

# Unambiguous Concept Mapping in Medical Domain

Pawel Matykiewicz, Wlodzislaw Duch, John Pesian



## Introduction

Neurocognitive approach to natural language processing (NLP) is inspired by the putative brain processes used by the brain to remember and understand concepts. To approximate formation of primed [1] semantic subnetwork that facilitates interpretation of the text, **graphs of consistent concepts (GCCs)** are constructed. The concept mapping algorithm presented here is based on this approach, although many other variants and applications are possible [2, 3, 4]. Concepts and relations from the Unified Medical Language System (UMLS), a large collection of medical ontologies, form the basis for semantic network construction. The algorithm for phrase sense disambiguation adds new relations and determines relations strength between concepts with the use of prior knowledge and the acquisition of new knowledge. Analysis of texts, independent of the purpose, requires three main steps:

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

These three steps roughly correspond to the function of three kinds of human memory [5]: recognition memory, semantic memory and episodic memory. NLP research usually ignores this fact, focusing on formal approaches (grammar, logics, statistical correlations). The long-term goal is to reach human-level competence in natural language processing.

## Method

Following definitions are used:  $N(CUI_i)$  – number of occurrence of a  $CUI_i$ , or concept unique identifier, in the relational table,  $C(CUI_i, CUI_j)$  – number of co-occurrences of  $CUI_i$  and  $CUI_j$  concepts in the relational table row,  $W = \{w_{ij}\}$  – matrix storing weights between  $i$ th and  $j$ th concept. The weights are defined as conditional probabilities:

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

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

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

where  $\Theta$  is the step function and  $\alpha < 1$  is a spontaneous decay parameter. Similar function has been considered in [6]. The main problem for spreading activation in networks without inhibition is to prevent the infinite growth of all node activities.  $\alpha$  decays should be sufficiently large to achieve this; all experiments in this paper are with  $\alpha = 0.73$ .

## Example

The example below shows a radiological dictation from *Corpus II* which was labeled with a *ICD-9-CM* code number 518.0 ("OTHER DISEASES OF LUNG: PULMONARY COLLAPSE"):

CLINICAL HISTORY: 9-month-29-day-old male with a history of cough.  
Rule out pneumonia.

PROCEDURE COMMENTS: None.

COMPARISON: XX/XX/XX.

FINDINGS: There is mild hyperinflation of the lungs with increased peribronchial markings most consistent with viral versus reactive airway disease. Hazy increased density is seen in the right middle lobe, left lower lobe which could represent subsegmental atelectasis. Hazy increased density is also noted at the lingula with partial effacement of the left heart contour which could represent atelectasis versus early pneumonia. No pleural effusion is noted. The cardiothymic silhouette is within normal limits.

Soft tissues and bony structures are unchanged.

IMPRESSION: Findings most consistent with viral versus reactive airway disease. Patchy atelectasis is associated. Lingular early infiltrate cannot be excluded.

9-month-29-day-old male with a history of cough. rule out pneumonia. there is mild hyperinflation of the lung with increased peribronchial marking most consistent with viral versus reactive airway disease. hazy increased density is seen in the right middle lobe, left lower lobe which could represent subsegmental atelectasis. hazy increased density is also noted at the lingula with partial effacement of the left heart contour which can represent atelectasis versus early pneumonia. no pleural effusion is noted. the cardiothymic silhouette is within normal limits. soft tissue and bony structure be unchanged. finding most consistent with viral versus reactive airway disease. patchy atelectasis be associated. lingular early infiltrate can be excluded.

FIGURE 1: Example of a filtered, normalized and mapped text. Mapped words are light and dark green.



FIGURE 2: Example of a semantic graph after 4 steps of spreading activation showing nodes with activation above 0.2 value. No learning has yet been used, 6 out of 13 ambiguous mappings have correct maximum activation.

