IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.7, Special Issue.1, Jan 2019 E-ISSN: 2347-2693

Using WordNet-based Semantic Relatedness Measure for Reducing Redundancy and Improving Multi-document Text Summarization

Santanu Dam[1*], Kamal Sarkar[2]

[1]Department of Computer Science, Future Institute of Engineering and Management, Kolkata, India
[2]Department of Computer Science, Jadavpur University, Kolkata, India

[*]Corresponding Author: sntndm@gmail.com, Tel.: 9339495102

Available online at: www.ijcseonline.org

Abstract—Multi-document text summarization (MDS) is a task to generate a single summary from a set of articles related to the same topic or event. Since each input article is related to the same topic or event, the generated summary contains redundant sentences or the sentences that contain almost similar information. This paper presents a sentence similarity measure for reducing redundancy in multi-document summary. Our proposed similarity measure combines the WordNet based semantic sentence similarity measure with the traditional cosine similarity measure. We have conducted our experiments using DUC 2004 benchmark multi-document summarization dataset to judge whether the proposed similarity measure is useful for redundancy removal and improving multi-document text summarization performance or not. Our experiments reveal that our proposed similarity measure is effective for reducing redundancy and improving multi-document text summarization performance.

Keywords—Text Summarization; WordNet; Semantic Relatedness measure; Hybrid Sentence Similarity Measure; Redundancy Removal

I. Introduction

Text summarization is an effective mechanism for managing a large volume of textual data available on the Internet. Text summarization presents a document or a set of documents in condensed form which enables readers to digest the main content of the document(s) very quickly. Though text summarization is a research problem which have already been addressed in various way by the researchers for the last many years, the solution with the acceptable accuracy is still unknown. Summary is usually produced in two different forms-extractive form and abstractive form. The most prior research works on text summarization [1] [2] [3] [4] produce summaries in extractive form. Though the researchers have also tried to devise a number of summarization methods that produce summaries in abstractive form, the existing abstractive summarization methods are not yet proven to be successful for producing grammatically correct abstractive summaries of size beyond that of ultra-summaries [5][6][7]. Though the recent process in deep learning based summarization approaches have shown relatively better performance compared to the earlier abstractive summarization approaches, the deep learning based approach to generic text summarization have been tested mostly on generating short or headline like summaries [8][9][10] from single documents. The main reason is that such kinds of approaches require a large amount of document-summary pairs for training. Since creating a large number of multi-document human provided abstracts is a tedious task, the large multi-document summarization corpus is not available till date. In this work, we develop a MDS system that uses a hybrid sentence similarity measure for handling redundancy. Our proposed similarity measure combines the semantic sentence similarity measure that considers WordNet based lexical semantic relatedness between terms, and the traditional cosine similarity measure. In this paper we have blended our proposed similarity method to remove redundancy in the generated summary from the given document set. The outcome is significantly improves the result. In the next section we will formally discuss thework done so far in this area. Our primary goal was to understand how we can effectively improve the quality of the summary which was generating more than one document. We proposed a hybrid similarity to improve the quality of the summary. So, in section we discuss the proposed hybrid similarity measure. The text summarization method is used in this paper are discuss in the section IV. Section V includes the experimental set up and result we obtained from our proposed approach also this section followed by comparison with the other existing system and finally the section VI concludes the paper.

II. Related Work

The earliest extractive summarization methods ranked sentences based on the combined score calculated based on a positional information, frequency of terms and/or cue phrases [11][12][13] [14][15] etc., remove redundant sentences[16] [17]and finally selects top *m* non-redundant sentences to create a summary. Redundancy is an important factor affecting text summarization performance because users like to see important but diverse information in the summaries. Maximal marginal relevance is a popular technique for redundancy removal while generating summaries [18]. Another type of approach used to handle redundancy is the clustering based approach which partitions sentences into multiple clusters based on sentence similarities and identify groups of similar sentences as themes and finally selects sentences one by one from clusters to produce the final summary [19][20][21][22][23][24][25][26]. Performances of the MMR based approach and the clustering based redundancy removal approach are mainly dependent on the sentence similarity measure used for measuring similarity between two sentences.

