Artif Intell Rev (2012) 37:1–41
DOI 10.1007/s10462-011-9216-z

Text summarisation in progress: a literature review

Elena Lloret · Manuel Palomar

Published online: 30 April 2011

Abstract This paper contains a large literature review in the research field of Text Summarisation (TS) based on Human Language Technologies (HLT). TS helps users manage the vast amount of information available, by condensing documents' content and extracting the most relevant facts or topics included in them. The rapid development of emerging technologies poses new challenges to this research field, which still need to be solved. Therefore, it is essential to analyse its progress over the years, and provide an overview of the past, present and future directions, highlighting the main advances achieved and outlining remaining limitations. With this purpose, several important aspects are addressed within the scope of this survey. On the one hand, the paper aims at giving a general perspective on the state-of-the-art, describing the main concepts, as well as different summarisation approaches, and relevant international forums. Furthermore, it is important to stress upon the fact that the birth of new requirements and scenarios has led to new types of summaries with specific purposes (e.g. sentiment-based summaries), and novel domains within which TS has proven to be also suitable for (e.g. blogs). In addition, TS is successfully combined with a number of intelligent systems based on HLT (e.g. information retrieval, question answering, and text classification). On the other hand, a deep study of the evaluation of summaries is also conducted in this paper, where the existing methodologies and systems are explained, as well as new research that has emerged concerning the automatic evaluation of summaries' quality. Finally, some thoughts about TS in general and its future will encourage the reader to think of novel approaches, applications and lines to conduct research in the next years. The analysis of these issues allows the reader to have a wide and useful background on the main important aspects of this research field.

Keywords Human language technologies · Text summarisation · Intelligent systems

E. Lloret (✉) · M. Palomar
Department of Software and Computing Systems, University of Alicante, Apdo. de correos, 99,
03080 Alicante, Spain
e-mail: elloret@dlsi.ua.es

M. Palomar
e-mail: mpalomar@dlsi.ua.es

1 Introduction

Human Language Technologies (HLT) cover a broad range of activities with the eventual goal of enabling people to communicate with machines using natural communication skills (Cole 1997). Moreover, the rapid growth of the Internet has resulted in a massive increase of available information in different formats (e.g. text, video, images) difficult to cope with. Consequently, intelligent applications based on HLT can be developed, such as information retrieval, text classification, text summarisation or sentiment analysis, to efficiently deal with all this information. In particular, Text Summarisation (TS), whose aim is to obtain a reductive transformation of source text to summary text through content condensation by selection and/or generalisation on what is important in the source (Spärck Jones 1999), is essential to efficiently manage such information, thus allowing users to save time and resources, as well as to find quickly the specific information they are looking for within documents.

Although it started in the late fifties (Luhn 1958), TS has experienced a great development in recent years, and a wide range of techniques and paradigms have been proposed to tackle this research field (Spärck Jones 2007). However, to produce a summary automatically is very challenging. Issues such as redundancy, temporal dimension, coreference or sentence ordering, to name a few, have to be taken into consideration especially when summarising a set of documents (multi-document summarisation), thus making this field even more difficult (Goldstein et al. 2000). Moreover, research attempting to overcome the lack of coherence that summaries often present has been fuelled in the last years, resulting in combined approaches that identify relevant content and merge it into new fragments of information (Barzilay and McKeown 2005; Zajic et al. 2008). It is also worth mentioning that as society changes, so does TS, adapting itself to new requirements. For instance, the Web 2.0 (social Web) has led to the emergence to new types of Websites, such as blogs, forums, or social networks, where anybody can express his/her feelings towards a topic, entity, product or service. This has resulted in a new type of summaries (sentiment-based) with the purpose of summarising users' opinions. Another important aspect of TS concerns its evaluation. This is also very challenging because it is not clear, even by humans, what type of information a summary should contain, as it has been shown in previous studies (Nenkova 2006). Depending on what the summary is intended for, the information will vary and to capture this automatically is very complicated. Methods for automatically evaluating summaries have been show to correlate well with human evaluation, as in the case of ROUGE (Lin 2004) or AutoSummENG (Giannakopoulos et al. 2008b). However, some issues regarding the quality of the summaries still remain, such as grammaticality, redundancy, or coherence.

In this paper, a review of the state-of-the-art in TS is carried out, focusing especially on the last decade and the new types of summaries that have appeared in recent years, such as sentiment-based summaries or update summaries. It is not expected to be a comprehensive review of all the systems and techniques that have been developed since the beginning of this research field, because there are already good surveys for such purposes (Spärck Jones 2007; Saggion 2008). On the contrary, the contribution of this paper is to provide a general overview of TS, emphasizing recent summary types, and how summaries, although not perfect, can be of great help for other systems based on HLT. Furthermore, the second contribution concerns the summarisation evaluation process, which is especially interesting, due to the room for improvement that it still allows. Therefore, we provide an analysis of the current methodologies and tools to evaluate summaries. Finally, we analyse future directions, identifying some new trends about how this research field is expected to progress in the next years.

The structure of the paper is as follows. Section 2 contains two parts. A general overview of the main TS approaches along the years is provided first (Sect. 2.1), and then a roadmap of the presence of TS in international evaluation campaigns is described (Sect. 2.2). Section 3 addresses the present of TS, where the description of new types of summaries, together with the summarisation techniques employed are explained (Sect. 3.1). Moreover, the analysis of TS in other scenarios different from the traditional newswire texts is also provided in Sect. 3.2. Further on, in Sect. 4, we study the applicability of text summaries in other intelligent systems based also in HLT and, in particular, how they can benefit applications such as information retrieval, question answering or text classification. The use of summaries for this purpose can be considered an extrinsic way of evaluating them. However, Sect. 5 is devoted to the evaluation, focusing mainly on the intrinsic evaluation, distinguishing between methods that assess the content of summaries (Sect. 5.1) from those which assess their quality (Sect. 5.2). Therefore, we describe the current evaluating methodologies (either semi-automatic or fully automatic), as well as their advantages and their main limitations. Finally, Sect. 6 provides an broad analysis in the form of conclusions, and give insights into the future directions of TS.

2 Text summarisation overview

The definition of summary provided in Sect. 1 is not strict. On the contrary, it allows a wide range of summary types depending on what is the summary intended for, how is the input, etc. One of the most well-known existing taxonomies, was proposed in Spärck Jones (1999), where three classes of context factors that influence summaries are taken into consideration: input, purpose and output factors. Input factors deal with aspects related to the source, such as genre, language, or register. The second ones, purpose factors, include audience and use, for example literary reviews or emergency alerts. Finally, output factors, focus on the style and coverage, and are normally driven by purpose factors. The taxonomy proposed by Spärck Jones (1999) is not the only one proposed to classify different summarisation factors. Hovy and Lin (1999) suggested also a similar taxonomy, where the types of summaries are classified according to the most relevant aspects. In the same way as the input, output and purpose factors described in the previous taxonomy, this classification distinguishes between characteristics of the source document, characteristics of the summary as a text, and characteristics of the summary usage. The main difference between both taxonomies is the fact that the latter takes into consideration specific factors concerning the coherence and the subjectivity level of the summary. Additionally, there is one more taxonomy suggested by Mani and Maybury (1999), classifying summarisation systems with regard to the approach adopted to generate summaries. Different approaches can be tackled to produce summaries, facing the problem from three levels: surface-, entity- or discourse-level. Surface-level approaches aim to represent information in terms of shallow features which are then selectively combined together to determine a salient function used to extract the most important information of a document. Such features comprise *thematic features* which take into consideration statistically salient terms, for example based on frequency counts; *location* accounts for the position of a specific unit (word, sentence, etc.) in a document (paragraph, section); *background* refers to the presence of terms from the title or headings in the text, the initial parts of the document, or a user's query; finally, *cue words and phrases* are expressions such as *"in summary"*, *"our investigation"*, *"in particular"*, or *"in conclusion"*. Entity-level approaches build an internal representation of the document or documents to model the entities and their relationships. These approaches tend to represent patterns of connectivity in the document and these relationships include, for example, *similarity* through vocabulary overlap; *proximity*, which refers to distance between text units; *thesaural relationships* among words like synonymy or part-of-relations; or *logical relations* such as contradiction, entailment or agreement. Lastly, discourse-level approaches model the global structure of the document. This includes features concerning the format of the document (such as document outlines or hypertexts), the threads of topics or subtopics developed in the document, or their attempt to capture the structure of different sorts of texts, for example, narrative or argumentative documents' structure.

The taxonomies proposed in Spärck Jones (1999) and Hovy and Lin (1999) deal with a very fine-grained granularity, in the sense that summaries are classified with respect to different criteria in accordance of their own nature, whereas the one suggested by Mani and Maybury (1999) groups summarisation approaches, as far as the type of features and techniques used to generate summaries is concerned. The main problem of having a taxonomy with such a fine-grained granularity arises when one wants to classify a summarisation system with regard to those criteria, since most systems may share several characteristics, increasing the difficulty in its classification and making it unclear. At the same time, the problem with Mani and Maybury's taxonomy is the purity of the suggested classification. Currently, systems rely on hybrid approaches (Mani and Maybury 1999), combining for example, discourse- and surface-level features.

Therefore, taking into account the aforementioned taxonomies, summarisation approaches can be characterized according to many features. Although it has traditionally been focused on text, the input to the summarisation process can also be multimedia information, such as images (Fan et al. 2008); video (He et al. 1999) or audio (Zechner and Waibel 2000), as well as on-line information or hypertexts (Sun et al. 2005). Furthermore, we can talk about summarising only one document (*single-document summarisation*) or multiple ones (*multi-document summarisation*). Regarding the output, a summary may be an *extract* (i.e. when a selection of "significant" sentences of a document is shown), *abstract*, when the summary can serve as a substitute to the original document and new vocabulary is added, or even a *headline* (or title). It is also possible to distinguish between *generic* summaries and *query-focused* summaries (also known as user-focused or topic-focused). The first type of summaries can serve as surrogate of the original text as they may try to represent all relevant facts of a source text. In the latter, the content of a summary is driven by a user need or a query. Concerning the style of the output, a broad distinction is normally made between two types of summaries. *Indicative* summaries are used to indicate what topics are addressed in the source text. As a result, they can give an brief idea of what the original text is about. The other type, *informative* summaries, are intended to cover the topics in the source text and provide more detailed information. Apart from these two types of summaries, another one can be also taken into account, i. e. *critical evaluative abstracts*. This kind of summaries focuses on expressing author's points of view about a specific topic or subject, and they include reviews, opinions, feedback, recommendations, etc., with a strong dependence in cultural interpretation (Mani 2001a). That is the reason why they are so difficult to produce automatically, and therefore most systems only attempt to generate either indicative or informative summaries, by just summarising what appears in the source document. In recent years, new types of summaries have appeared. For instance, the birth of the Web 2.0 has encouraged new types of textual genres, containing high degree of subjectivity, thus allowing the generation of *sentiment-based summaries*. *Update summaries*, which assume that the user has already a background and he/she needs only the most recent information about a topic, are another example of new summary type. More types of summaries are explained in Sect. 3.1. Finally, concerning the language of the summary, it can be distinguished between *mono-lingual*, *multi-lingual*, and *cross-lingual* summaries, depending on the number of languages dealt with. The cases where the input and the output language is

Table 1 Summarisation types according to several factors

MEDIA	Text
	Images
	Video
	Speech
	Hypertext
INPUT	Single-document
	Multi-document
OUTPUT	Extract
	Abstract
	Headline
PURPOSE	Generic
	Personalised
	Query-focused
	Update
	Sentiment-based
	Indicative
	Informative
	Critical
LANGUAGE	Mono-lingual
	Multi-lingual
	Cross-lingual

the same lead to mono-lingual summaries. However, if different languages are involved, the summarisation approach is considered multi-lingual or cross-lingual. For example, if a summarisation system produces a Spanish summary from one or more documents in Spanish, that is the case of a mono-lingual system. On the contrary, if it is able to deal with several languages, such as Spanish, English or German, and produces summaries in the same language as the input document was, we would have a multi-lingual summarisation system. Beyond these approaches, if the summary is in Spanish, but the original documents are in English, the summarizer would deal with cross-linguality, since the input and output languages are different. Table 1 summarizes the most common factors regarding summarisation.

In the remaining of this section, several well-known approaches for TS are going to be described (Sect. 2.1). The approaches will be grouped according to the predominant techniques employed. Furthermore, a brief review of international forums related to TS is also provided in Sect. 2.2, together with an analysis of the TS trends that has been changing over the years which are reflected in the proposed conference tracks.

2.1 Common approaches for generating summaries

The summarisation process can be decomposed into three main subtasks: *topic identification*, *topic interpretation*, and *summary generation* (Radev et al. 2002). The first stage determine a text's topic structure, that is, a representation indicating what topics are included in a text and how those topics change within the text. Once the main themes of a document have been identified, a second stage is needed in order to understand their meaning and therefore distinguish between relevant and irrelevant information. The final stage addresses the generation of a summary by merging and fusing the information previously identified. However, since the summary generation stage is not easy to tackle, most approaches only focus on the first two stages, by simply extracting the sentences as they appear in the documents, thus producing extracts as a consequence. Next, several extractive approaches are described. Specifically, we distinguish between five kinds of approaches suitable for TS, depending on the nature of the techniques employed, which are: *statistical-based* (Sect. 2.1.1); *topic-based* (Sect. 2.1.2); *graph-based* (Sect. 2.1.3); *discourse-based* (Sect. 2.1.4); and *machine learning-based* (Sect. 2.1.5).

2.1.1 Statistical-based approaches

Luhn (1958) used term frequency counts to produce summaries from scientific documents with the aim to determine the relevance of a sentence in a document. The underlying assumption is that the most frequent words are indicative of the main topic of a document. However, not all the words are taken into consideration. On the contrary, stop words, i.e. words without carrying any semantic information, such as “a” or “the”, are not used for computing the term frequency. Under the same assumption, a number of techniques based on term frequency counts have been employed in TS. For instance, (Lloret and Palomar 2009) use the frequency of words in combination with the length of noun phrases to compute the relevance of a sentence, outperforming the result of the state-of-the-art in single-document summarisation for newswire domain. In McCargar (2005), several statistical approaches, such as term frequency or inverse document frequency (*tf***idf*), are briefly analysed, as well as the potential problems this kind of features may have. The idea behind *tf***idf* is that frequent terms in a document are important only if they are not very frequent in the whole collection. This techniques has been also employed to score sentences for instance in Gotti et al. (2007). As it is claimed in Filatova and Hatzivassiloglou (2004) this kind of methods may be not sufficient for building high-quality summaries, and other types of knowledge, for instance events, semantic, topic-based or discourse information may be most appropriate to tackle TS. However, a deeper review of statistical techniques for TS is carried out in Orăsan et al. (2004) and Orăsan (2009), where it is shown that these techniques, despite being simple and not requiring a deep level of knowledge analysis, are appropriate for building good summaries. In addition to the aforementioned techniques, mutual information, information gain and residual inverse document frequency are also analysed. Mutual information can be used to measure the dependency or the common information between two words, whereas information gain is a good metric for deciding the relevance of an attribute, and in this case, it could perfectly apply to the terms or sentences in a document. Residual inverse document frequency is a variant of the inverse document frequency, which computes the term document frequency according to the *Poisson* distribution. Each technique establishes a manner of assigning weights to the words included in the document, and then, sentences are scored based on these weights, in order to determine their relevance. The approach suggested in Mori (2002) also employs information gain for determining the weight of document terms, and then use it for successfully summarising documents. The idea is to first build clusters of documents according to the similarity among them, and then compute the weight of each word in the clusters. The final summary will be produced by selecting those highest scored sentences on the basis on the weight of words it contains, previously computed using information gain.

