A sentence scoring method for extractive text summarization based on Natural language queries

R.V.V Murali Krishna[1] and Ch. Satyananda Reddy[2]

[1] I.T Department, G.V.P College of Engineering
Visakhapatnam, Andhra Pradesh 530048, India

[2] CS & SE Department, Andhra University
Visakhapatnam, Andhra Pradesh 530009, India

Abstract

The developments in storage devices and computer networks have given the scope for the world to become a paperless community, for example Digital news paper systems and digital library systems. A paperless community is heavily dependent on information retrieval systems. Text summarization is an area that supports the cause of information retrieval systems by helping the users to get their needed information. This paper discusses on the relevance of using traditional stoplists for text summarization and the use of Statistical analysis for sentence scoring. A new methodology is proposed for implementing the stoplist concept and statistical analysis concept based on parts of speech tagging. A sentence scoring mechanism has been developed by combining the above methodologies with semantic analysis. This sentence scoring method has given good results when applied to find out the relation between natural language queries and the sentences in a document.

Keywords: *Information retrieval systems, traditional stoplists, sentence scoring, statistical analysis, semantic analysis, parts of speech tagging*.

1. Introduction

Text summarization is the process of extracting important information from a given text. Based on the how this important information is presented to the user, two types of text summarization systems are defined [1]. They are 1) Extractive summarization system 2) abstractive summarization system. In Extractive summarization system important text segments of the original text are identified and presented as they are. In abstractive summarization original text is interpreted and is written in a condensed form so that the resulting summary contains the essence of the original text. The summary in extractive summarization contains the words and sentences of the original text. This may not happen in abstractive summarization system.

Stop lists play an important role in building search engines and text summarization systems. They help in filtering useful information from the original text. Traditional stoplists are those which are specific to a natural language and are primarily developed for use in a search engine. Since Text summarization is a complex task involving natural language processing, it uses natural language processing tools like Dictionaries, Thesaurus, Wordnet, POSTagger etc... . A Parts-Of-Speech Tagger (POS Tagger) is a piece of software that reads text in some language and assigns parts of speech to each word (and other token), such as noun, verb, adjective, etc., although generally computational applications use more fine-grained POS tags like 'noun-plural'. Wordnet is an on-line lexical reference system whose design is inspired by current psycholinguistic theories of human lexical memory. Second one is a Stanford Log-linear Part-Of-Speech Tagger. Term frequency is a statistical measure used in calculating relevance of a document. It tells something about the document as a whole with respect to a user query.

2. Motivation

The meaning of an English sentence lies in the noun phrase and verb phrase of that sentence. Most important elements of noun phrases are nouns and adjectives. The important elements of verb phrases are verbs and adverbs. So in pre processing stage of the sentence we have tagged the words of the sentence with their corresponding parts of speech and separated the words whose tags fall in the set (Nouns, adjectives, verb, and adverb). The traditional stop lists are not best suited for text summarization because they contain words whose parts of speech are pronouns, adverbs, adjectives, verbs. So when we apply these stop lists we may lose some important information. We have classified the stop words in some of the stop lists available on the World Wide Web based on their parts of speech. Table 1 gives quantified description of the classification. It shows 8 different stoplists and four classes of words i.e, Nouns, Adjectives, Adverbs, and Verbs.

Table 1: Total no of words in each stoplist and their classification based on Stanford POSTagger [2, 3, 4].

Stoplist	Total no of				
	Words In The Stoplist	Nouns	Adje-ctives	Ad - verbs	Verbs
Stoplist[8]	550	190	29	89	68
Stoplist[9]	571	106	41	118	82
Stoplist[10]	236	44	7	30	4
Stoplist[11]	61	5	4	13	4
Stoplist[12]	199	44	11	11	9
Stoplist[13]	425	104	40	62	81
Stoplist[14]	571	106	41	118	82
Stoplist[15]	429	104	43	65	79

Term frequency measure can be applied to calculate relationship between two sentences. It does not give acceptable results when the there is a big difference between the lengths of the sentences. Term frequency does not take care of the context of the words. Term frequency will give better results when it is clubbed with parts of speech tagging.

3. Proposed System

A sentence scoring method is built on the concepts of stop word removal, Semantic relationship and statistical relationship. Fig 1 shows the overall functionality of the system.