In thegraph-based summarization approaches [27] [28][29][30][31], the sentences in a document or a document set are represented as a graph where a sentence corresponds to a node of the graph and an edge between a pair of sentences is established if their similarity value crosses a predefined threshold value. The graph-based methods compute the importance of a sentence by considering global information on the graph, rather than depending onlyon redundancy removal using information local to sentences. The most existing graph based techniques mainly use the traditional cosine similarity measure for constructing the similarity graph [28][31].

Like centroid based text summarization [17], many existing summarization systems applytext similarity measure for either redundancy removal or graph construction or both. So, we hypothesize that using the more accurate sentence similarity measure, summarization performance can be improved. In this paper, we have presented a hybrid sentence similarity measure that blends our defined semantic similarity measure with the cosine similarity measure for redundancy removal.

III. OUR PROPOSED HYBRID SIMILARITY MEASURE

Sentence similarity measure has an important role in MMR(Maximum Marginal Relevance) [18] based redundancy removal method because, in this method, summary is generated incrementally starting with the top most ranked sentence and gradually selecting sentences in the summary from the ranked list if the sentence to be selected is sufficiently dissimilar with the already selected sentences. Our proposed similarity measure hybridizes two sentence similarity measures – (1) our defined semantic similarity measure that takes into account WordNet based lexical semantic relatedness between terms, and (2) the traditional cosine similarity measure. The rationale behind combining two types of similarity measures is that the cosine similarity measure alone is not sufficient to detect similarity between sentences due to the well- known data sparseness problem and term mismatch problem whereas our defined semantic similarity measure discussed later in this section cannot capture relative importance of the terms while finding similarity between sentences. So, combining both the similarity measures leads to producing more accurate hybrid similarity measure helping in reducing redundancy effectively from generated summaries.

A. Cosine Similarity

If sentences S1 and S2 are represented in the following vector form: S1= {w11, w12, w13,….,w1n}, S2= { w21, w22, w23,….., w2n}, where wi,j means the TFS-IDF value for the jth word in the i-th sentence and n is the vector length which is equal to the distinct number of words in the pair of sentences after deleting stop words, the cosine similarity between S1 and S2 is calculated using following formula [28][31]:

$$Cosine\_Sim(S_1, S_2) = \frac{S_1 \circ S_2}{|S_1||S_2|} \quad (1)$$

TFS-IDF value of a term is equal to the product of TFS and IDF, where TFS = how many times the term occurs in a sentence and IDF is calculated using the formula given in [28][31]:

$$IDF_i = \log\left(\frac{N}{n_i}\right) \quad (2)$$

Where N = size of corpus in terms of number of documents and n_i = the number of documents containing the word *i* at least once. IDF value is computed from a large corpus different from our test corpus.

B. Semantic Similarity Measure

We have defined semantic similarity between two sentences based on WordNet based lexical semantic relatedness between terms. WordNet [32] is a lexical network covering a large number of English words. It is networks of synonym sets (synsets) for verbs, adjectives, nouns and adverbs where each synset represents one lexical concept and words are interlinked with a variety of semantic and lexical relations [33]. Overall, WordNet can be represented as a graph where each vertex represents a synset (v), and each directed edge v->w represents that w is semantically or lexically related to v. It is basically a directed and acyclic graph having a root node, though not necessarily a tree since each synset may consist of several synonyms.

To compute semantic similarity between sentences, WordNet is consulted to know how much the words in the two sentences are semantically related. We have used lexical semantic related measures that use WordNet.

The main problem in using WordNet based relatedness between terms for computing sentence similarity is that semantic relatedness in WordNet is a concept which is more general than similarity [33] and WordNet maintains various relationships between terms. For example, “bank” and “trust company” are related because they are similar, but “hot” and “cold” are related because they are antonyms. So, we need to find how to use semantic relatedness between terms for finding sentence level semantic similarity useful for detecting sentence level redundancy and thus improving text summarization performance by reducing redundancy while selecting sentences for summary generation.

