On Definition of Automatic Text Summarization

Pashutan Modaresi  Stefan Conrad
Heinrich-Heine-University of Düsseldorf
Institute of Computer Science, Düsseldorf, Germany
{modaresi, conrad}@cs.uni-duesseldorf.de

ABSTRACT

Research in the continuously growing field of automatic text summarization is branched into extractive and abstractive approaches. Over the past few decades, major advances have occurred in extractive summarization and a smooth transition from extractive to abstractive approaches can be observed in recent years. Despite advances, a proper definition of automatic text summarization has been mainly neglected by researchers. In this work we emphasize on the importance of an appropriate definition of automatic text summarization. We review previous definitions on text summarization, investigate their properties and propose our own definition.

KEYWORDS

Text Summarization, Scientific Definition, Content Selection, Readability, Text Mining

1 INTRODUCTION

Modern research on automatic text summarization began almost 60 years ago with the work of Luhn [1] on automatic creation of literature abstracts. Over the years, much progress has been done in the development of algorithms to automatically summarize documents. Among the two major approaches of extractive and abstractive summarization, the first one has been investigated extensively in the literature. Despite satisfactory results in extractive summarization, researchers are focusing more and more on abstractive summarization in the recent years. Extractive summaries are usually created by concatenation of fragments of the source document with the addition of some post-processing. On the other hand, abstractive summaries are the result of rewriting or paraphrasing the source document, where a one-to-one mapping between the sentences of the source and target document is not always possible. Abstractive summarization is considered to be the natural way of summarizing performed by humans, which is one of the reasons for its popularity in the recent years.

As in other scientific disciplines, the first step to approach the automatic summarization problem is to define the problem itself. Previous research on extractive text summarization has revealed a disagreement in the community regarding the understanding and definition of the summarization problem. This can be acknowledged by the discrepancy between the various definitions of the problem proposed in the literature so far. In like manner, a similar flaw can also be observed in the field of abstractive text summarization. With this in mind, a need for a proper definition of automatic text summarization is being felt in the research community. This has been mainly neglected by researchers and consequently led to inconsistent foundations of this field. By inconsistent we mean that there is no single definition of a summary which has been agreed upon by researchers.

We impose several requirements on a proper definition of automatic text summarization:

- **Universality**: The definition should be valid for the known types of automatic text summarization. This includes indicative [2] and informative (according to functionality), single- [3] and multi-document [4] (according to input cardinality), hierarchical and flat [5] (according to output cardinality), extractive [6] and abstractive [7] (according to type), as well as generic, update [8] and query-guided (according to context).

- **Generality**: The definition should not apply any restrictions on the implementation details of the various stages of automatic
text summarization. This includes particularly the representation type of the source document, content selection, scoring, lexical selection or text realization.

- **Minimality**: The definition should be minimal, meaning that only a minimal number of properties or characteristics to reconstruct a text summary should be mentioned.

- **Exclusivity**: The definition should be exclusive, meaning that the definition allows degenerate cases that one may wish to exclude.

- **Repeatability**: The definition should be repeatable, meaning that applying the definition to a summary document as the input document, should either return a valid summary document, or prevent us from creating a new summary document if it is not possible.

In Section 2 we provide an overview of the existing definitions of automatic text summarization and investigate their properties. A commonly used concept in automatic text summarization is compression rate. The usage of this concept will be criticized in Section 3 and the concept of readability will be suggested as a substitution. Content selection as an important part of any summarization system will be discussed in Section 4. In Section 5 we propose our own definition of automatic text summarization and finally in Section 6 we will conclude our work.

2 RELATED WORK

Various definitions of automatic text summarization have been proposed in the literature. Despite some commonalities, these also include contradictions in some cases. Furthermore the proposed definitions are mostly applicable to a certain type of automatic text summarization and lack the properties introduced in Section 1. The lack of a proper definition of automatic text summarization can be due to a conservative attitude in the community, as Das and Martin state: “...it seems from the literature that any attempt to provide a more elaborate definition for the task [of automatic text summarization] would result in disagreement within the community”[3].

In this section we investigate the previous definitions of automatic text summarization proposed in the literature. We study their flaws and inspect the properties that make them inappropriate definitions of automatic text summarization. In fact Luhn did not propose a definition of text summarization, but rather he mentioned the purpose of a summary in the context of literature abstracts as: “the purpose of abstracts in technical literature is to facilitate quick and accurate identification of the topic of published papers. The objective is to save a prospective reader time and effort in finding useful information in a given article or report”[1].

Although this cannot be considered as a definition of automatic text summarization, but Luhn’s statement points to two important properties of a text summary. The first property is that the time and effort for reading a summary should be less than the one being consumed in reading the original document, and the second property is that a summary should accurately reflect the topic of the original document.

