NLU METHODOLOGIES FOR CAPTURING NON-REDUNDANT INFORMATION FROM MULTI-DOCUMENTS

A Survey

Michael T. Mills and Nikolaos G. Bourbakis
College of Engineering, ATRC, Wright State University, Dayton, Ohio 45435, U.S.A.

Keywords: Natural language understanding, Natural language processing, Document analysis.

Abstract: This paper provides a comparative survey of natural language understanding (NLU) methodologies for capturing non-redundant information from multiple documents. The scope of these methodologies is to generate a text output with reduced information redundancy and increased information coverage. The purpose of this paper is to inform the reader what methodologies exist and their features based on evaluation criteria selected by users. Tables of comparison at the end of this survey provide a quick glance of these technical attributes indicators abstracted from available information in the publications.

1 INTRODUCTION

Over the past several years, information has become so vast that professionals, such as medical doctors, have difficulty keeping up to date within their respective fields. Time is wasted reading redundant information from various documents. Needed information may also be lost in the process of summarization. Advanced methods of search, database technologies, data mining, and other areas have helped, but not enough to meet the growing need from these professionals.

For the past 40 years, researchers have advanced in automatically or semi-automatically capturing information from single and multiple documents into less redundant text, typically in the form of summaries. Several methodologies have been developed to advance the area of natural language processing in order to find solutions to this problem. However, no known methodology appears to capture the needed information and generate text with enough quality and speed to satisfy this need. Thus, this survey summarizes current methodologies, which deal with the removal of redundancy for documents retrieved from different resources. The purpose is to document the progress in natural language understanding research and how it can be applied to capturing concepts from multi-documents and producing non-redundant text while attempting to maximize coverage of the significant information needed by the user.

The methodologies under evaluation in this paper cover the following areas: (1) detection of important sentences, (2) concept extraction from text, (3) building concept graphs, (4) attribute and relation structures leading toward knowledge discovery from text, (5) increasing efficiency in the processes leading to concept representations, (6) generation of non-redundant text summaries, and (7) maximizing the readability (or coherence) of automatically generated or extracted text.

2 METHODS AND FEATURES

In this section we present a variety of methodologies classified according to their features. In particular this section covers the various groups: text relationship map with latent semantic analysis, extraction methods for text summarization, cluster summarization, formulated semantic relations, SPN representation for document understanding, concepts representation for text, learning ontologies from text, synthesis of documents, generation of semantically meaningful text using logic order, text generation methods, document structural understanding, and other relevant methods. The methods presented here will be compared and evaluated based on their maturity. The overall results are presented in section 3.

T. Mills M. and G. Bourbakis N. (2010).
NLU METHODOLOGIES FOR CAPTURING NON-REDUNDANT INFORMATION FROM MULTI-DOCUMENTS - A Survey.
In *Proceedings of the 5th International Conference on Software and Data Technologies*, pages 384-393
DOI: 10.5220/0002998603840393

2.1 Text Relationship Map with Latent Semantic Analysis (LSA)

Yeh et al present two methodologies, text relationship map (TRM) and latent semantic analysis (LSA), used together for text summarization. TRM uses feature weights to create similarity links between sentences forming a text relationship map (Yeh et al, 2008a).

Advantages: This methodology captures various features that help in calculating the similarity of sentences throughout one or more documents. The paper gives significant detail about the methodology.

Disadvantages: This methodology is based at the word level.

Yeh et al.'s LSA-based text relationship map (T.R.M.) approach derives semantically salient structures from a document. Latent semantic analysis (LSA) is used for extracting and inferring relations of words with their expected context (Yeh et al., 2008b).

Advantages: The paper gives significant detail about the methodology. Several features are used in the similarity calculation.

Disadvantages: This methodology is based at the word level. The LSA approach uses a Word-Sentence matrix that can get very large due to the number of words in a document or in multi-documents.

