ISBN 978-83-60810-27-9
ISSN 1896-7094

Is Shallow Semantic Analysis Really That Shallow? A Study on Improving Text Classification Performance

Przemysław Maciołek, Grzegorz Dobrowolski
AGH University of Science and Technology
Al. Mickiewicza 30, 30-059 Kraków, Poland
Email: {pmm,grzela}@agh.edu.pl

***Abstract*—The paper presents a graph-based, shallow semantic analysis-driven approach for modeling document contents. This allows to extract additional information about meaning of text and effects in improved document classification. Its performance is compared against the "legacy" bag-of-words and Schenker et al. approaches with $k-NN$ classification based on Polish and English news articles.**

I. Introduction

RESEARCH on computational linguistics has over 50 years of history. It is currently considered as an interdisciplinary field covering topics such as linguistics, psychology, cognitive science, artificial intelligence and others. While the research is done for more than a half of a century, there are still many basic open issues. One of such notable problems is relatively poor ability of machine text content understanding.

The state-of-the-art in modeling text meaning for document classification purposes is practically still basing on a statistical bag-of-words approach, developed in early 70's. In this method, single words are extracted and represented by a vector with their frequency.

Although many enhancements have been proposed to this approach, it has essential drawback: substantially limited amount of information that can be "captured" by such representation. Omission of information about words order is a striking example.

Numerous solutions leveraging graph-based representation of text were presented. One of such method was proposed by Schenker, Last, Bunke and Kandel [23], [22], [21]. The taken approach built document model as graph, where each sequentially found word was added as a next node. Also, links between nodes were tagged with information about section of the document where they were found (such as *title*, *body* or *link*). Thus instead of just calculating frequency of word occurrences, information about their order is also stored.

Other graph-based methods include graph node ranking method similiar to *PageRank* [15], analyzing structure of the document only (ignoring its contents) [8], extracting document features from previously built tree-like structures [10] and extracting features from Schenker *et al.* based graphs [14].

In this paper a new method is presented. Like in the Schenker *et al.* approach, document contents is represented using a graph model. However, the way the graph is built is different and takes into account part-of-speech information about each word. Also, semantic dictionary (such as *WordNet*) is used in the build process. The reasoning behind it is based on observations of language acquisition by children [24], [16] and statistical part-of-speech usage [12], [2].

The new method is compared to well-known Schenker *et al.* and bag-of-words approaches in document classification task. Test results are presented and commented in the final section of the paper.

II. Document Modeling Approaches

In this section, baseline text modeling methods are presented. As a prerequisite, they require processing raw document contents into a form that allows to extract relevant document features. The process might be summarized as an algorithm consisting of following steps:

1) Reading text from a source. Segmentation into sentences and words. Converting all letters to common case.
2) Removing stop-words. These are frequently found words, that do not provide substantial information about document contents (as they are commonly found in any kind of text). The stop-words list is arbitrary and might contain from zero to hundreds of such words. Typical elements are: *"the", "in", "he", "one", "of", "is"*, etc.
3) Word stemming. In this process, the inflected or derived words are reduced to their stems (non-changeable parts). This allows properly treating words of the same base used in different forms (e.g. *"thankfully"* → *"thank*"*, *"thanks"* → *"thank*"*, etc.).

Vector Space

The most commonly used approach for document representation is to count each word occurrence, calculate its frequency and put it into a vector space. This allows to easily find distance between any two documents (vectors) using a selected metric, such as *Euclidean* or *Jaccard* distance [13].

It is worth noting that while the general idea of vector space text representation has not changed during the last 40 years, many methods improving selected vector features quality have emerged, such as *Latent Semantic Analysis*.

978-83-60810-27-9/09/$25.00 © 2010 IEEE

Schenker Text-to-Graph Approach

Schenker *et al.* [23] proposed a method (actually a variant of more general *text-to-graph* approach) for building a graph from hypertext. The process first marks three sections of the text: *title*, *links* and *text*. Next, the graph is being built using following rules:

1) If word (stem) A occurs for the first time, then new node A is created.
2) If word B occurs after word A in section 1, a connection $A \rightarrow B$ is created with label 1.

