corpus or even to

use the Web as the source of additional data. Fine-grained categories do not map well between different sources and, in the case of the Web, there are no categories associated with documents at all.

The aim of this work is to explore different clustering algorithms in order to support automatic construction of a lexical semantic network of Polish lexemes. First off all, we want to inspect possibilities and difficulties in using standard clustering algorithms to group documents written in Polish language. Hierarchical clustering solution could be valuable aid in extraction of lexical relation between words. Ideally we would like to extract keywords for clusters and acquire hierarchical thesaurus out of cluster tree. On the other hand, construction of MSRs on text from different identified domains opens variety of possibilities for sense distinction inside a given MSR.

2 Document representation

The key aspect of clustering is the representation of documents. There are four main models for the task: boolean vectors, vector space model (henceforth VSM), graph-based and metric space.

Although VSM is often applied in the domain of clustering it has several drawbacks. The first and the foremost one: word order is discarded in VSM — a document is treated as a *bag of words*. All information encoded in the document structure is lost, e.g.: the title, the author, paragraphs, sentences and even syntactic relations between words (however by a transformation syntactic relations can be extracted and used as features). VSM also assumes that words are known before the construction of the model. Nevertheless, using this representation is very convenient as most data clustering algorithms can be applied with little or no modification to the domain of text documents. On the other hand some previous works showed that some semantic lexical relation can be extracted from text using bag of words approach (Landauer and Dumais, 1997; Konchady, 2006).

Another common approach to document representation is a representation based on the graph theory. There are two main ways to do that. Firstly, one can treat documents as nodes of a graph and existing explicit connections between them as arcs. This technique is especially suitable for clustering web documents with hyperlinks. Secondly, one can describe a document as a graph — words are nodes and arcs are relations between words (Schenker *et al.*, 2005). After the introduction of the notion of the document graph similarity Schenker *et al.* (2005) investigate standard clustering algorithms designed for VSM reimplemented in their representation.

The *metric space model* (Forster, 2006) is used in a situation when it is hard to describe clustering problem accurately in terms of other representations, but a formal notion of similarity between elements is defined. A similarity matrix is used to store elements. VSM can be transformed to metric space model by transforming document \times feature matrix into a document \times document similarity matrix. We chose the metric space model based on VSM as the starting point for the experiments presented in the paper.

2.1 Thematic segmentation

Agirre and Edmonds (2006, p. 124) showed that the one sense per discourse hypothesis is not always fulfilled, especially for words with fine grained sense distinction. A document can be segmented into semantically coherent fragments to relax the influence of the assumption. Another reason to perform segmentation is that the longer the document is, the larger is the probability of different domains being represented in it. Moreover, our experiments with thesauri construction based on documents hierarchy showed that methods extracting keywords produce worse results for long documents (see sec. 5).

The discourse segmentation problem can be cast as the identification of semantically consistent parts of a document. Our first experiments in discourse segmentation for Polish were done using existing tools: TextTiling (Hearst, 1997) and LCSeg (Galley *et al.*, 2003).

A seminal solution of the segmentation problem was proposed by Hearst (1997). She introduced *TextTiling* — an method for discourse segmentation based on assigning lexical score by counting word occurrences to every gap between blocks of texts. Each block consists of k pseudo-sentences, every pseudo-sentence consists of w words. Introduction of pseudo-sentences should eliminate influence of actual paragraphs on the algorithm results. Galley *et al.* (2003) used a similar approach, the most significant difference is in the method of counting lexical score — LCSeg calculates the score on the basis of lexical chains.

Preliminary experiments on artificially created documents showed that tested methods gives good results, but needs further tuning for Polish. Documents were created by merging few news articles from different domains altogether. For some cases both algorithms fail, e.g. a topic shift between two articles concerning art and space shuttles was not detected.

Before investing more resources into developing segmentation algorithms for Polish we tried clustering applied to a corpus of text segmented by rather naive approach. We based this approach on the assumption that paragraphs already present in text are semantically coherent. Documents are divided onto fragments consisting of multiple pre-existing paragraphs, but every fragment cannot contain more then k -words. Manual inspection of results of this algorithm showed that sometimes topic was split into two segments. This is not a serious problem, a clustering algorithm should join them later into one cluster.