2.2 Extraction Methods for Text Summarization

Ko and Seo present a hybrid sentence extraction method that uses some context information augmented with mainline statistical approaches to find important sentences in documents. This model combines two consecutive sentences into a bi-gram pseudo sentence representation to overcome feature sparseness (Ko and Seo, 2008).

Advantages: Test results of the hybrid sentence extraction approach showed that it out performed other approaches listed by a small percentage.

Disadvantages: What the authors (of the hybrid approach) call context information is limited to two consecutive (i.e., adjacent) sentences with no apparent global context capability. Generally, context implies more extensive surrounding information than groups of two adjacent sentences.

2.3 Cluster based Summarization

Moens et al. extract important sentences and detect redundant content across sentences. It uses generic linguistic resources and statistical techniques to detect important content from topics and patterns of themes throughout text (Moens et al., 2005).

Advantages: Moens et al. methodology provides a significant capability in automatically finding content from text and representing it by hierarchical topics and subtopics. This provides flexibility in selecting how much detail goes into the summary. From competitive testing at DUC 2002 and 2003, the performance of the methodology provided good results, even when compared with trained methodologies.

Disadvantages: Topic trees and themes are the main information sources to be captured using this methodology. Although these contribute to forming a summary, more queues could be added to enhance the accuracy of this approach. The authors discuss several improvements that could be made. This system incorporates several technologies to provide flexibility. It appears that system integration could be improved to make this a better product.

Radev et al. present a Cluster Centroid-Based summarization technique called MEAD that detects topics and tracks to evaluate the results. This methodology measures how many times a word appears in a document, and what percentage of all documents in a collection contains a given word. A cluster is a set of words that are statistically important to a cluster of documents and are used to identify important (or salient) sentences in a cluster (Radev et al., 2004).

Advantages: The authors state that the MEAD algorithms produced summaries similar in quality to summaries produced by humans for the same documents.

Disadvantages: Additional factors could be addressed to help provide higher quality output. Scores determined by using this methodology are limited to word frequency, position, and sentence overlap. More factors could be added to improve redundancy removal of the resulting summary output.

2.4 Chaining Lexically to Formulate Semantic Relations

Silber and McCoy propose an algorithm to improve

the execution time and space complexity of creating lexical chains from exponential to linear in order to make computation feasible for large documents. Lexical chains are created as an intermediate representation to extract the most important concepts from text to be used for generating a summary. An implementation of Lexical chains is evaluated as an efficient intermediate representative format. Silber and McCoy implicitly store every interpretation of source documents without creating each interpretation as a lexical chain, thus reducing the vast number of lexical chains from multiple word senses per noun instance (Silber and McCoy, 2002).

Advantages: Silber and McCoy's algorithm provide linear time for calculating lexical chains which is a big step from former exponential time complexity implementations they reference from 1997 implementations and earlier.

Disadvantages: Their focus is on efficiency of one part of the entire process. They leave some issues left for future work.

Manabu & Hajime provide lexical chaining based on a topic submitted by a user. Lexical chains are sequences of words related to each other that form a semantic unit. This procedure increases coherency and readability of resulting summaries which yields improved accuracy or relevance to the user. (This has an objective increasing coherency and readability of a generated text summary similar to Barzilay and Lapata but applies the lexical chaining methodology.) The methodology constructs lexical chains, calculates scores of the chains based on high connectivity with other sentences, and constructs clusters of words using the similarity score (Manabu & Hajime, 2000).

Advantages: This methodology provides a higher level calculation of semantic similarity and offers a potential increase in accuracy.

Disadvantages: Results showed improved accuracy but left possibilities of ignoring other useful information. More improvements need to be made.

Reeve et al. propose to use lexical chaining for concept chaining (distinguished from term chaining) to identify candidate sentences for extraction for use in generating biomedical summaries. This concept chaining process consists of text to concept mapping, concept chaining, identifying strong chains, identifying frequent concepts and summarizing. The resulting sentences are used to generate the summary (Reeve et al., 2006).

Advantages: Test results (90 % precision and 92 % recall) are high compared to results of other lexical chaining methodologies in this survey.

