Evaluation Of Automatic Text Summarizations Based On Human Summaries

Farshad Kiyomarsi a *

aIslamic Azad University Shahrekord Branch Iran

Abstract

The goal of this paper is to compare summaries generated by different automatic text summarization methods and those generated by human beings. To achieve this end, we did two series of experiments: in the first one, we employed automatically produced extractive summaries; in the second one, manually-produced summaries obtained by several English teachers were used. Our automatic summaries were obtained using Fuzzy method and Vector approach. Using Rouge evaluation system, we compared the manually-produced summaries and the automatically-produced ones. Rouge evaluation of generated summaries indicated the superiority of summaries produced by humans over the automatically produced summaries. On the other hand, the comparison between the generated summaries showed that summaries produced by Fuzzy method were much more acceptable and understandable compared to summaries produced by Vector approach. This can provide support for the replacement of manually generated summaries by summaries produced using Fuzzy method in certain cases where real time summaries are needed.

© 2015 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Keywords: evaluation, manually-produced, fuzzy method.

1. Introduction

Conventional technologies have become more and more insufficient for finding relevant information effectively. Nowadays, it is quite common that a keyword-based search on the Internet returns hundreds, or even thousands of hits, by which the user is often confused. Therefore, there is an increasing need for new technologies that can help
the user to sift through vast volumes of information, and to quickly identify the most relevant documents (Elhaddad, 2004). Document summarization refers to the task of creating document surrogates that are smaller in size but retain various characteristics of the original document (Eslami, 2002).

The information-overloading problem can be reduced by automatic summarization. Automatic text summarization is the technique in which a computer automatically creates such a summary. This process is significantly different from that of human based text summarization since human can capture and relate deep meanings and themes of text documents while automation of such a skill is very difficult to implement.

However, research into automatic text summarization has received considerable attention in the past few years due to the exponential growth in the quantity and complexity of information sources on the internet. Specifically, such text summarizer can be used to select the most relevant information from an abundance of text sources that result from a search by a search engine (Liu, 2001).

Many summarization models have been proposed previously. None of the models are entirely based on document structure, and they do not take into account of the fact that the human abstractors extract sentences according to the hierarchical document structure. While abstracts created by professionals involve rewriting of text, automatic summarization of documents has been focused on extracting sentences from text so that the overall summary satisfies various criteria: optimal reduction of text, coverage of document themes, and similar (Ferrier, 2001).

Automatic text summarization can be classified into two categories based on their approach:

1. Summarization based on abstraction
2. Summarization based on extraction.

Most of the works in this area are based on extraction method. In contrast to abstraction method which heavily utilizes computation power for natural language processing (NLP) with the inclusion of grammars and lexicons for parsing and generation, can be simply viewed as the process of selecting important excerpts (sentences, paragraph, etc.) from the original document and concatenating them into a more compact form. In other words, extraction is mainly concerned with judging the importance or the indicative power of each sentence in a given document (Liu, 2001).

Human abstractors extract the topic sentences according to the document structure from top level to low level until they have extracted sufficient information. The traditional automatic summarization techniques adopt the traditional salient features, but they consider the document as a sequence of sentences.

When human abstractors extract the sentences, they pay more attention to the range block with heading contains some bonus word such as “conclusion”, since they consider it as a more important part and more sentences are extracted. The cue feature of heading sentence is classified as rhetorical feature (Ferrier, 2001). We proposed the fuzzy analysis to consider the cue feature not only in sentence-level but also in paragraph and essay level.

With a large volume of text documents, presenting the user with a summary of each document greatly facilitates the task of finding the desired documents. A compact and concise summary enables the user to quickly get a rough idea of the document’s content, and to efficiently identify the documents that are most relevant to his/her needs.

The most common way to evaluate the informativeness of automatic summaries is to compare them with human-authored model summaries. For decades, the task of automatic summarization was cast as a sentence selection problem and systems were developed to identify the most important sentences in the input and those were selected to form a summary. It was thus appropriate to generate human models by asking people to produce summary extracts by selecting representative sentences. Systems were evaluated using metrics such as precision and recall (Salton, 1998), measuring to what degree automatic summarizers select the same sentences as a human would do.