2.1.2 Topic-based approaches

In Edmundson (1969) summaries are produced by means of cue word identification. This technique consists of determining the relevance of a sentence by means of the phrases or words it contains. Sentences containing phrases like "*in conclusion*" or "*the aim of this paper*" may be good indicators of relevant information. Moreover, other approaches, such as Boguraev anf Neff (2000), Neto et al. (2000), Angheluta et al. (2002), or Harabagiu and Lacatusu (2005) take profit of the advantages of combining topics' identification and segmentation. Particularly, in Harabagiu and Lacatusu (2005), the topic structure is characterized in terms of topics themes, which are representations of events that are reiterated throughout the document collection, and therefore represent repetitive information. Five different ways of representing topics are analysed: (1) via topic signatures. This idea comes from Lin and Hovy (2000), where it is assumed that the topic of a document can be represented using a set of terms; (2) via enhanced topic signatures. This differs from the previous one in the fact that the aim now is to discover relevant relations between two topic concepts; (3) via thematic signatures, which is carried out by segmenting documents using the TextTiling algorithm (Hearst 1997) first, and then assigning labels to themes to be able to rank them later; (4) via modelling the content structure of documents. The assumption here is that all texts describing a given topic are generated by a single content model (in this case a Hidden Markov Model). Finally, the last method to represent topics within a text is to use (5) templates, following the idea of the field of information extraction, identifying specific entities or facts. Furthermore, in Teng et al. (2008), a single-document summarisation approach is suggested which combines local topic identification with term frequency. The proposed methodology computes the sentence similarity first, and then performs the topic identification by doing sentence clustering. In a second step, sentences from local topics are selected according to the term frequency value. Moreover, not only topic words are used to detect relevant information within a document. In other approaches, (Kuo and Chen 2008), for instance, informativeness and event words are also taken into consideration in order to produce multi-document summaries. The underlying idea is that this kind of words indicate the important concepts and relationships, and can be used to detect relevant sentences within a set of documents. Furthermore, a temporal resolution algorithm is used, so that dates and other temporal expressions can be translated into calendrical forms. The identification of multiple themes within an heterogeneous collection of documents is addressed in Ando et al. (2005) by means of vector space representations; in particular, *Iterative Residual Rescaling* is used which it is proved to be suitable for building space models for linguistic objects.

2.1.3 Graph-based approaches

The use of graph-based ranking algorithms has been also shown to be effective in TS. Basically, the nodes of the graph represent text elements (i.e. normally words or sentences), whereas edges are links between those text elements, previously defined (for instance, semantic relations, such as synonymy). On the basis of the text representation as a graph, the idea is that the topology of the graph will reveal interesting things about the salient elements of the text, for example concerning the connectivity of the different elements. LexRank (Erkan and Radev 2004) is a multi-document summarisation system, in which all candidate sentences that can be potentially included in the summary are represented in a graph. In this graph representation, two sentences are connected if the similarity between them is above a predefined threshold. Then, once the network is built, the system finds the most central sentences by performing a random walk on the graph. In Mihalcea (2004), an analysis of several graph-based algorithms is carried out, evaluating also their application to automatic sentence extraction in the context of the TS. Furthermore, in Wan et al. (2007), an approach based on affinity graphs, for both generic and query-focused multi-document summarisation, is suggested. The idea here is to extract sentences with high information richness and novelty. This is achieved by taking into consideration the similarity between each pair of sentences, incorporating topic information, differentiating intra-document and inter-document links between sentences, and finally penalizing redundant information. In Giannakopoulos et al. (2008a) character and word n-gram graphs are used to extract relevant information from a set of documents, whereas in Plaza et al. (2008) graphs are built using concepts identified with Wordnet (Fellbaum 1998) and *is-a* relationships, which are then used to build a graph representation for each sentence in a document. This approach has been proven successfully in different domains, such as newswire, biomedical documents or image captions.

2.1.4 Discourse-based approaches

Besides all the previous mentioned techniques, it is also possible to face the summarisation problem from a linguistic point of view, for instance exploiting discourse relations. Rhetorical Structure Theory (RST) proposed in Mann and Thompson (1988) served as a basis for the summarisation approach developed in Marcu (1999), extending the rhetorical relations, and using this kind of discourse representation (nucleus and satellite relations, depending on how relevant the information is) to determine the most important textual units in a document. Furthermore, in Khan et al. (2005) the RST is combined within a generic summarizer in order to add linguistic knowledge to the summarisation process. Although the results obtained for this mixed approach do not improved the ones obtained by the generic summarizer, it was claimed that the drawback of this approach relied on the parser which could not detect all the RST relationships, otherwise linguistic knowledge could have improve the overall summarisation performance. Furthermore, in Cristea et al. (2005) an approach similar to RST is described, differing from the previous ones, in the lack of relation names and the use of binary trees. This summarisation approach is intended to exploit the coherence and cohesion of a document.

Cohesion and coherence are two of the main challenging issues for TS. Some approaches rely on the identification of such relations in order to improve the quality of the generated summaries. Cunha et al. (2007) combines statistical and linguistic techniques to prove that results improve with respect to use only one type of techniques. In Gonçalves et al. (2008), coreference chains are used to deal with referential cohesion problems that are frequent in the extractive summarisation approach. A post-processing system is developed in order to rewrite referential expressions in the most possible coherent way, and it is applied after the summary is generated, obtaining considerable improvements in comparison to the original summaries. In order to guarantee the coherence of a summary, a widespread approach is to use lexical or coreference chains. However, the use of coreference chains is not novel in TS. The first approaches can be found in Baldwin and Morton (1998), and Azzam et al. (1999). The main assumption is that the longest coreference chain indicates the main topic of the document, and shorter chains represent subtopics. Therefore, one possible strategy for building summaries is to select only those sentences related in the longest chain. This strategy helps to maintain the coherence of the text. A similar idea is to use *lexical chains*, which consists of determining sequences of semantic related words (for example, by concept repetition or synonymy relations). By using lexical chains, the main topics of a document can be also detected. This technique has been also widely used in summarisation, and approaches like the ones described in Barzilay and Elhadad (1999), Medelyan (2007) or Ercan and Cicekli

(2008), exploit them to produce summaries. It is worth mentioning that being able to identify all the entities that are connected within a document or across documents, prevent summaries from the common *dangling anaphora* phenomenon, thus producing more coherent resulting summaries (Elsner and Charniak 2008). This phenomenon consists of having words in a text (mostly pronouns) without its correct antecedent included in the summary. For example, if a summary contains the pronoun "he", but its antecedent (e.g. the president of Spain) is not mention in it, this would lead to an unclear summary with this specific type of problem, which would make the summary difficult to understand, or even incoherent. In order to reduce this problem some approaches combine the use of anaphora resolution in order to help TS (Orăsan 2004; Mitkov et al. 2007). In these approaches, documents are first processed to resolve anaphoric pronouns, and then a summarisation system is run in order to produce a summary, and determine whether an anaphora resolution systems improve the quality of summarisation or not. Due to the moderated performance of this kind of systems, this is hard to achieve, and contrary to the intuition, TS does not improve very much. However, in Orăsan (2007) an ideal anaphora resolution system was simulated, resolving the anaphoric relations in scientific documents manually, and it was proven that in such situations, summary's results improve noticeably. In Steinberger et al. (2007), it was stated that the improvement associated to TS, when using an anaphora resolution system, not only depends on the lower performance of the anaphora resolver, but also in the way anaphoric relations are used. As a consequence, they used the anaphoric relations from two different perspectives: on the one hand, to improve the quality of summaries, and on the other, to check the coherence of a summary, once it was already generated by checking if the coreference chains of the summary are sub chains of the ones identified in the source documents.

2.1.5 Machine learning-based approaches

The approaches that are next explained are based on machine learning algorithms to produce summaries. The first machine learning methods used in TS include binary classifiers (Kupiec et al. 1995), *Hidden Markov Models* (Conroy and O'leary 2001; Schlesinger et al. 2002), and *Bayesian* methods (Aone et al. 1998). However, a wide range of machine learning techniques can be used for TS. NetSum (Svore et al. 2007) bets on single-document summarisation and produces extracts from newswire documents based on neuronal nets, using RankNet (Burges et al. 2005) as a learning algorithm to score the sentences and extract the most important ones. Besides the common features based on keywords and sentence position, a new set of features based on Wikipedia[1] and query logs are also used in a way that for example, sentences containing query terms or Wikipedia entities, contain therefore important content. In Schilder and Kondadadi (2008), a query-focused multi-document summarizer is presented, named as FastSum, where sentences are ranked using a machine learning technique called *Support Vector Regression* (SVR), and *Least Angle Regression* for feature selection. SVR was used in summarisation before, in the approach described in Li et al. (2007), where word-, phrase-, semantic-based, as well as sentence position or name entities features were used to train the classifier automatically. Further on, the extracted features were combined, and then sentences were scored. In Wong et al. (2008), an extractive summarisation approach is presented, employing supervised and semi-supervised learning methods. The sentence features involved are grouped into different types—surface, content, relevance and event features—which include sentence position, number of words in a sentence, centroid and high frequent terms, or similarity between sentences, among others. Regarding the supervised approach, a

[1] http://www.wikipedia.org/.

Support Vector Machine (SVM) algorithm is used, whereas for the semi-supervised approach, a probabilistic SVM and a *Naïve Bayesian* classifier are co-trained to exploit unlabelled data. SVM technique was also used in Fuentes et al. (2007) to detect relevant information to be included in a query-focused summary, where structural, cohesion-based and query-dependent features were used for training.

The advantage of using machine learning for TS is that it allows to test easily the performance of a high number of features, for instance lexical, syntactic, statistical, etc. using then different machine learning paradigms for learning which are the most suitable ones. However, these approaches also need a big training corpus in order to be able to obtain conclusive results. Usually the corpus consists of a set of human-written summaries, or annotated source documents containing which sentences are important for the summary, and which not.

2.2 Relevant conferences and workshops

At the end of the 90's, the TIPSTER Text Summarisation Evaluation[2] (SUMMAC) was the first conference aimed at evaluating automatic summarisation systems, where text summaries were tested in document classification and question answering, in order to analyse whether they were suitable surrogates for full documents. A detailed explanation of this evaluation forum and how summaries were evaluated can be found in Mani et al. (2002). The National Institute for Informatics Test Collection for IR[3] (NTCIR) also developed a series of Text Summarisation Challenges (TSC) workshops, which included Japanese summarisation tasks in 2001 (TSC), 2002 (TSC2), and 2003 (TSC3). Besides these conferences, the important conferences that focused only in TS were the Document Understanding Conferences[4] (DUC) that were hold yearly from 2001 to 2007. In these conferences, different tasks were proposed over the years, taking into account new challenges and requirements for TS, forcing also systems to be dynamic and adaptable. Along the editions of the DUC conferences, it can be seen how the summarisation systems have progressed, as well as the different evaluation methodologies that have been proposed to evaluate the corresponding summaries. These changed from a complete manual evaluation, where assessors used the SEE evaluation environment[5] to facilitate the comparison of automatic and human-made summaries' content, to a fully automatic evaluation of the content using ROUGE (Lin 2004) and Basic Elements (Hovy et al. 2006).

The tasks involved in the DUC conferences also changed over the editions, starting at the beginning with generic single-document summarisation, and continuing further on with query-focused multi-document summarisation. An extense general overview of the major summarisation conferences, focusing particularly in DUC, can be found in Over et al. (2007). More specifically, the overview for some DUC editions is also provided in Over and Ligget (2002) and Dang (2006). These sorts of conferences are very useful to evaluate and compare automatic systems, and at the same time, they also provide a good set of corpora, comprising documents and model summaries, which are free available on demand.[6] Unfortunately, due to the fact that all the editions worked under the newswire domain, the data deals with a unique domain. Since 2008, DUC conferences are no longer organised, because they have

[2] http://www-nlpir.nist.gov/related_projects/tipster_summac/.

[3] http://research.nii.ac.jp/ntcir/outline/prop-en.html.

[4] http://www-nlpir.nist.gov/projects/duc/.

[5] http://www.isi.edu/licensed-sw/see/.

[6] http://www-nlpir.nist.gov/projects/duc/data.html.

become part of the Text Analysis Conference[7] (TAC), within which a summarisation track is included. In TAC 2008, two different tasks were proposed within the TS track. The first followed the same idea as the update summarisation task in the DUC 2007, which consisted of building summaries containing updated information with respect to a given set of news documents, whereas the second one, was a pilot task whose aim was to generate opinion summaries from blogs. In TAC 2009, the update summarisation task was kept but, on the other hand, instead of the opinion summarisation task, a new task concerning the automatic evaluation of summaries was proposed (*Automatically Evaluating Summaries Of Peers*[8]). The goal of this task is to automatically score a summary for a given metric that reflect summary content. In the current edition of TAC (2010), the task concerning evaluation is maintained but, in contrast, update summaries have been changed into guided summaries. The idea under this new kind of summaries is to encourage systems to use deeper semantic knowledge, building summaries that contain specific information about different aspects of a topic. For instance, if a set of documents are about an accident, we might be interested in information concerning when it occurred, why, where, etc.

Table 2 shows some features of the tasks involved in the previously mentioned conferences. Bold typed words indicate the novelties introduced in summarisation tasks over the years. In the first summarisation conference, SUMMAC (Mani et al. 1999), query-focused single-document summaries from newswire documents were evaluated. To perform this, two extrinsic evaluation tasks, and another one intrinsic were proposed. Extrinsic evaluation judges the quality of the summarisation based on how it affects the completion of some other task, whereas intrinsic evaluation measures a summary on its own. On the one hand, in the extrinsic evaluation, an *adhoc* task was suggested in which indicative summaries were evaluated with regard to whether they allowed to quickly determine the relevance of a document focused on a specific topic. Moreover, a categorisation task was also proposed, whose aim was to determine if generic summaries could effectively present enough information to allow a person to correctly categorise a document. On the other hand, regarding the intrinsic evaluation task (question-answering task), the goal was to measure the content of a summary with respect to which degree it contained answers to several topic-related questions. The tasks involved in the TSC workshops within NTCIR conferences (Fukushima and Okumura 2001; Okumura et al. 2004; Hirao et al. 2005) also dealt with the evaluation using intrinsic and extrinsic methods. In the first TSC, three tasks were proposed, consisting of producing summaries given a specific length. The differences between them were that in the first task, only important sentences had to be extracted whereas in the second, automatic generated summaries were compared to human-made ones. The third task involved extrinsic evaluation, and summaries where evaluated for information retrieval purposes. This task was very similar to the ad hoc task of SUMMAC conference. Multi-document summarisation was first included in TSC2, in which a single-document as well as multi-document tasks were defined. However, since then, multi-document summarisation became a central issue, and consequently, in the following conferences (TSC3), tasks were no longer addressed to produce summaries from only one input document. As time goes by, systems evolve and so does the requirements of summarisers, according to the needs of society. The changes of summarisation requirements and systems over the time line can also be seen at DUC conferences. At the beginning, the proposed tasks were aimed at producing generic summaries from a single or several input documents, but at the end, query-focused summaries and novelty were paid more attention to. One aspect to remark was the attempt to perform cross-lingual summarisation between

[7] http://www.nist.gov/tac/.

[8] http://www.nist.gov/tac/2009/summarisation/aesop.09.guidelines.html.

Table 2 Summarisation task features for each conference

Conference	Summarisation task requirements
SUMMAC[a]	Single-document, query-focused, news
TSC[b] (NTCIR)	Query-focused, generic, news
TSC2 (NTCIR)	Single and **multi-document**, generic, news
TSC3 (NTCIR)	Multi-document, generic, news
DUC-01[c]	Single and multi-document, generic, news
DUC-02	Single and multi-document, generic, news
DUC-03	Multi-document, **query-focused**, news
DUC-04	Single and multi-document, topic-oriented, news, **cross-lingual**
DUC-05	Multi-document, query-focused, news
DUC-06	Multi-document, query-focused, news
DUC-07	Multi-document, **update**, query-focused, news
TAC-08[d]	Multi-document, update, query-focused, **sentiment-based**, news & **blogs**
TAC-09	Multi-document, update, query-focused, news, **evaluation**
TAC-10	Multi-document, **guided**, query-focused, news, evaluation

[a] SUMMAC summary types and TIPSTER text summarisation evaluation conference
[b] *TSC* text summarisation challenges
[c] *DUC* document understanding conferences
[d] *TAC* text analysis conferences

English and Chinese in DUC (2004). This task consisted of producing either very short or short summaries in English from a set of documents that were previously automatic translated into English (its original language was Chinese). The concept of novelty and novel information was first addressed in the update task of DUC (2007). The goal of this task was to generate a summary from a cluster of related documents, but taking into account that some of those documents had been already read by users, so the information contained in them, did not need to appear in the summary. Regarding the domain of the documents, the DUC conferences were also focused in newswire documents. On the contrary, besides newswire documents, a well-known source of information on the Internet, i.e. blogs, was introduced to be dealt with, in the opinion summarisation task at the TAC (2008) conference.[9] However, due to the difficulty involved in the task itself and type of data (blogs), the opinion summarisation task was out of the scope of the TAC (2009), keeping only the update summarisation task and introducing a new task concerning the automatic evaluation of summaries. Finally, as it was aforementioned, in TAC (2010), the task concerning evaluation is kept, whereas a new kind of summaries are introduced, guided summaries.