Experiments in clustering mostly short segments showed weakness of the segmentation based approach. In order to utilise the possible advantage of short, semantically coherent segments one needs to define measure of their similarity which is not based on word frequencies (all of them being very small). Such features like morphosyntactic word associations can be explored (cf. Forster, 2006).

3 Clustering algorithms

There is plethora of clustering algorithms that differ in terms of the criterion function, efficiency, feasibility for textual data, etc. On the basis of a review of existing algorithms (Broda, 2007) we chose two algorithms for the further analysis. We took into account the following criteria: a possibility of building hierarchical

trees, feasibility for clustering documents, ability to detect a cluster of irregular shape. Another rich review can be found in (Jain *et al.*, 1999; Mazur, 2005).

3.1 ROCK

Guha *et al.* (2000) proposed ROCK (RObust Clustering using linKs) — a hierarchical clustering algorithm designed for handling large sets of non-numerical data using concepts of *neighbours* and *links*. Two documents d_1 and d_2 are neighbours if the similarity between them is greater than θ . $link(d_1, d_2)$ is defined as the number of common neighbours between d_1 and d_2 . A criterion function (eq. 1) in ROCK should maximize the number of links between documents in one cluster, and to prevent the documents being merged into one large cluster.

$$E_l = \sum_{i=1}^k n_i * \sum_{p_q, p_r \in C_i} \frac{link(p_q, p_r)}{n_i^{1+2f(\theta)}}, \quad (1)$$

where C_i is a group of size n_i , $n_i^{1+2f(\theta)}$ is the estimation of the number of links inside a cluster of the size n_i . The function $f(\theta)$ should depend on the dataset used, but Guha *et al.* (2000) argue that even errors in this estimation should not have large impact on the clustering quality.

ROCK starts with drawing a random sample from the data, that is clustered by using links. After the construction of the initial hierarchy all documents are assigned to it. For our needs clustering accuracy is more important than having each document assigned to some cluster, so we do not perform the last step. On the other hand, the performance is not our priority so we use all data points for clustering instead of using only a random sample.

General the idea of clustering by using links follows the agglomerative clustering scheme. First, all documents are in one-element clusters. Next pairs of the most similar clusters are iteratively merged together. The main difference from the other algorithms is how the decision concerning which clusters to merge is made. Guha *et al.* (2000) select for merging a cluster maximizing the *goodness measure* (1):

$$g(C_i, C_j) = \frac{link[C_i, C_j]}{(n_i + n_j)^{1+2f(\theta)} - n_i^{1+2f(\theta)} - n_j^{1+2f(\theta)}}, \quad (2)$$

where $link[C_i, C_j] = \sum_{d_q \in C_i, d_r \in C_j} link(d_q, d_r)$.

Because of clustering on the basis of links ROCK can produce clusters of unusual shapes in the situation of ‘not so well-separated’ clusters, which is the usual case in clustering documents. Even if two documents are similar, but do not share many common neighbours, then they will be separated by ROCK.

3.2 GHSOM

The Growing Hierarchical Self-Organizing Map (GHSOM) (Rauber *et al.*, 2002) is a natural extension of Kohonen *et al.* (2000) idea of Self Organizing Maps (SOM). SOM is an *artificial neural network* consisting of many neurons. Every neuron

consists of a weight vector and a vector of positions in the map. Training SOM is done in an unsupervised manner applying *winner takes all* strategy. Every document vector is delivered to the network input several times. For every input vector the similarity with the neuron weight vector is computed. Weights of the most similar neuron and its neighbourhood are updated to be even more similar to the input pattern. The learning algorithm is constructed in such a way, that the neighbourhood and the degree of the weight updating is decreasing over time.

GHSOM address one of the most important drawback of SOM — the a priori definition of the map structure. Rauber *et al.* (2002) proposed an algorithm for growing SOM both in a terms of the number of map neurons and the hierarchy. After the training stage of SOM mean quantization error for every neuron i (mqe_i) is calculated as the average distance of every document recognised by the neuron i to its weight vector. The average MQE_j for whole map on level j is computed, too. If $MQE_j \geq \tau_1 \cdot MQE_{j-1}$ then the additional row or column of neurons is added to the map and the training stage is repeated. In the other case the mqe_i for every neuron is compared to MQE_j . If $mqe_i \geq \tau_2 \cdot MQE_{j-1}$ then another layer of the map is created for documents recognised by the neuron i .

