Multiple inheritance problem in semantic spreading activation networks

Paweł Matykiewicz[1] and Włodzisław Duch[2]

[1] Division of Biomedical Informatics,
Cincinnati Children's Hospital Medical Center,
3333 Burnet Ave, Cincinnati OH 45220, USA,
pawel.matykiewicz@cchmc.org
[2] Department of Informatics, Faculty of Physics,
Astronomy and Informatics, Nicolaus Copernicus University,
ul. Grudziadzka 5, 87-100, Torun, Poland,
wduch@is.umk.pl

Abstract. Semantic networks inspired by semantic information processing by the brain frequently do not improve the results of text classification. This counterintuitive fact is explained here by the multiple inheritance problem, which corrupts real-world knowledge representation attempts. After a review of early work on the use of semantic networks in text classification, our own heuristic solution to the problem is presented. Significance testing is used to contrast results obtained with pruned and entire semantic networks applied to medical text classification problems. The algorithm has been motivated by the process of spreading neural activation in the brain. The semantic network activation is propagated throughout the network until no more changes to the text representation are detected. Solving the multiple inheritance problem for the purpose of text classification is similar to embedding inhibition in the spreading activation process – a crucial mechanism for a healthy brain.

Keywords: text classification, semantic networks, multiple inheritance problem, spreading activation networks, inhibition

1 Introduction

The neurocognitive approach to language has not led to practical algorithms [15] in natural language processing (NLP). Semantic networks, inspired by the work on human semantic memory, are a convenient way to store general information about the world. Using graphical notation network nodes are identified with concepts, and edges with semantic relations, allowing for direct logical inferences [19]. This seemed to be an obvious improvement over the most popular representation of texts in form of vectors built by counting the number of distinct words in a "bag-of-words" approach [12].

Scott and Matwin [21] experimented with bag-of-words, stemmed words, noun phrases, stemmed noun phrases, key phrases, stemmed key phrases, WordNet synonyms, and WordNet hypernyms. In WordNet [7], a huge lexical database of the English language, words are grouped together by their synonymy with basic semantic relations between them. An artificial intelligence scholar would call it one of the largest non-monotonic logical systems. Unfortunately, Scott and Matwin [21] experiments did not demonstrate the superiority of using synonyms and hypernyms over a bag-of-words. Only certain combinations of phrases and words improved results. In case of a general purpose semantic

network using synonyms and hypernyms does not help much. This was unacceptable and non-intuitive for many researchers. How can adding knowledge to a representation based on a simple word count degrade text classification performance? Isn't adding such associations the key to how brains work?

Let us think of a simple example: "Humans have two legs", "John is a human", "John has one leg". In this semantic network, "John" inherits multiple contradictory properties: having one leg and having two legs. Such a small semantic network exposes an important fact: in real life not all assertions about the world are true all the time. In such cases, how should the knowledge encoded in a semantic network be used for inference? While episodic memory may come to rescue [5], it would lead to significant complication of the algorithm. It took many years to find the answer to this problem. In essence, a general purpose semantic network must be pruned before it may be used to enhance a text representation. From the artificial intelligence and NLP perspective, one needs to solve the multiple inheritance problem before using a non-monotonic knowledge representation system.

Inspired by many findings relevant to neurolinguistics [1, 24, 6], our goal is to create a neurocognitive language processing approach inspired by the spreading neural activation over a large semantic brain network [5]. Thus far, we have developed a practical method for pruning semantic networks in a way that improves the results of text categorization. Our algorithm spreads activation from one term to the other, inferring facts not present in the text, but preserving only those facts that improve text classification, avoiding unnecessary inheritance. In this way relevant "pathways of the brain" [15] are discovered. **We will show that text classification with a pruned semantic network is significantly better than a baseline model, while using entire network does not lead to improvements.**

2 Background and Significance

Text classification is pursued by the statistical machine learning community; the term "multiple inheritance" comes from the field of artificial intelligence. Textbooks on artificial intelligence mention statistical learning, but the converse is rarely true. The multiple inheritance problem has been rarely addressed in literature on text classification. Google Scholar cites over 10,100 articles that mention "text classification" and some version of a semantic network[3]. Only 127 of them mention "multiple inheritance"[4]. That is not to say that the term "multiple inheritance" is unknown to the language processing community. There are over 1,600 articles on "multiple inheritance" and "semantic networks"[5]. Most of them, however, discuss semantic similarity and dismiss the multiple inheritance problem by taking maximal, average or minimal paths between two terms [18]. Semantic similarity will not be discussed in this paper, the focus will be on text classification using semantic networks.

Typical work in this field follows the following four steps: choose a semantic network, match words from text with the elements from the semantic network, expand the text representation by adding or replacing semantically related elements, and then classify documents using the ex-

[3] http://scholar.google.com/scholar?q="(text|document)-(categorization|classification|clustering)"+"umls|wordnet|cyc|opencyc|framenet|sumo"&as_ylo=1990&as_yhi=2012

[4] http://scholar.google.com/scholar?q="(text|document)-(categorization|classification|clustering)"+"multiple-inheritance"+"umls|wordnet|cyc|opencyc|framenet|sumo"&as_ylo=1990&as_yhi=2012

[5] http://scholar.google.com/scholar?q="multiple-inheritance"+"umls|wordnet|cyc|opencyc|framenet|sumo"&as_ylo=1990&as_yhi=2012

