Lang Resources & Evaluation (2018) 52:101–148
https://doi.org/10.1007/s10579-017-9399-2

ORIGINAL PAPER

The challenging task of summary evaluation: an overview

Elena Lloret[1] · Laura Plaza[2] · Ahmet Aker[3]

Published online: 2 September 2017

Abstract Evaluation is crucial in the research and development of automatic summarization applications, in order to determine the appropriateness of a summary based on different criteria, such as the content it contains, and the way it is presented. To perform an adequate evaluation is of great relevance to ensure that automatic summaries can be useful for the context and/or application they are generated for. To this end, researchers must be aware of the evaluation metrics, approaches, and datasets that are available, in order to decide which of them would be the most suitable to use, or to be able to propose new ones, overcoming the possible limitations that existing methods may present. In this article, a critical and historical analysis of evaluation metrics, methods, and datasets for automatic summarization systems is presented, where the strengths and weaknesses of evaluation efforts are discussed and the major challenges to solve are identified. Therefore, a clear up-to-date overview of the evolution and progress of summarization evaluation is provided, giving the reader useful insights into the past, present and latest trends in the automatic evaluation of summaries.

Keywords Text summarization · Evaluation · Content evaluation · Readability · Task-based evaluation

✉ Ahmet Aker
aker@is.inf.uni-due.de

Elena Lloret
elloret@dlsi.ua.es

Laura Plaza
lplaza@lsi.uned.es

[1] Universidad de Alicante, Alicante, Spain

[2] IR & NLP UNED, Madrid, Spain

[3] University of Duisburg-Essen, Duisburg, Germany

1 Introduction

Evaluation is an important part of any research task, since it can determine to what extent the investigated approaches are appropriate, allowing also the comparison between them. For some tasks dealing with the production of language, such as text summarization, machine translation or natural language generation, the evaluation is extremely complex, since there may not exist a unique possible output. Instead, we can find a high number of equivalent answers/outputs that also depend on pragmatic issues, such as the purpose of the text to be generated, the context in which a sentence/text will be used, the background of the person that is going to use that text, etc. Moreover, the concept of goodness in these tasks is fuzzy, since it may depend on several criteria and can vary among different assessors. Thus, it is crucial to design and investigate suitable methods that can be adapted to the task and to the type of text to be produced.

In the context of text summarization, there is a great variety of summary categorizations according to different issues (Sparck Jones and Galliers 1996; Lloret and Palomar 2012). The traditional and most common ones include the type of input and output (e.g., single- or multi-document summaries; extractive or abstractive summaries; headlines). In any case, automatically generated summaries have to be evaluated in order to assess the quality of the systems used for their generation. Sparck Jones and Galliers (1996) distinguish between *intrinsic* and *extrinsic* evaluation methods. Intrinsic evaluation assesses the coherence and the informativeness of a summary, whereas extrinsic evaluation assesses the utility of summaries in a given application context, for example, relevance assessment, reading comprehension, etc. The main challenge to be addressed in evaluation is the definition and use of a good metric, able to capture whether the summary is good enough. But the concept of good is very subjective and depends on a great number of issues, so existing metrics may be not suitable for all types of summaries. For instance, Sparck Jones and Galliers (1996) also identified the purpose of the summary as a criterion that should be taken into account when generating it. In this respect, a summary can be *indicative*, *informative*, *generic*, or *topic-oriented*, among others. Moreover, depending on the reader, a summary would fit or not within his/her needs. Therefore, the main challenge involves the subjectivity associated to the evaluation, which is very difficult to capture with automatic metrics. Most of the metrics proposed in the literature focused on intrinsic evaluation; however, their limitations have been discussed along the years and there would not be a perfect metric, having to complement the evaluation with manual assessment in most of the cases, which makes it necessary to continue carrying out research on the evaluation task.

In intrinsic evaluation of automatic summarization it is common to distinguish *reference* or *model* summaries from automated or *peer* summaries. Reference or model summaries are those summaries that will be considered correct, and normally refer to those summaries generated manually by humans. Peer summaries are the summaries to be evaluated that usually have been automatically produced.

While there has been a wide range of work on summary evaluation in the past two decades, the most influential work has been carried out within the challenges under the Document Understanding Conferences (DUC)[1] and the Text Analysis Conferences (TAC)[2], organized by the National Institute of Standards and Technology (NIST) in the US. Both conferences have contributed to the dissemination of recent results, definition of tasks and evaluation setups which have focused research investigations towards new directions in text summarization. These challenges have always included an intrinsic summarization evaluation. While the details of the tasks and the evaluation procedures have changed over time, there have always been several main components or criteria to be measured in the intrinsic evaluation. These concern: (i) the linguistic quality or *readability* of the peer summary; (ii) *informativeness* or the *content coverage* of the summary, in relation to an information need or topic that may have been expressed in advance; and (iii) the *non-redundancy* of the summary produced.

To assess readability, the peer summary is evaluated, for example, on how coherent it is, i.e. the summary is checked to see if it contains dangling anaphora or gaps in its rhetorical structure (Mani 2001). Assessing the readability of a summary is done manually. Humans are asked to assess various aspects of the readability of a peer summary by answering questions in terms of a five point scale. The scores for the peer summaries are compared to those for the reference summaries, which are assessed in the same way as the peer summaries (Mani 2001; Dang 2005, 2006). Such manual assessment of readability is labor intensive and thus expensive to conduct and difficult to repeat. This is the reason why, even though important progress is being achieved, there is still no standard method for assessing the coherence of summaries that is widely used by the research community.

To assess the summary's content a variety of approaches have been adopted within the document summarization community. Most of them revolve around the comparison of the peer summary with one or more reference summaries (using more than one reference summary helps to overcome the subjectiveness inherent in using a single reference produced by a single human summarizer). In essence, these approaches are variants of two broad types. In one, the reference summaries are analyzed into semantic chunks, roughly equivalent to simple propositions and variously called "elementary discourse units", "model units", "summary content units" or "factoids" (Teufel and van Halteren 2004), which are used to assess summary informativeness on identifying the aforementioned units in the reference summaries and determining the extent to which these are present in the peer summary. In these approaches, human judgements are made about the overlap between peer summaries and reference summaries in terms of the proportion of reference units found in peer summary. This type of evaluation was carried out in DUC for the first few years against a single reference summary. As the inadequacies of comparing against a single reference summary became apparent, the method was elaborated by a number of groups, the most popular being the Pyramid approach of Nenkova and Passonneau (2004), Nenkova et al. (2007). In the other type of

[1] http://www-nlpir.nist.gov/projects/duc/index.html.

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

approach, various forms of n-gram overlap between the peer and reference summaries are automatically computed and the peer is given a score that reflects its recall of reference n-grams. In this context, the most popular method is ROUGE (Lin 2004a), which will be explained in the next section. However, there is also an ongoing-effort in which interesting evaluation methods without relying on model summaries are also proposed (Cabrera-Diego et al. 2016; Ellouze et al. 2016) that could complement the already existing ones (Torres-Moreno et al. 2010b, a).

Regardless of the type of summary evaluation conducted (either manual or automatic), this task is highly difficult due to the subjectivity involved (Fiori 2014): first, because of the lack of agreement on the quality criteria that a summary must fulfill; second, because of the subjectivity of assessing the summarization criteria (the agreement between human evaluators has been reported to be quite low (Sparck Jones and Galliers 1996)); and third, because of the amount of effort required to evaluate the summaries (i.e., it is very time-consuming).

“De facto” standard evaluation of automatic summaries follows the guidelines of DUC and TAC evaluations. This mainly includes automatically computing ROUGE-2 and ROUGE-SU4 on four model summaries for evaluation of summary content coverage. A few number of works also include the manual computation of Pyramid scores (Nenkova et al. 2007) and overall responsiveness and readability (based on a set of questions to be answered by human judges), as it is done in the most recent editions of the TAC evaluations (Owczarzak and Dang 2011). Performance of automatic systems is compared to those of different baselines, such as using leading sentences from the document/more recent document or using publicly available summarizers such as MEAD (Radev 2001).

Moreover, in order to compare and evaluate different metrics in their ability and effectiveness to predict human judgements, as well as to better discern between human and peer summaries, statistical research works have been also conducted, taking advantage of the data and participant systems in evaluation fora. In Owczarzak et al. (2012a), a thorough assessment of automatic evaluation in summarization of news is carried out. Using methods introduced in Rankel et al. (2011), the authors aim to identify the best variant of ROUGE on several TAC editions, finding out that ROUGE 1 and 2 appear to best emulate human pyramid and responsiveness scores on four years of NIST evaluations (TAC data from 2008 to 2011). In Rankel et al. (2013) the authors focus on the ability of ROUGE to predict significant differences between top performing systems. They also reassess different ROUGE variants using the same datasets as in Owczarzak et al. (2012a), but this time to determine statistical significance between systems.

Although research in summarization evaluation is gaining increasing attention, more effort is needed in order to advance this complicated task. On the one hand, automatic metrics are usually employed to evaluate the quality of automatic summaries, but they mainly measure informativeness (Tratz and Hovy 2008). On the other hand, research in the automatic evaluation of readability is still very preliminary (Pitler and Nenkova 2008). However, thanks to the novel AESOP task organized within TAC competitions, great progresses have been achieved in the last years, providing a wide set of metrics and baselines to continue working in this text summarization subarea. Despite the increasing number of metrics and methods available for evaluating the content and readability of summaries, one can have difficulties in identifying and analyzing all of them, since they have been developed in different years and for different purposes.

Several survey articles, chapters and books dealing with the topic of summary evaluation have been already published (Torres-Moreno 2011; Nenkova and McKeown 2011; Torres-Moreno 2014; Gambhir and Gupta 2017), where different algorithms for automatic summarization evaluation are examined. Specifically in Gambhir and Gupta (2017) detailed information about the performance of different summarization systems that were evaluated with ROUGE participating the DUC/TAC conferences along several years is provided. However, we believe that the work in this paper is still necessary since it covers important gaps of previous ones. With our survey research work we aim to compile all the previous existing work done with regards to the evaluation of summaries, paying also attention to manual evaluation and crowdsourcing, which was the latest strategy of collaborative work for carrying out the evaluation of automatic summaries. Moreover, we present an extensive compilation of the corpora and datasets that are available for the research community to work within the summarization evaluation area.

This article is organized as follows. Section 2 presents a comprehensive list of metrics and systems for automatically evaluating summaries according to both content and readability. Section 3 reviews the different criteria and methodologies proposed so far to manually evaluate summaries. Section 4 describes the most popular corpora and datasets that have been specifically developed for summarization evaluation. Section 5 highlights the importance of assessing the evaluation methods and the most commonly adopted approaches. Section 6 describes the evaluation of automatic summaries in the context of a particular task, i.e. the aim is to measure how much help the summary provides for a human performing this particular task. In Sect. 7 we summarize the major events or conferences hosted for the purpose of summarization evaluation. Finally, Sect. 8 summarizes the main conclusions of this work.

2 Automatic evaluation

In automatic evaluation, summaries generated by automatic summarization systems are assessed by automated metrics, thus significantly reducing the evaluation cost. However, most of these metrics still need of some human effort since they rely on the comparison of automatic summaries with one or more human-made model summaries (either in the form of an abstractive/extractive summary or a ranking of sentences). As we will show in the following subsections, most of the automatic metrics proposed so far address the summary content evaluation, being the development of automatic metrics for readability evaluation still very preliminary.

2.1 Summary content

In summary content evaluation, the automated summaries are assessed based on how much relevant information they contain from the original document(s). In theory, a summary can contain all relevant information from the original document(s) when it is long enough, e.g., almost identical in length to the input document(s). Therefore an automatic summary must, given a length constraint, contain only the relevant information from the inputs which are most important (Hovy 2005). Related work has investigated various metrics to determine whether automatic summaries fulfill this requirement.

Earlier studies in text summarization adapted metrics from information retrieval (IR) such as **recall**, **precision** and **F-measure** as shown in the following equations to assess the content of the automatic summaries[3] (Edmundson 1969; Paice 1990; Kupiec et al. 1995; Marcu 1997; Salton et al. 1997; Ono et al. 1994; Donaway et al. 2000).

$$Recall = \frac{Human\ selected\ sentences \cap machine\ selected\ sentences}{Human\ selected\ sentences} \tag{1}$$

$$Precision = \frac{Human\ selected\ sentences \cap machine\ selected\ sentences}{Machine\ selected\ sentences} \tag{2}$$

$$F - measure = 2 * \frac{Recall * Precision}{Recall + Precision} \tag{3}$$

In these ways, an automatic summary (peer summary) is compared to a human-written one (reference model), and the common sentences between them are measured. Following Nenkova (2006), recall evaluates which portion of the sentences selected by a human are also identified by a summarization system, whereas precision is the fraction of these sentences identified by the summarization system that are correct. F-measure is the harmonic mean between precision and recall.

