SEMANTIC NETWORK BASED MECHANISMS FOR KNOWLEDGE ACQUISITION

Dariusz Ceglarek

Department of Applied Informatics, Poznan School of Banking, Poznan, Poland
dariusz.ceglarek@wsb.poznan.pl

ABSTRACT

This article summarizes research work started with the SeiPro2S (Semantically Enhanced Intellectual Property Protection System) system designed to protect resources from the unauthorized use of intellectual property. The system implements semantic network as a structure of knowledge representation and a new idea of semantic compression. As the author proved that semantic compression is viable concept for English, he decided to focus on potential applications. An algorithm is presented that employing semantic network WiSENet for knowledge acquisition with flexible rules that yield high precision results. Developed algorithm is implemented as a Finite State Automaton with advanced methods for triggering desired actions. Detailed discussion is given with description of devised algorithm, usage examples and results of experiments.

KEYWORDS

Semantic network, Semantic compression, Knowledge acquisition, Lexical relationships, Natural Language Processing

1. INTRODUCTION

Natural language is a very complex system which needs to be represented in a way that would be understandable for computer systems. One need to possess some structures that can represent a part of semantic knowledge. Choosing proper knowledge representation structure is very important determinant of the classification quality of text documents [3][11][15]. Semantic knowledge, as identified lexical relations between concepts, should be stored in an appropriate data structure in order to be utilized to refine Information Retrieval (IR) or Natural Language Processing (NLP) tasks and their results. a semantic network is a structure incorporating knowledge about all possible lexical relations between words. Lexical relations reflect the interdependences between the concepts. Semantic networks store information about similarity relations (like a thesaurus): word similarity, synonymy, antonymy; hierarchical relations (like a taxonomy): hypernymy, troponymy (lexical relations existing only for verbs) or hyponymy and meronymy or holonymy relations. Semantic network can incorporate connotations as well these are any other word associations. Using the graph theory terminology, semantic networks can be represented as directed graphs. Direction is crucial in case of hierarchical relations. Edges between concepts can be weighted as well in order to reflect strength of a relation. Semantic networks are the most advanced structures representing semantic knowledge of natural language [26]. Choosing proper knowledge representation structure is very important determinant of the classification quality of text documents. That is why their utilization in information retrieval systems should bring the biggest improvement in their effectiveness.

Sundarapandian et al. (Eds) : CSE, CICS, DBDM, AIFL, SCOM - 2013
pp. 105–118, 2013. © CS & IT-CSCP 2013 DOI : 10.5121/csit.2013.3311

This paper shows that improving the efficiency of NLP methods by utilizing sophisticated models grooves on inclusion of mechanisms reflecting and using information of lexical relations between concepts. Text documents are most often a subject of information retrieval or information extraction. The complexity of human natural language, however, negatively affects the results of classic algorithms implemented in NLP/IR systems. Retrieval methods are getting accommodated to identify lexical relations between words or phrases (i.e. relations between the meanings of words or the concepts they represent) and use this knowledge to compare and match documents with user's needs more accurately. Such knowledge is represented in structures like thesauri or semantic networks. Thesauri or semantic networks can be created manually Unfortunately, it is a very time-consuming task and needs an involvement of expert knowledge. This paper presents a method developed to automatically extract mentioned knowledge from a corpus of documents.

The information included in semantic network can be used in order to limit the number of keywords to describe a document, expand user queries or identify concepts if a word represents more than one meanings. Its greatest advantage is by supplying a system with the right meaning of the concept processed based on its contextual usage [4]. Benefits one can obtain by applying semantic nets in classification tasks were described by [1]. Commonly used semantic network in NLP systems for processing English is *WordNet* [10][20]. Its structure is organized around notion of synsets. Every *WordNet*'s synset contains words which are mutually synonyms. Relationships among synsets are hypernyms or hyponyms, when combined with previous data it is easily seen that whole WordNet acts as a thesaurus. The details of the adoption and motivation of transferring WordNet to a new format *WiSeNet* is discussed in [7]. In this paper were also enumerated various aspects and possible merits of applying the *WiSENet* semantic network.

The aim of this work is to present an application of previously introduced semantic network *WiSENet* (semantic network *WordNet* transferred into *SenecaNet* format introduced in [5]). Since earlier publications, developed semantic network has grown taking in account number of concepts. This was necessary action, as most of advanced operations that can be carried with the *WiSENet* cannot function well without extensive concept vocabulary. The most important was the recognition of named entity (proper names, geographical names, names of organizations etc.) and further acquisition them to semantic network what is possible using e.g. *shallow text processing* methods [2].

Taking into account, that some of readers may not be familiar with specifics of *WiSENet* a brief summary of its origin and capabilities is given.

