International Journal For Technological Research In Engineering
Volume 4, Issue 9, May-2017 *ISSN (Online): 2347 - 4718*

BIO-MEDICAL MINING AND SUMMARIZATION USING KEA-MEANS AND LEXICAL ANALYSIS

Himresh Chaudhary[1], Aakriti Sharma[2]
[1]M.Tech Scholar, [2]Assistant Professor
Computer Science, Swami Keshvanand Institute of Technology Management and Gramothan Jaipur

Abstract: Medical Documents are of crucial importance for all. And the documents related to medical analysis are too lengthy and if these documents get summarized then easy going analysis can be done. If the medical reports can be summarized it will save a lot of time of the doctor so it will also help doctor to see more number of patients. Summarizing the documents also involves the gathering the related lines and portion of the document, for which we have taken the concept of the Kea-Means for the clustering of the medical documents. We have used the concept of the WordNet Library to hold as the major dictionary for us for validating the words in the documents and using the concept of the lexical chains we have formulated the summary after evaluating the significance and the utility of the lexical chains. And for validating the creditability of the summary of the medical documents generated we have compared it with manually summarized medical documents to calculate the matching percentage known as recall.
Keywords: Clustering, Kea-means, Word Net, Medical Documents, Summarization

I. INTRODUCTION

Lexical chains can be used to model lexical cohesion in documents. A topic can be expressed within a representation formed of words contributing to the topic presentation. When we read a document, we immediately interpret the correct senses of words in that document. Meaning of each word seen in the document contributes to a topic. Lexical chains are sets of word senses that are related with each other. Let a document D be formed of word occurrences {w1...wi... wn}. These n words are only symbolic representations, meaning of the word can only be determined from the text with prior knowledge. Each word can have more than one sense. For example, word 'bank ' has 10 different senses defined in WordNet. A lexical chain in D is a set of word senses {ws32 , ws61 , ws410 , ws102 }, where wsij is the j'th sense of the word wi. The goal of a document clustering theme is to reduce intra-cluster distances between documents, whereas maximizing inter-cluster distances (using an acceptable distance live between documents). A distance live (or, dually, similarity measure) so lies at the center of document clustering. The big type of documents makes it nearly not possible to make a general algorithmic program which might work best just in case of every kind of datasets. Although clustering will be applied to several sorts of knowledge, the main target of this dissertation is on cluster text documents, a field noted within the literature as document clustering that may be a subfield of text mining. Document bunch deals with the unsupervised partitioning of a document assortment into meaningful teams supported their matter content, sometimes for the aim of topic categorization; i.e. documents in one cluster belong to a particular topic, whereas {different totally completely different completely different} clusters represent different topics. in contrast to document classification – that may be a supervised learning technique that needs previous information of document classes to coach a classifier, document bunch is an unsupervised learning technique that doesn't believe previous categorization information. Document bunch has several applications, like bunch of program results to gift organized and perceivable results to the user (e.g. Vivisimo1), bunch documents during a assortment (e.g. digital libraries), machine-driven (or semi-automated) creation of document taxonomies (e.g. Yahoo! and Open Directory styles), and economical data retrieval by that specialize in relevant subsets (clusters) instead of whole collections. News aggregation is turning into a typical application of document bunch, exemplified by the Google News service that uses document bunch to cluster news articles from multiple news sources, providing an automatic compilation of recent news. In this section we tend to review document bunch. We tend to begin by reviewing bunch algorithms, their quality and their shortcomings. We tend to then illustrate the restrictions of each algorithmic program by providing graphical examples showing why they could fail (i.e., create clusters that are "intuitively" wrong). Finally, we tend to discuss document-clustering analysis in Information Retrieval.

II. RELATED WORK

According to the paper "Efficient Text Summarization Using Lexical Chains,H. Gregory Silber,Kathleen F. McCoy" [9]
The increased in the growth of the net has resulted in huge amounts of information that has become tougher to access with efficiency. Web users need tools to manage this immense amount of information. The main goal of this analysis is to form an economical and effective tool that's able to summarize quite large documents quickly. This analysis presents a linear time algorithmic rule for finding out lexical chains that could be a technique of capturing the "aboutness" of a document. This technique is compared to previous, less efficient strategies of lexical chain extraction. They additionally give different strategies for extracting and evaluation lexical chains. They show that their technique provides similar results to previous analysis, however is considerably quite more efficient. This efficiency is important in web search applications where several quite large documents might have to be summarized promptly, and where the reaction time to the end user is very vital.