4 Experiments in clustering

Evaluation of clustering algorithms can be done in many ways (cf. Forster, 2006). We argue that for the domain of documents the most suitable evaluation method is by referring to some *external criteria*, e.g. the comparison of the results with some pre-existing categories created manually. We used several measures for evaluation to capture different aspects of created groups. For measuring how homogeneous clusters are, the purity was applied:

$$Purity(L, C) = \frac{1}{N} \sum_k \max_j |c_k \cap l_j|, \quad (3)$$

where $C = \{c_1, c_2, \dots, c_k\}$ is a set of clusters, a $L = \{l_1, l_2, \dots, l_j\}$ — a set of pre-existing categories. $Purity(L, C) \in \langle 0, 1 \rangle$, where 1 is the best case. A drawback of *Purity* is its preference for solutions with large number of groups. Assigning every document to a different cluster gives $Purity = 1$ (Manning *et al.*, 2007).

The *Rand Index* measures accuracy on the basis of decisions performed for the subsequent document pairs¹: $R_I = \frac{TP+TN}{TP+FP+FN+TN}$. One of its drawbacks is the equal treatment of false positives and negatives. We can use standard IR measures of: precision $P = \frac{TP}{TP+FP}$, recall $R = \frac{TP}{TP+FN}$ and the harmonic mean of precision and recall $F_\beta = \frac{(\beta^2+1)PR}{\beta^2P+R}$.

For the evaluation we used a news collection which is a part of the IPI PAN Corpus (Przepiórkowski, 2004). Two test corpora were isolated: DZP_{98} — 25 486 articles from the „Dziennik Polski” newspaper from the year 1998 and DZP_{04} — 7 776 articles from „Dziennik Polski”, January to April, 2004. DZP_{98} was divided into a few general categories: *Kraków*, *Economy*, *Sport*, *Magazine*, *Home News*, *Word News*. In DZP_{04} regional categories were added.

¹Single decision can be one of: TP - true positive, TN - true negative, FN - false negative, FP - false positive.

We evaluated only the first layer groups (GHSOM) and groups on the top level of hierarchy (ROCK) to make the results comparable. The ROCK measure of similarity was set to the cosine between document vectors weighted by two functions: $tf.idf_{t,d} = tf_{t,d} \times \log \frac{N}{df_t}$ ($tf_{t,d}$ is a frequency of word t in document d , df_t is a number of documents containing word t , N is total number of documents) and *logent*. *Logent* was used as a weighting scheme by Landauer and Dumais (1997) in LSA. It combines the logarithmic scaling with the entropy normalization.

Impact of the θ parameter for ROCK on DZP_{04} for *logent* is shown in Tab. 1 and for *tf.idf* in Tab. 2. In Tab. 3 and 4 similar results for DZP_{98} are shown. As expected increasing value of θ improves *Purity*, Rand Index and precision, but decreases recall. Surprisingly using *logent* for weighting documents results in better overall accuracy of ROCK. The best results are for $\theta \geq 0.75$.

Clustering on DZP_{04} has a very low precision and recall. Careful manual inspection of the groups showed that many documents are ambiguously categorized, e.g. many articles about sports events were assigned to regional categories and not to sport. We did not find any mixing of major topics in groups, e.g. there was no document from *Sport* and *Economy* categories joined together. Additionally, the algorithm found more categories than actually present in the corpus, e.g. different sport disciplines were extracted into a group. An important drawback of ROCK is that it sometimes produces very deep and unbalanced hierarchy.

TABLE 1: Evaluation of ROCK on DZP_{04} using *logent*, k is number of roots.

θ	<i>Purity</i>	R_I	P	R	$F_{\beta=0.1}$	$F_{\beta=1}$	$F_{\beta=2}$	k
0,65	0,15	0,09	0,08	0,99	0,09	0,16	0,32	3
0,70	0,22	0,29	0,09	0,78	0,09	0,17	0,31	135
0,75	0,64	0,83	0,29	0,33	0,29	0,31	0,33	165
0,80	0,79	0,84	0,52	0,24	0,52	0,33	0,27	73

TABLE 2: Evaluation of ROCK on DZP_{04} using *tf.idf*, k is number of roots.

