A TEMPLATE BASED ALGORITHM FOR AUTOMATIC SUMMARIZATION AND DIALOGUE MANAGEMENT FOR TEXT DOCUMENTS

Prashant G. Desai[1], Sarojadevi H[2], Niranjan N. Chiplunkar[3]

[1]*Lecturer, Department of Computer Science & Engineering, N.R.A.M.P., Nitte, Karnataka, India*
[2]*Professor & Head, Department of Computer Science, N.M.A.M.I.T., Nitte, Karnataka, India*
[3]*Principal, N.M.A.M.I.T., Nitte, Karnataka, India*

Abstract

This paper describes an automated approach for extracting significant and useful events from unstructured text. The goal of research is to come out with a methodology which helps in extracting important events such as dates, places, and subjects of interest. It would be also convenient if the methodology helps in presenting the users with a shorter version of the text which contain all non-trivial information. We also discuss implementation of algorithms which exactly does this task, developed by us.

Key Words: *Cosine Similarity, Information, Natural Language, Summarization, Text Mining*

---***---

1. INTRODUCTION

The world is witnessing a large quantity of digital information today. This digital information is available in different ways and from different sources. The World Wide Web has been the dominant source of digital information. This digital information has been presented to user mainly in two formats: structured information and unstructured information. Understanding and processing of structured information is quite easier due to its very basic nature that the information is arranged in a specific format. The same is not true with unstructured information since there is no specific format in which the information is arranged. The information no matter how significant it is, spread across the whole document. Therefore it is a challenging task to understand, process and extract the useful and significant information from unstructured documents such as emails, blogs, and research documents available in doc, pdf and txt formats. The ultimate goal of text mining is to discover useful information and make knowledge out of it.

This paper describes a natural language processing approach for identifying and extracting the important information from the unstructured text.

2. RELATED WORK

The work presented in [6] describes a model for answering the user inputted question, by referring to a large database collection of question-answer sets. The words in the question submitted by the user are expanded using the synonyms of respective words. Researchers have mainly concentrated on obtaining the synonyms of noun, verbs and adjectives. Database search for finding the answer is done after the question is expanded.

Researchers of [18] present a two-phased process for text summarization. In the 1st phase text from original document is analyzed for topic and passages are extracted. In the next phase the extracted text is understood with the help of WordNet and FrameNet based on which text is extracted, joined together before being presented to the user as output.

[23] is a research work for extracting answer from on-line psychological counseling websites. The information available from these websites is used as knowledge base. An index is created for making the search of key words efficient. Using this index based search candidate answers fetched for a user query is generated as response.

Three attributes' information is used for summarizing a document in [1]. Those 3 words are- field, associated terms and attribute grammars. The results of experiments conducted by the authors' show that high level accuracy is attained while implementing their concept.

The work illustrated by [5], formulates a set of rules for summarizing the text document. To start with algorithm analyses the structure of the natural language text. Later, the prescribed rules formulated by the researchers are applied to generate the summary.

[17] discusses a summarization technique based on fuzzy logic. Before applying the fuzzy logic, sentences from the original text document, are selected based on 8 pre-defined rules such as title feature, sentence length, term weight, sentence position, sentence to sentence similarity, proper noun, thematic word, and numerical data. These are implemented as simple IF-THEN rules in an inference engine module of the algorithm.

[9] illustrates the significance of artificial neural networks for finding key phrases from the text. Authors compare the key words extracted from neural networks, naïve-bayes and decision tree algorithms. Key words extracted from all these algorithms are then compared with publicly available key phrase extraction algorithm KEA. Laters authors are of the opinion that, neural networks perform better for extracting keywords.

A text mining approach is discussed in [22] for creating ontology from text belonging to a specific domain. The methodology adopted for building ontology uses data pre-processing, natural language analysis, concept extraction, semantic relation extraction and ontology building.

In the study of [14], it is learnt that quality of automated summary is dependent of key phrases. Therefore, the author focuses on locating high quality key words from the text. The procedure adopted for summarizing is pre-processing, key phrase extraction and then sentence selection.