This initial part of their implementation constructs an array of "meta chains" [9]. Every Meta chain contains a score and a data structure that encapsulates the meta-chain. The score is computed as every word is inserted into the chain. Whereas the implementation creates a flat illustration of the source text, all interpretations of the source text are implicit among the structure. Every line represents a semantic association [9] between 2 word senses. Every set of connected dots and lines represents a meta-chain. The gray ovals represent the list of chains to that a word will belong. The dashed box indicates the strongest chain in their illustration. Notice that in some senses of the word machine, it's semantically like friend, whereas in different senses, it's semantically like computer (i.e. within the same meta-chain). The algorithmic rule continues by making an attempt to search out the "best" interpretation from among their flat illustration. They consider the illustration as a group of transitively closed graphs whose vertices are shared. In figure, the sets of lines and dots represent 5 such graphs. The set of dots among an oval represent a single shared node. That's to mention, that whereas 2 of those graphs could share a node, the individual graphs aren't connected. The " best" interpretation are going to be the set of graphs that may be created from the initial set mentioned above, by deleting nodes from every of the graphs in order that no 2 graphs share a node, and also the overall " score" [9] of all the meta-chains is largest. According to paper "Using Lexical Chains for Text Summarization, Regina Barzilay and Michael Elhadad," They investigate one technique to supply a summary of an original text while not requiring its full semantic interpretation [11], however instead hoping on a model of the topic progression within the text derived from lexical chains. They present a new algorithmic program to find out lexical chains in a text, merging many robust knowledge sources: the WordNet thesaurus, a part-of-speech tagger, shallow parser for the identification of nominal teams, and a segmentation algorithmic program. Summarization is carried out in four steps: the initial step is, text is segmented, lexical chains are made, strong chains are marked or identified and vital sentences are extracted. They present in this paper empirical results on the identification of strong chains [11] and of important sentences. Preliminary results indicate that quality indicative summaries are made. Unfinished issues are then identified. Plans to deal with these short-comings are concisely presented. Summarization is the method of reducing a source text into a shorter version conserving its data content. It will serve many goals — from survey analysis of a scientific field to fast indicative notes on the general topic of a text. Resulting a high quality informative summary of an arbitrary text remains a challenge which needs full understanding of the text. Indicative summaries, which may be used to quickly decide whether a text is worth full for reading, are naturally easier to provide. In this paper they investigate a technique for the creation of such indicative summaries from arbitrary text [11]. Sparck Jones (Jones 1993) [11] describes summarization as a two-step process:

1. Building from the source text a source representation;
2. Summary generation — forming a summary representation from the source illustration built in the primary step and synthesizing the output summary text.

According to paper "Comparative Study of Text Summarization Methods, NikitaMunot and Sharvari S. Govilkar,2014" Text summarization is among one application of natural language processing and is now becoming much common for info condensation. Text summarization could be a method of reducing the size of original document and results a summary by holding necessary info of original document. This paper provides comparative study of varied text summarization strategies based on differing kinds of application. The paper discusses well 2 main classes of text summarization strategies these are extractive and abstractive summarization strategies [12]. The paper conjointly presents taxonomy of summarization systems and statistical and linguistic approaches [12] for summarization. Natural language processing (NLP) could be a field of computer science, artificial intelligence and linguistics involved with the interactions between computers and human language. Natural language processing could be a method of developing a system process and results language pretty much as good as human can turn out. The utilization of World Wide Web has increased and then the problem of info overload conjointly has increased. Therefore there's a requirement of a system that automatically retrieves, categorize and summarize the document as per users requirement. Document summarization is one attainable solution to the present problem.

