A Method for Semantic Relatedness Based Query Focused Text Summarization

Nazreena Rahman[1(✉)] and Bhogeswar Borah[2]

[1] Department of Computer Science and Engineering, Assam Kaziranga University, Jorhat 785006, Assam, India
nazreena@kazirangauniversity.in
[2] Department of Computer Science and Engineering, Tezpur University, Sonitpur 784028, Assam, India
bgb@tezu.ernet.in

Abstract. In this paper, a semantic relatedness based query focused text summarization technique is introduced to find relevant information from single text document. This semantic relatedness measure extracts the related sentences according to the query. The query focused text summarization approach can work on short query when the query does not contain enough information. Better summaries are produced by this method with increased number of query related sentences included. Experiments and evaluation are done on DUC 2005 and 2006 datasets and results show significant performance.

Keywords: Semantic relatedness · Query focused text summarization · Relevant information · Short query

1 Introduction

Text summarization finds information rich sentences for readers. The research area of text summarization is increasingly becoming popular due to the availability of huge amount of information. Text summarization presents the significant content to minimizing time and cost. It is considerably different from human summarization. Human summary can include significantly rich content and themes which is very difficult to include in case of automatic text summary. To find out the linguistic meaning of words and relations with other words, semantic measure is applied. Text summarization can be generic or user focused; generic summary summarizes the important content and query focused summary gives the summary specifically for user's interest. Extractive and abstractive methods are used to make summary. Abstractive method needs reformulation of sentences while extractive method extracts the sentences present in input text documents [1]. Here, we propose one semantic relatedness based text summarization method to extract semantically related sentences with the query.

Luhn in 1958 [2] first introduced text summarization by finding significant words from a text. Significant words are found by calculating the occurrence

B.U. Shankar et al. (Eds.): PReMI 2017, LNCS 10597, pp. 387–393, 2017.
https://doi.org/10.1007/978-3-319-69900-4_49

of a word in a text file. Based on the presence of significant words, sentences are ranked and extracted for summarization. In some recent approaches, Abadi et al. [3] (2015) used linguistic knowledge and expansion of content words. Content words includes noun, verb, adjective and adverb. The method finds semantic similarity between the content words along with the word-order similarity. Finally, they used combination model to select relevant sentences to the input query and also the sentences which are semantically very similar to the other high scoring sentences. We introduce semantic relatedness based query focused text summarization (SRQ) method to get well-defined summary according to the user's need. This SRQ method can work when the query words are not present in the input text. Present method can also perform when the query is short or does not contain enough information.

2 Proposed Semantic Relatedness Based Query Focused Text Summarization (SRQ Method)

Semantic relatedness measure: On the basis of semantic relatedness measure, important sentences are selected for summary purpose. In linguistics, semantics is the study of meaning and semantic relatedness gives the measure of how two words are related to each other. It is different from semantic similarity measure. Semantic similarity gives the measure of alikeness of two words or concepts and semantic relatedness gives more general concept than semantic similarity. For example, hand and finger are not semantically similar but they are semantically related. To find semantic relatedness between content words, WordNet is used. WordNet is a database used to find semantic relations (Miller 1998) [4] for English words. WordNet contains semantic network that defines different relations for content words. The following Table 1 gives different semantic relations for each content word present in WordNet database.

Hirst and St-Onge (HSO) [5] proposed one path based semantic relatedness measure using WordNet. Two words can be related in many ways like 'is-a', 'part-of', 'member-of' relations. For example, in Wordnet, hand and fingers are semantically related with 'part-of' relation. Semantic relatedness between two words includes all types of relations that are present in WordNet and finds the shortest path from the various semantic networks. They find the semantic relation between two content words by measuring the shortest path between them along with number of changes of direction in the shortest path. The following Fig. 1 shows the 'is-a' relation where shortest path and number of changes of direction between two words are (Hemorrhagic_fever and Respiratory_tract_infection) as found in WordNet:

Semantic relatedness between two words: Initially, pre-process the content words by doing stemming. The required method for finding semantic relatedness between two words is given in Eq. 1.

$$Score\,(w1, w2) = 2 * c - path\,length\,between\,w1\,and\,w2 - k\ * number\,of\ direction\,changes\,between\,w1\,and\,w2 \quad (1)$$