To reduce the graph size, only the n most relevant words are selected from the document for building the graph. Word occurrence counters are not incorporated (in the typical variant).

Definition 1: To calculate distance between any two graphs, following metrics are proposed:

$$dist_1(G_1, G_2) = 1 - \frac{|mcs(G_1, G_2)|}{max(|G_1|, |G_2|)} \quad (1)$$

$$dist_2(G_1, G_2) = 1 - \frac{|mcs(G_1, G_2)|}{|G_1| + |G_2| - |mcs(G_1, G_2)|} \quad (2)$$

$$dist_3(G_1, G_2) = 1 - \frac{|mcs(G_1, G_2)|}{|MCS(G_1, G_2)|} \quad (3)$$

where:
$mcs(G_1, G_2)$ - maximum common subgraph of graphs G_1 and G_2
$MCS(G_1, G_2)$ - minimum common supergraph of graphs G_1 and G_2
$|G|$ - size of graph G, defined as a sum of numbers of nodes (vertices) and edges: $|G| = |V| + |E|$

To calculate the distance, maximum common subgraph has to be found first. Finding such subgraph is, in general, a NP-complete problem. However, when taken into account that in the created graphs each node has a unique label, it is possible to construct an algorithm finding the solution with $O(|V|^2)$ computational complexity [6], [21].

The originally presented approach (*standard representation*) was further extended. One of the interesting variants was incorporating information about word counts into the graph model and discarding information about document section (such as *link*, *body* and *title*). Each node and edge was labeled with additional information about number of times each associated term appeared in the document (in case of nodes) and number of times two nodes were connected together (in case of edges). These numbers were stored as absolute values (*absolute frequency representation*) or as normalized values - divided by the maximum number of node occurrences in document (*relative frequency representation*).

Definition 2: Graph size used for models using frequency information is defined as:

$$|G|' = \sum_{i=1}^{|V|} v(i) + \sum_{j=1}^{|E|} e(i)$$

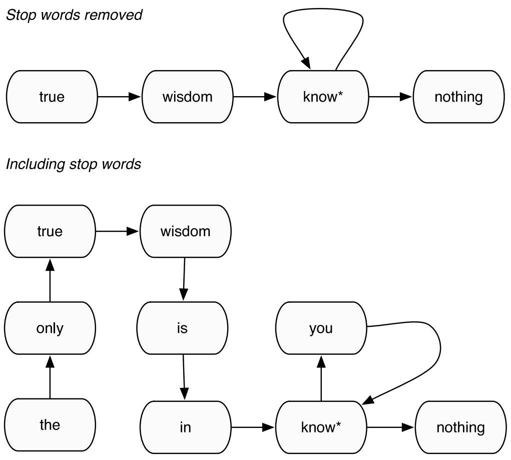

Fig. 1. Example of Schenker graph modeling for sentence *"The only true wisdom is in knowing you know nothing"*. Note: *know** is a stem of both *know* and *knowing*.

where:
$|V|$ - number of nodes in graph G,
$|E|$ - number of edges in graph G,
$v(i)$ - frequency associated to node i
$e(j)$ - frequency associated to edge j

According to the results obtained by Schenker, the *standard representation* provided generally better results for hypertext (HTML) documents in comparison to other *text-to-graph* algorithm variants. When information about links and title was not incorporated (such as when using *simple representation*) the performance slightly decreased. It might be expected that for documents with low number of hyperlinks (or without any of them) the frequency-type representation will provide the best results.

III. MOTIVATION

Almost any linguistics-related problem is still solved drastically better by humans than by machines. A human language is considered as one of the greatest achievements of evolution, practically unique for mankind. Even after so many years of research, we still possess relatively not too much knowledge about the way it really works. Basically, we do not even understand the mechanism in which the language is acquired (learned).