Disadvantages: Concept disambiguation is not implemented but planned for future work. Complexity appears not to be addressed. Internal evaluation was specifically toward quality of generated summary.

2.5 Stochastic Petri-net (SPN) Representations

Bourbakis and Manaris presented a paper on an SPN based Methodology for Document Understanding. They describe four levels of processing: **lexical** to enforce case (subject-verb) agreement, **syntactic** to combine words into sentences, **semantic** to assign meaning to words and sentences, and **pragmatic** to form context from relations to previous sentences, paragraphs, topics, and information from related data (Bourbakis and Manaris, 1998).

Advantages: The combination of augmented semantic grammars (ASGs) and SPNs in this methodology provides significant capability in not only capturing semantic meaning from text but extracting contextual and other available information to resolve ambiguities. The methodology suggested in this paper shows how SPNs, used with ASGs, can model a tremendous amount of interrelationships that exist in both text and imagery. It provides significant potential for extended areas such as knowledge abstraction and representation and extending their capabilities.

Disadvantages: SPNs have existed for a long time. However, the methodology presented in this paper illustrates the potential for SPNs to model technologies in ways that significantly enhance their modeling capabilities compared to conventional (main line) approaches in using SPNs.

2.6 Building Concept Representations from Text

Ye et al. propose a concept lattice to represent text understanding and to extract text from multiple documents and generate an optimized summary. The concept lattice provides indexing of local topics within a hierarchy of topics (Ye et al., 2007).

Advantages: The document concept lattice approach provides an efficient way to account for all possible word senses without calculating them all on line. This provides significant improvement in accuracy without the computational complexity.

According to the authors, the approach reduces complexity from $O(n^2)$ to O(1), i.e. linear.

Disadvantages: WordNet is required for this approach. New tools adopting this approach may be restricted to use WordNet, depending on any implementation dependent concerns.

Guo and Stylios investigate event indexing by applying cognitive psychology to create clusters for building concept representations from text. Their methodology extracts the most prominent content by lexical analysis at phrase and clause levels in multiple documents (Guo and Stylios, 2005).
Advantages: Working at the phrase or clause level is an advantage over word level. This reduces the number of possible combinations of pairs (phrase, sentence) instead of (word, sentence) for example. Multi-document capability is another plus for the user. Features such as actors, time/space displacements, causal chains, and intention chains add a significantly more capability to detecting sentence similarities. Reducing all this potentially multi-dimensional vector data to two dimensional index clustering is a significant savings in complexity, especially storage complexity.

Disadvantages: Dimension reduction can sometimes hide important vector component data.

Cimiano et al. formed concept hierarchies using formal concept analysis (FCA) through unsupervised learning. Their methodology automatically acquires (through learning) concept hierarchies from collections of text (corpus) (Cimiano et al., 2005).

Advantages: Automatic (unsupervised) leaning approach is a big plus, reducing the traditional manual work to near zero. The concept similarity calculation uses more characteristics that can result in greater accuracy of output text. The authors state "this is a first time approach." Similarity calculations are made at the concept and semantic level, using LSA.

Disadvantages: The approach appears to be integrated with the LoPar parser implementation, but benefits in capability are significant.

2.7 Learning Ontology from Text

Bendaould et al. used relational concept analysis (RSA) to formulate concepts through text-based ontology. This paper presents a semi-automatic methodology that builds ontology from a set of terms extracted from resources consisting of text corpora, a thesaurus for a particular domain, and syntactic patterns representing a set of objects (Bendaould et al., no year given).

Advantages: This is a very methodological treatment at the higher level concept representation. This methodology is more for building ontology and less on capturing the information from text, but has significant capability.

Disadvantages: Based on the methodology description, the computation could have high complexity.

Valakos et al. used machine learning to build and maintain concept representations called allergens ontology. Building ontologies include: selecting concepts, specifying their attributes and relations (between concepts), and filling (populating) their properties with instances (Valakos et al., 2006).

