TEXT SUMMARIZATION : AN OVERVIEW *

Elena Lloret
Dept. Lenguajes y Sistemas Informáticos
Universidad de Alicante
Alicante, Spain
elloret@dlsi.ua.es

Abstract

This paper presents an overview of Text Summarization. Text Summarization is a challenging problem these days. Due to the great amount of information we are provided with and thanks to the development of Internet technologies, needs of producing summaries have become more and more widespread. Summarization is a very interesting and useful task that gives support to many other tasks as well as it takes advantage of the techniques developed for related Natural Language Processing tasks. The paper we present here may help us to have an idea of what Text Summarization is and how it can be useful for.

Keywords: automatic text summarization; extracts and abstracts

1 Introduction

The World Wide Web has brought us a vast amount of on-line information. Due to this fact, everytime someone searches something on the Internet, the response obtained is lots of different Web pages with many information, which is imposible for a person to read completely. Although the attempts to generate automatic summaries began 50 years ago [40], in the recent years the field of automatic Text Summarization (TS) has experienced an exponential growth [25] [27] [46] due to these new tecnologies.

This paper addresses the current state-of-the-art of Text Summarization. Section 2 gives an overview of the field TS and we present the factors related to it. Section 3 explains the different approaches to generate summaries. In section 4 we present a number of Text Summarization systems existing today. Section 5 presents the common measures to evaluate those systems, whereas section 6 exposes the
tendency adopted these days in Text Sum-
marization. Finally, section 7 concludes
this paper and discusses future work.

2 What is TEXT SUM-
MARIZATION?

2.1 Definition and types

A summary can be defined as a text
that is produced from one or more texts,
that contains a significant portion of the
information in the original text(s), and
that is no longer than half of the original
text(s) [23]. According to [39], text sum-
marization is the process of distilling the
most important information from a source
(or sources) to produce an abridged version
for a particular user (or user) and task (or
tasks).

When this is done by means of a computer,
i.e. automatically, we call this Automatic
Text Summarization. Despite the fact that
text summarization has traditionally been
focused on text input, the input to the
summarization process can also be multi-
media information, such as images, video
or audio, as well as on-line information
or hypertexts. Furthermore, we can talk
about summarizing only one document or
multiple ones. In that case, this process is
known as Multi-document Summarization
(MDS) and the source documents in this
case can be in a single-language (monolin-
gual) or in different languages (translingual
or multilingual).

The output of a summary system may
be an extract (i.e. when a selection of
"significant" sentences of a document is
performed) or abstract, when the summary
can serve as a substitute to the original
document [15]. We can also distinguish
between generic summaries and user-
focused summaries (a.k.a query-driven).
The first type of summaries can serve as
surrogate of the original text as they may
try to represent all relevant features of
a source text. They are text-driven and
follow a bottom-up approach using IR
techniques. The user-focused summaries
rely on a specification of a user information
need, such a topic or query. They follow a
top-down approach using IE techniques.

Concerning the style of the output, a
broad distinction is normally made be-
tween indicative summaries, which are used
to indicate what topics are addressed in
the source text and they can give an brief
idea of what the original text is about,
and the informative summaries, which are
intended to cover the topics in the source
text [40][46].

2.2 Process of Automatic
Text Summarization

Traditionally, summarization has been de-
composed into three main stages [23]
[40][53]. We will follow the Sparck Jones
[53] approach, which is:

- **interpretation** of the source text to
  obtain a text representation,

- **transformation** of the text represen-
tation into a summary representation,
  and,
• finally, generation of the summary text from the summary representation

Effective summarizing requires an explicit and detailed analysis of context factors. Sparck Jones in [53] distinguishes three classes of context factors: input, purpose and output factors. We will described them briefly. For further information, see [53].

• Input factors. The features of the text to be summarized crucially determine the way a summary can be obtained. These fall into three groups, which are: text form (e.g. document structure); subject type (ordinary, specialized or restricted) and unit (single or multiple documents as input).

• Purpose factors. These are the most important factors. They fall under three categories: situation refers to the context within the summary is to be used; audience (i.e. summary readers) and use (what is the summary for?).

• Output factors. In this class we can group: material (i.e. content); format and style.