Extracts or Abstracts? When asked to write a summary of a text, people do not normally produce an extract of sentences from the original. Rather, they use their own wording and synthesis of the important information. Thus, exact match of system sentences with human model sentences, as required for recall and precision metrics, is not at all possible. As the field turns to the development of more advanced non-extractive summarizers, we will clearly need to move to a more sophisticated evaluation method which can handle semantic equivalence at varying levels of granularity. The focus of this research is to compare summaries produced by different automatic text summarization approaches with human summaries and see which method is the most reliable one regarding the generation of a good summary.

The two automatic summarization methods used in this research are vector approach and fuzzy method. These methods extract sentences from the original documents based on specific linguistic features that will be mentioned
in detail in the following chapters. These extracted sentences will be included in the intended summaries. Moreover, summaries of the original documents were produced manually by several English teachers. Finally, the generated summaries were compared using Rouge evaluation system and also some human judges (several university faculty members in English department).

2. Background

Nowadays, people need much more information in work and life, especially the use of internet make information more easily gained. So, automatic text summarization draws substantial interest since it provides a solution to the information overload problem people face in this digital era. Text summarization (TS) is the process of identifying the most salient information in a document or set of related documents and conveying it in less space (typically by a factor of five to ten) than the original text. In principle, TS is possible because of the naturally occurring redundancy in text and because important (salient) information is spread unevenly in textual documents. Identifying the redundancy is a challenge that hasn’t been fully resolved yet.

A number of evaluation techniques for summarization have been developed. They are typically classified into two categories. Intrinsic measures attempt to quantify the similarity of a summary with one or more model summaries produced by humans. Intrinsic measures include Precision, Recall, Sentence Overlap, Kappa, and Relative Utility. All of these metrics assume that summaries have been produced in an extractive fashion. Extrinsic measures include using the summaries for a task, e.g., document retrieval, question answering, or text classification.

Traditionally, summarization has been mostly applied to two genres of text: scientific papers and news stories. These genres are distinguished by a high level of stereotypical structure. Attempts to summarize other texts, e.g., fiction or email, have been somewhat less successful. Recently, summarization researchers have also investigated methods of text simplification (or compression). Typically, these methods apply to a single sentence at a time. Simple methods include dropping unimportant words (determiners, adverbs).

Summarizing is the process of dealing with a large amount of information by comprising only the essential martial. It often occurs in everyday communication and it is an important and professional skill for some people. So, automated summarizing functions are urgently needed. Automatic text summarization aims at providing a condensed representation of the content according to the information that the user wants to get. Traditionally, Information Retrieval systems rank and present documents based on measuring relevance to the user query. But not all the information retrieved are really useful to the user, and it always takes a lot of time to read and select before the user get what he wants. Automatic text summarization became an exciting topic in Information Retrieval since it presents the user with summaries of the matching documents which can help the user identify which documents are most relevant to the user’s needs in a very short period of time.

Text Summarization is not a new idea. Research on automatic text summarization has a very long history, which can date back at least 40 years ago, from the first system built at IBM. Several researchers continued investigating various approaches to this problem through the seventies and eighties, especially very recently. Many innovative approaches began to be explored such as statistical and information-centric approaches, linguistic approaches and the combination of them. In next section, we will deal with the main approaches, which have been used or proposed. Some of the major types of summary that have been identified include:

- **indicative vs. informative**: It depends on whether a summary contains preliminary information or contains the main content of the document (Mani, 1999).
- **Generic vs. query-based**: It depends on whether a summary is on the central subject matter of the document or on special matter related to a user’s query (Mani, 1999).
- **Single-document vs. multi-document**: It depends on whether the input is a single document or multiple documents (Mani, 1999).