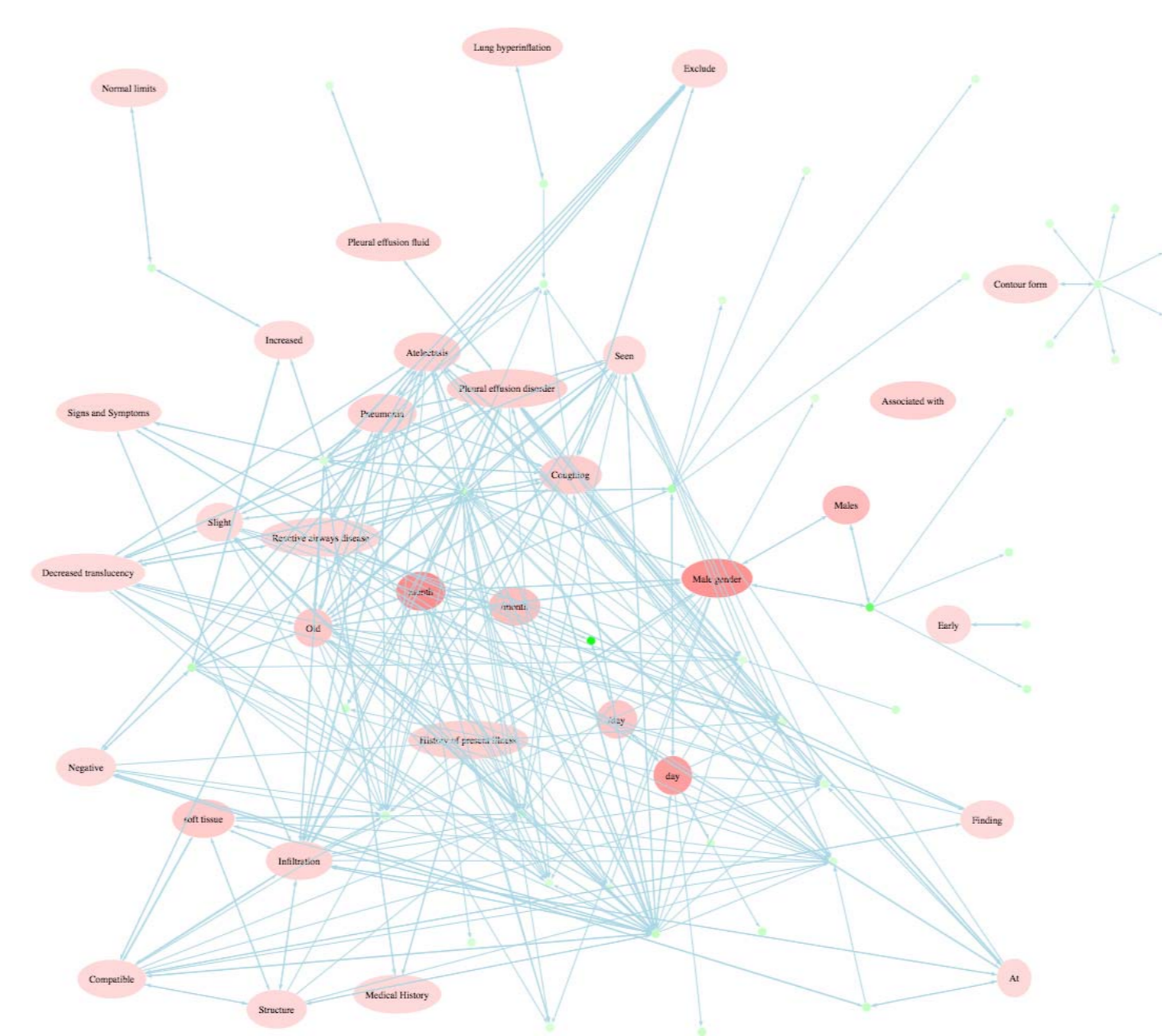


FIGURE 3: Example of a semantic graph after 4 steps of spreading activation showing nodes with activation above 0.2 value. *Corpus I* has been used as the learning set, 9 out of 13 ambiguous mappings have correct maximum activation.