Publication	Task	Semantic Network	Data set	Algorithm
Scott and Matwin [20]	Categorization	WordNet	Reuters	Ripper
—	—	WordNet	USENET	Ripper
—	—	WordNet	Digital Tradition	Ripper
Hotho et al. [11]	Clustering	WordNet	Reuters	K-center
Sedding and Kazakov [23]	Clustering	WordNet	Reuters	K-center
Mavroeidis et al. [17]	Categorization	WordNet	Reuters	SVM
—	—	WordNet	Amazon	SVM
Yoo and Hu [26]	Clustering	MeSH	PubMed	K-center
—	—	MeSH	PubMed	Hierarchical
—	—	MeSH	PubMed	Suffix Trees
Gabrilovich and Markovitch [9]	Categorization	Wikipedia	Reuters	SVM
—	—	Wikipedia	OHSUMED	SVM
—	—	Wikipedia	20 Newsgroups	SVM
—	—	Wikipedia	Movies Reviews	SVM
Bloehdorn and Hotho [2]	Categorization	WordNet	Reuters	AdaBoost
—	—	WordNet	OHSUMED	AdaBoost
—	—	MeSH	OHSUMED	AdaBoost
—	—	AgroVoc	FaoDoc	AdaBoost

Table 1. Text classification with semantic networks started in 1998 and continues today. Only the early work is summarized in this table. Multiple sources were used to create semantic networks, but none of the papers addressed the multiple inheritance problem.

panded representations (see Table 1). Using this scheme various research groups made important observations.

Scott and Matwin [20] showed that more general terms give better categorization than the less general terms. The optimal level of generalization, however, was different for each data set. Hotho et al. [11] showed the same for clustering. In addition, they concluded that it is better to keep terms from the lower levels of hierarchy rather than just replace them with terms from higher levels. Sedding and Kazakov [23] replicated the results of Hotho et al. [11], and then focused on mapping terms from WordNet. They found that ambiguities present in WordNet might render it useless when adding hypernyms. They concluded that part of speech tagging is insufficient for disambiguation of word senses. Taking the most frequent meaning, as Hotho et al. [11] did, is helpful but clearly not sufficient. Mavroeidis et al. [17] improved mapping text to the WordNet by using the Steiner tree cost. In addition, they studied the effect of sample size on levels of generalization. First, they found that adding hypernyms works better for small data, but the depth of generalization does not show any regularity. Second, they discovered that as the sample size is increased, the behavior of different depths stabilizes and converges, but at the cost of decreased improvements.

WordNet was not the only source of adding semantics. Yoo and Hu [26] used Medical Subject Headings (MeSH) for representing the text and improving clustering. Gabrilovich and Markovitch [9] used Wikipedia as a semantic network: Wikipedia's articles became concepts and links from the articles to most similar web pages became associative relations. Bloehdorn and Hotho [2] used WordNet, MeSH, and the United Nations Food and Agriculture multilingual agriculture thesaurus (AgroVoc) with marginal text categorization improvements.

Work in the early years of text classification with semantic networks lacked a mechanism vital to the healthy brain: inhibition. Google Scholar cites around 46,000 publications on inhibition in

brains[6]. Brain studies show that inhibition is crucial for normal functioning of associative memory [14], and too low inhibition may lead to epilepsy, schizophrenia and a "formal thought disorder" [16]. Surprisingly, the neurofunctional and neuroanatomical "lack of inhibition" has a long-lost brother in the field of artificial intelligence: the multiple inheritance problem in non-monotonic reasoning [3, p. 206]. A machine retrieves all related nodes from a semantic network with the same conviction as a patient with a formal thought disorder. Inheritance along all edges cannot be allowed because not every fact about the world is true or relevant in a given context. General solutions like "default logic", "circumscription", or "truth maintenance systems" require inference with negations, rules for overriding default values, closed-world assumption, or infinite computing power [19]. These requirements make them unsuitable for the large semantic networks that are currently available for automated text processing.

Evidently finding the ideal solution to the multiple inheritance problem is going to be quite difficult. In this paper an algorithm is proposed that removes just enough inference paths to significantly improve text classification performance. We contrast our approach with a scenario where no pruning of semantic network is performed, and problems due to the multiple inheritance cancel advantages of added semantics, making text classification improvements statistically not significant.

3 Databases and Document Collection

OHSUMED. The OHSUMED document collection, named after the Oregon Health and Science University School of Medicine, was created to benchmark information retrieval algorithms. It contains 348,566 PubMed papers published between 1987-1991 in 270 medical journals [10]. All papers have titles but only 233,445 have abstracts of an average length of 167 words. The papers have been manually indexed with 14,626 distinct Medical Subject Headings (MeSH). There are on average 253 papers per one MeSH. The inter-indexer consistency measured using 760 papers was between 61%-75% [8]. The challenge is to create an automated system that will do the indexing with competency comparable to human experts.

Researchers have created many such systems [25, 22, 13]. It is rare that someone would use all the data to develop and benchmark an algorithm but there is no consensus on how to split the data. One might say that the "Heart Diseases" (HD) subset is a common one. It has 12,417 training instances (years 1987-1990), 3,630 testing instances (1991 year), and 119 MeSH codes. The multiple inheritance problem is very complex so for clarity we have reduced the data set down to just ten MeSH codes. 4 of them, "endocarditis, bacterial", "aortic valve stenosis", "heart neoplasms", and "mitral valve stenosis" are used to develop the edge/node pruning algorithm and 5 of then, "mitral valve insufficiency", "atrial fibrillation", "aortic valve insufficiency", "cardiomyopathy, hypertrophic", and "heart arrest" are used for final benchmarking.