2.1 Automatic Text Summarization using a Machine Learning Approach

One of the approaches that has been recently used to perform automatic text summarization is the use of Machine Learning methods (Lin, 2002). The field of machine learning studies the design of computer programs able to induce
patterns, regularities, or rules from past experiences. Learner (a computer program) processes data representing past experiences and tries to either develop an appropriate response to future data, or describe in some meaningful way the data seen. A Machine Learning (ML) approach can be envisaged if we have a collection of documents and their corresponding reference extractive summaries. In some Machine Learning approaches, a trainable summarizer can be obtained by the application of a classical (trainable) machine learning algorithm in the collection of documents and its summaries. In this case the sentences of each document are modeled as vectors of features extracted from the text. The summarization task can be seen as a two-class classification problem, where a sentence is labeled as “correct” if it belongs to the extractive reference summary, or as “incorrect” otherwise. The trainable summarizer is expected to “learn” the patterns which lead to the summaries, by identifying relevant feature values which are most correlated with the classes “correct” or “incorrect”. When a new document is given to the system, the “learned” patterns are used to classify each sentence of that document into either a “correct” or “incorrect” sentence, producing an extractive summary.

A large variety of features can be found in the text-summORIZATION method using Machine Learning. These features include:

(a) **Mean-TF-ISF.** Since the seminal work of Luhn (Kiyomarsi, 2011) text processing tasks frequently use features based on IR measures (liu, 200; Kiani, 2002; Duck, 2006). In the context of IR, some very important measures are term frequency (TF) and term frequency–inverse document frequency (TF-IDF) (Jones, 1999). In text summarization we can employ the same idea: in this case we have a single document d, and we have to select a set of relevant sentences to be included in the extractive summary out of all sentences in d. Hence, the notion of a collection of documents in IR can be replaced by the notion of a single document in text summarization. Analogously the notion of document – an element of a collection of documents – in IR, corresponds to the notion of sentence – an element of a document – in summarization. This new measure will be called term frequency–inverse sentence frequency, and denoted TF-ISF(w,s) (Kiyomarsi, 2011). The final used feature is calculated as the mean value of the TF-ISF measure for all the words of each sentence.

(b) **Sentence Length.** This feature is employed to penalize sentences that are too short, since these sentences are not expected to belong to the summary (Kiani, 2002). We use the normalized length of the sentence, which is the ratio of the number of words occurring in the sentence over the number of words occurring in the longest sentence of the document.

(c) **Sentence Position.** This feature can involve several items, such as the position of a sentence in the document as a whole, its the position in a section, in a paragraph, etc., and has presented good results in several research projects (Liu, 2011; Kiani, 2002; Kiyomarsi, 2011; Neto, 2002; Duck, 2006). We use here the percentile of the sentence position in the document, as proposed by Nevill-Manning (Mani, 1999); the final value is normalized to take on values between 0 and 1.

(d) **Similarity to Title.** According to the vectorial model, this feature is obtained by using the title of the document as a “query” against all the sentences of the document; then the similarity of the document’s title and each sentence is computed by the cosine similarity measure (Jones, 1999).

(e) **Similarity to Keywords.** This feature is obtained analogously to the previous one, considering the similarity between the set of keywords of the document and each sentence which compose the document, according to the cosine similarity. For the next two features we employ the concept of text cohesion. Its basic principle is that sentences with higher degree of cohesion are more relevant and should be selected to be included in the summary (Eslami, 2002; Roak, 2006; Neto, 2002; Mani, 2001).

(f) **Sentence-to-Sentence Cohesion.** This feature is obtained as follows: for each sentence s we first compute the similarity between s and each other sentence s’ of the document; then we add up those similarity values, obtaining the raw value of this feature for s; the process is repeated for all sentences. The normalized value (in the range [0, 1]) of this feature for a sentence s is obtained by computing the ratio of the raw feature value for s over the largest raw feature value among all sentences in the document. Values closer to 1.0 indicate sentences with larger cohesion.