Advantages: Authors machine learning approach provides a way to capture new knowledge in the form of concepts, attributes, properties, and relations. They maintain (or update) the knowledge with what has been established. The approach includes lexical to semantic relations to transform lexical to semantic information which is a contribution toward proving concepts.

Disadvantages: Details about extraction of the information to form the concepts is not presented. The approach is specific to maintaining ontology within a medical (allergen) domain but its general principles could be applied to other applications.

Zhou and Su use machine learning to integrate evidence from internal (within the word) and external (context) to formulate named entity recognition. This method extracts and classifies text elements into predefined categories of information (Zhou and Su, 2005).

Advantages: This named entity recognition approach provides significant and useful detail that could be applied to information extraction from text. Machine learning is applied to recognizing named entities and is used with constraint recognition, Hidden Markov Models to determine tags, and mutual information to increase coverage of non-redundant information.

Disadvantages: This concept provides significant capabilities on the theoretical level but appears to need further development before product information with metrics is available.

Shunsfard and Barforoush propose an automatic ontology building approach, starting with a small

ontology kernel and implement text understanding to construct the ontology. Their model can handle multiple viewpoints, flexible to domain changes, and can build ontology from scratch without a large knowledgebase (Shunsfard and Barforoush, 2004).

Advantages: This system can create ontology from scratch by learning from text. This significantly reduces manual interaction to create and build ontology. This methodology is based on an integration of learning, clustering and splitting of concepts, similarity measures, and several other techniques that, together, form a unique capability that shows promise.

Disadvantages: The current implementation and testing has been limited to Persian text, but the authors plan to expand the system to other languages.

Hahn and Marko form concepts from text through machine learning of both grammars and ontologies and use evidence, or background knowledge, to steer refinement of generated text. This methodology is an integrated approach for learning lexical (syntactic) and conceptual knowledge as it is applied to natural text understanding (Hahn and Marko, 2002).

Advantages: Evidence within both lexical and conceptual hypotheses is used together to bound the resulting number of hypothesis search space to a manageable quantity. This refines the lexical and conceptual quality, thus increasing the accuracy of text understanding.

Disadvantages: Complexity of the approach can be extensive but tractable.

Loh et al. provides a text mining approach to form concepts from phrases and analyzes their distributios throughout a document. The approach combines categorization to identify concepts within text and mining to discover patterns by analyzing and relating concept distributions in a collection (Loh et al., 2003).

Advantages: This approach captures concepts from phrases, finds patterns from concept distributions, and discovers themes within a document by collecting concepts and generating centroids to represent the collections. Together, these features contribute to a knowledge discovery technique.

Disadvantages: This approach was developed for decision support systems and may have some features dedicated to that application.

Rajaraman and Tan constructed a conceptual knowledge base, called a concept frame graph, for mining concepts from text. A learning algorithm constructs the concept map which is guided by the user via supervised learning (Rajaraman and Tan, 2002).

Advantages: The approach captures conceptual knowledge from text by constructing a concept map to produce a knowledge base. This provides a high level representation including concepts, relations to other concepts, and relations to synonyms. Such representations can be used to reduce redundancy at the high, concept level. A clustering algorithm discovers word sense to reduce ambiguous words.

Disadvantages: The supervised learning in this approach may not be useful for applications requiring automatic (unsupervised) learning. This word sense disambiguation depends on a Wordnet tool, which may include some implementation dependency within the approach.

Pado and Lapata propose a general framework for semantic models that determines context in terms of semantic relations. Their algorithm constructs semantic space models from text annotated with syntactic dependency relations to provide a representation that contains significant linguistic information (Pado and Lapata, 2007).

Advantages: This methodology operates at the semantic level and finds context in terms of semantic relations and contains significant linguistic information. The authors state that their model provides a linear runtime performance. A GNU website is provided for a Java implementation of the general framework for semantic models.

Disadvantages: This proposed methodology will need time to mature after implementation.

