Nonambiguous Concept Mapping in Medical Domain

Paweł Matykiewicz[1,2], Włodzisław Duch[1,3], John Pestian[2]

[1] Department of Informatics, Nicolaus Copernicus University, Toruń, Poland
[2] Dept. of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, OH, USA
[3] School of Computer Engineering, Nanyang Technological University, Singapore
pawelm@phys.uni.torun.pl, Google: Duch, john.pestian@cchmc.org

Abstract. Automatic annotation of medical texts for various natural language processing tasks is a very important goal that is still far from being accomplished. Semantic annotation of a free text is one of the necessary steps in this process. Disambiguation is frequently attempted using either rule-based or statistical approaches to semantical analysis. A neurocognitive approach for a nonambiguous concept mapping is proposed here. Concepts are taken from the Unified Medical Language System (UMLS) collection of ontologies. An active part of the whole semantic memory based on these concepts forms a graph of consistent concepts (GCC). The text is analyzed by spreading activation in the network that consist of GCC and related concepts in the semantic network. A scoring function is used for choosing the meaning of the concepts that fit in the best way to the current interpretation of the text. ULMS knowledge sources are not sufficient to fully characterize concepts and their relations. Annotated texts are used to learn new relations useful for disambiguation of word meanings.

1 Introduction

The Unified Medical Language System (UMLS) is a collection of 88 medical knowledge sources. The most recent edition of UMLS (2005AB ed.) contains 1 196 265 unique concepts, each labeled by a Concept Unique Identifier (CUI), and 2 873 310 unique phrases (SUI) [1]. Annotation of texts requires mapping of noun phrases, words, abbreviations and acronyms discovered in the unstructured text to the unique UMLS concepts. The *MetaMap* software (MMTx) [2] is frequently used to discover UMLS concepts in texts. The software has been developed by experts in the *U.S. National Library of Medicine*, a part of the *National Institute of Health*. The *MetaMap* algorithm is rather slow and quite complicated. It is aimed at discovering all possible terms in the text without carrying much about ambiguity of the output. As a result some words are given many annotations, listing all possible meanings and various phrases they appear in, making the semantic search even more difficult than with the raw text.

The goal of our research is to overcome these drawbacks and create fast, precise and unambiguous concept mappings. There are many statistical, pattern recognition and syntactical approaches to the general word sense disambiguation (WSD) problem [3], but most experiments have been conducted on a small scale, while the number of medical concepts that need to be taken into account exceeds one million. Moreover, although some word meanings are easily distinguishable other are quite difficult to capture and

even human annotators agree only in no more than 80% [4]. Despite that fact experts have no problem with understanding medical or technical texts. The only system capable of language understanding at the human competence level is the human brain, and it should be the source of inspirations for development of semantic annotation systems.

General philosophy of our neurocognitive approach to natural language processing (NLP) is presented in the next section. To approximate formation of primed semantic subnetwork providing interpretation of the text graphs of consistent concepts (GCCs) are constructed. The concept mapping algorithm, presented in the third section, is based on this approach, although many other variants and applications are possible [5–7]. The algorithm for phrase sense disambiguation that adds new relations and determines relations strength between concepts with the use of prior knowledge and the acquisition of new knowledge are also discussed in this section. The last section contains conclusions and future plans.

2 Neurocognitive approach to NLP

Analysis of texts, independent of the purpose, requires three main steps:

- recognition of tokens, or mapping from strings of letters to unique terms;
- resolving ambiguities, grouping terms into phrases and mapping them to concepts;
- semantic representation of the whole text capturing relations among entities that are involved, facilitating inferences, and thus understanding and answering questions about its content.

These three steps roughly correspond to the function of three kinds of human memory [8]: recognition memory, semantic memory and episodic memory. NLP research usually ignores this fact, focusing on formal approaches (grammar, logics, statistical correlations). Neurocognitive approach to NLP follows inspirations from brain science focusing on approximated models of memory and other neural processes. The long-term goal is to reach human-level competence in natural language processing.