3 Approaches to Text Summarization

Although many different approaches to text summarization can be found in the literature [46], [55], in this paper we will only take into account the one proposed by Mani and Marbury (1999) [40]. This classification, based on the level of processing that each system performs, gives an idea of which traditional approaches exist. This can be suitable as a reference point from which many techniques can be developed. Based on the traditional approaches, summarization can be characterized as approaching the problem at the surface, entity, or discourse levels [40].

Surface level

This approach inclines to represent information taking shallow features and then selectively combining them together in order to obtain a salience function that can be used to extract information. Among these features, we have:

- **Thematic features** rely on word (*significant words*) occurrence statistics, so that sentences containing words that occur frequently in a text have higher weight than the rest. That means that these sentences are the important ones and they are hence extracted. Luhn (1958) [37] , who used the term frequency technique in his work, followed this approach. Before doing term frequency, a filtering task must be done using a stop-list words which contains words such as pronouns, prepositions and articles. This is the classical statistical approach. However, from a point of view of a corpus-based approach $td*idf$ measure (commonly used in information retrieval) is very useful to determine *keywords* in text.

- **Location** refers to the position in text, paragraph or any other particular section in the sense that they contain the target sentences to be included in the summary. This is usually genre-dependent, but there are two
basic general methods, which are lead-method and the title-based method. The first one consists of extracting only the first sentences, assuming that these are the most relevant ones, whereas the second considers that words in the headings or titles are positive relevant to summarization. Edmundson (1969) in [15] used this approach together with cue-word method which is explained later.

- **Background** assumes that the importance of meaning units is determined by the presence of terms from the title or headings, initial part of the text or a user’s query.

- **Cue words** and phrases, such as "in conclusion", "important", "in this paper", etc. can be very useful to determine signals of relevance or irrelevance. These kind of units can be detected automatically as well as manually.

**Entity level**
This approach attempts to build a representation of the text, modeling text entities and their relationships. The objective is to help to determine what is salient. This relations between entities include:

- **Similarity** occurs for example, when two words share a common stem, i.e. whose form is similar. This can be extended for phrases or paragraphs. Similarity can be calculated by vocabulary overlap or with linguistic techniques.

- **Proximity** refers to the distance between texts units. With that information is possible to establish entity relations.

- **Co-occurrence**: meaning units can be related if they occur in common texts.

- **Thesaural relationships among words** can be described as relationships like synonymy, hypernymy, part-of-relations (meronymy).

- **Coreference**. The idea behind coreference is that referring expressions can be linked so that, coreference chains can be built with coreferring expressions.

- **Logical relations** such as agreement, contradiction, entailment, and consistency.

- **Syntactic relations** are based on parse trees.

- **Meaning representation-based relations**, establishing relations between entities in the text as for example, predicate-argument relations.

**Discourse level**
The target of discourse level approaches is to model the global structure of the text and its relations in order to achieve communicative goals. The information that can be exploited at this level includes:

- **Format** of the document, such as hypertext markup or document outlines.

- **Threads of topics** as they are revealed in the text.
Rhetorical structure of text, representing argumentative or narrative structure. The idea behind this deals with the possibility to build the coherence structure of a text, so that the 'centrality' of textual units will reflect their importance.

To applied all these methods, two different appoaches can be taken. These techniques described so far can be developed by using linguistic knowledge or by applying machine learning techniques. Last ones have a support role, for example, in identifying the information to be applied at specific process stages such as interpretation or generation (for instance, they seem useful for training output sentence order).

4 Text Summarization Systems

Approaches presented so far are examples of pure techniques to apply, in order to develop summarization systems. The predominant tendency in current systems is to adopt a hybrid approach and combine and integrate some of the techniques mentioned before (e.g. cue phrases method combined with position and word frequency based methods in [24], or position, length weight of sentences combined with similarity of these sentences with the headline in [21]). As we have given a general overview of the classical techniques used in summarization and there is a large number of different techniques and systems, we are going to describe in this section only few of them briefly, considering systems as wholes. However, in table 1 some more systems are shown as well as their main features. To finish this section, the most recent approaches concerning summarization are mentioned.

Systems which have been selected to be broader described are the following: MEAD [51], WebInEssence [50], NeATS [36], GISTExter [20], and NetSum [54]

- MEAD [51]: This system was developed at the University of Michigan in 2001. It can produce both single and multi-document extractive summaries. The idea behind it is the use of the centroid-based feature. Moreover, two more features are used: position and overlap with the first sentence. Then, the linear combination of the three determines what sentences are most salient to include in the summary.