Maedche and Staab present a generic architecture for ontology learning which consists of components: ontology management (browse, validate, modify, version, evolve), resource processing (discover, import, analyze, transform input data), algorithm library, and coordination (interaction with ontology learning components for resource sharing and algorithm library access) (Maedche and Staab, 2004).

Advantages: The methodology finds semantic patterns and structures and concept pairs.

Disadvantages: As a new methodology, it will require time to mature into a product.

Dahab et al. discuss a methodology for constructing ontology from natural domain text using a semantic pattern-based approach. Their “TextOntoEx” tool

extracts candidate relations from text and maps them to meaning representations to help construct an ontology representation (Dahab et al., 2008).

Advantages: Provides semantic pattern formats for converted paragraphs.

Disadvantages: Manual editing is required for the library of semantic patterns.

2.8 Redundancy Synthesis

Bourbakis et al. presents a methodology for retrieving multimedia web documents and removal of redundant information from text and images Bourbakis et al. (1999).

Advantages: Out of the papers surveyed, this is the only methodology that provides an integrated similarity detection and redundancy removal of both paragraphs of text and corresponding images. This approach is also integrated with the authors' developed query language that includes Webpage (text) and image similarity criteria to yield increased definitive returns closer to the user's intended query.

Disadvantages: Since the time of the article, other authors have created new features for similarity detection. More text reduction opportunities should be possible with some of the newer features various authors have created. Counts and histograms of text components can detect paragraph similarities up to a certain point. By using approaches similar to this as a baseline, future developments in capturing the meaning from multiple documents should advance similarity detection, resulting in less text redundancy in the synthesized document.

Yang and Wang (2008), apply the hierarchical and redundancy sharing characteristics of fractal theory to increase the performance of text summarization when compared to non-hierarchical approaches (Yang and Wang, 2008).

Advantages: This hierarchical approach to summarization provides multiple levels of abstraction and takes advantage of fractal theory capabilities in representing multiple levels of hierarchy.

Disadvantages: More salient features could be added to make this approach more accurate.

Hilberg proposes an approach to produce and store higher levels of abstraction that represent sequences of words, and sentences in the higher (hidden) levels of a neural net (Hilberg, 1997).

Advantages: This proposal has some unique possibilities for representing abstraction and possibly extending it to paragraphs and documents.

Disadvantages: Getting this to work at a large enough scale (such as large or multi document) may be challenging. The learning of representative corpus of text may be computationally hard to make it work on a large enough scale to get beyond the prototype stage.

2.9 Generating Semantically Meaningful Text through Coherence and Logical Order

Barzilay and Lapata, by representing and measuring local coherence, provide a framework to increase readability and semantic meaning to automatically generated sentences such as a summary of multiple documents. The goal is to order sentences in a way that maximizes local coherence (Barzilay and Lapata, 2008).

Advantages: This methodology provides a needed capability to make generated text more coherent and readable. This entity distribution approach provides significant improvement in sentence meaning representation which can result in improved, automatically generated text. Results of testing showed increased accuracy.

Disadvantages: New approaches like this will need time to mature, but the benefits should be significant.

Stein et al. provide a methodology that clusters documents, uses extraction to find main topics and organizes the resulting information for a logical presentation of a summary of multiple documents. This is an interactive approach that focuses on summarizing news line documents (reducing text to 15%) (Stein et al., 2000).

Advantages: This methodology both summarizes multi-document text and is designed to provide a smooth flow of the summary to the reader. It clusters single document representative summaries with similar topics to reduce redundancy. It orders the generated summary for multiple documents based on paragraph similarity to minimize the jerkiness of topic changes from paragraph to paragraph. The result is improved readability.

Disadvantages: The multi-document summarizer currently uses simple similarity scoring approaches but plans to replace them with better performing ones.

Nomoto and Matsumoto provide a method to exploit diversity of concepts in text in order to evaluate information based on how well source documents are represented in automatically generated summaries (Nomoto and Matsumoto, 2003).