III. PROPOSED CONCEPT

KEA is associate degree algorithmic program for extracting key phrases from text documents. It is either used free assortment with a controlled vocabulary. Kea associate degree algorithmic program for mechanically extracting key phrases from text. Parrot identifies candidate key phrases victimization lexical strategies, calculates feature values for every candidate, and uses a machine- learning algorithmic program to predict that candidates are smart key phrases. Additionally, key phrases will facilitate users get plan concerning the content of a group, offer smart entry points into it. Within the specific domain of key phrases, there are 2 essentially completely different approaches: key phrase assignment and key phrase extraction. Key phrase assignment seeks to pick out the phrases from a controlled vocabulary that best describe a document.

1. Input the text file.
2. Extract the text file line by line.
3. Perform tokenization i.e. remove ',','; ', '.' And replace them by space.
4. Split the line on the basis of space to form array of word.
5. Remove stop word from array.
6. Perform stemming of words in the array to get the base form of each work. (Using Word Net).
7. Perform POS tagging to identify the verb, adverb, Noun, and pronoun using MAX Tagger.
8. Now we will form the words array containing only Noun and Proper Noun.

9. Now we will find unique word and their count from the above word array.
10. From the lexical chains (Synonyms, Hyponyms, Antonyms, Hyponyms) using Word Net API and RTI Word Net.
11. Find each chain and its chain length which is the number of words in the chain.
12. Now we will calculate significance of each chain using formula mentioned in the DOC.
13. (chain length/sum)*(Log(chain length / sum) / Log2)
14. Sum= Sum of all chain length.
15. Formula is used are in general for text summarization and referred "Dinkar paper".
16. Calculate utility of each chain.
17. Utility= significance of chain * chain length
18. Calculate the threshold value which is (sum of all chain utility/ total*2).
19. Find accepted chains which are greater than or equal to threshold value.
20. Now we will gather all the words in the accepted chains and we will find all files containing the words in the accepted chains and calculate frequency chart i.e. the line index and no. of words it contains (Match from accepted chains words) and arrange in descending order.
21. Fetched percentage of lines for summary generation.
22. And match will original summary to general recall.

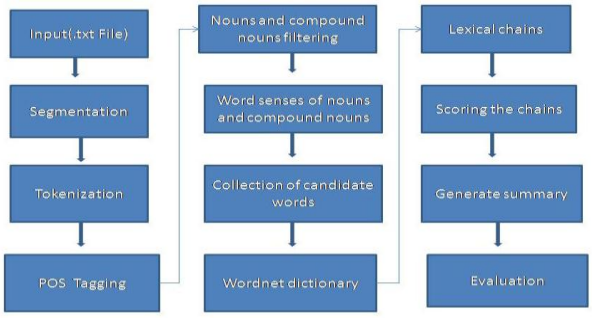

Fig 1. Work Flow Diagram

IV. IMPLEMENTATION

I am using Eclipse Java EE (Enterprise Edition) IDE (Integrated Development Environment) for Web Developers Version: Kepler Service Release 1 Build id: 20130919-0819 in my Dissertation.

Run the BIO Medical Using WorldNet of Proposed Work

STEP- 1 Open the file BioMed.java in eclipse.

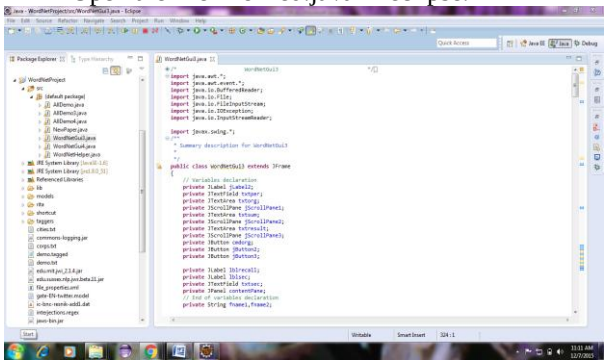

Fig 2. Eclipse IDE

STEP-2 Run the Program: Go to the toolbar→Run option →Run as Option → Select java application.