Regarding the number of participant groups, Table 3 shows this number for each edition's conference. It is worth realising how this number has risen over the years, showing the increasing interest in the TS research field. As a consequence, specific workshops focusing only on TS are being organised within important conferences, such as the Workshop on Multi-source, Multilingual Information Extraction and summarisation[10] (MIMIES), the Workshop on Language Generation and summarisation[11] (UCNLG+Sum), the Workshop

[9] http://www.nist.gov/tac/tracks/2008/summarisation/index.html.

[10] http://doremi.cs.helsinki.fi/mmies2/.

[11] http://www.nltg.brighton.ac.uk/ucnlg/ucnlg09/.

Table 3 Number of participant groups for each conference

Conference name (Year)	Number of participant groups
SUMMAC (1999)	16
TSC (2001)	9
TSC2 (2002)	8
TSC3 (2003)	9
DUC (2001)	15
DUC (2002)	17
DUC (2003)	21
DUC (2004)	22
DUC (2005)	31
DUC (2006)	34
DUC (2007)	32
TAC (2008)	38
TAC (2009)	39

on Web Search Result summarisation and Presentation (WSSP),[12] or the 1st International Workshop on Discovering, Summarising and Using Multiple Clusterings.[13]

3 Text summarisation in the current context

This section is divided into two subsections. On the one hand, a review of recent types of summaries is provided in Sect. 3.1, and on the other hand, TS for new scenarios is also briefly explained in Sect. 3.2. Apart from the classical summarisation types, such as single- or multi-document summaries, generic or query-focused, etc. there are several interesting novel types of summaries, where specific objectives are pursued (e.g. sentiment-based summaries). In addition to these types of summaries, the emergence of new scenarios is also interesting for TS. Rather than carrying out research into the same datasets traditionally based on newswire or scientific documents, in recent years, novel domains, such as literary text, patents, or blogs has been paid a lot of attention to. Therefore, we also address them in this section.

3.1 New types of summaries

In the remaining of this subsection, different kinds of summaries that have recently appeared are going to be described. It is worth stressing upon the fact that the kinds of summaries next described mostly focus on user's needs or attempt to efficiently deal with vast amounts of information. Regarding the former, we have selected personalised, update and sentiment-based summaries, since their common goal is to produce a summary, the content of which is determined directly by user requirements (the user has to delimit what type of information he is interested in). With respect to the latter, surveys and abstractive summaries are analysed, because they represent two good examples of summaries that have to be built employing

[12] http://www.wssp.info/2009.html.

[13] http://eecs.oregonstate.edu/research/multiclust/.

techniques that go beyond the mere concatenation of sentences, and currently it seems that this is the tendency and final goal of TS systems.

3.1.1 Personalised summaries

The purpose of this type of summaries is to provide a summary containing the specific information a user is interested in. This mean that different users may have different needs, so that summarisation systems have to determine the user profile before they select the relevant information that will be included in the finally summary.

In (Agnihotri et al. 2005) the user profile is determined by means of a statistical mapping method from the users' personality traits identified using personality tests to the content. The analysis performed over 59 users showed that only some traits such as gender and only some features (e.g. text) were of help for personalising the summary. The main drawback to this approach is the limited data it is used for the experimental set-up and due to the difficulty of the task, the experiments in another environment are very hard to extend or replicate. Regarding also personalised summaries, in Díaz and Gervás (2007) an approach to produce newswire summaries that contain relevant information for a given user profile is proposed. The idea is to select those sentences that are most relevant to a given user model. This is done by calculating the similarity between the user model for a specific individual and each one of the sentences in the document, so depending on which part of the model is chosen to compute the similarity, several possible personalised summaries can be obtained. The user-model is determined by the combination of the specific domain-features, a set of keywords, which reflects the information needs that do not change across the time, and a relevance feedback tier, which takes into account the changes given by users' feedback. The extensive set of experiments carried out shown the appropriateness of this type of summaries, achieving the summaries around 60% recall.

In Kumar et al. (2008), personalised summaries are generated based on the area of expertise and personal interests of a user. With this purpose, a user background model is developed taking as a basis the information found on the Internet with regard to a person, such as his/her personal Web page, or on-line publications. Once the user profile has been identified, the relevance of document's sentences are determined according to this profile. Two scoring functions, one for generic information and one for user specific are proposed. The first one relies on term frequency to extract the most relevant generic sentences, whereas the second one computes the probability of the generic sentences to contain also user specific information. Then, in the summary generation stage, the top ranked sentences are selected and extracted. Although this approach is very interesting, its main difficulty is that name entity disambiguation should be performed when looking specific information of a person on the Internet. This would be essential because it may happen that people share the same name but they are totally different (e.g. George Bush could refer either to the US president from 1989 to 1993 or to the US president from 2001 to 2009 (his son)).

In (Berkovsky et al. 2008), a preliminary user evaluation is conducted in order to assess different aspects of users attitudes towards personalised TS. Three experiments are suggested with the purpose of analysing this issue. Firstly, it is evaluated whether the personalisation of summaries has the desired effect on users or not. Then, the impact of summary lengths is analysed. Finally, the degree of faithfulness between the personalised summaries and the original documents is assessed. The conclusions derived from the analysis are very preliminary. It is shown that the more personalised information a summary contains, the better it will be preferred. It is also claimed that users prefer not too short nor too long summaries; however, no clues about which should be the optimal length are given.

3.1.2 Update summaries

Update summarisation attempts to generate summaries taking into consideration that users have a background knowledge of the topic they want to read about, so they are only interested in the most recent events related to that topic. This type of summaries emerged thanks to the task proposed at DUC and TAC conferences.

In order to generate update summaries, the approach described in Sweeney et al. (2008) consists of incorporating novelty to summaries, by minimizing the content overlap between a summary sentence and a potential candidate one. In Witte et al. (2007), Bellemare et al. (2008), Li et al. (2009) update summaries are built on the basis of cluster graph data structures, which are based on the context and on the set of documents that are going to be summarized. A sentence ranking scheme is proposed depending on the overlap between sentences from clusters and the context, so finally ranks are established and summaries are generated by selecting sentences from each rank. The approach suggested in Li et al. (2008), defines the concept of history (those documents already known by a reader), and introduces a new type of features (filtering features), which reflect that the summary is summarized with its history. Therefore, in such cases, filtering features can be calculated through two different similarity metrics to exclude those sentences which are similar to the history. One of these similarity metrics is based on the cosine distance formula, whereas the other one takes into consideration unigrams, bigrams and syntactic functions of the words and combines all of them linearly to obtain finally a similarity metric.

Machine learning algorithms are also exploit for generating update summaries. Schilder et al. (2008) relies on FastSum summarisation system (Schilder and Kondadadi 2008) which uses SVM, but features concerning related to new and old information, such as new/old entities, new/old word/document frequency are also taken into account. Such features penalise sentences that are similar to the ones from the previously selected ones. Moreover, in Fisher et al. (2009), similarity metrics are used as features within a supervised machine learning paradigm, a perceptron ranker, rather than being used to directly rank sentences. Together with these features, a discourse segmenter is also employed to determine potential sub-sentential units to be included in the summary as well. Both approaches obtained good results, remaining at the middle of the ranking among all participants.

In recent approaches, such as Liu et al. (2009) and Nastase et al. (2009), the background information is taken from Wikipedia articles. On the one hand, in the former approach, Wikipedia is used to produce a summary, taken the first paragraph of the entry related to a topic, and then computing the similarity between the potential summary sentences to the ones already contained in the Wikipedia-based summary. The sentences with lower similarity will be selected for the update summary. On the other hand, the latter uses Wikipedia to retrieve concepts that are discussed in the set of documents to summarise. This way, it is possible to predict which concepts are more likely to be found in well-formed summaries. Apart from Wikipedia knowledge, this approach also performs sentence compression to give the final summary an abstract nature. The results shown that Wikipedia is a useful resource to exploit due to the amount of information it contains and the way this information is structured.

3.1.3 Sentiment-based summaries

In recent years, the subjectivity appearing in documents has led to a new emerging type of summaries: sentiment-based summaries, which have to take into consideration the sentiment a person has towards a topic, product, place, service, etc. Consequently, TS and Sentiment Analysis (SA), also known as opinion mining, have to be combined together in order to produce this type of summaries. SA provides the sentiment associated to a document at different levels (document, fragment, sentence or even word-level) (Pang and Lee 2008), whereas TS identifies the most relevant parts of a document and build from them a coherent fragment of text (the summary). Regarding sentiment-based summaries, opinions have to be detected and classified first, according to their subjectivity (whether a sentence is objective or subjective, for instance), and then to their polarity (positive, negative or neutral). Further on, TS is in charge of determining which sentences will be included in the summary, and generate the final summary. Sentiment-based summarisation systems that participated in the *Opinion Summarization Pilot Task* of TAC 2008 conference, such as Conroy and Schlesinger (2008), He et al. (2008), Balahur et al. (2008), or Bossard et al. (2008) follow these steps.

However, out of the scope of the TAC competition, other interesting approaches can be found as well. For instance, in Beineke et al. (2004) machine learning algorithms are used to determine which sentences should belong to a summary, after identifying possible opinion text spans. The features found to be useful to locate opinion quotations within a text included location within the paragraph and document, and the type of words they contained. Similarly, in Zhuang et al. (2006) the relevant features (e.g. screenplay, actors for a movie) and opinion words and their polarity (whether a positive sentiment or a negative is) are identified, and then, after identifying all valid feature-opinion pairs, a summary is produced, but focusing only in movie reviews. Normally, on-line reviews contain also numerical ratings that users give when providing a personal opinion about a product or service. The approach described in Titov and McDonald (2008) proposed a *Multi-Aspect Sentiment* model. This statistical model uses aspect ratings to discover the corresponding topics and extract fragments of text. Moreover, in Lerman and McDonald (2009), an approach to produce contrastive summaries in the consumer reviews domain is suggested. Contrastive summarisation refers to the problem of generating a summary for two entities in order to highlight their differences, for example, different people's sentiments about several products. In order to produce this type of summaries they adapt the *Sentiment Aspect Match* model described in Lerman et al. (2009), originally designed to generate single product sentiment-based summaries. This model determines which sentences to extract comparing the average sentiment of a sentence with respect to the average sentiment of the specific entity, thus selecting the closest ones.

3.1.4 Survey summaries

This kind of summaries aims at providing a general overview of a particular topic or entity. They are generally long rather than short, because they attempt to capture the most important facts concerning to a person, for instance. Next, we focus on biographical summaries, survey summaries and Wikipedia articles as three of the most recently TS types that can be categorized within this group.

The challenge of producing summaries from biographies was presented in Zhou et al. (2004). The idea behind multi-document biography summarisation is to produce a piece of text containing the most relevant aspects of a specific person, answering questions, such as *"Who is Barack Obama?"* for example. To accomplish this task, several machine learning algorithms are used (*Naïve Bayes*, SVM, and *Decision Trees*) to classify sentences. Moreover, redundant information is removed in a later stage. A similar biographic summarisation system using also machine learning techniques is described in Biadsy et al. (2008). The difference with the aforementioned one is that a binary classifier is used to discriminate between biographical and non-biographical sentences, and then a SVM regression model is trained to reorder biographical sentences extracted using the Wikipedia as a corpus. The final stage of this approach is to employ a rewriting heuristic to create the final summaries.

Another interesting approach for this kind of TS is to use citations from articles. In Kan et al. (2002) it was shown that from bibliographic entries it was possible to produce an indicative summary. The main idea behind this assumption is that such entries contain informative as well as indicative information, for example, details about the resource or metadata, such as author or purpose of the paper). In their research, a big annotated corpus (2000 annotated entries) is developed for such purposes. Following the idea of generating summaries from this input information, in Qazvinian and Radev (2008) citations are analysed to produce a single-document summary from scientific articles. The final objective is to generated summarisation about a specific topic. Also, the work described in Mohammad et al. (2009) addresses this topic and consequently presents some preliminary experiments of the usefulness of citation text to automatically generate technical surveys. Three kinds of input are used (full papers, abstracts and citation texts), and already existing summarisation systems are taken into consideration to create such surveys, for instance LexRank (Erkan and Radev 2004), Trimmer (Zajic et al. 2007), and C-RR and C-LexRank (Qazvinian and Radev 2008). Among the conclusions drawn from the experiments, it was shown that multi-document technical survey creation benefits considerably from citation texts.

Different from these approaches, (Sauper and Barzilay 2009) suggest the automatically creation of Wikipedia articles using domain-specific templates which are induced from human-generated documents. For producing such articles, a search engine is employed to retrieve documents related to a topic, which are considered as the input of the summarisation process. Following the same structure of a Wikipedia article, the appropriate information for each section is determined by machine learning techniques, training excerpts based on how representative they are to a selected topic.

3.1.5 Abstractive summaries

To simplify the problem of summarisation, most approaches follow an extractive paradigm, by outputting the most relevant sentences of a document/s without doing any changes. Although it is a widespread method, the resulting summaries often present several problems with respect to their quality, such as the lack of coherence or *"dangling anaphor"*. The abstractive paradigm can solve these limitations, since it attempts to produce new text from the fragments of information or concepts identified as relevant. Despite not being a novel issue, in recent years, research in abstraction has been fuelled, due to the fact that information is repeated across documents, and specific ways of conveing and present the information are required. Methods such as sentence compression (Zajic et al. 2007), sentence fusion (Barzilay and McKeown 2005) or natural language generation (Radev and McKeown 1998) have been traditionally applied for the generation of abstracts.

In addition, several analysis have been conducted to understand how humans summarise (Jing and McKeown 2000; Jing 2002). As a consequence, the basic operations to transform source information into summary information are analysed. For instance, (Hasler 20007) claims that the technique humans do is to copy and paste the same material present in the source documents. However, some slightly changes are applied in most of the cases, and two types of operations, atomic and complex, are identified, involving deletion, insertion, replacement, reordering or merging (the first two are atomic operations while the last three are complex). From the evaluation carried out in terms of coherence, the results showed that 78% of the abstracts were more coherent than extracts. In (Fiszman et al. 2004) an approach to generate abstracts from biomedical documents is proposed. The main idea is to identify semantic predicates using SemRep[14] and then produce the summary in an schematic way. The summarisation process comprises the identification of such predicates, and the connectivity between them. Further on, the novelty and the salience of each predication is computed based on term frequency counts. Saggion (2009) suggests a novel approach to combine different fragments of information that have been extracted from one or more documents. From a predefined vocabulary (e.g. *to address*, *to indicate*, *to report*, etc.) the algorithm is able to decide which of these expressions is more appropriate for a sentence, depending on the content and the partial abstract generated. The motivation under this research is to study to what extent the addition of extra information not present in the source documents is useful and benefits the abstraction process. Using machine learning techniques and experimenting with different types of classifiers (e.g. *Decision Trees*), results showed that the best classifier, based on summarisation features, including linguistic, semantic, cohesive, discourse or positional information, is able to correctly predict 60% of the cases.

Natural Language Generation (NLG) is also applied for producing abstractive summaries. In (Yu et al. 2007) very short summaries are produced from large collections of numerical data. The data is presented in the form of tables, and new text is generated for describing the facts that such data represent. Firstly, the data has to be analysed, and understood before generating the descriptions. In particular, for the last step a NLG module is used, specifically accounting for three types of information to generate: background information, overall description, and most significant patterns found in the data collection. Belz (2008) also proposed a TS approach based on NLG to generate weather forecasts automatically, but focusing mainly on the NLG stage.

Other abstractive approaches rely on the use of templates to structure the information that has been previously identified, for instance using an information extraction system. Kumar et al. (2009) attempts to generate reports from the event information stored in databases from different domains (biomedical, sports, etc.). Human-written abstracts are used to determine the information to include in a summary, where some templates are generated and patterns to fill in such templates are identified in the texts. Similarly, in Carenini and Cheung (2008) patterns are also identified, but since the aim is to generate contrastive summaries, discourse markers indicating contrast such as *"although"*, *"however"*, etc. are also added to make the summary sound more naturally.