The system works as follows: first of all, MEAD uses the CIDR Topic Detection and Tracking system to identify all the articles related to an emerging event. CIDR produces a set of clusters. From each cluster a centroid is built. Then, for each sentence, three values are computed: the centroid score, which measures how close the sentence to the centroid is; the position score indicates how far is the sentence with respect to the beginning of a document; and finally, the overlap with the first sentence or title of the document by calculating
tf*idf between the given sentence and the first one. Then all these measures are normalized and sentences which are too similar to others are discarded (based on a cosine similarity measure). Any sentence that have not been discarded would be included in the summary.

- **WebInEssence** [50]: This system was also developed at the University of Michigan in 2001. It is more than a summarization system. It is a search engine to summarize clusters of related Web pages which provide more contextual and summary information to help users explore retrieval results more efficiently. A version of MEAD [51] was used in the development of this Web-based summarizer, so that the features used to produce extracts are the same as the ones used in MEAD. The overall architecture of the system can be decomposed into two main stages: the first one behaves as a Web-spider that collects URLs from the Internet and then it groups the URLs into clusters. The second main stage is to create a multi-document summary from each cluster using the MEAD centroid-algorithm.

- **NeATS** [36] was first developed in 2001 by the University of Southern California’s Information Sciences Institute. It is tailored to the genre of newspaper news. Its architecture consists of three main components: content selection, content filtering and content presentation. The goal of content selection is to identify important concepts mentioned in a document collection. The techniques used at this stage are term frequency, topic signature or term clustering. For content filtering three different filters are used: sentence position, stigma words and redundancy filter. To achieve the latter, NeATS uses a simplified version of MMR [9] algorithm. To ensure coherence of the summary, NeATS outputs the final sentences in their chronological order. From this system, iNeATS [35], i.e., an interactive multi-document summarization system that provides a user control over the summarization process, was later developed.

- **GISTexter** [20]: This system was developed in 2002 and produces single and multi-document extracts and abstracts by template-driven IE. The system performs differently depending on working with single document or multi-document summarization. For single-documents, the most relevant sentences are extracted and compressed by rules learned from a corpus of human-written abstracts. In the final stage, reduction is performed to trim the whole summary to the length of 100 words. When multi-document summarization has to be done, the system, based on **Information Extraction (IE) techniques**, uses IE-style templates, either from a prior
set (if the topic is well-known) or by ad-hoc generation (if it is unknown). The templates generated by CICERO IE system are then mapped into text snippets from the texts, in which anaphoric expressions are resolved. These text snippets can be used to generate coherent, informative multi-documents summaries.

- **NetSum** [54]. Different from the other approaches previous shown, NetSum, developed in 2007 by Microsoft Research Department, bets on single-document instead of multi-document summarization. The system produces fully automated single-document extracts of newswire articles based on neuronal nets. It uses machine learning techniques in this way: a train set is labeled so that the labels identify the best sentences. Then a set of features is extracted from each sentence in the train and test sets, and the train set is used to train the system. The system is then evaluated on the test set. The system learns from a train set the distribution of features for the best sentences and outputs a ranked list of sentences for each document. Sentences are ranked using RankNet algorithm [8].

After the brief description of the former systems, the reader can take a look at table 1, where a few more systems can be found. In order to understand what each column means, the following information is provided. In the first column (**SYSTEM [REF.], YEAR**) the name of the system with its reference and year is written; the second column (**# INPUTS**) distinguish between single document or multi-document summarization (both values are also possible. That means the system can perform both inputs). Next column (**DOMAIN SYSTEM**) indicates the genre of the input, that is, whether it is designed for domain-specific topics or for non-restricted domain. The fourth column (**FEATURES**) lists the main characteristics and techniques used in each system, and finally, last column (**OUTPUT**) represents whether the summary generated is either an extract or an abstract (with its variants. For some particular systems both values are also possible).
Table 1: Text Summarization Systems