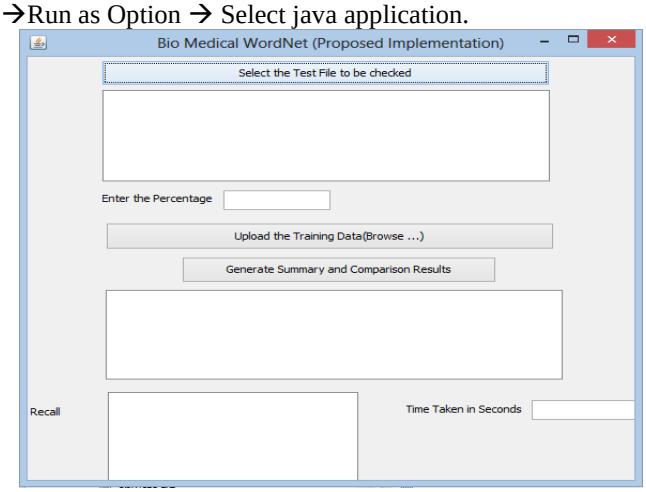

Fig 3. Proposed Main Screen

STEP-3 Select the test medical document

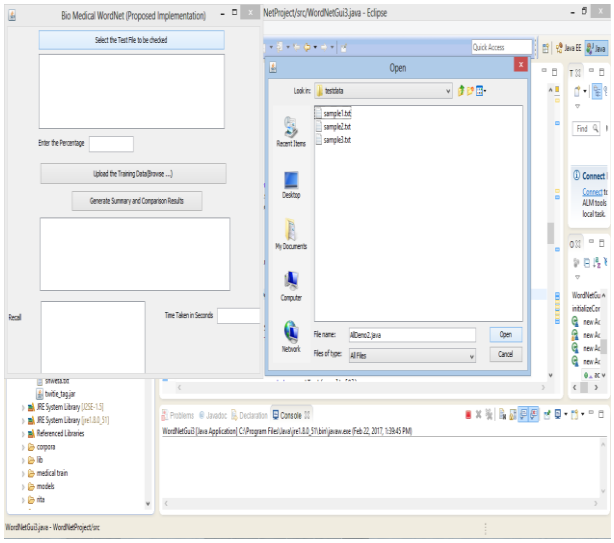

Fig 4. Open the Test File

STEP-4 Open the Sample Data

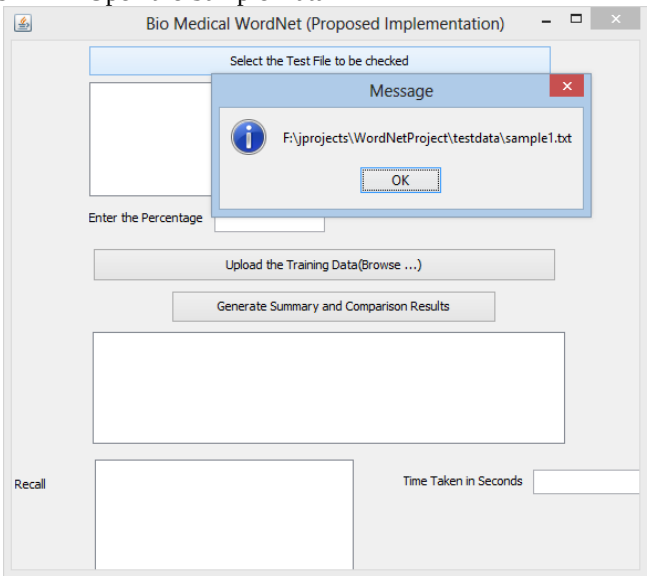

Fig 5. Open the Test File Message

STEP-5 Enter the Percentage of Summary (Sample1.txt)