3.2 New scenarios for text summarisation

Although most work in TS has traditionally focused on newswire (Gotti et al. 2007; Nenkova et al. 2005; Nenkova 2005), scientific documents (Jaoua and Hamadou 2003; Teufel and Moens 2002), or even legal documents (Saravanan et al. 2006; Cesarano et al. 2007), these are not the unique scenarios in which TS approaches have been tested on. Next, several new scenarios in which TS has been also applied are described. From the analysis of TS in such scenarios, it is worth stressing upon the fact that, although the nature of the documents is totally different across domains, and it may seem that each domain would need a different manner to tackle the TS process, in practice, the techniques employed do not experiment great changes. It can be seen that statistical or positional features are the preferred ones. In some cases, specific vocabulary is added, but generally minor changes are performed to adapt TS approaches to other scenarios. Hence, the following discussion may arise: *"Would it be better to develop generic systems for a wide range of scenarios, although with moderate performance, or to build very specific systems that could obtain a higher performance?"*. On

[14] http://skr.nlm.nih.gov/papers/index.shtml#Sem\discretionary-Rep.

the other hand, since some features, such as term frequency or inverse-document frequency can be considered domain-independent, an appropriate approach could be to combine these type of features with domain-specific ones within the same TS process. This would allow that, for each domain, the process could benefit from issues such as specific vocabulary or the structure of the documents, thus increasing its performance in the specific scenario with respect to a generic TS on their own.

3.2.1 Literary text

Attempts to summarise literary texts, either short stories (Kazantseva 2006) or longer texts, i.e. books (Mihalcea and Ceylan 2007) have been also addressed in recent years. In Mihalcea and Ceylan (2007), the difficulties of TS when it is addressed to book summarisation are analysed, building a benchmark, where the evaluation of book's summaries is specifically targeted. Moreover, several techniques for book summarisation are addressed as well, for instance text segmentation, suggesting a summarisation approach based on the already existing system MEAD (Radev et al. 2001), with some particular changes, in order to adapt the system to long-document summarisation. Further on, in Ceylan and Mihalcea (2009), two kinds of summaries are generated: objective and interpretative summaries. The former contains information about the events occurring in the books and their plot, whereas the latter attempts to capture the author's ideas and thoughts. From this analysis, it is found that approximately 48% of the objective summaries can be reconstructed by cut-and-paste operations from the original document. However, for interpretative summaries, this number decreases to only 25%. For short stories (Kazantseva 2006), indicative summaries are generated in order to help the user to decide whether to read or not the whole document. The relevant information to include is determined based on linguistic traits, such as grammatical tenses, temporal expressions, voice, meaning of the verb, and type of speech (direct or indirect).

3.2.2 Patent claims

The particular writing style of patents, although difficult to process due to the kind of language employed, has been also targeted for TS. An interesting approach performing at approximately 60% (F-measure) can be found in Mille and Wanner (2008) where a multi-lingual summarisation system for Spanish, English and French is developed taking profit of the structure of the patent claims and employing discourse and semantic features as well as dependency patterns for the summarisation process. They also perform linguistic simplification in order to give the resulting summaries an abstractive nature. A complete text mining approach for patent analysis, including a TS stage is proposed in Tseng et al. (2007). With respect to TS, the extractive process ranks sentences based on the frequency of keywords, similarity to the title of the patent claim, and the cue words it contain. Also, positional features are considered. Finally, all these features are combined in a linear way and the highest weighted sentences, up to a desired length, are selected to form the final summary. Key phrases are also identified and used as feature for determining the relevance of a sentence in Trappey and Trappey (2008). Moreover, clustering techniques are employed to obtain the information density of a sentence. However, different from the previous approaches, the novelty in this approach relies on incorporating domain-specific features based on phrases and topic sentences of a given patent document. Trappey et al. (2009) extend the previous work with ontological knowledge in order to retrieve the domain-specific keywords and phrases using concept hierarchies and semantic relationships. Results are evaluated in terms of compression and retention ratios.

On the one hand, the compression ratio is the ratio of the word counts between the summary and its original document. On the other hand, the retention ratio indicates the average value of the recall ratio and the precision ratio. Results show that the best compression ratio is 20%, which is line of state-of-the-art (Morris et al. 1992), and the use of ontologies improves the retention ration.

3.2.3 Image captioning

The need of producing short descriptions of images can be also seen as a TS problem, where a summary is produced from a set of related documents referring to an image annotated with geographical information. In (Deschacht and Moens 2007) image captions using the associated information related to an image are produced. They take profit of the immediate context of the image to extract such information, for instance, text in HTML tags. Their main purpose is to correctly detect and classify entities appearing in images, and then, calculate the salience of such entity with the final goal to produce a short annotation for the image. Similarly, Feng and Lapata (2008) suggest an image annotation model which is able to learn image captions from auxiliary documents and noisy annotations. The auxiliary documents are very useful for this task since they can provide important information related to the image, thus allowing to generate more accurate image descriptions. Aker and Gaizauskas (2009) propose an approach based on language models (n-grams) to generate 250 word-length summaries for image captions using the corpora described in Aker and Gaizauskas (2010). The results obtained are very encouraging, being later improved by means of dependency patterns (Ahmet and Gaizauskas 2010), which performed very close to the ones obtained using the first paragraph of Wikipedia articles as summary baseline. Furthermore, in Plaza et al. (2010), two TS approaches based on statistical features (term frequency and noun-phrase length) and semantic-graphs using WordNet concepts, are also tested within the same corpus. Results for both approaches were acceptable, obtaining around 10% in recall according to ROUGE-SU4 metric, and improving the language models approach originally proposed in Aker and Gaizauskas (2009).

3.2.4 Web 2.0 textual genres

Summaries from new textual genres, such as blogs (Balahur-Dobrescu et al. 2009; Lloret et al. 2009), reviews (Balahur and Montoyo 2008; Zhuang et al. 2006) or threads (Zajic et al. 2008; Balahur et al. 2009) can be also found in the literature. The summarisation techniques used within these approaches are in the line of the ones presented in Sect. 3.1.3, being the integration of sentiment analysis techniques essential for generating summaries from these new genres born with the social Web. Focusing on TS, in Balahur et al. (2009), the techniques employed are based on term frequency counts, whereas in Balahur-Dobrescu et al. (2009) summaries are generated using *Latent Semantic Analysis*. Other approaches, e.g. Zhuang et al. (2006) simply rely on the output of the SA system to group sentences according to their polarity without caring about any TS technique.

4 Combining text summarisation with intelligent systems

The goal of this Section is to present how summaries can help other systems, therefore analysing the applicability of TS within other intelligent systems. This can be considered as a manner of indirectly evaluating summaries, also known as extrinsic evaluation. The rapid growth of the information leads to an increase in the time, when one want to efficiently deal with it. Therefore, summaries can be a good way to allow systems to spend less processing time, if they are used instead of the whole document. Moreover, at the same time, summaries can be suitable for removing noisy information, thus keeping only the really important one. This aspects are derived from the definition of a summary itself, where a summary is a brief but accurate representation of the contents of a document or a set of them. The reason why this Section, although related somehow to evaluation, is independent of Sect. 5 is because here we are more interested in analysing in depth and stressing the usefulness of TS for other systems, such as the ones explained next. In particular, we focus on information retrieval, question answering and text classification.

4.1 Combining text summarisation with information retrieval

The goal of Information retrieval (IR) is to find material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers) (Manning et al. 2008). TS has been combined with IR from a double perspective. On the one hand, several approaches use summaries to benefit IR, for example at the indexing stage, improving the time to retrieve documents and its performance. On the other hand, other approaches take as input for TS the output of IR systems (i.e. the documents retrieved by the IR system) and in some cases, summaries instead of traditional snippets are provided as an output of the IR system. For instance, in Kan and Klavans (2002), summaries are employed to present an alternative visualisation of the documents coming from a standard IR framework. Moreover, the optimal length that a summary should have to be useful for users when using them as output of an search engine is analysed in Kaisser et al. (2008), concluding that the preferred length for the users depends on the type of the query. However, the most common approach is to combine IR and TS in the following manner: the documents related to a topic are retrieved first, and then, a summary taking into account these documents is generated. Therefore, IR helps to gather only relevant documents to a query, while TS selects the most important information from them. Radev and Fan (2000) proposes an domain-independent multi-document summariser, that generates summaries from Web search results. Similarly, SWEeT (Steinberger et al. 2008) relies on a search engine to retrieve relevant documents to a query from the Web, and then summarisation techniques based on *Latent Semantic Analysis* are used to identify and extract the most important sentences from the retrieved documents using, at the same time, cosine similarity to avoid redundancy in the final summaries. The QCS system (Dunlavy et al. 2007) also integrates a IR module but, instead of retrieving documents directly from the Internet, it does so from a static document collection. Once the relevant documents have been retrieved, the system clusters them according to their main topic, and finally a summary is produced for each cluster. The summarisation process is performed in two steps. Firstly, a single-document summary is generated for each document cluster, and then those extracted summary sentences are taken into account to produce the final summary. The way sentences are selected to become part of the summaries is by using a *Hidden Markov Model*, computing the probability of a sentence with regard to whether it is a good summary sentence or not.

Less research has been carried out to analyse how text summaries can be beneficial for IR. In Sakai and Sparck-Jones (2001) it was proven that generic summaries with a compression rate ranging from 10 to 30% were the most appropriate for the indexing stage in IR, concluding that a summary index was as effective as the full text index, for precision-oriented search. In Szlávik et al. (2006), whether summarisation was useful in interactive XML retrieval was investigated, thus providing summaries from XML elements in order to allow users to browse and judge XML documents more easily.

4.2 Combining text summarisation with question answering

Question Answering (QA) aims at automatically answering questions, either simple or complex, posed in natural language (Strzalkowski and Harabagiu 2007). Specific research where different TS approaches have been integrated into a QA system can be found in the literature. The approach shown in Mori et al. (2004) analyses the effectiveness of topic signatures in the multi-document QA summarisation context for a particular type of questions. The generated summaries were about people and contained the answer to the question *"Who is X?"*, where X is a person. It was found that, although topic signatures were able to capture information emphasized in the corresponding source texts, this was not sufficient, as human-made summaries also contained some details that were mentioned, despite not being emphasized.

An interesting approach to QA is presented in Demner-Fushman and Lin (2006), where TS and IR techniques are combined to provide answers to questions belonging to the medical domain. Questions like *"What is the best drug treatment for X?"* are tackled by identifying the drugs from a set of citations first, and then clustering the corresponding abstracts, so that a short extractive summary can be produced for each of them. The summaries are generated by outputting the title of the abstract, the main intervention, and the top-scoring sentence, which is determined using supervised machine learning techniques. Also in the medical domain, the BioSquash system (Shi et al. 2007) summarizes multiple biomedical documents answering a specific question. The system, based on a generic summarizer, has four main components. The *Annotator* module annotates the documents and the question with syntactic and shallow semantic information, and then the relations between concepts in the documents and the questions are determined by the *Concept Similarity* module. The remaining modules, the *Extractor* and *Editor* modules, focus on content selection and linguistic readability, respectively. The final result is a fluent summary relevant to a question concerning the biomedical domain.

Finally, the QAAS system (Torres-Moreno et al. 2009) has resulted from the integration of a TS system with a QA system. In this approach, a generic multi-document summarizer of several compression rates is coupled with a QA system, thus allowing the document search space to be reduced, compared to when the whole document is used. The results obtained show that the number of correct answers returned by the combined system increase. The generic summarizer is also used to identify informative textual zones in documents. However, the limitations of using generic summaries for QA are identified, and for this reason, the generic summarisation system is adapted to a query-focused one by doing query expansion and re-scoring the sentences selected by the generic summarizer according to the terms of the question.

Apart from these approaches, the inverse combination, QA applied to TS can be also found in the literature. In Mori et al. (2005), a QA engine is used to help determine sentence importance, which is calculated based on the scores produced by the QA system and a set of queries. The final objective is to generate a summary, and for this purpose, the QA engine is finally integrated into a generic multi-document summarizer.

4.3 Combining text summarisation with text classification

Text Classification (TC), also known as text categorisation, aims at automatically sorting a set of documents into categories from a predefined set (Sebastiani 2002). In Ker and Chen

(2000), summarisation features (i.e. position and word frequency) are used to categorise news according to different categories (e.g. “money”) reaching 82% for the precision value. Taking text summaries instead of full documents is the approach suggested in Shen et al. (2004), under the assumption that they may be a good noise filter. Since Web pages contain too much irrelevant information which can be detrimental for TC, summaries can extract the most important information, producing a new text, which is then used for classification purposes. The TS approaches used are based on term frequency and *Latent Semantic Analysis*. They carry out a large experimentation with more that 150,000 Web pages and 64 categories. The results obtained show that the proposed summarisation-based classification algorithm improves approximately 8.8% compared to full documents. The same idea and dataset is analysed in Shen et al. (2007) where, in addition, the optimal compression rate for summaries is studied. Summaries of 20 and 30% reach the best results. However in the range from 10 to 40%, it is proved that text summaries can improve the classification performance to some extent over full documents.

The rating-inference task can be seen as a particular type of TC. Its goal is to identify the author’s evaluation of an entity, product, service, etc. with respect to an ordinal-scale based on his/her textual evaluation of the entity (Pang and Lee 2005). Therefore, it can be considered as an opinion classification problem. Usually, opinions are classified with regard to two or three dimensions, subjective vs. objective, or positive vs. neutral vs. negative, respectively (Wilson et al. 2005). However, it is frequent to find texts where users give a score, depending on how much they liked or not a product, movie, restaurant, hotel, service, etc. which is normally associated with a scale rating (1 = worst,...5 = best). When tackling this task with short documents (reviews containing at most three sentences), the classification process achieves good results (74%) (Saggion and Funk 2009). On the contrary, it only reaches 32% when dealing with longer texts. As a consequence, and taking as a basis the assumption aforementioned concerning the suitability of TS for filtering noise, in Lloret et al. (2010) and Saggion et al. (2010), a preliminary set of experiments is carried out for predicting the rating of a review using text summaries instead of full documents. These experiments comprise the analysis of a wide range of summarisation types of different compression rates. In particular, generic, query-focused and sentiment-based summaries are studied, together with several types of baselines, including the first and the last sentences of a review. Although it is claimed that query-focused and sentiment-based summaries may be more appropriate for the rating inference task, this analysis has two main limitations. On the one hand, this task is very complex, since they work at a very fine-grained granularity, and for instance, the differences between a text rated as 4 from one rated as 5 can be very subtle. On the other hand, the dataset used is very small (only 89 reviews) to obtain strong evidence on what type of summaries could be more appropriate.

5 Text summarisation evaluation

Methods for evaluating systems can be broadly classified into two broad categories: *intrinsic* or *extrinsic* (Spärck-Jones and Galliers 1996). In the context of TS, the former assesses a summary itself, for example according to its information content, whereas the latter focus on testing the effectiveness of a summarisation system on other applications (e.g. IR).

Regarding the intrinsic evaluation, there are different methods that can be taken into account to evaluate a summary. According to Mani (2001a), it can be distinguished between evaluating the *quality* or the *informativeness* of a summary. Apart form these two, another one is proposed, *fidelity to the source*, which determines summary informativeness in the context of the source document, that is, if the summary contains the same or similar relevant concepts as the source document has. The problem with this method is how to account for the relevant concepts in the source document. However, despite having different intrinsic methodologies to evaluate summaries, the most common approach is to evaluate summary informativeness by comparing its content to a human-model one, which is considered a reference summary. Following this idea, several methods have been developed which are analysed in Sect. 5.1. These methods focus on the amount of information reflected in the summary, i. e. its content. However, due to the inherent subjectivity associated to summaries, it is not possible to build a fair gold-standard to decide that any automatic summary not similar to this would not be a good summary. The conception of what a good summary is varies a lot between different people, and also depends on what the summary is intended for. This and other problems regarding with this type of evaluation are also explained in Sect. 5.3. Furthermore, other evaluation approaches are more concerned with a qualitative evaluation, which aims at evaluating the quality of a summary with respect to different criteria, such as grammaticality or focus. Although these issues are very difficult to automate, several approaches attempting to provide this type of evaluation have emerged in recent years and they will be described in Sect. 5.2.

With respect to the extrinsic evaluation methods, several scenarios have been proposed as surface methods for summarisation evaluation inspired by different disciplines (Hassel 2007). Examples of these scenarios are: *The Shannon Game*, which aims at quantifying information content by guessing tokens, so that the original document can be recreated; *The Question Game*, in which the reader's understanding of the summary and ability to convey the main concepts is tested; *The Classification Game*, which consists of determining the category either for original documents or for summaries, and measure the correspondence between them; and *Keyword Association*, in which a list of keywords is provided and the goal of the task is to check whether summaries contain such words or not. Also, in Mani (2001a) different extrinsic evaluation methods are outlined, such as *relevance assessment*, in which subjects are asked to determine the relevance of a topic, whether in a summary or source document; or *reading comprehension*, which involves answering multiple-choice tests, having read either the summary or the whole document. Moreover, the application of TS in other types of systems, such as IR, QA, or TC previously described in Sect. 4 could be also considered as a manner to extrinsically evaluate summaries. Therefore, if the summaries are able to help other tasks, it can be assumed that they are good enough to be used to improve other applications, although they may be far from perfect with respect to issues such as coherence.

