Automated Document Summarization of News Articles

Project Report: CS 6320 Natural Language Processing (Spring 2015)

By: Ninaad Pai
Email: ndp140030@utdallas.edu
Date of Submission: April 29th 2015

Abstract

Automated Text summarization is an emerging technique for finding out the summary of a text document. Due to the massive amount of information increasing day by day on the Internet; it is difficult for the user to go through all the information available on web. Summarization techniques need to be used to reduce the user's time in reading the whole information available on web. In this project, I propose an automatic text summarization technique using statistical features by use of successive threshold for finding the summary i.e. important sentences from the given input text document. Here the sentences are selected for based on the weight of the sentence. The weight of the sentences is calculated based on the statistical and linguistic features. This approach assigns scores to the sentences by weighting the features like term frequency, word occurrences, phrases etc. In my approach, the number of sentences present in the summary would be reduced to two to three sentences from the original text.

---

1. Introduction

Natural language is any informal language that can be learnt by a person. It differs from formal languages like computer programming languages, which have a proper structure and syntax. Natural language processing is the understanding of context of these natural languages and/or generating a text in natural language.

Information overload is one of the most serious problems of the present-day web. There are various approaches to addressing this problem. To curb this overload is to improve search results and document classification resulting in better file structuring and maintenance. The main objective of text summarization is to reduce a textual document to a shorter content that retains all pivotal points of the original document.

Reading the entire article, dissecting it and separating the important ideas from the raw text take time and effort. Reading and understanding a regular sized article requires at least five minutes. Automatic summary software summarizes texts of 500-5000 words in a split second. This allows the user to read less data but still receive the most important information and make solid conclusions. Today's computers are far more powerful then the human mind - and it is most likely that computer will create a good summary before the human will have a chance to look at the article.

Several summarization software work in any language- an ability that exceeds the abilities of most humans. Since summarizers work on linguistic models they are able to summarize texts in most languages- from English to Russian- without the need for manual intervention. This makes them ideal for people who read and deal with multi-lingual knowledge, or for people who need to translate their information but wish to keep it as short as possible.

Some software summarizes not only documents but also web pages. This highly improves productivity as it speeds up the surfing process. Instead of reading full news articles that can be comprising of unimportant information to the user - the summaries of such web pages can be precise and accurate - but still 20% the size of the original article.

Particularly there are two chief approaches of conducting summarization: extraction and abstraction. Extraction methods selects a subset of existing words, phrases or sentences in the original content to form a summary. On the contrary, abstraction methods build an internal semantic representation and use natural language generation techniques to create a summary closer to what a human might generate.

The research regarding summarization is largely active and many summarization techniques have been proposed.

2. Previous Work

Most recent research in multi-document summarization (MDS) focuses on sentence selection for increasing coverage and does not consider coherence of the summary (Section 2.1). Although coherence has been used in ordering of summary sentences (Section 2.2), this work is limited by the quality of summary sentences given as input. In contrast, GFLOW incorporates coherence in both selection and ordering of summary sentences. G-FLOW can be seen as an instance of discourse driven summarization (Section 2.3). There is prior work in this area, but primarily for summarization of single documents. There is some preliminary work on the use of manually created discourse models in MDS. Our approach is fully automated.

2.1. Subset Selection in MDS

Most extractive summarization research aims to increase the coverage of concepts and entities while reducing redundancy. Approaches include the use of maximum marginal relevance (Carbonell and Goldstein, 1998), centroid-based summarization (Saggion and Gaizauskas, 2004; Radev et al., 2004), covering weighted scores of concepts (Takamura and Okumura, 2009; Qazvinian et al., 2010), formulation as minimum dominating set problem (Shen and Li, 2010), and use of sub-modularity in sentence selection (Lin and Bilmes, 2011). Graph centrality has also been used to estimate the salience of a sentence (Erkan and Radev, 2004). Approaches to content analysis include generative topic models (Haghighi and Vanderwende, 2009; Celikyilmaz and Hakkani Tur, 2010; Li et al., 2011b), and discriminative models (Aker et al., 2010). These approaches do not consider coherence as one of the desiderata in sentence selection. Moreover, they do not attempt to organize the selected sentences into an intelligible summary. They are often evaluated by ROUGE (Lin, 2004), which is coherence-insensitive. In practice, these approaches often result in incoherent summaries.

