Summarizing Documents with the Aid of a Knowledge Repository

Abdullah Bawakid

Faculty of Computing and Information Technology, University of Jeddah, Jeddah, Saudi Arabia
abawakid@uj.edu.sa

Abstract: This paper describes the problem of information expansion and highlights the need for a knowledge repository that can aid processing information in the field of Natural Language Processing. The focus of this paper is on automatic text summarization. We described the reasons behind our choice of using Wikipedia as a knowledge base. We illustrated how it was pre-processed to make it usable for the task of summarization. Additionally, we explained how it is possible to derive a term concepts vector and why it is useful for the task we have at hand. Also included are the general design steps that were taken into account when building the system and how it was implemented.

[Bawakid A. **Summarizing Documents with the Aid of a Knowledge Repository.** *Life Sci J* 2017;14(2):48-51]. ISSN: 1097-8135 (Print) / ISSN: 2372-613X (Online). http://www.lifesciencesite.com. 6. doi:10.7537/marslsj140217.06.

Keywords: Knowledge repository; Documents Summarization; Extraction

Introduction

It is evident that there has been a huge information expansion within the past few decades especially in an electronic form. The number of available websites and the data they contain has been increasing exponentially. It would be impossible for humans to process the available data without the aid of an electronic tool. Thus, this signifies the acute need for developing methods, algorithms and tools to aid in processing the newly created information in addition to the old. Typically, when the available information that needs to be processed is limited, humans would process them based on the task they desire to accomplish. For example, processing the information present in a document or several documents may simply require classifying them based on what they contain. The process of classification itself may be accomplished via several methods such as identifying the dominant themes within a document, or considering all of themes referenced in the document and comparing them against a list of pre-defined categories. Another possible method a human may desire to apply on a set of documents is clustering, where the main themes are identified and then the documents are clustered based on what they contain. Another example for how humans process a document is summarization.

The above mentioned methods for processing documents and the information they contain vary highly in reality from the simple description given above based on the task at hand and the type of information present at hand. Again, when there is an abundant amount of information, it is simply not possible for humans to process all of the information present without the assistance of an intelligent or semi-intelligent system. The most common example for such systems that almost everybody nowadays uses is web search engines. Web engines target the problem of finding relevant information to the user needs in the web. Another problem that many people face due to the overload of information is the existence of similar or non-relevant information in the targeted documents that also contain important information. The document can be either too long, or there can be a relatively large number of documents which a user may not have the time to process alone. Thus, the need for automatic summarization systems emerges with the existence of such a problem.

Summarization systems evaluate a documents or several documents by identifying the most significant and relevant parts to the user and presenting them in a condensed form. The summarization task may also involve measuring the similarities between information from one or more sources and deciding whether to present this information to the user and in what form it is presented. For instance, consider a user that is following a single re-occurring event on the web. A good summarization system would have to provide the user with a relevant summary and also updates as new information become available. In this paper we propose a method for automatic summarization that attempts to exploit the semantic information present in a document or multiple documents to identify its main themes and present the relevant themes to the user.

We propose a semantic-based summarizer that targets text documents. The summarizer works by utilizing a knowledge base to help with semantically processing the content of the document. The knowledge base chosen to implement the designed algorithms is Wikipedia. However, the methods can be adapted to be applied on any other similar knowledge base. The structure of this paper is as follow: the next section gives some background about summarization in general and other technique similar previously implemented which are similar to what we describe here. The following section illustrates our methodologies and findings. The last section is the conclusion.

Background

The task of automatic documents summarization has been targeted by the NLP for several decades. Among the first attempts in this direction was the work of [1] in which the authors used mainly the location of sentences and paragraphs to devise summaries in the form of indices. In [2], they used the position of sentences and the existence of specific cue terms to decide the importance of sentences and their significance in a generated summary. Another attempt for documents summarization was made in [3] where the authors suggested an implementation of maps for text relations that link various parts of text in different portions of a document or several documents. These maps would then be used in their procedure to highlight the main themes of the text at hand and thus extracting summaries in the form of statements. In essence, the words distributions in the document were examined for achieving their goal. In another work [4], summaries via sentences extraction was implemented through the use of rhetorical analysis based algorithms and they were applied to generate only single document summaries. In [5], the summaries were formed by simply selecting the first sentences in the paragraphs, particularly the first and last, to include in their generated summaries.