There are many existing lexical semantic relatedness measures such as Lin [34], Wup (Wu and Palmer) [35], path length [36] etc. For our present work, we have used Wup as a measure for finding lexical semantic relatedness between terms. Wup similarity measure [35] which Wu and Palmer (1994) call conceptual similarity measure is defined as follows:

$$wup(C_p, C_q) = \frac{2*d}{L_p+L_q+2d} \quad (3)$$

Where *wup(C_p, C_q)* computes similarity between two conceptsC_p andC_q, *d* indicates the depth of the least common subsumer (LCS) from the root of WorldNet hierarchy, L_p is the length of path between C_p and LCS, and L_q is the length of path between C_q and LCS. LCS is the lowest common node between the paths of the two senses C_p andC_q from the root. LCS (least common subsumer) node in WordNet hierarchy connects two concepts C_p and C_q. For example, the LCS of “canine” and “chap” is “organism” which is the lowest common node between the paths of these two senses from the root ofWorldNet hierarchy. According to this measure, semantic similarity between two concepts increases based on how much their LCS is closer to them and how far the LCS is from the root of the tree likeWorldNet hierarchy. This is based on the two assumptions. The first assumption considers that two concepts are closely to each other if they are closely related to the LCS, that is, when L_p and L_q have lower values. The second assumption considers whether the LCS is more general or not. If two concepts are related to the less general (more specific) LCS, they are closely related. This is captured by incorporating the parameter *d* in equation 3. ThusWup measure gives how much two words are semantically similar. Using word level semantic similarity value given by Wup measure, we have defined semantic similarity between sentences as follows:

$$Semantic\_Sim(S1, S2) = \frac{2*Match}{|S_1|+|S_2|} \quad (4)$$

Where match =# no of semantic and syntactic term-matches between S1 and S2. Two terms are taken as semantically matched if WordNet based Wup value for the pair of terms is greater than a predefined threshold (0.85 in our setting). When two terms are string identical, it is considered as syntactic match. Stop words are removed from sentences before computing similarity between them.

The reason for having higher threshold on Wup value in our current settings (threshold on Wup value is set to 0.85) is that words in the WordNet are related by the various relations of which the relation that tells us whether a pair of words are semantically similar or not is of our interest and our main objective is to reduce redundancy while generating a summary from a document set. We have observed that choosing higher threshold onWup value gives our desired relation between words. For example, with threshold value of 0.85, we have obtained the words, which are similar with the word “agencies”, are “police”, “Committee”, “Commission”, “Agency” etc. The optimal threshold value of 0.85 has been determined through experimentations.

C. Hybrid Similarity Measure

As we have mentioned earlier in this section, our proposed similarity measure hybridizes two sentence similarity measures – (1) our defined semantic similarity measurethat takes into account WordNet based lexical semantic relatedness between terms, and (2) the traditional cosine similarity measure.Two similarity measures have been combined in the following way.

$$Hybrid\_sim(S_1, S_2) = \frac{Cosine\_Sim\,(S_1, S_2) \; + Semantic\_Sim\,(S_1, S_2)}{2} \quad (5)$$

Where Cosine-Sim and Semantic_Sim are two different similarity measures discussed earlier in this section.

IV. TEXT SUMMARIZATION METHOD

Since our main objective is to judge impact of semantic similarity measure on text summarization performance, we have chosen a centroid based summarization method proposed in [17] for further improvement using our proposed similarity measure. In brief, we will present below the basic idea of centroid based summarization.

A. Existing Centroid based MDS System

The centroid based summarization system selects a set of important words from the cluster of input documents as the centroid. A term is included in the centroid if the TF*IDF value for the term is greater than a pre-defined value (this is 3 in our settings). Here TF is defined as the average occurrences of the term in the input document set and IDF is calculated using the formula given in subsection 2.A. After calculating the centroid for an input document-set, sentences in the documents are ranked based on their cosine similarities with the centroid. Since the similarity value is used here only for ranking sentences, instead of computing traditional cosine similarity, a similarity of a sentence with the centroid is computed using the formula [17] given below and it is used as the score of a sentence.

$$Score(S) = \sum_{W \in (S \cap C)} TF(W) * IDF(W) \quad (6)$$

Where S∩C= the set of words common between the sentence S and the centroid C.
Originally, the centroid based summarization has been a part of the MDS system called MEAD, which is publicly available multi-document summarizer [37] considered as a hard baseline for multi-document summarization task.