Our lack of knowledge in this area does not allow creating a general tool that would render a text understanding performance comparable to humans. However, while we do not (yet) see the "whole picture", a lot of observations about language development were collected during the years. This knowledge may help to improve computational linguistics mechanisms.

Verbs Absence During Language Acquisition

An important observation is that children learn primarily nouns, even if they can observe other parts of speech with similar frequency [24]. There is a huge disproportion: when child dictionary contains between 20-50 words, as much as

45% of them are nouns, and only 3% of them are verbs. While these numbers were observed for English language, in case of other languages the situation is very similar [3].

This finding might be linked as a consequence of the fact, that nouns are basically used for object labeling (such as "mom", "dad", "bed", "cat") and thanks to this they are relatively firm. The child usually knows elements (classes) which are named by nouns. On the other hand, the verbs might describe an action, state or occurrence and might have different meaning depending on the arguments and context used. Also, they might be often found in sentences describing abstract ideas. In effect, an advance of child development is required for possessing skills necessary for understanding and using verbs.

Polysemy and Parts of Speech Distribution

The WordNet 3.0 [7] database contains more than 155,000 unique strings, which are categorized either as nouns (117798 unique strings), verbs (11529), adjectives (21479) or adverbs (4481). Each string is assigned to one or more synsets (groups of words with similar meaning).

As it might be observed, there's notable difference in ratios of polysemous to monosemous words for different parts of speech (see table I). It is noticeably higher for verbs in comparison to other parts of speech.

It is reversed in case of nouns. Even if the absolute number of them is much higher than any other part of speech, the ratio is very low. Similar findings might be also found when analyzing large text corpora for other languages [12], [2].

TABLE I
WORDNET POLYSEMY STATISTICS [7]

POS	AVERAGE POLYSEMY (INCLUDING MONOSEMOUS WORDS)
NOUN	1.24
VERB	2.17
ADJECTIVE	1.40
ADVERB	1.25

Summary of Observations

An analogy can be observed: capabilities of machine text processing (classification, information retrieval, etc.) might be compared to skills of a young child, which does not yet posses general knowledge about the surrounding world.

Based on the presented observation, a following hypothesis might be suggested: the sub-optimal methodology for unstructured text processing is similar to the one observed for children language capabilities. That is, an emphasis should be put on information that are not troubled by ambiguity (such as nouns). Also, because there are many ways an action can be represented with different verbs, a method for generalizing their meaning (for example using synonyms from semantic dictionary) could improve machine text representation.

IV. SHALLOW SEMANTIC ANALYSIS FOR BUILDING A GRAPH MODEL

The method presented in this section is a special case of a more general solution, which is a member of family of methods based on the shallow semantic analysis approach for building a graph model. The algorithm presented in this paper is a variant, that was experimentally found to be sub-optimal for a document classification task. It is based on observations presented in the previous section and simplifications in predicate-argument sentence decomposition.

As a prerequisite, a number of steps must be performed on the input text. The complete process of transformation contains following phases:

1) Reading raw text from a source.
2) Text segmentation and removal of non-characters.
3) Tagging parts of speech; *Stanford POS Tagger*[1] [27], [26] is used for English and *TAKIPI*[2] [17] for Polish documents. The cited accuracy is approx. 97% for *Stanford POS Tagger* and 93.4% for *TAKIPI* (in the latter case, including gender and case resolution).
4) Finding synsets using semantic dictionary; *WordNet*[3] [7] and *plWordNet*[4] [18] are used for this purpose.
5) Word stemming. In case of English documents *Snowball Stemmer*[5] was used. For Polish, *Morfologik*[6] was chosen.

