Proceedings of International Joint Conference on Neural Networks, Orlando, Florida, USA, August 12-17, 2007

Neurolinguistic Approach to Vector Representation of Medical Concepts

Włodzisław Duch, *Senior Member, IEEE*, Paweł Matykiewicz and John Pestian

***Abstract*—Putative brain processes responsible for understanding language are based on spreading activation in semantic networks, providing enhanced representations that involve concepts not found directly in the text. Approximation of this process is of great practical and theoretical interest. Vector model should reflect activations of various concepts in the brain spreading through associative network. Medical ontologies are used to select concepts of specific semantic type and add to each of them related concepts, providing expanded vector representations. The process is constrained by selection of useful extensions for the classification task. Short hospital discharge summaries are used to illustrate how this process works on a real, very noisy data. Results show significantly improved clustering and classification accuracy. A practical approach to mapping of associative networks of the brain involved in processing of specific concepts is presented.**

I. Introduction

Semantic internet and semantic search in free text databases require automatic tools for annotation. In medical domain terabytes of text data are produced and conversion of unstructured medical texts into semantically-tagged documents is urgently needed because errors may be dangerous and time of experts is costly. Our long-term goal is to create tools for automatic annotation of unstructured texts, especially in medical domain, adding full information about all concepts, expanding acronyms and abbreviations and disambiguating all terms. This goal cannot be accomplished without solving the problem of word meaning and knowledge representation. So far the only systems that can deal with linguistic structures are human brains. Neurocognitive approach to linguistics "is an attempt to understand the linguistic system of the human brain, the system that makes it possible for us to speak and write, to understand speech and writing, to think using language …" [1].

Although neurocognitive linguistics approach has been quite interesting and fruitful in understanding the neuropsychological language-related problems it has not been so far that useful in creation of algorithms for text interpretation, and it is still exotic in the natural language processing (NLP) community. The basic premise is rather simple: each word in analyzed text is a part of an associative network where activation spreads and the states of the networks facilitate semantic interpretation of the text, yet unraveling these "pathways of the brain" [1] is quite difficult.

Connectionist approach to natural language has been introduced in the influential PDP books [2]. Application of constrained spreading activation techniques in information retrieval [3], semantic search techniques [4] and word sense disambiguation [5] have recently been made. In this paper neurolinguistic inspirations are used to find useful approximations to spreading of brain activity during text comprehension. A database of short medical documents divided into ten categories is used for illustration of this approach. In the next section neurocognitive inspirations for NLP are presented, in the third section medical data used in experiments are described, followed by the algorithm that creates semantic description vectors, and by applications of this algorithm to categorization of medical texts. Discussion of results and their wider implications closes this paper.

II. Neurocognitive inspirations

How are words and concepts represented in the brain? The neuroscience of language in general, and word representation in the brain in particular, is far from being complete, but the cell assembly model of language has already quite strong experimental support [6]. In this model words are represented as strongly linked subnetworks of microcircuits that bind articulatory and acoustic representations of spoken words. The meaning of the word comes from extended network that binds related perceptions and actions. Various neuroimaging techniques confirm existence of such semantically extended networks. Psycholinguistic experiments show that acoustic speech input is quickly changed into categorical, phonological representation. A small set of phonemes is linked together in ordered string by a resonant state representing word form, and extended to include other brain circuits defining semantic concept. Phonological processing precedes semantic by about 90 ms [6].

To recognize a word in a conscious way activity of its subnetwork must win a competition for an access to the working memory [7]-[9]. Hearing a word activates strings of phonemes priming (increasing the activity) all candidate words and non-word combinations. Polysemic words proba-

Wlodzislaw Duch is with the Department of Informatics, Nicolaus Copernicus University, Grudziądzka 5, 87-100 Toruń, Poland and has been with the School of Computer Engineering, Nanyang Technological University, 639798 Singapore while working on this paper; contact: Google: Duch.

Paweł Matykiewicz and John Pestian are with the Department of Biomedical Informatics, Children's Hospital Research Foundation, Cincinnati, OH, USA; emails: John.Pestian@cchmc.org

bly have a single phonological representation that differs only by semantic extension. In people who can read and write visual representation of words in the recently discovered Visual Word Form Area (VWFA) in the left occipitotemporal sulcus is strictly unimodal [7]. Adjacent lateral inferotemporal multimodal area (LIMA) reacts to both auditory and visual stimulation and has cross-modal phonemic and lexical links. It is quite likely that the auditory word form area also exists [7][8], the homolog of the VWFA in the auditory stream located in the left anterior superior temporal sulcus; this area shows reduced activity in developmental dyslexics. Such word representations help to focus symbolic thinking. Context priming selects extended subnetwork corresponding to a unique word meaning, while competition and inhibition in the winner-takes-all processes leaves only the most active candidate network. Semantic and phonological similarities between words should lead to similar patterns of brain activations for these words.