B. Modified Centroid based MDS System

To judge the effectiveness of our proposed similarity measure, we have considered the centroid based summarization method discussed above for further improvement by adding to it a redundancy removal component that uses our proposed similarity measure.

Though MEAD combines centroid feature with other features such as position, overlap with the first sentences while ranking sentences, for our implementations, we have considered only the centroid feature due to its domain independence nature. We have not considered positional and first sentence overlap features since they are highly specific to the news domain.

In our approach, our proposed hybrid similarity measure is used for calculating similarity between a sentence and the centroid, and sentences in the input document set are assigned scores based on the degree of similarities to the centroid. Then the sentence which has the highest similarity value is selected first and sentences from the ranked list are selected one by one in the summary if they have sufficient dissimilarity with the sentencesalready selected in the summary. In other words, the candidate sentence is not selected in the summary if the similarity of the candidate sentence with each of the previously chosen sentences exceeds a predefined threshold value (we set its value to 0.45 for the best results).

V. EXPERIMENTS AND RESULTS

Since the most existing work on multi-document summarization focused on generic summarization using the DUC 2004 (Task 2) dataset, we have considered it as the most appropriate one for evaluating the summarization performance. The DUC 2004 dataset is a collection of 50 input folders for multi-document summarization task where each folder contains approximately 10 news documents related to a topic. According to specifications of DUC 2004 (task2), the task was to create a 665-byte summary (approximately 100 words) for each input. The organizers (NIST) of DUC 2004 shared task (http://duc.nist.gov) released this dataset along with the reference summaries used for evaluating systems participated in the shared task.

For summary evaluation, our system generated summary is compared with a set of reference summaries released by NIST. ROUGE (version 1.5.5) package [38] has been used for this purpose. ROUGE evaluates summary by comparing the system generated summary with the set of human created (reference) summaries and calculates ROUGE-N scores based on N-gram overlap between the system-generated summaries and the human summaries. Accordingly, ROUGE-1 score is calculated based on unigram overlap, ROUGE-2 score is calculated based on bigram overlap and so on. We have considered ROUGE-1 and ROUGE-2 recall scores for comparing our systems. We have used recall scores for evaluation because we have set the option –b 665 in the ROUGE evaluation tool which takes the first 665 bytes of the system generated summary for evaluation.

a. Results

To prove how much our proposed similarity measure is effective, we have designed our experiments in two ways-(1) existing centroid-based summarization system with centroid feature and traditional cosine similarity measure for redundancy removal, (2) our modified centroid-based summarization system with centroid feature and our proposed similarity measure for redundancy removal. The results obtained by our developed models are shown inTable 1 and Table 2.

We have shown ROUGE-1 score in Table 1 and ROUGE-2 score in Table 2. As we can see from both the tables, performance of the centroid based summarization system is improved with our proposed sentence similarity measure used for redundancy removal. Since our proposed similarity measure considers WordNet based semantic relatedness between terms, the obtained results prove the usefulness of carefully chosen semantic similarity measure while removing redundancy from the generated summaries.

TABLE 1. SUMMARY OF ROUGE-1 RECALL SCORES FOR OUR DEVELOPED SYSTEMS IN DUC 2004 TASK 2

SYSTEMS	ROUGE-1 recall score
Modified Centroid based system	**0.3795 [0.3733 - 0.3850]**
Existing Centroid based system	0.3746 [0.3688 - 0.3799]
Lead baseline	0.3210 [0.3147- 0.3278]

The difference of the ROUGE scores obtained by two different systems is also significant at 95%confidence intervals. The min-max range shown in the brackets after each score shown in Table 1 and Table 2 indicate that both the minimum score and maximum score obtained by our proposed system are better than the system it is compared to.