The UMLS Metathesaurus. The Unified Medical Language System (UMLS)[7] is a set of tools, websites and databases created and maintained by the National Library of Medicine, a division of U.S. National Institutes of Health. The UMLS has two main components: implementation resources (software) and knowledge sources (databases). We are interested just in one knowledge source - Metathesaurus - and one implementation resource - MetaMap. In particular, we used the 2009AB version of the Metathesaurus as a source of medical semantic data and the 2009 version of MetaMap Transfer (MMTx)[8] to map PubMed abstracts and titles to UMLS Metathesaurus medical

[6] http://scholar.google.com/scholar?q="inhibition"+"brain"+"fmri|erp"&as_ylo=1990&as_yhi=2012

[7] http://www.nlm.nih.gov/research/umls/

[8] http://metamap.nlm.nih.gov/

concepts. After parsing the HD data set, MetaMap discovered 21,127 unique concepts out of the 2,181,062 available in the Metathesaurus. Every concept had to be part of one root branch of the semantic network: "clinical finding", "body structure", "substance", "procedure", or "pharmaceutical", otherwise the concept was discarded.

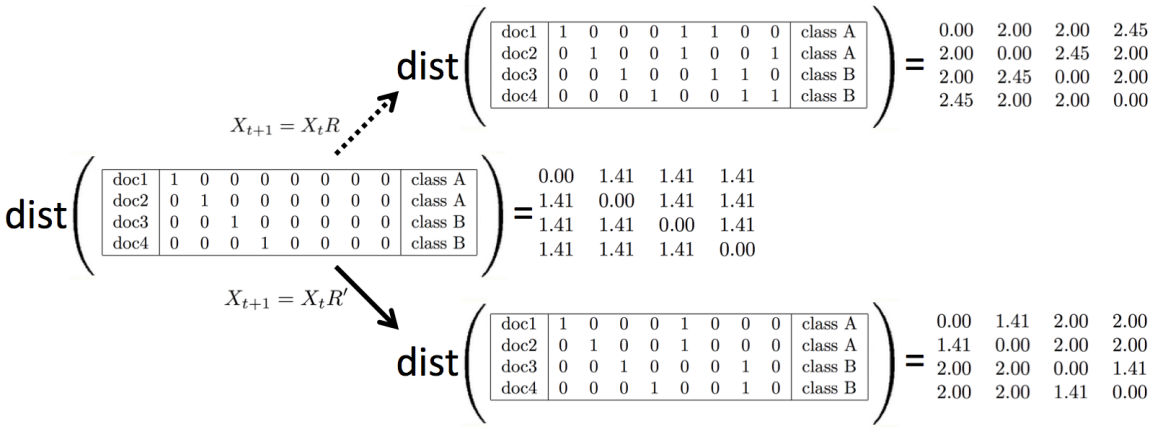

Fig. 1. Example of spreading activation matrices using semantic network with and without a solution to the multiple inheritance problem. The top right matrices show the features space and distances after entire semantic network has been applied (Figure 2 with all nodes and edges). The bottom right matrices show the feature space and distances after solving the multiple inheritance problem (Figure 2 without the dotted nodes and edges). Documents cluster according to the class labels only if the pruned semantic network is used.

The 2009AB version of the UMLS Metathesaurus is a conglomerate of 101 individual biomedical semantic networks, also called "source vocabularies". Each sub-network has its own set of concepts and relations; when these are combined, it contains 26,762,104 relations. We followed the 21,127 concepts present in the HD data set along the following edge types: "other related" (RO), "related and possibly synonymous" (RQ), "similar or like relationship' (RL), "children" (CHD), "parent" (PAR), "broader" (RB), "narrower" (RN), and "source asserted synonymy" (SY). After 14 steps of spreading activation, we reached all 2,131,301 semantically related concepts using 11,250,022 distinct connections[9]. As a solution to the multiple inheritance problem, we proposed an algorithm that reduces the 11,250,022 connections to a bare minimum that improves automated indexing of PubMed citations.

4 Problem Identification and Methods of Solution

The Multiple Inheritance Problem. Let's start with an illustrative text categorization problem. There are 4 documents, each containing just one of the following medical terms: "aortic valve insufficiency", "aortic valve stenosis", "mitral valve insufficiency", "mitral valve stenosis". Let's say that the first two documents belong to the class "A" and the other two to the class "B". The vector

[9] CHD = 2,137,767; PAR = 2,137,767; RB = 1,087,501; RL = 34,066; RN = 1,087,501; RO = 5,304,808; RQ = 287,280; SY = 49,846.

space representation would look like the first matrix on the left in Figure 1. There would be equal distances between all documents, offering no learning generalization. Let's assume now that the four terms come from a semantic network like the one in Figure 2. As with every medical dictionary, a disease can be categorized by a location or by a pathophysiology. That is the case in our "small world" example: each disease inherits two concepts. Even though the inheritance by location and by pathophysiology is always true, it is not always relevant to the categorization task at hand. If unpruned network is used ($X_{t+1} = X_t R$) the distances calculated for enhanced representation do not lead to good clusters (upper right matrix in Figure 1), and will lower the chances of correct classification.

