Document summarization using Wikipedia

Krishnan Ramanathan, Yogesh Sankarasubramaniam, Nidhi Mathur, Ajay Gupta

HP Laboratories
HPL-2009-39

Keyword(s):
Single Document Summarization, Wikipedia, ROUGE

Abstract:

Although most of the developing world is likely to first access the Internet through mobile phones, mobile devices are constrained by screen space, bandwidth and limited attention span. Single document summarization techniques have the potential to simplify information consumption on mobile phones by presenting only the most relevant information contained in the document. In this paper we present a language independent single-document summarization method. We map document sentences to semantic concepts in Wikipedia and select sentences for the summary based on the frequency of the mapped-to concepts. Our evaluation on English documents using the ROUGE package indicates our summarization method is competitive with the state of the art in single document summarization.

External Posting Date: February 21, 2009 [Fulltext] Approved for External Publication
Internal Posting Date: February 21, 2009 [Fulltext]

Published and presented at the First International Conference on HCI, Allahabad, India. Jan 20-23, 2009

© Copyright the First International Conference on HCI

Document summarization using Wikipedia

Krishnan Ramanathan, Yogesh Sankarasubramaniam, Nidhi Mathur, Ajay Gupta

HP Laboratories
24, Salarpuria arena, Hosur Road,
Adugodi, Bangalore, India
{krishnan_ramanathan,yogesh,nidhim,ajay.gupta}@hp.com

Abstract. Although most of the developing world is likely to first access the Internet through mobile phones, mobile devices are constrained by screen space, bandwidth and limited attention span. Single document summarization techniques have the potential to simplify information consumption on mobile phones by presenting only the most relevant information contained in the document. In this paper we present a language independent single-document summarization method. We map document sentences to semantic concepts in Wikipedia and select sentences for the summary based on the frequency of the mapped-to concepts. Our evaluation on English documents using the ROUGE package indicates our summarization method is competitive with the state of the art in single document summarization.

Keywords: Single Document Summarization, Wikipedia, ROUGE

1 Introduction

Mobile phones will be the onramp to the Internet for a large fraction of the world's population. However, Internet access on the move often happens in attention deficit situation and the user is capable of assimilating lower amount of information in a mobile context. Hence, the user interaction has to be adapted to the mobile scenario by presenting only the most relevant information to the user. Consequently, many mechanisms have been employed to simplify information presentation on mobile phones of which summarization is one [1]. Most research on document summarization has focused on multi-document summarization; this is more relevant for news sites where documents from multiple news agencies are available. Single document summarization is more relevant to simplifying information consumption on the Internet and on mobile phones, but has received lesser attention [2]. In this paper, we describe a novel, language independent, single document summarization system that uses Wikipedia for sentence selection.

2 Related work

The first paper on summarization appeared in 1958 [7]. Kupiec et.al. [9] proposed summarization by sentence extraction in SIGIR 1995. Today, there are broadly four approaches to summarization. The first uses heuristics for rating sentences (e.g. rate sentences that contain document title words higher). The second approach is corpus based and uses TF*IDF of words in the corpus to identify important words [1]. The third approach uses the structure of text, for instance the method of lexical chains [6]. The final approach is the knowledge-rich approach, our method falls into this category.

Microsoft Word has a summarization algorithm based on a statistical approach. It works only for English documents, this limits its usability. Newsinessence [4] is a multi-document summary of news. A language independent summarization algorithm based on a graph based ranking of sentences (to identify sentence similarity) is presented in [2]. In [1], the authors present five methods for summarizing parts of web pages for handheld devices using a combination of keyword extraction and text summarization.

3 The proposed method

Our approach to summarization is a proxy based approach that processes the document enroute to a mobile device and sends only the summary to the mobile device. Most summarization systems today extract parts of original documents and output them as summaries. Sentence extraction [9] is the most popular way of creating summaries. In this section, we describe a new approach to document sentence extraction using Wikipedia and its application to generating summaries.

3.1 Mapping sentences to Wikipedia concepts

Wikipedia has grown to become the largest encyclopedia with over 2 million articles. Our technique is based on using the Wikipedia corpus to find the document topic [10]. We first map individual sentences in the document to Wikipedia concepts. For doing this, the entire Wikipedia corpus is indexed using the Lucene engine. The sentence is then input as a query to the Lucene engine. The titles of the Wikipedia documents are extracted from the results to the query (“hits” in Lucene terminology). This process is illustrated in figure 1.