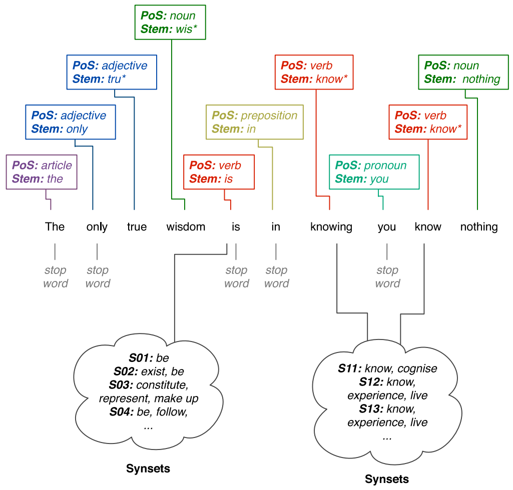

Fig. 2. Example of sentence decomposition

The processed document contents is sequentially analyzed, word after word. Graph representing the text is built using the following set of rules:

1) If a word A is a noun or adjective, and a node with label A does not yet exist, it is created. In case the word occurs rarely in the whole data set (e.g. less than

[1] http://nlp.stanford.edu/software/tagger.shtml
[2] http://nlp.ipipan.waw.pl/TaKIPI/
[3] http://wordnet.princeton.edu
[4] http://plwordnet.pwr.wroc.pl/
[5] http://snowball.tartarus.org/
[6] http://morfologik.blogspot.com

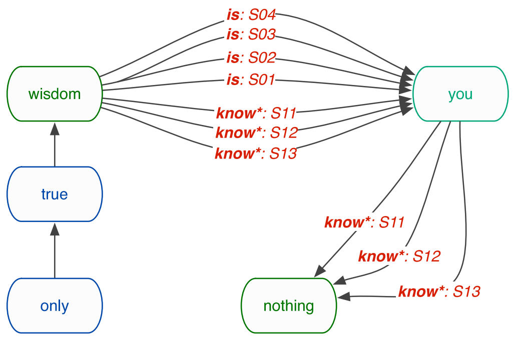

Fig. 3. Example of shallow semantic analysis based text-to-graph transformation result

predefined number of times), it is being replaced by its synsets.

2) If a noun or adjective B is found after another noun or adjective A, then a connection with empty label is created between nodes labeled with A and B.
3) If a verb C was found between nouns or adjectives A and B, then all synsets (sets of words with similar meaning and the same part-of-speech) of C are being added as connection between nodes A and B, labeled with synset identifier.
4) If adjective E is found between two nouns D and F, an additional direct connection between D and F is created.

New distance measures are proposed. They take into account not only the single maximum common subgraph, but all nodes and edges that were found to be common:

Definition 3: To calculate distance between any two shallow semantic analysis created graphs following measures will be used:

$$dist_4(G_1, G_2) = 1 - \frac{|G_1 \cap G_2|}{|G_1| + |G_2| - |G_1 \cap G_2|} \quad (4)$$

$$dist_5(G_1, G_2) = 1 - \frac{|V_1 \cap V_2|}{|V_1| + |V_2| - |V_1 \cap V_2|} \quad (5)$$

where:
$|V_a|$ - number of nodes in graph G_a.

The second of the new metrics ($dist_5$) takes into account number of nodes rather than the size of graph. It might be considered that such approach allows to find how many common subjects are raised in any two compared documents, especially giving the way the graph is being build, as the nodes are mostly constituted by nouns.

V. Tests

The analysed methods are tested using typical automated document classification scenario. It is based on assumption that having examples of documents from different classes it should be possible to automatically assign correct classes to previously unseen documents, as the distance between similar documents (that is from similar classes) should be generally minimal.

In presented scenario, $k - NN$ (k-Nearest Neighbors) classification algorithm is used. Each of the tested document collections is randomly split into two subsets: *training* and *test*. Next, for each document in *test* set, k nearest documents from *training* set are found. After that, a voting is performed among them. The most common occurring category is assigned to the *test* document.