TABLE 2. SUMMARY OF ROUGE-2 RECALL SCORES FOR OUR DEVELOPED SYSTEMS IN DUC 2004 TASK 2

SYSTEMS	ROUGE-2recall score

Modified Centroid based system	**0.0951 [0.0913 - 0.0992]**
Existing Centroid based system	0.0933 [0.0895 - 0.0973]
Lead baseline	0.0638 [0.0592-0.0685]

In Table 1 and Table 2, we have also shown the results of DUC 2004 baseline defined for generic multi-document summarization shared task. Lead baseline was to take the first 665 byte of the most recent document in the input as a summary.

b. Comparison to Some other Existing Systems

In this section, we havepresented performance comparison ofour developed systems and the systems that participated in DUC 2004 shared task -Task 2. We have compared our proposed MDS system with few top DUC systems.For evaluating the DUC systems,we have used the ROUGE package and the peer summaries released on the DUC conference website (http://duc.nist.gov).Table 3 shows comparisons of our proposed system with the DUC systems.

TABLE 3. COMPARISONS OF ROUGE SCORES FOR FEW TOP DUC SYSTEMS AND DUC BASELINE SYSTEM (LEAD) WITH OUR PROPOSED SYSTEM

Systems	ROUGE-1 recall score
DUC system with peer code 65	0.3822
Our proposed summarization system	0.3795
DUC system with peer code 104	0.3744
DUC system with peer code 35	0.3743
DUC lead baseline	0.3242

On comparing the ROUGE scores, we find that the proposed system performs better than the system which was regarded as the second best (peer code 104) in DUC 2004 shared task. Though our proposed system performs slightly worse than the best system CLASSY with peer code 65 shown at row1 in Table 3, our proposed system differs from the best system in many ways: the system CLASSY has used Hidden Markov Model (HMM), which requires a large training data set. HMM states are determined based on empirical testing. In addition, CLASSY applied sentence trimming using some linguistic rules at the preprocessing stage. So, compared to the best system CLASSY, our system is simple, easily implementable and portable.

VI. CONCLUSIONS AND FUTURE WORK

In this study, we have investigated whether WordNet based lexical semantic relatedness between terms is useful in defining sentence similarity measure that effectively removes redundancy while generating multi-document summaries.

Our experiments reveal that WordNet based semantic relatedness between terms should be carefully used in defining sentence level sentence similarity because WordNet maintains various kinds of semantic relatedness, but a few of them are useful in defining sentence similarity measure. In this study, we have controlled this factor by setting a higher threshold on the value of semantic relatedness measured by Wup [35]. Though our defined semantic similarity measure being combined with cosine similarity measure helped inhandling redundancy and improving summarization performance, it was not as accurate as we thought it could be. We feel that defining the appropriate sentence similarity measure for redundancy removal is not an easy task. However, our obtained results are encouraging because our study reveals that the traditional cosine similarity alone is not sufficient for redundancy removal and our proposed hybrid similarity measure can be used for the sake of improving summarization performance.

In our present study, we have studied with only Wup measure for measuring semantic relatedness between terms. We will investigate in future how other WordNet based word similarity measures such as Lin [34], Path length [36] can affect multi-document summarization performance. The contextual word level similarity measure using Word2Vec model [39] can also be studied in this context.

We have also planned to apply our proposed hybrid similarity measure for other kinds of summarization tasks such as query focused summarization and opinion summarization.

REFERENCES

[1] J. Goldstein, V. Mittal, J. Carbonell, and M.Kantrowitz,"Multi-document summarization by sentence extraction," In Proceedings of the 2000 NAACL-ANLP Workshop on Automatic summarization, pp. 40-48, Association for Computational Linguistics.

[2] V. K. Gupta, T. J Siddiqui,"Multi-document summarization using sentence clustering," In Intelligent Human Computer Interaction (IHCI), 2012 4th International Conference on , pp. 1-5, IEEE, 2012.