5.1 Informativeness evaluation

A lot of effort has been done to develop evaluation tools, that can be used to assess the informativeness of a summary at different levels. The well-known information retrieval metrics, precision, recall and F-measure (Van Rijsbergen 1981) have been also adapted to TS. Recall evaluates which portion of the sentences selected by a human are also identified by a system, whereas precision is the fraction of these sentences identified by a system that are correct (Nenkova 2006). F-measure is a combination of both precision and recall. However, using these metrics, it is possible that two equally good extracts are judged very differently. *Relative Utility* (Radev and Tam 2003) was proposed as another metric for evaluation, in which multiple judges rank each sentence in the input with a score, giving them a value between

0 to 10, with respect to its suitability for inclusion in a summary. Therefore, the higher the sentences were ranked, the more suitable for a summary were.

In the light of this, an information-based summarisation metric, called *factoid score*, is suggested in Teufel and Halteren (2004), so automatic summaries are evaluated regarding factoids (atomic information units which represent the meaning of a sentence). The idea is to use several reference summaries as gold standard and measure the information overlap among them. The Pyramid method (Passoneau 2010) follows a similar philosophy. Its goal is to identify information with the same meaning across different human-authored summaries, which is called *Summary Content Units* (SCU). Each SCU has a weight depending on the number of human assessors, who expressed the same information, and these weights follow a specific distribution, allowing important content to be differentiate form less important one. In order to avoid the laborious task to annotate manually all the SCU, an attempt to automatically performed this is suggested in Fuentes et al. (2005). However, the manual effort need to detect factoids or SCU along a collection of summaries, is still one main disadvantage of both methods.

Inspired in BLEU (Papineni et al. 2002), a tool for automatic evaluate the output of a machine translation system, a tool for automatically evaluating a summary is developed in Lin (2004). Its name, ROUGE, stands for *Recall-Oriented Understudy for Gisting Evaluation*, and it obtains precision, recall, and F-measure.[15] This tool relies on n-gram co-occurrence, and the idea behind it is to compare the content of a summary with one or more reference summaries, and see the number of n-gram of words they all have in common. Different types of n-grams can be computed, such as unigrams (ROUGE-1), bigrams (ROUGE-2) or longest common subsequence (ROUGE-L). The hypothesis of this method is that two texts that have a similar meaning, they must also share similar words or phrases.

Under the assumption that the best similarity metric should be the one that best discriminates between manual and automatically generated summaries, QARLA, an evaluation framework described in Amigó et al. (2005), is suggested. Having a set of reference summaries, a set of automatic ones, and a set of similarity metrics, the framework provides the following types of measures: QUEEN, which is an estimation of the quality of an automatic summary; KING, as an estimation of the quality of a similarity metric; and finally, JACK, as a measure to indicate the reliability of the automatic summaries set. In this approach, a total number of 59 different similarity metrics are used, including recall, precision, sentence length and frequency and grammatical distribution metrics. Recall and precision metrics are based on ROUGE n-gram overlap method, combining several n-gram options with different pre-processing steps.

Moreover, in order to tackle and improve the drawbacks derived from comparing fixed n-gram words, a new evaluation framework is proposed in Hovy et al. (2006). The underlying idea of this method is to split a sentence into very small units of content, called *Basic Elements*[16] (BE), which are defined as triplets of words consisting of a head and a modifier or argument, with its relation to the head (*head* | *modifier* | *relation*). Its goal is to allow greater flexibility for the matching of different equivalent expressions. Further on, to address the shortcomings of different but equally good expressions, *ParaEval* (Zhou et al. 2006) is developed. Its objective is to provide a summarisation evaluation method, facilitating the detection of paraphrase matching. The paraphrase detection is performed according to three-level strategy. First, multi-word paraphrases between phrases in the reference summaries,

[15] The latest version, ROUGE-1.5.5, allows us to obtain precision, recall and F-measure. Previous versions only computed the recall value.

[16] http://www.isi.edu/publications/licensed-sw/BE/index.html.

as well as in the automatic summaries are identified. Then, for those fragments that do not matched, the method tried to find synonyms between single-words, and if this also fails, simple lexical matching is finally performed.

In Branny (2007), text grammars are employed to automatically evaluate text summaries. A text grammar is a way of describing a valid text structure in a formal way (Van Dijk 1972), and it takes into consideration surface and deep structure by means of relations between sentences (microstructures) and the structure of the text as a whole (macrostructure), respectively. Under the assumption that vocabulary overlapping is not enough to measure the informativeness of a summary, this approach relies on a list of propositions previously identified, and the humans have to decide whether each proposition is relevant or not and establish several groups in order to face the problem of quantifying the informativeness of each proposition. Then, three scores are proposed, based on *informativity* (how many propositions are present in the summary), *misinformation* (misleading statements of the summary are detected) and *t-grammaticality* (which is related to the correctness of the sentences based on orthographical or grammatical issues, as well as coherence problems). The application of this method on human-written and automatic summaries shows that human summaries get higher scores than automatic ones, as it is expected. However, the main drawback of this method is that, although results show that human summaries are better than automatic ones, it is not possible to know how well would it correlate with human evaluation. Moreover, human intervention is required for identifying propositions and evaluate the amount of misinformation and ungrammaticality summaries have, which is very costly and time-consuming and it would not be easily scalable.

AutoSummENG[17] (Giannakopoulos et al. 2008b) is another automatic method recently developed which has been proven to have high correlation with human judgements. This method differs from the others in three main aspects: (1) the type of statistical information extracted; (2) the representation chosen for this extracted information, and (3) the method used to calculate the similarity between summaries. Here, the comparison between summaries is carried out by building first n-gram character graphs, and then comparing their representations to establish a degree of similarity between the graphs. Moreover, its methodology is language-neutral, so it is expected to work in other languages, as well.

Owczarzak (2009) suggest DEPEVAL(summ), which is a dependency-based metric. The idea here is similar to Basic Elements, and it consists of comparing dependency triples extracted from automatic summaries against the ones from model summaries. The main difference with Basic Elements is the parser used. Whereas Basic Elements uses Minipar,[18] DEPEVAL(summ) is tested with different parsers, for instance the Charniak parser.[19]

GEMS (Generative Modelling for Evaluation of Summaries) (Katragadda 2010) suggests the use of signature terms in order to analyse how they are captured in automatic summaries. The signature terms are calculated on the basis of part-of-speech tags, such as nouns or verbs; query terms and terms of reference summaries. The distribution of the signature terms is calculated in the source document and then the likelihood of a summary being biased towards such signature-terms is obtained.

On the other hand, different evaluation methodologies have been proposed specifically for other languages apart from English, such as for Chinese using the HowNet resource. HowNet[20] is an electronic knowledge system for English and Chinese, and differs from other

[17] AUTOmatic SUMMary Evaluation based on N-gram Graphs: http://www.iit.demokritos.gr/~ggianna/.

[18] http://webdocs.cs.ualberta.ca/~lindek/minipar.htm.

[19] ftp://ftp.cs.brown.edu/pub/nlparser/.

[20] http://www.keenage.com/.

Table 4 Types of informativeness evaluation methods

Approach	Automatic	Semi-automatic
Relative utility		✓
Factoid score		✓
Pyramid method		✓
ROUGE	✓	
QARLA	✓	
BE	✓	
Text grammars		✓
ParaEval	✓	
AutoSummENG	✓	
DEPEVAL(summ)	✓	
GEMS	✓	
HowNet eval.	✓	

existing lexical databases, for instance WordNet,[21] in the way in which word similarity is computed. Moreover, HowNet provides richer information and each concept is represented unambiguously by their definition and association links to other concepts. It is a well-known resource for the Chinese language, and has been applied to many approaches. In (Wang et al. 2008) an approach for evaluating summaries based on HowNet is proposed. Despite the fact that this method is also based on n-gram co-occurrence statistics, its main novelty is the use of HowNet to compute word similarity, so that synonyms can be also taken into consideration. In addition, it is also shown that a few quality metrics could also be detected to some extent, such as conciseness or sequence ordering.

Table 4 depicts the aforementioned methods distinguishing between automatic and semi-automatic methods. The latter will refer to those methods, such as Pyramid, which need some kind of human annotations. It is worth stressing upon the fact that all of the automatic methods, except the one relying on text grammars, rely on reference human-written summaries that are used to evaluate automatic ones, which can lead to some problems if the automatic summary differs greatly from the human-written with respect to the vocabulary it contains.

Research in automatic content evaluation of summaries has been gained special attention thanks to the AESOP track first proposed at TAC 2009, which encourages research into this issue. Improved versions of the aforementioned approaches were proposed in the context of this task, such as Giannakopoulos and Karkaletsis (2009) or Conroy et al. (2009), which experiment with different approaches based on ROUGE ideas. Due to success in participation of the task, it is again considered for the new edition of the conference (TAC 2010).

5.2 Quality evaluation

Generally speaking, one shortcoming of the existing evaluation methods is that they only assess the quality of a summary according to its content, and they do not take into consideration other important aspects, such as coherence or non-redundancy. The evaluation concerning the quality of a summary taking into consideration other issues different from its informativeness, has always been in mind of the researchers. The goal of the FAN Protocol described in Minel et al. (1997) was to assess the quality of an abstract independently from

[21] http://wordnet.princeton.edu/.

the source text and the information it contained. Four criteria were proposed: (1) number of anaphora deprived of referents; (2) rupture of textual segments; (3) presence of tautological sentences; and (4) legibility of the abstract. All these criteria were evaluated manually by two jurors. In the light of this, another protocol was also proposed: the MLUCE Protocol. The idea behind this protocol is to enable potential users to evaluate summaries, depending on what they wanted the summary for. For instance, if they want the summary just to decide to read the whole document or not, or on the contrary, to serve as a synthesis of the source document. Again, the evaluation is carried out manually as it is for the FAN Protocol. Attempts to evaluate indicativeness and acceptability have also been addressed in Saggion and Lapalme (2000). The former measures whether the summary is able to extract the topics of the document, whereas the latter determines if a selected sentence by a summarisation system is adequate compared to what humans would have selected, so human intervention is needed to evaluate this criterion.

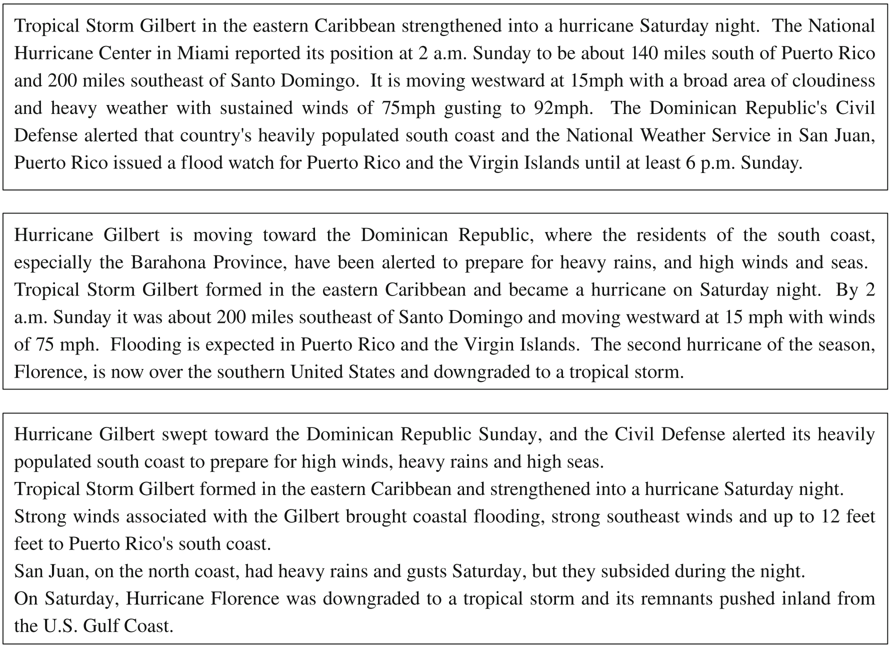
More recently, in Conroy and Dang (2008), the need of having tools which assess content as well as other linguistic aspects is addressed. In the DUC and TAC conferences, summaries are evaluated with respect five linguistic quality questions (*Grammaticality*, *Non-redundancy*, *Referential clarity*, *Focus*, and *Structure and Coherence*) which do not involve any comparison with a reference summary. This type of evaluation is manually performed by expert human assessors, who score the quality of a summary according to a five-point scale. Furthermore, in the literature we can find some studies to predict text quality through the analysis of various readability factors (Pitler and Nenkova 2008). The idea here is to analyse the quality of a text by means of different criteria including vocabulary, syntax, or discourse, in order to account for the correlation between those factors and human readability ratings previously gathered. Each criterion is modelled in a different way, for instance, vocabulary is represented in terms of unigrams, and syntax is modelled via features, such as average number of noun-phrases or average number of verb-phrases. Results show that when combining all proposed readability factors, the prediction obtains an accuracy close to 90%, and therefore this idea could be applied and extended to evaluate the quality of a summary. Other approaches, such the ones presented in Barzilay and Lapata (2005), Lapata and Barzilay (2005), or Hasler (2008) focus on modelling local coherence from different perspectives, with the purpose of representing and measuring text coherence. Approaches range from the *Centering Theory* (Grosz et al. 1995) to the development of syntactic and semantic models, in order to capture the distribution of entities in the text, or the degree of connectivity across sentences, respectively. Attempts to automatically evaluate the grammaticality of a summary have been explored in Vadlapudi and Katragadda (2010b). N-gram models, in particular unigrams, bigrams, trigrams and the longest common subsequence are used for capturing this aspect. In addition, this problem is considered as a classification problem, where summary sentences are classified into classes on the basis of their acceptability. The acceptability parameter is estimated using trigrams. The proposed methods are evaluated in the same way as were summaries evaluated in DUC or TAC conferences. Results obtained correlate well (85% at most) with respect to the already existing manual evaluations. Furthermore, in Vadlapudi and Katragadda (2010a), structure and coherence aspects are also investigated on the basis of lexical chains and the semantic relatedness of two entities. Results using this approach achieve a 70% using Spearman's correlation.

5.3 Limitations of the existing evaluation methods

Although the aforementioned methods really help to evaluate automatic summaries, there are several questions related to the evaluation that remain still unsolved. Therefore, some shortcomings of the evaluation task in general can be addressed as well as some specific problems that evaluation tools have.

First of all, to have a collection of documents and their corresponding summaries is a very costly task. On the one hand, different humans would write different summaries, ones containing more abstraction, others synthesizing more the information, and so on. This means that it is possible to have several valid summaries, although different in content. Figure 1 shows three different examples of summaries (A, B, and C). In addition, their corresponding original document can be seen in Fig. 2. These documents refer to document AP880911-0016 taken from the DUC 2002 corpus. It is worth noting that summaries A and B, at the top of the figure, respectively, are produced by expert humans, whereas summary C is generated automatically using a system that extracts the most relevant sentences from a text, employing textual entailment to remove redundancy and word frequency as weighting scheme, giving more importance to those sentences which contain words with high frequency (Lloret et al. 2008). As a consequence, summaries A and B have an abstract nature, whilst the summary produced automatically has been produce following an extractive approach. As can be seen from the given examples, none of them are identical, although they contain similar information. Some vocabulary is shared among all the summaries (*"tropical storm Gilbert in the eastern Caribbean"*), but other facts are expressed using different words (*''Puerto Rico issued a flood watch for Puerto Rico"*, *"flooding is expected in Puerto Rico"*, *"Gilbert brought coastal flooding [...] to Puerto Rico's south coast"*). In order to evaluate the content of each summary, we computed ROUGE scores taking as a reference another human-written summary different from the ones shown in Fig. 1. As far as the recall value for ROUGE-1 is concerned, 46% is obtained for the automatic summary, whereas the results for the human-written summaries are 53 and 49%, respectively. Although these values are very close, it means that the first human summary would be the best one, but as can be seen, all the summaries contain relevant information. Moreover, it may happen that if these summaries were evaluated manually by different people, results might vary, due to the inherent subjectivity of the summary evaluation process. Therefore, it is a very delicate matter to decide which of these summaries is the best, and how we can take such a decision. Consequently, choosing human-authored summaries as a gold standard may not be the optimal solution to the summarisation evaluation problem.