A sudden insight (Aha!) experience accompanies solutions of some problems. Studies using functional MRI and EEG techniques contrasted insight with analytical problem solving that did not required insight [10]. About 300 ms before the Aha! moment a burst of gamma activity was observed in the Right Hemisphere anterior Superior Temporal Gyrus (RH-aSTG). One can conjecture that this area is involved in higher-level abstractions that can facilitate indirect associations [11]. The RH has only an imprecise view of the left hemisphere (LH) activity, generalizing over similar concepts and their relations. This activity represents abstract concepts, corresponding to categories higher in ontology, but also captures complex relations in concepts that have no name, but are useful in reasoning and understanding. For example, “left kidney” sounds correct, but “left nose” does not. The feeling arising from understanding may be connected to the left-right hemisphere activation interplay. Associations at higher level of abstraction in the RH are passed back to facilitate LH activations that form intermediate steps in language interpretation. High-activity gamma burst projected to the left hemisphere prime LH subnetworks with sufficient strength to form associative connections linking the problem statement with partial or final solution. This is a universal mechanism that operates in case of difficult problems as well as in understanding of complex sentences.

The qualitative picture is thus quite clear: chunking of terms and their associations corresponds to patterns of activations defining more general concepts in hierarchical way. The main challenge is how to use inspirations from neurocognitive linguistics to create practical algorithms for NLP.

III Discharge summaries

The data used in this paper comes from the Cincinnati Children’s Hospital Medical Center, a large pediatric academic medical center with over 750,000 pediatric patient encounters per year and terabytes of medical data in form of raw texts, stored in a complex, relational database [12]. Processing of medical texts requires resolving ambiguities and mapping terms to the Unified Medical Language System (UMLS) Metathesaurus concepts [13]. Prior knowledge generates expectations of a few concepts and inhibition of many others, a process that statistical methods of natural language processing [14] based on co-occurrence relations approximate only in a very crude way.

Discharge Summaries, contain brief medical history, current symptoms, diagnosis, treatment, medications, therapeutic response and outcome of hospitalization. Several labels (topics) may be assigned to such texts, such as medical subject headings, names of diseases that have been treated, or billing codes. Two documents with the same labels may contain very few common concepts. Our previous work [15] has been focused on defining useful feature spaces for categorization of such documents, selecting 26 semantic types (a subset of 135 semantic types defined in ULMS) that may contribute to document categorization. Similarity measures that take into account *a priori* knowledge of the topics were introduced in a model that tried to capture expert intuition using a few parameters. Preparation of the database and the pre-processing steps have already been described in [15] and due to the lack of space are not repeated here. Summary given in Table I allows to relate disease names to class numbers displayed in Fig. 1 and Fig. 2.

TABLE I.
INFORMATION ABOUT DISEASES USED IN THE STUDY

Disease name	No. of records	Average size (bytes)
1. Pneumonia	609	1451
2. Asthma	865	1282
3. Epilepsy	638	1598
4. Anemia	544	2849
5. Urinary tract infection (UTI)	298	1587
6. Juvenile Rheumatoid Arthritis	41	1816
7. Cystic fibrosis	283	1790
8. Cerebral palsy	177	1597
9. Otitis media	493	1420
10. Gastroenteritis	586	1375

Among 4534 discharge summary records “asthma” is the most common, covering 19.1% of all cases. Summary discharges are usually dictated and contain frequent misspelling and typing errors, punctuation errors, large number of abbreviations and acronyms. Categories assigned to these documents are not mutually exclusive and an expert reading such texts would not come close to the 100% classification accuracy, but for illustration purposes this division will be sufficient. The bag-of-words representation of such documents leads to very large feature spaces, many strongly correlated features (terms forming concepts), and extremely sparse representation. Unified Medical Language System (UMLS) [13] is a collection of many medical concept ontologies that are used to discover useful concepts and their relations, enabling semantic smoothing.