<table>
<thead>
<tr>
<th>SYSTEM</th>
<th># INPUTS</th>
<th>DOMAIN SYSTEM</th>
<th>FEATURES</th>
<th>OUTPUT</th>
</tr>
</thead>
</table>
| Luhn [37], 1958         | single-document| domain-specific          | - term filtering and word frequency is carried out (low-frequency terms are removed)  
|                         |                | domain-specific (technical articles)  
|                         |                | - sentences are weighted by the significant terms they contained  
|                         |                | - sentence segmentation and extraction is performed                   | extracts¹ |
| Edmundson [15], 1969    | single-document| domain-specific          | - techniques used in this approach are: word frequency, cue phrases, title and heading words and sentence location  
|                         |                | domain-specific (articles)  
|                         |                | - it uses a corpus-based methodology                                  | extracts |
| ADAM [48], 1975         | single-document| domain-specific          | - Cue phrases  
|                         |                | domain-specific (chemistry)  
|                         |                | - term frequencies  
|                         |                | - sentence selection and rejection                                     | indicative abstracts² |
| ANES [7], 1995          | single-document| domain-specific          | - term and sentence weighting (tf*idf)  
|                         |                | domain-specific (news)  
|                         |                | - non-anaphora resolution  
|                         |                | - first sentences are added to the summary                             | extracts |
| Barzilay & Elahadad [4], 1997 | single-document | unknown | - topic identification of the text by grouping words into lexical chains³  
|                         |                |  
|                         |                | - sentence extraction is helped by strong chains identification  
|                         |                | - non-anaphora resolution                                              | extracts |

¹Although the output in Luhn’s work is called abstract, it is more correct to say extract, as sentences from the source document take part into the summary.

²Output sentences are edited to produce somewhat different to the original ones, but not new sentences are generated.

³Lexical chains are sequences of related terms grouped together by text cohesion relationships (e.g. synonymy or holonomy)
<table>
<thead>
<tr>
<th>SYSTEM</th>
<th># INPUTS</th>
<th>DOMAIN SYSTEM</th>
<th>FEATURES</th>
<th>OUTPUT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boguraev &amp; Kennedy [6], 1997</td>
<td>single-document</td>
<td>domain-independent</td>
<td>- linguistic techniques to identify salient phrasal units (topic stamps)</td>
<td>capsule</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- content characterisation methods to reflect global context (capsule overview)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- anaphora resolution</td>
<td>overview⁴</td>
</tr>
<tr>
<td>DimSum [3], 1997</td>
<td>single-document</td>
<td>unknown</td>
<td>- it uses corpus-based statistical NLP techniques</td>
<td>extracts</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- multi-word phrases are extracted automatically</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- conceptual representation of the text is performed</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- discourse features of lexical item within a text (name aliases, synonyms, and morphological variants) are exploited</td>
<td></td>
</tr>
<tr>
<td>Marcu [41], 1997</td>
<td>single-document</td>
<td>domain-specific (news)</td>
<td>- it uses text coherence models</td>
<td>extracts</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- RST⁵ trees are built</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- this kind of representation is useful to determine the most important units in a text</td>
<td></td>
</tr>
<tr>
<td>SUMMARIST [24], 1998</td>
<td>single-document</td>
<td>domain-specific (news)</td>
<td>- symbolic concept-level world knowledge is combined with NLP processing techniques</td>
<td>extracts</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- stages for summarization are divided in: topic identification, interpretation and generation</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- it is a multi-lingual system</td>
<td></td>
</tr>
</tbody>
</table>

⁴A capsule overview is not a conventional summary, i.e. it does not attempt to output document content as a sequence of sentences. It is a semiformal representation of the document.

⁵Rhetorical Structure Theory
<table>
<thead>
<tr>
<th>SYSTEM [REF.], YEAR</th>
<th># INPUTS</th>
<th>DOMAIN SYSTEM</th>
<th>FEATURES</th>
<th>OUTPUT</th>
</tr>
</thead>
<tbody>
<tr>
<td>SUMMONS [44], 1998</td>
<td>multi-document</td>
<td>domain-specific (online news)</td>
<td>-its input is a set of templates from the Message Understanding Conference -key sentences from an article are extracted using statistical techniques and measures - planning operators such as contradiction, agreement or superset are used to synthesize a single article</td>
<td>extracts and abstracts</td>
</tr>
<tr>
<td>FociSum [28], 1999</td>
<td>single-document</td>
<td>domain-independent</td>
<td>- it merges information extraction (IE) with sentence extraction techniques - the topic of the text (called foci in this system) is determined dynamically from name entities and multiwords terms</td>
<td>extracts</td>
</tr>
<tr>
<td>MultiGen [5], 1999</td>
<td>multi-document</td>
<td>domain-specific (news)</td>
<td>- it identifies and synthesizes similar elements across related text from a set of multiple documents - it is based on information fusion and reformulation - sets of similar sentences are extracted (themes)</td>
<td>abstracts</td>
</tr>
<tr>
<td>Chen &amp; Lin [26], 2000</td>
<td>multi-document</td>
<td>domain-specific (news)</td>
<td>- it produces multilingual (only English and Chinese) news summaries - monolingual and multilingual clustering is done - meaning units detection such as topic chains or linking elements is performed - similarity between meaning units is measured</td>
<td>extracts</td>
</tr>
</tbody>
</table>