Advantages: This approach provides an improvement in clustering on the information level. The paper provides detailed analysis of its approach verses other traditional approaches and favorable test results including a favorable comparison with human summarization.

Disadvantages: Disadvantages have yet to be found. The authors present this approach as novel, at least in the 2003 timeframe.

Marco et al. improved reading order of automatically generated text. The approach is implemented in a system and is designed to analyze heterogeneous documents (Marco et al., 2002).

Advantages: This approach is implemented in a system that captures the physical and logical layout of generic documents.

Disadvantages: Most of the discussion focuses on the physical portions of a document and the reading order considers large chunks of what is on a page of a document. It applies more to the big (mostly physical) view of a document, little toward the actual knowledge or understanding level.

2.10 Text Generation Methodologies

Dalianis uses aggregation before generating text to eliminate redundant text in documents before they can be paraphrased (generated) into natural language. This methodology provides aggregation at the syntax level (Dalianis, 1999).

Advantages: This approach provides four types of aggregation with rules which should provide more information for generating significantly less redundant summaries.

Disadvantages: An update to this paper could provide a more accurate indicator for the state of this methodology.

2.11 Document Processing & Understanding

Aiello et al. presents a methodology to capture the structural layout and logical order of text blocks within several documents and represents this information in connected graphs (Aiello et al., 2002)

Advantages: This document level methodology captures physical layout of partitioned text blocks spanning over multiple documents with a complexity of O (n^4).

Disadvantages: This only provides top level information about a set of documents. Without being used in conjunction with other methodologies discussed in this survey, the information provided does not include information from within text blocks. Information within text blocks is a needed feature addressed by other methodologies.

2.12 Other Relevant Methodologies

Feldman et al. describes a natural language processing (NLP) system, called the LitMiner system, that uses semantic analysis to mine biomedical literature (Feldman et al., 2003).

Advantages: Although this paper addresses the biomedical domain, it is quite useful in providing the various steps and different methodologies for text mining, plus describing in detail the specific system with good evaluation results for this type of system.

Disadvantages: Several of the key elements are interlinked with the biomedical domain. However, several of the methodologies presented appear to be applicable to several domains. (Different data bases and tools would be needed.) The system described requires some pre-processing and is a semi-automatic process with a visualization system.

Neustein uses sequence analysis to improve natural language understanding from conversations. A goal of this analysis of sequence packages (or frames) of speech is to uncover important information that might otherwise get unnoticed (Neustein, 2001).

Advantages: The proposed sequence analysis would address context dependency in natural language, especially in speech context. Success in this kind of analysis should provide benefits toward reducing ambiguity in natural language processing and understanding.

Disadvantages: This discussion is basically a proposed approach to a difficult problem area and didn't appear to be implemented at the time the paper was written. Little details of the approach were presented at the time this paper was published.

Capability Definitions 1.

C1	Topics	C6	Context	C11	Chains (Lex, Sem, Con)	C16	Statistical
C2	Concepts	C7	Aggregation	C12	Hierarchical	C17	Word Sense
C3	Relations	C8	Overlap	C13	Learning	C18	Large Document
C4	Semantic	C9	Clusters	C14	Detect Themes	C19	Multi-Document
C5	Hierarchical	C10	Quarry	C15	Answer Evaluation		

Table 1: Comparing Key Capabilities/Approaches in Survey.

Authors	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19
Aiello																			√
Barzilay																			
Bendaould		√		√								√							
Bourbakis1	√	√	√	√		√				√						√		√	√
Bourbakis2	√	√	√	√		√				√						√		√	√
Cimiano		√			√	√			√										
Dahab		√	√	√															
Dalianis																			
Feldman																			
Guo									√		√								
Hahn-1													√						
Hilberg																			
Ko						√	√			√						√			
Liddy																			
Loh																			
Manabu	√	√		√					√		√								
Marco																			
Meadche																			
Moens	√				√				√							√			
Neustein																			
Nomoto																			
Pado			√	√		√													
Radev	√								√							√			√
Rajaraman		√											√						
Reeve		√		√		√					√			√					
Shunsfard													√						
Silber									√		√						√	√	
Stein																			
Valakos		√	√										√						
Yang												√		√					
Ye												√			√				√
Yeh				√				√	√							√			√
Zhou																			
Union All	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√

3 COMPARATIVE TABLE OF METHODOLOGIES AND APPROACHES

The following table captures some of the main features and approaches over a global comparison of papers throughout the survey. The intent of this comparison is to provide a collective picture of what main capabilities exist from the papers in this survey.