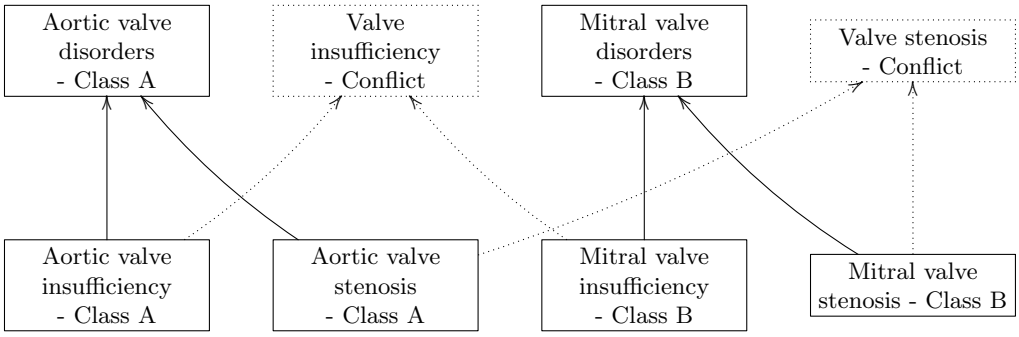

Fig. 2. Semantic network with an imposed document/term classification task. First, **relative frequency** is used to assign class to a node (see Figure 1). Second, edges that connect nodes belonging to different classes are identified. Third, conflicting nodes and edges (marked with dotted lines) are removed. This procedure prunes the semantic network, solving the multiple inheritance problem in text classification tasks.

On the other hand, if the **relative frequency** of a medical term in a class is used to denote its **belongingness**, we would find that certain edges connect medical terms from opposite classes. When that happens, documents from opposite classes become more similar and less distinguishable. In our example, connections to "valve insufficiency" and "valve stenosis" come from opposite classes. This situation can be repaired by removing at least one edge connecting "valve insufficiency" and at least one edge connecting "valve stenosis". Removing both edges will allow for removing also the nodes "valve insufficiency" and "valve stenosis". This leads to a reduced semantic network shown in Figure 2, without the elements marked with dotted lines. Spread activation ($X_{t+1} = X_t R'$) in the pruned network leads to a representation of documents from the same class grouped tighter than documents from opposite classes (lower right matrix Figure 1). Is there a way to do this programmatically on a larger scale?

Edge pruning. The conflict and non-conflict edge types shown in Figure 2, can be generalized for any binary classification problem. Let's call the positive class "**this**" and the negative class "**other**". The **relative frequency** will be used to assign class to a node. If we include spreading activation, it will give us four types of nodes: an **old** node belonging to the class "**this**", an **old** node belonging to the class "**other**", a **new** node belonging to the class "**this**", and a **new** node belonging to the class "**other**". If we exclude feedback loops, we will have eight edge types that connect the four node types, as shown in Figure 3. The learning process in a given context (here text categorization) should empirically determine which semantic associations should be inhibited.

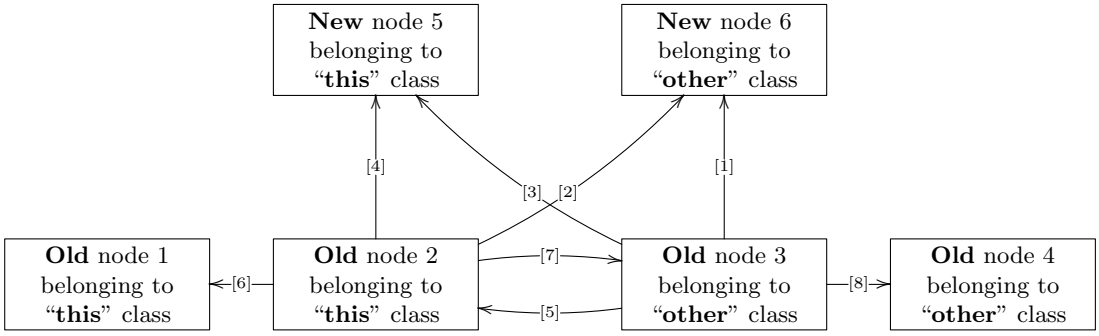

Fig. 3. Generalized semantic network with an imposed document/term classification task. Just by using **relative frequencies** we identify four types of nodes that receive activation and eight types of edges that carry the activation. We can empirically check which edge-pruning procedure improves the text classification task and use it as a heuristic solution to the multiple inheritance problem. Edges are enumerated in the order of their performances in Table 3.

Therefore in the training phase a check is made to see if any improvement will result from removal of selected edge types. This will indirectly show whether the distances between vectors representing documents change in favor of or against the two-class separation. This is not an ideal solution, where all unnecessary nodes are removed, but it offers sufficient separation. Such pruning method allows for unattended spreading of concept activations.

What is meant here by "unattended spreading activations"? Let's say that f is a Heaviside step function. Our goal of pruning the semantic network is to get a map, $X_{t+1} = f(X_t R)$, that can be applied to the data iteratively until the document/term matrix stops changing $X_{n+1} = X_n$ for $n \gg 0$. This means that the pruning process has to be iterative. Let's say that P^i is a function that removes one type of edge, then $R' = P^i(X_{t+1}, X_t, R)$. If $R' \neq R$, then some edges were removed, and we need to reset the training matrix $X_t := X_0$ to its initial state. We keep applying $P^i(X_{t+1}, X_t, R)$ and resetting X_t until $R' = R$ for all t. Once the pruning procedure P^i is completed, the next procedure, P^j, is done, and so on. Which pruning procedures improve text categorization and the order in which pruning is applied should be empirically determined.