[16] provides guidelines which are implemented by them for achieving text mining independent of the language of original text. Important points made by the authors' are to formulate the rules independent of the language, minimizing the use of language specific resources, if the use of such resources become inventible, keep them in a modular representation so that adding new ones become easier, for cross-lingual applications, represent the docuSDment in a neutral language to overcome the complexity of dealing with multiple languages.

3. METHODOLOGY

This methodology is implemented in two phases: 1. Text Pre-Processing 2. Information Extraction. Soon after the document is chosen by the user, document goes to file classification module. This module identifies whether the input document is a pdf or MS-Word document before reading the contents.

Phase 1: Text Pre-Processing

This part of the implementation includes the modules shown below:

1.Syntactic analysis
2. Tokenization
3. Semantic analysis
4. Stop word removal
5. Stemming

Phase 2: Information extraction

This part of the implementation includes the modules shown below:

1. Training the dialogue control
2. Knowledge base discovery
3. Dialogue Management
4. Template based summarization
5. Domain intelligence suites

Figure 1 and Figure 2 show the building blocks of our system architecture.

Phase1: Text Pre-Processing

3.1 Syntactic Analysis

The unstructured text does not have any pre-defined format in which data is organized. Hence, it is very much required to decide where a particular piece of data begins and where exactly it ends. The task of syntactic analysis module is exactly this, i.e., to decide beginning and ending of each sentence in a document. Presently we assume that full stop symbol marks the end of sentence. Any string of characters up to full stop symbol is treated as one sentence. We do this with help of java class string tokenizer.

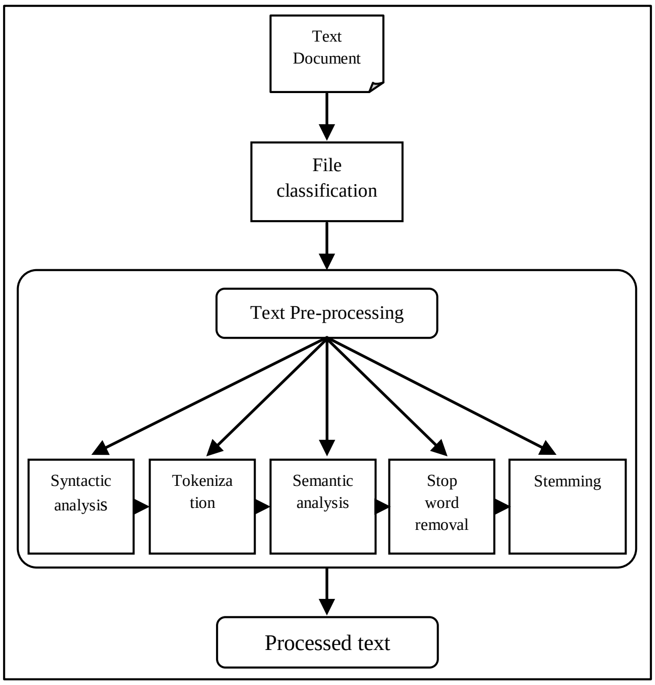


Fig -1: Text Pre-Processing modules

3.2 Tokenization

The task of tokenizer is to break the sentence into tokens. The broken pieces of a sentence may include words, numbers and punctuation marks. This step is very much essential for next level of processing the text. [15].

3.3 Semantic Analysis

Semantic analysis module understands the part of each word played in a sentence. Then it assigns a tag to every word such as noun, verb, adjective, adverb and so on. This process of understanding the part played by each word and then assigning an appropriate tag is called Part-Of-Speech tagging or POS tagging. In order to implement this module, Stanford University [19] POS tagger is used

3.4 Stop Word Removal

Some of the words appear in the natural language text more often but have very little significance when we take over all meaning of the sentence. Such words are termed stop words. These words are removed before the document is considered for next level of processing. [10]

3.5 Stemming

Stemming is the process of finding base form of a certain word in the text document [4, 3]. Any text document may include repetition of same word but in different forms of the grammar such as word being present in past tense, or in present tense and sometimes in present continuous tense and so on. To avoid all these forms of the same considered as different words, stemming is performed. Presently porter stemmer [11] is used for stemming.