---

6SUMMarizing Online NewS articles
8for more detailed information see[44]
<table>
<thead>
<tr>
<th>SYSTEM</th>
<th># INPUTS</th>
<th>DOMAIN SYSTEM</th>
<th>FEATURES</th>
<th>OUTPUT</th>
</tr>
</thead>
<tbody>
<tr>
<td>CENTRIFUSER [30] [29], 2001</td>
<td>multi-document</td>
<td>domain-specific</td>
<td>- it produces query-driven summaries</td>
<td>extracts</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(health-care articles)</td>
<td>- clustering is applied by SIMFINDER tool</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- it is based on document topic tree</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(each individual document is represented by a tree data structure)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- composite topic trees are designed</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(they carry topic information for all articles)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- query mapping using a similarity function enriched with structural</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>information from the topic trees is done</td>
<td></td>
</tr>
<tr>
<td>Cut &amp; Paste [22], 2001</td>
<td>single-document</td>
<td>domain-independent</td>
<td>- it uses sentence reduction and sentence combination techniques</td>
<td>abstracts&lt;sup&gt;9&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- Key sentences are identified by a sentence extraction algorithm that covers</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>this techniques: lexical coherence,</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>tf*idf score, cue phrases and</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>sentence positions</td>
<td></td>
</tr>
<tr>
<td>MEAD [51], 2001</td>
<td>multi-document</td>
<td>domain-specific</td>
<td>- it is based on sentence extraction through the features: centroid score,</td>
<td>extracts</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(news)</td>
<td>position and overlap with the first sentence</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- sentences too similar to others are discarded</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- experiments with CST&lt;sup&gt;10&lt;/sup&gt; and Cross-document subsumption have been made</td>
<td></td>
</tr>
</tbody>
</table>

<sup>9</sup>In this case, summaries are generated by reformulating the text of the original document

<sup>10</sup>Cross-Document Structural Relationships proposes a taxonomy of the informational relationships between documents in clusters of related documents. This concept is similar to Rhetorical Structure Theory (RST)
<table>
<thead>
<tr>
<th>SYSTEM</th>
<th># INPUTS</th>
<th>DOMAIN SYSTEM</th>
<th>FEATURES</th>
<th>OUTPUT</th>
</tr>
</thead>
</table>
| NeATS<sup>11</sup> [36], 2001 | multi-document | domain-specific (news) | - to select important content it uses: sentence position, term frequency, topic signature, term clustering  
- to avoid redundancy it employs MMR<sup>12</sup> technique [9]  
- to improve cohesion and coherence stigma words and time stamps are used | extracts     |
| NewsInEssence [49], 2001   | multi-document | domain-specific (online news) | - clusters are built through CIDR Topic detection and tracking component  
- it is based on the Cross-document Structure Theory (CST)  
- its summaries are produced by MEAD [51] | personalized extracts |
| WebInEssence [50], 2001    | multi-document | domain-independent    | - it is a Web-based summarization and recommendation system  
- it employs centroid-based sentence extraction technique  
- it uses similar techniques to NewsInEssence [49] but applied to Web documents | extracts and personalized summaries |
| COLUMBIA MDS [43], 2002    | multi-document | domain-specific (news) | - it is a composite system that uses different summarizers depending on the input: MultiGen for single events or DE<sup>MS</sup><sup>13</sup> for multiple events or biographical documents  
- statistical techniques are used | extracts and abstracts |