Text Categorization. After final representation is generated classification is done using the support vector machines (SVM) with cosine kernel. Cosine kernel SVM is equivalent to a linear SVM if all vectors are normalized to the unit length. Normalization converts scalar products to cosine measures. This way SVM becomes insensitive to very short or very long documents. Linear and cosine kernel SVMs have only one parameter that has to be tuned by the user: "cost" of regularization. This parameter has been optimized checking results for its values from 2^{-1} to 2^5 with $2^{0.25}$ increments. All features are binary: the text either mentions the concept or it does not. The best models were selected using a 16-fold cross-validation, with classification quality measured by the F_1 score, a harmonic mean between precision and recall. Once the best model with the best pruning procedure was selected, it has been used for the final testing. All spreading activation models were compared to a model without the feature space enrichment and tested for significance.

Final testing. The statistical significance of the F_1 score improvement is measured using a paired t-test [4]. For each classification label we have a total of 17 F_1 scores for the baseline model and the same number for the enhanced model, resulting in 17 pairwise comparisons. We used "endocarditis, bacterial", "aortic valve stenosis", "heart neoplasms", and "mitral valve stenosis" to find the best

REL Type	F_1^{CV} (δ^{CV})	**Edges (Nodes)**
RL + RB + SY	0.6286 (-0.1212)	129,003 (41,972)
RB + SY + PAR	0.6477 (-0.1021)	225,796 (51,672)
RL + RB	0.6523 (-0.0976)	116,592 (38,544)
RL	0.7064 (-0.0435)	35,231 (23,897)
RL + SY	0.7091 (-0.0408)	41,868 (25,870)
SY + PAR	0.7103 (-0.0395)	167,126 (43,354)
RB + PAR	0.7356 (-0.0143)	207,727 (47,888)
RB + SY	0.7488 (-0.0010)	78,939 (31,266)
RB	0.7498 (-0.0001)	68,270 (28,294)
SY	0.7509 (+0.0010)	24,437 (20,459)
PAR	**0.7758 (+0.0252)**	152,559 (40,134)

Table 2. Performance of various UMLS relation types after 60 steps of spreading activation without the edge-pruning technique. This table shows the SVM macro F_1 using 16-fold cross-validation, improvement when compared to a model with no spreading activation (δ), and the number of unique edges and unique nodes used during the 60 steps of activation. The best-performing relationship is PAR (parent), and it has been used for network pruning experiments.

pruning procedure. Then we added "mitral valve insufficiency", "atrial fibrillation", "aortic valve insufficiency", "cardiomyopathy, hypertrophic", "cardiomyopathy, congestive" and "heart arrest" to see if the improvement generalized over different labels, a total of ten labels. Thus, the t-test across all experiments has 170 pairwise comparisons. The Pearson correlation coefficient is used to see if the data is improved by the same factor across different classes. If the baseline model is correlated with the enhanced model there is a stable improvement.

Node A →	**Node B**	F_1^{CV} (δ^{CV})	**Edges (Nodes)**
Old Other →	Old Other [8]	0.7485 (-0.0014)	18,786 (18,782)
Old This →	Old Other [7]	0.7487 (-0.0012)	18,942 (18,844)
Old This →	Old This [6]	0.7535 (+0.0037)	19,375 (18,926)
Old Other →	Old This [5]	0.7593 (+0.0095)	19,618 (18,879)
Old This →	**New This** [4]	**0.7736 (+0.0237)**	147,857 (39,269)
Old Other →	**New This** [3]	**0.7740 (+0.0242)**	117,064 (32,946)
Old This →	**New Other** [2]	**0.7746 (+0.0247)**	139,621 (38,039)
Old Other →	**New Other** [1]	**0.7757 (+0.0258)**	63,957 (24,436)

Table 3. Performance of PAR relationship after 60 steps of spreading activation with eight types of edge removal procedures. Edge types are defined in Figure 3. This table shows the SVM macro F_1 using 16-fold cross-validation, improvement when compared to a model with no spreading activation (δ), and the number of unique edges and unique nodes used during the 60 steps of activation. The best four edge types were chosen for permutation experiments.

Concept space visualization. The changes to the semantic network rely on assigning medical concepts to a class based on the **relative frequency** measure. We use the class belongingness to identify edges connecting nodes from different classes. We can visualize the process. Each medical

concept is represented by two **relative frequencies**[10]: $\mathrm{rf}^3_{\mathrm{this}}$ and $\mathrm{rf}^3_{\mathrm{other}}$. If the semantic network is separating two classes well, we should see concepts travel to the top-left corner and the bottom-right corner ($\mathrm{rf}^3_{\mathrm{this}} \gg \mathrm{rf}^3_{\mathrm{other}}$ or $\mathrm{rf}^3_{\mathrm{this}} \ll \mathrm{rf}^3_{\mathrm{other}}$). If, on the other hand, the semantic network does not separate classes, then most concepts will have similar **relative frequencies** ($\mathrm{rf}^3_{\mathrm{this}} \approx \mathrm{rf}^3_{\mathrm{other}}$) and will lie along the $x = y$ line.

Node A	**Node B**	F_1^{CV} (δ^{CV})	**Edges (Nodes)**
(1) **Old Other**	**New This** [2]		
(2) **Old Other**	**New Other**[1]		
(3) **Old This**	**New Other**[3]		
(4) **Old This**	**New This** [4]	**0.7857 (+0.0358)**	51,405 (21,684)

Table 4. The best sequence of edge removal calculated using macro F_1 on classes "endocarditis, bacterial", "aortic valve stenosis", "heart neoplasms", and "mitral valve stenosis". Four out of eight removal procedures from Table 3 were permuted and the best sequence was chosen for final testing. Removing all edges that connect medical concepts that did not appear in any of the training documents worked best (edge types numbered 1-4 in Figure 3).