The text after going through all these modules is fed into the second phase called Information extraction module.

Phase2: Information extraction
This module takes the pre-processed text for discovering important information using machine learning algorithms designed and implemented by us. The process of the algorithm begins with training the model.

3.6 Training The Dialogue Control

Here the administrator of the module trains the system. During training, the system learns important or index terms, named entities such as name of the persons, places, temporal formats according to rules. These rules are defined and discussed in our previous paper [12]. Intelligence of the algorithm increases with every training. The concepts learnt during training are stored in the knowledge base of the system.

3.7 Knowledge Base Discovery

Knowledge base discovery is the task of creating intelligence from structured (e.g., database, XML, etc.), and unstructured text (e.g., documents, blogs, emails, etc.) [8, 7]. Our research is purely focused on processing unstructured text. Hence, knowledge discovery here means the process of deriving intelligent information and storing from unstructured text.

The knowledge base is capable of storing important terms such as index terms, name of the persons, name of the locations, in an efficient may. By calling the knowledge base design as efficient, we mean that, structure is formed in such a way that all the terms are stored together yet in a distinguishable format. There by reducing the need for creating multiple storage structures for storing different category terms, reducing the search time and thus improving the overall performance of the algorithm.

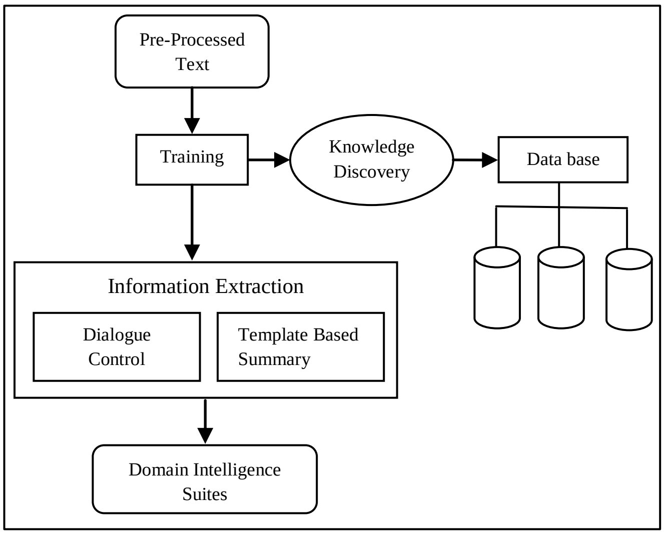


Fig -2:Information Extraction modules

3.8 Dialogue Management

Dialogue is an interaction occurring between two parties: either between 2 humans or between human and machine [21]. This paper discusses implementation of an algorithm for a dialogue between human and machine. The dialogue management module is a program which offers interaction between human and the computer. With this module the end user can request the information using natural language text. The dialogue control module which built upon the training model, accepts the user request, understands, refers the knowledge base and then produces the answers which possibly contain the sought information. Algorithm below illustrates the functioning of dialogue management.

1. Choose the input document
2. User is presented with 2 choices:
    1) Event related options
    2) Writing a customized query
3. With choice-1 events such as important dates, locations, subjects of interest available in the document are extracted.
4. With choice-2 user has the provision for writing a customized query
5. Tokenize the query given in natural language text
6. Identify and extract the index terms (which are also called as index words, cue words, key words and so on)
7. Search the knowledge base using the index terms.
7.1 While searching the answers, algorithm considers the synonyms of every index word.
7.2 Word Net [13] is used for collecting the synonyms of a term.
8. Answers which have the index terms appearing in them are short listed.
9. Then sequence of appearance of index terms in answer and query is matched
10. Answers are presented to user along with score for every answer.
11. Score of the answer is count of the index terms appearing in sequence in which they appear in query.