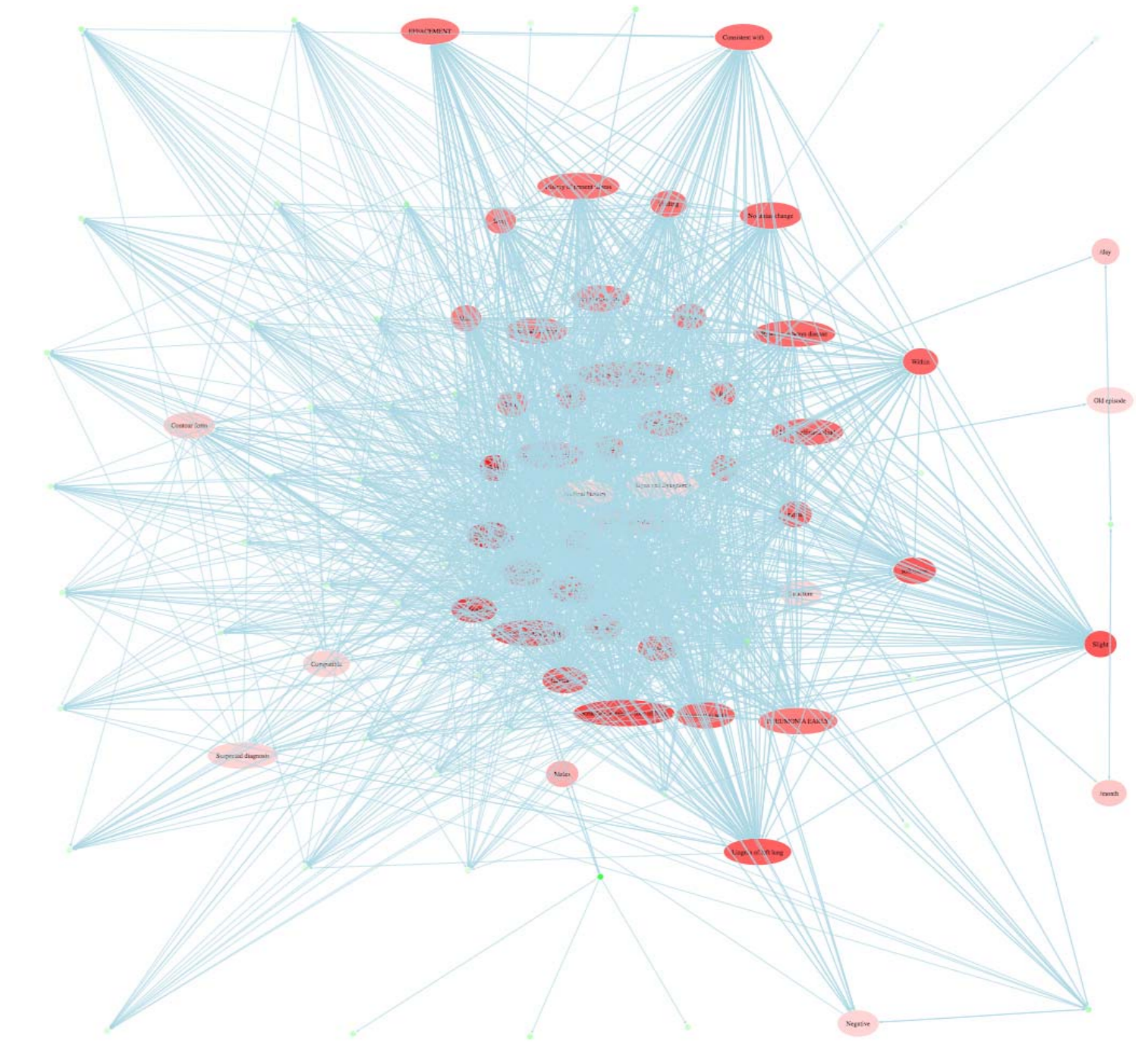


FIGURE 4: Example of a semantic graph after 4 steps of spreading activation showing nodes with activation above 0.2 value. *Corpus II* has been used as the learning set, 12 out of 13 ambiguous mappings have correct maximum activation.

For knowledge acquisition Cincinnati Children's Hospital Medical Center clinical texts from radiology have been used. These texts were dictated by physicians and changed into a text form by medical voice recognition software. 60 of these documents were chosen for manual annotation and divided in two parts. Every set of documents has 30 chest x-ray dictations for 6 different *ICD-9-CM* codes.

## Conclusion

Unfortunately even with the use of UMLS ontologies and relations analysis of clinical texts showed that many relations needed for correct annotations are still missing. A software tool for enriching UMLS relation by creating manually annotated texts and learning from them has been created. Experiments performed on two small corpuses showed significant influence of additional knowledge on the disambiguation performance of the GCC graphs.

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

	no training	training	
		<i>Corpus I</i>	<i>Corpus II</i>
<i>Corpus I</i>	79%	96%	86%
<i>Corpus II</i>	57%	64%	79%

FIGURE 5: Comparison of GCC disambiguation accuracies with and without additional training.

Neurocognitive approach to the NLP is very useful for nonambiguous medical concept mapping, but additional knowledge in form of concepts and their relations is needed to achieve human-level performance.

## References

- [1] T.P. Mcnamara, "Semantic Priming; Perspectives from Memory and Word Recognition (Essays in Cognitive Psychology)", Psychology Press, UK, 2005
- [2] J.P. Pesian, L. Itert, C. Andersen, W. Duch, "Preparing Clinical Text for Use in Biomedical Research." Journal of Database Management 17(2), 1-11, 2006.
- [3] J. Szymanski, T. Sarnatowicz, W. Duch, "Towards Avatars with Artificial Minds: Role of Semantic Memory". Journal of Ubiquitous Computing and Intelligence (in print).
- [4] W. Duch, J. Szymanski, T. Sarnatowicz, "Concept description vectors and the 20 question game". In: Intelligent Information Processing and Web Mining, Eds. M.A. Klopotek, S.T. Wierzchon, K. Trojanowski, Advances in Soft Computing, Springer Verlag, pp. 41-50, 2005.
- [5] J.R. Anderson, "Learning and Memory". J. Wiley and Sons, NY 1995.
- [6] C. Rocha, D. Schwabe, M. P. Aragao, A hybrid approach for searching in the semantic web. In: Proc. of the 13th Int. Conf on World Wide Web 2004, New York, NY, USA, pp. 374-383.