3.1 Algorithm for Stop Word Removal (Stopword_Remover)

Input: Sentence
Output: Keywords list

Step 1: Parse the sentence in to words based on Standard English language tokens [16]
Step 2: Tag the words with their corresponding parts of Speech [16].
Step 3: Add the words to the keywords list whose tag is Noun/verb/adverb/adjective.

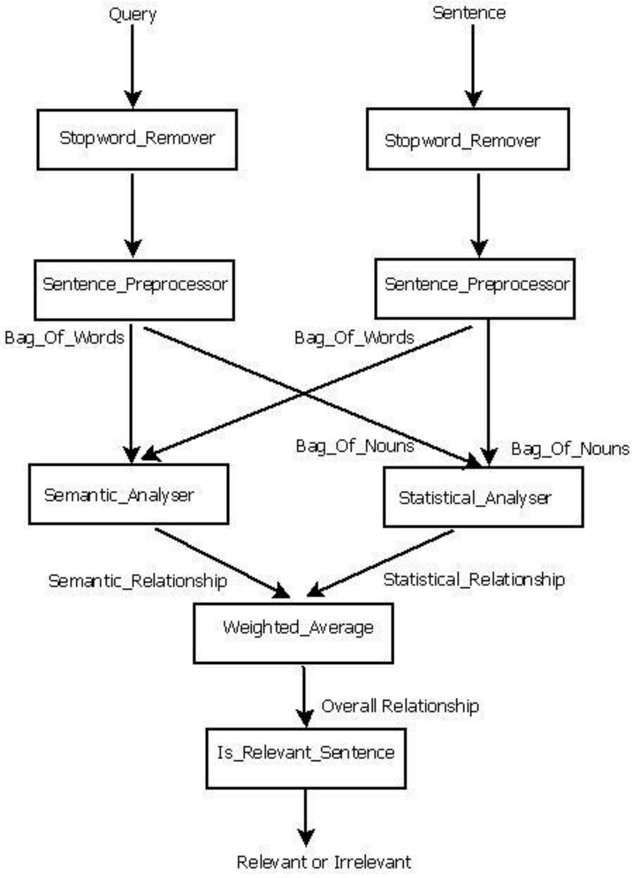

Fig. 1: functionality of Sentence scoring method.

3.2 Algorithm for Sentence Processing (Sentence_Processor)

Input: keywords list
Output : Bag_Of_Words, Bag_of _Nouns

Step1: Pick a word in the keywords list and find synonyms [5,6,7] for it.
Step2: Add the words and its synonyms to a list called as Bag_Of_Words
Step3: Repeat step 1 until all the words in keywords list are picked.
Step4: Add the words in keywords list to Bag_of _Nouns whose tag is noun[16]

3.3 Algorithm for Semantic analysis (Semantic_Analyser)

Input: Bag_Of_Words of Query (BOW_Q) and Bag_Of_Words of sentence (BOW_S)
Output: Semantic relationship (Sem_Rel)

Step 1:

Semantic relationship (Sem_Rel) = set of common words in both the lists/ set of all unique words in the union of both the lists.

Sem_Rel = (BOW_Q ∩ BOW_S) / (BOW_Q U BOW_S)

3.4 Algorithm for statistical analysis (Statistical_Analyser)

Input: Bag_Of_Nouns of Query (BON_Q) and Bag_Of_Nouns of sentence (BON_S)
Output: Statistical relationship (Stat_Rel)

Step 1:

Statistical relationship (Stat_Rel) = set of common words in both the lists / set of all unique words in the union of both the lists.

Stat_Rel = (BON_Q ∩ BON_S) / (BON_Q U BON_S)

3.5 Algorithm for Weighted Average (Weighted_Average)

Input: Sem_Rel and Stat_Rel
Output: Weighted_Average

Step 1:

Weighted_Average = (Sem_Rel + 2 * ((Stat_Rel *100) / S_words.length))/3
where S_words.length contains count of the words in the sentence.

