International Journal of Intelligent Information Systems
2013; 2(2) : 26-33
Published online April 2, 2013 (http://www.sciencepublishinggroup.com/j/ijiis)
doi: 10.11648/j.ijiis.20130202.11

Knowledge acquisition for expanding semantic network

Dariusz Ceglarek

Poznan School of Banking, Poznan, Poland

Email address:
dariusz.ceglarek@wsb.poznan.pl (D. Ceglarek)

To cite this article:
Dariusz Ceglarek. Knowledge Acquisition for Expanding Semantic Network. *International Journal of Intelligent Information Systems.* Vol. 2, No. 2, 2013, pp. 26-33. doi: 10.11648/j.ijiis.20130202.11

Abstract: This article presents the issues of knowledge management, in particular knowledge acquisition. The 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. Detailed discussion is given with description of devised algorithm, usage examples and results of experiments.

Keywords: Semantic Network, Semantic Compression, WiseNet, Knowledge Acquisition, Lexical Relationships, Natural Language Processing, Knowledge Representation Structures

1. Introduction

Natural language is a very complex system which needs to be represented in a way that would be understandable for computer systems (the main task realized by computational linguistics 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][10][14]. Semantic knowledge, as identified lexical relations between concepts, should be stored in an appropriate data structure in order to be utilized to refine Natural Language Processing (NLP) tasks and their results. a semantic network is a structure incorporating knowledge about all possible semantic relations between words. Semantic networks store information about similarity relations (like a thesaurus): word similarity, synonymy, antonymy; hierarchical relations (like a taxonomy): hypernymy, troponymy (for verbs only) 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. Choosing proper knowledge representation structure is very important determinant of the classification quality of text documents [21]. That is why their utilization in information retrieval systems should bring the biggest improvement in their effectiveness. 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. 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 [9][19]. Its structure is organized around notion of synset. 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 [4]. In this paper were enumerated various aspects and possible merits of the WiSENet semantic network.

The aim of this work is to present a new mechanism working as application of the previously introduced semantic network WiSENet (popular English semantic network WordNet transferred into SenecaNet format introduced in [4]).

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 and further acquisition them to semantic network what is possible using e.g. *shallow text processing* methods.

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

To begin with, the WiSENet derived its content from the ,ost popular english semantic network 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 synset structure and turn it into a graph where nodes represent concepts and graph vertices denote relation of hypernymy/hyponymy. This enabled devised algorithms to easily follow relations among particular concepts found in real life textual data. Restructuring was carried out in a lossless manner (algorithm is given in [4]).

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 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 terms. Descriptors are chosen to represent a set of synonyms or hyponyms in the processed passage. Decision is made taking into account relations among the terms 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 and various objects. WordNet in general does not miss most general categories of entities, yet a lot of highly specialized 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 discovered that WiSENet is very useful in discovering concepts that represent some specialization of other concepts by employing specially prepared rules.

WiSENet can be applied to a set of procedures, that aim 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.

If one is to prepare 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 terms than can denote a managerial position in some company.

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

William (Bill) H. Gates is chairman of Microsoft Corporation
Richard K. Matros has been Chairman and CEO of Sun Healthcare
Larry Ellison has been CEO of Oracle Corporation
Amit Singhal is Senior Vice President and a Google Fellow
Jeffrey Epstein as its new chief financial officer
Prof. Dr. Dr. h.c. mult. Martin Winterkorn - Member of the Board of Management of Volkswagen AG,with responsibility for 'Group Research and Development'
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.

One is ready to observe a vast number of possibilities when it comes to word order in researched material. Furthermore, the given list of search terms 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 expression that will match more than 99% valid emails. As with practice comes experience, more important issue with regular expression ([13] demonstrated that regular expressions

can be converted into non-deterministic finite automata) is its sensitivity to word order permutations.

When one is to consider grammars, one has to 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 user exert the least amount of effort in 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 terms as executive, person and company name will be of tremendous help.

When one is to start with a corpus of some textual data, one can filter it through envisioned method and come up with elements that become candidates to extend current knowledge base. Found elements 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. Effects of the process are reflected onto it, thus subsequent usage yields better result than previous ones.