<sup>11</sup>Next Generation Automated Text Summarization  
<sup>12</sup>Maximal Marginal Relevance  
<sup>13</sup>Dissimilarity Engine for Multidocument Summarization
<table>
<thead>
<tr>
<th>SYSTEM [REF.], YEAR</th>
<th># INPUTS</th>
<th>DOMAIN SYSTEM</th>
<th>FEATURES</th>
<th>OUTPUT</th>
</tr>
</thead>
</table>
| Copeck et al. [12], 2002 | single-document | domain-specific (biographies) | - it uses *machine learning* techniques  
- *keyphrases* are extracted and ranked  
- text is segmented according to sentences that talk about the same topic | extracts |
| GISTexter [20], 2002 | single and multi-document | domain-specific (news) | - for single-document summarization, it extracts key sentences automatically using the technique of single-document decomposition  
- for multi-document summaries, it relies on CICERO IE system to extract relevant information by applying templates that are determined by the topic of the collection | extracts and abstracts |
| GLEANS [13], 2002 | multi-document | unknown | - it performs document mapping to obtain a database-like representation that explicits their main entities and relations  
- each document in the collection is classified into one of these categories: *single person*, *single event*, *multiple event* and *natural disaster*  
- the system is IE based | headlines, extracts and abstracts |
| NTT [21], 2002 | single-document | unknown | - it employs the *Support Vector Machine* (SVM) machine learning technique to classify a sentence into relevant or non-relevant  
- it also uses the following features to described a sentence: *position*, *length*, *weight*, *similarity* with the headline and *presence* of certains verbs or prepositions | extracts |

continue on next page
Table 1 – continued from previous page

<table>
<thead>
<tr>
<th>SYSTEM</th>
<th># INPUTS</th>
<th>DOMAIN SYSTEM</th>
<th>FEATURES</th>
<th>OUTPUT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Karamuftuoglu [31], 2002</td>
<td>single-document</td>
<td>unknown</td>
<td>- it is based on the extract-reduce-organize paradigm</td>
<td>extracts</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- as a pattern matching method it uses <em>lexical links</em> and <em>bonds</em>¹⁴</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- sentences are selected by SVM technique</td>
<td></td>
</tr>
<tr>
<td>Kraaij et al. [33], 2002</td>
<td>multi-document</td>
<td>unknown</td>
<td>- it is based on probabilistic methods: <em>sentence position</em>, <em>length</em>, <em>cue phrases</em></td>
<td>headlines and extracts</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- for headline generation, noun phrases are extracted</td>
<td></td>
</tr>
<tr>
<td>Lal &amp; Reuger [34], 2002</td>
<td>single-document</td>
<td>unknown</td>
<td>- it is built within the GATE¹⁵ framework</td>
<td>extracts</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- it uses simple Bayes classifier to extract sentences</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- it resolves anaphora using GATE’s ANNIE¹⁶ module</td>
<td></td>
</tr>
<tr>
<td>Newsblaster [42], 2002</td>
<td>multi-document</td>
<td>domain-specific (online news)</td>
<td>- news articles are clustered using <em>Topic Detection and Tracking</em> (TDT)</td>
<td>extracts</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- it is a composite summarization system</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(it uses different strategies depending on the type of documents in each cluster)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- it uses similar techniques to [43]</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- thumbnails of images are displayed</td>
<td></td>
</tr>
<tr>
<td>SumUM [52], 2002</td>
<td>single-document</td>
<td>domain-specific (technical articles)</td>
<td>- shallow syntactic and semantic analysis</td>
<td>abstracts</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- concept identification and relevant information extraction</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- summary representation construction</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>and text regeneration</td>
<td></td>
</tr>
</tbody>
</table>

¹⁴A lexical link between two sentences is a word that appears in both sentences. When two or more lexical links between a pair of sentences occur, a lexical bond between them is constituted.

¹⁵General architecture for Text Engineering, University of Sheffield

¹⁶A Nearly New Information Extraction System
Table 1 – continued from previous page