2.2. Sentence Reordering

A parallel thread of research has investigated taking a set of summary sentences as input and reordering them to make the summary fluent. Various algorithms use some combination of topic-relatedness, chronology, precedence, succession, and entity coherence for reordering sentences (Barzilay et al., 2001; Okazaki et al., 2004; Barzilay and Lapata, 2008; Bollegala et al., 2010). Recent work has also used event-based models (Zhang et al., 2010) and context analysis (Li et al., 2011a). The hypothesis in this research is that a pipelined combination of subset selection and reordering will produce high-quality summaries. Unfortunately, this is not true in practice, because sentences are selected primarily for coverage without regard to coherence. This methodology often leads to an inadvertent selection of a set of disconnected sentences, which cannot be put together in a coherent summary, irrespective of how the succeeding algorithm reorders them. In our evaluation, reordering had limited impact on the quality of the summaries.

2.3. Coherence Models and Summarization

Research on discourse analysis of documents provides a basis for modeling coherence in a document. Several theories have been developed for modeling discourse, e.g., Centering Theory, Rhetorical Structure Theory (RST), Penn Discourse TreeBank (Grosz and Sidner, 1986; Mann and Thompson, 1988; Wolf and Gibson, 2005; Prasad et al., 2008). Numerous discourse-guided summarization algorithms have been developed (Marcu, 1997; Mani, 2001; Taboada and Mann, 2006; Barzilay and Elhadad, 1997; Louis et al., 2010). However, these approaches have been applied to single document summarization and not to MDS. Discourse models have seen some application to summary generation in MDS, for example, using a detailed semantic representation of the source texts (McKeown and Radev, 1995; Radev and McKeown, 1998). A multi-document extension of RST is Cross-document Structure Theory (CST), which has been applied to MDS (Zhang et al., 2002; Jorge and Pardo, 2010).

3. Proposed Approach

The following algorithm will consists of the proposed method:
1. Pre-Process the given document, segmenting the given document into sentences and then segment the each sentence into words.
2. Carry out an analysis of stop words, and then apply stop word removal and stemming procedure.
3. Assign a Score for each sentence in a document based on features as follows:

i) Assigning a weight for each word based on its level of importance.

ii) Calculate the total weight of each word (Twt) of a sentence by applying all feature's weight:

$$Twt = ((TermFrequency_{wt}(word)))/\ Total\ no\ of\ features.$$

iii) Calculate the total score of a sentence by summing the total weights of a word:

$$Score = (Twt(\dashv) + Twt(\dashv) + Twt(\dashv) + \cdots \ldots)/(Total\ no\ of\ words)$$

The Frequency Summarizer tokenizes the input into sentences then computes the term frequency map of the words. Then, the frequency map is filtered in order to ignore very low frequency and highly frequent words, this way it is able to discard the noisy words such as determiners, that are very frequent but don't contain much information, or words that occur only few times. And finally, the sentences are ranked according to the frequency of the words they contain and the top sentences are selected for the final summary.

4. Used Data Set

In this project, I have implemented a text summarizer using the NLTK library, which is operated on news articles extracted from BBC news feed available on the World Wide Web. The approach extracts one or more sentences that cover the main topic of the original document using the idea, that if a sentence contains the most recurrent words in a text, it probably covers most of the topics of text document.
Using a data set from the internet helps improve the result analysis as the news feed on the internet changes daily according to new articles published on the website. News feed websites such as BBC, Google, Washington Post, etc. provide large number of classified articles according to type of content which makes it easier to provide summaries of articles that a user wants to read.