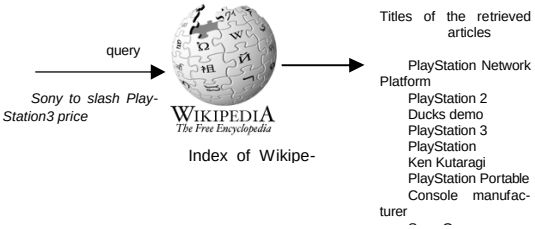

Figure 1 Querying a Wikipedia index

The above step is repeated for each sentence in the document and the number of "hits" for each Wikipedia concept is accumulated in a data structure (C++ multi-map).

3.2 Construction of the bipartite graph

Document sentences are mapped to semantic concepts in Wikipedia by virtue of query "hits" using the Lucene engine as described previously. This mapping can be captured as a bipartite graph, with one set of nodes (or vertices) denoting the document sentences and the other set of nodes denoting the Wikipedia concepts. An edge between a sentence node and a concept node indicates a mapping between the corresponding document sentence and Wikipedia concept, while the absence of an edge indicates that there is no mapping. Figure 2 illustrates this bipartite graph for a small document of three sentences.

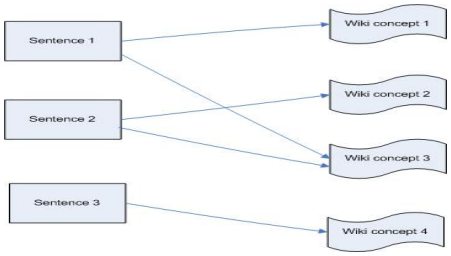

Figure 2 Construction of the sentence-Wikipedia concept bipartite graph

Let G denote the connection matrix of the bipartite graph. The matrix G is of size MXN in general, where M is the total number of sentences and N is the number of concepts in the given document. The goal of the summarization algorithm is to use G to derive the summary S.

3.3 The summarization algorithm

After the entire document has been processed and the bipartite graph has been created in the manner outlined above, we identify the Wikipedia concepts that got "hit" multiple times by different sentences in the document. The larger the number of hits, the more that particular concept is relevant to the summary and hence the sentences pointed to by that concept. The sentences in the document that mapped to the Wikipedia concepts with the largest number of hits are selected and output as the summary of the document. In our current system, the user can specify two thresholds for selecting concepts, concepts with hits above the maximum threshold and below the minimum threshold are excluded and sentences that map to them will not be included in the summary.

Figure 2 illustrates the process for a small document of three sentences. Wiki concept 3 is hit by both sentence 1 and 2. Assuming both the min and max threshold was set at 2 hits, sentence 1 and 2 are chosen as the two sentence summary of this three sentence document.

More concretely, we first compute the sum of the columns of the sentence-concept bipartite graph (the in-degree of the columns). We then apply a user specified threshold to select concepts whose column sums are above the specified threshold. These concepts are considered central to the document. This essentially amounts to selecting concepts that are mapped to by the highest number of sentences. For the graph of figure 2, the matrix is shown below

	C1	C2	C3	C4
S1	1	0	1	0
S2	0	1	1	0
S3	0	0	0	1

The sum of the elements in column 3 is the maximum (equals 2) and hence only concept 3 would be chosen as the representative concept if the threshold was 2 concepts. For column 3, row 1 and row 2 have non-zero entries and hence sentence 1 and sentence 2 are chosen as the summary.

4 Evaluation

To give a flavour of the summaries generated by our system, we now reproduce the summary produced by our system for a small document of 13 sentences using the algorithm outlined in section 3.3. This document was chosen from a news article.

Original document -

Running nose. Raging fever. Aching joints. Splitting headache. Are there any poor souls suffering from the flu this winter who havent longed for a pill to make it all go away? Relief may be in sight.

Researchers at Gilead Sciences, a pharmaceutical company in Foster City, California, reported last week in the Journal of the American Chemical Society that they have discovered a compound that can stop the influenza virus from spreading in animals. Tests on humans are set for later this year.The new compound takes a novel approach to the familiar flu virus. It targets an enzyme called neuraminidase, that the virus needs in order to scatter copies of itself throughout thebody. This enzyme acts like a pair of molecular scissors that slices through the protective mucous linings of the nose and throat. After the virus infects the cells of the respiratory system and begins replicating, neuraminidase cuts the newly formed copies free to invade other cells. By blocking this enzyme, the new compound, dubbed GS 4104, prevents the infection from spreading.

