Summarizing Long Historical Documents Using Significance and Utility Calculation Using Wordnet

Aakanksha Jain\(^1\) & Atul Gaur \(^2\)
\(^1\)Siddhi Vinayak College of Sci. & Hr. Education, Alwar
\(^2\)Siddhi Vinayak College of Sci. & Hr. Education, Alwar

Abstract: The current technology of automatic text summarization confers a critical part in the information retrieval and text grouping, and it gives the best answer for the information over-burden issue. When mulling over the size and number of documents which are accessible on the Internet and from alternate sources, the necessity for a very effective device on which produces usable summaries is clear. We show a superior algorithm utilizing lexical chain computation. The algorithm one which makes lexical chains a computationally possible for the user. Furthermore, utilizing these lexical chains the user will produce a summary, which is a great deal more successful contrasted with the arrangements accessible and furthermore nearer to the human created summary.

1. Introduction

A summary might be characterized as a text that is made from one or a considerable measure of texts that contains a major segment of the information inside the original text, which isn’t any more drawn out than half of the initial text. Text summarization [1] is the technique for refining the premier essential information from a source (or sources) to give a short form to a particular user (or users) and task (or tasks).

At the point when this should be possible by method for a pc, i.e. automatically, they call this Automatic Text summarization. Regardless of the established truth that text summarization has historically been focused on text input, the contribution to the summarization strategy additionally can be multi-media data, similar to pictures, video or sound, likewise as on-line information or hypertexts. Additionally, they will allude summarizing only one document or numerous ones. For this situation, this technique is comprehended as Multi-document summarization (MDS) [1] and furthermore the source documents for this situation are regularly in and the exceptionally single-language (monolingual) or a few languages (trans-lingual or multilingual).

The output of a summary system is additionally a concentrate (i.e. at the point when a choice of "critical" sentences of a document is performed) or dynamic, when the summary will work as a substitute to the initial document. They will conjointly distinguish between generic summaries and user-focused summaries (i.e. query-driven). The rest kind of summaries will function as a surrogate of the initial text as they will try to represent all relevant options of a source text. They’re text-driven and follow a bottom-up approach using (Information Retreival) IR techniques [1]. The user-focused summaries consider a specification of a user info want, such a subject or query. They follow a top-down approach using IR techniques.

2. Importance and Relevance of the Study

Summarization approaches are commonly, as specified, isolated into 2 bunches, text extraction and text deliberation. Text extraction intends to recognize the primary important entries in at least one documents, regularly utilizing standard factually based information retrieval methods expanded with pretty much shallow normal language preparing and heuristics. These entries, regularly sentences or expressions, are then removed and stuck along to make a non-repetitive summary that is shorter than the initial document with as meager data misfortune as could be expected under the circumstances.

Text deliberation, being the more troublesome task, is to dissect the initial text in a profound phonetic way, translate the text semantically into a formal delineation, and find new more brief ideas to clarify the text so create a fresh out of the plastic new shorter text, a dynamic, with comparative information content. The parsing and translation of a text is an old examination zone that has been explored for quite a long while. Around there we've a wide range of methods and systems beginning from word by word parsing to logical talk parsing and additionally a ton of measurable methodologies or a blend of all.
H. Gregory Silber, Kathleen F. McCoy[2] Describes 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.

This initial part of their implementation constructs an array of “meta chains” [2]. 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 [2] 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. 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” [2] of all the meta-chains is largest.

Fig1: Implementation proposed in paper- “Efficient Text Summarization Using Lexical Chains”

Regina Barzilay and Michael Elhadad ,[6] they investigate one technique to supply a summary of an original text while not requiring its full semantic interpretation [3], 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 [3] and of important sentences. Preliminary results indicate that quality indicative summaries are made. Unfinished issues are then identified. Plans to deal with these shortcomings are concisely presented.

Fig 2: Visual Representation of Lexical Chains

Nikita Munot and Sharvari S. Govilkar, 2014 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 [5].

The paper conjointly presents taxonomy of summarization systems and statistical and linguistic approaches [5] 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.

3. Proposed Concept

We utilized 4 Sample Data to produce base paper and proposed work base summary and contrast with human created summary and discover the estimation of review, aggregate words coordinated and time taken to create summary.

Lexical chains try to identify cohesion links between parts of text by identifying relations holding
between their words. Two bits of text are thought to be lexically related in the event that they utilize similar words, as well as on the off chance that they utilize semantically related words. This is a way to obtain certain structure of a text based on the distribution of its content.

Lexical chains provide a representation of text that has been widely used in a variety of Information Retrieval and Natural Language Processing tasks, from text segmentation to Word Sense Disambiguation, term weighting for IR tasks, topic detection, detection of malapropisms, hypertext generation, detecting topic and sub-topic shifts in texts, analysis of the structure of texts to compute their similarity, Passage Retrieval, Question & Answering and Automatic Summarization.