However, the main problem with such an approach is that comparing the system outputs to a single human-written summary is too subjective. Other sentences within the source document(s) may be of same relevance than the ones included in the single model summary; but, since such sentences are not included in the model summary, peer summaries containing such sentences will be scored low although they are as good as the automatic summaries containing sentences from the model summary. For instance, if the model summary contains sentences {1 2} and is compared to two automated summaries containing sentences {1 2} and {1 3}, an automated summary containing exactly the same sentences will be scored higher than {1 3}. Nevertheless, it can happen that the sentences 2 and 3 are both equally relevant but it happened that the sentence 2 was selected by the human instead of sentence 3. To overcome this problem, Jing et al. (1998) propose to use multiple human-written summaries generated by different human subjects and construct an "ideal" summary from these multiple model summaries. An ideal summary is constructed by taking the majority opinion from the multiple summaries. For instance, if five human subjects are used to generate model summaries, a sentence

[3] The main reason for the use of IR oriented metrics was that the earlier summarization systems were mainly extractive and it was enough to judge them based on how much sentences identified by humans as relevant they retrieved.

selected by three or more humans is regarded as the majority opinion and taken to the ideal summary; otherwise not. According to Jing et al. (1998) the recall and precision metrics can still fail to judge system summary fairly using the ideal summary. This can happen, for instance, when the sentences 2 and 3 from the example above are favoured almost identically by the five humans (2 was selected by three humans and 3 by only two). Based on the majority voting, the sentence 2 is taken to the ideal summary but sentence 3 is excluded. Assuming sentence 1 is already in the ideal summary, if the sentence 2 is included it will have {1 2} again. With this setting, the same problem occurs as with only one human constructed model summary case.

Radev and Tam (2003) proposed a metric called **relative utility** to overcome the shortcoming of one human constructed or the ideal model summary. This method allows multiple judges to rank each sentence in the source document with a score, giving it a value ranging from 0 to 10, which determines its suitability for a summary. The higher the relative utility number is, the more relevance is given to the sentence. Therefore, summaries containing different sentences with the same relative utility weights are considered equally good. Only summaries containing sentences with higher relative utility scores are better or scored higher than summaries with less higher relative utility weights.

However, assigning relative utility scores to every sentence in the input document is tediOther methods have been also investigated with the aimous. It gets more tedious when the input consists of multiple documents. To avoid this very labour intensive task there was a shift from full sentence comparison to comparison of smaller units within the sentence. In this process, sequences of words from the model summaries are extracted and compared to sequences of words taken from the automatic summaries. The intuition behind this is that syntactically different sentences (i.e. sentences which are not exactly the same when compared using surface level exact matching) can still carry units (e.g. a sequence of some words) which are identical in all of them and they may also make the syntactically different sentences as equally relevant. Furthermore, if we look at the summary level two summaries which are similar in meaning then they must share similar sequences of words. When those sequences are compared to each other, one can capture their similarity. For determining the quality of an automated summary based on sequences of words the most popular system is called **ROUGE** (Lin 2004b). Its name stands for *Recall-Oriented Understudy for Gisting Evaluation* and it was inspired by BLEU (Papineni et al. 2002), which is a method for automatically evaluating the output of a machine translation system.

ROUGE compares automatically generated summaries against several model reference summaries (i.e. human-created). In this way, it estimates the coverage of appropriate concepts in an automatically generated summary. Several ROUGE metrics can be calculated. ROUGE-1 to ROUGE-4 (shown in Eq. 4) give recall scores from uni-gram (a single word) to four-gram (four contiguous words) overlap between the automatically generated summaries and the reference summaries. Sequences of overlapping words that do not immediately follow each other are captured by ROUGE-L. In ROUGE-L gaps in word sequences are ignored so that, for instance, *A B C D G* and *A E B F C K D* are counted as having the common sequence $A\ B\ C\ D$. ROUGE-W allows the longest common sub-sequences to be controlled/weighted. ROUGE-SU4 allows bi-grams to consist of non-contiguous words, with a maximum of four words between two words in the bi-grams.

$$ROUGE-N = \frac{\sum\limits_{S\in\{GoldStandard\ Summaries\}} \sum\limits_{gram_n \in S} Count_{match}(gram_n)}{\sum\limits_{S\in\{GoldStandard\ Summaries\}} \sum\limits_{gram_n \in S} Count(gram_n)} \tag{4}$$

where n is the length of n-gram, $gram_n$; $Count_{match}(gram_n)$ is the maximum number of n-gram co-occurring in system summary and a set of gold standard summary.

ROUGE was used in DUC conferences starting in 2004 to assess the quality of single and multi-document summarization systems and is now used in TAC conferences. However, although ROUGE is the *de facto* evaluation system for automatically generated summaries, it has been criticized because it only performs string matching between the summaries and does not take the meaning expressed in single words or sequences of words (n-grams) into consideration. In Sjöbergh (2007), it was shown that a very poor summary could easily get high ROUGE scores. In order to prove this claim, a simple summarization method was developed, using a greedy word selection strategy. Although the generated summaries were not good from a human's point of view, they obtained good results for some ROUGE metrics (for example, a recall score of 41% for ROUGE-1, which is acceptable in the state-of-the-art in this research field). In addition, the correlation between ROUGE and model summaries was shown to be lower than it was claimed, especially in some summarization types, such as in speech summarization (Liu and Liu 2008b). Despite the need to have model summaries beforehand when using ROUGE, various researchers have shown that there is significant correlation between ROUGE scores and approaches based on human comparison of semantic content units (indeed this was necessary for ROUGE to win acceptance). It has been reported that ROUGE correlates highly with human judgments on DUC 2001–2003 data (Lin 2004b) on system level. Depending on the data and the ROUGE metric used, the correlations varied between 0.49 and 0.90. Louis and Nenkova (2008) and Passonneau et al. (2005) also report that ROUGE correlates highly (around 90%), with Pyramid as well as responsiveness scores (around 90%) indicating that ROUGE is a low cost choice for obtaining similar results as manual evaluations. More details about the pyramid method as well as responsiveness evaluations are given below in Sect. 3.

Similarly, Hong et al. (2014) showed that state-of-the-art systems get similar average ROUGE scores but produce very different summaries, which evidences that more sensitive evaluation measures are needed. In the same line, Schluter (2017) shows that, according to ROUGE, there has been no substantial improvement in performance of summarization systems in the last decade, stressing the fact that perfect scores of ROUGE are impossible to reach even for humans.

However, there have been also studies aiming to address the drawbacks of ROUGE. For instance, in order to overcome with the difficulty of obtaining a set of model summaries, He et al. (2008) suggest an alternative method based on ROUGE

(**ROUGE-C**) that allows to evaluate a summary comparing it directly to the source document, given that some query-focused information is also provided. In ROUGE-C the peer summaries are treated as model ones and the original document(s) as peer summaries, and standard ROUGE is run over this setting. As other evaluation tools that do not rely on reference summaries, ROUGE-C avoids the need to have model summaries, that are difficult and time-consuming to obtain. Moreover, it correlates well with methods that depend on human summaries, so this also validates the method, and proves it usefulness for the research community.

To address the "meaning" problem, evaluation methods which rely on dependency parsing for representing the information in peer and model summaries haven been proposed. **Basic Elements** (BE) (Hovy et al. 2006) is such an evaluation methodology. The underlying idea of this method is to split a sentence into very small units of content in order to allow greater flexibility for matching different equivalent expressions. The small units are called basic elements and are defined as triplets of words consisting of a head, a modifier or argument, and the relationship between both (head–modifier–relation). An improved version of this evaluation tool was later developed in Tratz and Hovy (2008). It was called **Basic Elements with Transformations for Evaluation** (BEwT-E) and its philosophy was the same as for BE. However, whereas BE used a predefined and static list of paraphrases for matching equivalent expressions, BEw-T-E automates this stage of the process proposing a set of rules capable of identifying abbreviations, prepositional phrases, nominalizations or synonyms, among others. The main drawback of this method concerns the use of several language-dependent preprocessing modules for parsing and cutting the sentences. As a consequence, parser resources in other languages rather than English would be a requirement for using it when summaries in different languages have to be evaluated. The BEwT-E was one of the strongest performers among the systems that participated in the TAC 2009 AESOP track, achieving the best performance according to the Spearman metric when evaluated on the TAC 2009 update summaries and showing a high correlation with overall responsiveness and modified Pyramid score. **DEPEVAL(summ)** (Owczarzak 2009) is also a dependency-based metric. The idea here is similar to BE, and similarly, it compares dependency triples extracted from automatic summaries against the ones from model summaries. The main difference with BE is that a different parser is employed. Whereas BE uses Minipar[4], DEPEVAL(summ) is tested with the Charniak parser.[5]

Other methods have been also investigated with the aim of addressing what we have named the "meaning" problem. Zhou et al. (2006), for instance, applies the idea of paraphrases to capture the common meaning between peer and model summaries. The authors obtain the paraphrases through the MOSES statistical machine translation tool kit (Koehn et al. 2007), which produces a phrase table using parallel data. In this phrase table, source language (e.g. English) phrases (including single word to multi-word phrases) are aligned with some probability to target language (e.g. Chinese) phrases. Zhou et al. (2006) group

[4] http://ai.stanford.edu/~rion/parsing/minipar_viz.html.

[5] https://github.com/BLLIP/bllip-parser.

English phrases together which are assigned to the same Chinese phrase. Every group of phrases are treated as paragraphs. After this step, a three-tier comparison between an automated and a model summary is performed. In the first tier, the aim is to find multi-word paraphrases from the model summary which are also found in the automated summary. In the second tier, single word paraphrases (synonym words) are determined from the remaining text of the model summary. Finally, in the third tier, words which have simple lexical matches are collected. The summary is then assigned a score which is the ratio of matched model summary words to the total number of words in the model summary.

Steinberger et al. (2009) propose a content-based metric that measures the amount of content shared between a pair of texts (e.g., summaries) on the basis of the average semantic similarity between the set of concepts within the first (model) text and the set of concepts within the second text that is formally defined as follows:

$$avg\_sim(C_m, C_s) = \sum_{c_m in C_m, c_s in C_s} \frac{^{max} c_m, c_s[sim(c_m, c_s)]}{|C_m|} \tag{5}$$

where C_m is the set of concepts contained in the model summary, C_s is the set of concepts within the system summary and $sim(c_m, c_s)$ is Resnik's semantic similarity measure using a taxonomy (Resnik 1995). This content-based metric is next combined with other features such as unigram and bigram recall using a weighted linear combination. In a further experiment, the authors expanded all taxonomy concepts found in each summary (model and peer) with all their IS-A ancestors and used a named-entity disambiguator and a geo tagger to identify and disambiguate persons, organizations and geographic places. All this information is combined using a linear function as follows:

$$Sc = \alpha \times uni(M; Sys) + \beta \times bi(M; Sys) + \gamma \times cpts(M; Sys) + \delta \times ents(M; Sys) \tag{6}$$

where Sc is the score, $uni(M; Sys)$ is the recall of unigrams, $bi(M; Sys)$ is that of bigrams, $cpts(M; Sys)$ is the recall of taxonomy concepts and $ents(M; Sys)$ is the recall of entities.

In the recent years, the emergence of new types of text representation, such as word, sentence, paragraph or document embeddings has allowed the improvement and adaptation of similarity scoring methods, such as ROUGE. For instance, in Ng and Abrecht (2015a), ROUGE-WE is proposed as a variant of ROUGE that uses pre-trained word embeddings (in this case word2vec[6] implementation). The main focus is that, instead of measuring lexical overlaps, as ROUGE traditionally does, word embeddings are integrated in the approach, so that the semantic similarity of the words used in summaries and peers can be computed instead. The use of embeddings has been shown to be useful in different natural language processing tasks (Collobert et al. 2011) and, in particular, for semantic similarity since this type

[6] https://www.tensorflow.org/tutorials/word2vec.

of representation is more accurate and it can take into account the context in which the text appears (Yin and Schütze 2015). The results showed good correlations with human judgements (i.e., Pyramid scores) according to different metrics (e.g., the Pearson, Spearman and Kendall rank coefficient) achieving the best correlations for ROUGE-WE1 for Spearman (0.9138) and Kendall (0.7534).

In contrast to the previous methods, the **GEMS** (Generative Modeling for Evaluation of Summaries) approach of Katragadda (2010) suggests the use of signatures terms to analyze how they are captured in peer summaries. Signature terms (also known as topic signatures) are word vectors related to a particular topic. They are calculated on the basis of part-of-speech tags, such as nouns or verbs; query terms and terms of model summaries. The distribution of the signature terms is computed first in the source document and then the likelihood of a summary being biased towards such signature terms is obtained to determine how informative the peer summary is. The main difficulty associated to this approach is to have lists of signature terms belonging to a topic that could serve to determine the important content of the source document, and consequently be used to assess the information contained in the peer summary.