Recognition memory helps to ignore most spelling errors. As long as the first and the last letter of the word is not changed even severely distorted texts containing *wrods wtih many paris of letres trasnpoesd* is read without much troubles, a phenomenon that is of interest to spammers and cognitive scientists. It is rather obvious that context and anticipation plays a major role in correct recognition. Although we do not consider problems at the recognition level here unstructured medical texts need a lot of data cleaning. Lexical Systems Group of the US National Library of Medicine has developed a spelling suggestion tool Gspell and the SPECIALIST lexicon containing many biomedical terms and general English words, that Gspell is using to provide suggestions. Without the use of context and understanding the topic of the text Gspell makes many spelling suggestions, although humans recognize a single term, frequently paying no attention to the misspellings. It is clear that recognition memory cannot be separated from other memory systems, doing much more than just searching for similar terms in the lexicon. Reading text leads to priming effects: expectation and anticipation of a few selected words, and inhibition of many others that do not come to our mind.

The *semantic priming* (SP) phenomenon has been known in cognitive psychology since more than 30 years (see the review in [9]). Each word excites brain subnetworks that encode different meanings of that word [10]. In such coding identical phonological representations of words may be shared among several concepts without leading to any problems. Words that have been processed earlier (context) have already activated many brain subnetworks, increasing the probability of a particular meaning of the new concept, and inhibiting all other meanings. This competition, leading to inhibition of subnetworks coding alternative meanings of the word, makes it hard to think about alternatives when one of the meanings (interpretations) fits really well to the current context. Statistical language processing models applied to a large text corpus used for training allow for prediction of the next word in a sequence with high reliability [11], partially capturing this anticipation, although statistical algorithms do not approximate well real brain processes behind this phenomenon. Anticipation may help to disambiguate word senses, facilitating the mapping of terms into concepts.

Semantic memory encodes in the activity of brain's subnetworks information about objects and concepts, together with their properties and relations. Formal models of semantic networks, computational structures inspired by psychological ideas about semantic memory, are known since more than 30 years [8, 12–14]. Semantic networks are used in artificial intelligence as knowledge representation tools [15, 16] and may provide a model to approximate functions of biological semantic memory (SM). Each node in semantic network is in fact a subnetwork, with similarities and associations between concepts resulting in sharing some parts of the subnetworks. Activations of semantic subnetworks are responsible for semantic priming, building an episode that may be memorized and retrieved later, reinstating a particular configuration of brain activities, or an episode.

Episodic memory is based on semantic relations of concepts found in the analyzed text, understanding or rough categorization of the text topic, and binding different entities in a specific way for this particular text. Our assumption here is that a simple model of episodic memory may be provided by priming the semantic network during text reading, forming an active subnetwork of main concepts found in the text, their mutual relations and their relations to concepts forming background knowledge that has not been explicitly mentioned. One way to achieve it is to scan the text to find main, unambiguous concepts that form the skeleton of the active subnetwork, and add other concepts selecting the meaning that increase overall consistency. In this way graphs of consistent concepts (GCCs) may be formed, capturing the meaning of the text and facilitating its unambiguous annotation. Some relations are not defined directly at the level of relations between individual concepts, but at the higher ontological level (this is probably done by the non-dominant hemisphere, where representations of abstract concepts may be stored [17]. UMLS contains *Semantic Network*, but it has only 132 highest level broad subject categories (Semantic Types).

An alternative to the network model is provided by vector space models, for example the High Dimensional Semantic Space memory model [18]. An ambiguous word, for example cold, corresponding to several different concepts (UMLS Methathesaurus has 6 senses of cold) is represented by 6 different vectors. Each vector component measures statistical co-occurrences of each particular word sense to all other lexicon terms,

defining how likely it is that this term will appear in the context window of a given word. To select the correct sense the context window is used to find how similar is a given vector representation to the current context. This approach has been used with some success in general word sense disambiguation tasks. Vector models may be understood as an approximation to the activation of network nodes, and the search for consistent concepts that is used in the GCC algorithm may be formulated as the search for vectors that create smallest volume.