To begin with, the *WiSENet* semantic network derived its whole content from the *WordNet*. The decision was based on overall number of words and potential for further development and restructuring.

The most important fact is that, author had to dismantle a synset structure and turn it into a graph where nodes represent concepts and vertices denote lexical relation of hypernymy/hyponymy. This enabled devised algorithms to easily follow relations among particular concepts found in real life textual data (generally in unstructured text files). Restructuring was carried out in a lossless manner (the algorithm is given in [7]).

Additionally, the *WiSENet* proved useful in combination with frequency dictionaries developed for a number of various domains. These frequency dictionaries allow for highly efficient disambiguation of concepts stored in the *WiSENet*. To some point, frequency dictionary coupled with semantic network resembles human cognition when confronted with decisions concerning disambiguation. New structure aided by domain frequency dictionaries proved to work well, results of application of the *WiSENet* to semantic compression for English were highly satisfactory.

Semantic compression is a process throughout which reduction of dimension space (used for indexing) occurs. The reduction entails some information loss, but in general it aims not to degrade quality of results thus every possible improvement is considered in terms of overall impact on the quality. Dimensions' reduction is performed by introduction of descriptors for viable concepts. Descriptors are chosen to represent a set of synonyms or hyponyms in the processed passage. Decision is made taking into account relations among the concepts and their frequency in context domain.

2. MOTIVATING SCENARIO FOR KNOWLEDGE ACQUISITION

As mentioned earlier, it was observed that the *WiSENet* lacks a great number of concepts that are to be met in various textual data. Those most impeding experiments are originating from general culture. Vast majority of identified missing concepts are proper names of various entities. For sake of clarification, by proper names author understand names of people, organizations, geographical names and various objects (e.g. names of products). The *WordNet* in general does not miss most general categories of entities, yet a lot of highly specialised concepts is not present. As the *WordNet* was not devised for text processing tasks previous statement is offered not as a criticism but as an observation.

Stating the above author decided to invest effort in expanding the *WiSENet*. What is more important, this effort surpasses traditional methods of bulk import of all available resources and their later refactoring to match initial structure of to be extended semantic net.

It was observed that the *WiSENet* is very useful in discovering concepts that represent some specialization of other concepts by employing specially prepared rules.

The *WiSENet* semantic network can be applied to a set of procedures, that aim is to extract information from some textual data. As is well known in the domain of text processing, there should be manually prepare a set of rules that trigger when given order of elements is met. A great disadvantage to anyone who has to prepare this set of rules is that one is in need of specifying them in a manner that enumerates every plausible variant of a rule.

For preparing a set of rules that enabling to retrieve information from the data, one should begin with investigation of domain. Let's assume, that the whole process should supply its invoker with new data on people that hold managerial positions at various companies. First of all, one should issue an recognizance query to a search engine of his choice, probing for concepts than can denote a managerial position in some company.

It can be easily checked, that querying with search concepts such as: chairman, CEO, chief executive officer, managing director, manager; shall bring results similar to following ones from the Table 1.

Table 1. Examples of persons with managerial positions

William (Bill) H. Gates is chairman of Microsoft Corporation
Richard K. Matros has been Chairman and CEO of Sun Healthcare
Novartis AG chairman Daniel Vasella steps down from the company he helped build over 25 years, he leaves behind [...]
Larry Ellison has been CEO of Oracle Corporation
Amit Singhal is Senior Vice President and a Google Fellow
TerreStar Corp. (TSTR) President and *CEO* Jeffrey Epstein as its new chief financial officer
Brian McBride joined Amazon UK as Managing Director

Amazon CEO Jeff Bezos

David Drummond joined Google in 2002. Today is senior vice president

[...] said Bill McDermott, co-CEO of SAP

Mr Krishan Dhawan, Managing Director of Oracle India Pvt. Ltd.

Jonathan Ian Schwartz (born October 20, 1965) was President and CEO at Sun Microsystems prior to its acquisition by Oracle

Source: own elaboration

It is easy to observe a vast number of possibilities when it comes to word order in researched material. Furthermore, the given list of search concepts is far from completion.

Standard methods of local pattern matching dictate creation of rules that trigger when exact number of tokens of right characteristics is found. Apart from great effort investment spent on rule creation, they are prone to misfiring when slightest change of word order occurs.