3.9 Template Based Summarization

Text summarization is the process of putting together meaningful text present in a document in a condensed format. Template based summarization in our system is designed to perform this operation.

Here user has the freedom of choosing what should be present in the summary. In other words, user prepares template based on which summary is generated. This prepared template can include POS tags like Noun, Verb, Adverb, etc, and the sequence in which user wants them to appear in the document. User is not confined to only one pattern of this kind but can have as many patterns as possible. Once the POS patterns are finalized, user can decide whether to include named entities and dates in the summary expected. This completes the step of preparing the template. The template based summary module takes into account all these requisites of the user while it generates summary.

4. EXPERIMENTS AND RESULTS

4.1 Evaluation Metrics

There are many similarity measures to decide the closeness of two text documents [2] such as Cosine similarity, Euclidean distance, Jaccard coefficient, Pearson Correlation

Coefficient, Averaged Kullback-Leibler Divergence. The experimental study carried out by [20] is of the opinion that Cosine similarity is better suited for text documents. This cosine similarity measure is used to evaluate automatically obtained summary. While evaluation metrics such as Precision, Recall, and F-Measure are used to evaluate the dialogue management interface.

4.2 Cosine Similarity

If D = {d1, . . . , dn} is a set of documents and T = {t1, . . . ,tm} the set of unique terms appearing in D. The m-dimensional vector is used represent the document $\overrightarrow{t_d}$. The frequency of term t ∈ T in document d ∈ D is denoted by tf(d, t). The vector of a document d is represented by the equation $\overrightarrow{t_d}$ = (tf(d, t1), . . . , tf(d, tm)). Given two documents 'a' and 'b' whose term vectors are represented by $\overrightarrow{t_a}$ and $\overrightarrow{t_b}$ respectively, their cosine similarity is calculated by the formula

$$SIM_C(\overrightarrow{t_a}, \overrightarrow{t_b}) = \frac{\overrightarrow{t_a} \cdot \overrightarrow{t_b}}{|\overrightarrow{t_a}| \times |\overrightarrow{t_b}|}$$

4.3 Precision

Precision measures the number of correctly identified items as a percentage of the number of items identified. It is formally defined as

$$\text{Precision (P)} = \frac{\left(\text{Correct} + \frac{1}{2}\text{Partial}\right)}{(\text{Correct} + \text{Spurious} + \text{Partial})}$$

4.4 Recall

Recall measures the number of correctly identified items as a percentage of the total number of correct items.

Recall is formally defined as

$$\text{Recall(R)} = \frac{\left(\text{Correct} + \frac{1}{2}\text{Partial}\right)}{(\text{Correct} + \text{Missing} + \text{Partial})}$$

4.5 Measure

The F-measure is often used in conjunction with Precision and Recall, as a weighted average of the two. If P and R are to be given equal weights, then we can use the equation

$$\text{F1=} \frac{(P * R)}{0.5*(\mathrm{P}+\mathrm{R})}$$

4.6 Experimental Setup

Ten text documents belonging mainly to research documents and conference papers (CFP) are chosen for evaluating the results. We manually generated a summary for each of these ten documents. These manually created summaries are also called reference summaries. The reference or manual summaries for the research papers are generated on the basis of contents having important parts-of-speech (POS) such as Noun, Verb, Adverb, Proper nouns and so on. The manual summaries for conference papers are chosen on the basis of text containing important dates, subjects of interests and names of the locations. Templates are prepared for generating automated summary in such a way that, templates contain patterns having sequences of POS which were chosen for preparing manual summaries. This is done mainly to verify the accuracy of the automated summary. That is how humans prepare the manual summary and algorithm prepares the summary with same references. Then we compared the template based summary generated by the system, with the respective manual summaries. In order to compare system summary with the manual summary cosine similarity measure is used, as both of these summaries are stored as text documents. Table 1 displays the results in terms of cosine similarity values.