Memory-based process are rarely acknowledged in natural language processing research. Semantic Knowledge Representation (SKR) project at the National Library of Medicine (NLM) has a very ambitious goal [19], although it is only loosely inspired by psychological ideas about semantic memory, rather then being a model of semantic memory as implemented by the brain. Yet it is obvious that without recognition, semantic and episodic memory understanding texts would not be possible. Each concept has numerous properties and relations that are encoded in the structure of subnetworks that encode it in the brain [10]. Our goal should be to approximate some of these processes. This leads to the extension of the idea of semantic networks [12, 13, 15, 16], providing a model to approximate functions of biological semantic memory (SM). Concepts should be represented by distributed subnetworks that contain phonological representations and by semantic extensions of these representations, linking to all the properties of a given concept and to all concepts that may be associated in some way with them. Two main processes in such networks are spreading activation and competition. Competitive processes should not be considered as the "winner-takes-all" only, as there are many winners and the activity of the whole subnetwork providing consistent interpretation of the text being analyzed is growing.

The challenge is to collect sufficient knowledge about concepts and their relations that allows humans to understand language and to interpret texts. The largest collection of ontologies combined by specialists is available in medical domain. UMLS ontologies have hierarchical structures and thus do not provide strong concept descriptions. An expert knows much more about basic medical concepts than can be found in the UMLS. Thus UMLS may serve only as a poor approximation to the real semantic memory. Relations contained in the UMLS may be represented as excitatory connections between concepts. UMLS (2005AB edition) contains 4 435 387 unique relations but it is not a comprehensive medical encyclopedia, so a lot of relations are missing. It is not clear how to score most relations because only co-occurrence relations in UMLS are numerical and all others are logical. Lack of knowledge about concept properties and relations is the major obstacle to annotate in an unambiguous way clinical texts.

To approximate formation of primed semantic subnetwork providing interpretation of the text, graphs of consistent concepts (GCCs) are constructed. An algorithm for adding new relations and determining relations strength between concepts with the use of prior knowledge is described below

3 Concept Mapping Algorithm

UMLS includes three main modules used in our approach: GSPELL, a tool for spelling correction; and two UMLS Knowledge Sources: *Specialist Lexicon*, a general lexi-

con that includes both common English words and biomedical vocabulary; *Methathesaurus*, describing biomedical and health-related concepts, a very large, multi-lingual and multi-purpose vocabulary. Overall the whole UMLS installation needs 26 GB of storage space. The GCC algorithm uses only part of UMLS. For normalizing and varying terms following files from the SPECIALIST LEXICON are used: `DM.DB, SM.DB, LRABR, LRAGR, LRNOM, LRSPL`. For binding concepts with phrases and sources the following file are used: `MRCONSO.RRF, MRXNS_ENG.RRF`.

In order to map a noun and verb phrases to concepts *TreeTagger* software [20] is used to annotate text with parts of speech (POS). Every word (EUI) in a text is mapped to its normalized form (WUI) - a singular noun. Unique string identifiers (SUI) are composed in turn from normalized words (WUI). Every phrase (SUI) has on average 2.4 different CUIs associated with it. The following schema for mapping is used:

$$EUI \mapsto WUI \mapsto SUI \mapsto CUI$$

Figure 1 presents a simplified schema for mapping a text to concepts. Words in that example are already normalized.