III ENHANCED CONCEPT REPRESENTATION

Three basic methods to improve representation of the texts in document categorization may be used: selection, expansion and the use of reference topics. Reference knowledge from background texts has recently been used to define topics (prototypes) [15]. Here the focus is on expansion of the feature space using prior knowledge. A standard practice in the document categorization is to use term frequencies tf_j for terms $j = 1 \ldots n$ in document D_i of length l_i=$|D_i|$ calculated for all documents that should be compared. Term frequencies are then transformed to obtain features in such a way that in the feature space simple metric relations between vectors representing these documents reflect their similarity. In document categorization we are interested in distribution of a given term among different categories. Words that appear in all documents may appear frequently, but carry little information that could be used for document categorization. In the *tf* x *idf* weighting scheme the uniqueness of each term is inversely proportional to the number of categories C_j, j=1..K this term appears in. The logarithm of ratio log(K/cf_j) is used as an additional factor in term weights:

$$s_j\left(D_i\right) = round\left(10\frac{1+\log tf_j\left(D_i\right)}{1+\log l\left(C_j\right)}\log\frac{K}{cf_j}\right) \qquad (1)$$

where $l(C_j)$ is the average length of documents from the class C_j. If the term i appears in all documents it does not contribute to their categorization and therefore $s_j(D_i)$=0 for all i. Additional normalization of all vectors in the document puts them on a sphere with a unit radius Z_i=s_i /$||s_i||$. This normalization tends to favor shorter documents. More sophisticated normalization method have been introduced [14] but all such normalization schemes treat each term separately and do no approximate specific distribution of brain's activations over related terms.

A set of terms t_i defines a feature space, with each term represented by a binary vector composed of zeros and a single 1 bit. In a unit hypercube this corresponds to vertices that lie on the coordinate axes. In this space documents D_j defined by term frequencies tf_i are also points defined by $tf_i(D_j)$ vectors with integer components. All term vectors are orthogonal to each other. Correlation between different terms is partially captured by latent semantic analysis, but features that are linear combination of terms have no clear semantics. Most terms from a large dictionary have zero components for all documents in typical document database, defining a null space.

Agglomerative hierarchical clustering methods with typical normalizations and similarity measures perform poorly in document clustering because the original representation is too sparse and the nearest neighbors of a document belong in many cases to different classes [16]. This is quite evident in the multi-dimensional scaling representation of our discharge summary collection shown in Fig.1. The simplest extension of term representation is to replace single terms by a group of synonyms, using for example Wordnet synsets (wordnet.princeton.edu). In this way the document text itself is expanded by new synonymous terms. This extends the non-null part of the feature space, simulating some of the spreading activation processes in the brain and increasing similarity of the documents that use different words to describe the same topic. In the binary approximation vector $\mathbf{X}(t_j)$ representing term t_j has zero elements except for $\mathbf{X}(t_k)$=1 for k=j and for those terms that are in the synset. The vector is multiplied by the term frequency tf_j in a given document D_i and may be normalized in a standard way.

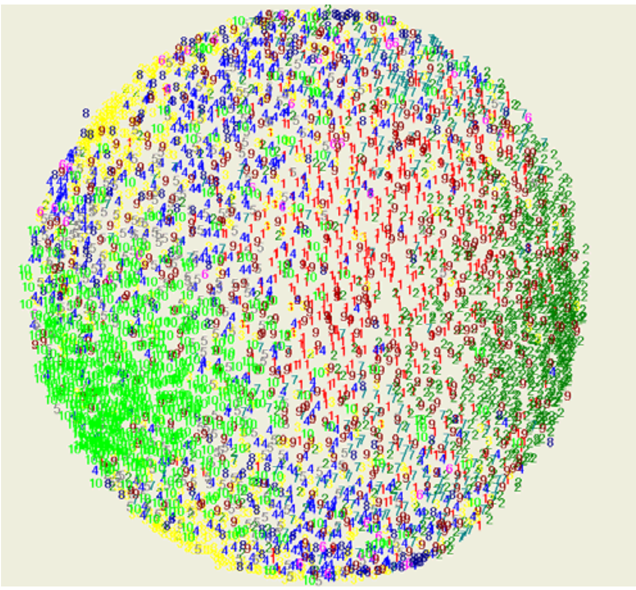

Fig. 1. MDS representation of 4534 summary discharges documents, showing little clusterization.