Table -1: Cosine Similarity Indices

Length	Document Category	Manual Summary Length	System Summary Length	Cosine Similarity Value
168	Technical / Research Document	114	52	0.534
215	CFP	114	246	0.696
221	Technical / Research Document	168	140	0.81
236	Technical / Research Document	156	35	0.56
311	Technical / Research Document	172	226	0.844
325	Technical / Research Document	85	72	0.788
348	CFP	174	295	0.6666
512	Technical / Research Document	265	246	0.785
676	Technical / Research Document	165	379	0.84
1328	Technical / Research Document	151	571	0.656

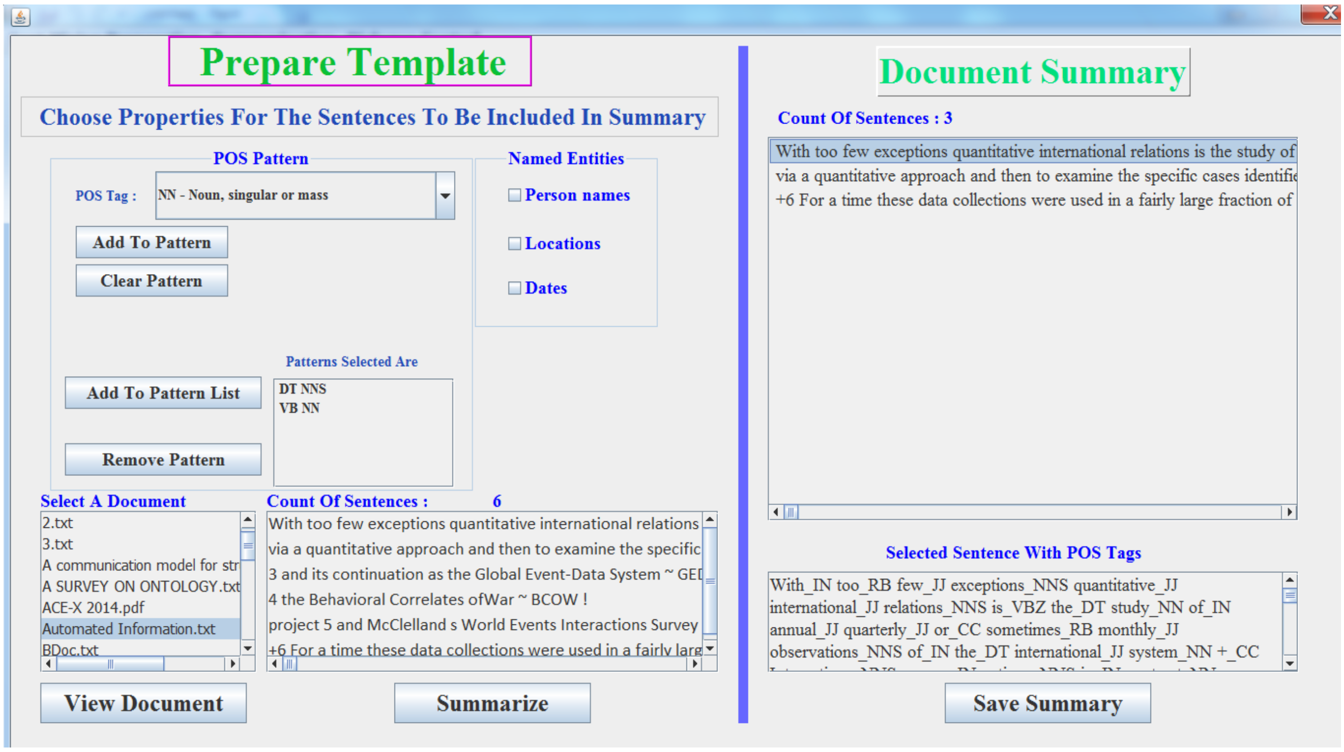
Table 2 illustrates results analysis for the dialogue management interface, tested on 02 technical documents and 02 conference papers from the chosen dataset.