Table 1. Different semantic relations in WordNet

Different relations for Noun:		
Relation-type	Meaning	Example
Hypernym	Gives superordinate term	flower → angiosperm
Hyponym	Gives subordinate term	flower → african_daisy
Member Meronym	From group to their member	university → graduate_ school
Part Meronym	From whole to part	house → loft
Has-Instance	From concept to instance	wood → lignin
Member Holonym	From member to group	people → world
Part Holonym	From part to whole	face → head
Stuff-Of	From instance to concept	wood → beam
Antonym	Gives the opposite word	winner → looser
Different relations for Verb:		
Hypernym	From a verb to superordinate verb	run → travel_rapidly
Troponym	Gives manner relation	sleep → nap
Entails	A verb follows logically another verb	step → walk
Antonym	Gives the opposite word	start → stop
Different relations for Adjective:		
Antonym	Gives the opposite word for adjective	able → unable
Different relations for Adverb:		
Antonym	Gives the opposite word for adverb	kindly → unkindly

Here, $c = 8$ and $k = 1$ are considered as constants. If two words are same then the maximum semantic relatedness value of HSO will be 16 and minimum value is 0 [6]. We tested semantic relatedness score with different threshold values. Based on performance, the method uses average or higher semantic relatedness score by taking the threshold value as 8.

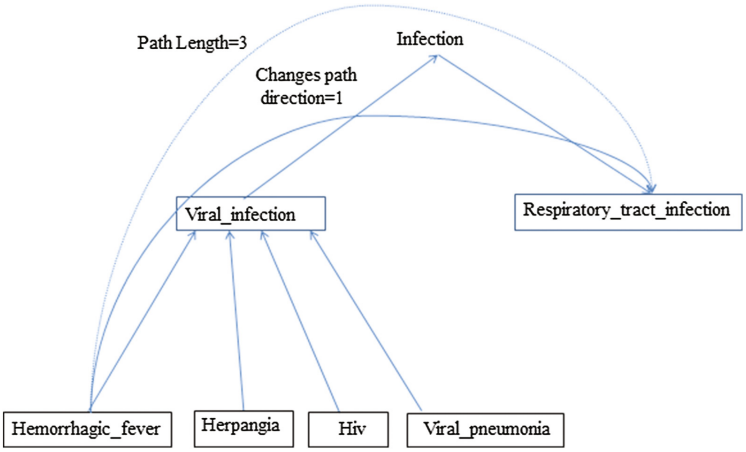

Fig. 1. Fragment WordNet concept hierarchy showing the path and direction changes of paths between $Hemorrhagic\_fever$ and $Respiratory\_tract\_infection$

Semantic relatedness between two sentences: To find out the semantic related two sentences, semantic relatedness is calculated for each of the content word of the first sentence S_1 with all the content words present in the second sentence S_2 and the maximum score is taken. After finding score for every word in the sentence S_1 with the words in S_2, we take maximum score as the score for S_1. The method to find semantic relatedness for the sentence S_1 with respect to S_2 is given in Eq. 2:

$$Score\,(S_1, S_2) = \max_{w1 \in S_1, w2 \in S_2} (score\,(w1, w2)) \tag{2}$$

Important sentence selection: Now, in query focused text summarization, we have a query with input text documents. Before applying semantic relatedness in SRQ method, we give priority to the sentences on the basis of following nine criteria to be considered as important sentences for the text summarization purpose. Semantic relatedness is calculated only for the important sentences.
Title Word Matching: If the words present in a sentence also occur in the title or heading of a text document, then that sentence can be considered as an important one.
Proper Noun: Proper noun or entity name gives more importance to a sentence. Hence, we take out the proper noun containing sentences.
Numerical Data: Presence of numerical data in a sentence always contains rich information.
Thematic Word: Thematic word means word that occur in a text file more frequently. Presence of thematic word makes the sentence important. We find top ten most frequent words from the text file and take out those sentences where any thematic word is present.

Noun Phrase: Presence of noun phrases in a sentence makes the sentence important. The method uses chunkparser to find noun phrases [7].
Font-based Word: Sentences containing words appearing as uppercase, bold, italics or underlined fonts are normally considered as more meaningful.
Cue Phrase: Sentences containing any cue phrase such as in conclusion, this letter, this report, summary, argue, purpose, development are most likely to be in summary.
Sentence Length: It is considered as longer sentence contains more information.
Sentence Position: Important sentences are usually present at the first and the last of the paragraph. We consider the first and the last sentences from paragraphs.

Semantic relatedness is calculated between the input text title (S_t) and an important sentence (S_i) present in input text document by using Eqs. 1 and 2. Again semantic relatedness is measured between query (S_q) and an important sentence (S_i) using the same Eqs. 1 and 2. We will consider those sentences where score is equal or above the defined threshold value.