A better approximation to spreading activation in brain networks is afforded by soft evaluation of similarity of different terms. Distributional hypothesis assumes that terms similarity is the result of similar linguistic contexts [14]. However, in medical domain and other specialized areas it may be quite difficult to reliably estimate similarity on the basis of co-occurrence because there are just too many concepts and without systematic, structured knowledge statistical approaches will always be insufficient. Semantic smoothing for language modeling emerged recently as an important technique to improve probability estimations using document collections or ontologies [16]. Smoothing techniques assign non-zero probabilities to terms that do not appear explicitly in a given text. This is usually done by clustering terms using various (dis)similarity measures used also in filtering of information [17], or other measures of similarity of probability distributions [18]-[20], such as the Expected Mutual Information Measure (EMIM) [19].

In comparison of medical documents only specific concepts that belong to selected semantic types are used and thus there is no problem with shared common words. Documents from different classes may still have some words in common (e.g. basic medical procedures), but the frequency normalization will de-emphasize their importance. Discharge summaries from the same class may in some cases use completely different vocabulary. Wordnet synsets are not useful for very specific concepts that have no synonyms.

Semantic networks allow for concept disambiguation [5][21], but even using huge UMLS resources collection of relations is not sufficient to create semantic networks. These relations are based on co-occurrences and do not contain any systematic description of the ontological concepts. Ontology itself may be used, with parent-child, related and possibly synonymous, is similar to, has a narrower or broader relationship, has sibling relationship, being most useful relations for semantic smoothing. Several parents for each concept may be found in UMLS. Global approach to smoothing may simply use some of these relations to enhance the bag of words representation of texts, adding to each term a set of terms that come from different relations – this will be called a "term coset". These cosets for different terms may partially overlap as many terms may have the same parents or other relations. Such terms will be counted many times and thus will be more important. Straightforward use of relationships may be quite misleading. For example, many concepts related to body organs map to "Body as a whole - General disorders" concept that belongs to "Disease or Syndrome" type. Adding such trivial concepts will make all medical documents more similar to each other.

To avoid this type of problems one should either characterize more precisely the types of relations that should be used, or to score each new term individually looking at its usefulness for various tasks. The simplest scoring indices that will improve discrimination may be based on Fisher's criterion, mutual information, or other feature ranking indices [17]. From computational perspective it is much less costly to add all terms using specific relations and then use feature ranking to reduce the space.

Vector representation of terms have been created using the database of discharge summaries, but their use is not limited to the analysis of the database. The following algorithm to create them has been used:

1. Perform the text pre-processing steps: stemming, stop-list, spell-checking, either correcting or removing strings that are not recognized.
2. Use MetaMap [22] with a very restrictive settings to avoid highly ambiguous results when mapping text to UMLS ontology, try to expand some acronyms.
3. Use UMLS relations to create first-order cosets; add only those types of relations that lead to improvement of classification results.
4. Reduce dimensionality of the first-order coset space, but do not remove original (zero-order) features; any feature ranking method may be used here [17].
5. Repeat steps 3 and 4 iteratively to create second- and higher-order enhanced spaces.
6. Create $\mathbf{X}(t_i)$ vectors representing concepts.

Vectors $\mathbf{X}(t_j)$ representing terms t_j have zero elements except for $\mathbf{X}(t_k)=1$ for $k=j$ and for those terms that are in the cosets for a given term. They are highly dimensional and may be normalized to the unit length $\|\mathbf{X}(t_k)\|=1$ without loss of information; any metric may now be used to compare them. The non-zero coefficients of these vectors show connections between related terms. Iterative character of the algorithm leads to non-linear effects, feedback loops are strengthening some connections. Vectors representing terms are biased towards the data that has been used to create them and to the task used to define their usefulness, but with many labels and diverse text categories they may be useful in many applications. In medical document categorization a single specific occurrence of a concept may be an important indicator of the document category. The Latent Semantic Indexing (LSI) [14] will miss it, finding linear combinations of terms that do not have clear semantics.

IV EXPERIMENTS WITH MEDICAL RECORDS

The initial number of candidate words was 30260, including many proper names, spelling errors, alternative spellings, abbreviations, acronyms, etc. Out of 135 UMLS semantic types only 26 are selected (e.g. Antibiotics, Body Organs, Disease), ignoring more general types (e.g. Temporal or Qualitative Concepts) [15]. Each document has been processed by the MetaMap software [22] and concepts of the predefined semantic types have been filtered leaving 7220 features found in discharge summaries. After matching these features with *a priori* knowledge derived from medical textbook a relatively small subset of 807 unique medical concepts has been designed [15].