Another statistical method is the **AutoSummENG** proposed by Giannakopoulos et al. (2008). This method is based on n-grams graphs, and takes into account within-a-window co-occurrence of either word n-grams or character n-grams. In this type of approaches, it is always desired to use the lemma form of the same word, i.e. to convert all different forms of a word to its lemma (e.g. convert *goes*, *going* to *go*). This requires lemmatizers which are not available for many languages but just for few languages such as English, German, etc. To keep this approach language independent, the authors simply use character n-grams (sequences of characters) in the co-occurrence computation. E.g., if two character n-grams are taken then *go* is always extracted from our examples *goes* and *going* which is identical with the lemma of both words. This method has been shown to have higher correlation with human judgements than ROUGE. In Giannakopoulos and Karkaletsis (2011a), the AutoSummENG method is applied over the TAC 2011 AESOP task data, along with the Merge Model Graph (MeMoG) variation. The first method creates a n-gram graph representation of the text to be evaluated, and another graph for the model summary. The Value Similarity metric is computed and used to compare the similarity of the evaluated text to each model summary. The average of these similarities is considered to represent the overall performance of the summary text. The second method, instead of comparing the graph representation of the evaluated summary text to the graph representation of individual model texts and averaging over them, calculates the merged graph of all model texts (MeMoG variation). Then, it compares the evaluated summary graph to this overall model graph. Both methods offered very good results in different aspects of the evaluation. In Giannakopoulos and Karkaletsis (2013), the authors propose the use of a combination of the graph-based methods mentioned above and other machine learning approaches (linear regression) to better estimate the final grade for automatic summaries. The authors call the approach the NPowER method (N-gram graph Powered Evaluation via Regression). For evaluation, they use the data generated within the AESOP tasks of 2009 and 2010, and compute Pearson, Spearman and Kendal correlations with respect to responsiveness and Pyramid. They show that the combined use of evaluation methods outperforms the individual ones.

One of the best systems that participated in the AESOP 2010 track and that deserves to be mentioned is CLASSY (Conroy et al. 2010). It extended ROSE (ROUGE Optimal Summarization Evaluation) (Conroy and Dang 2008a) to include new content and linguistic features. The following features were tested: term overlap, normalized term overlap, two redundancy scores, number of sentences, term entropy and sentence entropy. Three feature-combining methods were developed and tested: robust regression and non-negative least squares, which predict a manual evaluation score such as pyramid or overall responsiveness; and a canonical correlation method that predicts a weighted average of the manual scores. In order to determine which of these linguistic features and ROUGE-type features should be included in the model, they made use of the TAC 2008 and TAC 2009 data for training/testing different models.

In Conroy et al. (2011), an improved version of the CLASSY 2010 system was presented that included six variations of measures on bigrams (defined as follows by the authors):

1. ROUGE-2, the consecutive bigram score.
2. ROUGE-SU4, the bigram score that allows for a skip distance of up to 4 words.
3. Bigram coverage score. This score is similar to ROUGE-2 but does not take the frequency that the bigram occurs in either the model summaries or in the summary to be scored. A credit of $\frac{i}{n}$ for a bigram is given if i out of n model human summaries included that bigram.
4. Unnormalized ROUGE-2. The score is essentially ROUGE-2 without the normalization for the length of the summaries.
5. Bigram coverage, as measured by a point to point comparison. This score is similar to the 3rd score; however, it is computed comparing one summary to another as opposed to one summary to 3 or 4 summaries.
6. Unnormalized ROUGE-2 as measured by a point to point comparison. This score is a point to point version of score 4.

They used the TAC 2009 and TAC 2010 data to train the model. Three methods were used for feature selection and weighting: canonical correlation, robust least squares and non-negative least squares. Again, the CLASSY system was one of the best ranking systems in all evaluation metrics.

Although the use of models summaries (normally human ones) is quite common, some authors have been working toward the **automatic evaluation of summaries without using references**, which is one of the more challenging strategies nowadays. Louis and Nenkova (2008, 2009b, 2009a) were among the pioneering ones. Louis and Nenkova (2009b) present different experiments directed to evaluate summary quality without using human model summaries (or using just few of them). They analyzed several similarity metrics to compare a summary with its input, such us information-theoretic metrics (e.g., the Kullback–Leibler divergence and Jensen–Shannon divergence between vocabulary distributions of the input and summary), vector space similarity (cosine similarity on the TF*ID representations of the input documents and summaries), generative model (comparing word distributions of the input and summary), topic signatures (e.g., the percentage of summary content words which match the input topic words, the percentage of input topic words that also appear in the summary and the cosine overlap between inputs topic words and summary content words), and the regression-based combination of all above mentioned features. These input-summary similarity metrics obtain correlations of about 0.70 with manual pyramid scores on the TAC 2009 data. They also investigated whether system-produced summaries can be used to improve predictions of summary quality when few or no human summaries are available and found that using only a collection of system summaries in place of gold standard allows for a correlation of 0.90 with manual pyramid scores. Other research works under the same topic can be found in Torres-Moreno et al. (2010a), Torres-Moreno et al. (2010b) and Saggion et al. (2010). In these latter studies, the authors analyze the correlation of rankings of text summarization systems using evaluation methods with and without human models. The comparison made is applied to various well-established content-based evaluation measures in text summarization and within several summary types (e.g. generic, focus-based, multi-document, single-document). Specifically, the research is carried out using a content-based evaluation framework called FRESA[7] (FRamework for Evaluating Summaries Automatically) to compute a variety of divergences among probability distributions. FRESA provides a tool for automatically evaluating text summaries and is multilingual, thus working for French, Spanish, English, and German. Recently, Cabrera-Diego et al. (2016) proposed a trivergent model to evaluate summaries without human references. The model is based on three elements: the summary to be evaluated, its source document and a set of other summaries from the same source. The core of the approach relies on the probability distributions of the vocabulary in these elements using n-grams (unigrams, bigrams, and skip-grams). In their experiments, the set of other summaries are extracts created by humans, so this may be a limitation of this method, since this type of summaries are costly to obtain, and relying on other peer summaries may introduce inaccuracies to the results. To assess their method, the authors compute the correlation with respect to ROUGE, obtaining around 0.75 for Kendall and 0.90 for Spearman when the size of the corpora increase (the experiments were performed within a range from 3 to 170 documents). In comparison with FRESA tool, FRESA obtains lower correlation values, but shows a more stable behaviour regardless the corpus size.

Different from the aforementioned methods, an implementation of Van Dijk's theories about discourse analysis (Van Dijk 1972) is presented in Branny (2007). This approach relies on **text grammars**. A text grammar is a way of describing a valid text structure in a formal way, and it takes into consideration the surface and deep structure of sentences by means of their relationships (microstructure) 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 identifies first a list of propositions. Then, humans have to

[7] This tool can be downloaded at: http://fresa.talne.eu/.

Table 1 Summary of content based automatic evaluation methods

Study by	Units used for assessment	Assesment method	Refs needed	Semantic?	Use of external rsources	ML used
Edmundson (1969), Paice (1990), Kupiec et al. (1995), Marcu (1997), Salton et al. (1997), Ono et al. (1994), Donaway et al. (2000), Jing et al. (1998), Radev and Tam (2003)	Sentence	Assessment by exact match through precision, recall and F-measure	Yes	No	No	No
Lin (2004b)	n-grams	ROUGE metrics (recall, precision, f-measure)	Yes	No	No	No
He et al. (2008)	n-grams	ROUGE-C (metrics as in ROUGE)	No	No	No	No
Hovy et al. (2006), Tratz and Hovy (2008)	Basic elements triples (head, modifier, relation)	Scoring by matching basic elements	Yes	Yes	Require dependency parser (Minipar)	No
Owczarzak (2009)	Triples of words from parse tree	Scoring by matching triples extracted from parse tree	Yes	Yes	Require dependency parser (Charniak parser)	No
Zhou et al. (2006)	n-grams	Scoring through overlap of n-grams	Yes	Yes	MOSES SMT system Koehn et al. (2007)	No
Steinberger et al. (2009)	Concepts, n-grams	Recall	Yes	Yes	Named entity tagging and Resnik similarity (Resnik 1995)	Linear regression
Ng and Abrecht (2015a)	n-grams	As in ROUGE	Yes	Yes	Word embeddings	Deep learning
Katragadda (2010)	n-grams	Likelihood	Yes	No	Domain dependent signature words	No

Table 1 continued

Study by	Units used for assessment	Assesment method	Refs needed	Semantic?	Use of external rsources	ML used
Giannakopoulos et al. (2008), Giannakopoulos and Karkaletsis (2011a, b, 2013)	Chracter/ word n-grams	n-grams graph overlap	Yes	Yes	No	Linear regression
Conroy et al. (2010), Conroy et al. (2011)	Document/term	Learns score based on term statistics, document features, redundancy, etc.	Yes	No	No	Regression and non-negative least squares, canonical correlation
Louis and Nenkova (2009b)	Words	Cosine similarity, Kullback–Leibler divergence and Jensen–Shannon, generative model, topic signatures, etc.	No	No	Topic signatures	Linear regression
Cabrera-Diego et al. (2016), Torres-Moreno et al. (2010a), Saggion et al. (2010)	n-grams	probability distributions of n-grams	No	No	Different system summaries	No

The columns indicate how the methods differ from each other: *units used for assessment* indicates of what level the assessment is performed, *assessment method* captures the metric used for assessment, *refs needed* indicates whether the method needs any gold standard data or not, *semantic* refers whether the method captures semantic similarity or not, *use of external resources* refers whether the method relies on any additional resources and finally *ML* used indicates whether the method uses any machine learning decide whether each proposition is relevant or not for a summary. Further on, three scores are proposed, based on: (i) information overlap (how many propositions are present in the summary); (ii) misinformation (misleading statements detected in the summary); and (iii) 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 model and peer summaries shows that human summaries get higher scores than automatic, as it would have been expected. The main shortcoming of this method is that it is not possible to know how well it would correlate with human evaluation. Moreover, human intervention is required for identifying propositions and evaluating the amount of misinformation and grammar issues summaries have, which is very costly and time-consuming. Finally, due to the complexity of the method, it would not be easily scalable.

Finally, Table 1 lists the approaches described earlier by highlighting attributes such as the textual unit that is used as basis for evaluation, the assessment method, whether external resources are required, etc.

2.2 Evaluation strategies for non-English

Although most methods have been developed for English, other evaluation methodologies have been proposed specifically for languages such as Chinese or Swedish. **HowNet**[8] (Dong and Dong 2003) is an electronic knowledge resource for English and Chinese similar to WordNet, but differing from it in the way in which word similarity is computed. Moreover, HowNet provides richer information and each concept is represented unambiguously by its definition and association links to other concepts. It is a well-known resource for Chinese, and has been used in many approaches also for the evaluation of peer summaries, such as in Wang et al. (2008). 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 also be taken into consideration. In addition, the authors also claim that this approach could be also used for detecting a few quality metrics to some extent, such as conciseness or sequence ordering.

Saggion et al. (2002) suggested a framework for evaluating different types of summaries both in English and Chinese. The method used only relied on vocabulary overlap by means of cosine similarity. Moreover, model summaries were also needed in order to be compared with peer summaries.

Specific evaluation tools and resources for Scandinavian languages (mostly Swedish and Norwegian) have been also developed. Dalianis and Hassel (2001) developed a newswire corpus useful for evaluating summaries in Swedish (KTH extract corpus) which contains a set of documents together with the corresponding extracts manually written. In addition, Hassel (2004) proposed an evaluation framework (KTH extract tool). This tool is capable to compute some statistics with regard to the source documents and the summaries. For instance, how close a summary is with respect to a model one, or which text units appear more frequently

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

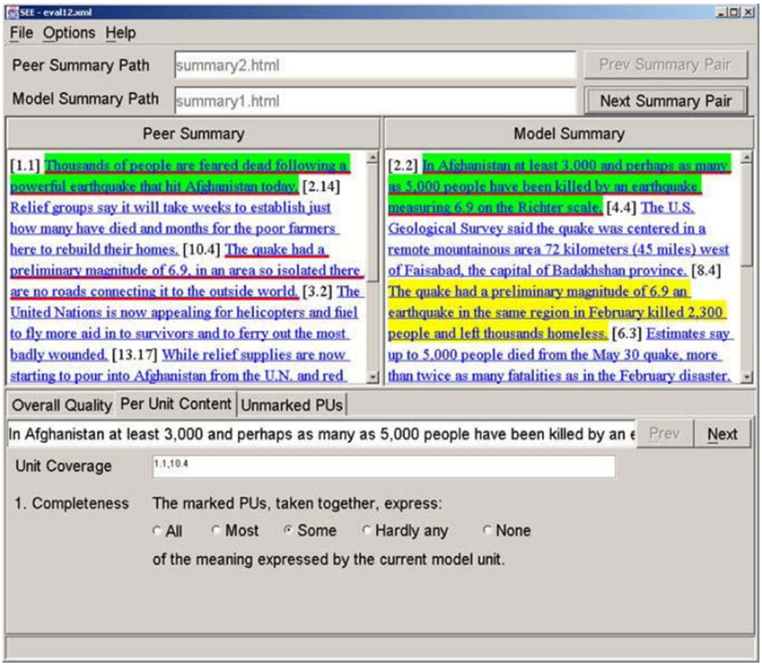
in model summaries. In a similar way, a corpus and a set of evaluation resources for the Norwegian language are suggested in Liseth (2004).