The utility of a lexical chain $L$ to a document $D$ is defined as

$$\text{util}(L,D) = \text{sig}(L) \sum_{k=0} \text{util}(L,D)$$

Like for chain $L_1$: $L_1$ has some significance and multiplied with the summation of the words that are included in chain $L_1$. Like a chain $S$: Computer-device-machine-memory

And candidate words are 10, than for this chain only $\text{sig}(S)[1+1+1+1+0+0+0+0+0+0]$.

### 3.1 Computing the Score of Chain

Now we need to find out the strong chains among them.

**Strong chains (scores) = Average (scores) + 2*standard deviation.**

Where standard deviation is applied on the utility of each chain. And average of each utility of scores has been taken.

If Number of the chains which has greater than to utility than average (scores) + 2*standard deviation, would be considered as strong chains.

### 3.1.4 Computing the Threshold value

This formula is used for computing the accepted chains and the formula is as follows,

$$\text{Threshold} = \sum_{k=0} \text{related}(w,L)/(\text{TotalChains} * 2)$$

### 4. Steps for proposed concepts

There are two steps behind this proposed concept

#### 4.1 Analysis of text summarization

**Fig 4: Analysis of Text Summarization**

**START**

READ THE LINES OF TEXT FROM THE FILE TO BE SUMMARIZED

REPEAT THE STEPS TILL ALL LINES PROCESSED

PERFORM THE TOKENIZATION IN ORDER TO REMOVE THE PUNCTUATIONS, COMMAS ETC...

PERFORM THE STEMMING IN ORDER TO OBTAIN THE BASE FORM USING THE WORDNET

PERFORM THE TAGGING OF EACH WORD IN THE LINE, NN FOR NOUN ETC...

A
5. Implementation and Result

5.1 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 software

Step 1: run the software to summarize the text

Step 2: select the sample data file

Step 3: enter the amount of percentage by which you want to summarize

Step 4: to compare our result we so upload the human generated summary

Figure 5.1 Run the Text summarization

Figure 5.2 selecting the input file

Figure 5.3 entering the percentage

Figure 5.4 Input of human generated file
Step 5 generating result of proposed work and base work

We have compared our proposed work with base paper work. All the steps are similar in that software also and we have generated the result. Here, we are presenting snapshot to see the difference between both of the softwares.

Similarly, we utilized 4 Sample Data to produce base paper and proposed work base summary and contrast with human created summary and discover the estimation of review, aggregate words coordinated and time taken to create summary. We utilized 30% of the lexical chains for summary of test information and created a table to share the result.

Table 1 Comparison Result Data

<table>
<thead>
<tr>
<th>Data</th>
<th>Recall of BasePaper</th>
<th>Recall of Proposed Work</th>
<th>Words Matched BasePaper</th>
<th>Words Matched Proposed Work</th>
<th>Total time taken (in sec) BasePaper</th>
<th>Total time taken (in sec) Proposed Work</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data 1</td>
<td>0.54</td>
<td>0.56</td>
<td>77</td>
<td>79</td>
<td>214</td>
<td>145</td>
</tr>
<tr>
<td>Data 2</td>
<td>0.44</td>
<td>0.72</td>
<td>37</td>
<td>61</td>
<td>112</td>
<td>48</td>
</tr>
<tr>
<td>Data 3</td>
<td>0.41</td>
<td>0.64</td>
<td>63</td>
<td>68</td>
<td>128</td>
<td>35</td>
</tr>
<tr>
<td>Data 4</td>
<td>0.53</td>
<td>0.64</td>
<td>102</td>
<td>123</td>
<td>105</td>
<td>32</td>
</tr>
</tbody>
</table>

Figure 5.5 proposed work result generation

Figure 5.6 generating result of base work

Figure 5.7 recall value of Summary generation
6. Conclusion and future work

In this thesis, we presented a method to find out the lexical chains as an efficient intermediate representation of our document. Along with WordNet API, our method also included the nouns and proper nouns in the computation of lexical chains. And the statistical calculations in our proposed methodology resulted in the better output as compared to the base paper.

The future scopes still exists in the following degrees:-

Document clustering: Document clustering is the main step forward towards the identification of the various representations in a multi-document collection. Accurate and better similarity measure imparts an important role in the process of determining the overall efficiency of the clusters or parts of the document. I have compute the similarity measure and calculations based on the overlap of nouns (used in same sense) and proper nouns between two words and the segments. I will in my further studies try to extend this similarity measure including verb, adverb etc.

Multi-document summarization: Multi-document summarization is still a very complex and difficult task. I will try to extend my research to generate the summaries related to multiple documents at a time.

6. References