Regarding the above examples, the difficulty increases when a summary produced following an extractive approach has to be compared with another generated by an abstractive one. Different studies have proven that, if humans had to decide which sentences from documents were most relevant to belong to the final summary in order to produce extracts, they would also disagree in which sentences best represent the content of a document. Therefore, the low agreement between humans is a problem when their summaries are used for comparison. In Donaway et al. (2000) it was shown the variance in the recall value, depending on which human summary was taken as reference for comparison with the automatic one. This problem was also stated in Mani (2001b). Furthermore, the semantic equivalence between different nouns, for example by means of synonymy, or expressions, when there are various ways to express the same idea, is another drawback of the evaluation (Nenkova 2006), because most methods only perform a superficial analysis, and do not take into consideration the semantic meaning of phrases. Concerning automatic methods, some criticisms against the ROUGE tool were made. On the one hand, in Sjöbergh (2007) the fact that a summary could be easily produced in order to get high ROUGE scores was addressed. To prove this hypothesis a simple summarisation method was developed, using a greedy word selection strategy, and although the generated summaries were not good from a human's point of view, they obtained

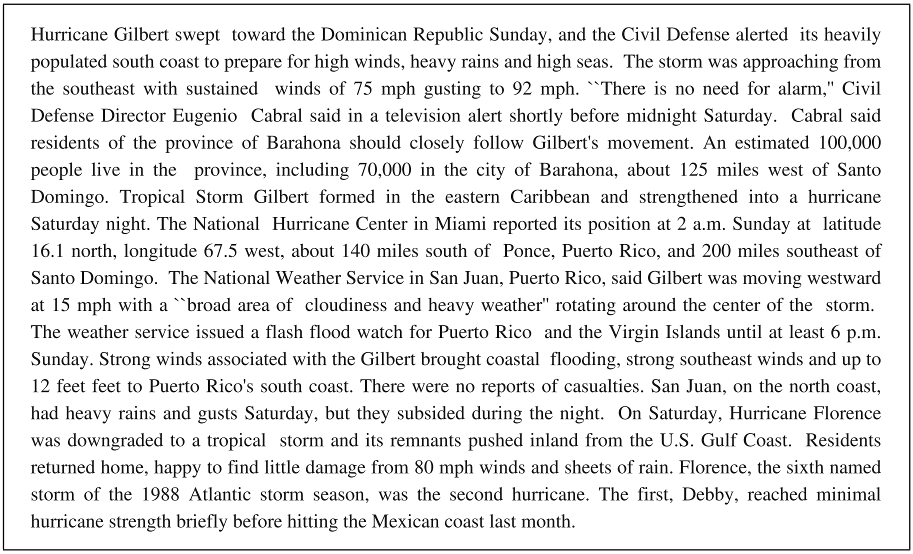


Tropical Storm Gilbert in the eastern Caribbean strengthened into a hurricane Saturday night. The National Hurricane Center in Miami reported its position at 2 a.m. Sunday to be about 140 miles south of Puerto Rico and 200 miles southeast of Santo Domingo. It is moving westward at 15mph with a broad area of cloudiness and heavy weather with sustained winds of 75mph gusting to 92mph. The Dominican Republic's Civil Defense alerted that country's heavily populated south coast and the National Weather Service in San Juan, Puerto Rico issued a flood watch for Puerto Rico and the Virgin Islands until at least 6 p.m. Sunday.

Hurricane Gilbert is moving toward the Dominican Republic, where the residents of the south coast, especially the Barahona Province, have been alerted to prepare for heavy rains, and high winds and seas. Tropical Storm Gilbert formed in the eastern Caribbean and became a hurricane on Saturday night. By 2 a.m. Sunday it was about 200 miles southeast of Santo Domingo and moving westward at 15 mph with winds of 75 mph. Flooding is expected in Puerto Rico and the Virgin Islands. The second hurricane of the season, Florence, is now over the southern United States and downgraded to a tropical storm.

Hurricane Gilbert swept toward the Dominican Republic Sunday, and the Civil Defense alerted its heavily populated south coast to prepare for high winds, heavy rains and high seas.
Tropical Storm Gilbert formed in the eastern Caribbean and strengthened into a hurricane Saturday night.
Strong winds associated with the Gilbert brought coastal flooding, strong southeast winds and up to 12 feet feet to Puerto Rico's south coast.
San Juan, on the north coast, had heavy rains and gusts Saturday, but they subsided during the night.
On Saturday, Hurricane Florence was downgraded to a tropical storm and its remnants pushed inland from the U.S. Gulf Coast.

Fig. 1 Examples of human and automatic summaries (summaries A, B and C, respectively)



Hurricane Gilbert swept toward the Dominican Republic Sunday, and the Civil Defense alerted its heavily populated south coast to prepare for high winds, heavy rains and high seas. The storm was approaching from the southeast with sustained winds of 75 mph gusting to 92 mph. ``There is no need for alarm," Civil Defense Director Eugenio Cabral said in a television alert shortly before midnight Saturday. Cabral said residents of the province of Barahona should closely follow Gilbert's movement. An estimated 100,000 people live in the province, including 70,000 in the city of Barahona, about 125 miles west of Santo Domingo. Tropical Storm Gilbert formed in the eastern Caribbean and strengthened into a hurricane Saturday night. The National Hurricane Center in Miami reported its position at 2 a.m. Sunday at latitude 16.1 north, longitude 67.5 west, about 140 miles south of Ponce, Puerto Rico, and 200 miles southeast of Santo Domingo. The National Weather Service in San Juan, Puerto Rico, said Gilbert was moving westward at 15 mph with a ``broad area of cloudiness and heavy weather" rotating around the center of the storm. The weather service issued a flash flood watch for Puerto Rico and the Virgin Islands until at least 6 p.m. Sunday. Strong winds associated with the Gilbert brought coastal flooding, strong southeast winds and up to 12 feet feet to Puerto Rico's south coast. There were no reports of casualties. San Juan, on the north coast, had heavy rains and gusts Saturday, but they subsided during the night. On Saturday, Hurricane Florence was downgraded to a tropical storm and its remnants pushed inland from the U.S. Gulf Coast. Residents returned home, happy to find little damage from 80 mph winds and sheets of rain. Florence, the sixth named storm of the 1988 Atlantic storm season, was the second hurricane. The first, Debby, reached minimal hurricane strength briefly before hitting the Mexican coast last month.

Fig. 2 Example of a source document

good results in some ROUGE values, for example a recall score of 41% for ROUGE-1.[22] On the other hand, the correlation between ROUGE and human summaries was shown to be lower than it was claimed, especially in other types of summarisation, for instance in

[22] Although this value could seem not to be very high, it is quite good with regard to the state-of-the-art.

speech summarisation (Liu and Liu 2008). Regarding the Basic Elements method, its main drawback is that it uses some language-dependent preprocessing modules for parsing and cutting, so it would be not very easy to apply it to other languages, especially if they lack such resources.

With respect to the quality evaluation, the suggested methods are still at their early steps. Performing this kind of evaluation is really difficult and develop good metric that correlate well with human assessment is not a trivial issue. Normally, expert judges are asked to evaluate summaries using a 5-point scale, comprising qualitative values with respect to a question or statement. For example, in the evaluation performed at DUC 2001, the questions had this form:

To what degree does the summary say the same thing over and over again?

1. Quite a lot; most sentences are repetitive
2. More than half of the text is repetitive
3. Some repetition
4. Minor repetitions
5. None; the summary has no repeated information

What does exactly mean *"some repetition"* or *"minor repetition"*? It would be possible to map into a quantitative scale the first and the last claim, for example, none would mean any repetition at all, but the boundaries between the middle ones are very subtle. Moreover, statements like the ones in the example contains a degree of subjectivity, which is not possible to capture automatically. All this issues make the task of evaluating a summary's quality really challenging and difficult to tackle from an automatic point of view.

6 Future directions in text summarisation

Since the late 50's, summarisation methodologies and systems have experimented great advances. New approaches are developed, taking also into account linguistic aspects in some cases, which allows an automatic summary to be something else than a mere joining of sentences. Furthermore, these advances have led to new types of summaries (e.g. update or personalised summaries) and new scenarios where summaries play a crucial role (e.g. patent claims, blogs, reviews). However, there is still a lot of room for improvement, especially due to the large amounts of available data in different formats, and the rapid development of the technology, which brings new challenges for this research field, such as multi-document, multi-lingual or multimedia summarisation. Another challenging issue is the evaluation, which raises the following questions: *"is it possible to evaluate a summary in an objective way?"*; *"is it fair to compare automatic summaries against human-written?"*; *"how could the quality of a summary be assessed automatically?"*.

The aim of this paper was to conduct a research into the state-of-the-art in TS, focusing especially in the new types of summaries and scenarios raised in the last years. However, in order to provide some basic background information about this research field, a brief review of well-known summarisation approaches was also conducted, grouping these approaches into different groups as far as the techniques or algorithms employed was concerned. Moreover, international forums and conferences with regard to TS that have been taken place along the years were described, in order to provide a historical perspective of this research field, highlighting how it has evolved, and the attention that the research community has paid to it. It is very interesting to see the impact of TS in the community research, since it can be concluded that, despite its age and difficulty, it attracts more and more researchers. An important part of this literature review deals with the current state-of-the-art of TS, where it can be seen how it has been adapted to the new society requirements. This has resulted in new types of summaries, and new scenarios in which summarisation can be of great help. Furthermore, the combination of TS with other intelligent systems, in particular, IR, QA or TC is also analysed, since text summaries can improve the performance of other systems, being very appropriate to use them in combination with other applications. Finally, a survey in TS cannot avoid the evaluation problem, and consequently, a special attention is given to this issue as well. Although there are several evaluation metrics and methods, which are able to evaluate either the informativeness of a summary (if the summary contains the right information), or its quality (if the information is clearly expressed and legible), there are still some challenging aspects, allowing a lot of room from improvement regarding this topic. However, it is essential to analyse the current evaluation methods, and their limitations, in order to give some insights for future evaluations.

The analysis conducted in this paper allows us to have basic information about the past of TS, the current state-of-the-art, and possible trends for the future. As far as the TS approaches is concerned, it is worth mentioning that over the years, existing approaches are changing. For instance, new machine learning algorithms are proposed for tackling TS; however, the features used do not change too much with respect to the ones already existing (e.g. term frequency, part-of-speech, position). What it seems to be changing fast over the years is the types of summaries, as the society has to adapted to new user requirements. Whereas at the beginning generic and single-document summarisation were one of the most important types, currently multi-document summarisation and even multi-lingual has gained great importance due to the vast amount of information we have to deal with. This can also be seen in the international forums devoted to TS, which, year after year, the proposed tracks are updated. Moreover, in the last years, there is a tendency towards the generation of summaries with specific purposed, such as sentiment-based or personalised summaries. It is worth stressing upon the fact that the generation of abstracts is also becoming very important, and as long as the current existing techniques improve, it is possible to employ them to generate summaries that are closed to the human-written ones. Also, new scenarios different from the traditional ones, such as newswire, or scientific papers are also increasing, and currently we can find summarisation in a wide range of genres and document types, for instance, literary documents, blogs, etc.

The evaluation of a summary, either automatic or human-written, is a delicate issue, due to the inherent subjectivity associated to the process. In this paper, we presented two different types of evaluation, intrinsic and extrinsic, depending on whether we want to assess the summary itself or its performance for meeting the goals of other applications. The evaluation carried out by the current methods and tools is mostly intrinsic. Moreover, it can be distinguished between informativeness and quality evaluation. As it could be seen, most of the current automatic tools evaluate the content of a summary, and only a few approaches attempt to determine whether the generated summary has high quality or not, with regard to different criteria, such as grammaticality or coherence. The existing methods have some limitations, in the sense that they mostly relied on vocabulary overlap between an automatic and a model summary. On the other hand, regarding the quality evaluation, in recent years there has been a surge of novel approaches for automating this process, which is a further step towards this research field, since this type of evaluation has been performed completely manually by expert judges so far. Although they are still very preliminary, they represent a good starting point to tackle this issue. In order to successfully face the problem of qualitative scales, we should first research into methods capable of mapping qualitative aspects (e.g. minor repetitions) into quantitative (e.g. *how many repetitions (in numbers) should be present in order to be considered as minor? 10% of the text?, 5% of the named entities?*).

Some conclusions about the tendencies of TS for the future can be drawn. As we previously said, the society requirements change, and the information grows at an exponential rate, which forces TS to adapt to new needs. For the next years, multi-document and multilingual summarisation will be essential, since the same information can appear in a high number of documents but also in different languages. And it is worth mentioning that this information has to be presented in a coherent way and going beyond the concatenation of sentences. Therefore, abstractive paradigms or at least hybrid ones will become one of the main challenges to solve. Hybrid approaches would be capable of identifying and selecting relevant fragments of information, and then merge, compress or delete such information in order to generate new summary information. As a consequence, it is possible to take into account the benefits of extractive and abstractive approaches together. Moreover, since nowadays users play a crucial role on the Internet, sentiment-based summaries, personalised and update summaries will be also very important, because a summary should provide the exact information a user requires. How to present such information is another issue that is becoming more and more important in the sense that, traditionally, the input and the output of a TS system is text. However, this tendency is changing and we can find many approaches summarising other types of input, such as meetings, or video, or producing the output in a format different from text.[23] This would consist for instance in taking text as input, but on the contrary presenting the summary in another format, for instance, by means of statistics, tables, graphics, visual rating-scales, etc. which would allow users to visualize the results immediately, and maybe find the information they are interested in more quickly. Besides, these visual representations could be also complemented by text summaries. Concerning the evaluation, the main issues to focus on is the intrinsic evaluation, carrying out research in novel method and ways to assess the summary, both according to the information it contains and how it is presented. Due to the fact that the evaluation process contains a high degree of subjectivity, we still do not know whether the process could be fairly automated, because a good criterion for distinguishing what to consider relevant and what not, should be first defined. The same happens for evaluating the quality of a summary. Although some guidelines are set for carrying out this task, there is always a margin of subjectivity, which generate different results when the same summary is evaluated by two humans. Apart form this, as long as semantic methods make progress it will be more feasible to account for equivalent expressions. This will help automatic methods, since it would be possible to determine if the summary contains appropriate content or not.

Finally, despite having more that 50 years old, TS is still alive with a great interest among the research community. Indeed, this research field is very dynamic, since it is continuously adapting itself to the new needs and challenges. Although the performance of TS is still moderate and the generated summaries are far from perfect, it can be seen how the combination with other systems leads to the improvement of the overall performance of the combined system, helping to develop even more intelligent systems. The evaluation of a summary presents also great challenges, which are being tackled for years, and improvements are obtained over past approaches. All the possibilities that TS offers together with the extensive application it has in the real world, make it an interesting research field to conduct research into, and this survey provided a good point to start with in order to acquire a general overview of all the main issues with respect to it.

[23] In this paper, we only focused on text, and consequently, although there exists other input/output formats, they were out of the scope of this paper.

Acknowledgments This research has been supported by the project PROMETEO “Desarrollo de Técnicas Inteligentes e Interactivas de Minería de Textos” (2009/119) and the project reference ACOMP/2010/286 from the Valencian Government, as well as by the Spanish Government (grant no. TIN2009-13391-C04-01). Elena Lloret is funded by the FPI program (BES-2007-16268) from the Spanish Ministry of Science and Innovation, under the project TEXT-MESS (TIN2006-15265-C06-01).

