Neural Networks 21(10), 1500-1510, 2008

Neurolinguistic Approach to Natural Language Processing with Applications to Medical Text Analysis

Włodzisław Duch[a,b,*1], Paweł Matykiewicz[b,c] and John Pestian[c]

[a]*Department of Informatics, Nicolaus Copernicus University, Grudziądzka 5, 87-100 Toruń, Poland*
[b]*School of Computer Engineering, Nanyang Technological University, 639798 Singapore*
[c]*Department of Biomedical Informatics, Children's Hospital Research Foundation, Cincinnati, OH, USA*

Abstract

Brain processes responsible for understanding language are approximated by 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. Snapshots of activations of various concepts in the brain spreading through associative network may be captured in a vector model. Medical ontologies are used to identify concepts of specific semantic type in the text, and add to each of them related concepts, providing expanded vector representations. To avoid rapid growth of the extended feature space after each step only the most useful features that increase document clusterization are retained. 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. Although better approximations to the spreading of neural activations may be devised a practical approach presented in this paper helps to discover pathways used by the brain to process specific concepts.

Keywords: Natural language processing; Semantic networks; Spreading activation networks; Medical ontologies; vector models in NLP

1. 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 proposed in [1] has been quite 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 represented as a state of an associative network, and the spreading of activation creates network states that facilitate semantic interpretation of the text. Unfortunately unraveling the "pathways of the brain" [1] proved to be quite difficult.

Connectionist approach to natural language has been introduced in the influential PDP books [2] but has not been developed beyond models providing qualitative explanations of linguistic phenomena. Much later some applications of constrained spreading activation techniques in information retrieval [3], semantic search techniques [4] and word sense disambiguation [5] have been made. In this paper neurolinguistic inspirations are used to search for useful approximations to spreading of brain activity during text comprehension, connecting this approach with standard, vector-based NLP techniques. A database of short medical documents (hospital discharge summaries) divided into ten categories is used for illustration of

[1] Corresponding author: for contact information Google: W. Duch.

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 section introducing an algorithm that creates semantic description vectors, and by the section that shows applications of this algorithm to clusterization and discovery of topics in medical texts. Discussion of presented results and their wider implications closes this paper.

2. 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 understood, but the cell assembly model of language has already quite strong experimental support [6][7], and agrees with broader mechanisms responsible for memory [8]. In the cell assembly (or neural clique) model words (or general memory patterns) are represented by 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, activating sensory, motor and premotor cortices [6]. 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. Hearing a word phonological processing creates localized attractor state whose activity spreads in about 90 ms to extended areas defining its semantics [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]-[10]. Hearing a word activates strings of phonemes priming (decreasing the threshold for activity) all candidate words and some non-word phoneme combinations. Polysemic words probably 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 [11]. It is quite likely that the auditory word form area also exists [7][9]. It may be a 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 [12]. About 300 ms before the Aha! moment a burst of gamma activity was observed in the Right Hemisphere anterior Superior Temporal Gyrus (RH-aSTG). This has been interpreted as "making connections across distantly related information during comprehension (…) that allow them to see connections that previously eluded them" [13]. Bowden *et al.* [12] performed a series of experiments that confirmed the EEG results using fMRI techniques. One can conjecture that this area is involved in higher-level abstractions that can facilitate indirect associations [13].

It is probable that the initial impasse in problem solving is due to the inability of the processes in the left hemisphere, focused on the precise representation of the problem, to make progress in finding a solution. 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 among concepts, relations that have no name, but are useful in reasoning and understanding. For example, "left kidney" sounds correct, but "left nose" seems strange, although we do not have a concept for spatially extended things that "left" applies to. The feeling arising from understanding words and sentences may be connected to the left-right hemisphere activation interplay. Most of RH activations do not have phonological components; the activations result from diverse associations, temporal dependencies and statistical correlations that create certain expectations. It is not clear what brain mechanism is behind the signaling of this lack of familiarity, but one can assume that interpretation of text is greatly enhanced by "large receptive fields" in the RH, which can constrain possible interpretations, help in the disambiguation of concepts and provide ample stereotypes and prototypes that generate various expectations.

Associations at higher level of abstraction in the RH are passed back to facilitate LH activations that form intermediate steps in language interpretation and problem solving. The LH impasse is removed when relevant activations are projected back from the less-focused right hemisphere, allowing new dynamical associations to be formed. 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. An emotional component is needed to increase the plasticity of the brain and remember these associations. The "Aha!" experience may thus result from the activation of the left hemisphere areas by the right hemisphere, with a gamma burst helping to bring relevant facts to the working memory, making them available for conscious processing. This process occurs more often when the activation of the left hemisphere decreases (when the conscious

efforts to solve the problem are given up). Sometimes it leads to a brief feeling that the solution is imminent, although it has not yet been formulated in symbolic terms, a common feeling among scientist and mathematicians. The final step in problem solving requires synchronization between several left hemisphere states, representing transitions from the start to the goal through intermediate states. This seems to be a universal mechanism that should operate not only in solving difficult problems, but also on much shorter time scale in understanding of complex sentences. The feeling "I understand" signifies the end of the processing and readiness of the brain to receive more information.