To test usefulness of the new method for document classification problem, the following document collections (both English and Polish) have been chosen:

- *wiadomosci24.pl* - containing 1500 short articles from one of the leading internet news services in Poland. Each document is tagged with 1 to 5 out of 50 tags, such as: *Gdansk, Festival, Politics, Warsaw, Money, Culture, ...*
- *Rzeczpospolita*[7] - a 800 documents subset of randomly chosen articles published in 2002 in one of major Polish newspapers. They are tagged with one of eight tags: *World, Culture, Law, Publicism-Commentary, Sport, Poland, Economy, Plus Minus*.
- *Reuters*[8] - a subset of 1800 articles randomly chosen from the *famous* Reuters-21578 "Lewis Split" data set. Each document is tagged with at least one out of 32 categories, such as: *wheat, trade, acq, ship, money-fx, etc.*
- *PDDP K-series*[9] [1] - a subset of 800 randomly chosen (out of 2340) *Yahoo!* articles originally extracted by Daniel Boley and used by various researchers for testing document classification performance (including Schenker). Each of them is tagged with 1 of 20 classes such as: *politics, tech, entertainment, business, etc.*

Benchmarks

Quality of results is typically measured using *precision* and *recall*. Instead of presenting both of these numbers, it is a common practice (used also in this paper) to present their harmonic mean also known as *F-measure*:

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \quad (6)$$

As typically there are more than two classes in total being categorized, two approaches might be used for calculating the average results. The first one is calculating the precision and recall for each category and taking an average. It's called *macro-averaging*. The second one (*micro-averaging*) gives equal weight to every document (rather than category). In such case, precision and recall numbers are calculated for each document and averaged.

Results

For each document collection, considered here as a separate test case, documents are randomly split into training and test sets. Only documents from the ten most frequently occurring tags are selected.

[7] http://www.cs.put.poznan.pl/dweiss/rzeczpospolita
[8] http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html
[9] http://www-users.cs.umn.edu/~boley/ftp/PDDPdata/

The method parameters (such as maximum number of nodes in a graph, maximum number of features in bag-of-words approach, minimum word count to be selected as a feature, k-number, etc.) were optimized to achieve best results for each of the methods. In other words - multiple tests with varying parameters were run and the best case results are presented.

The tables and figures presented below show results for test/training set ratio of 0.5, so there is equal number of test and training documents. Tests were performed also for other ratios (in $0.1 - 0.9$ range) and relative results were similar (i.e. shallow semantic analysis based approach provided better results than other tested methods).

For *bag-of-words* method both *Jaccard* and *cosine* similarity (with *tf-idf*) were tested. Results for the better performing measure are presented. This is also the case for both graph methods - only the best performing metrics are presented in the results, even if all were tested.

The Schenker approach was implemented according to the description found in [23], [22], [21].

TABLE II
Results - Rzeczpospolita

		Micro	Macro
		Avg. F1	Avg. F1
Bag-of-words	Jaccard	0.68	0.65
Schenker	$dist_3$	0.73	0.70
	$dist_4$	0.70	0.66
Random		0.30	0.16
Shallow semantic analysis	$dist_4$	0.74	0.69
	$dist_5$	0.72	0.69

Polish Texts - Rzeczpospolita: Documents in this collection were tagged with the lowest number of tags in comparison to other sets. In effect, the categories assigned were more general. The only geographic tags were *world* and *Poland*.

TABLE III
Results - wiadomosci24.pl

		Micro	Macro
		Avg. F1	Avg. F1
Bag-of-words	Jaccard	0.46	0.47
Schenker	$dist_3$	0.47	0.47
	$dist_4$	0.46	0.49
Random		0.17	0.17
Shallow semantic analysis	$dist_4$	0.50	0.51
	$dist_5$	0.52	0.50

Polish Texts - wiadomosci24.pl: Lower (in comparison to other test cases) absolute results for *wiadomosci24.pl* articles might be related to the way the tags were assigned, as there were many disambiguations found. For example - document was tagged as *Sport* while in fact it was more related to other tags available for the set, such as: *Lodz, Football.*

English Texts - Reuters-21578: Reuters results provide highest absolute values. There are many reasons for this. One of them is the fact that the original collection was analyzed and "cleaned up" - so it might be expected that many incorrectly assigned tags were removed and some of the disambiguations fixed. Also, it was used as a training set for Stanford part-of-speech tagger, thus it should be most correctly tagged.