5. Results and Analysis

I evaluated the system with documents containing at least 300 to 400 words. The summary generation is carried out as:

A. Preprocessing the document:
In the pre-processing stage the document mainly includes three steps:

a) Identification of sentences/word boundary: Generally in English language the sentence boundary is identified when periods occur, similarly using space that separates the words.

b) Stop word Elimination: Common words with no semantics and which do not provide important information for the final summary are eliminated.

c) Stemming: The purpose of stemming is to obtain the stem or radix of each word, which helps to improve its frequency.

B. Calculating weight of words using term frequency:
Term frequency depends on the occurrence of the word the text document using $log(n)$ where n denotes number of times a word occurs. The extracted statistical features weights

are involved in the assignment of word weight. The sum of the term frequency and default feature weights will be the final weight of a particular word weight.

C. Calculating the weight of the sentence:
Once all the term weights are computed then calculating the mean or average of all the term weight of the particular sentence can attain the sentence weight. Now these sentences along with their calculated weight are to be arranged as pair.

D. Generating the summary using iterative threshold:
Calculating the average weight of all the sentences would initially set the value of threshold. Then we select the sentences that satisfy the initial threshold; then again we calculate the threshold value by averaging the weight of selected sentences that satisfy the initial threshold. This procedure is repeated until number of sentences in summary would become equal to the number of paragraphs of a document.

E. Sample Experiment:
The experiment was conducted on different news articles and then analyzed by using manual summaries. Also, the summary generated by available summarizers such as Microsoft and online summarizers. Below is shown a sample text and its summary sentences.

Consider the text given below as a sample news article:

Ed Miliband has accused David Cameron and other world leaders of failing to stand by Libya, contributing in part to the crisis in the Mediterranean. The Labour leader said the UK had repeated the same mistakes "in post-conflict planning" for Libya as were made in Iraq and the current refugee situation should have been anticipated. Conservatives denounced the remarks. Mr Cameron called them "ill-judged". But Mr Miliband rejected claims he had politicised the issue as "nonsense". Setting out his foreign policy priorities in a speech in London, Mr Miliband also said Mr Cameron had presided over the "biggest loss of influence in a generation" and placed the UK's future in the European Union in doubt. The BBC's assistant political editor Norman Smith said Labour were making clear that they were not blaming the prime minister for the recent deaths in the Mediterranean. But Lib Dem leader Nick Clegg said any suggestion of "political point-scoring" on the back of a "total human tragedy" was "pretty distasteful".
Mr Miliband voted in favour of UN-authorised air strikes against former Libyan leader Muammar al-Gaddafi in 2011, designed to stop the slaughter of Libyan civilians in Benghazi.

The intervention led to the collapse of the Gaddafi regime but the country has descended into chaos since then. An estimated 800 people died when their boats sank off the Libyan coast on Sunday while more than 35,000 people are thought to have crossed from Africa to Europe this year, many of them being transited through Libya and departing from there. In a speech in London, the Labour leader suggested that the UK and the wider international community had let Libya down. It was meant to be about Ed Miliband's vision beyond the purely domestic. It became, in part, a row over whether or not he was

accusing David Cameron of being in some way culpable for the deaths of migrants in the Mediterranean. First there were the briefings and the counter-briefings by the unelected spin-doctors. Then those seeking elected office weighed in, with Tory representatives accusing Mr Miliband of being absolutely offensive; and their Labour opponents insisting the other side was manufacturing a row. And all this before the man who wants to govern had uttered a word. And all because of 29 words in bold in a Labour briefing document.