The above table shows capabilities from various approaches. Intuitively, as more pertinent information is captured, higher quality (minimal redundancy and maximum information coverage) should result. However, most of the performance qualities are not addressed. This may be due to the overall maturity of the technical area which is currently striving for accuracy as measured in the Document Understanding Conferences (DUC) that some of the authors reference. Performance time characteristics, other than computational complexity, appear to be a future effort.

4 CONCLUSIONS

This survey revealed very little commonality among the methodologies that were found. However, the methodologies were able to be categorized into some general headings. The papers covered in the survey did not include enough maturity information that could be used for comparison. A resulting conclusion suggests that this area of natural language processing has not matured enough to provide this kind of product information.

Methodologies that were tested provided precision and recall results and some included complexity. Most were theoretical. According to a definition found on the Oracle web site, precision measures how well non-relevant information is screened (not returned), and recall measures how well the information sought is found.

A few of the most capable methodologies show promise in providing an approximately optimized, minimum redundancy with maximum information coverage. However, more research needs to be performed in natural language understanding before maturity of these methodologies can transform into high volume, commercial products. Normally, providing the more capability to produce accurate text comes with a computational (time and space) complexity price, especially when heuristics are involved. Some of the concept graphical approaches, chain, meta-chains, and hierarchical approaches provided impressive opportunities to compress and optimize resulting text. Finding an efficient methodology to accomplish all this would be a significant step toward eventual technical maturity.

REFERENCES

Aiello, M., Monz, C., Todoran, L., Worring, M., 2002. Document understanding for a broad class of documents, *Int. Journal on Document Analysis Recognition*.

Barzilay, R., Lapata, Mirella, 2008. Modeling Local Coherence: An Entity-Based Approach, *Association for Comput Linguistics*, pages 34.

Bendaould, R., Hacene, M.R., Toussaint, Y., Delecroix, B., Napoli, A., Text-based ontology construction using relational concept analysis, (http://simbad.u-strasbg.fr/simbad/sim-fid)

Bourbakis, N., Manaris, R., 1998. An SPN based Methodology for Document Understanding, *IEEE International Conference on Tools for Artificial Intelligence*, Tapei, Taiwan, pages 10-15.

Bourbakis, N., Meng, W., Zhang, C., Wu, Z., Salerno, N. J., Borek, S., 1999. Removal of Multimedia Web Documents and Removal of Redundant Information, *International Journal on Artificial Intelligence Tools (IJALT),* Vol. 8, No. 1, pages 19-42, World Scientific Pubs.

Cimiano, P., Hotho, A. Staab, S., 2005. Learning Concept Hierarchies from Text Corpora using Formal Concept Analysis, *Journal of Artificial Intelligence Research*, Vol. 24, pages 305-339.

Dahab, M. D., Hassan, H. A., Rafea, A., 2008. TextOntoEx: Automatic ontology construction from natural English text, *Expert Systems with Applications*, Vol. 34, pages 1474-1480.

Dalianis, H., 1999. Aggregation in Natural Language Generation, *Computational Intelligence*, Vol. 15, No. 4, pages 31.

Feldman, R., Regev, Y., Hurvitz, E., Finkelstein-Landau, M., 2003. Mining the biomedical literature using semantic analysis and natural language processing techniques, *BIOSILICO*, Vol. 1, No. 2, pages 12.

Guo, Yi, Stylios, G., 2005. An intelligent summarization system based on cognitive psychology, *Information Sciences* 174, pages 1-36.