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

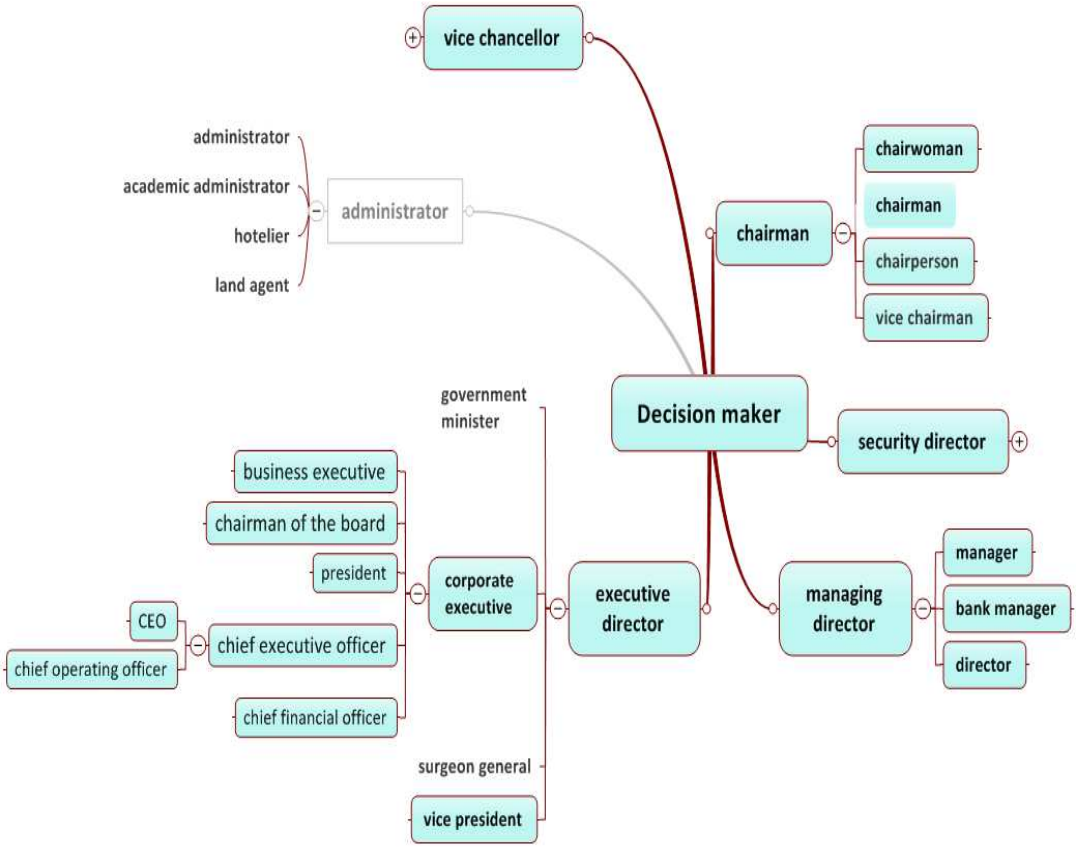

Figure 1. *Excerpt from WiSENet showing concepts related to decision maker. Elements filled in blue constitute corporate decision makers.*

4. Algorithm of Matching Rules

Before applying algorithm for matching rules there is necessary to carry out text-refinement process (from un-structured text document input to a structure containing stacked sequentially descriptors of concepts found in the input document). 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 or French (multiple noun declination forms

and verb conjugation forms).

Synonyms need to be represented with concept descriptors using semantic network. It allows correct similarity analysis and also increases classification algorithms efficiency without loss in comparison quality [8].

Abstracting process faces another problem here, which is polysemy. One word can represent multiple meanings, so the apparent similarity need to be eliminated. It is done by concept disambiguation, which identifies word meaning depending on its context, is important to ensure that no irrelevant documents will be returned in response to a query [11], [15], [16], [22]. Disambiguation method based on lexical relations from semantic network examines word context to determine its meaning, resulted in 82\% accuracy. It seems that only linguistic analysis methods can exceed 90\% accuracy, while human experts are able to recognize correct meaning of 96,8\% of polysemic words.

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

The final effect of 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 (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 semantic network. The first important step in algorithm is 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 network. To shorten processing time, one can specify number of levels that 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. Bag of concepts is implemented as a Finite State Automaton 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 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 term normalization.