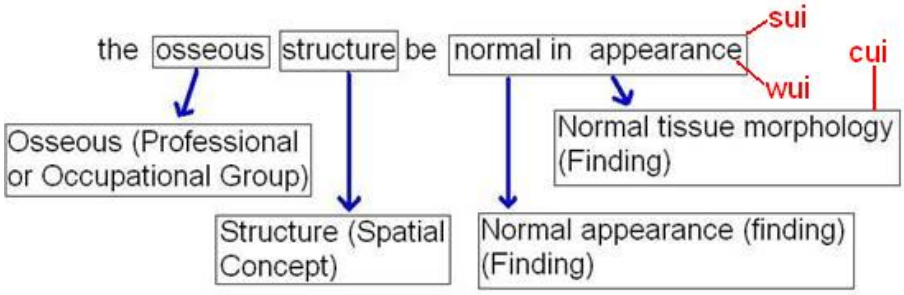

Fig. 1. A simplified schema for mapping normalized words to concepts.

In order to map phrases (SUI) to concepts (CUI) following algorithm was used:

```
1. Assign part of speech tags to every token.
2. Map all the words to their normalized forms.
3. Scan normalized words from the end of the text.
  3a. If a POS tag matches one of the symbols:
    CD,RB,JJ,N,VV,LS,SYM
    start scanning the text from the current position
    towards beginning of the text,
    add words to a phrase that match mentioned POS tags
    until there is a phrase that is not in the UMLS.
  3b. Resume after position where last UMLS
  phrase was found.
4. Finish when at the beginning of the text.
```

This is a very fast and simple mapping algorithm. The following text from ultrasonography dictation has been mapped to a concept space:

```
Fever, left flank pain, pyelonephritis. The right kidney is
normal in sonographic appearance with no evidence of scarring,
hydronephrosis or calculi. It measures XXXX cm which is normal
for patient's age. The left kidney is enlarged. It measures xx
cm in length. No focal areas of abnormal echogenicity or scarring
are seen. No hydronephrosis or calculi are identified. Images of
the bladder demonstrate no abnormality. Enlargement of the left
kidney which may suggest acute pyelonephritis. This could also
represent a normal variant. Normal appearing right kidney.
```

The algorithm found 30 concepts in this text, 11 of which are ambiguous. Next step is to find which $SUI \mapsto CUI$, or phrases map to concepts. In order to disambiguate SUI phrases relational table from UMLS (`MRREL.RRF` file) is used. This file contains 5 499 792 unique relations, with the same relations found in many different ontologies. Some relations are rather peculiar and appear only in one specialized ontology, while important relations are found in many ontologies. A weight matrix for all relations is constructed. Following definitions are used: $N(CUI_i)$ – number of occurrence of a CUI_i concept in the relational table, $C(CUI_i, CUI_j)$ – number of co-occurrences of CUI_i and CUI_j concepts in the relational table row, $W = \{w_{ij}\}$ – matrix storing weights between ith and jth concept. The weights are defined as conditional probabilities:

$$w_{ij} = P(j|i) = \frac{C(CUI_i, CUI_j)}{N(CUI_i)} \tag{1}$$

Once a text is mapped to a set of ambiguous concepts a graph of consistent concepts is created, with nodes corresponding to concepts and edges to relations. Each node corresponding to the concept found in the text has an initial activity $a_i(t = 0)$ that spreads to other nodes according to W matrix:

$$a_i(t+1) = \alpha a_i(t) + \sum_j w_{ij} H(a_j(t)) \tag{2}$$

where H is the Heaviside step function and α is a spontaneous decay parameter. Similar function was considered in [21]. The main problem for spreading activation in networks without inhibition is to prevent the infinite growth of all node activities. α decays should be sufficiently large to achieve this; all experiments in this paper are with $\alpha = 0.73$.

Propagation of activations in the semantic network should lead to the decay of concepts that are not supported by other active concepts. After a few iterations only most consistent concepts forming the GCC graph should have activations above certain threshold. These concepts should give the right sense of a phrase (SUI). Figure 2 shows an example of the GCC graph after 4 iterations.