4.7 Screen Shots

Figure 3 shows the template prepared and the summary generated by the system. The template is prepared with two patterns :

1. A “Determiner” POS tag followed “Plural Noun”
2. A “Base form of Verb” POS tag followed by a “Singular Noun”.

The original document has 6 sentences. After processing this original document with reference to the template prepared by the user, system has produces an automated summary which contains only 3 sentences of significance. This is shown in figure 3.



Patterns Selected Are:
DT NNS
VB NN

Select A Document:
2.txt
3.txt
A communication model for str
A SURVEY ON ONTOLOGY.txt
ACE-X 2014.pdf
Automated Information.txt
BDoc.txt

Count Of Sentences : 6

With too few exceptions quantitative international relations
via a quantitative approach and then to examine the specific
3 and its continuation as the Global Event-Data System ~ GEI
4 the Behavioral Correlates ofWar ~ BCOW !
project 5 and McClelland s World Events Interactions Survey
+6 For a time these data collections were used in a fairly larg

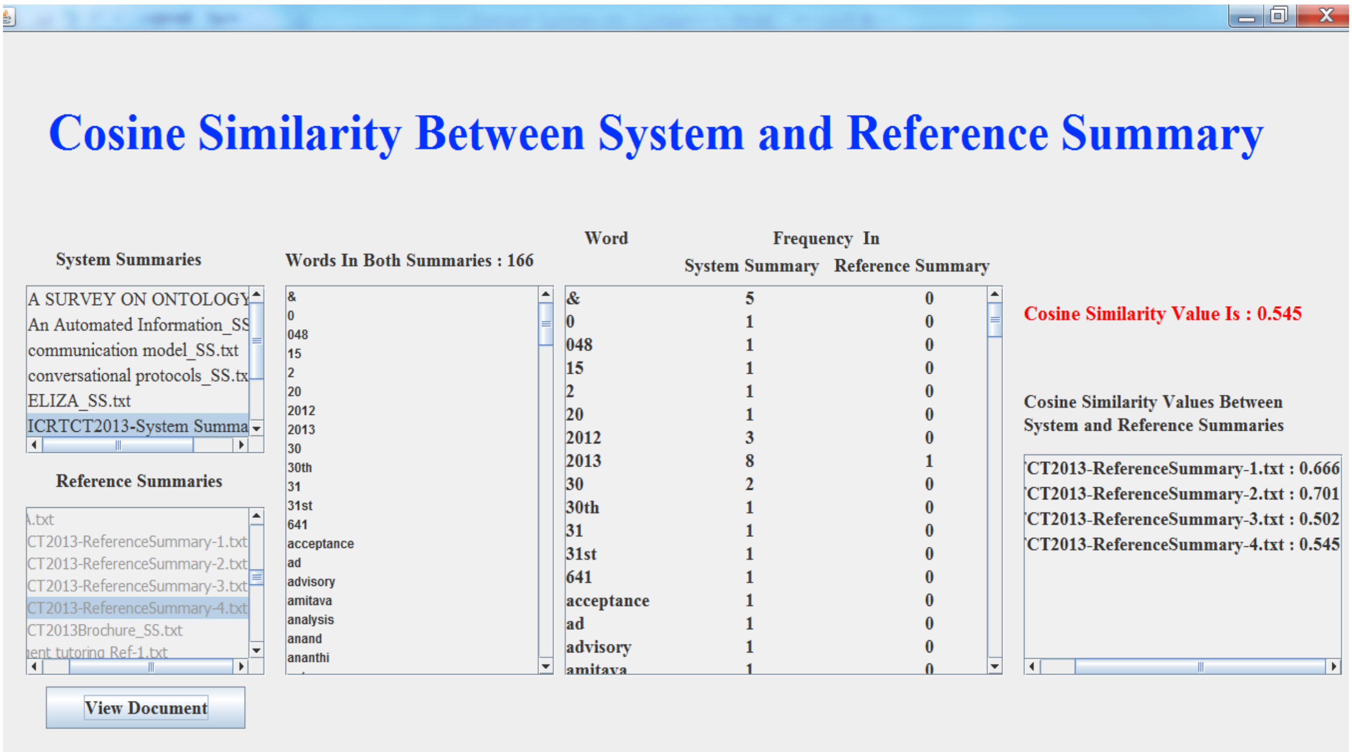
Count Of Sentences : 3

With too few exceptions quantitative international relations is the study of
via a quantitative approach and then to examine the specific cases identifie
+6 For a time these data collections were used in a fairly large fraction of