Good examples of local pattern matching are regular expressions and text processing automata. While tremendous tools they might induce considerable effort when applied to information extraction. First of all, it was observed that regular expressions tend to fail in information retrieval task, not because their inefficiency but due to users being overwhelmed by their syntax. To exemplify above lets point out that, one has to be an experienced user to produce regular

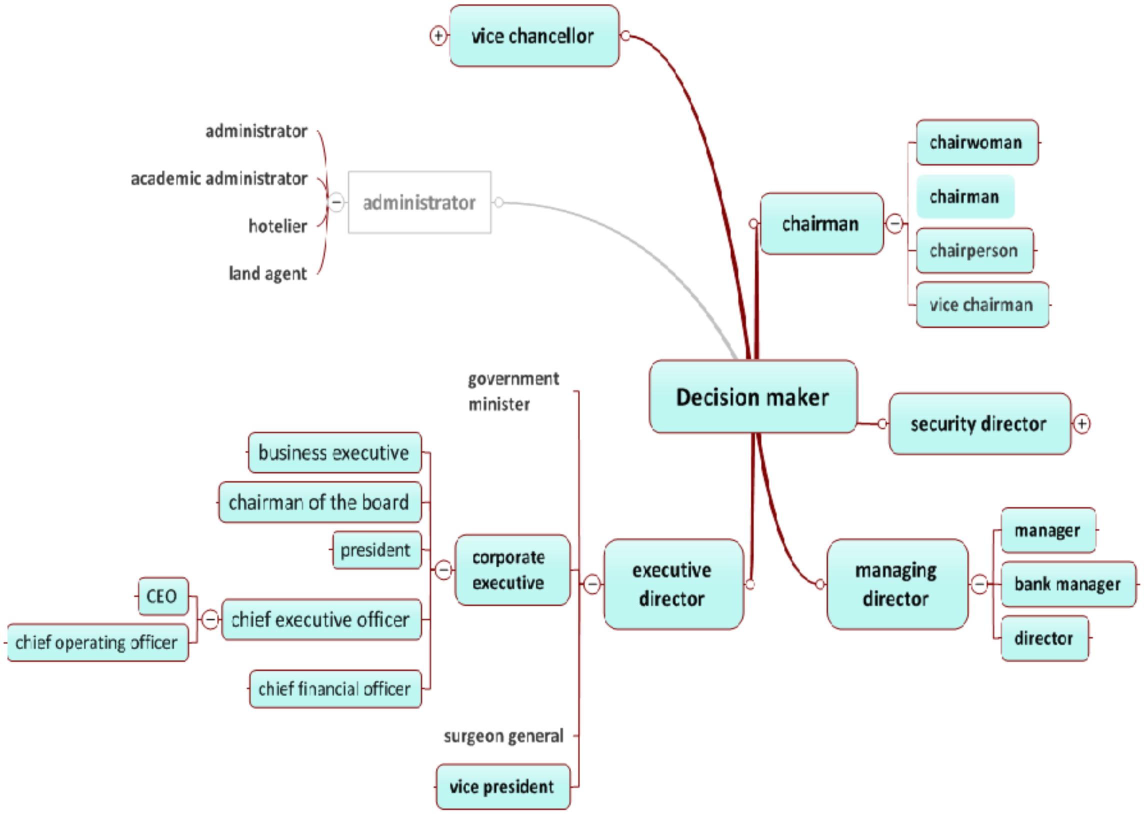

Figure 1**:** Excerpt from the *WiSENet* taxonomy showing concepts related to decision maker. Elements filled in blue constitute corporate decision makers. Source: own elaboration

expression that will match more than 99% valid emails. As with practice comes experience, more important issue with regular expression ([19] demonstrated that regular expressions can be converted into non-deterministic finite state automata - FSA) is its sensitivity to word order permutations.

Considering grammars, there should be remember that they will have to face the challenge of an alphabet that is finite but actual number of symbols cannot be counted a priori. One has to process whole corpora to enumerate all alphabet's symbols. When processing a language such as English this can be troublesome, as there is no known boundaries of resources that should be processed.

Ideal solution to above mentioned issues, shall combine flexibility and ease of use. Flexibility shall be understood as ability to adapt to natural permutations in a word order of processed text. Ease of use shall make a user exert the least amount of effort in a formulation of his information needs.

3. APPLICATION OF WISENET

Coming back to introduced motivation scenario, there is easily to observe that given results of recognizance query share common structure. This structure shall be treated as case analysis which leads to introduction of method designed by the author to automate information retrieval in this specific task.

Every result contains some information on person, its position (managerial one) and some company. Whether there is a task to build a datastore of data on managers in some kind of industry, a method that works with such high level query concepts as executive, person and company name will be of tremendous help.

When there should be start with a corpus of some textual data, it can be filtered it through envisioned method and come up with elements that become candidates to extend current knowledge base. Found elements in textual data can be new relations among already stored data, or new more general/specific concepts directly in relation with existing ones. The whole process of acquisition of new concepts and relations bases on the *WiSENet* semantic network structure. Effects of the process are reflected onto it, thus subsequent usage yields better result than previous ones.