θ	<i>Purity</i>	R_I	P	R	$F_{\beta=0.1}$	$F_{\beta=1}$	$F_{\beta=2}$	k
0,65	0,15	0,09	0,08	0,99	0,09	0,16	0,32	3
0,70	0,16	0,13	0,09	0,94	0,1	0,17	0,31	50
0,75	0,4	0,81	0,16	0,25	0,16	0,19	0,22	209
0,80	0,64	0,89	0,46	0,09	0,44	0,16	0,11	170

GHSOM results are presented in Tab. 5. We fixed the number of neurons in the network to evaluate impact of the net structure on clustering. Because of the weight random initialization all results are given as an average from five runs of the algorithm for every parameter configuration. Surprisingly, results for different runs are almost the same. Changes in weighting scheme, training time and corpus did not produce significantly different results. Those observations suggest, that used measures of quality cannot capture all properties of the algorithm. Manual inspection of the clusters showed that with the increase of the training time

TABLE 3: Evaluation of ROCK on DZP_{98} using *logent*, k is number of roots.

θ	<i>Purity</i>	R_I	P	R	$F_{\beta=0.1}$	$F_{\beta=1}$	$F_{\beta=2}$	k
0,65	0,3	0,25	0,23	0,96	0,23	0,38	0,6	113
0,70	0,59	0,56	0,31	0,58	0,31	0,41	0,5	608
0,75	0,91	0,75	0,93	0,11	0,87	0,2	0,14	553
0,80	0,96	0,68	0,82	0,05	0,82	0,09	0,06	344

TABLE 4: Evaluation of ROCK on DZP_{98} using *tf.idf*, k is number of roots.

θ	<i>Purity</i>	R_I	P	R	$F_{\beta=0.1}$	$F_{\beta=1}$	$F_{\beta=2}$	k
0,65	0,3	0,24	0,23	0,97	0,23	0,38	0,6	68
0,70	0,41	0,41	0,24	0,68	0,24	0,35	0,5	489
0,75	0,76	0,75	0,44	0,75	0,43	0,17	0,12	793
0,80	0,89	0,75	0,84	0,02	0,6	0,04	0,02	698

TABLE 5: Evaluation of GHSOM on DZP_{04} using *tf.idf*.

Structure	<i>Purity</i>	R_I	P	R	$F_{\beta=0.1}$	$F_{\beta=1}$	$F_{\beta=2}$
7x7	0,88	0,22	0,83	0,03	0,65	0,05	0,03
6x6	0,88	0,23	0,83	0,04	0,7	0,08	0,05
5x5	0,88	0,23	0,79	0,05	0,69	0,09	0,06
4x4	0,88	0,25	0,82	0,07	0,75	0,14	0,09
3x3	0,88	0,29	0,8	0,14	0,77	0,23	0,16
2x2	0,88	0,38	0,81	0,28	0,8	0,4	0,3

documents are placed more evenly on the map and that *logent* resulted in better topically separated neuron groups. We noted that there is more false positives decision made by GHSOM, but it produces more balanced hierarchy then ROCK.

5 Towards thesaurus extraction

We wanted to label document groups clustered in a hierarchical tree with *representative words*. Words describing groups of documents closer to the root of the tree should be more general than words describing documents in the leafs. Ideally we would like to obtain hierarchical thesaurus out of the created labels.

Keyword extraction can be done in a supervised or unsupervised manner. As the supervised algorithms requires manually constructed resources, we applied only unsupervised methods. They try to capture some statistical properties of words occurrences to identify words which are the best descriptors of the given document. The statistics can be counted locally, using data from a single document only or estimated from large body of text. To benefit from merits of both local and global strategies we extended the method proposed by Indyka-Piasecka (2004) with Matsuo and Ishizuka (2004) algorithm into a hybrid keyword extraction method.

Indyka-Piasecka (2004) assigns a weight w to every word t that occurs in every document of a group. Additionally, words are filtered on the basis of their *document frequency* df_t , that is a number of documents in which the word t occurred. Both seldom and often words are not good discriminator of document content (cf. Indyka-Piasecka, 2004). Weight w is calculated by using two weighting schemes: *tf.idf_{t,d}* and *cue validity* $cv = \frac{tf_{group}}{tf}$, where tf is abbreviation of term frequency.