[3] K. Sarkar,"A Keyphrase-Based Approach to Text Summarization for English and Bengali Documents," International Journal of Technology Diffusion (IJTD), 5(2), pp. 28-38, 2014.

[4] K. Sarkar, "Sentence clustering-based summarization of multiple text documents," Int. J. Comput. Sci. and Commun. Tech, 2(1), pp. 225-235, 2009.

[5] K. Sarkar, and S. Bandyopadhyay,"Generating headline summary from a document set," In International Conference on Intelligent Text Processing and Computational Linguistics, pp. 649-652, Springer Berlin Heidelberg. 2005

[6] M. Banko, V. Mittal and M. Witbrock,"Headline generation based on statistical Translation," In Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL-2000), Hong Kong, pp. 318–325, 2000.

[7] D. Zajic, B. Dorr and R. Schwartz, "Automatic Headline Generation for Newspaper Stories," Workshop on Automatic Summarization. Philadelphia, PA, pp. 78-85, 2002.
[8] A. Rush, S. Chopra, and J. Weston,"A Neural Attention Model for Abstractive Sentence Summarization," In Procedings of EMNLP 2015
[9] R. Nallapati, B. Zhou, and C. Santos,"Abstractive Text Summarization Using Sequence-to- RNNs and Beyond," In Computation and Language, 2016.
[10] R. Nallapati, F. Zhai, and B. Zhou,"SummaRuNNer: A Recurrent Neural Network based Sequence Model for Extractive Summarization of Documents,", In Computation and Language, 2016.
[11] P. B. Baxendale,"Man-made index for technical literature—An experiment," IBM Journal of Research and Development, 2(4), pp. 354–361, 1958.
[12] H. P. Edmundson,"New methods in automatic extracting. Journal of the Association for Computing Machinery," 16(2), pp. 264–285, 1969.
[13] H. P Luhn,"The automatic creation of literature abstracts," IBM Journal of Research Development, 2(2), pp. 159–165, 1958.
[14] K. Sarkar,"Using domain knowledge for text summarization in medical domain," International Journal of Recent Trends in Engineering, 1(1), pp. 200-205, 2009.
[15] K. Sarkar, M. Nasipuri, and S. Ghose, " Using machine learning for medical document summarization," International Journal of Database Theory and Application. vol. 4, pp. 31-49, 2011.
[16] K. Sarkar, K. Saraf, A. Ghosh,"Improving graph based multidocument text summarization using an enhanced sentence similarity measure," In Recent Trends in Information Systems (ReTIS). IEEE 2nd International Conference on, pp. 359-365, 2015. IEEE
[17] D.R. Radev, H. Jing, M. Styś and D.Tam,"Centroid-based summarization of multiple documents," Information Processing & Management, 40(6), pp.919-938, 2004.
[18] J. G. Carbonell, J. Goldstein,"The use of MMR, diversity-based re-ranking for reordering documents and producing summaries," In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia, pp. 335–336, 1998
[19] D. Marcu, L. Gerber,"An inquiry into the nature of multi-document abstracts, extracts, and their evaluation," In Proceedings of the NAACL-2001 Workshop on Automatic Summarization, Pittsburgh, June. NAACL, pp. 1–8, 2001.
[20] E. Boros, P. B. Kantor, D. J. Neu,"A Clustering Based Approach to Creating Multi-Document Summaries," In Proceedings of the 24th ACM SIGIR Conference, LA, 2001.
[21] H. Hardy, N. Shimizu, T. Strzalkowski, L. Ting, G. B. Wise, and X. Zhang,"Cross-document summarization by concept classification," In Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Finland, pp. 121–128, 2002.
[22] M. F. Moens, C. Uyttendaele, and J. Dumortier, "Abstracting of legal cases: the potential of clustering based on the selection of representative objects," Journal of the American Society for Information Science, 50 (2), pp. 151-161, 1999.
[23] K. Sarkar,"Sentence clustering-based summarization of multiple text documents," Int. J. Comput. Sci. and Commun. Tech, 2(1), pp. 225-235, 2009.