Selected Sentence With POS Tags

With_IN too_RB few_JJ exceptions_NNS quantitative_JJ international_JJ relations_NNS is_VBZ the_DT study_NN of_IN annual_JJ quarterly_JJ or_CC sometimes_RB monthly_JJ observations_NNS of_IN the_DT international_JJ system_NN +_CC

Fig -3 Summary generated based on a template



System Summaries:
A SURVEY ON ONTOLOGY
An Automated Information_SS
communication model_SS.txt
conversational protocols_SS.tx
ELIZA_SS.txt
ICRTCT2013-System Summa

Reference Summaries:
A.txt
CT2013-ReferenceSummary-1.txt
CT2013-ReferenceSummary-2.txt
CT2013-ReferenceSummary-3.txt
CT2013-ReferenceSummary-4.txt
CT2013Brochure_SS.txt
ent tutoring Ref-1.txt

Words In Both Summaries : 166
&
0
048
15
2
20
2012
2013
30
30th
31
31st
641
acceptance
ad
advisory
amitava
analysis
anand
ananthi

Word	Frequency In System Summary	Frequency In Reference Summary
&	5	0
0	1	0
048	1	0
15	1	0
2	1	0
20	1	0
2012	3	0
2013	8	1
30	2	0
30th	1	0
31	1	0
31st	1	0
641	1	0
acceptance	1	0
ad	1	0
advisory	1	0
amitava	1	0

Cosine Similarity Value Is : 0.545

Cosine Similarity Values Between System and Reference Summaries

CT2013-ReferenceSummary-1.txt : 0.666
CT2013-ReferenceSummary-2.txt : 0.701
CT2013-ReferenceSummary-3.txt : 0.502
CT2013-ReferenceSummary-4.txt : 0.545

Fig -4 Cosine similarity index between system and reference summaries.

Table -2: Result Analysis of Dialogue Management Interface

Index Words Used for information extraction Submitted	Document Category	Expected Information	Information Extracted	Correct	Partial	Spurious	Missing	Precision	Recall	F Measure
Text, Summarization	Technical research paper	Information from the document on Text, Summarization	All the information present in the document having the input words	23	8	8	0	0.69	0.87	0.19
Dates	CFP	Dates embedded inside the text	Dates present in the document	1	0	0	0	1.00	1.00	0.25
Places		Name of Places	Places present in the text	2	0	0	1	1.00	0.67	0.20
Subjects of Interest		Techincal subjects	Areas of Interest	11	1	6	0	0.64	0.96	0.19

5. CONCLUSIONS

Text mining is a trending research area in the present scenario. The simple reason for which it is being considered as trending topic is the vast availability of information in the form of natural language. In this paper, we presented algorithms for automatic text summarization and dialogue management. The focus of automatic text summarization algorithm is on the requirements of user before original text is summarized while the algorithm used for dialogue management establishes interactions between user and the computer. The main contribution of work reported here is the template based algorithm for text summarization and dialogue management. The experiments conducted during the implementation of work have produced encouraging results. These results indicate that the accuracy of both automatic summarization and dialogue management algorithms is more than70%. Hence, it can be concluded that the performance of algorithmic implementation of our work is satisfactory.

REFERENCES

[1] Abdunabi UBUL, EI-Sayed ATLAM, Kazuhiro MORITA, Masao FUKETA, Junichi AOE, “A Method for Generating Document Summary using Field Association Knowledge and Subjectively Information”, Proceedings of IEEE, 2010

[2] Anna Huang, “Similarity Measures for Text Document Clustering”, proceedings of NZCSRSC, 2008, pp 49-56

[3] Brill, E. “A Simple Rule-Based Part-of-speech Tagger”. Proceedings of 3rd Conference on Applied Natural Language Processing, 1992, pp.152–155.