Matsuo and Ishizuka (2004) used a three-step-process to calculate a word weight. First, all words in a document are reduced to the base morphological form and filtering on the basis of term frequency and stoplist is done. Next, they cluster words in a document using two algorithms. If two words have a similar distribution, then they belong to the same group. As a similarity measure of probability distribution between words w_1 and w_2 they use the Jensen-Shannon divergence:

$$J(w_1, w_2) = \log 2 + \frac{1}{2} \sum_{w' \in C} \{h(P(w'|w_1) + P(w'|w_2)) - h(P(w'|w_1) - P(w'|w_2))\}, \quad (4)$$

where $h(x) = x \log(x)$ and $P(w'|w_1) = \text{frequency}(w', w_1) / \text{frequency}(w_1)$. When $J(w_1, w_2) > 0.95 \cdot \log 2$, the words are put into one group. Words are also clustered when they follow a similar co-occurrence pattern with the other words. This can be measured by *Mutual Information*:

$$MI(w_1, w_2) = \log \frac{P(w_1, w_2)}{P(w_1)P(w_2)} = \log \frac{N_{total} \cdot \text{freq}(w_1, w_2)}{\text{freq}(w_1) \cdot \text{freq}(w_2)}, \quad (5)$$

where N_{total} is total number of words in a document. If $MI(w_1, w_2) > \log 2$ then words belong to the same group.

After constructing clusters the weight w is assigned to each word using χ^2 for testing if there is a bias in word occurrences with group. Combining the methods of Indyka-Piasecka (2004) and Matsuo and Ishizuka (2004) word weight is calculated in the following way:

$$w_t = \alpha \cdot \min_{tf.idf_t} + \beta \cdot cv_t + \gamma \cdot \chi_t^2, \quad (6)$$

where $\min_{tf.idf_t}$ is a minimum of term frequency for documents in cluster, α , β , γ are parameters controlling impact of every measure on the final value. Words with the highest weights are used as labels for groups.

We evaluated this approach using plWordNet (Derwojedowa *et al.*, 2007) — we compared hyponymy hierarchy of the Polish wordnet with the automatically created thesaurus. This approach failed, only 86 relation of hyponymy were present in thesaurus (less than 1% of all relations). Clustering whole documents might be a factor for low accuracy, but experiments with segments decreased quality of clustering. On the other hand, methods for keyword extraction developed primarily for IR are not suitable for the extraction of relations between words describing different groups of documents.

Nevertheless, the extracted group labels are very descriptive. For example a group of documents about “interventionist purchase of grain and harvest on the area of Małopolska” are labelled with: *zboże (grain)*, *pszenica (wheat)*, *tona*

(*tonne*), *rolnik* (*farmer*) and *agencja* (*agency*). Another possible use of extracted words is measuring degree of polysemy, because different meanings of words occurs in different branches of hierarchy. Labels also will help us in choosing which cluster to use for training domain MSR.

6 Conclusions and further research

We have explored feasibility of document clustering for supporting semi-automatic methods of construction of lexical semantic networks for Polish. After selecting appropriate clustering algorithms we have applied them to a large corpus of Polish news documents. ROCK achieved precision of 93% in reconstruction of the manually built categories. Although recall is low, purity of cluster is more important for further processing than ability to assign every document to a group.

We have made a preliminary experiments in an atypical approach to the extraction of lexical semantic relations from a hierarchy of document clusters. Even if this approach did not yield a fully satisfying results, as a side effect we have labelled the hierarchy in descriptive way which will be helpful in the future. Low accuracy of keyword extraction techniques is caused partially by documents describing more than one topic. That is why we have examined addition of an extra step of dividing documents into semantically coherent fragments. Unfortunately this lead to a decrease in the quality of clustering.

We plan to extend this work by designing better way for extracting keywords from document groups. On the other hand, using additional features can lead to better quality of clustering algorithms on document fragments. Last but not least, we will join approaches based on clustering documents and measures of distributional similarity based on words co-occurrences.