In 1995 Maybury defined an effective summary as “[a text that] distills the most important information from a source (or sources) to produce an abridged version of the original information for a particular user(s) and task(s).”[9].

By mentioning that a summary is produced from a source (or sources), Marbury covers the cases for single document and multi-document text summarization. Moreover the property that a summary is produced for a particular user(s) and task(s) can be interpreted as if the definition also covers the query-guided and generic cases. The most important property that Maybury’s definition lacks is exclusivity. The same definition could also be applied to the task of keyword extraction. Although keyword extraction is occasionally also considered to be a text summarization task, but in general it is a distinct branch of text mining, as different from text summarization, the target document in keyword extraction is a collection of keywords and
not a text document consisting of coherent sentences.

In his book *Automatic Summarization*, states Mani in 2001 that “a summary is a document containing several text units (words, terms, sentences or paragraphs) that are not present in the source document.”[10]. Consider a source document $d$ and a target document $t = d \cup \{s_1, \ldots, s_n\}$ constructed by addition of $n$ sentences to the source document. Clearly $t$ is not a summary document and thus Mani’s definition lacks the universality property of a proper definition.

The same problem also applies to the in the 2001 proposed definition of Sakai and Spark-Jones where they define a summary to be “a reductive transformation of a source text into a summary text by extraction or generation”[11]. The above definition, cannot be applied to the query-guided summaries and thus lacks the universality property of a definition.

In 2002, Radev et al. defined a summary to be “a text that is produced from one or more texts, that conveys important information in the original text(s), and that is no longer than half of the original text(s) and usually significantly less than that”[12]. This definition lacks the generality property of a proper definition, as a restriction on the size of the output document is applied.

An example of a recent attempt in defining automatic text summarization is the work of Torres-Moreno in 2014 where he defines an automatic summary as “a text generated by a software, that is coherent and contains a significant amount of relevant information from the source text. Its compression rate $\tau$ is less than a third of the length of the original document”[13].

Torres-Moreno’s definition points to an important property of the summary text, which is its coherence. The definition does not concretize relevant information and is not applicable to query-guided summaries, resulting in a lack of universality property. It also lacks the generality property by introducing compression rate as a part of the definition.

In Table 1 the properties of the discussed definitions are summarized. For the sake of readability, following abbreviations are used in the table: U: Universality, G: Generality, M: Minimality, E: Exclusivity, R: Repeatability

<table>
<thead>
<tr>
<th></th>
<th>U</th>
<th>G</th>
<th>M</th>
<th>E</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>Luhn [1]</td>
<td>X</td>
<td>✓</td>
<td>✓</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Maybury [9]</td>
<td>X</td>
<td>✓</td>
<td>✓</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Mani [10]</td>
<td>X</td>
<td>✓</td>
<td>✓</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Radev et al. [12]</td>
<td>X</td>
<td>X</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Torres-Moreno [13]</td>
<td>X</td>
<td>X</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

In the following section a commonly observed property of summaries, namely compression rate will be discussed and criticized as an unnecessary part of automatic summarization definition.

### 3 READABILITY

Usually in the context of automatic summarization we speak about compression rate $\tau$ which is defined as the ratio between the length of the summary and the length of the source document [13]:

$$\tau = \frac{|\text{summary}|}{|\text{source}|},$$

where $|$ is the length of the document in characters, sentences or words.

Various thresholds in the literature have been suggested for $\tau$. In [14] a summary is defined to be a text which is not longer than the half of the source document. At the same time, in [15] the optimal compression ratio for a summary is defined to be between 15% and 30%.

The use of compression ratio as an essential part of the definition has several disadvantages. The first one is that a direct comparison of the length of a summary and the length of the source document is not always possible. This comparison is commonly made based on the character, word, or sentence length. The choice between words, characters or sentences is normally made arbitrarily without any specific reasoning and typically depending on the underlying data set.
By using the count of sentences in source and summary document for the calculation of compression rate, the length of a sentence is completely ignored. So two sentences $s_1$ and $s_2$ with $|s_1| \gg |s_2|$ in the summary document will contribute the same amount to the computation of $\tau$.

On the other hand, computing the compression rate based on the length of documents in characters does not always return reliable information. As an example, consider two summaries $t$ and $t' = t + \text{”.”}$ where in $t'$ a punctuation mark is inserted at the end of $t$. According to the formula of compression rate, the compression rate of the first summary $t$ is less than the compression rate of the second one $t'$, although the summaries do not differ significantly and a human evaluator would not even notice the difference between the summary documents.