Relative frequency snapshots might not be enough to see a divergent or convergent trend, but if we follow the centers of the relative frequencies and connect them with arrows, the trend becomes apparent. If the arrows point outward, then the trend confirms separation by spreading activation. If the arrows are parallel to the $x = y$ line, then there is no separation trend, and spreading activation causes more harm than good.

5 Results

Spreading activation without the edge-pruning technique. First experiments had to determine which semantic relationships and their combinations yield the best results. Table 2 shows that only parent relationships ("is-a") improve classification performance. Other relationships or combinations tried did not improve the results. This finding is consistent with the work already published (Table 1). The vector feature space increases in size from 21,127 concepts to 40,134 concepts. Surprisingly, it takes almost 60 iteration steps before the feature space stabilized (lower right graph in Figure 4). Sixty multiplications of such huge matrix, even in a sparse format, is computationally demanding. It would be impractical for the full OHSUMED data set and impossible for the full PubMed database. If we look carefully at the relative frequency pathways, we notice a peculiar behavior where two classes initially diverge but then collapse (lower right graph in Figure 4). This is also congruent with others' work, where they would find a step of iteration with the largest separation, for example step 9, and use that for testing, sometimes without much success [20, 11, 23, 17]. Other authors also reported that each class requires a different number of iterations, so this would not be a good source of generalization.

Figure 4 and Table 5 support evidence that 152,559 parent-child relations are enough to cause very complex behavior. There is some improvement in performance but not statistically significant (p-value=0.01259 at best, p-value=0.02618 overall). The improvement is almost random because

[10] The power = 3 greatly enhances signal for concepts with $\mathrm{rf}_{class} \approx 1$.

Class name (size)	F_1^{CV} (δ^{CV})	F_1^{TEST} (δ^{TEST})	p-value
aortic valve insufficiency (239)	0.6026 (-0.0267)	0.6226 (-0.0697)	0.91325
aortic valve stenosis* (341)	0.7725 (+0.0096)	0.6621 (-0.0281)	0.25232
atrial fibrillation (222)	0.6511 (+0.0463)	0.6713 (+0.0559)	0.05404
cardiomyopathy, congestive (253)	0.6009 (+0.0171)	0.6092 (-0.0211)	0.23898
cardiomyopathy, hypertrophic (192)	0.7799 (+0.0507)	0.7640 (+0.0140)	0.01259
endocarditis, bacterial* (310)	0.8242 (+0.0182)	0.7211 (+0.0017)	0.19099
heart arrest (405)	0.6952 (-0.0071)	0.6966 (-0.0234)	0.68066
heart neoplasms* (197)	0.7729 (+0.0275)	0.6512 (+0.1032)	0.17259
mitral valve insufficiency (295)	0.6007 (-0.0338)	0.6087 (+0.0259)	0.86898
mitral valve stenosis* (172)	0.7335 (+0.0485)	0.7627 (+0.0448)	0.07377
across all experiments	0.7034 (+0.0150)	0.6770 (+0.0103)	0.02618

Table 5. Final results using PAR relations without edge-pruning procedures. PAR relations without pruning offer poor improvement of the F_1 score on the cross-validation and the test sets. None of the improvements offers statistical significance when a paired t-test was used to compare models with and without the semantic enhancement. *Data used for finding the best types of semantic relationships and the best pruning procedures.

it does not correlate well with the baseline model. The Pearson correlation coefficient between the baseline model and the enhanced model over all 170 runs of SVM is 0.66; it ranges between 0.29 and 0.81 depending on the class label. In summary, this means that 152,559 relations react differently to different classes, need more computational time and are not a reliable source of background knowledge.

Spreading activation with edge-pruning technique. This is uncharted territory. At the start 152,559 parent-child relations are included. Spreading activation and removing one edge type at a time requires restart of the process each time there is a change to the semantic network. After 60 iterations the algorithm stops. This means that $X_{t+1} = f(X_t R')$ and rf_{class} must be calculated on average between 73 and 1,082 times, depending on the edge type from Table 3. After that the four best-performing edge types are used and the order in which they are being applied to the semantic network is permuted. The best sequence of pruning (edge type 2, then 1, then 3, and then 4, see Table 4) requires on average 451 $X_{t+1} = f(X_t R')$ and rf_{class} operations, but reduces the initial 152,559 edges to a more modest 51,405, cutting the number of active concepts by half.

Figure 4 and Table 6 offer evidence that 51,405 parent-child relations create a predictable behavior. Spreading activation stabilizes around the 30th iteration. It has slightly better separation around ten iterations. After that, the relative frequency centers move back (upper right graph in Figure 4), but not nearly as much as in the case of spreading activation without edge pruning. The improvement is statistically significant in the case of four out of ten labels (best p-value 0.00088), three of which were not used during the best pruning sequence-seeking process. The improvement across all 170 subsets is statistically significant (p-value=0.00003). The Pearson correlation coefficient between the baseline model and the enhanced model over all 170 runs of SVM is 0.72 and ranges between 0.17 and 0.87, depending on the class label. In summary, the 51,405 parent-child relations offer good performance improvement, need less computational resources, and are a good source of background knowledge.