Such observations may be used as inspirations for neurocognitive models. Distal connections between the left and right hemispheres require long projections, and therefore neurons in the right hemisphere may generalize over similar concepts and their relations. Distributed activations in the right hemisphere form configurations that activate extended regions of the left hemisphere. High-activity gamma bursts projected to the LH prime its subnetworks with sufficient strength to allow for synchronization of groups of neurons that create distant associations. In problem solving, this synchronization links brain states coding the initial description D of the problem with its partial or final solutions S. Such solutions may initially be difficult to justify, they become clear only when all intermediate states T_k between D and S are transversed. If each step from T_k to T_{k+1} is an easy association, a series of such steps is accepted as an explanation. An RH gamma burst activates emotions, increasing the plasticity of the cortex and facilitating the formation of new associations between initially distal states. The same neural processes should be involved in sentence understanding, problem solving and creative thinking.

According to these ideas, approximation of the spreading activation in the brain during language processing should require at least two networks activating each other. Given the word $w = (w_f,w_s)$ with phonological (or visual written) component w_f, an extended semantic representation w_s, and the context *Cont*, the meaning of the word results from spreading activation in the left semantic network *LH* coupled with the right semantic network *RH*, establishing a global state $\Psi(w,Cont)$. This state changes with each new word received in sequence, with quasi-stationary states formed after each sentence is understood. It is quite difficult to decompose the $\Psi(w,Cont)$ state into components, because the semantic representation w_s is strongly modified by the context. The state $\Psi(w,Cont)$ may be regarded as a quasi-stationary wave, with its core component centered on the phonological/visual brain activations w_f and with quite variable extended representation w_s. As a result the same word in a different sentence creates quite different states of activation, and the lexicographical meaning of the word may be only an approximation of an almost continuous process. To relate states $\Psi(w,Cont)$ to lexicographical meanings, one can clusterize all such states for a given word in different contexts and define prototypes $\Psi(w_k,Cont)$ for different meanings w_k.

The high-dimensional vector model of language popular in statistical approach to natural language processing [14] is a very crude approximation that does not reflect essential properties of the perception-action-naming activity of the brain [6][7]. The process of understanding words (spoken or read) starts from activation of the phonological or grapheme representations that stimulate networks containing prior knowledge used for disambiguation of meanings. This continuous process may be approximated through a series of snapshots of microcircuit activations $\phi_i(w,Cont)$ that may be treated as basis functions for the expansion of the state $\Psi(w,Cont) = \Sigma_i\ \alpha_i\ \phi_i(w,Cont)$, where the summation extends over all microcircuits that show significant activity resulting from presentation of the word w. The high-dimensional vector model used in NLP measures only the co-occurrence of words $\mathbf{V}_{ij} = \langle \mathbf{V}(w_i),\mathbf{V}(w_j)\rangle$ in some window, averaged over all contexts. A better approximation of the brain processes involved in understanding words should be based on the time-dependent overlap between states $\langle \Psi(w_1,Cont) \mid \Psi(w_2,Cont)\rangle = \Sigma_{ij}\ \alpha_i\,\alpha_j\ \ \langle \phi_i(w_1,Cont) \mid \phi_j(w_2,Cont)\rangle$. Systematic study of transformations between the two bases: activation of microcircuits ϕ_i and activation of complex patterns $\mathbf{V}(w_i)$, has not yet been done for linguistic representations in the brain. Analysis of memory formation in mice hippocampus [8] in terms of combinatorial binary codes signifying activity of neuronal cliques goes in this direction. The use of wave-like representation in terms of basis functions to describe neural states makes this formalism similar to that used in quantum mechanics, although no real quantum effects are implied here.

Spreading activation in semantic networks should provide enhanced representations that involve concepts not found directly in the text. Approximations of this process are of great practical and theoretical interest. The model should reflect activations of various concepts in the brain of an expert reading such texts. A few crude approximations to this process may be defined. First, semantic networks that capture many types of relations among different meanings of words and expressions may provide space on which words are projected and activation spread. Each node w in the semantic network represents the whole state $\Psi(w,Cont)$ with various contexts clustered, leading to a collection of links to other concepts found in the same cluster that capture the particular meaning of the concept. Usually only the main differences among the meanings of the words with the same phonological representation are represented in semantic networks (meanings listed in thesauruses), but the fine granularity of the meanings resulting from different contexts may be captured in the clusterization process and can be related to the weights of connections in semantic networks. The spreading activation process should involve excitation and inhibition, and "the winner takes most" processes. Current models of semantic networks used in NLP are only vaguely inspired by the associative processes in the brain and do not capture such details [14]-[16].

Such crude approximation to the spreading activation processes leads to an enhancement of the initial text being analyzed by adding new concepts linked by episodic, semantic or hierarchical ontological relations. The winner-takes-all processes

lead to inhibition of all but one concept that has the same phonological word form. Locally this may be represented as a subnetwork (graph) of consistent concepts centered around a prototype for a given word meaning $\Psi(w_k,Cont)$, linking it to words in the context. Such approach has been applied recently to disambiguate concepts in medical domain [17]. The enhanced representations are very useful in document clusterization and categorization, as is illustrated using short medical texts in the next section. Vector models may be related to semantic networks by looking at snapshots of the activation of nodes after several steps of spreading the initial activations through the network. In view of the remarks about the role of the right hemisphere, larger "receptive fields" in the linguistic domain should be defined and used to enhance text representations. This is much more difficult because many of these processes have no phonological component and thus have representations that are less constrained and have no directly identifiable meaning. Internal representations formed by neural networks are also not meaningful to us, as only the final result of information processing or decision making can be interpreted in symbolic terms. Defining prototypes for different categories of texts, clusterizing topics or adding prototypes that capture some *a priori* knowledge useful in document categorization [19], is a process that goes in the same direction.