The performance of several classifiers has been evaluated on different versions of transformed data, including the most common and widely used text smoothing methods. Feature ranking based on Pearson's linear correlation coefficients (CC) have been performed to estimate feature/class correlations (other ranking methods, including Relief, did not give better results). In experiments with the kNN and SVM classifiers discriminating one class against all others it has been noticed that the CC threshold as small as 0.05 dramatically decreases accuracy, but for 0.02 the decrease is within 10-fold crossvalidation variance. Similar results are obtained with other feature spaces and classifiers.

In [15] 6 different normalizations of concept frequencies have been used, but results did not differ on more than a standard deviation (about 2%). The best 10-fold crossvalidation results for kNN do not exceed 52%, for SSV decision trees [23] (as implemented in the Ghostminer package [24] used for all calculations) about 43%, and for the linear SVM method 60.9% (Gaussian kernel gives very poor results). In all calculations reported here variance of the test results was below 2%. A new method based on similarity evaluation and the use of *a priori* knowledge applied to this data [15] gave quite substantial improvement, reaching 71.6%.

Detailed interpretation of results with topic-oriented *a priori* knowledge is not quite straightforward. Adding ontological relations and creating cosets for 807 terms selected as primary features allows for more much more detailed analysis. Using mostly the parent, broader and “related and possibly synonymous” relationships the first order space with 2532 has been created. In this space rather simple SSV decision tree model improved by about 6%, reaching 48.6±2.0%, with similar improvement from linear SVM that reached about 65%, with balanced accuracy (average accuracy in each class) reaching about 63%. Inspecting the most important features using Pearson’s correlation coefficient shows that this improvement may be largely attributed to mappings of various pharmacological substances to common higher-order concepts, for example Dapsone (of the Pharmacologic Substance type), becomes related to Antimycobacterials, Antimalarials, Antituberculous and Antileprotic, Sulfones and other drug families, making the diseases treated by these agents more similar. Two drugs with different names, Dapson and Vioxx, are related via the Sulfone, so although the name differ similarity is increased.

Taking all primary and first-order features with correlation coefficients CC>0.02 and repeating the expansion generates second-order space with 2237 features that provide even more interesting relations. Feedback loops in which term *A* has term *B* in its coset, and *B* has *A* in return, become possible. Bacterial Infections comes from many specific infections: Yersinia, Salmonella, Shigella, Actinomycosis, Streptococcal, Staphylococcal and other infections, increasing similarity of all diseases caused by bacteria. In this space accuracy of the linear SVM (C=10) is improved to 72.2±1.5%, with balanced accuracy 69.1±2.8%. Feature selection with CC>0.02 leaves 823 features, degrading the SVM results (C=32) only slightly to 71.5±1.8% and the balanced accuracy to 68.1±2.1%. Even better results have been obtained using the Feature Space Mapping (FSM) neurofuzzy network [25] that was used with Gaussian functions and a target learning accuracy of 80%: with 70-80 functions in each crossvalidation partitions it reached 73.8±2.4% with balanced accuracy of 69.6±2.3%. This shows that in each class separately expected accuracy is about 70%. Other methods of feature ranking, such as Relief or methods based on entropy [17] did not improve these results.

The improvement in classification accuracy with the second-order space is clearly reflected in better clusterization of the multidimensional scaling (MDS) representation [26] of similarity among documents shown in Fig. 2. For Pneumonia (Class 1) one cluster is observed in the upper part of this figure, and a rather diffused cluster in the lower part. Upon closer examination of source documents and the new coset terms it becomes clear that the second cluster contains documents with cases that are hard to qualify uniquely as pneumonia, as the patient have also several other problems and the diagnosis is uncertain. There are quite a few problems with the type of expansion before dimensionality reduction. Some drugs, for example Acetaminophen, are related to 15 specific concepts in the UMLS ontology, all of the type: Acetaminophen 80 mg Chewable Tablet. Perhaps for medical doctor writing prescriptions this is a unit of information, but obviously such detailed concept appear rarely and thus are removed by feature selection, while general concepts, such as Acetaminophen (Pharmacologic Substence) are left. In effect most important features tend to come from the second-order cosets.