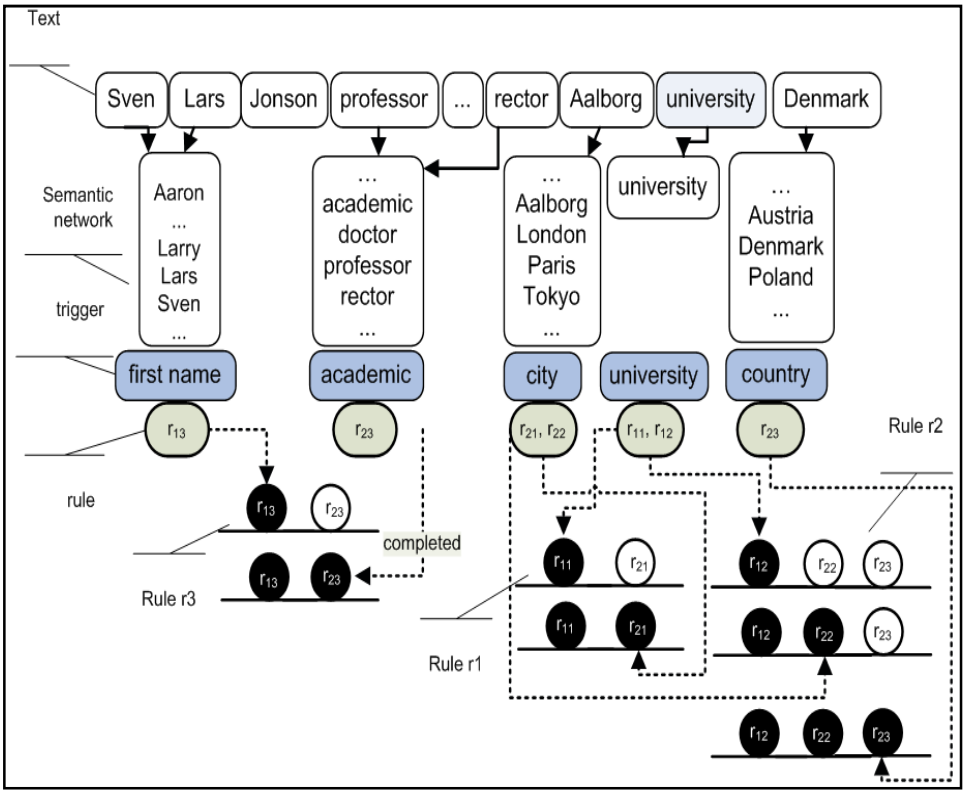

Figure 2. *Process of matching rules from Example 1.*

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:

Example 1	
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

Examples are taken from one of the experiments performed with biology 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 bag of concepts.

Procedure matching rules holds internal registers, that store rules that are actively valid with given bag of concepts. To give an example, please consider a set of three rules:

rule 1 : university, city (all hyponyms)

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

rule 3 : first name (all hyponyms), academic

Given exemplary text fragment:

SVEN LARS CASPERSEN, Professor of Economics, President of the World Rector's Association, Rector of Aalborg University (Denmark) (1999)

Procedure shall match and matches previously defined rules:

rule number 1 with university → university, Aalborg → city, new concept: Aalborg University

rule number 2 with university → university, Aalborg → city, Denmark → country, new concept: Aalborg University

rule number 3 with Sven Lars → first name, professor → academic, new concept: Sven Lars Caspersen = professor(Aalborg University)

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 user to accept match or reject it. When 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 R do
for all Word, Relations Rule do
N = SN.getNeighbourhood(Word, Relations)
for all Word 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 SN.getTriggers(oldWord) do
Rule:unhit(Word)
push(Bag, Word)
for all Rule 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 was used to perform an experiment on biology related data. Test corpus consisted on 2,589 documents. A total number of words in documents was over 9 million. Author decided to search for specialists and their affiliations in test corpus. This converges with motivating scenario, as the WiSENet was enriched by both specialists (and their fields of interest), universities, institutes and research centers. In experiment were used following rules:

***Table** 1. Rules used in experiment.*

rule 1	first name (all hyponyms), professor (all hyponyms), institute (all hyponyms)
rule 2	first name (all hyponyms), professor (all hyponyms), university (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 bag of concepts was set at 8 elements. Additionally, all rules were to match exactly all concepts. Rules that were used for the experiment are shown in Table 1.