The *WiSENet* network stores a corporate executive as a concept. This concept has other concepts in lexical relation, such as its hypernym and various hyponyms. A list of most important is given in Figure 1.

4. ALGORITHM OF MATCHING RULES

Before applying algorithm for matching rules there is necessary to carry out text-refinement process (starting from unstructured text document input and resulting finally a structure containing stacked sequentially descriptors of concepts found in the input document). An action that make up the process of text-refinement in documents starts from extracting lexical units (tokenization), and further text refinement operations are: elimination of the words from the so-called information stop-list, the identification of multiword concepts, bringing concepts to the main form by lemmatization or stemming. It is particularly difficult task for highly flexible languages, such as Polish, Russian or French (multiple noun declination forms and verb conjugation forms).

Synonyms need to be represented with concept descriptors using a semantic network. It allows correct similarity analysis and also increases classification algorithms efficiency without loss in comparison quality [14].

Abstracting process faces another problem here, which is polysemy. One word/phrase can represent multiple meanings, so the apparent similarity need to be eliminated. The problem is distinguishing words/phrases senses: for a polysemous word/phrases, we aim to have one node per sense in the resulting network, merging all occurrences to the correct node. This corresponds to unsupervised Word Sense Induction and Discrimination [28]. It is done by a procedure of concept disambiguation, which identifies word/phrase meaning depending on its context, is important to ensure that no irrelevant documents will be returned in response to a query [16], [17], [23]. Concept disambiguation entails indicating the appropriate meaning for ambiguous concepts, which results in obtaining information, as inferred from the documents, which better matches information needs. Disambiguation method based on lexical relations from the semantic network examines word context to determine its meaning, resulted in 82 % accuracy. It seems that only linguistic analysis methods can exceed 90 % accuracy [25], while human experts are able to recognize correct meaning of 96,8 % of polysemous words/phrases [24].

The last operation in the text refinement procedure is a generalization of concepts using semantic compression.

The final effect of the refinement procedure is the structure of documents containing ordered descriptors of concepts derived from the input document. This structure can be stored as an abstract (containing data for creating index) of the document, and then use the algorithm for discovering new concepts or new lexical relationships between concepts already existing in the *WiSENet*.

Devised algorithms uses ideas already mentioned in previous publications. All operations are performed with the *WiSENet* as a semantic net. The first important step in the algorithm is a procedure that unwinds rule into all hyponyms stored inside the network. This operation can be of considerable cost in terms of execution as it has to traverse all possible routes from chosen concept to terminal nodes in the network. After completion a list of rules is obtained, listing every possible permutation of concepts from the semantic network. To shorten processing time, there should be specify a number of levels that the procedure shall descend in its course of execution.

Next phase of the algorithm is to step through textual data in order to find matches on computed rules. Stepping through is done by employing bag of concepts approach. The bag of concepts is implemented as a Finite State Automaton (formally the automaton is a transducer) with advanced methods for triggering desired actions. At any state, it check whether any of the rules to be matched is completed. Discussion covering details of implementation is beyond the scope of this article. Nevertheless, it can be visualized as a frame passing through textual data. With every shift towards end of text fragment, concepts inside frame are used to check whether they trigger any of the rules obtained in the first phase. Size of the bag is chosen by researcher, yet performed experiments show that best results are obtained for a bag of size from 8 to 12 when rules are 2 to 5 concepts long.

Bag of concepts is very important idea, as it tolerates mixins and concept order permutations. All matchings are performed after initial text processing phase is performed. Text processing phase consist of well known procedures such as applying stop list and words/terms normalization.

A mixin is in this case a passage of text that serves some purpose to original text author, yet it separates two or more concepts that exist in one of the computed rules. Consider following examples placed in the Table 2.

Table 2: Examples of matching rules.

Rule: disease (all hyponyms), therapy (all hyponyms)	
Match in: *chemotherapy drug finish off remaining cancer*	
Matched concepts	therapy → chemotherapy, disease → cancer
Mixin	drug finish off remaining
Match in: *gene therapy development lymphoma say woods*	
Matched concepts	therapy → gene therapy, disease → lymphoma
Mixin	development
Match in: *cancer by-bid using surgery chemotherapy*	
Matched concepts	therapy → chemotherapy, disease → cancer
Mixin	by-bid using surgery
Match in: *Encephalitis is an acute infection and inflammation of the brain where therapy is supportive treatment*	
Match concepts	disease → acute infection, therapy → supportive treatment
Mixin	encephalitis, brain

Source: own elaboration

Examples are taken from one of the experiments performed with biology and medicine corpus. It can be observed, that bag of concepts performs well in various cases, it handles long mixins and concept permutation. Additional observation shall be made as concepts being hyponyms to those in the original example rule were matched (as referenced earlier).