Associative memory processes involved in spreading activation are also related to creativity, as has already been noticed a long ago [20]. Experimental support for the ideas described above may be found in pairwise word association experiments using different priming conditions. Puzzling results from using nonsensical words were observed [21] for people with high compared to those with low creativity levels. Analysis of these experiments [22] reinforces the idea that creativity relies on associative memory, and in particular on the ability to link distant concepts together. Adding neural noise by presenting nonsensical words in priming leads to activation of more brain circuits and facilitates in a stochastic resonance-like way a formation of distal connections for not obvious associations. This is possible only if weak connections through chains involving several synaptic links exist, as is presumably the case in creative brains. For simple associations the opposite effect is expected, with strong local activations requiring longer times for the inhibitory processes to form consistent interpretations. Such experiments show that some effects cannot be captured at the symbolic level. It is thus quite likely that language comprehension and creative processes both require subsymbolic models of neural processes realized in the space of neural activities, reflecting relations in some experiential domain, and therefore cannot be modeled using semantic networks with nodes representing whole concepts. Recent results on creation of novel words [22] give hope that some of this process can be approximated by statistical techniques at the morphological level.

Qualitative picture that follows through this (admittedly speculative) analysis 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. A practical algorithm will be presented below and applied to categorization of short medical texts.

3. Discharge summaries

The data used in this paper comes from the Cincinnati Pediatric Corpus (IRB approval 0602-37) developed by the Biomedical Natural Language Processing group at 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 [23]. Processing of medical texts requires resolution of ambiguities and mapping terms to the Unified Medical Language System (UMLS) Metathesaurus concepts [24]. 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 [19] 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. Detailed preparation of the database and the pre-processing steps have already been described in [19], therefore it is not repeated here. Summary given in Table I allows to relate disease names to class numbers displayed in Fig. 1 and Fig. 2.

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) [24] is a collection of many

medical concept ontologies that are used to discover useful concepts and their relations, enabling semantic smoothing.

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* (*J.R.A.*)	41	1816
7. *Cystic fibrosis*	283	1790
8. *Cerebral palsy*	177	1597
9. *Otitis media*	493	1420
10. *Gastroenteritis*	586	1375

4. 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) for text fragments [19], but here no prior knowledge is assumed. 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:

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

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 methods 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 [14], 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 [25]. 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, but this approach captures only one type of relations between words. 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.

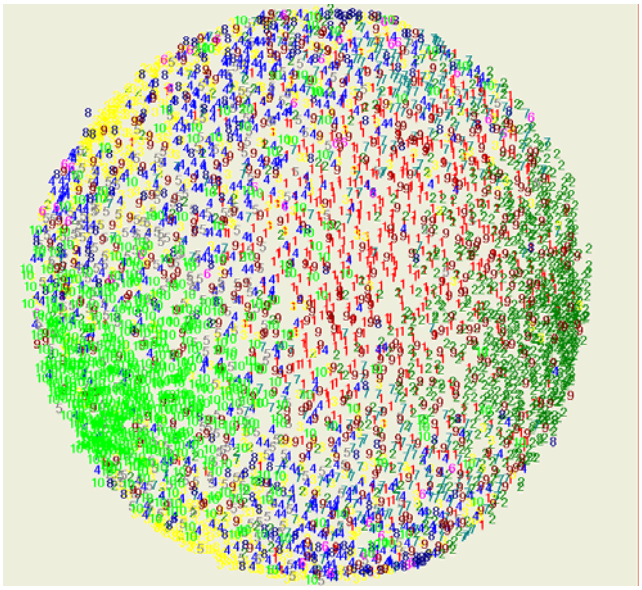

Fig. 1. MDS representation of 4534 summary discharges documents, showing little clusterization.

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 [25]. 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 [26], or other measures of similarity of probability distributions [27]-[29], such as the Expected Mutual Information Measure (EMIM) [28].

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][14][16][17], 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 many such relations as: 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 a general "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 Person's correlation, Fisher's criterion, mutual information, or other feature ranking indices [26]. 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 general 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 [30] 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 [26].
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.

This general algorithm should help to discover associations that are important from the categorization point of view, and thus understanding of text topics. It may help to build semantic network with useful associations. The algorithm may have different variants. Although ULMS ontology is huge it is difficult to manage, as it contains many strange concepts that are not useful in our application. Therefore a smaller MeSH (Medical Subject Headings) ontology [31], a subset of the whole ULMS, may been used as an alternative, a collection of keywords used to describe papers in medical domain. All MeSH terms are relevant and clinical free text may be mapped to the MeSH concepts using the same MetaMap (MMTx) software as has been used for mapping to ULMS [30]. Other medical ontologies, such as SnowMed, may be used instead of MeSH. Different dimensionality reduction may be used in practice, but here only Pearson's correlation coefficient will be used for ranking.