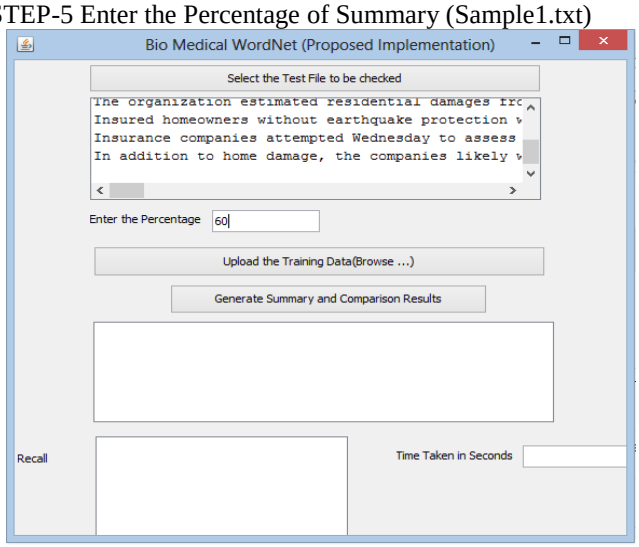

Fig 6.. Precentage

STEP-6 Upload the standard medical document

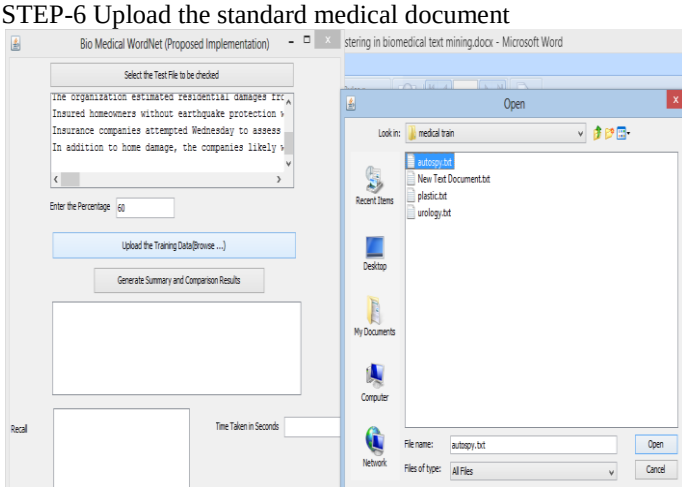

Fig 7. Standard Document

STEP- 7 Generate the Summary & Compare the Results

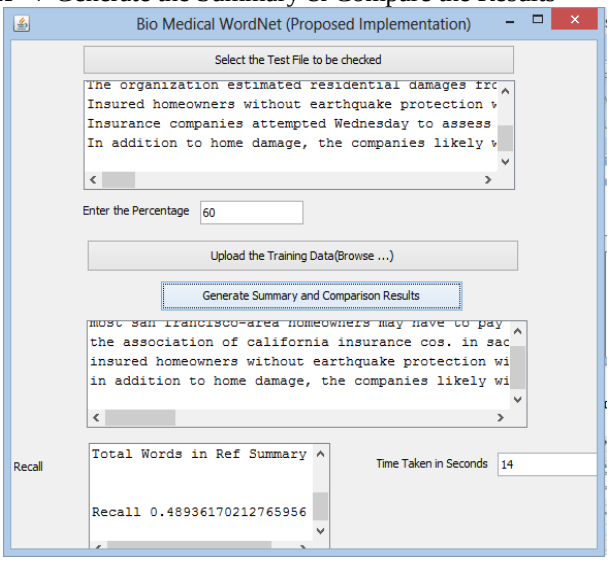

Fig 8. Summary Result

V. TEST RESULT ANALYSIS

In this we have taken the Sample Text document and the master document form the same to generate the recall. And together we have created the base paper algorithm implementation for the comparison on the basis of the recall and the time taken to complete the overall process.