2.3 Readability

Although the previously explained evaluation methods are useful to assess the quality of a summary, they mainly provide information regarding its informativeness. The way the information contained in the summary is presented is also very important, since it also affects the summary's quality. This is crucial for determining how helpful a summary is when a user reads it. Previous works have highlighted the danger of divorcing evaluations of summary content from linguistic quality (Conroy and Dang 2008b), showing that, while ROUGE strongly correlates with human responsiveness, there is a gap in responsiveness between humans and systems that is not accounted for by ROUGE and that is frequently caused by the truncation of the last sentence made by most automatic summarizers.

Different studies have investigated metrics and methods to automatically assess the automated summaries with respect to different linguistic and readability aspects of the summaries, including text coherence, grammmaticality, summary indicativeness and sentence acceptability.

Text coherence is an essential characteristic that summaries should account for. However, it is very difficult to correctly measure it. Pioneering attempts to find automatic approaches to model and evaluate the coherence of a text can be found in Barzilay and Lapata (2005), Lapata and Barzilay (2005a) and Barzilay and Lapata (2008). These works provided a correlation analysis between human coherence assessments and semantic relatedness between adjacent sentences, and analyzed how mentions of the same entity in different syntactic positions are spread across adjacent sentences. Hasler (2008) adapted the Centering Theory (Grosz et al. 1995) to the automatic assessment of coherence by using guidelines describing how best to post-edit the automatically produced extract to transform it into a readable and coherent abstract. In Pitler et al. (2010), several classes of metrics to capture various aspects of well-written text are presented, including word choice, the reference form of entities, local coherence and sentence fluency. To measure each of these aspects, different general and summary specific features are proposed, such as cohesive devices, adjacent sentence similarity, Coh-Metrix, word co-occurrence patterns, and entity-grid. In Lin et al. (2012) and Wang et al. (2013), discourse relations extracted from summary sentences are used to assess the readability of automatic summaries. In Rankel et al. (2012), both content coverage and linguistic quality of automatic summaries are measured. Linguistic features correspond to the grammaticality, readability, and flow, including any impact of redundancy, of the summary. In Christensen et al. (2013), a graph-based multi-document summarizer is presented which estimated the coherence of the summaries that are generated by using textual clues (such as noun references, discourse markers ("however", "but", etc.) and co-reference mentions) to identify pairs of sentences that have a relationship. In Smith et al. (2012), a cohesive extractive single-document summarizer is presented. Sentences providing the most references to other sentences and that other sentences



Peer Summary Path: summary2.html

Model Summary Path: summary1.html

Peer Summary

[1.1] Thousands of people are feared dead following a powerful earthquake that hit Afghanistan today. [2.14] Relief groups say it will take weeks to establish just how many have died and months for the poor farmers here to rebuild their homes. [10.4] The quake had a preliminary magnitude of 6.9, in an area so isolated there are no roads connecting it to the outside world. [3.2] The United Nations is now appealing for helicopters and fuel to fly more aid in to survivors and to ferry out the most badly wounded. [13.17] While relief supplies are now starting to pour into Afghanistan from the U.N. and red

Model Summary

[2.2] In Afghanistan at least 3,000 and perhaps as many as 5,000 people have been killed by an earthquake measuring 6.9 on the Richter scale. [4.4] The U.S. Geological Survey said the quake was centered in a remote mountainous area 72 kilometers (45 miles) west of Faisabad, the capital of Badakhshan province. [8.4] The quake had a preliminary magnitude of 6.9 an earthquake in the same region in February killed 2,300 people and left thousands homeless. [6.3] Estimates say up to 5,000 people died from the May 30 quake, more than twice as many fatalities as in the February disaster.

In Afghanistan at least 3,000 and perhaps as many as 5,000 people have been killed by an e

Unit Coverage: 1.1,10.4

1. Completeness The marked PUs, taken together, express:

○ All ○ Most ● Some ○ Hardly any ○ None

of the meaning expressed by the current model unit.

Fig. 1 Summarization evaluation environment (SEE)

are referring to, are considered the most important and are therefore extracted for the summary.

Attempts to automatically evaluate the **grammaticality** of a summary have been explored in Vadlapudi et al. (2010a). 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 summaries were evaluated in DUC or TAC. Results obtained correlate well (85% at most) with respect to the already existing manual evaluations. Furthermore, in Vadlapudi and Katragadda (2010c), structure and coherence aspects are also investigated on the basis of lexical chains and the semantic relatedness of two entities. Results achieve a 70% agreement with human assessments when measured using the Spearman's correlation.

The evaluation of **summary indicativeness** and **sentence acceptability** was also addressed in Saggion and Lapalme (2000). On the one hand, *indicativeness* measures whether the summary is able to extract the topics of the document. The authors focus on scientific papers, and therefore, *indicativeness* is computed by comparing the terms appearing in the summary to the ones included in the abstract this type of documents already contain. Using the abstracts already given in the document avoids the costly task of producing model summaries again; but there is a limitation regarding this issue, since not all documents contain an abstract, so in these cases human need would be necessary. On the other hand, *acceptability* determines if a selected sentence by a summarization system is adequate compared to what humans would have selected, in terms of some criteria such as "good grammar", "correct spelling" and "brevity and conciseness". In this case, human intervention is needed to evaluate this criterion.

Conroy and Dang (2008a) address the need of having tools to assess the content as well as other linguistic aspects in summaries. For this reason, **ROSE** (ROUGE Optimal summarization Evaluation) was developed. This tool is based on ROUGE but, in order to account for linguistic aspects, the idea behind it is to find which ROUGE metrics better correlate with the overall responsiveness criteria manually evaluated in DUC and TAC conferences (see Sect. 3 for overall responsiveness evaluation).

Despite the challenges involved in automating quality criteria for evaluating summaries, the number of approaches attempting to automate some of these criteria, such as grammaticality or coherence, has increased considerably. Consequently, research in the evaluation of the summary's readability is advancing beyond the content assessment only. Recently, the method presented in (Ellouze et al. 2016) proposes the adaptation of ROUGE scores to evaluate also the structure and the grammaticality of a summary. This adaptation consists of comparing the peer summary with respect to one or several source documents, instead of using reference summaries. Then, other features to build the model include the use of readability metrics, such as FOG, or Flesh Reading Ease, or local coherence features, that are computed using common similarity metrics, such as cosine or Levenshtein distance. The authors build a model using machine learning techniques which can predict the linguistic quality of a summary that correlates with the score assigned by human judges in TAC 2008 conference.

3 Manual evaluation

The manual evaluation of a summary is not a trivial task. On the one hand, a lot of human effort is involved to be able to assess either the content or the readability of a summary. On the other hand, the inherent subjectivity of the evaluation may lead to the fact that the agreement between assessors is not reached, thus assigning totally different scores for the same summary. In order to overcome with this limitation, different strategies and methodologies have been proposed to perform this task, in which human expert judgements rely on a specific criteria with a pre-fixed scale of evaluation values (e.g., 3 or 5 valued Likert scale). The objective of this section is to provide an overview of the different methodologies and strategies that have been proposed along the years to manually evaluate a summary. A revision of the methods for evaluating the content of a summary is first provided (Sect. 3.1), followed by the revision of the strategies adopted to assess a summary's readability (Sect. 3.2).

3.1 Summary content

In the first DUC editions, evaluation of summary content was performed mostly manually. NIST assessors used a modified version of the Summary Evaluation Environment (SEE) (Lin 2001) to support the process (see Fig. 1). Using SEE, the assessors compared the systems text (the peer text) to the ideal (the model text) (Lin and Hovy 2002). In SEE, each text is decomposed into a list of units (e.g., sentences) and displayed in separate windows. SEE provides interfaces for assessors to judge both the content and the quality of summaries. The evaluator reads the peer summary and then makes overall judgements as to the peer summary's content, grammaticality, cohesion and organization. To measure content, assessors step through each model unit, mark all system units sharing content with the current model unit (shown in green highlight in the model summary window), and specify that the marked system units express *all*, *most*, *some*, *hardly any* or *none* of the content of the current model unit.

A few years later, Teufel and van Halteren (2004) proposed the **Factoid method**. This method compares the overlap of atomic information units (factoids) extracted from summaries. A factoid represents the meaning of a sentence. Factoids are based on the idea of *information nuggets*, which are facts which help humans to assess automatic summaries by checking whether the automatically generated summary contains the fact or not (Voorhees 2003). For instance, for the sentence *The police have arrested a white Dutch*, the following factoids are generated:

- A suspect was arrested.
- The police did the arresting.
- The suspect is white.
- The suspect is Dutch.

In this method, the idea is to use several model summaries as gold standard and measure the information overlap among them, identifying the associated factoids and assigning them a weight based on the degree of agreement found. After that, an automatic summary is evaluated with respect to the number of factoids it contains, and their associated weights are employed to score the summary.

From DUC 2006, the **Pyramid method** (Nenkova and Passonneau 2004; Nenkova et al. 2007) has been used to evaluate summary content, along with the ROUGE metrics. The Pyramid method is semi-automatic, as it requires part of the core evaluation task to be completed by humans. The method was created under the assumption that no single best model summary exists. Its main idea is to create a gold-standard based on a comparison between human-written summaries in terms of Summary Content Units (SCUs). From a set of model summaries, the authors manually identify similar sentences. From these similar sentences, SCUs are generated and ranked in a pyramid model. The pyramid model has *n* levels, where *n* is the number of model summaries. The levels are labeled in ascending order from 1 to *n*. SCUs are ranked in the pyramid according to their occurrence in the model summaries. The resulting set of SCUs is what is called a "pyramid". For instance, if a SCU occurs in 3 of the 4 model summaries then this SCU will be placed in the 3rd

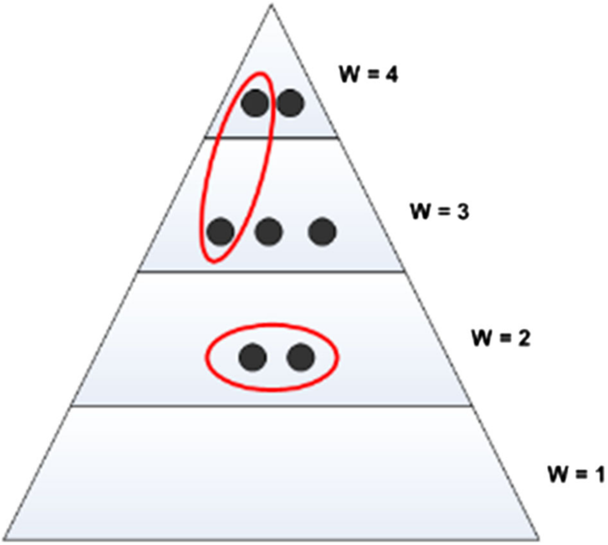


Fig. 2 Example of a pyramid with SCUs identified and marked

Table 2 Evaluation questions in DUC 2002

1. About how many gross capitalization errors are there?
2. About how many sentences have incorrect word order?
3. About how many times does the subject fail to agree in number with the verb?
4. About how many of the sentences are missing important components (e.g. the subject, main verb, direct object, modifier) causing the sentence to be ungrammatical, unclear, or misleading?
5. About many times are unrelated fragments joined into one sentence?
6. About how many times are articles (a, an, the) missing or used incorrectly?
7. About how many pronouns are there whose antecedents are incorrect, unclear, missing, or come only later?
8. For about how many nouns is it impossible to determine clearly who or what they refer to?
9. About how times should a noun or noun phrase have been replaced with a pronoun?
10. About how many dangling conjunctions are there (“and”,“however”...)?
11. About many instances of unnecessarily repeated information are there?
12. About how many sentences strike you as being in the wrong place because they indicate a strange time sequence, suggest a wrong cause-effect relationship, or just don’t fit in topically with neighboring sentences?

level of the pyramid. The exact formula to evaluate a peer summary that has X SCUs against n model summaries is shown in Eq. 7.

$$Max = \sum_{i=j+1}^{n} i * |T_i| * (X - \sum_{i=j+1}^{n} |T_i|) \tag{7}$$

where $j = max_i(\sum_{t=i}^{n} |T_t| \geq X)$, T_x is the tier at level x.