(g) **Sentence-to-Centroid Cohesion.** This feature is obtained for a sentence s as follows: first, we compute the vector representing the centroid of the document, which is the arithmetic average over the corresponding coordinate values of all the sentences of the document; then we compute the similarity between the centroid and each sentence, obtaining the raw value of this feature for each sentence. The normalized value in the range [0, 1] for s is obtained by computing the ratio of the raw feature value over the largest raw feature value among all sentences in the
document. Sentences with feature values closer to 1.0 have a larger degree of cohesion with respect to the centroid of the document, and so are supposed to better represent the basic ideas of the document. For the next features an approximate argumentative structure of the text is employed. It is a consensus that the generation and analysis of the complete rhetorical structure of a text would be impossible at the current state of the art in text processing. In spite of this, some methods based on a surface structure of the text have been used to obtain good-quality summaries (Duck, 2006). To obtain this approximate structure we first apply to the text an agglomerative clustering algorithm. The basic idea of this procedure is that similar sentences must be grouped together, in a bottom-up fashion, based on their lexical similarity. As result a hierarchical tree is produced, whose root represents the entire document. This tree is binary, since at each step two clusters are grouped. Five features are extracted from this tree, as follows:

(h) Depth in the tree. This feature for a sentence \( s \) is the depth of \( s \) in the tree.

(i) Indicator of main concepts. This is a binary feature, indicating whether or not a sentence captures the main concepts of the document. These main concepts are obtained by assuming that most of relevant words are nouns. Hence, for each sentence, we identify its nouns using a part-of-speech software (Ferrier, 2001). For each noun we then compute the number of sentences in which it occurs. The fifteen nouns with largest occurrence are selected as being the main concepts of the text. Finally, for each sentence the value of this feature is considered “true” if the sentence contains at least one of those nouns, and “false” otherwise.

(j) Occurrence of proper names. The motivation for this feature is that the occurrence of proper names, referring to people and places, are clues that a sentence is relevant for the summary. This is considered here as a binary feature, indicating whether a sentence \( s \) contains (value “true”) at least one proper name or not (value “false”). Proper names were detected by a part-of-speech tagger (Ferrier, 2001).

(k) Occurrence of non-essential information. Some words are indicators of non-essential information. These words are speech markers such as “because”, “furthermore”, and “additionally”, and typically occur in the beginning of a sentence. This is a binary feature, taking on the value “true” if the sentence contains at least one of these discourse markers, and “false” otherwise. The above features are used in Machine Learning approach to summarize the texts based on artificial intelligence.

3. Implementation and computational results

3.1 Implementation of computer methods (Vector and Fuzzy method)

In this part we try to implement two computer summarization methods such as Vector approach and Fuzzy approach. For each document, a summary was produced using one of the following two approaches: (1) An automatically-generated summary, formed by using Vector and Fuzzy method. This kind of summary is called an “ideal automatic summary”. (2) A manually-generated summary, produced by several English teacher by selecting the most relevant sentences of the text. This is called an “ideal manual summary”. Using ROUGE evaluation system, we compare the resulting summaries with summaries produced by human beings. Furthermore; the summaries will be compared by some human judges.

3.2 The Used Attribute in Text Summarization based on computer methods

We concentrate our presentation in two main points: (1) the set of employed features; and (2) the framework defined for the trainable summarizer, including the employed classifiers. A large variety of features can be found in the text-summarization literature. In our proposal we employ the following set of features which are used in the text-summarization method using Machine Learning and were mentioned in previous part:

(a) Mean-TF-ISF, (b) Sentence Length, (c) Sentence Position, (d) Similarity to Title, (e) Similarity to Keywords, (f) Sentence-to-Sentence Cohesion, (g) Sentence-to-Centroid Cohesion. (h) Referring position in a given level of the tree (positions 1, 2, 3, and 4), (i) Indicator of main concepts. (k) Occurrence of proper nouns, (l) Occurrence of anaphors, and (m) Occurrence of non-essential information.