Class name (size)	F_1^{CV} (δ^{CV})	F_1^{TEST} (δ^{TEST})	p-value
aortic valve insufficiency (239)	0.5924 (-0.0370)	0.6733 (-0.0190)	0.99374
aortic valve stenosis* (341)	0.7787 (+0.0158)	0.6853 (-0.0048)	0.15238
atrial fibrillation (222)	0.6731 (+0.0683)	0.7172 (+0.1019)	0.00088**
cardiomyopathy, congestive (253)	0.6363 (+0.0526)	0.6590 (+0.0287)	0.00240**
cardiomyopathy, hypertrophic (192)	0.7879 (+0.0587)	0.8235 (+0.0735)	0.00334**
endocarditis, bacterial* (310)	0.8149 (+0.0089)	0.7273 (+0.0078)	0.34398
heart arrest (405)	0.6886 (-0.0136)	0.6667 (-0.0533)	0.79991
heart neoplasms* (197)	0.8019 (+0.0565)	0.7229 (+0.1749)	0.01222
mitral valve insufficiency (295)	0.6262 (-0.0083)	0.6452 (+0.0624)	0.56264
mitral valve stenosis* (172)	0.7471 (+0.0621)	0.8062 (+0.0883)	0.00429**
across all experiments	0.7147 (+0.0264)	0.7127 (+0.0460)	0.00003

Table 6. Final results using PAR relations with the best edge-pruning procedure. Edge pruning offers good improvement of the F_1 score on the cross-validation and the test sets. Four out of ten data sets achieved statistically significant improvement when a paired t-test was used to compare models with and without the semantic enhancement. *Data used for finding the best types of semantic relationships and the best pruning procedure. **Data with statistically significant categorization improvement.

6 Conclusion and Discussion

Language competence at the human level may require detailed neurocognitive models that combine several kinds of memory: recognition, semantic, episodic and short-term working memory, in addition to the iconic spatial and other types of imagery that goes beyond representation based on verbal concepts. Such systems, requiring embodied cognition, are not practical at present. It is therefore worthwhile to identify and solve specific problems that pose a challenge to the current NLP approaches.

Semantic network stores default and commonsense knowledge. Multiple inheritance problem can be solved by adding inhibition to network links. The network is pruned to adjust it to the current knowledge, avoiding confusion and contradictions. The algorithm presented in this paper identified PAR relations as the only one that lead to significant improvements. Although the number of medical concepts in our experiments has been limited the role of inhibition of some associations has been clearly demonstrated. Understanding practical applications of inhibition in the design of semantic memory shows the way to applications of the same techniques to other types of memories implemented by other types of networks.

Experiments with classification of medical collections of texts show that adding inhibition indeed in many cases leads to significant improvements of results. This is merely one way of pruning semantic networks. Insights from granular information processing imply that a dynamic balancing of semantic generality and specificity could be a useful approach for subsequent refinements of the proposed method.

Acknowledgments. Authors would like to thank Drs. John P. Pestian, Imre Solti, Lawrence Hunter, K. Bretonnel Cohen, Karen M. Stannard, Guergana K. Savova, and Alan R. Aronson for their interest in this article.

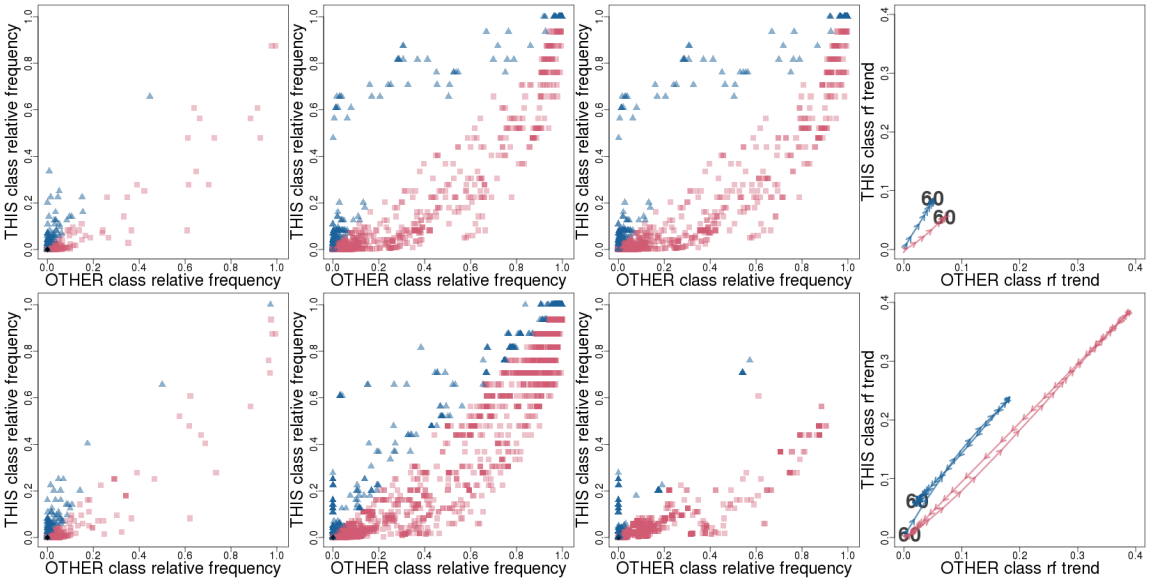

Fig. 4. Relative frequencies of the UMLS Metathesaurus concepts as they change with spreading activation steps. The X-axis has relative frequencies corresponding to the class "other" and the Y-axis has relative frequencies corresponding to the class "heart neoplasms". First three images show spreading activation on 1, 10 and 60 steps of the 1991 citation year data set. The top images show spreading activation using PAR relations that were pruned using the best edge-pruning procedure from Table 4. The bottom images show spreading activation using PAR relations that were not pruned in any way. The two images on the right show the 60-step pathway of the relative frequency centers as they move outward or inward and then settle down and stabilize around the 30th iteration with the pruning and around the 50th iteration without the pruning.