TABLE IV
Results - Reuters 21578

		Micro	Macro
		Avg. F1	Avg. F1
Bag-of-words	Jaccard	0.87	0.73
Schenker	$dist_3$	0.87	0.71
	$dist_4$	0.87	0.72
Random		0.46	0.13
Shallow semantic analysis	$dist_4$	0.89	0.78
	$dist_5$	0.88	0.75

TABLE V
Results - PDDP K-series

		Micro	Macro
		Avg. F1	Avg. F1
Bag-of-words	Jaccard	0.81	0.78
Schenker	$dist_3$	0.83	0.79
	$dist_4$	0.84	0.80
Random		0.22	0.15
Shallow semantic analysis	$dist_4$	0.85	0.81
	$dist_5$	0.84	0.81

English Texts - PDDP K-series: PDDP collection was a subject of tests by Schenker *et al.* Thus it was interesting to see how the shallow semantic analysis approach will compare to it. It is worth noting that size of the document collection used here is about twice as large as the collection size used by Schenker [22].

Analysis

For each of the test cases and benchmarks, except one (macro averaged F1 measure for *Rzeczpospolita* articles), shallow semantic analysis method presents results better than both the vector space and Schenker approaches. The improvement is most notable for *wiadomosci24.pl* and *Reuters* articles. The probable cause of this is the fact that *Rzeczpospolita* and *PDDP K-series* collections have too general tags assigned. For example - in case of *PDDP K-series* there is a single tag related to *business*, while for *Reuters* document collections there are: *acq*, *trade*, *money-fx* and others. It is similar for *Rzeczpospolita* vs. *wiadomosci24.pl* collections. The first one has only 8 possible tags, while the latter has 50 classes.

VI. Conclusions

A new method for building graph representation of text is presented. With a help of the part-of-speech tagger and semantic dictionary it is performing a shallow semantic analysis of input document producing a model reminiscent of semantic net. Preliminary experiments have been performed for both Polish and English documents to check its practical usability.

The obtained results show that the proposed approach has a slight edge over Schenker and bag-of-words approaches. This suggests that the new method is able to produce graphs better representing the latent meaning of document, even if it effectively uses only some of the terms from the original text representation.

It is important to note that the proposed algorithm, presented in section IV, is a current variant rather than a final version. Depending on the features of target documents and regimes, it might be accordingly modified and extended.

As for the metrics, it has been found that $dist_3$ produces the best results for Schenker method (with a close call for $dist_4$). For shallow semantic analysis approach, $dist_4$ gives the strongest performance. While the new proposed metrics work generally well for both graph methods, the $dist_1$, $dist_2$ and $dist_3$ produce poor results for the shallow semantic analysis.

The presented method is a subject of intensive research. The current focus is directed on two aspects. The first one is leveraging the *Anaphora Resolution* in the method, which should effect in even better classification results. The second is extracting weighted features from the graph (e.g. as presented by Markov *et al.* [14]) and using them with more sophisticated classifier (such as SVM).

References

[1] D. Boley. Principal direction divisive partitioning. *Data Mining and Knowledge Discovery*, 2:325–344, 1997.

[2] L. Borin and K. Prütz. Through a glass darkly: Part-of-speech distribution in original and translated text. *Language and Computers*, 15, 2000.

[3] M.C. Caselli, E. Bates, P. Casadio, J. Fendon, L. Fenson, L. Sanderl, and J. Weir. A cross-linguistic study of early lexical development. *Cognitive Development*, 1995.

[4] Tommy W. S. Chow, Haijun Zhang, and M. K. M. Rahman. A new document representation using term frequency and vectorized graph connectionists with application to document retrieval. *Expert Syst. Appl.*, 36(10):12023–12035, 2009.

[5] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. Wiley, 1991.

[6] Peter J. Dickinson, Horst Bunke, Arek Dadej, and Miro Kraetzl. On graphs with unique node labels. In Hancock and Vento [9], pages 13–23.