References

Agnihotri L, Kender JR, Dimitrova N, Zimmerman J (2005) User study for generating personalized summary profiles. In: Proceedings of the IEEE international conference on multimedia and expo (ICME). pp 1094–1097

Ahmet A, Gaizauskas R (2010) Generating image descriptions using dependency relational patterns. In: Proceedings of the 48th annual meeting of the association for computational linguistics

Aker A, Gaizauskas R (2009) Summary generation for toponym-referenced images using object type language models. In: Proceedings of the international conference on recent advances in natural language processing (RANLP-2009)

Aker A, Gaizauskas R (2010) Model summaries for OPTlocation-related images. In: Proceedings of language resources and evaluation

Amigó E, Gonzalo J, Peñas A, Verdejo F (2005) QARLA: a framework for the evaluation of text summarization systems. In: ACL ’05: proceedings of the 43rd annual meeting on association for computational linguistics. pp 280–289

Ando R, Boguraev B, Byrd R, Neff M (2005) Visualization-enabled multi-document summarization by Iterative Residual Rescaling. Nat Lang Eng 11(1):67–86

Angheluta R, Busser RD, Francine Moens M (2002) The use of topic segmentation for automatic summarization. In: Proceedings of the ACL-2002 post-conference workshop on automatic summarization. pp 66–70

Aone C, Okurowski ME, Gorlinsky J (1998) Trainable, scalable summarization using robust NLP and machine learning. In: Proceedings of the 36th annual meeting of the association for computational linguistics and 17th international conference on computational linguistics, vol 1. pp 62–66

Azzam S, Humphreys K, Gaizauskas R (1999) Using coreference chains for text summarization. In: Proceedings of the ACL’99 workshop on coreference and its applications

Balahur A, Montoyo A (2008) Multilingual feature-driven opinion extraction and summarization from customer reviews. In: Proceedings of 13th international conference on applications of natural language to information systems. pp 345–346

Balahur A, Lloret E, Ferrández O, Montoyo A, Palomar M, Muñoz R (2008) The DLSIUAES team’s participation in the TAC 2008 tracks. In: Proceedings of the text analysis conference (TAC)

Balahur A, Lloret E, Boldrini E, Montoyo A, Palomar M, Martinez-Barco P (2009) Summarizing threads in blogs using opinion polarity. In: Proceedings of the international workshop on events in emerging text types (eETTs). pp 5–13

Balahur-Dobrescu A, Kabadjov M, Steinberger J, Steinberger R, Montoyo A (2009) Summarizing opinions in blog threads. In: Proceedings of the Pacific Asia conference on language, information and computation conference. pp 606–613

Baldwin B, Morton TS (1998) Dynamic coreference-based summarization. In: Proceedings of the third conference on empirical methods in natural language processing (EMNLP-3)

Barzilay R, Elhadad M (1999) Using lexical chains for text summarization. In: Advances in automatic text summarization. pp 111–122

Barzilay R, Lapata M (2005) Modeling local coherence: an entity-based approach. In: Proceedings of the 43rd annual meeting of the association for computational linguistics (ACL’05). pp 141–148

Barzilay R, McKeown KR (2005) Sentence fusion for multidocument news summarization. Comput Linguist 31(3):297–328

Beineke P, Hastie T, Manning C, Vaithyanathan S (2004) An exploration of sentiment summarization. In: Proceedings of the AAAI spring symposium on exploring attitude and affect in text: theories and applications

Bellemare S, Bergler S, Witte R (2008) ERSS at TAC 2008. In: Proceedings of the text analysis conference (TAC)

Belz A (2008) Automatic Generation of Weather Forecast Texts Using Comprehensive Probabilistic Generation-space Models. Nat Lang Eng 14(4):431–455

Berkovsky S, Baldwin T, Zukerman I (2008) Aspect-based personalized text summarization. In: Proceedings of the 5th international conference on adaptive hypermedia and adaptive web-based systems. pp 267–270

Biadsy F, Hirschberg J, Filatova E (2008) An unsupervised approach to biography production using Wikipedia. In: Proceedings of ACL-08: HLT. pp 807–815

Boguraev BK, Neff MS (2000) Discourse segmentation in aid of document summarization. In: Proceedings of the 33rd Hawaii international conference on system sciences, vol 3. p 3004

Bossard A, Généreux M, Poibeau T (2008) Description of the LIPN systems at TAC 2008: summarizing information and opinions. In: Proceedings of the text analysis conference (TAC)

Branny E (2007) Automatic summary evaluation based on text grammars. J Digit Inf 8(3). http://journals.tdl.org/jodi/article/viewArticle/232

Burges C, Shaked T, Renshaw E, Lazier A, Deeds M, Hamilton N, Hullender G (2005) Learning to rank using gradient descent. In: Proceedings of the 22nd international conference on Machine learning. pp 89–96

Carenini G, Cheung JCK (2008) Extractive vs. NLG-based abstractive summarization of evaluative text: the effect of corpus controversiality. In: Proceedings of the fifth international natural language generation conference, ACL 2008. pp 33–40

Cesarano C, Mazzeo A, Picariello A (2007) A system for summary-document similarity in notary domain. International Workshop on Database Expert Syst Appl:254–258

Ceylan H, Mihalcea R (2009) The decomposition of human-written book summaries. In: Proceedings of the 10th international conference on computational linguistics and intelligent text processing (CICLing ’09). pp 582–593

Cole R (ed) (1997) Survey of the state of the art in human language technology. Cambridge University Press, Cambridge

Conroy J, Schlesinger J (2008) CLASSY at TAC 2008 Metrics. In: Proceedings of the text analysis conference (TAC)

Conroy JM, Dang HT (2008) Mind the gap: dangers of divorcing evaluations of summary content from linguistic quality. In: Proceedings of the 22nd international conference on computational linguistics (Coling 2008). pp 145–152

Conroy JM, O’leary DP (2001) Text summarization via hidden Markov models. In: SIGIR ’01: proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval. pp 406–407

Conroy JM, Schlesinger JD, O’Leary DP (2009) CLASSY 2009: summarization and metrics. In: Proceedings of the text analysis conference (TAC)

Cristea D, Postolache O, Pistol I (2005) Summarisation through discourse structure. In: Proceedings of the computational linguistics and intelligent text processing, 6th International conference (CICLing 2005). pp 632–644

Cunha ID, Fernández S, Velázquez-Morales P, Vivaldi J, SanJuan E, Moreno JMT (2007) A new hybrid summarizer based on vector space model, statistical physics and linguistics. In: MICAI 2007: advances in artificial intelligence. pp 872–882

Dang HT (2006) Overview of DUC 2006. In: The document understanding workshop (presented at the *HLT/NA-ACL*). Brooklyn, New York, USA

Demner-Fushman D, Lin J (2006) Answer extraction, semantic clustering, and extractive summarization for clinical question answering. In: Proceedings of the 21st international conference on computational linguistics and the 44th annual meeting of the association for computational linguistics. pp 841–848

Deschacht K, Moens MF (2007) Text analysis for automatic image annotation. In: Proceedings of the 45th annual meeting of the association of computational linguistics. pp 1000–1007

Díaz A, Gervás P (2007) User-model based Personalized Summarization. Inf Process Manag 43(6):1715–1734

Donaway RL, Drummey KW, Mather LA (2000) A comparison of rankings produced by summarization evaluation measures. In: Proceedings of NAACL-ANLP 2000 workshop on automatic summarization. pp 69–78

Dunlavy DM, O’Leary DP, Conroy JM, Schlesinger JD (2007) QCS: A system for querying, clustering and summarizing documents. Inf Process Manag 43(6):1588–1605

Edmundson HP (1969) New methods in automatic extracting. In: Mani I, Maybury M (eds) Advances in automatic text summarization. pp 23–42

Elsner M, Charniak E (2008) Coreference-inspired coherence modeling. In: Proceedings of ACL-08: HLT, short papers. pp 41–44

Ercan G, Cicekli I (2008) Lexical cohesion based topic modeling for summarization. In: Proceedings of the 9th international conference in computational linguistics and intelligent text processing. pp 582–592

Erkan G, Radev DR (2004) LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. J Artif Intell Res (JAIR) 22:457–479

Fan J, Gao Y, Luo H, Keim DA, Li Z (2008) A novel approach to enable semantic and visual image summarization for exploratory image search. In: MIR '08: proceeding of the 1st ACM international conference on multimedia information retrieval. pp 358–365

Fellbaum C (1998) WordNet: an electronical lexical database. The MIT Press, Cambridge

Feng Y, Lapata M (2008) Automatic image annotation using auxiliary text information. In: Proceedings of ACL-08: HLT. pp 272–280

Filatova E, Hatzivassiloglou V (2004) Event-based extractive summarization. In: Marie-Francine Moens SS (ed) Text summarization branches out: proceedings of the ACL-04 workshop. pp 104–111

Fisher S, Dunlop A, Roark B, Chen Y, Burmeister J (2009) OHSU summarization and entity linking systems. In: Proceedings of the text analysis conference (TAC)

Fiszman M, Rindflesch TC, Kilicoglu H (2004) Abstraction summarization for managing the biomedical research literature. In: Moldovan D, Girju R (eds) HLT-NAACL 2004: workshop on computational lexical semantics. pp 76–83

Fuentes M, González E, Ferrés D, Rodríguez H (2005) QASUM-TALP at DUC 2005 automatically evaluated with a pyramid based metric. In: The document understanding workshop (presented at the *HLT/EMNLP* annual meeting)

Fuentes M, Alfonseca E, Rodríguez H (2007) Support vector machines for query-focused summarization trained and evaluated on pyramid data. In: Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions. pp 57–60

Fukushima T, Okumura M (2001) Text summarization challenge: text summarization evaluation at NTCIR workshop 2. In: Proceedings of the second NTCIR workshop meeting on evaluation of chinese and japanese text retrieval and text summarization. pp 9–13

Giannakopoulos G, Karkaletsis V (2009) N-GRAM GRAPHS: representing documents and document sets in summary system evaluation. In: Proceedings of the text analysis conference (TAC)

Giannakopoulos G, Karkaletsis V, Vouros G (2008a) Testing the use of n-gram graphs in summarization sub-tasks. In: Proceedings of the text analysis conference (TAC)

Giannakopoulos G, Karkaletsis V, Vouros G, Stamatopoulos P (2008) Summarization System Evaluation Revisited: N-gram graphs. ACM Trans Speech Lang Process 5(3):1–39

Goldstein J, Mittal V, Carbonell J, Kantrowitz M (2000) Multi-document summarization by sentence extraction. In: NAACL-ANLP 2000 workshop on automatic Summarization. pp. 40–48

Gonçalves PN, Rino L, Vieira R (2008) Summarizing and referring: towards cohesive extracts. In: DocEng '08: proceeding of the eighth ACM symposium on document engineering. pp 253–256

Gotti F, Lapalme G, Nerima L, Wehrli E (2007) GOFAISUM: a symbolic summarizer for DUC. In: The document understanding workshop (presented at the *HLT/NAACL*)

Grosz BJ, Weinstein S, Joshi AK (1995) Centering: A Framework for Modeling the Local Coherence of Discourse. Comput Linguist 21(2):203–225

Harabagiu S, Lacatusu F (2005) Topic themes for multi-document summarization. In: SIGIR '05: proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval. pp 202–209

Hasler L (2007) From extracts to abstracts: human summary production operations for computer-aided summarisation. In: Proceedings of the RANLP 2007 workshop on computer-aided language processing (CALP). pp 11–18

Hasler L (2008) Centering theory for evaluation of coherence in computer-aided summaries. In: (ELRA) ELRA (ed) Proceedings of the sixth international conference on language resources and evaluation (LREC'08)

Hassel M (2007) Resource lean and portable automatic text summarization. PhD thesis, Department of Numerical Analysis and Computer Science, Royal Institute of Technology

He L, Sanocki E, Gupta A, Grudin J (1999) Auto-summarization of audio-video presentations. In: MULTIMEDIA '99: proceedings of the seventh ACM international conference on multimedia (Part 1). pp 489–498

He T, Chen J, Gui Z, Li F (2008) CCNU at TAC 2008: proceeding on using semantic method for automated summarization. In: Proceedings of the text analysis conference (TAC)

Hearst MA (1997) TextTiling: segmenting text into multi-paragraph subtopic passages. Comput Linguist 23(1):33–64

Hirao T, Okumura M, Fukusima T, Nanba H (2005) Text summarization challenge 3—text summarization evaluation at NTCIR workshop 4. In: Proceedings of the fourth NTCIR workshop on research in information access technologies information retrieval, question answering and summarization. pp 407–411

Hovy E, Lin CY (1999) Automated multilingual text summarization and its evaluation. Technical report Information Sciences Institute, University of Southern California

Hovy E, Lin CY, Zhou L, Fukumoto J (2006) Automated summarization evaluation with basic elements. In: Proceedings of the 5th international conference on language resources and evaluation (LREC)

Jaoua M, Hamadou AB (2003) Automatic text summarization of scientific articles based on classification of extract's Population. In: Proceedings of computational linguistics and intelligent text processing, 4th international conference. pp 623–634

Jing H (2002) Using hidden Markov modeling to decompose human-written summaries. Comput Linguist 28(4):527–543

Jing H, McKeown KR (2000) Cut and paste based text summarization. In: Proceedings of the 1st North American chapter of the association for computational linguistics Conference. pp 178–185

Kaisser M, Hearst MA, Lowe JB (2008) Improving search results quality by customizing summary lengths. In: Proceedings of ACL-08: HLT. pp 701–709

Kan MY, Klavans JL (2002) Using librarian techniques in automatic text summarization for information retrieval. In: JCDL '02: proceedings of the 2nd ACM/IEEE-CS joint conference on digital libraries. pp 36–45

Kan MY, Klavans JL, Mckeown KR (2002) Using the annotated bibliography as a resource for indicative summarization. In: Proceedings of the language resources and evaluation conference. pp 1746–1752

Katragadda R (2010) GEMS: generative modeling for evaluation of summaries. In: Proceedings of the 11th international conference on computational linguistics and intelligent text processing, CICLing. pp 724–735

Kazantseva A (2006) An approach to summarizing short stories. In: Proceedings of the student research workshop at the 11th conference of the European chapter of the association for computational linguistics. pp 55–62

Ker SJ, Chen JN (2000) A text categorization based on summarization technique. In: Proceedings of the ACL-2000 workshop on recent advances in natural language processing and information retrieval. pp 79–83

Khan AU, Khan S, Mahmood W (2005) MRST: a new technique for information summarization. In: The second world enformatika conference, WEC'05. pp 249–252

Kumar C, Pingali P, Varma V (2008) Generating personalized summaries using publicly available web documents. In: Proceedings of the 2008 IEEE/WIC/ACM international conference on web intelligence and international conference on intelligent agent technology. pp 103–106

Kumar M, Das D, Agarwal S, Rudnicky A (2009) Non-textual event summarization by applying machine learning to template-based language generation. In: Proceedings of the 2009 workshop on language generation and summarisation (UCNLG + Sum 2009). pp 67–71

Kuo JJ, Chen HH (2008) Multidocument Summary Generation: Using Informative and Event Words. ACM Trans Asian Lang Inf Process (TALIP) 7(1):1–23

Kupiec J, Pedersen J, Chen F (1995) A trainable document summarizer. In: SIGIR '95: proceedings of the 18th annual international ACM SIGIR conference on research and development in information retrieval. pp 68–73

Lapata M, Barzilay R (2005) Automatic evaluation of text coherence: models and representations. In: Proceedings of the 19th international joint conference on artificial intelligence. pp 1085–1090

Lerman K, McDonald R (2009) Contrastive summarization: an experiment with consumer reviews. In: Proceedings of human language technologies: the 2009 annual conference of the North American chapter of the association for computational linguistics, companion volume: short papers. pp 113–116

Lerman K, Blair-Goldensohn S, McDonald R (2009) Sentiment summarization: evaluating and learning user preferences. In: Proceedings of the 12th conference of the European chapter of the ACL (EACL 2009). pp 514–522

Li S, Ouyang Y, Wang W, Sun B (2007) Multi-document summarization using support vector regression. In: The document understanding workshop (presented at the *HLT/NAACL*). Rochester, New York USA

Li S, Wan W, Wang C (2008) TAC 2008 update summarization task of ICL. In: Proceedings of the text analysis conference (TAC)

Li S, Wang W, Zhang Y (2009) Tac 2009 update summarization of icl. In: Proceedings of the text analysis conference (TAC)

Lin CY (2004) ROUGE: a package for automatic evaluation of summaries. In: Proceedings of ACL text summarization workshop. pp 74–81

Lin CY, Hovy E (2000) The automated acquisition of topic signatures for text summarization. In: Proceedings of the 18th conference on computational linguistics. pp 495–501

Liu F, Liu Y (2008) Correlation between ROUGE and human evaluation of extractive meeting summaries. In: Proceedings of ACL-08: HLT, short papers. pp 201–204

Liu M, Yu B, Fang F, Sun H (2009) TAC 2009 update summarization task of WUST. In: Proceedings of the text analysis conference (TAC)

Lloret E, Palomar M (2009) A gradual combination of features for building automatic summarisation systems. In: Proceedings of the 12th international conference on text, speech and dialogue (TSD). pp 16–23

Lloret E, Ferrández O, Muñoz R, Palomar M (2008) A text summarization approach under the influence of textual entailment. In: Proceedings of the 5th international workshop on natural language processing and cognitive science (NLPCS 2008). pp 22–31

Lloret E, Balahur A, Palomar M, Montoyo A (2009) Towards building a competitive opinion summarization system: challenges and keys. In: Proceedings of the NAACL. Student Research Workshop and Doctoral Consortium. pp 72–77

Lloret E, Saggion H, Palomar M (2010) Experiments on summary-based opinion classification. In: Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text. pp 107–115

Luhn HP (1958) The automatic creation of literature abstracts. In: Advances in automatic text summarization. pp 15–22

Mani I (2001) Automatic summarization. John Benjamins Publishing Co. Amsterdam, Philadelphia, USA