Based on this equation, a summary is regarded as good if it contains a large number of the higher-level SCUs. Summaries containing more SCUs from the lower levels than from the higher levels are considered poor summaries as they are less informative.

For example, if there are four reference summaries, a SCU appearing in all summaries can be thought of as one of the most important ideas and would receive a weight of 4. A SCU appearing in just one reference summary would be regarded as less important, and would receive a weight of 1. A pyramid is formed because the tiers descend with the SCUs assigned the highest weight at the top, and the SCUs with the lowest weight appearing in the bottom-most tiers. The fewest SCUs would appear in the topmost tier since fewer concepts would be present in all reference summaries (see Fig. 2). An annotation tool[9] was also developed to facilitate the task of marking SCUs. This method has been applied not only in DUC conferences but also in very recent works, especially related to the generation of abstractive summarization (Khan et al. 2015). Moreover, attempts to automate the Pyramid evaluation approach were proposed in (Harnly et al. 2015) and (Passonneau et al. 2013), respectively. In both studies, the focus was on automating and improving the scoring of a summary. In the former, dynamic programming techniques were used to find an optimal candidate contributor set of a summary based on different clustering methods and similarity metrics that go beyond n-grams. In the latter, three automated pyramid scoring procedures were proposed, obtaining the best results for the one that was based on distributional semantics. The authors found that the results obtained with this approach correlated best with manual pyramid scores, and had higher precision and recall for content units than other methods depending on string matching.

In DUC 2005 to 2007 editions, NIST assessors manually assigned a **responsiveness** score to each summary, which indicated the amount of information in the summary actually helps to satisfy the information need expressed in the topic statement, in the context of a topic-oriented summarization task (Dang 2005). To this end, assessors were given a topic statement and a simple user profile, along with a file containing a number of summaries that contribute toward satisfying the information need expressed in the topic. Assessors should read the topic statement and all the associated summaries, and then grade each summary according to how responsive it is to the topic in relation to the others (1 = worst, 5 = best).

3.2 Readability

As already told, efforts for the development of automatic evaluation methods have mostly focused on evaluating the summary's content. In contrast, evaluation of readability has been performed almost exclusively manually.

DUC and TAC conferences manually assess each summary for readability. In the first DUC editions, NIST assessors used the SEE environment (Lin 2001) to measure readability by rating the summaries according to their overall

[9] http://www1.cs.columbia.edu/~ani/DUC2005/Tool.html.

Table 3 Corpora for text summarization (details)

Data set	Doc. type	Summary type	Summarization task	# documents	# clusters	# model summaries
DUC 2001	News	Multi-doc	Generic	309	30	Abstracts (400, 200, 100, 50 words)
DUC 2002	News	Multi-doc Single-doc	Generic	567	59	Abstracts (10, 50, 100, 200 words) and extracts (400, 200 words)
DUC 2003 (Task 1)	News	Single-doc	Generic	734	60	Abstracts (10 words)
DUC 2003 (Task 2)	News	Multi-doc	Topic-oriented	298	30	Abstracts (100 words)
DUC 2004 (Tasks 1)	News	Single-doc	Generic	500	50	Abstracts ($\leq$75 bytes)
DUC 2004 (Tasks 2)	News	Multi-doc	Generic	500	50	Abstracts ($\leq$665 bytes)
DUC 2005	News	Multi-doc	Query-focused, user-oriented	1593	50	Abstracts (250 words)
DUC 2006	News	Multi-doc	Query-focused, user-oriented	1250	50	Abstracts (250 words)
DUC 2007 (Main task)	News	Multi-doc	Query-focused, user-oriented	1125	45	Abstracts (100 words)
DUC 2007 (Update task)	News	Multi-doc	Updated summary	250	10	Abstracts (100 words)
TAC 2008 (Update task)	News	Multi-doc	Updated summary	960	48	Abstracts (100 words)
TAC 2008 (Opinion task)	Blogs	Multi-doc	Query-focused	609	25	–
TAC 2009 (Update task)	News	Multi-doc	Updated	880	44	Abstracts (100 words)
TAC 2010	News	Multi-doc	Guided summary, topic-oriented	920	46	Abstracts (100 words)
TAC 2011	News	Multi-doc	Guided summary, topic-oriented	880	44	Abstracts (100 words)
MSE 2005	News	Multi-doc	Multilingual	–	25	Abstracts
MSE 2006	News	Multi-doc	Multilingual	–	24	Abstracts

Table 3 continued

Data set	Doc. type	Summary type	Summarization task	# documents	# clusters	# model summaries
MultiLing 2011	Wikinews	Single-doc	Multilingual	–	7	Abstracts
MultiLing 2013	Wikipedia	Multi-doc	Multilingual	1200	40	Abstracts
MultiLing 2015, 2017	Wikipedia	Multi-doc Single-doc	Multilingual	1140	38	Abstracts
CAST Corpora	News+articles	Single-doc	Generic	163	–	Extracts
BC3	Emails	Multi-doc	Generic	261	40	Extracts and abstracts
Multilingual Summ. Eval.	Comments	Multi-doc	Multilingual	20	4	Extracts
Image Captioning	Websites	Multi-doc	Topic-oriented	9210	307	Abstracts (200 words)
CMU Movie Summary	Movie plots	Single-doc	Generic	42306	–	Abstracts (700 words)
CONCISUS	Web pages	Multi-doc	Event-oriented	184	4	Abstracts

grammaticality by answering a set of questions (see Table 2 for the questions posed in DUC 2002 (Over and Liggett 2002)).

In most recent editions, expert judges are asked to evaluate summaries according to five aspects/criteria: grammaticality, non-redundancy, referential clarity, focus and coherence/structure. The definition of such aspects that are provided by the organization are given below:

- *Grammaticality* The summary should have no datelines, system-internal formatting, capitalization errors or obviously ungrammatical sentences (e.g., fragments, missing components) that make the text difficult to read.
- *Non-redundancy* There should be no unnecessary repetition in the summary. Unnecessary repetition might take the form of whole sentences that are repeated, or repeated facts, or the repeated use of a noun or noun phrase (e.g., Bill Clinton) when a pronoun (he) would suffice.
- *Referential clarity* It should be easy to identify who or what the pronouns and noun phrases in the summary are referring to. If a person or other entity is mentioned, it should be clear what their role in the story is. So, a reference would be unclear if an entity is referenced but its identity or relation to the story remains unclear.
- *Focus* The summary should have a focus; sentences should only contain information that is related to the rest of the summary.
- *Structure and coherence* The summary should be well-structured and well-organized. The summary should not just be a heap of related information, but should build from sentence to sentence to a coherent body of information about a topic.

Each question was assessed on a five-point scale: *1. Very poor*, *2. Poor*, *3. Barely acceptable*, *4. Good*, *5. Very good*. The problem associated to this type of evaluation is that humans could understand concepts such as “repetitions”, but this would be very difficult for computers. In this case, it would be possible to map the outermost values into a quantitative scale (i.e., “Quite a lot”, and “None”). For instance, “None” would mean no repetition at all, but the boundaries in the middle are very subtle. Moreover, this sort of statements contain a degree of subjectivity, which is not possible to capture automatically. All these issues make the task of evaluating a summary’s quality very challenging and difficult to tackle from an automatic point of view.

4 Golden standard data

As it has been previously shown, the automatic evaluation of summaries normally needs reference summaries to be used for comparison with the peer ones. However, building these reference or model summaries is not a trivial task. Therefore, the availability of corpora and datasets specifically developed for summarization is of great help for the research community, allowing different summarization approaches to be fairly compared against the same reference summaries. Table 3 presents the most popular datasets for evaluation of automatic summarization. This table shows the most relevant features of these datasets, including the type of documents to be summarized, the characteristics of the model summaries that are provided and the summarization tasks that they aim to address. The datasets are discussed in detail next.

Thanks to international evaluation campaigns, such as DUC and TAC, as well as the effort done by the research community, several summarization corpora and datasets are available for academic and research purposes. The corpora used at DUC and TAC conferences deal mainly with generic newswire documents gathered from several press agencies. The model summaries provided are either extracts or abstracts written by humans. These model summaries vary in content and length, depending on the proposed task in each conference edition (e.g. single-document or multi-document summarization, and from 50 to 250 words). Apart from newswire, other types of documents, such as blogs and posts were employed in more recent years. In particular, a new collection of documents pertaining to the *Blog06*[10] was used as corpora for generating summaries. In this case, instead of providing complete model summaries, humans were asked to select fragments of information that were more relevant to the task, since summaries were evaluated using the pyramid method (Nenkova and Passonneau 2004; Nenkova et al. 2007) (see Sect. 3).

Concerning evaluation fora for multilingual summarization, regardless of the specific tasks proposed within the DUC conferences (e.g., DUC 2004 for English and Arabic summarization), one of the pioneering was the Multilingual Summarization Evaluation (MSE) organised within the Translingual Information Detection, Extraction and Summarization (TIDES) program. The purpose was the generation of multi-document summaries from a mixture of English and English translations from Arabic documents. In particular, given a cluster of topic-related documents in English, the task was to create a multi-document 100 word summary of the cluster as a response to a “wh” question. A complete an detailed description of this task is provided in (Stewart 2008, and CLASSY summarizer, the best performing system, is further explained in (Schlesinger et al. 2008). After this initiative, in the recent years a set of bi-annual conferences also addressing multilingual summarization for a wide variety of languages appeared. These were known as MultiLing, and took place in the years 2011, 2013, 2015 and 2017. MultiLing started as a pilot task within TAC conferences, but after the interest of the research community in this task, it became a workshop in prestigious conferences, such as ACL in 2013, SIGDIAL in 2015 or EACL in 2017. The main objective of the proposed tasks was to encourage research into language-independent or multilingual summarization systems, that had to prove and validate this at least for two different languages from a set of more than 35, including Hinidi, Czech, or Hebrew. In these conferences, the data was extracted from either Wikinews or Wikipedia about heterogeneous topics, and it was clustered with respect to the language of the article. As reference summaries, the first paragraph of the articles was considered as the abstract.

[10] http://ir.dcs.gla.ac.uk/test_collections/blog06info.html.

Out of these international evaluation fora, other corpora that are also useful for the evaluation of automatic summaries can be found. The CAST Project Corpus (Hasler et al. 2003) consists of 163 documents, comprising newswire and articles about popular science. This corpus differs from others in that, apart from containing information about the importance of a sentence in a document, it also indicates which fragments of a sentence can be removed without affecting the sense of the sentence. This fine-grained annotation is very useful for evaluating the conciseness and coherence of the summaries.

The *AMI Meeting Corpus* (Carletta et al. 2005) was developed as part of the AMI project[11] and it consists of 100 hours of meeting recordings in English. Although it is not specifically for text summarization, it can also be adapted for this type of summaries, and it provides abstractive and extractive human-written summaries as well.

The *BC3: British Columbia Conversation Corpus* (Ulrich et al. 2008) is a corpus specifically developed for the task of email summarization. It consists of 3,222 annotated sentences extracted from 40 email threads. For each thread, extractive and abstractive summaries are created by three annotators.

Apart from the fora specifically addressing multilingual summarization, there exist another independent corpora for this purpose as well. The *Multilingual summary evaluation data* from the Joint Research Centre (JRC) (Turchi et al. 2010) is a set of documents related to four topics (genetic, the-conflict-between-Israel-and-Palestina, malaria, and science-and-society). Each cluster contains the same 20 documents in seven languages (Arabic, Czech, English, French, German, Russian and Spanish). In addition, the relevant sentences of each document are manually annotated, and as a consequence, this dataset is very appropriate for evaluating single- or multi-document, as well as multi-lingual extractive summarization systems.

Also for multi-lingual summarization, particularly for English and German but in the context of image captioning generation, Aker and Gaizauskas (2010) developed a corpus of 932 human-written abstractive summaries that describe the most relevant facts of object types found in Wikipedia. For instance, given the object *zoo*, model summaries for *Edinburgh Zoo*, or *London Zoo* are provided. The model summaries were collected first for English and then automatically translated to German. In order to assure that the translation was correct, a manual post-editing process was carried out, where the wrong translated sentences were corrected.

The *ESSEX Arabic summarization Corpus* (El-Haj et al. 2010) was created using a crowdsourcing service (i.e., Amazon's Mechanical Turk). This corpus includes 153 Arabic articles and 765 human-written extractive summaries.

The *CMU Movie Summary Corpus* (Bamman et al. 2013) provides a collection of 42,306 movie plot summaries and metadata about popular movies. The information for creating this corpus comes from two different resources: Wikipedia and Freebase.[12] The first resource is employed for searching the Wikipedia entry corresponding to a specific movie, and extracting the text included in the "plot"

[11] http://www.amiproject.org/.