Identification of precise topics, or in our case subtypes of disease, may proceed in a slightly different way. First, all very rare concepts (say, appearing in less than 1% of all documents) may be removed, and all documents that contain very few concepts used as features (say, less than 1%) are also removed, reducing uncertainty of clusterization. This approach would not be appropriate for document categorization, when all documents should be categorized, and even rare concepts may be useful, but for topic identification it should increase reliability of the task. Second, semantic pruning should be applied, leaving only those concepts that belong to specific semantic types. These types may be selected using filter methods, removing all features of specific type and checking classification accuracy. Below only those MeSH concepts that belong to 26 ULMS semantic types will be used. In this task binary representation is sufficient; it has worked well in categorization of summary discharges [19] and has been sufficient in analysis of activations of neural cliques [8]. ULMS relations are used to create new features. The number of documents $N(t)$ and the number of features $n(t)$ may both change during iterations. Each document D_{ij}, $i=1..N(t)$ is represented by a row of $j=1..n(t)$ binary features, therefore the whole vector representation of all documents in iteration t is given by a matrix $\mathbf{D}(t)$ with $N(t)$x $n(t)$ dimensions. ULMS contains relations R_{ij} between concepts i and j; selecting only those concepts i that have been used as features to create matrix $\mathbf{D}(t)$ and concepts j that are related to i and representing existence of each relation as a Kronecker δ_{ij} a binary matrix $\mathbf{R}(t)$ is created. Multiplying the two matrices $\mathbf{D}(t)\mathbf{R}(t) = \mathbf{D'}(t)$ gives an expanded matrix $\mathbf{D'}(t)$ with new columns defining enhanced feature space. These columns contain integer values indicating to which document the new concept is associated. For class $k = 1 .. K$ a binary vector $\mathbf{Ck}_i = \delta_{ik}$, $i= 1..N(t)$ serves as a class indicator of all documents. To evaluate the usefulness of the candidate features the Pearson correlation coefficient between these columns and all vectors that are class indicators are calculated. Only those candidate features with the highest absolute value of the correlation coefficients are retained and after removal of some $\mathbf{D'}(t)$ matrix columns and binarization of the remaining ones it is converted to a new current matrix $\mathbf{D}(t+1)$. For binary vectors there may be several features with the same largest correlation coefficient.

5. 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) [19]. Each document has been processed by the MetaMap software [30] 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 [19].

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 re-

sults). 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 [19] 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 [31] (as implemented in the Ghostminer package [36] 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 [19] 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 [37] 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 [26] 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 [38] 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.

MDS shows some artifacts due to the large number of points and high dimensionality of the data. Therefore a more detailed study to identify interesting clusters of documents that result from a subclass of a disease has been pursued. Unfortunately medical textbooks do not present such subtypes and their prevalence, mentioning instead quite rare cases and procedures that few doctors will ever encounter, therefore such background knowledge is hard to find, otherwise it could be used as a prior knowledge. Below we have focused on searching for interesting clusters that represent a clinical prototype of a particular subtype of the disease (clinotype).

Two classes *Pneumonia* and *Otitis Media*, that show strongest mixing. In addition *J.R.A.* has been added as the third class, a rare disease with small number of cases. There are 610 discharge summaries with *Pneumonia* as initial diagnosis, 493 summaries with *Otitis Media* and only 41 discharge summaries with *J.R.A*. as initial diagnosis. The MeSH (Medical Subject Headings) ontology [31] has been used. The summary discharges texts have been mapped on the 2007 MeSH concepts using the MetaMap (MMTx) software [30].

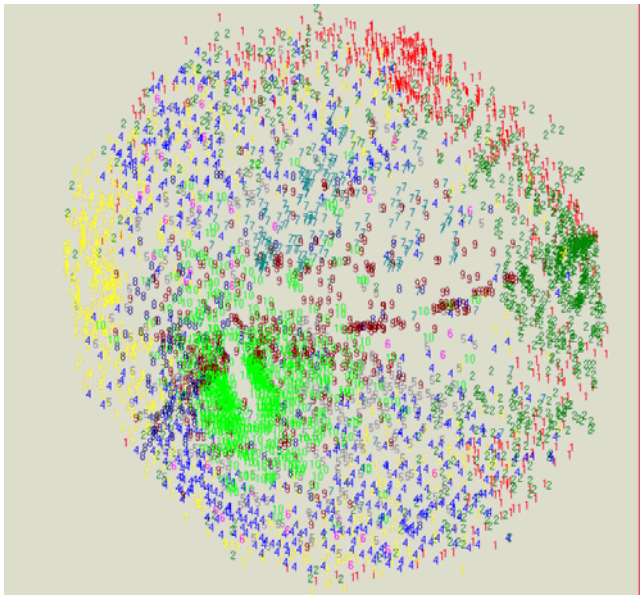

Fig. 2. Clusterization of document database after two steps of spreading activation using UMLS ontology.

Resulting concepts found in summary discharges have then been used to create a binary matrix, where rows correspond to documents and columns to concepts. This created an initial matrix with 1144 rows and 2874 columns. The number of documents and the number of concepts has then been reduced by removing all rows and columns that had less than 1% or more than 99% of non-zero values. In effect documents that have too few concepts to be reliably handled and concepts that are very rarely used have been deleted. This is justified because the goal here is to look for common clinotypes, in other applications all documents and rare concepts should be carefully analyzed. This frequency-based pruning reduced the number of columns (concepts) to 570 and the number of rows to 1002, including 542 discharge summaries from the *Pneumonia*, 426 from *Otitis Media* and 34 discharge summaries from the *J.R.A.* class.

In the next step semantic pruning has been performed. MeSH concepts were filtered through the 26 UMLS semantic types. This reduced the number of concepts that have been used as features to a modest 224 and the number of documents to 908. Semantic types that contributed more than 10 concepts, in order of their prevalence, are: *Disease or Syndrome, Pharmacological Substance, Sign or Symptom, Antibiotic, Therapeutic or Preventive Procedure, Body Part, Organ, or Organ Component, Diagnostic Procedure, Body Location or Region, Body Substance, Biologically Active Substance,* and *Finding.* 4 semantic types did not appear in the text left at this stage: *Anatomical Abnormality, Biological Function, Clinical Drugs, Laboratory or Test Result*, and *Vitamins*. However, concepts of these semantic types appear frequently in medical documents for other diseases.