Hahn, Udo, Marko, K. G., 2002. An integrated, dual learner for grammars and ontologies, *Data & Knowledge Engineering*, Vol.42, p 273-291.

Hilberg, W., 1997. Neural networks in higher levels of abstraction, *Biological Cybernetics*, 76, pp. 23-40.

Ko, Y., Seo, J., 2008. An effective sentence-extraction technique using contextual information and statistical approaches for text summarization, *Pattern Recognition Letters* 29, p 1366-1371.

Loh, S., De Oliveria, J, Gameiro, Mauricio, 2003. Knowledge Discovery in Texts for Constructing

Decision Support Systems, *Applied Intelligence,* 18, pp. 357-366.

Manabu, O., Hajime, M., 2000. Query-Based Summarization Based On Lexical Chaining, *Computational Intelligence*, Vol. 16, 4,), pp. 8.

Marco, A., Monz, C., Todoran, L., Worring, M., 2002. Document understanding for a broad class of documents, *International journal on Document Analysis and Recognition*, Vol. 5, pages 1-16.

Meadche, A., Staab, S., 2004. Ontology Learning, *Handbook on Ontologies*, Pages 18.

Moens, M.F, Angheluta, R., Dumortier J., 2005. Generic technologies for single- and multi-documents summarization, *Information Processing and Management*, Vol. 41, pages 569-586.

Neustein, A., 2001. Using Sequence Package Analysis to Improve Natural Language Understanding, *Int. Journal of Speech Technology*, Vol. 4, pages 31-44.

Nomoto, T, Matsumoto, Y, 2003. The diversity-based approach to open-domain text summarization, *Information Processing and Management*, 39 pages 363-389.

Pado, S., Lapata, M., 2007. Dependeny-Based Constuction of Semantic Space Models, *Association for Computational Linguistics,* pages 40.

Radev, D. R., Jing, H., Stys, M. Tam, D., 2004. Centroid-based summarization of multiple documents, *Information Processing and Management*, 40, pages 919-938.

Rajaraman, K, Tan, 2002. A-H, Knowledge Discovery from Texts: A Concept Frame Graph Approach, *CIKM* 2002, pages 3.

Reeve, L., Han, H., Brool, A.D., 2006. BioChain: Lexical Chaining Methods for Biomedical Text Summarization, *SAC* 2006, ACM, pages 5.

Shunsfard, M.,et.al., 2004. Learning ontologies from natural language texts, *International J. Human-Computer Studies*, 60, pages 17-63.

Silber, H. G., McCoy, K., 2002. Efficiently Computed Lexical Chains as an Intermediate Representation for Automatic Text Summarization, *Association for Computational Linguistics*, pages 10.

Stein, C.S., Strzalkowski, T., Wise, G.B., 2000. Interactive, Text-Based Summarization of Multiple Documents, C*omputational Intelligence*, Vol. 16, Nov. 4, pp.8.

Valakos, A.G., Karkaletsis, V., Alexopoulou, D. Papadimitriou, E., Spyropoulos, C.D., Vouros, G., 2006. Building an Allergens Ontology and Maintaining it using Machine Learning Techniques, *Computers in Biology and Medicine Journal*, pages 32.

Yang, C. C., Wang, F. L., 2008. Hierarchical Summarization of Large Documents, *Journal of the American Society for Information Science and Technology*, Vol. 59, Num. 6, pages 887-902.

Ye, Shiren, Chua, T-S, Kan, M-Y., Qiu, L., 2007. Document concept lattice for text understanding and summarization, *Information Processing and Management*, Vol. 43, pages 1643-1662.

Yeh, J-Y., Ke, H-R., Y, W-P, Meng, I-H., 2005. Text summarization using a trainable summarizer and latent semantic analysis, *Information Processing and Management*, Vol. 41, pages 75-95.

Zhou, G., Su, J., 2005. Machine learning-based named entity recognition via effective integration of various evidences, *Natural Language Engineering*, Vol. 11, No. 2, pages 189-206.