[24] G. C. Stein, A. Bagga, and G. B. Wise, "Multi-Document Summarization: Methodologies and Evaluations," In Conference TALN 2000, Lausanne, 2000.
[25] V. Hatzivassiloglou, J. Klavans, and E. Eskin,"Detecting test similarity over short passages: Exploring linguistic feature combinations via machine learning," In Proceedings of EMNLP,1999.
[26] V. Hatzivassiloglou, J. L. Klavans, M. L. Holcombe, R. Barzilay, M-Y. Kan, and K. R. McKeown,"SimFinder: A Flexible Clustering Tool for Summarization," NAACL, Workshop on Automatic Summarization, Pittsburgh, PA, 2001.
[27] R. Barzilay, N. Elhadad, K. R. McKeown,"Sentence ordering in multidocument summarization," In Proceedings of the first international conference on Human language technology research, Association for Computational Linguistics, pp. 1-7, 2001.
[28] G. Erkan, D. R. Radev,"LexRank: graph-based lexical centrality as salience in text summarization," Journal of Artificial Intelligence Research, pp. 457-479, 2004.
[29] R. Mihalcea, P. Tarau,"TextRank: Bringing order into texts," In Proceedings of EMNLP2004, 2004.
[30] R. Mihalcea, P. Tarau, "A language independent algorithm for single and multiple document summarization," In Proceedings of IJCNLP 2005.
[31] G. Erkan, D. R. Radev,"LexPageRank: Prestige in Multi-Document Text Summarization," In Proceedings of EMNLP, 2004.
[32] C. Fellbaum, editor, "WordNet: An Electronic Lexical Database," The MIT Press, Cambridge, MA, . 1998.
[33] A. Budanitsky and G. Hirst, "Evaluating WordNet-based measures of lexical semantic relatedness," Computational Linguistics, 32(1), pp.13-47, 2006.
[34] D. Lin. "An Information-Theoretic Definition of Similarity," Proc. Int'l Conf. Machine Learning, July 1998.
[35] P. Resnik,"Using information content to evaluate semantic similarity," In Proceedings of the 14th International Joint Conference on Artificial Intelligence, pages 448–453, Montreal, Canada, 1995.
[36] D. R. Radev, T. Allison, S. Blair-Goldensohn, J. Blitzer, A.Celebi, S. Dimitrov, E. Drabek, A. Hakim, W. Lam, D. Liu and J. Otterbacher,"MEAD-A Platform for Multidocument Multilingual Text Summarization," In LREC, 2004.
[37] C. Y. Lin, "ROUGE: A package for automatic evaluation of summaries," In Proc. Workshop Text Summarization Branches Out, PostConf.Workshop ACL, pp. 25–26, 2004.
[38] T.Mikolov, I.Sutskever, K.Chen, G.S. Corrado, and J.Dean,"Distributed representations of words and phrases and their compositionality," In Advances in neural information processing systems, pp. 3111-3119, 2013
[39] Z.Wu and M.Palmer, "Verb semantics and lexical selection," In Proceedings ofthe 32nd Annual Meeting of the Associations for ComputationalLinguistics, pages 133–138, Las Cruces, NewMexico, 1994

Authors Profile

Mr. Santanu Dam Completed M. Tech from Jadavpur University, India in 2010 He is currently pursuing Ph.D. and working as assistant professor in Department of Computer Science and Engineering, of, Future Institute of Engineering and Management, Sonarpur, Kolkata since 2008. His main research work focuses on Natural Language Processing, Machine Learning, Text Summarization, Text Mining, Cloud Computing. He has 13 years of teaching experience.

Prof K. Sarkar received his B.E degree in Computer Science and Engineering from the Faculty of Engineering, Jadavpur University in 1996. He received the M.E degree and Ph.D. (Engg) in Computer Science and Engg. from the same University. In 2001, he joined as a lecturer in the Department of Computer Science & Engineering, Jadavpur University, Kolkata, where he is currently aprofessor. His research interest includes Natural Language Processing, Machine Learning, Text Summarization, Text Mining, Speech Recognition.