References

- Eneko AGIRRE and Philip EDMONDS (2006), *Word Sense Disambiguation: Algorithms and Applications (Text, Speech and Language Technology)*, Springer-Verlag New York, Inc., Secaucus, NJ, USA, ISBN 1402048084.
- Bartosz BRODA (2007), Mechanizmy grupowania dokumentów w automatycznej ekstrakcji sieci semantycznych dla języka polskiego.
- J. CURRAN (2004), *From Distributional to Semantic Similarity*, Ph.D. thesis.
- Magdalena DERWOJEDOWA, Maciej PIASECKI, Stanisław SZPAKOWICZ, and Magdalena ZAWISŁAWSKA (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*, pp. 169–178, Universität Tübingen.
- Richard FORSTER (2006), *Document Clustering in Large German Corpora Using Natural Language Processing*, Ph.D. thesis, University of Zurich.
- Michel GALLEY, Kathleen R. MCKEOWN, Eric FOSLER-LUSSIER, and Hongyan JING (2003), Discourse Segmentation of Multi-Party Conversation, in Erhard HINRICHS and Dan ROTH, editors, *Proceedings of the 41st Annual Meeting of ACL (ACL-03)*, pp. 562–569, Sapporo, Japan.
- Sudipto GUHA, Rajeev RASTOGI, and Kyuseok SHIM (2000), ROCK: A Robust Clustering Algorithm for Categorical Attributes, *Information Systems*, 25(5):345–366.

- Marti A. HEARST (1997), TextTiling: Segmenting Text into Multi-Paragraph Subtopic Passages, *Computational Linguistics*, 23(1):33–64.
- Agnieszka INDYKA-PIASECKA (2004), *Modele użytkownika w internetowych systemach wyszukiwania informacji*, Ph.D. thesis, Politechnika Wrocławska.
- A. K. JAIN, M. N. MURTY, and P. J. FLYNN (1999), Data clustering: a review, *ACM Computing Surveys*, 31(3):264–323.
- T. KOHONEN, S. KASKI, K. LAGUS, J. SALOJRVI, J. HONKELA, V. PAATERO, and A. SAARELA (2000), Self organization of a massive document collection, *IEEE Transactions on Neural Networks*, 11:574–585.
- Manu KONCHADY (2006), *Text Mining Application Programming*, Charles River Media.
- T. LANDAUER and S. DUMAIS (1997), A Solution to Plato’s Problem: The Latent Semantic Analysis Theory of Acquisition., *Psychological Review*, 104(2):211–240.
- Dekang LIN (1998), Automatic Retrieval and Clustering of Similar Words, in *COLING 1998*, pp. 768–774, ACL.
- Christopher D. MANNING, Prabhakar RAGHAVAN, and Hinrich SCHÜTZE (2007), *Introduction to Information Retrieval*, Cambridge University Press, to appear.
- Yutaka MATSUO and Mitsuru ISHIZUKA (2004), Keyword extraction from a single document using word co-occurrence statistical information., *International Journal on Artificial Intelligence Tools*, 13(1):157–169.
- Dariusz MAZUR (2005), *Metody grupowania i ich implementacja do eksploracji danych postaci symbolicznej*, Ph.D. thesis, Politechnika Śląska.
- Patrick Andre PANTEL (2003), *Clustering by committee*, Ph.D. thesis, Edmonton, Alta., Canada, Canada, adviser-Dekang Lin.
- Maciej PIASECKI, Stanisław SZPAKOWICZ, and Bartosz BRODA (2007), Automatic Selection of Heterogeneous Syntactic Features in Semantic Similarity of Polish Nouns, in *Proceedings of the Text, Speech and Dialogue Conference*.
- Adam PRZEPIÓRKOWSKI (2004), *The IPI PAN Corpus: Preliminary version*, Institute of Computer Science, Polish Academy of Sciences, Warsaw.
- A. RAUBER, D. MERKL, and M. DITTENBACH (2002), The growing hierarchical self-organizing maps: exploratory analysis of high-dimensional data.
- Adam SCHENKER, Horst BUNKE, Mark LAST, and Abraham KANDEL (2005), *Graph-Theoretic Techniques for Web Content Mining*, World Scientific Publishing Co. Pte. Ltd.