The initial weights created from UMLS relations help to disambiguate phrases only in a limited way. The UMLS is a big but quite general knowledge base, and it

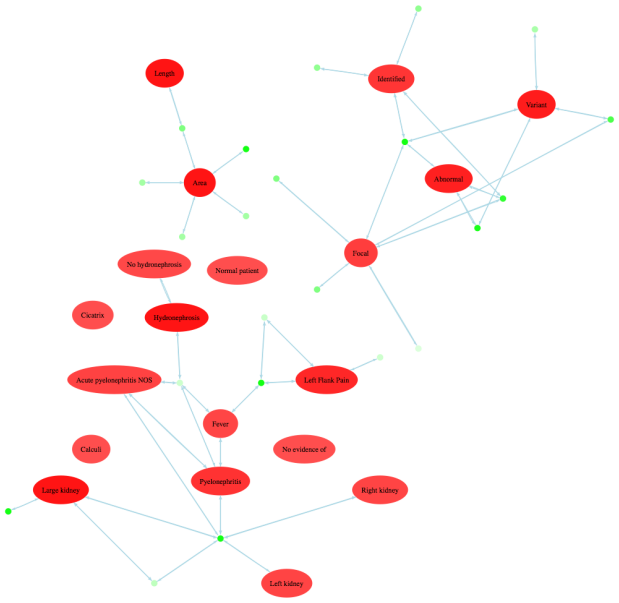

Fig. 2. Example of a graph of consistent concepts after 4 iteration of spreading activations.

frequently lacks more specific knowledge. Enriching UMLS relations means simply adding $N(CUI_i)$ and $C(CUI_i, CUI_j)$ for all pairs of concepts from an annotated text. For pilot purposes two small radiology corpuses were created. For knowledge acquisition Cincinnati Children's Hospital Medical Center clinical texts from radiology were used. These texts are dictated by physicians and changed into a text form by medical voice recognition software. 60 of those documents where chosen for manual annotation and divided in two parts. Every set of documents has 30 chest x-ray dictations for 6 different diseases. Special web application was created with an easy to use interface that allows a specialist to annotate a text. Figure 3 shows the main interface done with Asynchronous JavaScript And XML (AJAX) technology [22].

In order to check the usefulness of this approach pilot project accuracy measure that focuses only on the ambiguous mappings was used. If the maximally activated CUI corresponds to the manually chosen CUI then a correct recognition is counted. Overall *Corpus I* has 140 ambiguous phrases and *Corpus II* 301 ambiguous phrases. Table 3 shows comparison of accuracies with no training, training using Corpus I and training using Corpus II. The second corpus seems to be much more difficult to learn but overall results are promising.

The initial weights were able to give maximum activation to only 64% of correct concepts but manually adding the relations found in the radiology corpus to this toy example gave perfect disambiguation in all cases. Figure 4 shows the GCC graph after adding new co-occurrence relations. This figure presents more compact and consistent graph. Only the right senses of the phrases (SUIs) have maximum activation potential.

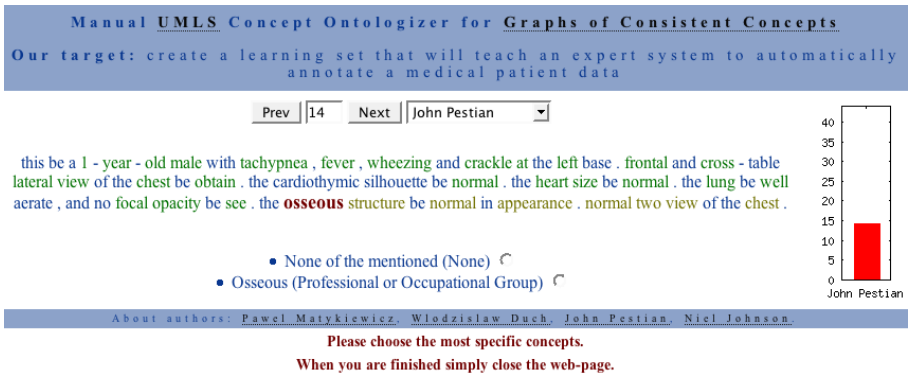

Fig. 3. Influence of learning on CUI selection.