Enhancement of the feature space constructed in such a way is done by iterative spreading of activation. Not all connections lead to activation of new concepts. After every iteration frequency pruning and semantic pruning is conducted. Activation is spread in a non-linear process that tries to approximate complex process of excitations and inhibitions of associated concepts in the brain of an expert. Since UMLS provides information about excitatory connections only an additional source of knowledge is needed to inhibit most of the concepts. This is done by looking at the usefulness of each new feature activated by the semantic relations for understanding the category structure of the document set. Let $\mathbf{A}(t)$ be the state of a binary document/concept matrix after the spreading activation step t. If UMLS has relation between concept i that is included in our space and concept j that is not yet included then a new column $\mathbf{A}(t)_j$ will be created. After scanning all relations $\mathbf{A}(t')$ matrix is created. Some form of ranking should now be done to leave only the most relevant features. Although this could be done in many ways [26] here Pearson's correlation coefficient between new features and each class is calculated and only the concepts that has highest correlation coefficient with respect to class that this document belongs to is chosen.

Thus enhancements of the feature space are followed by reductions, leaving in the end only the most informative features, including many features that do not appear in the original documents. Approximately 40 new features are added after each iteration of this algorithm, 294, 328, 360, 400, 438, 483, 526, 570, 617, 671, 727 ... For example, in the first step *Chronic Childhood Arthritis* adds *Sulfasalazine*, *Tolmetin*, *Aurothioglucose*, while *Acetylcysteine* is added by *Cystic Fibrosis, Pneumonia*, and *Lung diseases*; *Lateral Sinus Thrombosis* is added by *Otitis Media*, and *Brain* Concepts. In the second iterations these new concepts add further concepts, for example *Sulfasalazine* adds 3 concepts, *Ureteral obstruction, Porphyrias,* and *Bladder neck obstruction*, while the *Urinary tract infection* and *Otitis Media* add *Sulfisoxazole* and *Sulfamethoxazole*. For about 200 concepts there are about 70.000 ULMS relations described as “related or other than structural”, but only about 500

of them belong to one of the 26 semantic types selected here. Out of those 500 only about 35-40 that correlate most with predefined classes are actually added in each iteration.

How is this procedure influencing the quality of the clusters? Distance measure used is based on dissimilarity of Pearson's correlation coefficients. As a clustering technique Ward's algorithm has been used [32]. It partitions the data set in a manner that minimizes the loss associated with each grouping, and quantifies this loss in a form that is readily interpretable. At each step of the analysis, the union of every possible cluster pair is considered, and the two clusters whose fusion results in minimum increase of the 'information loss' measure are combined. Information loss is defined by Ward in terms of an error sum-of-squares criterion. As a measure of clustering quality silhouette width [33] has been used, as implemented by Brock *et al.* [34]. The silhouette width is the average of each observation's silhouette value, a measure of the degree of confidence in the clustering assignment of a particular observation. Let $a(\mathbf{X}_i)$ denote the average dissimilarity of vector $\mathbf{X}_i$ to all vectors in its own cluster and let $b(\mathbf{X}_i)$ be the average dissimilarity to the second best cluster for the vector $\mathbf{X}_i$. Silhouette value for the vector $\mathbf{X}_i$ is defined as:

$$Sil(\mathbf{X}_i) = \frac{b(\mathbf{X}_i) - a(\mathbf{X}_i)}{\max\left(a(\mathbf{X}_i), b(\mathbf{X}_i)\right)} \tag{2}$$

and the total silhouette value is simply the average value for all vectors. Well-clustered observations have values near 1 and poorly clustered observations having values near −1.

As a result of consistent use of frequency pruning, semantic pruning and conceptual pruning at every step of activation quality of clusters is increased. Even though the feature space is increased at approximately constant rate the increase of quality and convergence of the algorithm with growing number of clusters (shown in Fig. 3) are far from linear, but after 12 iterations changes are minimal. The number of candidate concepts before pruning is quite large (activation may spread to many related concepts) and increases significantly, from 21186 to 34156 after 9 iterations.

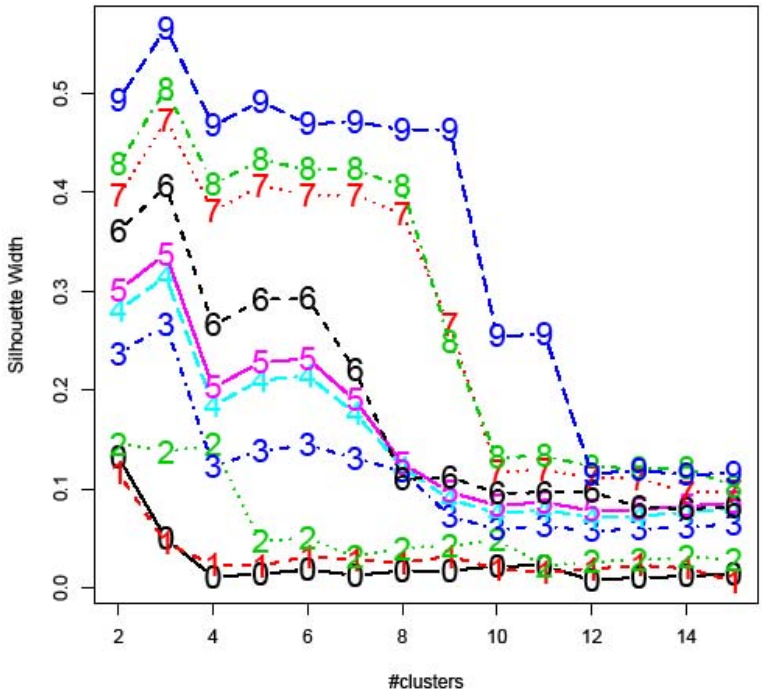

Fig. 3. Relation between the number of clusters and the quality of clustering in the first 15 iterations.