All experiments performed took into account possibility of matching more than single rule. Thus a mechanism for triggering a set of rules was devised and was signaled earlier along with the bag of concepts.

A procedure matching rules holds special internal registers, that store rules that are actively valid with given bag of concepts and actual results of filtering textual data. To give an example, please consider a set of three following rules:

rule 1 : university, city (all hyponyms)

rule 2: university, city (all hyponyms), country (all hyponyms)

rule 3 : first name (all hyponyms), academic (all hyponyms)

Given is exemplary text fragment:

> A team of chemists led by chemistry professor David Giedroc from Indiana University (in Bloomington, USA) described a previously unknown function of a protein they now know is responsible for protecting a major bacterial pathogen from toxic levels of copper. Co-author with Giedroc on the paper is professor Michael J. Maroney of the University of Massachusetts. The results were published Jan. 27 in Nature Chemical Biology.

Procedure shall match and matches previously defined rules:

The procedure shall match and matches previously defined rules:

rule number 1: with university → university, Bloomington → city, new concept: *Indiana University in Bloomington*

rule number 2: with university → university, Bloomington → city, USA → country, new concept: *Indiana University in Bloomington*

rule number 3: with David → first name, professor → academic, new concept: *David Giedroc = professor(Indiana University University in Bloomington)*

rule number 3: with Michael → first name, professor → academic, new concept: *Michael J. Maroney = professor(University of Massachusetts)*

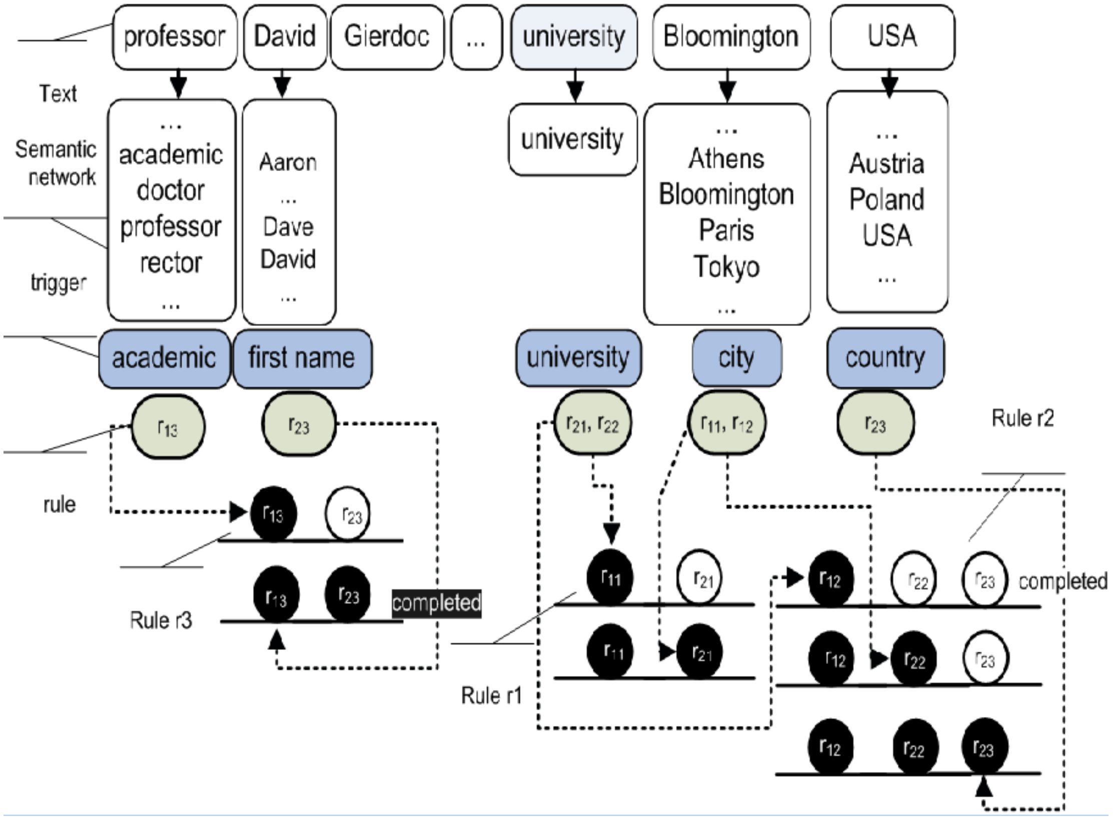

Figure 2: Process of matching rules from Example 1. Source: own elaboration

When a complete rule or its part (one can decide whether he is interested in total matches all partial ones) is mapped, it is presented to a user to accept match or reject it. When the bag of concepts drop earlier concepts and is filled with new ones, rules that were not matched are dropped from register of valid rules.