[7] Ch. Fellbaum, editor. *WordNet - An Electronic Lexical Database*. The MIT Press, 1998.

[8] Peter Geibel, Ulf Krumnack, Olga Pustylnikov, Alexander Mehler, Helmar Gust, and Kai-Uwe KÃijhnberger. Structure-sensitive learning of text types. In Mehmet A. Orgun and John Thornton, editors, *Australian Conference on Artificial Intelligence*, Lecture Notes in Computer Science, pages 642–646. Springer, 2007.

[9] Edwin R. Hancock and Mario Vento, editors. *Graph Based Representations in Pattern Recognition, 4th IAPR International Workshop, GbRPR 2003, York, UK, June 30 - July 2, 2003, Proceedings*, volume 2726 of *Lecture Notes in Computer Science*. Springer, 2003.

[10] Chuntao Jiang, Frans Coenen, Robert Sanderson, and Michele Zito. Text classification using graph mining-based feature extraction. In *SGAI Conf.*, pages 21–34, 2009.

[11] Jurafsky, Daniel, Martin, and H. James. *Speech and Language Processing (2nd Edition) (Prentice Hall Series in Artificial Intelligence)*. Prentice Hall, 2 edition, 2008.

[12] G. Leech, P. Rayson, and A. Wilson. *Word Frequencies in Written and Spoken English: based on the British National Corpus*. 2001.

[13] C. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 1 edition, 2008.

[14] A. Markov, M. Last, and A. Kandel. The hybrid representation model for web document classification. *Int. J. Intell. Syst.*, 23(6):654–679, 2008.

[15] Rada Mihalcea and Paul Tarau. TextRank: Bringing order into texts. In *Proceedings of EMNLP-04 and the 2004 Conference on Empirical Methods in Natural Language Processing*, 2004.

[16] E. Newport, H. Gleitman, and L. Gleitman. Mother i'd rather do it myself: Some effects and non-effects of maternal speech style. *Talking to Children: Language Input and Acquisition*, 1977.

[17] M. Piasecki. Polish tagger TaKIPI: Rule based construction and optimisation. *Task Quarterly*, 11(1–2):151–167, 2007.

[18] Maciej Piasecki, StanisÅĆaw Szpakowicz, and Bartosz Broda. *A Wordnet from the ground up*. Oficyna wydawnicza Politechniki WrocÅĆawskiej, WrocÅĆaw, Polska, 2009.

[19] S. Pinker, L. R. Gleitman, and M. Liberman (Eds.). *An Invitation to Cognitive Science, Vol. 1 Language*. The MIT Press, 2 edition, 1995.

[20] M. F. Porter. An algorithm for suffix stripping. *Program*, 1980.

[21] A. Schenker, H. Bunke, M. Last, and A. Kandel. *Graph-Theoretic Techniques for Web Content Mining (Machine Perception and Artificial Intelligence) (Series in Machine Perception and Artificial Intelligence)*. World Scientific Publishing Co., Inc., River Edge, NJ, USA, 2005.

[22] A. Schenker, M. Last, H. Bunke, and A. Kandel. Classification of web documents using a graph model. In *Proceedings of the Seventh International Conference on Document Ana lysis and Recognition*, 2003.

[23] A. Schenker, M. Last, H. Bunke, and A. Kandel. A comparison of two novel algorithms for clustering web documents. In *Second International Workshop on Web Document Analysis*, Edinburgh, UK, 2003.

[24] J. Snedeker and L. Gleitman. *Weaving a Lexicon*, chapter Why it is hard to label our concepts, pages 255–293. Bradford Book, 2004.

[25] A. Strehl, J. Ghosh, and R. Mooney. Impact of similarity measures on web-page clustering. In *AAAI-2000: Workshop of Artifical Intelligence for Web Search*, 2000.

[26] K. Toutanova, D. Klein, C. Manning, and Y. Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of HLT-NAACL 2003*, 2003.

[27] K. Toutanova and C. D. Manning. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 2000.