In the Vector approach, the system consists of the following main steps:

(1) the system extracts the individual sentences of the original documents, using one the approaches
analysed in (Radeu, 1998), in this work it was used the regular expression approach; (2) each sentence is associated with a vector of predictor attributes (features), whose values are derived from the content of the sentence; (3) each sentence is also associated with one of the following two classes: **Summary** (i.e., the sentence belongs to the summary) or **Not-Summary** (i.e., the sentence does not belong to the summary).

This procedure allows us to cast text summarization as a classification, supervised learning problem. As usual in the classification task, the goal of the classification algorithm is to discover, from the data, a relationship (say, an IF-THEN classification rule) that predicts the correct value of the class for each sentence based on the values of the predictor attribute for that sentence.

More precisely, this casting leads to the following steps for solving a text summarization problem: (1) The system constructs a training set where each example (record) corresponds to a sentence of the original documents, and each example is represented by a set of attribute values and a known class. (2) A classification algorithm is trained to predict each sentence’s class (**Summary** or **Not-Summary**) based on its attribute values. (3) Given a new set of documents, the system produces a test set with predictor attributes in the same format as the training set. However, the values of the classes are unknown in the test set. (4) Each sentence in the test set is classified, by the trained algorithm produced in step (2), in one of the two classes: **Summary** or **Not-Summary**.

In the fuzzy method, a fuzzy inference system extracts sentences to be included in the summary. The analysis of the parameters important in summarization is done by designed fuzzy analyzers based on human perception of this problem. This text summarization system consists of (1) the text pre processor which extracts information needed for fuzzy analysis and (2) the analyzers which contain fuzzy inference systems. Weighted score of excellent sentences is computed for each sentence and the scores of relevance are ranked. Starting with the highest score, the sentences for which the relevance score is higher than the threshold value set are included in the summary. The process continues until the ratio of compression satisfies the limitation set initially. A MATLAB simulation based model has been developed for this approach. The advantage of this method over the Vector approach is that linguistic variables and human perception are taken into consideration. Comparison of performance measures indicates the superiority of the approach as compared to the commercially available summarizers.

### 4. Rouge Evaluations

The evaluation of the quality of a generated summary is a key point in summarization research. Traditionally evaluation of summarization involves human judgments of different quality metrics, for example, coherence, conciseness, grammaticality, readability, and content (Mani 2001). However, even simple manual evaluation of summaries on a large scale over a few linguistic quality questions and content coverage as in the Document Understanding Conference (DUC)would require over 3,000 hours of human efforts. This is very expensive and difficult to conduct in a frequent basis. Therefore, how to evaluate summaries automatically has drawn a lot of attention in the summarization research community in recent years. For example, Saggion et al. (2002) proposed three content-based evaluation methods that measure similarity between summaries. These methods are: *cosine similarity*, *unit overlap* (i.e. unigram or bigram), and *longest common subsequence*. However, they did not show how the results of these automatic evaluation methods correlate to human judgments.

To overcome these shortcomings, we employed a package, ROUGE, for automatic evaluation of summaries. ROUGE stands for Recall-Oriented Understudy for Gisting Evaluation. It includes measures to automatically determine the quality of a summary by comparing it to other (ideal) summaries created by humans.

Here the ROUGE-scores (Lin, 2002), (Lin, 2003) using different settings for the ROUGEeval (Lin, 2003) software and for the summarizer are presented. The agreements between the human written model summaries are also reported. As you see in the following tables, both the Rouge evaluation system and human judges gave a better score to human summaries than the automatically produced summaries.

| Table 1: Recall and Precision values in percentage for different ROUGE-scores |
|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| System          | Rouge 1         | Rouge 2         | Rouge 3         | Rouge 4         | Rouge L         | Rouge W-1.2     |
| Vector Method   | 21.0            | 5.8             | 1.9             | 0.7             | 19.8            | 6.2             |
|                 | 21.0            | 5.8             | 1.8             | 0.7             | 19.8            | 10.1            |
Table 1: Recall and Precision values in percentage for different ROUGE-scores. 100 documents from DUC 2004, each with four human written 10% summaries, were used. Summaries were truncated to 885 bytes when evaluating. Evaluated using ROUGE 1.5.5.

| Fuzzy Method | 28.2, 7.6, 2.6, 0.9, 25.2, 10.3, 29.6, 7.8, 2.6, 0.9, 25.3, 18.9 |
| Human | 32.5, 9.3, 3.0, 1.1, 28.8, 12.4, 32.5, 9.3, 3.0, 1.1, 28.9, 20.1 |

Table 2: four human written 20% summaries.