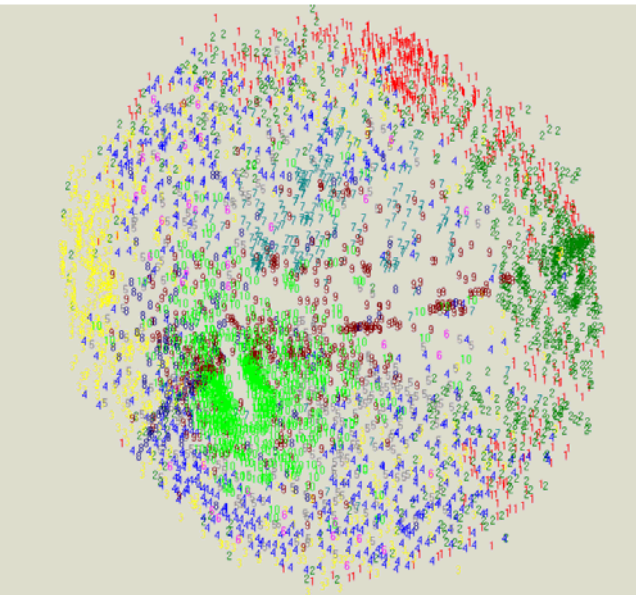

Fig. 2. Clusterization of document database after two steps of spreading activation using UMLS ontology.

V. Conclusions and wider implications

The use of background knowledge in computational intelligence is an important topic that may be approached from different perspectives. Without such knowledge analysis of texts, especially texts in technical or biomedical domains, is almost impossible. Neurolinguistic inspirations may be quite fruitful, leading to approximations of processes that are responsible for text understanding in human brains. Large-scale semantic networks and spreading activation models may be constructed starting from large ontologies. Vector representations of concepts may be regarded as a snapshot of dynamic activity patterns, defining connections with other concepts.

Creating useful numerical representation of various concepts is an interesting challenge. For medical applications such vectors may be created by expansion of each term to form a cosets using relationships provided by UMLS ontologies. To end up with a useful representation the utility of each new relation has to be checked, or a whole class of concepts based on some specific relations may be added to the coset and then pruned to remove concepts that are not useful in text interpretation. Association rules may be quite helpful here [27]. A crude version of such approach has been presented here and already using the second-order expansions gave excellent result on a very difficult problem of summary discharge categorization. This is one of the approaches to enhance UMLS ontologies by vectors that represent these concepts in numerical way and could be used in variety of tasks.

Finding optimal enhanced feature spaces and simplest decompositions of medical records into classes using either sets of logical rules or minimum number of prototypes in the enhanced space is an interesting challenge. “Optimal” may here depend on a wider context as the meaning of a concept depends on the depth of knowledge an expert has. For example, family physician may understand some concepts in a different way than a cardiologist, but it should be possible to capture both perspectives in concept description vectors. Visualization not only helps to see subtypes of the same category but also to identify various problems. Looking at documents that are far from the main cluster they should belong to, one may ask what type of associations allows humans to make correct assignment. This helps to create new relations or remove some misleading associations from the system.

A lot of knowledge that medical doctors gain through the years of practice is frequently never verbalized. Prototypes representing clusters of documents describing medical cases may be treated as a crude approximation to the activity of neural cell assemblies in the brain of a medical expert who thinks about a particular disease. This may perhaps be observed in clusterization of these documents if a proper space is defined. Clusters in Fig. 2 may be interpreted in this way, although MDS mapping to two dimensions only has to introduce many distortions. Finding such subclusters will require good database of clinical texts describing typical examples, and such databased does not exist. It is relatively easy to collect information about rare cases that are subject to scientific investigation, but not the common ones. Such analysis could help in training of young medical doctors by presenting optimal sets of cases for each specific cluster. It could also be potentially useful in more precise diagnoses. With sufficient amount of documents optimization of individual feature weights could also be attempted.

Although much remains to be done before unstructured medical documents and general web documents will be fully and reliably annotated in an automatic way, a priori knowledge certainly will be very important. Creating better approximations to the representation and the use of knowledge in this process is a great challenge for CI.

Acknowledgment

W. Duch thanks the Polish Committee for Scientific Research, research grant 2005-2007, for support.

References

[1] S. Lamb, Pathways of the Brain: The Neurocognitive Basis of Language. Amsterdam & Philadelphia: J. Benjamins Publishing Co. 1999.