References

[1] Billingsley, R.L., McAndrews, M.P., Crawley, A.P., Mikulis, D.J.: Functional MRI of phonological and semantic processing in temporal lobe epilepsy. Brain 124(6), 1218–1227 (2001)

[2] Bloehdorn, S., Hotho, A.: Boosting for text classification with semantic features. In: Advances in Web Mining and Web Usage Analysis. vol. 3932, pp. 149–166 (2006)

[3] Crevier, D.: AI: the tumultuous history of the search for artificial intelligence. Basic Books, New York, NY (1993)

[4] Demsar, J.: Statistical comparisons of classifiers over multiple data sets. The Journal of Machine Learning Research 7, 1–30 (2006)

[5] Duch, W., Matykiewicz, P., Pestian, J.: Neurolinguistic approach to natural language processing with applications to medical text analysis. Neural Networks 21(10), 1500–1510 (2008)

[6] Duffau, H., Gatignol, P., Mandonnet, E., Peruzzi, P., Tzourio-Mazoyer, N., Capelle, L.: New insights into the anatomo-functional connectivity of the semantic system: a study using cortico-subcortical electrostimulations. Brain 128(4), 797–810 (2005)

[7] Fellbaum, C.: WordNet. Wiley Online Library (1999)

[8] Funk, M.E., Reid, C.A.: Indexing consistency in MEDLINE. Bulletin of the Medical Library Association 71(2), 176–183 (1983)

[9] Gabrilovich, E., Markovitch, S.: Overcoming the brittleness bottleneck using wikipedia: Enhancing text categorization with encyclopedic knowledge. In: Proceedings of the 21st National Conference on Artificial Intelligence. pp. 1301–1306 (2006)
[10] Hersh, W., Hickam, D.: Use of a multi-application computer workstation in a clinical setting. Bulletin of the Medical Library Association 82(4), 382–389 (1994)
[11] Hotho, A., Staab, S., Stumme, G.: Wordnet improves text document clustering. In: Proc. of the Semantic Web Workshop, 26th Annual International ACM SIGIR Conf. pp. 541–544 (2003)
[12] Joachims, T.: A probabilistic analysis of the rocchio algorithm with tfidf for text categorization. In: Proceedings of the Fourteenth International Conference on Machine Learning. pp. 143–151. ICML '97, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1997)
[13] Joachims, T.: Text categorization with support vector machines: Learning with many relevant features. In: Machine Learning: ECML-98, pp. 137–142. Lecture Notes in Computer Science, Springer Berlin Heidelberg (1998)
[14] Khader, P., Knoth, K., Burke, M., Ranganath, C., Bien, S., Rosler, F.: Topography and dynamics of associative long-term memory retrieval in humans. Journal of Cognitive Neuroscience 19(3), 493–512 (2007)
[15] Lamb, S.M.: Pathways of the Brain: The Neurocognitive Basis of Language. John Benjamins Publishing Company (1999)
[16] Leeson, V.C., Simpson, A., McKenna, P.J., Laws, K.R.: Executive inhibition and semantic association in schizophrenia. Schizophrenia Research 74(1), 61–67 (2005)
[17] Mavroeidis, D., Tsatsaronis, G., Vazirgiannis, M.: Word sense disambiguation for exploiting hierarchical thesauri in text classification. In: Knowledge Discovery in Databases: PKDD 2005. Lecture Notes in Computer Science, vol. Volume 3721/2005, pp. 181–192. Springer (2005)
[18] Pedersen, T., Patwardhan, S., Michelizzi, J.: Wordnet::similarity: Measuring the relatedness of concepts. In: Demonstration Papers at HLT-NAACL 2004. pp. 38–41. ACL, Stroudsburg, PA, USA (2004)
[19] Russell, S.J., Norvig, P., Davis, E.: Artificial intelligence: a modern approach. Prentice Hall, Upper Saddle River, NJ (2010)
[20] Scott, S., Matwin, S.: Text classification using WordNet hypernyms. In: Use of WordNet in Natural Language Processing Systems: Proceedings of the Conference. pp. 38–44. ACL, Somerset, New Jersey (1998)
[21] Scott, S., Matwin, S.: Feature engineering for text classification. ICML 99, 379–388 (1999)
[22] Sebastiani, F.: Machine learning in automated text categorization. ACM Comput. Surv. 34(1), 1–47 (2002)
[23] Sedding, J., Kazakov, D.: WordNet-based text document clustering. In: COLING 2004 3rd Workshop on Robust Methods in Analysis of Natural Language Data. pp. 104–113. COLING, Geneva, Switzerland (2004)
[24] Tivarus, M.E., Ibinson, J.W., Hillier, A., Schmalbrock, P., Beversdorf, D.Q.: An fMRI study of semantic priming: modulation of brain activity by varying semantic distances. Cognitive and behavioral neurology: official journal of the Society for Behavioral and Cognitive Neurology 19(4), 194–201 (2006)
[25] Yang, Y., Pedersen, J.: A comparative study on feature selection in text categorization. In: Fisher, D. (ed.) Proceedings of the Fourteenth International Conference on Machine Learning. pp. 412–420. Morgan Kaufmann Publishers (1997)
[26] Yoo, I., Hu, X.: A comprehensive comparison study of document clustering for a biomedical digital library MEDLINE. In: International Conference on Digital Libraries Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries. pp. 220–229. IEEE Press (2006)