[12] https://www.freebase.com/.

Table 4 Where to find corpora for text summarization

Corpora	Source
DUC corpus	http://duc.nist.gov/data.html
TAC corpus	http://www.nist.gov/tac/data/index.html
MultiLing corpus	http://multiling.iit.demokritos.gr/
CAST corpus	http://clg.wlv.ac.uk/projects/CAST/corpus/
AMI Meeting Corpus	http://corpus.amiproject.org/
BC3	http://www.cs.ubc.ca/nest/lci/bc3/download.html
JRC Multilingual summary evaluation data	https://ec.europa.eu/jrc/en/language-technologies
Image Captioning Corpus	http://staffwww.dcs.shef.ac.uk/people/A.Aker/
ESSEX Arabic summarization Corpus	http://www.lancaster.ac.uk/staff/elhaj/corpora.htm
CMU Movie Summary Corpus	http://www.ark.cs.cmu.edu/personas/
CONCISUS Corpus	http://www.taln.upf.edu/pages/concisus/index.html

subsection of the entry; whereas the latter, Freebase, is used to add extra information to the summary in the form of metadata (e.g., who the director was, its year, etc.). It is important to note that the model summaries contained in this corpus are longer than the ones developed in DUC/TAC (more than 700 word-length in comparison to the 250 word-length that is normally used in these evaluation fora).

The *CONCISUS Corpus* (Saggion and Szasz 2012) was created with the purpose of helping both summarization and information extraction systems. In particular, this corpus is available in English and Spanish, and contains comparable event summaries of four domains: aviation accidents, rail accidents, earthquakes, and terrorist attacks. For each domain, the summaries are intended to provide key information about the event: what happened, when, where, who was involved, where it took place, etc. Given the multilinguality of this corpus, it is very useful to evaluate multi-lingual or cross-lingual summarizers.

Finally, Table 4 provides the source where each of the gold standard corpora can be downloaded or requested.

4.1 Crowdsourcing evaluation

Due to the difficulty associated to the creation of reference summaries, crowdsourcing services have become a good alternative to recruit, in an easy and quick manner, users that want to perform different tasks (e.g., manual summarization evaluation), in return for money as a reward for the job done. Given the popularity of crowdsourcing services in recent years, this section aims to describe how crowdsourcing services have been used for summarization.

In particular, Amazon Mechanical Turk[13] (MTurk) and Crowdflower[14] provide a framework that allows users (requesters) to define and upload Human Intelligence

[13] https://www.mturk.com/mturk/welcome.

[14] https://www.crowdflower.com/.

Tasks (HITS). These HITS are then performed by other humans (turkers or workers), who are rewarded with the corresponding amount of money associated to the task. Therefore, we could think, for instance, that generating model summaries would be relatively easy and fast using crowdsourcing services. However, it has been shown in Gillick and Liu (201) that one has to be very careful with the annotations provided by these services, since they are not always as good as they should be. The quality of the results has to be checked and therefore, when using this type of services, it is very important to ensure that turkers are suitable for the task, as well as to check that they do not give random answers. For this reason, MTurk, itself, provides a facility to assist quality control. Requesters can attach various requirements to their task in order to force turkers to meet such requirements before they are allowed to work (Tang and Sanderson 2010). For instance, the percentage of the accepted tasks a turker has completed can be used in order to decide if it is worthy to allow such turker to perform the tasks.

Focusing on summarization, crowdsourcing has been not as explored as for other applications, such as machine translation (Callison-Burch 2009). The subjectivity involved in the evaluation of summaries and the limitations associated to the quality of annotations was shown in Gillick and Liu (201), where the difficulty of obtaining the same readability results for peer summaries as in TAC 2009 with non-expert judges in contrast to expert ones was investigated. Quality control policies were first established, in order to assure that only turkers with a 96% HIT approval could perform the task. In addition, if the task was finished under 25 seconds, their work was rejected. Concerning the amount of money it was paid, different compensation levels were analyzed, finding out that lower compensations ($0.70 per HIT) obtained higher quality results. It seemed that this compensation level attracted turkers less interested in making money and more conscious of their work. Assessors were asked to read a topic and a description along with two different reference summaries, and provide a score between 1 (very poor) and 10 (very good) for each of the following criteria: (1) overall quality and (2) linguistic quality. Regarding the results obtained, average for TAC assessors was 4.13 for overall quality and 4.99 for linguistic quality, while MTurkers showed averages of 6.41 and 6.79 respectively. Also, the MTurk evaluation presented higher variability. Whereas TAC assessors could roughly agree on what makes a good summary, obtaining a standard deviation of 1.0 (ranging from 3.0 to 6.5), the standard deviation computed for turkers' results was doubled, obtaining a value of 2.3 (ranging from 1.0 to 9.5). As can be noticed, in this case, non-expert evaluation differed a lot from the official one, and therefore, MTurk was not of great help. However, El-Haj et al. (2010) showed the appropriateness of using MTurk for collecting a corpus of single-document model summaries from Wikipedia and newspaper articles in Arabic. These summaries were produced by extracting the most relevant sentences of the documents and not taking more than half of the sentences in the source documents. Finally, 765 model summaries were gathered. These summaries were then used to evaluate the corresponding automatic ones produced by several existing Arabic summarization systems using different evaluation approaches, such as ROUGE or AutoSummENG. In this case, MTurk facilitated the process of gathering a big number of model summaries.

More recently, in Lloret et al. (2013) a deep study of the use of crowdsourcing services for automatic summarization was carried out. Different short tasks were proposed for identifying relevant information from source documents, that could be later used to generate multi-document summaries. Quality control mechanisms were also established in order to prevent malicious workers from doing the tasks incorrectly. Even though for some of the experiments conducted, the results were not satisfactory, the analysis performed for determining the reason of the low performance led to the fact that the difficulty of the task itself had more influence than the amount of money paid for each task. Regarding the money to be paid, there was not a consensus in previous research works. Aker et al. (2012a) showed that high payments lead to better results, however Mason and Watts (2010) and Feng et al. (2009) argued that higher payments attracted more spammers, thus resulting in a decrease of quality in the job performed. This was confirmed by the experiments proposed in Lloret et al. (2013), where the amount of money paid for the same task was increased through small intervals. Despite this increase, the results did not improve, thus showing that there was not any relationship between the money paid and the quality of the results. The main conclusion drawn in Lloret et al. (2013) was that the major issue in obtaining high quality results was the level of difficulty of a task. In this manner, easier and faster tasks resulted in better results, while if a task required a lot of time to be completed and was difficult to perform, the workers lost their motivation on the task and the results were negatively affected.

As a conclusion, it seems that crowdsourcing evaluation is only adequate for easy, fast and clearly defined tasks that do not require of any expert knowledge or understanding of complex working criteria. Only in such scenario, and given that control mechanisms are established, can the results be trustable. For most complex tasks, such as the evaluation or the generation of textual summaries, and in the light of the results of previous investigations, crowdsourcing evaluation does not seem appropriate.

5 Assessing the effectiveness of evaluation metrics and approaches

The assessment of the automatic evaluation metrics standardly used in summarization research is essential to trust in the evaluation results, as well as to better understand and interpret them.

In order to evaluate the effectiveness of automatic evaluation metrics, Lin and Hovy (2003) proposed two criteria:

1. Automatic evaluations should correlate highly, positively, and consistently with human assessments.
2. The statistical significance of automatic evaluations should be a good predictor of the statistical significance of human assessments with high reliability.

Similarly, Owczarzak et al. (2012b) established two aspects of evaluation that should be paid special attention:

1. Significant difference: Ideally, all system comparisons should be performed using a test for statistical significance. As both manual metrics and automatic metrics are noisy, a statistical hypothesis test is needed to estimate the probability that the differences observed are what would be expected if the systems are comparable in their performance. When this probability is small (by convention 0.05 or less) we reject the null hypothesis that the systems performance is comparable.
2. Type of comparison: Established manual evaluations have two highly desirable properties: (1) they can tell apart good automatic systems from bad automatic systems and (2) they can differentiate automatic summaries from those produced by humans with high accuracy.

The most common way to assess the effectiveness of an automatic evaluation method (as in other natural language processing task) is to compute the **correlation** between the automatic prediction of the evaluation method and the manual evaluation of human judges. Correlation reflects how the value of one variable changes when the value of another variable changes. Depending on the change direction, two types of correlation exist: direct correlation (positive correlation), if both variables change in the same direction; and indirect correlation (negative correlation), if both variables change in opposite directions.

To this end, the general process is to first score the set of summaries generated using the automatic method and then assess each of the summaries by human judges using one or several metrics (e.g., responsiveness or readability). The evaluation system scores are then compared to the human assessments to see how well they correlate. Correlation is usually computed at the system level as the average of a number of summaries using three different coefficients: (1) Pearson correlation, (2) Spearman rank coefficient, and (3) Kendall rank coefficient (Ng and Abrecht 2015b).

- *Pearson correlation* measures the degree of the relationship between linearly related variables. The following formula is used to calculate it:

  $$r = \frac{N\sum XY-(\sum X\sum Y)}{\sqrt{[N\sum x^2-(\sum x)^2][N\sum y^2-(\sum y)^2]}}$$

  where r = Pearson r correlation coefficient; N = number of values in each data set; xy = sum of the products of paired scores; x = \,sum of x scores; y = \,sum of y scores; x^2 = sum of squared x scores; y^2= sum of squared y scores;
- *Spearman rank coefficient* is a non-parametric test that is used to measure the degree of association between two variables. It does not make any assumptions about the distribution of the data. The following formula is used to calculate the Spearman rank correlation:

  $$\rho = 1 - \frac{6\times\sum d_i^2}{n(n^2-1)}$$

  where ρ = Spearman rank correlation; d_i = the difference between the ranks of corresponding values x_i and y_i; n = number of values in each data set

- *Kendall tau rank coefficient* is a non-parametric test that measures the strength of dependence between two variables. The following formula is used to calculate the value of Kendall rank correlation:

  $$\tau = \frac{n_c - n_d}{\frac{1}{2}n(n-1)}$$

  where n_c = number of concordant; n_d = number of discordant.

The key difference between the Pearson correlation and the Spearman/Kendall correlation, is that the former assumes that the variables being tested are normally distributed while the latter are rank-based, thus not caring about whether correlation is linear or not. The latter two measures are however non-parametric and make no assumptions about the distribution of the variables being tested. Therefore, the Pearson correlation coefficient is usually employed to measure correlation with the scores while the Spearman/Kendall coefficients are used to measure correlation with the rankings. Higher values of the coefficients show higher (linear/non linear) correlation between the variables.

Table 5 shows the correlation values for some relevant automatic evaluation systems described in Sect. 2.1 in comparison with the Pyramids or ROUGE methods. The values have been extracted from the studies conducted in (Ng and Abrecht 2015a; Giannakopoulos and Karkaletsis 2011b; Torres-Moreno et al. 2010a; Cabrera-Diego et al. 2016). Please note that they are provided just as a reference to guide the reader of how different automatic methods correlate with human or other methods, but they are not comparable among them, since they have not been tested under the same corpus or conditions.

In the AESOP tracks, evaluation metrics are also tested for their **discriminative power**, i.e., the extent to which each metric can detect statistically significant differences between summarizers. The assumption is that a good automatic metric will make the same significant distinctions between summarizers as the manual metrics (and possibly add more), but will not give a contradicting ranking to two summarizers (i.e., infer that Summarizer X is significantly better than Summarizer Y when the manual metric infers that Summarizer Y is significantly better than Summarizer X) or lose too many of the distinctions made by the manual metrics.[15] In AESOP tracks, Pearson's, Spearman's, and Kendall's correlations between the summarizer-level scores produced by each submitted metric and the manual metrics (Overall Responsiveness and Pyramid) are also computed.

6 Task based evaluation

In task based evaluation, the idea is to assess a summary based on a task and measure how much help the summary provides for a human performing this particular task. This type of evaluation is also known as **extrinsic evaluation**. The extent to which a summary is helpful or not will depend on several aspects, such as

[15] TAC 2010 AESOP track. http://www.nist.gov/tac/2010/summarisation/AESOP.2010.guidelines.html.