4. Results

Input Set 1:

Query :	where is ramu going
Sentence :	ramu is going to school

Output : Similarity obtained based on proposed sentence scoring method : 50.0

Input Set 2:

Query :	who sat beside the car
` Sentence :	ramu sat beside the window

Output : Similarity obtained based on proposed sentence scoring method : 13.3

5 Result analysis

The results obtained are compared with a sentence scoring method built on the concepts of statistical anaylsis, semantic analysis and stoplist. The stoplist was built by Gerard Salton and Chris Buckley for the experimental SMART information retrieval system at Cornell University[15]

Table 2: Percentage of similarity between Query and Sentence

Input Set No.	Similarity obtained by the	
	proposed sentence scoring method	sentence scoring method based on a fixed Stoplist[15]
1	50.0	38.0
2	13.3	52.0

In table 2 the results show that there is a difference between the two sentence scoring methods in terms of similarity assessment. After the analysis, it can be said that the proposed system has given acceptable and more accurate results.

In Input Set No 1 the word “going” is an important word in both the sentences (Query and Sentence). But since it is listed as a stopword, it is not taken into consideration while calculating the similarity between the sentences. As Table 1 indicates there are so many words like that which are kept in the stoplists.

In Input Set No 2 there are two nouns in the sentence and there is one noun in the query. Normal statistical frequency calculation methods do not care for parts of speech of the words. Here the words “sat” and “beside” are matching in the given sentences (query and sentence). The resulting similarity is based on those two matched words. But the sentences are having nouns which are not at all matching. These two sentences are dissimilar. The proposed method has given very less similarity between the sentences. So statistical frequency combined with parts of speech has more precision.

6. Conclusions

Traditional stoplists that are used by search engines should not be used for sentence preprocessing in text summarization. Because they contain words which play an important role in fetching accurate data from a nonstructured database. The sentence scoring methods are very much dependent on keywords in the sentences and these keywords can be obtained by using parts of speech tagging. A sentence scoring method based on the proposed stoplist methodology will give better similarity results for sentences involving natural language queries.

Acknowledgments

I am deeply indebted to the almighty for giving his blessings.

References

[1] S. Gholamrezazadeh, M. A. Salehi, and B. Gholamzadeh. “A comprehensive survey on text summarization systems”. In Proceedings of CSA '09, 2009, pp. 1—6.
[2] http://nlp.stanford.edu/software/tagger.shtml
[3] Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer. ”Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network”. In Proceedings of HLT-NAACL 2003, pp. 252-259.
[4] Kristina Toutanova and Christopher D. Manning. “Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger”. In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000)”,2000, pp. 63-70.
[5] George A. Miller ,” WordNet: A Lexical Database for English”. Communications of the ACM Vol. 38,1995, No. 11: pp.39-41.
[6] Christiane Fellbaum, ”WordNet: An Electronic Lexical Database. Cambridge”, MA: MIT Press,1998.
[7] Princeton University "About WordNet." WordNet. Princeton University. 2010. <http://wordnet.princeton.edu>
[8] http://www.thebananatree.org /stoplist.html
[9] http://jmlr.csail.mit.edu/papers /volume5/lewis04a/a11-smart-stop-list/english.stop
[10] http://nlp.cs.nyu.edu/GMA_files /resources/english.stoplist
[11] http://www.acrobatfaq.com/tbx/index/tinderbo/configur/stoplist.html
[12] http://www.d.umn.edu/~tpederse/Group01/WordNet/words.txt
[13] http://frakes.cs.vt.edu/stoplist.html
[14] ftp://ftp.cs.cornell.edu/pub/smart/english.stop
[15] http://www.lextek.com/manuals/onix/stopwords2.html
[16] Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. “Building a large annotated corpus of English: the penn treebank.” *Comput. Linguist.* 19, 2,1993, pp. 313-330.
[17] Eduard Hovy and Chin Yew Lin, “Automated text summarization in SUMMARIST”, MIT Press, 1999, pp. 81–94.
[18] Mani, I., “Automatic Summarization” , John Benjamin’s Publishing Co. 2001,pp.1-22.
[19] Michael W. Berry and Jacob Kogan, “Text Mining: Applications and Theory ”, John Wiley & Sons, Ltd, 2010,

Mr. R.V.V. Murali Krishna is a faculty of information technology in G.V.P College of engineering (Autonomous), Visakhapatnam..He is currently pursuing Ph.D in Andhra University.

Dr. Ch. Satyananda Reddy has obtained Ph.D in Computer Science and engineering. He is a regular faculty in computer science and systems engineering department of Andhra University, Visakhapatnam. He has taught several subjects in the areas of computer science and information technology. He is specialized in Software Engineering, Software Cost Estimation, Object Oriented Analysis & Design, Web Technologies, Data Base Management Systems, and Operating Systems