Algorithm in pseudocode is presented in listing 1

Algorithm 1: Algorithm for matching rules using WiSENet and bag of concepts

```
//attach rule triggers to concepts in semantic network
mapRulesToSemNet(SN, R[])
for all Rule $\in$ R do
    for all Word, Relations $\in$ Rule do
        N = SN.getNeighbourhood(Word, Relations)
        for all Word $\in$ N do
            SN.createRuleTrigger(Word, Rule)
        end for
    end for
end for
// Phase 2: text processing: tokenization, phrases, stop list
T = analyzeText(Input)
foreach Word in T do
    if count(Bag) = size(Bag) then
    //First, deactivate rules hits for a word
    //that drops out from bag of words
    oldWord = pop(Bag)
    end if
for all Rule $\in$ SN.getTriggers(oldWord) do
    Rule:unhit(Word)
    push(Bag, Word)
    for all Rule $\in$ SN.getTriggers(Word) do
        //take all relevant rules and activate word hit
        Rule.hit(Word)
        if Rule.hitCount = Rule.hitRequired then
            //report bag of words when hits reaches required number
            report(Rule, Bag)
        end if
    end for
end for
```

SN - Semantic Network

R - semantic relation pattern

5. EXPERIMENT

Devised algorithm (automaton) was used to perform an experiment on biology related data. The main aim of the experiment was to discover new concepts and insert them in the *WiSENet* semantic Netowrk structure. The test corpus of documents consisted on 2589 documents. A total number of words in a document's corpus was over 9 million. Author decided to prepare query for searching for specialists and their affiliations. This converges with motivating scenario, as the *WiSENet* semantic network was enriched by both specialists (and their fields of interest), universities, institutes and research centers.

Experiment used following rules:

rule 1: first name (all hyponyms), professor (all hyponyms), university (all hyponyms)

rule 2: first name (all hyponyms), professor (all hyponyms), institute (all hyponyms)

rule 3: first name (all hyponyms), professor (all hyponyms), research center (all hyponyms)

rule 4: first name (all hyponyms), professor (all hyponyms), department (all hyponyms)

rule 5: first name (all hyponyms), professor (all hyponyms), college

Size of the bag of concepts was set at 8 elements. Additionally, all rules were to match exactly all concepts.

Out of 1326 documents where concept ”professor” was found, prepared rules matched 445 text fragments. This gives a recall rate of 33,56%. Precision of results was 84,56%. This level is found to be very satisfactory, especially taking into account that due to algorithm nature there can be duplicates of matched text fragments (due to multiple triggering of rules inside current bag of concepts).

Table 3 demonstrates sample results. Please notice, that match on its own does not discover new concepts. Rules present potential fragments that with high likelihood contain new concepts that can be included into semantic network.

In addition, experiment resulted in 471 concepts that were previously unknown to the *WiSENet* network. The context and the type of rules that matched text fragments led to extremely efficient updates of the network. The author developed a new version of the devised algorithm, in which the goal is to discover new lexical relations between concepts. So, the same mechanism of creating rules can be applied to discover and store new lexical relationships using rules containing concepts already stored in the semantic network.

Further experiments performed with this new version showed that the same corpus and the same rules can be used to acquisition new lexical relationships between concepts already stored in the semantic network.

Table 3: Sample results of experiments with rules based on *WiSENet* on corpus of biology related documents. Discovered concepts are written under matches.

text fragment	match/discovered concept	rule
explain senior author Douglas Smith Md professor department neurosurgery director	Douglas professor department **Douglas Smith**	5
Feb proceedings national academy of sciences researcher University of Illinois entomology professor Charles Whitfield postdoctoral	University of Illinois professor Charles **Charles Whitfield**	1
design function biological network she-bop visiting professor Harvard University Robert Dicke fellow visiting	professor Harvard University Robert **Robert Dicke**	1
modify bacteria Thomas Wood professor --Artie-- --McFerrin-- department chemical engineering have	Thomas professor department **Thomas Wood**	5
Matthew --Meyerson-- professor pathology Dana --Farber-- cancer institute senior associate	Matthew professor institute **Matthew Meyerson**	2
an assistant professor medical oncology Dana --Farber-- cancer institute researcher broad assistant	professor Dana institute **Dana Farber**	2
sun mat professor emeritus Robert --Hodson-- all university Georgia Robert Edwards	professor Robert university **Robert Hodson**	1
vacuole David Russell professor molecular microbiology --Cornell's-- college veterinary medicine colleague	David professor college **David Russell**	4
chemistry professor David Giedroc from Indiana University (in Bloomington, USA) described a previously unknown function of a protein	David Professor university **Davic Giedroc**	1
resistant cell professor Peter --Sadler-- chairman chemistry department University of Warwick lead research project	professor Peter University of Warwick **Peter Sadler**	1
said first author Quyen Nguyen doctorate assistant professor surgery si tan University of California San Diego school of medicine	Nguyen assistant professor University of California **Quyen Nguyen**	1
scientist --Sirtris-- co-author founder prof David --Sinclair-- Harvard Medical School published consecutive	professor David Harvard Medical School **David Sinclair**	1