Table 5 Best correlation values for automatic summarization evaluation methods, measured with Pearson r (P), Spearman ρ (S), and Kendall τ (K) coefficients

Method	Correlated with	P	S	K
ROUGE-1	Pyramids	0.9661	0.9085	0.7466
ROUGE-2	Pyramids	0.9606	0.8943	0.7450
ROUGE-SU4	Pyramids	0.9806	0.8935	0.7371
ROUGE-WE1	Pyramids	0.9492	0.9138	0.7534
ROUGE-WE2	Pyramids	0.9765	0.8984	0.7439
ROUGE-WE-SU4	Pyramids	0.9783	0.8808	0.7198
AutoSummENG	Pyramids	0.8420	0.9320	0.7980
MeMoG	Pyramids	0.9510	0.9240	0.7760
FRESA	Pyramids	–	0.8500	–
FRESA	ROUGE	–	0.7500	0.6300
Trivergent model	ROUGE	–	0.9200	0.7600

the time needed to process the document, its length, or the users' preferences (Hand 1997). The SUMMAC evaluation conference (Mani et al. 1999) was one of the first international fora where summaries were evaluated extrinsically within the task of question answering and categorization. Sixteen systems participated in this evaluation showing that the evaluation of text summarization was very effective in relevance assessment for these tasks. Moreover, it was shown that short summaries (less than 20% compression rate) allowed for relevance assessment almost as accurate as with full documents.

The INEX forum[16] (INitiative for the Evaluation of XML retrieval) was also created with the aim of promoting the evaluation of focused retrieval by providing large test collections of structured documents, uniform evaluation measures, and a forum for organizations to compare their results. Within this evaluation and from 2010, there is a track in this forum called Tweet Contextualization, which aims at providing an automatic summary that explains the tweets. This requires combining multiple types of processing, from information retrieval to multi-document summarization including entity linking. For this task, evaluation considers both informativeness and readability. From 2013, INEX takes place in the framework of CLEF (Conference and Labs of the Evaluation Forum). The FRESA evaluation system mentioned in Sect. 2.1 is currently used in CLEF-INEX.

There is also a number of applications where the integration or the use of automatic summaries have been shown to be appropriate, so therefore, they are evaluated in the context of these applications, such as information retrieval (Tombros and Sanderson 1998; Perea-Ortega et al. 2013; Alhindi et al. 2013), question answering tasks (Teufel 2001; Wu et al. 2004; Lloret et al. 2011; Jimeno-Yepes et al. 2013), report generation or synthesis tasks (Amigo et al. 2004; McKeown et al. 2005), or more recently, to obtain formative feedback (Labeke et al. 2013b, a; Field et al. 2013), or to manage clinical information (Zhu and Cimino 2013).

[16] http://inex.mmci.uni-saarland.de/.

Summaries applied to **information retrieval** have been normally used from a double perspective. On the one hand, summaries of the retrieved documents can be provided as a manner to reduce the time users need to decide whether a retrieved document is interesting or not for their purposes. On the other hand, summaries can be used as an intermediate stage within the information retrieval process, in order to speed the search process. Both approaches have been investigated by the research community. In Tombros and Sanderson (1998) the search and retrieval of relevant documents is first performed, and then, a query-based summary is generated taken into account the initial search query and the document retrieved. The usefulness of the summaries was evaluated by measuring users' speed and accuracy in identifying relevant documents, this being later compared to the traditional output of the information retrieval systems. The use of summaries was assessed according to the following criteria: (i) the recall and precision of the system; (ii) the speed users took to perform the judgement of the documents; (iii) the need of the users to seek assistance from the full text of the retrieved documents; and (iv) the users' opinion about the utility of the summaries. Different metrics were analyzed for evaluating each of the previous aspects, obtaining as conclusions that: (i) users preferred using query-based summaries rather than the traditional output of the information retrieval systems; (ii) the time for performing a relevance judgement was quicker when using summaries; (iii) summaries also provided the users with the necessary information, so they did not need to access the original documents; and (iv) users were more satisfied with the query-based summaries than with the full documents, since summaries were brief and more focused to the search query.

In Perea-Ortega et al. (2013), a deep analysis about the influence of a text summarization approach when it is integrated within a geographical information retrieval system is conducted. Both generic and geographic summaries of different compression ratios are studied, in order to decide whether summaries are beneficial as an intermediate stage of information retrieval processes. The idea was to determine the optimal summary type and size that could improve both the indexing time as well as the precision of the retrieved documents. Slight improvements were only obtained for some types of the proposed summaries ranging from 60–80% compression rate, particularly for those based on geographical information which took into account the geographic entities detected in the document collection.

Furthermore, Aker et al. (2012b) investigate the idea of **using summaries for indexing images** pertaining to geo-located entities. In this work they evaluated image retrieval effectiveness contrasting conditions when captions generated by a multi-document summarizer are used to index images and when existing image captions found in Flickr[17] are used. The generated captions were evaluated by user assessments and subjective measures. The best results were achieved when the automatically generated summaries were combined with existing keyword captions, i.e., Flickr captions. Differently, Alhindi et al. (2013) investigate the use of profile-based summarization with the goal of providing contextualization and interactive support for enterprise searches. The experimental setup relies on the hypothesis that profile-based summarization can help a user when searching for a company's

[17] https://www.flickr.com/.

documents, and guide him or her to the right documents more easily. In this work, summaries are generated for being used instead of snippets in the retrieval's result. Several summarization approaches were tested for this purpose, and the results indicate that a more personalized summary is perceived more positively by the users.

Other common task where summarization is extrinsically evaluated is within **question answering systems**. In Teufel (2001), the aim of the summaries is to describe the contribution of a scientific article with respect to previous work. Therefore, the proposed evaluation framework assesses the quality of the generated summaries asking questions about how related a specific paper is to the previous existing work. In this manner, the proposed summarizer selects the sentences to be included using machine learning techniques, and its results are compared to three baselines as well as the full text. Specifically, the baselines include: (i) a summarizer that selects random sentences; (ii) a summarizer that provides keyword lists; and (iii) a reference summary written by humans. For the evaluation, 24 users manually assessed the summaries based on a list of 5 queries (e.g., *how useful did you find the information you were given to solve this task?*). The results show that although the proposed summarizer obtained very low results when it was intrinsically evaluated, in this type of task-based evaluation, the summaries are proven to be useful. Moreover, whereas keyword lists and random sentences do not provide enough information to allow users to establish relations between the summaries and previous papers, the proposed summaries are able to provide the right information as full papers, besides not showing any statistically significant difference between the system output and the reference models when they are evaluated in the proposed question-answering framework.

In Wu et al. (2004), the usefulness of a summary for a question answering task is also evaluated. The underlying hypothesis is that a summary can help to provide an answer to a question. This summary is generated by taking into account the title of the retrieved document plus the extraction of the three most indicative sentences. The evaluation framework was designed to assess the user's interactions with the systems, as well as the user's preferences. The results obtained showed that summaries helped users to pose fewer questions, read fewer documents, and find more answers.

More recently, in Lloret et al. (2011), the COMPENDIUM summarizer was adapted and integrated with a Web-based question answering system that used search engine snippets to extract answers to specific questions. In this approach, snippets were substituted by summaries for finding the answers. The whole approach consisted of four stages: i) question analysis, ii) information retrieval, iii) summarization and iv) answer extraction. The approach was evaluated over a set of 100 factual questions, testing its performance with and without integrating the summarization module. The results showed that, by integrating the summarization stage in this question answering process, an average improvement of 30% was obtained.

The BioASQ challenge (Balikas et al. 2015) also evaluated the task of automatically **summarizing biomedical texts** as part of a question answering system. With the aim of facilitating access to biomedical literature, one of the tasks focused on providing exact answers to English questions written by biomedical experts along with a paragraph-sized summary answer. Evaluation of the paragraph-sized summaries was accomplished both manually and automatically (see (Balikas et al. 2013) for a detailed description). For the manual evaluation, experts where asked to score each summary in the interval [1-5] for the following criteria: information recall, information precision, information repetition and readability. Automatic evaluation was performed by computing ROUGE-2 and ROUGE-SU4 metrics.

It is worth mentioning that intrinsic evaluation of biomedical text summarization has primarly employed ROUGE as the evaluation framework (Reeve et al. 2007; Plaza et al. 2010; Plaza 2014), while others have used precision and recall metrics (Fiszman et al. 2009; Elhadad et al. 2005; Chen and Verma 2006; Berlanga Llavori et al. 2012). Manual evaluation is also present in several works.

Another specific scenario where automatic summarization has been evaluated is **multi-lingual and cross-lingual summarization**. The MultiLing Workshops (Giannakopoulos and Karkaletsis 2013) proposed a multi-lingual, multi-document summarization task that aimed to promote research in multi-lingual summarization. Depending on the edition, the number of languages varies. For instance, the 2013 edition included a total of 10 languages (Arabic, Czech, English, French, Greek, Hebrew, Hindi, Chinese, Romanian and Spanish). The competition consisted in two tracks. The first track aimed at building language-independent summarization algorithms to be applied to the summarization of news topics, while the second track was related to the automatic summarization evaluation of multiple language documents. The research question posed was whether an automatic measure is enough to provide a ranking of systems. Some of the methods that were presented to the workshops included AutoSummENG, MeMoG and NPowER, and the main conclusion extracted was that automatic evaluation of summaries in different languages is a very difficult task that needs to be further investigated.

In the MultiLing 2015 Workshop (Kabadjov et al. 2015), two new multi-lingual summarization tasks were proposed (Call Centre Conversation Summarization and Online Forum Summarization). The first one consisted in automatically generating summaries of spoken conversations in the form of textual synopses that should inform on the content of a conversation and might be used for browsing a large database of recordings. The second one, Online Forum Summarization task, consisted in linking online user comments in both English and Italian to the specific points within the text of the article the comment refers to, as well as to provide a set of labels for the link to capture the agreement of the comment (e.g., in favour, against) and the sentiment (e.g., positive or negative) with respect to the comment target. The evaluation was performed via crowdsourcing services, based on the IR-inspired pooling based schema used in TREC, so that links that were not proposed by any system were deemed irrelevant, and the remaining were classified according to the number of systems that proposed them and were later validated by CrowdFlower contributor.

In the last edition of MultiLing (MultiLing 2017 Giannakopoulos et al. (2017)) the same tasks as previous years were proposed, with the future intention to also organize a headline generation task. In this fora, Ellouze et al. (2017) presented a new multilingual text summary evaluation method based on machine learning on different features that aims to predict human overall responsiveness for English and Arabic summaries. Features used include: different ROUGE scores, AutoSummENG, MeMoG and NPoweR scores, SIMetrix scores, and several syntactic features such as the number of noun phrases, verbal phrases and prepositional phrases.

Regarding the task of **synthesis and reporting**, one of the first extrinsic evaluations can be found in McKeown et al. (2005), where automatic multi-document summaries generated by the Newsblaster application as news reports were evaluated to check whether they could help users' performance of an online news browsing application. The experiments conducted and the results obtained confirmed the benefit of research in multi-document summarization applied to other tasks. In particular, in this research work, a user evaluation was therefore defined, containing a set of questions that were focused on asking if the summaries produced were useful and helpful for the users. Examples of such questions were: *"Do summaries help the user find information needed to perform a report writing task?"* or *"In the context of a news browser, what is the comparison of information quality in this task, and user satisfaction, when users have access to Newsblaster summaries versus minimal or human summaries?"*. This manner, 45 English spoken subjects carried out the evaluation of summaries in the context of this application. The results obtained showed that subjects produced better quality reports using a news interface, where support with automatic summaries produced using the Newsblaster summarizer was provided. In addition, it was shown that users were also more satisfied when multi-document summaries were generated.

The task of **information synthesis** was defined in Amigo et al. (2004) as the task of extracting and organizing related pieces of information appearing in several relevant documents with the aim of generating a comprehensive, non-redundant report that meets an information need. Therefore, this task slightly differs from multi-document query-focused summarization in the sense that the reports are longer and contain more information. The evaluation of the reports was done according to different evaluation metrics, such as sentence similarity, ROUGE, document similarity, etc., based on the hypothesis that a good similarity metric would be able to distinguish between manual and automatic reports. Moreover, it was shown that the state-of-the-art metrics used for summarization did not perform equally to the generated reports, thus highlighting the issue that both tasks were not identical.

Automatic summaries have been also applied to support the task of **essay generation** that is included in student's formative processes. In this manner, automatic summaries are employed to help students write their essays. This has been studied in Labeke et al. (2013b), Labeke et al. (2013a) and Field et al. (2013), where a computer-based summarization application was developed to generate feedback on free-text essays, as a means of helping students to identify and detect patterns in the texts. In particular, the authors experimented with summarization strategies based on the extraction of key words and sentences, under the hypothesis

Table 6 Summary of content based evaluation methods

Event name	Focus
SUMMAC	Evaluation of single document summaries (news summaries)
TSC[a]	Similar to SUMMAC. However, also multi-document summarization was included.
DUC 2001 till 2007	Focus is single and multi-document summarization. Also different types of summarization was evaluated (i.e. query based, topic-oriented, cross-lingual, etc.)
TAC 2008 till to date	Similar to DUC conference but has richer tasks: slot filling; entity linking; sentiment-based; temporal-based; cross-lingual summarization of bio-medical documents; etc.
MultiLing 2011 till to date	Focus on multilingual summarization tasks: single-document; multi-document; Online Forum summarization; call center conversation summarization

[a] http://research.nii.ac.jp/ntcir/ntcir-ws3/tsc/cfp-en.html

that the quality and position of these key words and sentences could provide feedback for the users, in the sense that he or she could have an idea about how complete and well-structured the essay is. In order to verify and validate if the proposed summarization techniques could be useful for obtaining feedback, an empirical informal evaluation was conducted to know the users' perceptions and observations about the proposed system. In this evaluation, it was concluded that the system was useful for students to help seeing the main ideas of the essay and to detect whether an essay was perhaps not conveying the relevant ideas that were intended by the student. The general feel of the users was that the system would be a valuable tool for essay drafting.