Perhaps the most reliable measure among the introduced ones is the length of documents in words. However, to only consider the length of documents in word has its own drawbacks. One main drawback is that the complexity of the words will be completely ignored. This does not cause remarkable problems in the case of extractive summarization, but in the context of abstractive summarization where paraphrasing and lexical selection are typical procedures, this may make the measure inconsistent. From the other side, depending compression rate on the words (tokens), makes the comparison between the compression rates of different algorithms a tedious task. Different approaches use different techniques for tokenization of the underlying text and this results in the fact that compression rate will be highly dependent on the underlying tokenization algorithms.

The second problem with the concept of compression rate is the need to define a specific threshold. This threshold is usually selected without any specific reasoning and mostly the selection is done in a way that the algorithm will return the best possible results for the underlying data set. As already discussed various thresholds are suggested in the literature and this arises the question whether specifying a hard-coded threshold is reasonable and consistent with the natural way humans perform summarization?

Alternatively, readability of a text seems to be a suitable substitution for the compression rate (note that in this work we do not use readability in its common sense meaning but rather we define it as a measure). In general, we expect from a summary to be more readable than the source document. The concept of readability can capture various dimensions such as the consumed time for reading the summary, its cohesiveness or the complexity of the vocabulary in it.

The introduction of the readability may bring from one side more vagueness into the definition of the automatic text summarization, but from the other side, it will distract the focus from the compression rate that in the late researches was heavily regarded as a key factor, resulting in the ignorance of the other dimensions connected to the readability of a summary. Much research is already done to measure the readability of a text. This includes for example the Flesch Reading-Ease Score [16] that considers the average sentence length and the average word length in syllables:

$$FRES = 206.835 - 1.015 \left( \frac{\text{total words}}{\text{total sentences}} \right) - 84.6 \left( \frac{\text{total syllables}}{\text{total words}} \right),$$  

(2)

or the Gunning Fog Index [17] that considers the average sentence length and a list of hard words (words with more than two syllables) in the text:

$$GFI = 0.4 \left[ \left( \frac{\text{words}}{\text{sentences}} \right) + 100 \left( \frac{\text{complex words}}{\text{words}} \right) \right].$$

(3)

Of course a direct takeover of the existing readability scores in the field of automatic text summarization is not appropriate and more sophisticated multi-factorial scores considering reading time or cohesion of the summaries have to be designed.

Some of the properties that should be required for a readable summary document are:

- Time: The amount of consumed time for reading a summary document should be
less that the amount of time that has to be consumed for reading the original document.

- Length: The length of a summary document should be less than the length of the original document (this can be modelled using the compression rate).

- Cohesiveness: A summary document should be at least as cohesive as the original document.

- Word Complexity: The average complexity of the words used in the summary text should be less than the average complexity of the word in the original document.

Note that the proposed properties for a readable summary are a subset of all possible properties that a readable summary document can exhibit. The intention of introducing properties for readability is to emphasize that length of a summary document (as considered in compression rate) is only one of the crucial dimensions in text summarization and other dimensions (specifically amount of time consumed for reading the summary document in comparison to the original document) have to be considered too.

Finally instead of compression ratio we suggest the use of readability ratio \( q = \frac{\rho(s)}{\rho(t)} \) to compare the ease of reading of the source document \( \rho(s) \) to the summary document \( \rho(t) \).

Beside the readability, content selection is another vital part of any definition on automatic text summarization. This will be discussed in more details in the following section.

4 CONTENT SELECTION

The decision which content to include in a summary is a critical one. The reason for this is that a summary will be finally read by a reader or a group of readers with diverse expectations from the content of the summary. Each reader has its own subjective preferences and expectations and creating a summary that fulfills all these subjective expectations is in practice impossible. With this is mind, summaries are commonly classified into generic and query-guided summaries. Generic summaries are the ones that estimate user’s information need. In contrast guided summaries include the user’s information need and ignore other irrelevant parts of the source document [18].

The user queries form a set \( Q \) of concepts, aspects, keywords or entities formulated by the user, representing the user’s needs. Generic summaries shall be considered as a subcategory of query-guided summaries where the user query is an empty set \( Q = \emptyset \).

We claim that a generic summary is of lesser use and almost impossible to evaluate manually. Consider a document \( d \) with its corresponding summary \( t \), where \( Q = \emptyset \). Now consider two users \( U_1 \) and \( U_2 \) aiming to manually evaluate the quality of content selection in the summary \( t \). At this stage users will answer the question whether the summary contains the most relevant (significant) information of the source document or not. This is exactly the place where the subjective preferences of \( U_1 \) and \( U_2 \) will be formulated on the fly. Thus for \( U_1 \) we will have \( Q_1 = \{ q_1, \ldots, q_n \} \) and for \( U_2 \) we will have \( Q_2 = \{ q'_1, \ldots, q'_m \} \).

Having two different sets of queries (probably formulated unconsciously by the users) will make the comparison of the scores calculated for \( t \) by \( U_1 \) and \( U_2 \) inconsistent.