Fig. 3 shows that enhancing feature space not only creates better quality clusters but justifies larger number of tighter clusters then there are classes. Ward clusterization knowing that there are 3 classes tries to create 3 clusters, but for larger number of clusters the quality remains also high. After six iterations 6 clusters seem to be optimal, but after nine iterations even 9 clusters are acceptable. 5 clusters have been extracted after 4 iterations, as shown in the MDS visualization, Fig. 4.

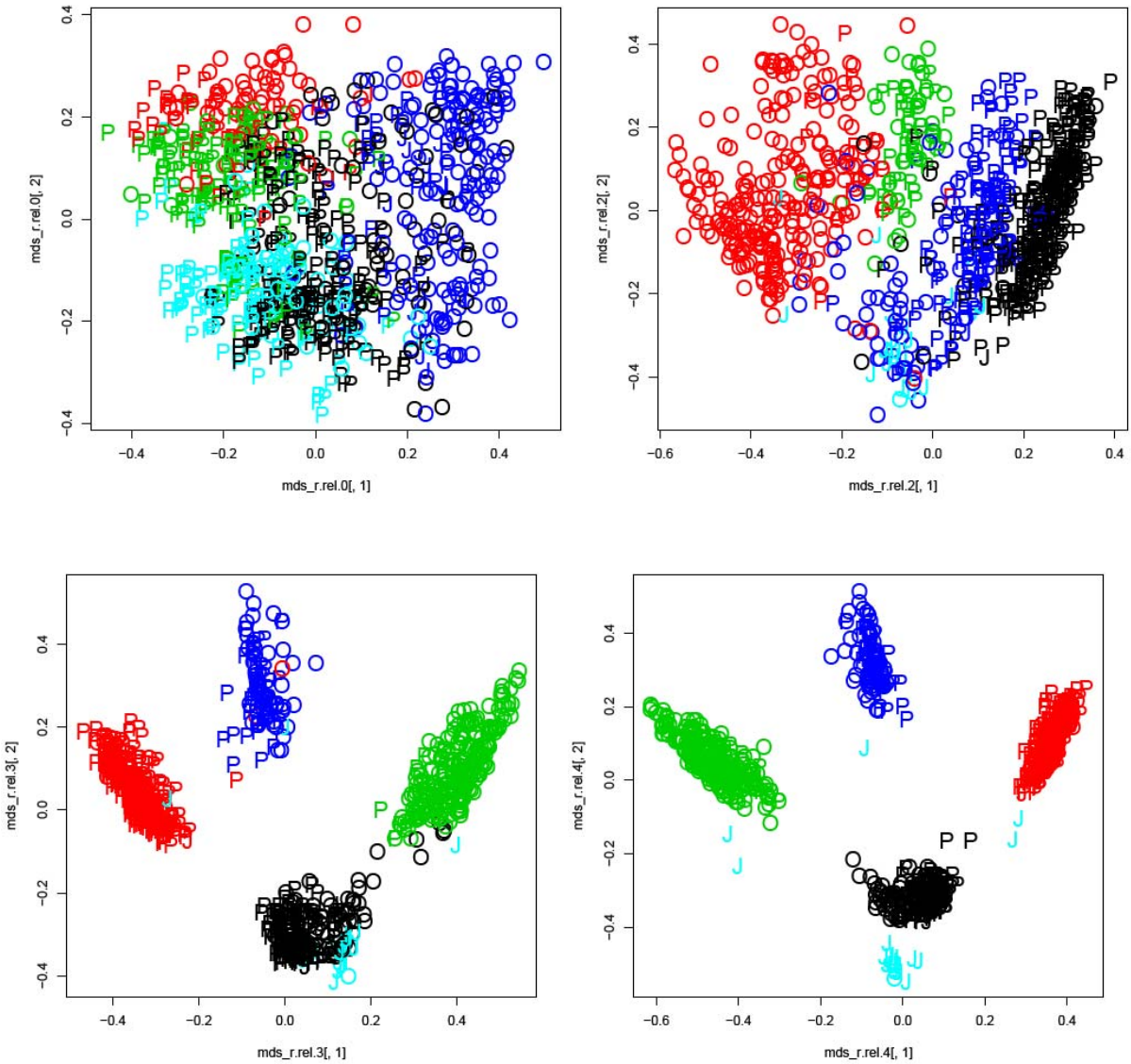

Fig. 4. MDS representation of 3 selected classes (P, O, J), top left: original data, followed by iterations 2, 3 and 4.

Initially 5 clusters have been distinguished, the first dominated by 199 P (*pneumonia*) documents, with 68 *otitis media* (O) and 6 *J.R.A* (J) documents, the second with 40 P, 62 O, and 1 J, the third with 127 P and 26 O, the fourth with 28 P, 163 O and 22 J, and the fifth with 103 P and 63 O documents. After 4 iterations a small *J.R.A* cluster (26 cases) with a single *otitis media* case is left at the bottom, above it a cluster with mixed P (131) and O (70) and 2 J cases, on the right a cluster with clear dominance of P (304), 17 O and 1 J case, on the left a cluster with dominance of O (240) and a few P (17), and a top cluster with 54 O and 45 P. There is no reason why these clusters should become pure, as some documents represent clearly mixed cases, with two or more diseases described, or with different subtypes (clinotypes) of the disease. What is important is that these clusters are quite distinct and thus may be analyzed by domain experts that should be able to describe such clinotype and distinguish it from others.