[2] D.E. Rumelhart and J.L. McClelland (eds), Parallel Distributed Processing: Explorations in the Microstructure of Cognition Vol. 1: Foundations, Vol. 2: Psychological and Biological Models. MIT Press, Cambridge, MA, 1986.

[3] F. Crestani, Application of Spreading Activation Techniques in Information Retrieval. Artificial Intelligence Review 11:453-482, 1997.

[4] F. Crestani, P.L. Lee, Searching the web by constrained spreading activation. Information Processing & Management 36:585-605, 2000.

[5] G. Tsatsaronis, M. Vazirgiannis, I. Androutsopoulos, Word Sense Disambiguation with Spreading Activation Networks Generated from Thesauri, 20th Int. Joint Conf. in Artificial Intelligence (IJCAI 2007), Hyderabad, India, pp. 1725-1730.

[6] F. Pulvermuller (2003) The Neuroscience of Language. On Brain Circuits of Words and Serial Order. Cambridge University Press.

[7] S. Dehaene, L. Cohen, M. Sigman and F. Vinckier, The neural code for written words: a proposal. Trends in Cognitive Science, 9: 335-341, 2005

[8] S. Dehaene and L. Naccache, Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. Cognition, 79:1-37, 2001

[9] W. Duch, Brain-inspired conscious computing architecture. Journal of Mind and Behavior 26(1-2), 1-22, 2005.

[10] E.M. Bowden, M. Jung-Beeman, J. Fleck, J. Kounios, New approaches to demystifying insight. Trends in Cognitive Science 9, 322-328, 2005.

[11] W. Duch, Creativity and the Brain. In: A Handbook of Creativity for Teachers. Ed. Ai-Girl Tan, World Scientific Publishing (in print).

[12] J. Pestian, B. Aronow, K. Davis, Design and Data Collection in the Discovery System. Int. Conf. on Mathematics and Engineering Techniques in Medicine and Biological Science, 2002.

[13] UMLS Knowledge Sources, 13th Edition – January Release. Available: http://www.nlm.nih.gov/research/umls

[14] C.D. Manning and H. Schütze, Foundations of Statistical Natural Language Processing MIT Press, Cambridge, MA 1999.

[15] L. Itert, W. Duch, J. Pestian, Influence of *a priori* Knowledge on Medical Document Categorization, IEEE Symposium on Computational Intelligence in Data Mining, IEEE Press, April 2007

[16] X. Zhou, X. Zhang and X. Hu, Semantic Smoothing of Document Models for Agglomerative Clustering, 20th Int. Joint Conf. on Artificial Intelligence (IJCAI 2007), India 2007 (in print).

[17] W. Duch, Filter Methods. In: Feature extraction, foundations and applications. Eds: I. Guyon, S. Gunn, M. Nikravesh, L. Zadeh, Studies in Fuzziness and Soft Computing, Physica-Verlag, Springer 2006, pp. 89-118

[18] P. Cimiano, A. Hotho and S. Staab, Learning Concept Hierarchies from Text Corpora using Formal Concept Analysis, Vol. 24, 305-339, 2005.

[19] W.W. Bein, J.S. Coombs, K. Taghva, A Method for Calculating Term Similarity on Large Document Collections, Int. Con. on Information Technology: Computers and Communications, 2003, p. 199-207.

[20] Y. Li, A.B. Zuhair and D. McLean, An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources. IEEE Transactions on Knowledge and Data Engineering, 15(4):871-882, 2003.

[21] P. Matykiewicz, W. Duch, J. Pestian, Nonambiguous Concept Mapping in Medical Domain, Lecture Notes in Artificial Intelligence 4029, 941-950, 2006.

[22] MetaMap, available at http://mmtx.nlm.nih.gov

[23] K. Grąbczewski and W. Duch, The separability of split value criterion, 5th Conf. on Neural Networks and Soft Computing, Zakopane, Poland, 2000, pp. 201-208.

[24] Ghostminer data mining software, www.fqspl.com.pl/ghostminer/

[25] W. Duch and G.H.F. Diercksen, Feature Space Mapping as a universal adaptive system. Computer Physics Communications 87: 341-371, 1995.

[26] E. Pękalska and R.P.W. Duin, The dissimilarity representation for pattern recognition: foundations and applications, New Jersey; London: World Scientific, 2005.

[27] M.-L. Antonie, O.R. Zaiane, Text document categorization by term association. Proc. of IEEE Int. Conf on Data Mining (ICDM), 2002, pp. 19- 26.