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 2 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. Context and type of rules that matched text fragments led to extremely efficient updates of the network.

***Table 2**. Sample results of experiments with rules based on the 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
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

6. Conclusions

Work presented in this article continues research efforts started with presentation of Semantically Enhanced Property Protection System SeiPro2S [2]. 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 plagiary detection task and employing semantic compression is strong resilient to false-positives examples of plagiarism (see [5]), which is may an issue in case of using competitive algorithms. 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 [6]. 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 (transducer) what is necessary to increase performance quality as WordNet realizing Natural Language Processing (NLP) 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 the WiSENet. Experiments performed with devised algorithm for rule matching showed that envisioned flexibility and precision are available.

As reported in the experiment section, due to reasonably high precision on results, unknown concepts can be easily added, thus realizing 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] R. Baeza-Yates, B. Ribeiro-Neto: "Modern Information Retrieval", ACM Press, Addison-Wesley Longman Publishing Co., New York, 1999.

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

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

[4] D. Ceglarek, K. Haniewicz K. and W. Rutkowski: "Semantically Enchanced Intellectual Property Protection System - SEIPro2S", 1st International Conference on Computational Collective Intelligence, Springer Verlag Berlin Heidelberg, 2009, pp. 449—59.

[5] D. Ceglarek, K. Haniewicz and W. Rutkowski: "Semantic compression for specialized Information Retrieval systems", In: Studies in Computational Intelligence, vol. 283, Springer Verlag, Berlin Heidelberg, 2010, pp. 111–121.

[6] D. Ceglarek, K. Haniewicz and W. Rutkowski: "Quality of semantic compression in classification", Lecture Notes in Artificial Intelligence, vol. 6421, Springer-Verlag, Berlin-Heidelberg, 2010, pp. 162–171.

[7] D. Ceglarek, K. Haniewicz and W. Rutkowski: "Robust Plagiary Detection Using Semantic Compression Augmented SHAPD", ICCCI 2012 Conference, LNCS, , Springer Verlag, Berlin Heidelberg, 2012, pp. 308–317.

[8] D. Ceglarek: "Architecture of the Semantically Enhanced Intellectual Property Protection System", In: Lecture Notes in Artificial Intelligence - Computer Recognition System 5, Springer Verlag, Berlin Heidelberg, 2013.

[9] C. Fellbaum: "WordNet - An Electronic Lexical Database", The MIT Press, May 1998.

[10] C. Goddard and A.C. Schalley: "Semantic Analysis", ed. N. Indurkhya F. Damerau In: Handbook of Natural Language Processing, Chapman Hall/CRC, 2010, pp. 93-121.

[11] J. Gonzalo et al.: "Indexing with WordNet Synsets can improve Text Retrieval", 1998.

[12] A. Hotho, S. Staab and S. Stumme: "Explaining Text Clustering Results using Semantic Structures", In: Principles of Data Mining and Knowledge Discovery, 7th European Conference PKDD 2003, 2003.

[13] A. Hotho, A. Maedche and S. Staab: "Ontology-based Text Document Clustering", In: Proceedings of the Conference on Intelligent Information Systems, Zakopane, Physica/Springer, 2003.

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

[15] L. Khan, D. McLeod and E. Hovy: "Retrieval effectiveness of an ontology-based model for information selection", 2004.

[16] R. Krovetz R. and W. B. Croft.: "Lexical Ambiguity and Information Retrieval", 1992.

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

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

[19] G. A. Miller: "Wordnet: a lexical database for English", Commun. ACM 38, 1995, pp. 39–41.

[20] M. E. Califf and R. J. Mooney: "Bottom-up relational learning of pattern matching rules for information extraction". Journal of Mach. Learn. Res. 4, Dec. 2003, pp. 177-210.

[21] F. Ricceri: "Intellectual Capital and Knowledge Management. Strategic management of knowledge resources", Routledge Francis & Taylor Group, New York 2008

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