For the Tories, it's been an opportunity, once again, to question whether Ed Miliband has what it takes to be prime minister. For Labour, it's been a chance to try and portray their leader as a man who'll be at ease representing the UK abroad. And for the electorate, the speech and the spat have been a reminder that the challenges of Europe, migrants and the so-called Islamic State await whoever occupies No 10 once voters have delivered their verdict. "David Cameron was wrong to assume that Libya was a country whose institutions could be left to evolve and transform on their own," he said. "The tragedy is that this could have been anticipated. It should have been avoided. "And Britain could have played its part in ensuring the international community stood by the people of Libya in practice rather than standing behind the unfounded hopes of potential progress only in principle." A briefing note released by Labour ahead of the speech sparked a row after it suggested Mr Miliband would say the refugee crisis and tragic scenes in the Mediterranean this week were in part a direct result of the failure of post-conflict planning for Libya. Asked after his speech whether he was directly pinning the blame on Mr Cameron, Mr Miliband said the Conservatives were trying to "whip up a storm" and his position was "absolutely clear". He said the deaths were the result of people traffickers, repeating his view that a failure of post-conflict planning was "responsible for some of the situation we see in Libya".

He was later asked during a BBC Radio 1 Newsbeat event why he had not asked about Libya during Prime Minister's Questions for four years. Mr Miliband said he had raised the issue in the Commons in February. Speaking on a campaign visit to Lincoln, Mr Cameron said leaders needed to demonstrate "clarity, consistency and strength" in the face of a "dangerous and uncertain world". He added: "People will look at these ill-judged remarks and they will reach their own conclusions."

Fig 1. Actual article from BBC news feed

Ed Miliband: UK failures 'contributed to Libya crisis' - BBC News
Setting out his foreign policy priorities in a speech in London, Mr Miliband said Mr Cameron had presided over the "biggest loss of influence in a generation" and placed the UK's future in the European Union in doubt.
Policy guide: Where the parties stand Speaking on BBC Radio 5 live, Mr Clegg said a considerable amount of thought had gone into how to stabilise Libya after Gaddafi's fall, contrasting this with the aftermath of the 2003 Iraq invasion for which he said there had been "no planning at all".

Fig 2. Summary of news feed

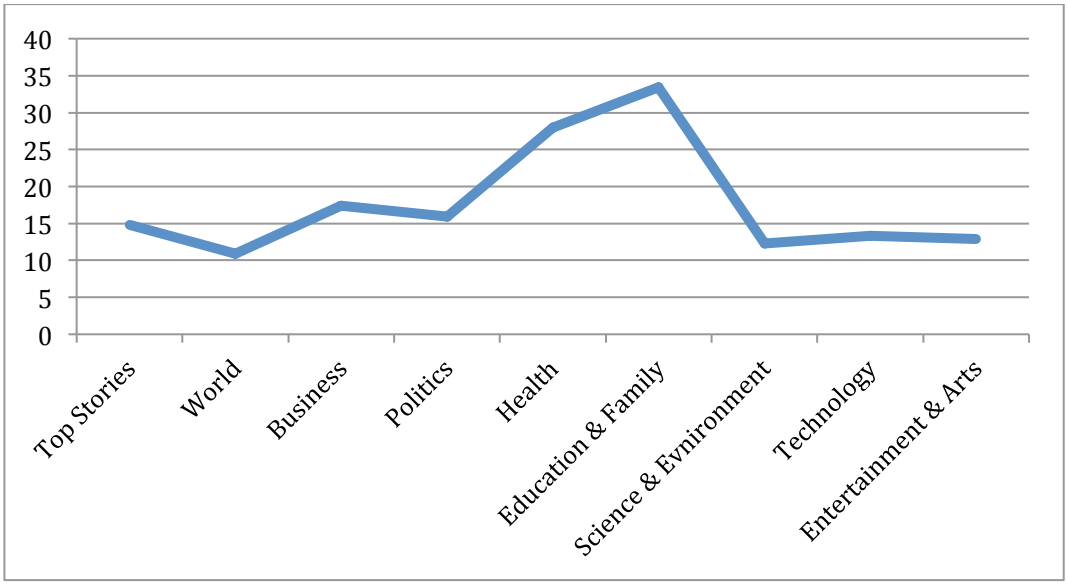

Fig 3. Chart for category v/ time (in seconds) for summarization of 10 news articles for each category

After using the program to perform document summarization on more than one document (or article in this scenario), following chart is given to depict time required (in seconds) for different genres of news articles. By observing Figure 3 we can deduce that news articles such as health, education, science, etc. a lot of varied phrases are used along with various proper nouns, which induces confusion in allotment of weightage and importance of word or phrase in order to generate a successful summary.