Clinotypes consist of a set of medical concepts that appear as clusters. A particular concept belongs to a given clinotype if it has high Pearson's correlation coefficient with documents (or with representative) from the cluster denoting this clinotype. As an example 5 clinotypes after 4 iterations of spreading activation are presented, showing new concepts from enhanced feature space. Only the most important concepts are showed, with concepts contributed from the original space listed first, followed by those from iterative enhancements, with the iteration number in which they appear first given in parenthesis.

Clinotype 1, bottom cluster in Fig. 4, obviously related to *J.R.A.*: Original space contributed *Chronic Childhood Arthritis, Methotrexate, Knee, Joints, Hip region structure and Rehabilitation therapy*, all with correlation coefficients above 0.30;
Enhanced Space: *Porphyrins* (4), *Porphyrias* (2), *Cyproheptadine* (4), *Tolmetin* (1), *Chlorpheniramine* (4) and *Ureteral obstruction* (2).

Clinotype 2: above the bottom cluster in Fig. 4, has been quite mixed and the feature space has not been expanded by our procedure. Original space: *Pneumonia, Bacterial; Pneumonia, Pneumococcal; Obstetric Delivery, Cefuroxime, Influenza, Ibuprofen* with relatively weak correlation coefficients 0.06-0.10. Interpretation of this cluster is not clear.

Clinotype 3: right cluster in Fig. 4. Original space contributes: *Pneumonia, Chest X-ray, Oxygen, Azithromycin, Cefuroxime, Coughing*, with correlation coefficients between 0.8 to 0.15.
Enhanced Space: *Deslanoside (3), Amyloid (3), Colchicine (3), Amyloidosis (2), Digitoxin (2) and Amyloid Neuropathies* (3). This cluster is related to pneumonia.

Clinotype 4: left cluster in Fig. 4. Original Space: *Otitis Media, Erythema, Tympanic membrane structure, Amoxicillin, Eye* and *Ear structure*, with correlation coefficients from 0.78 to 0.16.
Enhanced Space contributes: Respiratory Tract Fistula (4), Trimethoprim-Sulfamethoxazole Combination (4), Sulfisoxazole (2), Nocardia Infections (3), Sulfamethoxazole (2), Chlamydiaceae Infections (4). This is obviously related to *otitis media*.

Clinotype 5: top cluster in Fig. 4. Original Space: *Otitis Media, Pneumonia, Urinary tract infection, Virus Diseases, Respiratory Sounds, Chest X-ray*, with correlation coefficients between 0.40 and 0.10.
Enhanced Space: *Respiratory Tract Fistula (4), Trimethoprim-Sulfamethoxazole Combination (4), Sulfisoxazole (2), Nocardia Infections (3), Sulfamethoxazole (2) and Chlamydiaceae Infections (4)*. This is another type of *otitis media* inflammation.

6. Conclusions and wider implications

The use of background knowledge in natural language processing 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. General inspirations for neurocognitive linguistics have been outlined, drawing on recent experiments with priming, insight and creativity. A novel theoretical approach linking NLP vector models and description of brain activations has been outlined. Vector representations of concepts may be regarded as a snapshot of dynamic activity patterns, defining connections with other concepts. However, more work on relations between spreading activation networks, semantic networks and vector models is needed.

Creating useful numerical representation of various concepts is an interesting challenge. Large-scale semantic networks and spreading activation models may be constructed starting from large ontologies. 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 also be helpful here [39]. A crude version of such approach has been presented here and already using the second-order expansions gave quite good results 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 using prototypes. 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 be observed in clusterization of these documents if a proper space is defined. Clusters in Fig. 2 and 4 may be interpreted in this way, although MDS mapping to two dimensions only has to introduce many distortions. It is relatively easy to collect information about rare cases that are subject to scientific investigation, but not the subtypes of the common ones. Finding such subclusters, or identifying different subtypes of disease, is an interesting goal that may potentially help to train young 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 activity of the brain representing concepts and making inferences during sentence comprehension is a great challenge for neural modeling. In medical domain ontologies, relations between concepts, and classification of semantic types enables useful approximations to the neurolinguistic processes, while in general domains resources of this sort are still missing.