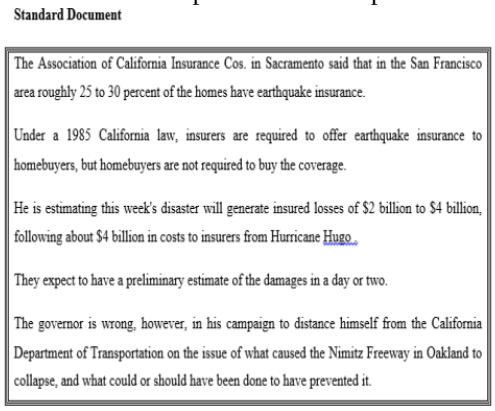

Fig 9. Standard Document

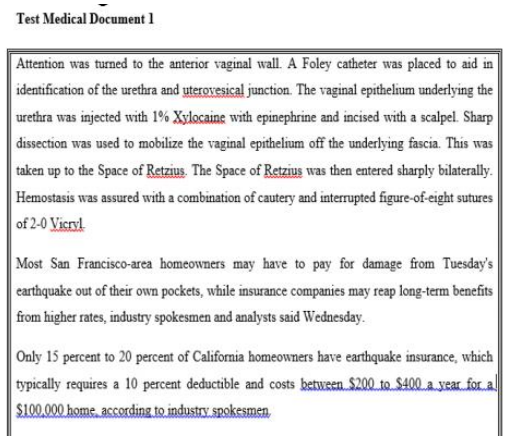

Fig 10. Test Document

Result:
Proposed Approach
Precision: .489 Times Taken: 14 sec
Base Approach
Precision: .475 Time Taken; 55 sec

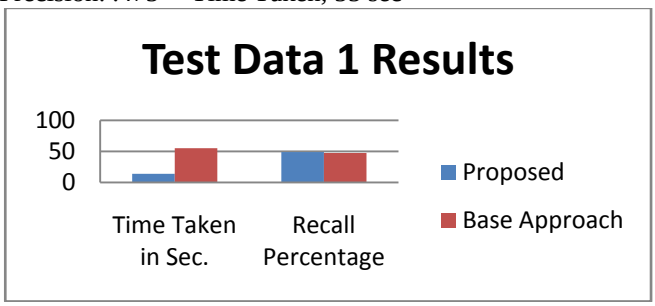

Fig 5.1 Graphical Result Representation Test Data 1

VI. CONCLUSION

Document clustering is being studied from several decades however still it is removed from a trivial and resolved downside.

The challenges are:

1. Choosing applicable options of the documents that ought

to be used for clustering.
2. Choosing an applicable similarity live between documents.
3. Choosing an applicable bunch methodology utilizing the higher than similarity live.
4. Implementing the clustering algorithmic program in an economical method that creates it possible in terms of needed memory and mainframe resources.
5. Finding ways that of assessing the standard of the performed bunch.

Document clustering has initial been investigated in info Retrieval chiefly as a method of rising the performance of search engines by pre-clustering the whole corpus.

The kea-means clustering algorithmic program is our planned new clustering methodology that improves the K-means algorithmic program by combining it with the Kea key phrase extraction algorithmic program. The Kea-means clustering tries to resolve the most downside of K-means that the amount of total clusters is pre-specified earlier. In Kea-means algorithmic program, documents square measure clustered into many teams like K-means; however the amount of clusters is set mechanically by the algorithmic program heuristically by mistreatment the extracted key phrases.

In our work, we have used the standard library WordNet for generating the lexical chains and find the summary of the medical documents and comparing the summary for documents with the manually generated summary and calculating and evaluating the recall.

REFERENCES

[1] Martin Ester, Hans-Peter Kriegel, Jrg Sander, and XiaoweiXu. A density-based algorithm for discovering clusters in large spatial databases with noise. In 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96), pages 226–231, Portland, Oregon, 1996. AAAI Press.

[2] T. Kohonen. Self-Organizing Maps.Springer, Berlin, 1995.

[3] M. Steinbach, G. Karypis, and V. Kumar.A comparison of document clustering techniques. KDD-2000 Workshop on Text Mining, August 2000.

[4] Brian S. Everitt, Sabine Landau, and MorvenLeese.Cluster Analysis. Oxford University Press, fourth edition, 2001.

[5] P. Willet. Recent trends in hierarchical document clustering: A critical review. Information Processing and Management, 24:577-597, 1988.

[6] Jardine, N. and van Rijsbergen, C. J. The use of hierarchical clustering in information retrieval.Information Storage and Retrieval, 7:217-240, 1971.

[7] Salton, G. Cluster search strategies and the optimization of retrieval effectiveness. In Salton, G. (ed), The SMART Retrieval System, Prentice-Hall, Englewood Cliffs, N.J., 223-242, 1971.

[8] Croft, W. B. Organizing and searching large files of documents. Ph.D. Dissertation, University of Cambridge, 1978.

[9] Griffiths, A., Luckhurst, H. C. and Willet, P. Using inter-document similarity information in document retrieval systems. Journal of the American Societyfor Information Science, 37:3-11, 1986.