In some other systems, others aspects of the documents and their text was investigated. When examining two sentences, it is possible that the two sentences may present the same meaning even though they carry different words. It is therefore essential that a process is devised in which sentences terms are not simply compared directly with each other without looking into the meaning they carry. An approach that attempts to automatically avoid this shortcoming was presented in [6]. Their algorithm was used to evaluate summaries and decide how similar two summaries through statistical based measures taking into account a domain independent paraphrase table that was previously built. This table was constructed from large bilingual corpora previously built with the aid of machine translated tools. In [7], another algorithm was presented to compute the semantic relations between different text fragments. They used a four-step process that examines the lexical chains between sentences and identifies the most important ones with the help of WordNet.

The frequency of a document terms and phrases were also considered in other summarization systems. In [8], a textual entitlement method was proposed and utilized in a summarization system. The performance they obtained indicated a %6.78 improvement in the accuracy of the summarization system when textual entitlement was utilized. In other work as in [9], graphs were used effectively to generate a representation of the text to be summarized. Some other systems were not entirely extractive, but rather abstractive. They differ mainly in their attempt to modify the sentences of the original text in an attempt to create a condensed and smaller version of them. An example for such a system is [10] where they attempted to merge common phrases into sentences with the aid of a statistical-based module they developed utilizing what they referred to as Sentences Fusion. Contrary to all of the above-mentioned methods, we propose a system in this paper that utilizes an external knowledge base. We describe in the next section the design and general implementation of our system.

Methodology

When a person evaluates a text document for the purpose of creating a summary, it is typical that this person attempts to read the text to comprehend it first before attempting the creation of a summary. Through reading and document comprehension, the person is able to tell what the main theme of the document is. The reader is able to tell how relevant the different parts of the read document are to its main theme after understanding the meaning of the whole document text. This process requires that the person should previously have acquired understanding the language of the documents text. In some cases such as when there are scientific or medical-related articles, it may be necessary to even have some background knowledge about the content of the document before being able to understand it, and eventually provide a summary.

The hypothesis we have is that systems that are purposely built for automatically summarizing documents would have some similarity in what they need to humans. In particular, we hypothesize that at the very least background knowledge is required to substitute humans language understanding and semantically link between the different portions of the documents text. It is therefore necessary to choose or build a suitable knowledge base that can be used with Natural Language Processing tasks, and especially documents summarization. One such suitable knowledge base is WordNet. It has the advantage of being machine readable and includes different levels of relationships between many of its entities. However, it is limited in its scope and expandability. New emerging concepts may not be necessarily included in WordNet. It is also possible to build a new knowledge or expand WordNet. However, the prohibitive cost that is anticipated for its construction and maintainability will be a challenge to overcome. These approaches may not preferable, especially with the existence of other alternatives such as Wikipedia.

Wikipedia is a large open source knowledge base. It is also the largest encyclopedia known to mankind. It has the advantages of being actively maintained by the web community. It also covers many aspects from diverse and different domains with a relatively good coverage. Additionally, emerging concepts and events are usually added in a short time span after they take place. Its articles structure is not uniform but many of its articles are attached to previously defined categories creating a semi-hierarchical structure for many portions of the encyclopedia. Due to its structure, it is not possible to use in its original state without applying a pre-processing stage first. In our system, after retrieving the latest dump of Wikipedia, the pre-processing starts by first parsing and cleaning the text of each article. We remove stop words, tags and markups which we deemed as unnecessary such as the articles tables. We also dismissed the short articles. We also marked the categories each article belongs to and saved them in our constructed database. Since each article discusses one main topic or event, we treated each article as a concept and considered the title of the article as the concept name. The content of the article are used to create a vector linking each term with its most representative concepts in a decreasing order using the term frequency-Inverse term frequency measure.