Source: own elaboration

6. Conclusions

In this article have been presented an approach for building or expanding large-scale semantic networks automatically from text, employing deep semantic processing with appropriate mechanisms (finite state automaton).

The work presented in this article continues research efforts started with presentation of Semantically Enhanced Property Protection System *SeiPro2S* [5]. The *SeiPro2S* system has proved to be a efficient tool in checking whether submitted content is not an unauthorized copy. *SeiPro2S* makes it possible to not only find direct copying, but also to find passages that rephrase the copied content with another set of words, thus reproducing the original thought. Designed *SHAPD2* algorithm is highly efficient in plagiarism detection task and employing semantic compression is strong resilient to false-positives examples of plagiarism (see [9]), which is may an issue in case of using competitive algorithms [27].

The SHAPD2 algorithm has extremely low computational complexity estimated as linearithmic and uses technique of hashing whole sentences. The final architecture of the *SeiPro2S* system and its functionality has been obtained by introducing new mechanisms which effectiveness was established thanks to performed experiments and was described in [9].

After realizing vision of semantic compression for English and presenting results, author decided to focus on applications enabling network expansion with new concepts and new lexical relationships using specially constructed automata (which is functionally a transducer) what is necessary to increase performance quality as the *WordNet* realizing NLP or computational linguistics tasks.

Rules created with the *WiSENet* are interesting application, that has great potential for future development, as it helps to expand body of knowledge represented by *WiSENet*. Experiments performed with devised algorithm for rule matching showed that envisioned flexibility and precision are available. Further experiments performed with devised algorithm showed that the same corpus and the same rules can be used to acquisition new lexical relationships between concepts already stored in the semantic network.

As reported in the experiment section, due to reasonably high precision on achieved results, unknown concepts can be easily added, thus realizing a vision of knowledge acquisition with the *WiSENet*.

Future work will focus on further addition of previously unknown concepts to the *WiSENet* along with restructuring of relations among them. Author believes that there are even more useful applications of semantic compression and plan to experiment with them and share experiments' results.

REFERENCES

[1] Baeza-Yates, R., Ribeiro-Neto, B. (1999), "Modern Information Retrieval", ACM Press, Addison-Wesley Longman Publishing Co., New York

[2] Becker, M., Drozdzynski, W., Krieger, H. U., Piskorski, J., Shaafer, U., Xu, F. (2002), *SProUT - shallow processing with unification and typed feature structures*, Proceedings of the International Conference on Natural Language Processing, ICON-2002

[3] Blackburn, P., Bos, J. (2005), *Representation and Inference for Natural Language*, A First Course in Computational Semantics, CSLI Publications

[4] Brachman, R. J., Levesque, H. J. (2004), *Knowledge Representation and Reasoning*, Morgan Kaufmann, p. 381

[5] Ceglarek, D., Haniewicz, K., Rutkowski, W. (2009), *Semantically Enchanced Intellectual Property Protection System - SEIPro2S*, 1st International Conference on Computational Collective Intelligence,. Nguen N.T. (ed.) In: Lecture Notes in Artificial Intelligence - vol. 5796: Computational Collective Intelligence - Technologies and Applications, pp. 449-59, Springer Verlag Berlin Heidelberg

[6] Ceglarek, D., Haniewicz, K., Rutkowski, W. (2010), *Semantic compression for specialized Information Retrieval systems*, Studies in Computational Intelligence, vol. 283 of Lecture Notes in Artificial Intelligence, pp. 111–121, Springer Verlag, Berlin Heidelberg

[7] Ceglarek, D., Haniewicz, K., Rutkowski, W. (2010), *Quality of semantic compression in classification*, Lecture Notes in Artificial Intelligence - vol. 6421: Computational Collective Intelligence - Technologies and Applications, pp. 162–171, Springer Verlag, Berlin Heidelberg