6. Conclusion

In this project, I proposed a summarization technique, which selects the important sentences based on statistical features. The number of sentences of this proposed summary is produced far less extensive than the actual article while still retaining the desired point to be made by the document. It was achieved by using successive threshold. Calculating the average weight of all the sentences would initially set the value of threshold. This procedure is repeated until number of sentences in summary would become equal to the number of paragraphs of a document. This approach gives satisfied

results when compared with commercial online summarizer and or proprietary products like the Microsoft summarizer. I expect that this method will pave the way for developing an efficient tool for text summarization. In future, adding some advanced features such as sentence-to-sentence cohesion, biased words etc. so that we could achieve still better precision and recall to increase the effectiveness of a summary.

7. References

1. Hingu, D. ; Shah, D. ; Udmale, S.S., "Automatic text summarization of Wikipedia articles", 2015 International Conference on Communication, Information & Computing Technology (ICCICT),

2. Inderjeet Mani, "Advances in Automatic Text Summarization", MIT Press, Cambridge, MA, USA, 1999.

3. Jade Goldstein, Mark Kantrowitz, Vibhu Mittal, and Jaime Carbonell, "Summarizing text documents: sentence selection and evaluation metrics", ACM SIGIR, 1999, pp 121–128.

4. E.H. Hovy and C.Y. Lin, "Automated Text Summarization in SUMMARIST", Proceedings of the Workshop on Intelligent Text Summarization, ACL/EACL-97. Madrid, Spain, 1997.

5. Kirill Kireyev," Using Latent Semantic Analysis for Extractive Summarization", In Proceedings of Text Analysis Conference, 2008.

6. Saeedeh Gholamrezazadeh, Mohsen Amini Salehi ,Bahareh Gholamzadeh A Comprehensive Survey on Text Summarization Systems , Computer Science and its Applications, 2009. CSA '09. 2nd International Conference on 10-12 Dec. 2009 .

7. Md. Mohsin Ali, Monotosh Kumar Ghosh, and Abdullah-AlMamum,"Multi-documentTextSummarization: SimWithFirst Based Features and Sentence Co-selection Based Evaluation", International Conference on Future Computer and Communication 2009

8. Cowie, J., Mahesh, K., Nirenburg, S., and Zajaz, R.,"MINDS Multilingual INteractive document summarization", In Working Notes of the AAAI Spring Symposium on Intelligent Text Summarization (pp. 131– 132). Menlo Park, CA: AAAI, 1998.

9. Vishal Gupta, Gurpreet Singh Lehal "A Survey of Text Summarization Extractive Techniques'', Journal Of Emerging Technologies In Web Intelligence, Vol. 2, Num. 3, August 2010.

10. WordSumamriser www.microsoft.com/education/autosummarize.mspx.

11. Ladda Suanmali, Naomie Salim and Mohammed Salem Binwahlan," Fuzzy Logic Based Method for Improving Text Summarization", (IJCSIS) International Journal of Computer Science and Information Security, Vol. 2, No. 1, 2009.

12. Karel Ježek, Josef Steinberger "Automatic Text Summarization (The state of the art 2007 and new challenges)".

13. Lin, C.Y.and Hovy, "Identify Topic by Position", in Proceedings of 5th Conference on Applied Natural Language Processing, March. 1997.

14. Jagadeesh J, Prasad Pingali, Vasudeva Varma "Sentence Extraction Based Single Document Summarization" Workshop on Document Summarization, March, 2005, IIIT Allahabad.

15. Niladri Chatterjee, Shiwali Mohan "Extraction-Based Single Document Summarization Using Random Indexing" 19th IEEE International Conference on Tools with Artificial Intelligence.