After applying the pre-processing stage, we should have the term concepts vector ready. This vector is used mainly for computing the semantic relatedness among different text extracts. These text portions can range from single terms, to sentences, or even a group of sentences. The resolution of comparison accuracy varies depending on the number of themes covered by the compared text portions. The algorithm we designed and implemented takes this into account when employing the term concepts vector to form summaries. The algorithm also considers other aspects of Wikipedia and takes them into account when performing comparisons between different text fragments. The computed semantic relatedness in our system considers not only the term concepts vector, but also the hierarchy of the articles which these concepts belong to.

The summarizer in our system is extractive, and works by selecting the most significant and representative sentences in the original documents to form the summary. The user is able to supply a query that may guide the system when forming the summary by shifting its focus on sentences with themes most relevant to the supplied query. The length of the summary is also controlled by the user. There is also a redundancy checking module implemented in the summarizer. The main tasks of the redundancy checking module are twofold: ensuring that any sentence added to the summary does not in its whole contain repetitive information which was already covered in the previously added sentences to the summary. The second task is deciding how much tolerance the system should give to significant sentences which partly contain repetitive information that is already existent in the summary. This is mainly affected by the maximum chosen length of the summary, the length of the source documents text, the length of the source document sentences, and the information they contain.

Conclusion

There is a need for an approach that captures the semantic relationships between the different segments of a document text and identifies its main themes. Employing a knowledge base for this task is a step in that direction. However, it is important to select a suitable knowledge base and define how its content is interpreted in the NLP domain. We presented in this paper our choice for a knowledge base, Wikipedia. We also illustrated how its content was preprocessed and extracted for use in automatic documents summarization. We also described how this knowledge base can be used for finding the semantic relatedness between text portions of varying length.

References

1. P. B. Baxendale, “Man-made index for technical literature - an experiment,” IBM J. Res. Dev., vol. 2, no. 4, pp. 354 – 361, 1958.
2. H. P. Edmundson, “New Methods in Automatic Extracting,” J. ACM JACM, vol. 16, pp. 264–285, Apr. 1969.
3. G. Salton, J. Allan, C. Buckley, and A. Singhal, “Automatic Analysis, Theme Generation, and Summarization of Machine-Readable Texts,” Science, vol. 264, no. 5164, pp. 1421–1426, Jun. 1994.
4. D. Marcu, “From Discourse Structures to Text Summaries,” in In Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization, 1997, pp. 82–88.
5. R. Brandow, K. Mitze, and L. F. Rau, “Automatic condensation of electronic publications by sentence selection,” Inf. Process. Manag., vol. 31, no. 5, pp. 675–685, Sep. 1995.
6. L. Zhou, C. Lin, D. S. Munteanu, and E. Hovy, “PARAEVAL: Using paraphrases to evaluate summaries automatically,” in IN:

PROCEEDINGS OF HLT-NAACL, 2006, pp. 447–454.

7. R. Barzilay and M. Elhadad, “Using Lexical Chains for Text Summarization,” in In Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization, 1997, pp. 10–17.
8. E. Lloret, O. Ferrández, R. Muñoz, and M. Palomar, “A Text Summarization Approach Under the Influence of Textual Entailment,” in Proceedings of the 5th International Workshop on Natural Language Processing and Cognitive Science (NLPCS 2008), 2008, pp. 22–31.
9. X. Wan and J. Xiao, “Towards a Unified Approach Based on Affinity Graph to Various Multi-document Summarizations,” in Research and Advanced Technology for Digital Libraries, L. Kovács, N. Fuhr, and C. Meghini, Eds. Springer Berlin Heidelberg, 2007, pp. 297–308.
10. R. Barzilay and K. R. McKeown, “Sentence Fusion for Multidocument News Summarization,” Comput Linguist, vol. 31, no. 3, pp. 297–328, Sep. 2005.

2/25/2017