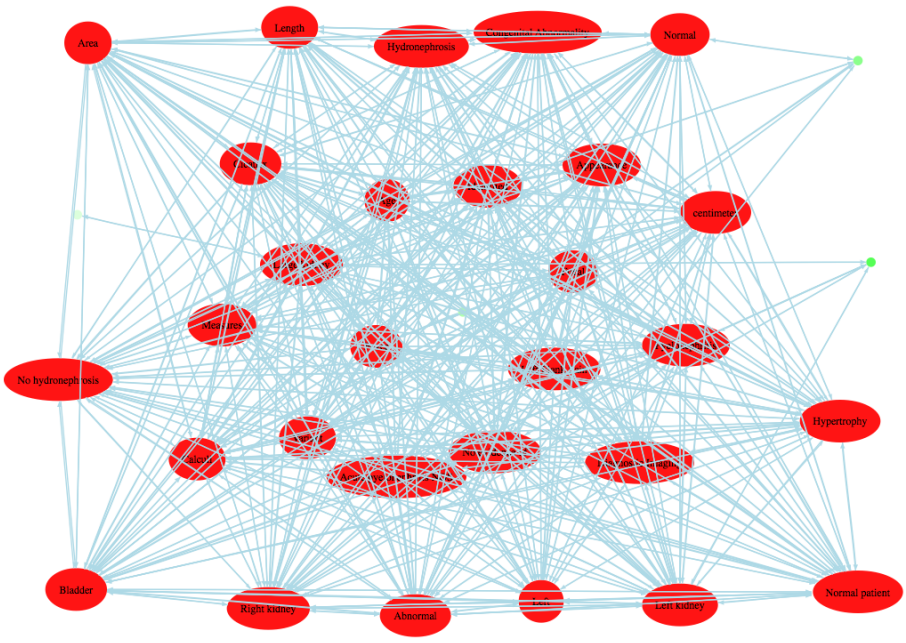

Fig. 4. Example of a graph of consistent concepts with enriched UMLS relations.

	no training	training	
		Corpus I	Corpus II
Corpus I	79%	96%	86%
Corpus II	57%	64%	79%

Table 1. Comparison of GCC disambiguation accuracies with and without additional training.

4 Conclusions

General neurocognitive approach to the natural language processing has been described and an algorithm for nonambiguous medical concept mapping, based on this approach, has been presented. Unfortunately even with the use of ULMS parsed clinical texts showed that many relations needed for correct annotations are still missing. A software tool for enriching UMLS relation by creating manually annotated texts and learning from them has been presented. Experiments performed on two small corpuses showed significant influence of additional knowledge on the disambiguation performance of GCC graphs.

Annotation of unstructured medical texts is quite difficult. Sometimes the UMLS mappings do not make sense for medical experts. In such cases they have additional CUI to choose: *None of the mentioned* (Fig. 3). This means that none of the CUI that are mapped to a SUI should be included in the graph (or they should at least have a very small activation). This and many other issues remain still to be investigated.

GCCs are a promising tool for Natural Language Processing tasks. They provide a better approximation to brain processes then vector models, yet computationally they are relatively simple, using only a single vector with node activations and a weight matrix estimating strength of relations. There may be many variants of GCC-based algorithms, with different strategies for initial node activations and subsequent activation spreading, weighting of relations, and overall consistency scoring evaluations. To add new knowledge large manually annotated corpus should be created, and better concept descriptions created using medical textbooks and dictionaries (the problem of automatic creation of semantic memory has been considered in [6]). Combining ideas from cognitive neuroscience with ideas from medical information retrieval literature algorithms that reach human level performance should finally be achieved.