<table>
<thead>
<tr>
<th>SYSTEM [REF.], YEAR</th>
<th># INPUTS</th>
<th>DOMAIN SYSTEM</th>
<th>FEATURES</th>
<th>OUTPUT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alfonseca &amp; Rodríguez[1], 2003</td>
<td>single-document</td>
<td>domain-specific (articles)</td>
<td>- relevant sentences identification (with a genetic algorithm) - relevant words and phrases from identified sentences are extracted - coherence is kept for the output</td>
<td>very short extracts (10 words)</td>
</tr>
<tr>
<td>GISTSumm [47], 2003</td>
<td>single-document</td>
<td>domain-independent</td>
<td>- it is based on the gist(^{17}) of the source text - it uses statistical measures: <em>keywords</em> to determine what the <em>gist sentence</em> is - by means of the gist sentence, it is possible to build coherent extracts</td>
<td>extract</td>
</tr>
<tr>
<td>K.U. Leuven [2], 2003</td>
<td>single and multi-document</td>
<td>unknown</td>
<td>- <em>topic segmentation</em> and <em>clustering</em> techniques are used for multi-document task - topic segmentation, <em>sentence scoring</em> (weight, position, proximity to the topic) and <em>compression</em> are used for single-document summarization</td>
<td>extracts</td>
</tr>
<tr>
<td>Univ. Lethbridge [10], 2003</td>
<td>single and multi-document</td>
<td>unknown</td>
<td>- it performs topic segmentation of the text - it computes lexical chains for each segment - sentence extraction techniques are performed - it uses heuristics to do some surface repairs to make summaries coherent and readable</td>
<td>extracts</td>
</tr>
<tr>
<td>LAKE [14], 2004</td>
<td>single-document</td>
<td>domain-specific (news)</td>
<td>- it exploits keyphrase extraction methodology to identify relevant terms in the document - it is based on a supervised learning approach - it considers linguistic features like name entity recognition or multiwords</td>
<td>very short extracts</td>
</tr>
</tbody>
</table>

\(^{17}\)The most important passage of the source text
<table>
<thead>
<tr>
<th>SYSTEM [REF.], YEAR</th>
<th># INPUTS</th>
<th>DOMAIN SYSTEM</th>
<th>FEATURES</th>
<th>OUTPUT</th>
</tr>
</thead>
</table>
| MSR-NLP Summarizer [56], 2004 | multi-document | domain-specific (news) | - its objective is to identify important events as opposed to entities  
- it uses a graph-scoring algorithm to identify highly weighted nodes and relations  
- summaries are generated by extracting and merging portions of logical forms | extracts |
| CATS [16], 2005 | multi-document | domain-specific (news) | - it analyzes which information in the document is important in order to include it in the summary  
- it is an Answering Text Summarizer  
- statistical techniques are used to computed a score for each sentence as well as temporal expression and redundancy are solved | extracts |
| CLASSY [11], 2005 | multi-document | domain-specific (news) | - it is a query-based system  
- it is based on Hidden Markov Model algorithm for sentence scoring and selection  
- it classifies sentences into two sets: those ones belonging to the summary and those ones which not | extracts |
| QASUM-TALP [17], 2005 | multi-document | domain-specific (news) | - it is a query-driven system  
- summary content is selected from a set of candidate sentences in relevant passages  
- summaries are produced in four steps:  
(1) collection pre-processing,  
(2) question generation,  
(3) relevant information extraction  
(4) and summary content selection | extracts |
| ERRS [57], 2007 | single and multi-document | domain-specific (news) | - it is basically a heurisitic-based system  
- all kinds of summaries are generated with the same data structure: Fuzzy Coreference Cluster Graph | extracts |

continue on next page
<table>
<thead>
<tr>
<th>SYSTEM [REF.], YEAR</th>
<th># INPUTS</th>
<th>DOMAIN SYSTEM</th>
<th>FEATURES</th>
<th>OUTPUT</th>
</tr>
</thead>
</table>
| FemSum [18], 2007  | single and multi-document | domain-specific (news) | - the system is aimed to provide answers to complex questions  
- summaries are produced taking into account a syntactic and a semantic representation of the sentences  
- it uses graph-representation to establish relations between candidate sentences  
- it is composed of three language independent components:  
  \- RID (Relevant Information Detector),  
  \- CE (Content Extractor),  
  \- SC (Summary Composer) | extracts |
| GOFAISUM [19], 2007 | multi-document | domain-specific (news) | - it is only based on a symbolic approach  
- basically, the techniques used are \( tf-idf \) and syntactic pruning  
- sentences with the highest score are selected to build the summary | extracts |
| NetSum [54], 2007  | single-document | domain-specific (news) | - it is based on machine learning techniques to generate summaries  
- it uses a neuronal network algorithm to enhance sentence features  
- the three sentences that best matches the document’s highlights are extracted | extracts |
From the systems described above, it is possible to notice that each system performs different methodologies to produce summaries. Furthermore, the inputs and the genre can be different too. That gives us an idea of how developed the state-of-the-art is and the number of different approaches that exist to tackle this field of research.