[8] Ceglarek, D., Haniewicz, K., Rutkowski, W. (2012), *Robust Plagiary Detection Using Semantic Compression Augmented SHAPD*, 4th 1st International Conference on Computational Collective Intelligence - ICCCI 2012, pp. 308–317, Lectures Notes in Computer Science, Springer Verlag, Berlin Heidelberg

[9] Ceglarek, D. (2013), *Architecture of the Semantically Enhanced Intellectual Property Protection System*, Lecture Notes in Artificial Intelligence - Computer Recognition System 5, pp. 12, Springer Verlag, Berlin Heidelberg

[10] Fellbaum, C. (1998), *WordNet - An Electronic Lexical Database*, The MIT Press

[11] Goddard, C., Schalley, A.C. (2010), *Semantic Analysis*, edit. Indurkhya N. & Damerau F. In: Handbook of Natural Language Processing, , pp. 93-12, Chapman Hall/CRC

[12] Gonzalo, J., Vardejo, F., Chugur I., Cigarrán, J.M. (1998), *Indexing with WordNet Synsets can improve Text Retrieval*, Proceedings of the COLING/ACL '98 Workshop on Usage of WordNet for NLP, pp. 38-44

[13] Hotho, A., Staab, S., Stumme, S. (2003), *Explaining Text Clustering Results using Semantic Structures*, Principles of Data Mining and Knowledge Discovery, 7th European Conference PKDD 2003

[14] Hotho, A., Maedche, A. , Staab, S. (2003), *Ontology-based Text Document Clustering*, Proceedings of the Conference on Intelligent Information Systems, Zakopane, Physica/Springer

[15] Keikha, M., Razavian, N. S., Oroumchian, F., Razi, H. S. (2008), *Document Representation and Quality of Text: An Analysis*, ed.. M. W. Berry M. Castellanos, Survey of Text Mining II: Clustering, Classification, and Retrieval, pp. 219-232, Springer Verlag, Berlin Heidelberg

[16] Khan, L., McLeod, D., Hovy, E. (2004), *Retrieval effectiveness of an ontology-based model for information selection*, The VLDB Journal, volume 13 (1), pp. 71-85, Springer Verlag, Berlin Heidelberg

[17] Krovetz, R., Croft, W. B. (1992), *Lexical Ambiguity and Information Retrieval*, ACM Transactions on Information Systems, vol. 10(2), pp. 115-141

[18] Frakes, W. B., Baeza-Yates, R. (1992), "Information Retrieval: Data Structures and Algorithms", Prentice Hall

[19] McNaughton, R., Yamada, H. (1960), *Regular expressions and state graphs for automata*, IRE Transactions on Electronic Computers EC-9(1), pp. 39–47

[20] Miller, G. A. (1995), *Wordnet: a lexical database for English*, Communications of the ACM, vol. 38 (11), pp. 39–41

[21] Califf, M. E., Mooney, R. J. (2003*), Bottom-up relational learning of pattern matching rules for information extraction*, Journal of Machine Learning Resources, vol. 4, pp. 177-210

[22] Sinha, R., Mihalcea, R. (2007), *Unsupervised graph-basedword sense disambiguation using measures of word semantic similarity*, pp. 363–369, In: ICSC

[23] Zhu, J., Hovy, E.H. (2007), *Active Learning for WordSense Disambiguation with Methods for Addressingthe Class Imbalance Problem*. In: Proceedings of the EMNLP Conference, Prague

[24] Gale, W., Church, K., Yarowsky, D. (1992), *A Method for Disambiguating Word Senses in a Large Corpus*, pp. 415--439, Computers and the Humanities. vol. 26

[25] Sanderson, M. (2000), *Retrieving with Good Sense*, Information Retrieval, pp. 45--65

[26] Sowa, J. (1991), "Principles of Semantic Networks", Morgan Kaufmann

[27] Zhang, Q., Zhang, Y., Yu, H., Huang, X. (2010), *Efficient partial-duplicate detection based on sequence matching*, In: SIGIR '10: Proceeding of the 33rd international ACM SIGIR Conference on Research and development in information retrieval, pp. 675--682, New York, NY, USA, ACM

[28] Navigli, R. (2009), *Word sense disambiguation: A survey*. ACM Computational Survey 41(2), pp. 1–69

Author

Dr. D. Ceglarek is a Assistant Professor of Information Technology, Poznan School of Banking, Poznan, Poland. He earned a Ph.D in Clustering of information from Department of Information Systems, Poznan University of Economics, Poznan, Poland in 1997. He has published more than thirty of research papers in the area of classification of information, plagiarism detection, natural language processing and semantic technology. He has served as Assistant Professor in Department of Information Systems, Poznan University of Economics during 1997 to 2007.