Acknowledgement: WD is grateful for the support by the Polish Committee for Scientific Research, research grant 2005-2007.

References

1. UMLS Knowledge Server web site: http://umlsks.nlm.nih.gov
2. MetaMap web site: http://mmtx.nlm.nih.gov
3. M. Stevenson, “Word Sense Disambiguation: The Case for Combinations of Knowledge Sources”, The University of Chicago Press, Chicago IL, 2003
4. P. Edmonds, and A. Kilgarriff, “Introduction to the special issue on evaluating word sense disambiguation systems”, Journal of Natural Language Engineering, 8(4), 279-291, 2002.

5. J.P. Pestian, L. Itert, C. Andersen, W. Duch, “Preparing Clinical Text for Use in Biomedical Research.” Journal of Database Management 17(2), 1-11, 2006.
6. J. Szymanski, T. Sarnatowicz, W. Duch, “Towards Avatars with Artificial Minds: Role of Semantic Memory”. Journal of Ubiquitous Computing and Intelligence (in print).
7. W. Duch, J. Szymanski, T. Sarnatowicz, “Concept description vectors and the 20 question game”. In: Intelligent Information Processing and Web Mining, Eds. M.A. Klopotek, S.T. Wierzchon, K. Trojanowski, Advances in Soft Computing, Springer Verlag, pp. 41-50, 2005.
8. J.R. Anderson, “Learning and Memory”. J. Wiley and Sons, NY 1995.
9. T.P Mcnamara, “Semantic Priming; Perspectives from Memory and Word Recognition (Essays in Cognitive Psychology)”, Psychology Press, UK, 2005
10. F. Pulvermller, “The Neuroscience of Language. On Brain Circuits of Words and Serial Order”. Cambridge Uni. Press, 2003.
11. R. Hecht-Nielsen, “Cogent confabulation”. Neural Networks 18(2): 111-115, 2005.
12. A.M. Collins and M.R. Quillian, “Retrieval time from semantic memory”. Journal of Verbal Learning and Verbal Behavior 8, 2407, 1969.
13. A.M. Collins, E.F. Loftus, “A spreading-activation theory of semantic processing”. Psychological Reviews 82, 40728, 1975.
14. J.L. McClelland and T.T. Rogers, “The Parallel Distributed Processing Approach to Semantic Cognition”. Nature Reviews Neuroscience 4, 310-322, 2003.
15. J.F. Sowa, ed. “Principles of Semantic Networks: Explorations in the Representation of Knowledge”. Morgan Kaufmann Publishers, San Mateo, CA, 1991.
16. F. Lehmann, ed. “Semantic Networks in Artificial Intelligence”. Pergamon Press, Oxford, 1992.
17. W. Duch, “Computational Creativity”. World Congres on Computational Intelligence, Vancouver, 16-21 July 2006
18. C. Burgess, “Representing and resolving semantic ambiguity: A contribution from high-dimensional memory modeling”. In: Gorfein, D.S. (Ed.), On the Consequences of Meaning Selection: Perspectives on Resolving Lexical Ambiguity. APA Press, 2001.
19. T.C. Rindflesch and A.R. Aronson. “Semantic processing for enhanced access to biomedical knowledge”. In: V. Kashyap, L. Shklar (eds), Real World Semantic Web Applications, IOS Press, pp. 157-172, 2002.
20. Tree-Tagger web site: http://www.ims.uni-stuttgart.de/ /projekte/corplex/TreeTagger/DecisionTreeTagger.html
21. C. Rocha, D. Schwabe, M. P. Aragao, A hybrid approach for searching in the semantic web Source, International World Wide Web Conference archive Proceedings of the 13th international conference on World Wide Web, New York, NY, USA
22. AJAX Sun web site http://developers.sun.com/channel/01_06/index.jsp?cid=59754