In the latest ACL (ACL’07) conference\(^\textsuperscript{18}\) attempts to summarize entire books [45] have been made. The argument to support this idea is that most of studies have been focused in short documents, specially in news reports and very little effort has been done on summarization of long documents, like books. Generating book summaries can be very useful for the reader to choose or discard a book only by looking at the extract or abstract of that book. On the contrary, in the previous year, european ACL conferences (EACL’06)\(^\textsuperscript{19}\) summarization of short fiction stories were also investigated [32] arguing that summarization is the key issue to determine whether to read a whole story or not.

Techniques employed in recent years are very similar to the classical ones but they have to be adapted to each particular kind of system and its objectives. Improvements in machine learning techniques have allowed that they can be used to train and develop summarization systems these days as well. NetSum [54] which was presented in ACL’07 is an example of a system that uses machine learning algorithms to perform summarization.

5 Measures of Evaluation

Methods for evaluating automatic text summarization can be classified into two categories: intrinsic or extrinsic methods [38]. The first one measures the system’s performance on its own, whereas the extrinsic methods evaluate how summaries are good enough to accomplish the purpose of some other specific task, e.g. filtering in information retrieval or report generation. An assessment of a summary can be done in different ways. Several examples, like Shannon Game or Question Game can be found in [23]. In summary evaluation programmes such as SUMMAC\(^\textsuperscript{20}\), DUC\(^\textsuperscript{21}\) or NTCIR\(^\textsuperscript{22}\) automatic generated summaries (extracts or abstract) are evaluated mostly instrinsically against human reference or gold-standard summaries (ideal summaries). The problem is to establish what an ideal summary is. Humans know how to sum up the most important information of a text. However, different experts may disagree in considering which information is the best to be extracted. Automatic evaluation programmes have therefore been developed to try to give


\(^{19}\)The 11th Conference of European Chapter of the Association for Computer Linguistics was held in Trento, Italy, April 3rd-7th 2006, http://eacl06.iste.it/


\(^{21}\)http://duc.nist.gov

\(^{22}\)http://research.nii.ac.jp/ntcir/
an objective point of view of evaluation. Systems like SEE\textsuperscript{23}, ROUGE\textsuperscript{24} or BE\textsuperscript{25} have been created to help to this task.

6 The Evolution of Text Summarization Approaches

Throughout the recent years summarization has experienced a remarkable evolution. Due to the evaluation programmes that take place every year, the field of Text Summarization has been improved considerably. For example, the tasks performed in The Document Understand Conferences (DUC) have changed from simple tasks to more complex ones. At the beginning, efforts were done to generate simple extracts from single documents usually in English. Lately, the trend has evolved to generate more sophisticated summaries such as abstracts from a number of documents, not just a single one, and in a variety of languages. Different tasks have been introduced year after year so that, apart from the general main task, it is possible to find tasks consisting of producing summaries from a specific question or user-need, or just to generate a summary from updated news. Finally, concerning to the evaluation, the tendency has moved on to extrinsic evaluation, i.e. how useful the task is in order to help other tasks, rather than intrinsic evaluation. However, this kind of evaluation is also important to measure linguistic quality or content responsiveness, so manual evaluation is still performed by humans, together with automatic evaluation systems like BE, SEE or ROUGE introduced in the previous section. The evolution of summarization systems has not finished yet. There is still a great effort to do to achieve good and high quality summaries, either extracts or abstracts.

7 Conclusion and Future Work

In this paper, we have described a general overview of automatic text summarization. The status, and state, of automatic summarising has radically changed through the years. It has specially benefit from work of other asks, e.g. information retrieval, information extraction or text categorization. Research on this field will continue due to the fact that text summarization task has not been finished yet and there is still much effort to do, to investigate and to improve. Definition, types, different approaches and evaluation methods have been exposed as well as summarization systems features and techniques already developed. In the future we plan to contribute to improve this field by means of improving the quality of summaries, and studying the influence of other neighbour tasks techniques on summarization.

References

[1] Alfonseca, E., Rodríguez, P. Descrip-


[34] Lal, P., Rueger, S. Extract-based Summarization with Simplification. In Workshop on Text Summarization (In conjunction with the ACL 2002 and including the DARPA/NIST sponsored DUC 2002 Meeting on Text Summarization), Philadelphia, 2002.


[38] Mani, I. Summarization evaluation: An overview. In Proceedings of the