<table>
<thead>
<tr>
<th>System</th>
<th>Rouge 1</th>
<th>Rouge 2</th>
<th>Rouge 3</th>
<th>Rouge 4</th>
<th>Rouge L</th>
<th>Rouge W-1.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vector</td>
<td>23.6, 6.3, 1.7, 0.5, 22.3, 6.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Method</td>
<td>23.6, 6.3, 1.8, 0.5, 22.4, 10.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fuzzy Method</td>
<td>30.0, 7.9, 2.5, 0.8, 26.2, 10.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Human</td>
<td>31.5, 8.1, 2.7, 0.8, 27.5, 17.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2: As in Table 3, but with four human written 20% summaries.

The results are shown in Tables 1 to 4. Scores are given for ROUGE-1, word overlap between a system generated summary and human written “gold standard” summaries; ROUGE-L, longest common word subsequence between the system summary and the “gold standard” summaries; and ROUGE-W, also the longest common word subsequence, but weighted to give higher scores to words occurring consecutively.

Table 3: four human written 30% summaries.

<table>
<thead>
<tr>
<th>System</th>
<th>Rouge 1</th>
<th>Rouge 2</th>
<th>Rouge 3</th>
<th>Rouge 4</th>
<th>Rouge L</th>
<th>Rouge W-1.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vector</td>
<td>32.6, 6.9, 2.4, 1.1, 31.8, 7.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Method</td>
<td>31.6, 6.6, 2.3, 1.0, 33.1, 13.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fuzzy Method</td>
<td>39.4, 11.8, 3.8, 0.8, 30.5, 10.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Human</td>
<td>36.4, 11.6, 3.5, 0.7, 27.5, 17.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3: As in Table 1, but with four human written 30% summaries.

Table 4: four human written 40% summaries.

<table>
<thead>
<tr>
<th>System</th>
<th>Rouge 1</th>
<th>Rouge 2</th>
<th>Rouge 3</th>
<th>Rouge 4</th>
<th>Rouge L</th>
<th>Rouge W-1.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vector</td>
<td>34.4, 7.3, 2.5, 1.2, 33.2, 7.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Method</td>
<td>33.2, 7.0, 2.4, 1.1, 32.1, 13.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fuzzy Method</td>
<td>41.4, 13.3, 2.9, 0.9, 32.7, 10.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Human</td>
<td>38.2, 12.1, 2.6, 0.8, 29.5, 17.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4: As in Table 1, but with four human written 40% summaries.

Tables 1 to 4 show the total system performance in terms of precision for different compression rates in case of all models for English articles. As you see in the tables, humans produce the best summaries and Fuzzy Model gives better results than Vector approach since Fuzzy Model has a good capability to model arbitrary densities.

The fuzzy method does not perform as well as human abstractors, though it sometimes at least has higher ROUGE-3 and ROUGE-4 scores than the Vector method. The reason the fuzzy method is not as strong as human
abstractors is likely that there is insufficient data to gather reliable fuzzy statistics in the source text. ROUGE-L and ROUGE-W give fuzzy method higher scores than the simple vector texts. Removing stop words also tends to affect the vector summaries less than the fuzzy generated texts that tend to score highly by including appropriate amounts of common stop words. Stemming does not seem to have that much of an effect.

Adapting the summary to the evaluation procedure gives large effects. Favoring short words is for instance bad when evaluating without stop words. Removing inflections or stop words when summarizing if these are disregarded in the evaluations also improves recall, though of course the precision drops quite a lot, since the summary is in effect longer in words despite being the same number of bytes.