Mani I (2001b) Summarization evaluation: an overview. In: Proceedings of the North American chapter of the association for computational linguistics (NAACL). Workshop on Automatic Summarization

Mani I, Maybury MT (1999) Advances in automatic text summarization. The MIT Press, Cambridge

Mani I, House D, Klein G, Hirschman L, Firmin T, Sundheim B (1999) The TIPSTER SUMMAC text summarization evaluation. In: Proceedings of the ninth conference on European chapter of the association for computational linguistics. pp 77–85

Mani I, Klein G, House D, Hirschman L, Firmin T, Sundheim B (2002) SUMMAC: a text summarization evaluation. Nat Lang Eng 8(1):43–68

Mann WC, Thompson SA (1988) Rhetorical structure theory: Toward a functional theory of text organization. Text 8(3):243–281

Manning CD, Raghavan P, Schtze H (2008) Introduction to information retrieval. Cambridge University Press, New York, NY, USA

Marcu D (1999) Discourse trees are good indicators of importance in text. In: Advances in automatic text summarization. pp 123–136

McCargar V (2005) Statistical Approaches to Automatic Text Summarization. Bull Am Soc Inf Sci Technol 30(4):21–25

Medelyan O (2007) Computing lexical chains with graph clustering. In: Proceedings of the ACL 2007 student research workshop. pp 85–90

Mihalcea R (2004) Graph-based ranking algorithms for sentence extraction, applied to text summarization. In: Proceedings of the ACL 2004 on interactive poster and demonstration sessions. p 20

Mihalcea R, Ceylan H (2007) Explorations in automatic book summarization. In: Proceedings of the joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL). pp 380–389

Mille S, Wanner L (2008) Multilingual summarization in practice: the case of patent claims. In: Proceedings of the 12th European association of machine translation conference. pp 120–129

Minel JL, Nugier S, Piat G (1997) How to appreciate the quality of automatic text summarization? Examples of FAN and MLUCE protocols and their results on SERAPHIN. In: Proceedings of intelligent scalable text summarization workshop in conjunction with the European chapter of the association of computational linguistics (EACL). pp 25–30

Mitkov R, Evans R, Orasan C, Ha LA, Pekar V (2007) Anaphora resolution: to what extent does it help NLP applications? In: Proceedings of the 6th discourse anaphora and anaphor resolution colloquium. pp 179–190

Mohammad S, Dorr B, Egan M, Hassan A, Muthukrishan P, Qazvinian V, Radev D, Zajic D (2009) Using citations to generate surveys of scientific paradigms. In: Proceedings of human language technologies: The 2009 annual conference of the North American chapter of the association for computational linguistics. pp 584–592

Mori T (2002) Information gain ratio as term weight: the case of summarization of IR results. In: Proceedings of the 19th international conference on computational linguistics. pp 1–7

Mori T, Nozawa M, Asada Y (2004) Multi-answer-focused multi-document summarization using a question-answering engine. In: COLING ’04: proceedings of the 20th international conference on computational linguistics. pp 439–445

Mori T, Nozawa M, Asada Y (2005) Multi-answer-focused multi-document summarization using a question-answering engine. ACM Trans Asian Lang Inf Process (TALIP) 4(3):305–320

Morris AH, Kasper GM, Adams DA (1992) The Effect and Limitations of Automatic Text Condensing on Reading Comprehension Performance. Inf Syst Res 3(1):17–35

Nastase V, Milne D, Filippova K (2009) Summarizing with encyclopedic knowledge. In: Proceedings of the text analysis conference (TAC)

Nenkova A (2005) Automatic text summarization of newswire: lessons learned from the document understanding conference. In: Proceedings of the American association fro artificial intelligence (AAAI). pp 1436–1441

Nenkova A (2006) Summarization evaluation for text and speech: issues and Approaches. In: INTERSPEECH-2006, paper 2079-Wed1WeS.1

Nenkova A, Siddharthan A, McKeown K (2005) Automatically learning cognitive status for multi-document summarization of newswire. In: HLT '05: proceedings of the conference on human language technology and empirical methods in natural language processing. pp 241–248

Neto JL, Santos A, Kaestner CAA, Freitas AA (2000) Generating text summaries through the relative importance of topics. In: IBERAMIA-SBIA '00: proceedings of the international joint conference, 7th Ibero-American conference on AI. pp 300–309

Okumura M, Fukusima T, Nanba H, Hirao T (2004) Text Summarization Challenge 2 text summarization evaluation at NTCIR workshop 3. SIGIR Forum 38(1):29–38

Orăsan C (2004) The influence of personal pronouns for automatic summarisation of scientific articles. In: Proceedings of the discourse anaphora and anaphor resolution colloquium. pp 127–132

Orăsan C (2007) Pronominal anaphora resolution for text summarisation. In: Proceedings of the recent advances on natural language processing. pp 430–436

Orăsan C (2009) Comparative Evaluation of Term-Weighting Methods for Automatic Summarization. J Quant Linguist 16(1):67–95

Orăsan C, Pekar V, Hasler L (2004) A comparison of summarisation methods based on term specificity estimation. In: Proceedings of the fourth international conference on language resources and evaluation (LREC2004). pp 1037–1041. Available at:http://clg.wlv.ac.uk/papers/orasan-04a.pdf

Over P, Ligget W (2002) Introduction to DUC: an intrinsic evaluation of generic news text summarization systems. In: The document understanding workshop

Over P, Dang H, Harman D (2007) DUC in Context. Inf Process Manag 43(6):1506–1520

Owczarzak K (2009) DEPEVAL(summ): dependency-based evaluation for automatic summaries. In: Proceedings of the joint conference of the 47th annual meeting of the ACL and the 4th international joint conference on natural language processing of the AFNLP. pp 190–198

Pang B, Lee L (2005) Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales. In: Proceedings of the association of computational linguistics. pp 115–124

Pang B, Lee L (2008) Opinion Mining and Sentiment Analysis. Found Trends Inf Retr 2(1–2):1–135

Papineni K, Roukos S, Ward T, Zhu WJ (2002) BLEU: a method for automatic evaluation of machine translation. In: Proceedings of 40th annual meeting of the association for computational linguistics. pp 311–318

Passoneau RJ (2010) Formal and Functional Assessment of the Pyramid Method for Summary Content Evaluation. Nat Lang Eng 16(2):107–131

Pitler E, Nenkova A (2008) Revisiting readability: a unified framework for predicting text quality. In: Proceedings of the 2008 conference on empirical methods in natural language processing. pp 186–195

Plaza L, Díaz A, Gervás P (2008) Concept-graph based biomedical automatic Summarization Using Ontologies. In: Coling 2008: Proceedings of the 3rd textgraphs workshop on graph-based algorithms for natural language processing. pp 53–56

Plaza L, Lloret E, Aker A (2010) Improving automatic image captioning using text summarization techniques. In: Proceedings of the 13th international conference on text, speech and dialogue (TSD)

Qazvinian V, Radev DR (2008) Scientific paper summarization using citation summary networks. In: Proceedings of the 22nd international conference on computational linguistics (Coling 2008). pp 689–696

Radev DR, Fan W (2000) Automatic summarization of search engine hit lists. In: Proceedings of the ACL-2000 workshop on recent advances in natural language processing and information retrieval. pp 99–109

Radev DR, McKeown KR (1998) Generating Natural Language Summaries from Multiple on-line Sources. Comput Linguist 24(3):470–500

Radev DR, Tam D (2003) Summarization evaluation using relative utility. In: CIKM '03: proceedings of the 12th international conference on information and knowledge management. pp 508–511

Radev DR, Blair-Goldensohn S, Zhang Z (2001) Experiments in single and multi-document summarization using MEAD. In: First document understanding conference. pp 1–7

Radev DR, Hovy E, McKeown K (2002) Introduction to the Special Issue on Summarization. Comput Linguist 28(4):399–408

Saggion H (2008) Automatic summarization: an overview. Revue franaise de linguistique appliquée XIII(1). pp 63–81

Saggion H (2009) A classification algorithm for predicting the structure of summaries. In: Proceedings of the 2009 workshop on language generation and summarisation (UCNLG+Sum 2009). pp 31–38

Saggion H, Funk A (2009) Extracting Opinions and Facts for Business Intelligence. RNTI E-17:119–146

Saggion H, Lapalme G (2000) Selective analysis for automatic abstracting: evaluating indicativeness and acceptability. In: Proceedings of content-based multimedia information access (RIAO). pp 747–764

Saggion H, Lloret E, Palomar M (2010) Using text summaries for predicting rating scales. In: Proceedings of the 1st workshop on computational approaches to subjectivity and sentiment analysis (WASSA)

Sakai T, Sparck-Jones K (2001) Generic summaries for indexing in information retrieval. In: SIGIR '01: proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval. pp 190–198

Saravanan M, Ravindran B, Raman S (2006) Improving legal document summarization using graphical models. In: Proceedings of legal knowledge and information systems—JURIX 2006: the 19th annual conference on legal knowledge and information systems. pp 51–60

Sauper C, Barzilay R (2009) Automatically generating wikipedia articles: a structure-aware approach. In: Proceedings of the joint conference of the 47th annual meeting of the ACL and the 4th international joint conference on natural language processing of the AFNLP. pp 208–216

Schilder F, Kondadadi R (2008) FastSum: fast and accurate query-based multi-document summarization. In: Proceedings of ACL-08: HLT, short papers. pp 205–208

Schilder F, Kondadadi R, Leidner JL, Conrad JG (2008) Thomson reuters at TAC 2008: aggressive filtering with FastSum for update and opinion summarization. In: Proceedings of the text analysis conference (TAC)

Schlesinger JD, Okurowski ME, Conroy JM, O'Leary DP, Taylor A, Hobbs J, Wilson H (2002) Understanding machine performance in the context of human performance for multi-document summarization. In: Proceedings of the DUC 2002 workshop on text summarization

Sebastiani F (2002) Machine Learning in Automated Text Categorization. ACM Comput Surv 34(1):1–47

Shen D, Chen Z, Yang Q, Zeng HJ, Zhang B, Lu Y, Ma WY (2004) Web-page classification through summarization. In: SIGIR '04: proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval. pp 242–249

Shen D, Yang Q, Chen Z (2007) Noise Reduction through Summarization for Web-page Classification. Inf Process Manag 43(6):1735–1747

Shi Z, Melli G, Wang Y, Liu Y, Gu B, Kashani MM, Sarkar A, Popowich F (2007) Question answering summarization of multiple biomedical documents. In: CAI '07: proceedings of the 20th conference of the Canadian society for computational studies of intelligence on advances in artificial intelligence. pp 284–295

Sjöbergh J (2007) Older Versions of the ROUGEeval Summarization Evaluation System were Easier to Fool. Inf Process Manag 43(6):1500–1505

Spärck Jones K (1999) Automatic summarizing: factors and directions. In: Advances in automatic text summarization. pp 1–14

Spärck Jones K (2007) Automatic Summarising: The State of the Art. Inf Process Manag 43(6):1449–1481

Spärck-Jones K, Galliers JR (eds) (1996) Evaluating natural language processing systems, an analysis and review, lecture notes in computer science, vol 1083. Springer, Berlin

Steinberger J, Poesio M, Kabadjov MA, Ježek K (2007) Two Uses of Anaphora Resolution in Summarization. Inf Process Manag 43(6):1663–1680

Steinberger J, Jezek K, Sloup M (2008) Web topic summarization. In: Proceedings of the 12th international conference on electronic publishing. pp 322–334

Strzalkowski T, Harabagiu S (2007) Advances in open domain question answering (Text, Speech and Language Technology). Springer-Verlag New York, Inc., Secaucus, NJ, USA

Sun JT, Shen D, Zeng HJ, Yang Q, Lu Y, Chen Z (2005) Web-page summarization using clickthrough data. In: SIGIR '05: proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval. pp 194–201

Svore KM, Vanderwende L, Burges CJ (2007) Enhancing single-document summarization by combining RankNet and third-party sources. In: Proceedings of the joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL). pp 448–457

Sweeney S, Crestani F, Losada DE (2008) Show me more: Incremental length summarisation using novelty detection. Inf Process Manag 44(2):663–686

Szlávik Z, Tombros A, Lalmas M (2006) Investigating the use of summarisation for interactive XML retrieval. In: SAC '06: Proceedings of the 2006 ACM symposium on applied computing. pp 1068–1072

Teng Z, Liu Y, Ren F, Tsuchiya S, Ren F (2008) Single document summarization based on local topic identification and word frequency. In: MICAI '08: proceedings of the 2008 seventh Mexican international conference on artificial intelligence. pp 37–41. http://dx.doi.org/10.1109/MICAI.2008.12

Teufel S, Halteren Hv (2004) Evaluating information content by factoid analysis: human annotation and stability. In: Proceedings of the 2004 conference on empirical methods in natural language processing. pp 419–426

Teufel S, Moens M (2002) Summarizing Scientific Articles: Experiments with Relevance and Rhetorical Status.. Comput Linguist 28(4):409–445

Titov I, McDonald R (2008) A joint model of text and aspect ratings for sentiment summarization. In: Proceedings of ACL-08: HLT. pp 308–316

Torres-Moreno JM, St-Onge PL, Gagnon M, El-Bze M, Bellot P (2009) Automatic summarization system coupled with a question-answering system (QAAS). NLP News Computing Language. http://arxiv.org/abs/0905.2990v1

Trappey A, Trappey C, Wu CY (2009) Automatic patent Document Summarization for Collaborative Knowledge Systems and Services. J Syst Sci Syst Eng 18(1):71–94

Trappey AJC, Trappey CV (2008) An R&D Knowledge Management Method for Patent Document Summarization. Ind Manag Data Syst 108(2):245–257

Tseng YH, Lin CJ, Lin YI (2007) Text Mining Techniques for Patent Analysis. Inf Process Manag 43(5):1216–1247

Vadlapudi R, Katragadda R (2010a) On automated evaluation of readability of summaries: capturing grammaticality, focus, structure and coherence. In: Proceedings of the NAACL HLT 2010 student research workshop. pp 7–12

Vadlapudi R, Katragadda R (2010b) Quantitative evaluation of grammaticality of summaries. In: Proceedings of the 11th international conference on computational linguistics and intelligent text processing, CICLing. pp 736–747

Van Dijk TA (1972) Some aspects of text grammars. *A study in Theoretical Linguistics and Poetics*, La Haya-parís, Mouton

Van Rijsbergen CJ (1981) Information retrieval. Elsevier, Amsterdam

Wan X, Yang J, Xiao J (2007) Towards a unified approach based on affinity graph to various multi-document summarizations. In: Proceedings of the 11th European conference. pp 297–308

Wang C, Long L, Li L (2008) HowNet based evaluation for Chinese text summarization. In: Proceedings of the international conference on natural language processing and software engineering. pp 82–87

Wilson T, Wiebe J, Hoffmann P (2005) Recognizing contextual polarity in phrase-level sentiment analysis. In: Proceedings of the empirical methods in natural language processing. pp 347–354

Witte R, Krestel R, Bergler S (2007) Generating update summaries for DUC 2007. In: The document understanding workshop (presented at the *HLT/NAACL*)

Wong KF, Wu M, Li W (2008) Extractive summarization using supervised and semi-supervised learning. In: Proceedings of the 22nd international conference on computational linguistics (Coling 2008). pp 985–992

Yu J, Reiter E, Hunter J, Mellish C (2007) Choosing the Content of Textual Summaries of Large Time-series Data Sets. Nat Lang Eng 13(1):25–49

Zajic D, Dorr BJ, Lin J, Schwartz R (2007) Multi-candidate reduction: Sentence compression as a tool for document summarization tasks. Inf Process Manag 43(6):1549–1570

Zajic DM, Dorr BJ, Lin J (2008) Single-document and multi-document summarization techniques for email threads using sentence compression. Inf Process Manag 44(4): 1600–1610

Zechner K, Waibel A (2000) DiaSumm: flexible summarization of spontaneous dialogues in unrestricted domains. In: Proceedings of the 18th conference on computational linguistics. pp 968–974

Zhou L, Ticrea M, Hovy E (2004) Multi-document biography summarization. In: Proceedings of the conference on empirical methods in natural language processing. pp 434–441

Zhou L, Lin CY, Munteanu DS, Hovy E (2006) ParaEval: using paraphrases to evaluate summaries automatically. In: Proceedings of the human language technology/North American association of computational linguistics conference. pp 447–454

Zhuang L, Jing F, Zhu XY (2006) Movie review mining and summarization. In: CIKM '06: proceedings of the 15th ACM international conference on information and knowledge management. pp 43–50