The summary produced by our system is as follows -

"Are there any poor souls suffering from the flu this winter who haven't longed for a pill to make it all go away? Relief may be in sight. Researchers at Gilead Sciences, a pharmaceutical company in Foster City, California, reported last week in the Journal of the American Chemical Society that they have discovered a compound that can stop the influenza virus from spreading in animals. The new compound takes a novel approach to the familiar flu virus. It targets an enzyme called neuraminidase, that the virus needs in order to scatter copies of itself throughout the body. After the virus infects the cells of the respiratory system and begins replicating, neuraminidase cuts the newly formed copies free to invade other cells"

Our system produced a summary of five sentences for the original 13 sentence document, all these sentences mapped to just one Wikipedia concept "Influenza". The sentences marked yellow in the original document are the ones removed by the sentence selection algorithm implemented in the summarizer.

We also evaluated our system on the DUC 2002 single document summarization task. In this task, there are 567 news articles and an expert written summary for each article. The evaluation was done using the ROUGE package [3]. We report the ROUGE_1 average recall numbers [5] at the 95 % confidence interval and for the first 100 words of each summary; these correspond to the ngram (1, 1) setting of ROUGE. We also used the "-m" option of the ROUGE toolkit for stemming the words. Summaries were generated for different thresholds for the Wikipedia concepts. The results are shown in table 1 below

Wikipedia concepts threshold	ROUGE recall
2	0.4680
3	0.4586
MAX	0.4368

Table 1 Recall results using ROUGE

The Wikipedia concept threshold is the number of hits a concept should receive before it is eligible to vote for a sentence. The MAX evaluation is done by using only a single concept with the highest number of hits. The best system in the DUC 2002 task (system S28) had a ROUGE-1 recall of 0.4804 [11], the above result with 2 concepts would place our system third in the DUC 2002 top performing systems. The

recall score dropped with a three concept threshold, this was mainly because 56 documents in this case had zero length summaries (no Wikipedia concept was mapped to 3 times by sentences in these documents).

5 Discussion and future work

The main limitation of this method is that a sentence could get chosen in the summary by virtue of getting mapped to only one concept. This concept might be a very generic concept pertaining to a high level topic (e.g. Sports). We would like multiple concepts to have a say in whether a sentence should be chosen. The other limitation is that this method does not offer an easy way to control the size of the summary. For instance, if the column score for a concept sums to N and we wish to have a summary of size M where M < N, the baseline method does not offer a principled way of doing this (we could use heuristics like compute the row sum of the sentences and order them by their row sums). We plan to overcome this limitation by making better use of the bipartite graph. In particular, we plan to devise an algorithm based on the intuition that important sentences in the graph map to important concepts and vice versa. Finally, we wish to evaluate the efficacy of our method in a multi-document summarization scenario such as in [8].

References

1. Orkut Buyukkokten et.al, Seeing the whole in parts: Text summarization for web browsing on handheld devices, WWW 2001.
2. Rada Mihalcea and Paul Tarau, A language independent algorithm for single and multiple document summarization, IJCNLP 2005.
3. ROUGE package for evaluating summaries, http://berouge.com/default.aspx.
4. Newsinessence, http://lada.si.umich.edu:8080/clair/nie1/nie.cgi
5. C.Y. Lin and E.H. Hovy, Automatic evaluation of summaries using n-gram co-occurrence statistics. In Proceedings of Human Language Technology Conference (HLT-NAACL 2003), Edmonton, Canada
6. R. Barzilay, M. Elhadad, Using lexical chains for text summarization, Proceedings of the ACL workshop on intelligent scalable text summarization, 1997, pp.10-17.
7. H.P. Luhn, The automatic creation of literature abstracts, IBM journal, April 1958.
8. P. Nguyen et.al., Summarization of multiple user reviews in the restaurant domain, Microsoft Research technical report, MSR-TR-2007-126.
9. Julian Kupiec, Jan Pedersen and Francine Chen, A Trainable Document Summarizer, SIGIR 1995.
10. E. Gabrilovich and S. Markovich, Overcoming the brittleness bottleneck with Wikipedia: Enhancing Text Categorization with Encyclopedic Knowledge, Proc. of the AAAI conference, 2006.
11. X. Wan, J. Yang and J. Xiao, Incorporating cross document relationships between sentences for single document summarization, ECDL 2006, LNCS 4172, pp. 403-414, 2006.