[10] Van Rijsbergen, C. J. Information Retrieval, Butterworths, London, 1979.

[11] Cutting, D. R., Karger, D. R, Pedersen, J. O. and Tukey, J. W. Scatter/Gather: A cluster-based approach to browsing large document collections. In Proceedings of the 15th International ACM SIGIR Conference on Research andDevelopment in Information Retrieval (SIGIR'92), 318-329, 1992.

[12] Border, A. Z., Glassman, S. C., Manasse, M. S. and Zweig, G. Syntactic clustering of the Web. In Proceedings of the Sixth International Web WideWorld Conference (WWW6), 1997.

[13] Krulwich, B. and Burkey, C. (1996) "Learning user information interests through the extraction of semantically significant phrases." AAAI Spring Symposium on Machine Learning in Information Access, Stanford, CA; March.

[14] Munoz, A. (1996) "Compound key word generation from document databases using a hierarchical clustering ART model." Intelligent Data Analysis, 1 (1).

[15] Steier, A.M. and Belew, R.K. (1993) "Exporting phrases: A statistical analysis of topical language." Proc Symposium on Document Analysis and Information Retrieval, 179-190.

[16] Munoz, A. (1996) "Compound key word generation from document databases using a hierarchical clustering ART model." Intelligent Data Analysis, 1 (1).

[17] Luhn, H.P. (1958). The automatic creation of literature abstracts. I.B.M. Journal of Research and Development, 2 (2), 159-165.

[18] Edmundson, H.P. (1969). New methods in automatic extracting.Journal of them Association for Computing Machinery, 16 (2), 264-285.

[19] Marsh, E., Hamburger, H., and Grishman, R. (1984). A production rule system for message summarization.InAAAI-84, Proceedings of the American Association for Artificial Intelligence, pp. 243-246. Cambridge, MA: AAAI Press/MIT Press.

[20] [20] Paice, C.D. (1990). Constructing literature abstracts by computer: Techniques and prospects. Information Processing and Management, 26 (1), 171-186.

[21] Paice, C.D., and Jones, P.A. (1993). The identification of important concepts in highly structured technical papers. SIGIR-93: Proceedings of the 16th Annual International ACM SIGIR Conference onResearch and Development in Information Retrieval, pp. 69-78, New York: ACM.

[22] Johnson, F.C., Paice, C.D., Black, W.J., and Neal, A.P. (1993). The application of linguistic processing to automatic abstract generation.Journal of Document and Text Management, 1, 215-241.

[23] [23] Salton, G., Allan, J., Buckley, C., and

Singhal, A. (1994). Automatic analysis, theme generation, and summarization of machine-readable texts.Science, 264, 1421-1426.

[24] Salton, G., Allan, J., Buckley, C., and Singhal, A. (1994). Automatic analysis, theme generation, and summarization of machine-readable texts.Science, 264, 1421-1426.

[25] Kupiec, J., Pedersen, J., and Chen, F. (1995). A trainable document summarizer. In E.A. Fox, P. Ingwersen, and R. Fidel, editors, SIGIR-95: Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 68-73, New York: ACM.

[26] Brandow, R., Mitze, K., and Rau, L.R. (1995). The automatic condensation of electronic publications by sentence selection.Information Processing and Management, 31 (5), 675-685.

[27] Jang, D.-H., and Myaeng, S.H. (1997). Development of a document summarization system for effective information services.RIAO 97 Conference Proceedings: Computer-Assisted InformationSearching on Internet, pp. 101-111. Montreal, Canada.

[28] Johnson, F.C., Paice, C.D., Black, W.J., and Neal, A.P. (1993). The application of linguistic processing to automatic abstract generation.Journal of Document and Text Management, 1, 215-241.

[29] Brandow, R., Mitze, K., and Rau, L.R. (1995). The automatic condensation of electronic publicationsby sentence selection.Information Processing and Management, 31 (5), 675-685.

[30] MUC-5. (1993). Proceedings of the Fifth Message Understanding Confer ence.California: Morgan Kaufmann.

[31] Soderland, S., and Lehnert, W. (1994). Wrap-Up: A trainable discourse module for information extraction. Journal of Artificial Intelligence Research, 2, 131-158.