Extracting Summary: To create the summary, common sentences are obtained from calculating semantic relatedness between text title and important sentences ($score\,(S_t, S_i)$) and query and important sentences ($score\,(S_q, S_i)$). To find out the set of sentences related to the title, the method uses Eq. 3.

$$T = \{s \mid s \in S_i,\ score\,(S_t, S_i) \geq 8\} \tag{3}$$

Similarly, to find out the set of sentences related to the query, the method uses Eq. 4.

$$Q = \{s \mid s \in S_i,\ score\,(S_q, S_i) \geq 8\} \tag{4}$$

Finally, summary can be found using the following method:

$$Summary_{\,sentences} = T \cap Q \tag{5}$$

3 Experiments

We use DUC 2005 and DUC 2006 datasets (http://duc.nist.gov), where each topic contains a query and a set of input text documents. Each text document contains newspaper or newswire information in English. DUC 2005 and 2006 datasets are particularly used for query-based text summarization purpose. Queries are based on real world complex questions, where answers not only contain date, name or quantity. Here, each dataset contains 50 documents and length of each summary has been restricted to 250 words only.

To evaluate the performance of SRQ method with other existing methods, ROUGE toolkit [8] is used. ROUGE compares similarity between candidate summary and reference summary. Candidate summary means summary produced from different methods and reference summary comes from DUC datasets.

This ROUGE consists of set of metrics, such as ROUGE-N (n-gram co-occurrence statistics), ROUGE-L (longest common subsequence), ROUGE-W (weighted longest common subsequence), ROUGE-S (skip-bigram co-occurrence statistics) and ROUGE-SU4 (skip-bigram based on maximum skip distance of 4, plus unigram). We compare our results with top-performing DUC 2005 and 2006 systems where systems have done their experiments particularly for query-based text summarization. Here, recall value of ROUGE-1 (unigram-based), ROUGE-2 (bigram-based) and ROUGE-SU4 are used for our experiment purpose. The following Figs. 2 and 3 shows the comparison of different ROUGE values of existing systems with SRQ method and finds that SRQ performs well in comparison with these existing systems.

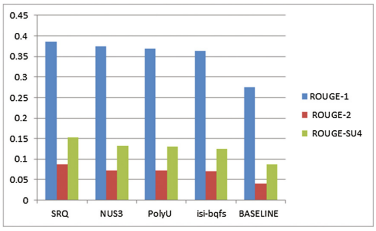

Fig. 2. Experimental results on DUC 2005 datasets

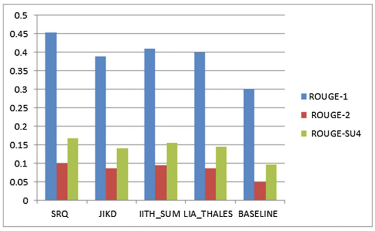

Fig. 3. Experimental results on DUC 2006 datasets

4 Conclusion and Future Work

The paper has presented a query focused text summarization method based on semantic relatedness. This SRQ method performs well for short query. The method is tested with different participating methods in DUC 2005 and DUC 2006 and gives better results. In future we can incorporate effective redundancy removal technique to get more query relevance and information rich summary.

References

1. Damova, M., Koychev, I.: Query-based summarization: a survey (2010)
2. Luhn, H.P.: The automatic creation of literature abstracts. IBM J. Res. Dev. **2**(2), 159–165 (1958)
3. Abdi, A., Idris, N., Alguliyev, R.M., Aliguliyev, R.M.: Query-based multi-documents summarization using linguistic knowledge and content word expansion. Soft Comput. **21**(7), 1–17 (2015)
4. Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K.J.: Introduction to wordnet: an on-line lexical database. Int. J. Lexicography **3**(4), 235–244 (1990)
5. Hirst, G., St-Onge, D., et al.: Lexical chains as representations of context for the detection and correction of malapropisms. In: WordNet: An Electronic Lexical Database, vol. 305, pp. 305–332 (1998)

6. Patwardhan, S., Banerjee, S., Pedersen, T.: Using measures of semantic relatedness for word sense disambiguation. In: Gelbukh, A. (ed.) CICLing 2003. LNCS, vol. 2588, pp. 241–257. Springer, Heidelberg (2003). doi:10.1007/3-540-36456-0_24
7. Bird, S., Klein, E., Loper, E.: Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. O'Reilly Media Inc., Sebastopol (2009)
8. Lin, C.-Y.: Rouge: a package for automatic evaluation of summaries. In: Text Summarization Branches Out: Proceedings of the ACL-2004 Workshop, vol. 8, Barcelona, Spain (2004)