The system generated summaries used the most readable options, so no stemming was performed and all stop words were left in the text. Summaries generated using other options are even less readable. The human readers assigned three scores to each summary, one for “text flow”, one for “understandability” and one for “overall impression”. These represent roughly how easy it is to read the summary, if the reader understands what the summary is about and finally if the reader subjectively thinks this is a good summary. All scores were given on a scale from 1 (very bad) to 5 (very good).

Table 5: Manual evaluation of text quality.

<table>
<thead>
<tr>
<th></th>
<th>Text Flow</th>
<th>Understandability</th>
<th>Overall Impression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vector Method</td>
<td>1.9</td>
<td>1.7</td>
<td>1.2</td>
</tr>
<tr>
<td>Fuzzy Method</td>
<td>3.5</td>
<td>3.6</td>
<td>3.1</td>
</tr>
<tr>
<td>Human</td>
<td>4.1</td>
<td>4.2</td>
<td>4.1</td>
</tr>
</tbody>
</table>

The results show that all the judges gave the best score to the summaries produced by humans. The difference of scores between summaries obtained using fuzzy method and manually produced summaries is not very great. On the other hand, there is considerable difference between the scores given to two different automatically generated summaries. In other words, summaries produced using fuzzy method had a higher score and much closer to human generated summaries. This indicates that fuzzy method worked better in parts of the sentence which contained uncertainty due to the use of fuzzy quantities. Therefore by using fuzzy approach in text summarization, we can improve the effect of available quantities for choosing sentences used in the final summaries. In order word, we can make the summaries more intelligent.

5. Conclusion

As mentioned earlier, the goal of this paper was to compare summaries generated by different automatic text summarization methods and those generated by human beings. To achieve this end, we did two series of experiments: in the first one, we employed automatically produced extractive summaries; in the second one, manually-produced summaries obtained by several English teachers, were employed. Our automatic summaries were obtained using Fuzzy method and Vector approach. Using Rouge evaluation system, we have compared the manually-produced summaries and the automatically-produced ones. Also, summaries were evaluated by some human judges, some university faculty members in English department. Both Manual evaluation and Rouge evaluation of generated summaries indicated a superiority of summaries produced by humans over the automatically produced summaries.

On the other hand, The difference between the quality of human generated summaries and summaries produced by Fuzzy method was not considerable. This can provide support for the replacement of manually generated summaries by summaries produced using Fuzzy method in certain cases where real time summaries are needed. Besides, When the purpose is to get a general idea of the text or when we have a long text to read, automatic summaries, especially those generated by Fuzzy method are more economical, more appropriate and more efficient. Therefore, automatically generated summaries can be produced much faster than human summaries and they are more economical. Furthermore, if we employ a reliable method in summarizing the original document, automatic summaries can be as useful as human summaries.

Anyhow, automatically generated summaries may not be as coherent and intelligent as human summaries since
humans can think and decide on the best option. But, most of the time readers are able to understand the summaries using their common sense and make the summaries coherent in their mind. So, automatically generated summaries, provided that a good summarization method such as Fuzzy method is used, can be a good replacement for human summaries and make the task of dealing with large amount of information much easier and faster. Automatic text summarization methods can generate a summary of the original text that allows the user to obtain the main pieces of information available in that text, but with a much shorter reading time.

Moreover, With a large volume of text documents especially on the internet, presenting the user with a summary of each document greatly facilitates the task of finding the desired documents. A compact and concise summary enables the user to quickly get a rough idea of the document’s content, and to efficiently identify the documents that are most relevant to his/her needs.

References


Fishar S., Roark B (2006) Query-focused summarization by supervised sentences ranking and skewed word distribution , In proceedings of DUC.


Louisa Ferrier (2001), A Maximum Entropy Approach to Text Summarization, School of Artificial Intelligence , Division of Informatics , University of Edinburgh.


Noemie Elhadad (2004), User-Sensitive Text Summarization Thesis Summary, American Association for Artificial Intelligence (www.aaai.org), USA.