By letting \( Q \) to be predefined, the vagueness of the phrases in the definition, such as a text containing significant amount of information, or a text containing important information in the original text will automatically disappear.

The question is now how to handle the case \( Q = \emptyset \)? More specifically, how should phrases such as important or significant information be interpreted, although no user preferences are pre-defined?

It is of course very restrictive to let the set of queries \( Q \) to be a non-empty set. This will lead to the lack of the universality property in the definition as the generic summaries can not be covered by such a restriction.

By letting the set \( Q \) to be an empty set, the concretization of the keywords such as “important” or “significant” in the definition
will be postponed to the implementation of the algorithm itself. As an example, consider the work of Luhn [1]. For selecting the most important sentences in the document, Luhn followed an approach based on the frequencies of the words. Although Luhn’s approach is classified under the generic summaries, but a query such as $q = \{ \text{sentences containing most frequent words in document} \}$, could also be used. Another example is the work of Edmundson et al. [19] that used the presence of cue words for content selection. This can also be formulated as a query set such as $q = \{ \text{"significant"}, \text{"impossible"}, \text{"hardly"} \}$. In this way the evaluation of the summaries by multiple human evaluators will also be possible, as the evaluators will all be using the same set of queries.

5 DEFINITION OF AUTOMATIC TEXT SUMMARIZATION

Having told the drawbacks of the previous definitions of automatic text summarization and after discussing important aspects of any automatic text summarization system, we propose our own definition:

Definition. Given a set $Q$ of queries and a set $K$ representing a knowledge base, automatic text summarization is a reductive transformation of a collection of documents $D$ with $|D| > 0$ into a single or multiple target documents, where the target document(s) are more readable that the documents in $D$ and contain the relevant information of $D$ according to $Q$ and $K$.

The above definition exhibits the required properties of a proper definition as discussed in Section 1.

The definition can be applied to indicative and informative summaries. Furthermore the set $D$ is defined as a collection of source documents with cardinality greater than or equal to one which covers both single and multi-document summarization. The set $Q$ consists of queries in the form of phrases, entities, sentences or keywords and it can also be an empty set. By this, both query-guided and generic summaries are covered. This all results in that the proposed definition has the universality property. The introduction of a knowledge base $K$ in the definition, covers the case of update summaries where the users need is only to get update information about a specific topic in a summarized manner. Note that $K$ can also be an empty set leading to the case of generic and query-guided summaries.

In the proposed definition it is also allowed to output multiple documents as the result of the summarization process. By this, the case of hierarchical summaries can be covered where the summary documents are ordered from more general and abstract ones, to more specific and detailed ones.

In the proposed definition, no restriction is applied to the implementation of the algorithms. Summarization is defined as a reductive transformation, meaning that the target document should be always shorter than the source document and by the introduction of the concept of readability and elimination of the concept of compression rate, the generality property of the definition is guaranteed.

The proposed definition also has the minimality property, as the elimination of any property in the definition will cause to the failure of the reconstruction of a text summary in a specific scenario.

We claim that the proposed definition has also the exclusivity property as the definition is not applicable to relevant fields such as keyword extraction, natural language translation or topic detection.

By use of the concept of readability, also the repeatability property of a proper definition is guaranteed. It is stated that the target document is more readable than the source document. Given a readable source document, it is always guaranteed that the target document is also readable. Assuming a source document consisting of two sentences and a target document consisting of one sentence produced from the source document, the question is now if the definition is still valid if we apply it to the target document consisting of one sentence? In other words, is it possible to summarize the target doc-
ument one more time, assuming the proposed definition? The requirement that the target document should always be more readable than the source document, results in the desired situation that some documents are not summarizable. This is especially the case where the source document is very short and compact enough that only a paraphrasing but not a summarization is possible.

6 CONCLUSION AND OUTLOOK

In this work we focused on a proper definition of automatic text summarization that has been neglected by many researchers in the community. We proposed the properties universality, generality, minimality, exclusivity, and repeatability (Section 1). Based on these properties, various existing definitions of automatic summarization in the literature have been investigated and criticized (Section 2).

We also discussed important aspects of a proper definition of text summarization such as readability (Section 3) and content selection (Section 4).

Finally in Section 5 we proposed our own definition of automatic text summarization and showed that the proposed definition exhibits all the properties of a proper definition introduced in Section 1. The proposed definition is by no means considered as a gold standard, however in the opinion of the authors it lacks many of the drawbacks of the previous definitions in the community.

Many other features such as the language of a summary, its coherence or the way it has to be evaluated have not been discussed in this work as a part of a proper definition. A more detailed investigation of the existing work on automatic text summarization is needed to examine the need for a more complex definition of automatic summarization.

Similar to any other definition, our proposed definition is also volatile in time and with respect to the community’s feedback.

REFERENCES