[4] Church, K.W., “A Stochastic parts program and noun phrase parser for unrestricted text”. Proceedings 1st Conference on Applied Natural Language Processing, ANLP, pp. 136–143. ACL, 1988.

[5] C. Lakshmi Devasenal, M. Hemalatha, “ Automatic Text Categorization and Summarization using Rule Reduction”, Proceedings of ICAESM, 2012, pp594-598

[6] Dang Truong Son, Dao Tien Dung, “Apply a mapping question approach in building the question answering system for vietnamese language”, Proceedings of Journal of Engineering Technology and Education, 2012, pp 380-386

[7] en.wikipedia.org/wiki/Knowledge_extraction.

[8] Jantima Polpinij , “Ontology-based Knowledge Discovery from Unstructured Text”, Proceedings of IJIPM, Volume4, Number4, 2013, pp 21-31

[9] Kamal Sarkar, Mita Nasipuri,Suranjan Ghose, “Machine Learning Based Keyphrase Extraction: Comparing Decision Trees, Naïve Bayes, and Artificial Neural Networks”, Proceedings of JIPS, 2012, pp693-712

[10] M.Suneetha, S. Sameen Fatima, “Corpus based Automatic Text Summarization System with HMM Tagger”, Proceedings of IJSCE, ISSN: 2231-2307, Volume-1, Issue-3, 2011, pp 118-123

[11] Porter stemmer, http://www.tartarus.org/~martin/PorterStemmer

[12] Prashant G. Desai , Sarojadevi H. , Niranjan N. Chiplunkar, “Rule-Knowledge Based Algorithm for Event Extraction”, Proceedings of IJARCCE, ISSN (Online) : 2278-1021, ISSN (Print) : 2319-5940, Volume 4, Issue 1, 2015, pp 79-85

[13] Princeton University. Word Net, http://wordnet.princeton.edu/

[14] Rafeeq Al-Hashemi, “Text Summarization Extraction System (TSES) Using Extracted Keywords”,

Proceedings of International Arab Journal of e-Technology, Volume. 1, Number. 4, 2010, pp 164-168

[15] Raju Barskar, Gulfishan Firdose Ahmed, Nepal Barska, "An Approach for Extracting Exact Answers to Question Answering (QA) System for English Sentences", Proceedings of ICCTSD, 2012, pp 1187 – 1194

[16] Ralf Steinberger, Bruno Pouliquen, Camelia Ignat, "Using language-independent rules to achieve high multilinguality in Text Mining", Proceedings of Mining Massive Data Sets for Security, 2008, pp 217-240

[17] Rucha S. Dixit, Prof. Dr.S.S.Apte, " Improvement of Text Summarization using Fuzzy Logic Based Method", Proceedings of IOSRJCE, ISSN: 2278-0661, ISBN: 2278-8727 , Volume 5, Issue 6, 2012, pp 05-10

[18] Shuhua Liu, "Enhancing E-Business-Intelligence-Service: A Topic-Guided Text Summarization Framework", Proceedings of the Seventh IEEE International Conference on E-Commerce Technology, 2005

[19] Stanford University POS Tagger

[20] Subhashini, Kumar, "Evaluating the Performance of Similarity Measures Used in Document Clustering and Information Retrieval", Proceedings of ICIIC, 2010, pp 27 - 31

[21] Trung H. BUI, Multimodal Dialogue Management - State of the art. January 3, 2006

[22] Xing Jiang,Ah-Hwee Tan "Mining Ontological Knowledge from Domain-Specific Text Documents ", Proceedings of the Fifth IEEE International Conference on Data Mining, 2005

[23] Yuanchao Liu1, Ming Liu, Zhimao Lu, Mingkai Song, " Extracting Knowledge from On-Line Forums for Non-Obstructive Psychological Counseling Q&A System", International Journal of Intelligence Science, 2012, pp 40-48