Acknowledgment

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

References

[1] Lamb, S. (1999). *Pathways of the Brain: The Neurocognitive Basis of Language*. Amsterdam & Philadelphia: J. Benjamins Publishing Co.
[2] Rumelhart, D.E. & McClelland, J.L. (eds), (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition Vol. 1: Foundations, Vol. 2: Psychological and Biological Models*. Cambridge, MA: MIT Press.
[3] Crestani, F. (1997). Application of Spreading Activation Techniques in Information Retrieval. *Artificial Intelligence Review 11*, 453-482.
[4] Crestani, F., & Lee, P.L. (2000). Searching the web by constrained spreading activation. *Information Processing & Management 36*, 585-605.
[5] Tsatsaronis, G. Vazirgiannis, M., & Androutsopoulos, I. (2007) Word Sense Disambiguation with Spreading Activation Networks Generated from Thesauri, in *20th Int. Joint Conf. in Artificial Intelligence* (IJCAI 2007), Hyderabad, India, pp. 1725-1730.
[6] Pulvermuller, F. (2003), *The Neuroscience of Language. On Brain Circuits of Words and Serial Order*. Cambridge, UK: Cambridge University Press.
[7] Dehaene, S., Cohen, L. Sigman, M. & Vinckier, F. (2005) The neural code for written words: a proposal. *Trends in Cognitive Science 9*, 335-341.
[8] Lin, L., Osan, R., & Tsien, J.Z. (2006). Organizing principles of real-time memory encoding: neural clique assemblies and universal neural codes. *Trends in Neuroscience 29(1)*, 48-57.
[9] Dehaene, S., & Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. *Cognition 79*, 1-37.
[10] Duch, W. (2005). Brain-inspired conscious computing architecture. *Journal of Mind and Behavior 26(1-2)*, 1-22.
[11] Gaillard, R., Naccache, L., Pinel, P., Clémenceau, S., Volle, E.., Hasboun, D., Dupont, S., Baulac, M., Dehaene, S., Adam, C., & Cohen, L. (2006). Direct intracranial, FMRI, and lesion evidence for the causal role of left inferotemporal cortex in reading. *Neuron 50*, 19-204.
[12] Bowden, E.M., Jung-Beeman, M., Fleck, J. & Kounios, J. (2005). New approaches to demystifying insight. *Trends in Cognitive Science 9*, 322-328.
[13] Jung-Beeman, M., Bowden, E.M., Haberman, J., Frymiare, J.L., Arambel-Liu, S., Greenblatt, R., Reber, P.J., &. Kounios, J. (2004). Neural activity when people solve verbal problems with insight. *PLoS Biology 2*, 500-510.
[14] Manning, C.D. & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing* Cambridge, MA: MIT Press.
[15] Sowa J.F. (Ed), (1991). *Principles of Semantic Networks: Explorations in the Representation of Knowledge*. San Mateo, CA: Morgan Kaufmann Publishers.
[16] Lehmann F. (Ed), (1992). *Semantic Networks in Artificial Intelligence*. Oxford, Pergamon.
[17] Matykiewicz, P., Duch, W., & Pestian, J. (2006). Nonambiguous Concept Mapping in Medical Domain, *Lecture Notes in Artificial Intelligence 4029*, 941-950.
[18] Duch, W. (2007) Creativity and the Brain. In: A Handbook of Creativity for Teachers. Ed. Ai-Girl Tan, Singapore: World Scientific Publishing, pp. 507-530.
[19] Itert, L. Duch, W. & Pestian, J. (2007). Influence of *a priori* Knowledge on Medical Document Categorization, *IEEE Symposium on Computational Intelligence in Data Mining*, IEEE Press, pp. 163-170.
[20] Mednick, S.A. (1962). The associative basis of the creative process. *Psychological Review 69*, 220–232.
[21] Gruszka, A., & Nęcka, E. (2002). Priming and acceptance of close and remote associations by creative and less creative people. *Creativity Research Journal 14,* 193-205.
[22] Duch W., & Pilichowski, M. (2007). Experiments with computational creativity. *Neural Information Processing - Letters and Reviews 11*, 123-133.
[23] Pestian, J. Aronow, B. & Davis, K. (2002). Design and Data Collection in the Discovery System. In *Int. Conf. on Mathematics and Engineering Techniques in Medicine and Biological Science*.
[24] UMLS Knowledge Sources, 13th Edition – January Release. Available: http://www.nlm.nih.gov/research/umls
[25] Zhou, X., Zhang, X. & Hu X., (2007). Semantic Smoothing of Document Models for Agglomerative Clustering, *20th Int. Joint Conf. on Artificial Intelligence (IJCAI 2007)*, Hyderabad, India, pp. 2922-2927.
[26] Duch, W. (2006). 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, pp. 89-118.
[27] Cimiano, P. (2006). *Ontology Learning and Population from Text. Algorithms, Evaluation and Applications*. Springer.
[28] Bein, W.W., Coombs, J.S., & Taghva, K. (2003). A Method for Calculating Term Similarity on Large Document Collections. *Int. Conf. on Information Technology: Computers and Communications*, pp. 199-207.
[29] Li, Y., Zuhair, A.B. & McLean, D. (2003). An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources. *IEEE Transactions on Knowledge and Data Engineering 15(4)*, 871-882.
[30] MetaMap, available at http://mmtx.nlm.nih.gov
[31] Medical subject headings, MeSH, National Library of Medicine, URL: http://www.nlm.nih.gov/mesh/.
[32] Ward, J.H. (1963). Hierachical grouping to optimize an objective function. *J. American Statistical Association 58*, 236-244.
[33] Rousseeuw, P.J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math. 20*, 53–65.
[34] Brock, G., Pihur, V., Datta, S., & Datta. S., see http://cran.r-project.org/src/contrib/Descriptions/clValid.html
[35] Grąbczewski, K., & Duch, W. (2000). The separability of split value criterion, 5th Conf. on Neural Networks and Soft Computing, Zakopane, Poland, pp. 201-208.
[36] Ghostminer data mining software, www.fqspl.com.pl/ghostminer/
[37] Duch, W. & Diercksen, G.H.F. (1995). Feature Space Mapping as a universal adaptive system. *Computer Physics Communications 87*: 341-371.
[38] Pękalska, E. & Duin, R.P.W. (2005). *The dissimilarity representation for pattern recognition: foundations and applications*. New Jersey; London: World Scientific.
[39] Antonie, M.-L. & Zaiane, O.R. (2002). Text document categorization by term association. *Proc. of IEEE Int. Conf on Data Mining (ICDM),* pp. 19- 26.