Finally, it is worth mentioning the work of (Martschat and Markert 2017), where a variant of ROUGE that allows for evaluation of **timeline summarization** is presented. This novel metric takes into account both temporal and semantic similarity of daily summaries.

7 Summarization evaluation events

In the previous sections we highlighted different strategies used to perform summary quality assessment. There have been also mentions about events hosted campaigns to motivate the research on the evaluation topic. In the following table we summarize all events from the pioneering to the very recent ones (Table 6).

8 Conclusions

In this article we presented an overview of the relevant issues concerning the evaluation of summaries. The most widespread approaches for assessing either the content of summaries and their readability have been analyzed, outlining their advantages and disadvantages. In addition, we introduced recent crowdsourcing services, such as Mechanical Turk, which can be of help for evaluating summaries or collecting large amounts of data in a relatively easy, fast, and cheap way. We also summarized golden standard data available for the purpose of summary evaluation and outlined summary evaluation activities performed under specific task scenarios.

Evaluating a summary, either manually or automatically, is not a trivial issue. The manual evaluation involves human effort for determining to what extent a summary is good with respect to specific criteria (information contained, grammaticality, coherence, etc.). This is very costly and time-consuming, especially if lots of summaries have to be evaluated. In addition, the subjective nature of manual evaluation may lead to different summary results depending on the assessor, even though strict guidelines are provided to carry out the evaluation process.

Most of the evaluation methods presented rely on model summaries, that have been written by humans. Different studies suggest that, when humans had to decide the most relevant sentences from documents in order to produce summaries, they frequently disagree in which sentences best represent the content of a document (Spärck Jones 2007; McKeown et al. 2001). However, although the low agreement between humans is a problem, new evaluation methods and tools (such as the Pyramid scheme previously described) appear to offer a way to reduce disagreement.

Another problem is the semantic equivalence between different nouns, for example by means of synonymy, or expressions, when there are various ways to express the same idea. This may lead to situations where two equivalent summaries (in content) that have been written using different vocabulary, are assessed differently by the summarization method. Although traditional methods usually performed a superficial analysis and did not take into consideration the semantic meaning of phrases, more recent investigations based on methods such as BE (Basic Elements (Hovy et al. 2006)) and SCU (Summary Content Units (Nenkova and Passonneau 2004)) are actively working toward covering this gap to produce more accurate evaluations of summaries.

With respect to the readability evaluation, the methods proposed so far are still at their early stages. Ideally, this type of evaluation should be independent of the source documents, since the grammaticality, cohesion and clarity of the source do not determine those of the summary. However, this is not completely true since automatic summaries are mostly generated by extract-based summarization systems. Even humans frequently take some sentences or phrases verbatim from the source when writing a summary (Banko and Vanderwende 2004) and only apply compression and rewritten when the desired compression rate is too high. Moreover, evaluation results in DUC and TAC competitions have shown a high correlation between linguistic quality and content-based measurements such as responsiveness (Conroy and Dang 2008b). Fortunately, during the last years, however, research in summarization evaluation is paying increasing attention to readability issues such as grammaticality and focus (Vadlapudi and Katragadda 2010b), and cohesion (Lapata and Barzilay 2005b).

Crowdsourcing services, such as MTurk, can be used for evaluating a summary; for instance, asking humans either to write model summaries, or evaluate existing summaries. Although they can provide fast and relative inexpensive mechanisms to carry out tasks that are simple for humans but very difficult for computers and require a lot of human effort, there are also some disadvantages related to these services. Some issues concerning the quality of the task performed by the turkers arise, since some turkers will be probably enrolled in a task only for the money, providing non-sense answers in order to decrease the time they spend with the task, but at the same time, increasing their rate of pay, being able to finish more tasks. Regarding this, research on how to account for the quality of the results provided by these services, as well as methods for ensuring such quality, would be needed.

Despite the considerable progress in the evaluation of summaries in recent years, there is still a lot of room for improvement. On the one hand, most content-oriented evaluation tools are based on content overlap, which presents a bias toward lexical similarity that may lead to unfair penalties when abstractive summaries are evaluated. However, it is interesting to mention that, in spite of this a priori disadvantageous situation, abstractive human summaries usually get significantly higher ROUGE scores (Liu and Liu 2008a).

On the other hand, the inherent subjectivity associated to the evaluation process poses greater challenges to this research subfield. State-of-the-art approaches mainly focus on intrinsic evaluation, in particular, in novel methods to assess either a summary content or its quality. To fully automate this process is very difficult, and for this reason new research about this topic can be considered as emerging research. However, as long as semantic methods improve (for instance, with distributional semantics methods, such as word, sentence, paragraph or document embedding representations), to account for equivalent expressions will be more feasible. In this manner, approaches will not rely on model summaries as much as they currently do for evaluating the informativeness of a summary. Furthermore, to be aware of the techniques and approaches existing from other research areas, such as essay scoring, can also help to achieve improvements in the evaluation of summaries. Similar to what occurred with ROUGE, which was inspired in a method for evaluating machine translation systems (i.e., BLEU), the success of the techniques employed for scoring essays could be perfectly investigated for evaluating also text summaries.

As the main conclusion drawn from this research and in order to provide some guidance for users interested in summary evaluation, it is worth noting that, althoug ROUGE is still the most common tool used for content evaluation, that not means it is the unique and best one. It would be essential to use other methods as well, such as AutoSummENG, MeMoG, FRESA or the Trivergent model, that have been shown good correlation with human evaluations. Depending on the setting and the type of summarization addressed, researchers should have to decide hich method would be more appropriate, without forgetting that at this moment, all these methods complement (and not replace) human evaluation, so manual evaluation to assess the quality of generated summaries should be also necessary. This would help advance the state of the art in summarization and summarization evaluation, derived from the results and analysis obtained from different setting and comparison benchmarks. Regarding readability assessment, most research works take as a basis the linguistic evaluation carried out in DUC/TAC conferences, and finally, as far as the extrinsic evaluation is concerned, to assess the influence of automatic summaries within the context of a specific task would be the type of evaluation that is less costly and guarantees that the summaries are suitable for that task, without having to inspect its content and linguistic quality.

Finally, given the inherent subjectivity involved in the summarization tasks and the variety of settings and types, which makes its evaluation a very challenging problem, the new evaluation methods should take into account more pragmatic issues, for instance, the purpose for which the summary was produced, who the intended user would be (e.g. a human or maybe an automatic process) as well as the user satisfaction in case the final user is a person.

Acknowledgements This research is partially funded by the European Commission under the Seventh (FP7-2007-2013) Framework Programme for Research and Technological Development through the SAM (FP7-611312) project; by the Spanish Government through the projects VoxPopuli (TIN2013-47090-C3-1-P), Vemodalen (TIN2015-71785-R), RESCATA (TIN2015-65100-R) and REDES (TIN2015-65136-C2-2-R), the Generalitat Valenciana through project DIIM2.0 (PROMETEOII/2014/001), and the Universidad Nacional de Educación a Distancia through the project "Modelado y síntesis automática de opiniones de usuario en redes sociales" (2014-001-UNED-PROY).

References

Aker, A., El-Haj, M., Albakour, M.-D., & Kruschwitz, U. (2012a). Assessing crowdsourcing quality through objective tasks. In *Proceedings of the eighth international conference on language resources and evaluation (LREC-2012)*. European Language Resources Association (ELRA), Istanbul, Turkey (pp. 1456–61).

Aker, A., Fan, X., Sanderson, M., & Gaizauskas, R. (2012b). Investigating summarization techniques for geo-tagged image indexing. In *Advances in information retrieval: 34th European conference on information retrieval (ECIR), Barcelona, Spain* (pp. 472–75).

Aker, A., & Gaizauskas, R. (2010). Model summaries for location-related images. In *Proceedings of the 7th language resources and evaluation conference*.

Alhindi, A., Kruschwitz, U., & Fox, C. (2013). A pilot study on using profile-based summarisation for interactive search assistance. In P. Serdyukov, P. Braslavski, S. Kuznetsov, J. Kamps, S. Rger, E. Agichtein, I. Segalovich & E. Yilmaz, E. (Eds.), *Advances in information retrieval. Vol. 7814 of Lecture Notes in Computer Science, Springer, Berlin* (pp. 672–75). doi:10.1007/978-3-642-36973-5_57.

Amigo, E., Gonzalo, J., Peinado, V., Peñas, A., & Verdejo, F. (2004). An empirical study of information synthesis task. In *Proceedings of the 42nd meeting of the association for computational linguistics (ACL'04), Main Volume, Barcelona, Spain* (pp. 207–14).

Balikas, G., Krithara, A., Partalas, I., & Paliouras, G. (2015). Bioasq: A challenge on large-scale biomedical semantic indexing and question-answering. In *Multimodal retrieval in the medical domain, Workshop at ECIR*.

Balikas, G., Partalas, I., Kosmopoulos, A., Petridis, S., Malakasiotis, P., & Pavlopoulos, I., et al. (2013). *Bioasq evaluation framework specifications. Project deliverable D4.1*. http://bioasq.org/sites/default/files/PublicDocuments/BioASQ_D4.1-EvaluationFrameworkSpecification_final.pdf.

Bamman, D., O'Connor, B., & Smith, N. A. (2013). Learning latent personas of film characters. In *ACL (1)*. The Association for Computer Linguistics (pp. 352–61).

Banko, M., & Vanderwende, L. (2004). Using n-grams to understand the nature of summaries. In *Proceedings of HLT-NAACL 2004: Short Papers*. HLT-NAACL-Short '04. Association for Computational Linguistics, Stroudsburg, PA, USA (pp. 1–4). http://dl.acm.org/citation.cfm?id=1613984.1613985.

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. 141–48).

Barzilay, R., & Lapata, M. (2008). Modeling local coherence: An entity-based approach. *Computational Linguistics*, *34*(1), 1–34.

Berlanga Llavori, R., Ramírez Cruz, Y., & Gil García, R. (2012). A framework for obtaining structurally complex condensed representations of document sets in the biomedical domain. *Procesamiento del Lenguaje Natural*, *49*, 21–8.

Branny, E. (2007). Automatic summary evaluation based on text grammars. *Journal of Digital Information*, *8*(3).

Cabrera-Diego, L. A., Torres-Moreno, J., & Durette, B. (2016). Evaluating multiple summaries without human models: A first experiment with a trivergent model. In *Natural language processing and information systems—21st international conference on applications of natural language to information systems, NLDB 2016, Salford, UK*, June 22–24, 2016, Proceedings (pp. 91–101).

Callison-Burch, C. (2009). Fast, cheap, and creative: evaluating translation quality using Amazon's mechanical turk. In *Proceedings of the 2009 conference on empirical methods in natural language processing* (pp. 286–95).

Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kraaij, W., Kronenthal, M., Lathoud, G., Lincoln, M., Lisowska, A., McCowan, I., Post, W., Reidsma, D., & Wellner, P. (2005). The AMI meeting corpus. In L. P. J. J. Noldus, F. Grieco, L. W. S. Loijens & P. H. Zimmerman (Eds.), *Proceedings of the measuring behavior 2005 symposium on "annotating and measuring meeting behavior"*.

Chen, P., & Verma, R. (2006). A query-based medical information summarization system using ontology knowledge. In *Proceedings of the IEEE symposium on computer-based medical systems* (pp. 37–42).

Christensen, J., Mausam, S. S., Soderland, S., & Etzioni, O. (2013). Towards coherent multi-document summarization. In *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: Human language technologies*. Association for Computational Linguistics, Atlanta, Georgia (pp. 1163–1173). http://www.aclweb.org/anthology/N13-1136.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, *12*, 2493–2537.

Conroy, J. M., & Dang, H. T. (2008a). 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)*. Coling 2008 Organizing Committee, Manchester, UK (pp. 145–52).

Conroy, J. M., & Dang, H. T. (2008b). Mind the gap: Dangers of divorcing evaluations of summary content from linguistic quality. In *Proceedings of the 22nd international conference on computational linguistics—Volume 1*. COLING '08. Association for Computational Linguistics, Stroudsburg, PA, USA (pp. 145–152). http://dl.acm.org/citation.cfm?id=1599081.1599100.

Conroy, J. M., Schlesinger, J. D., Kubina, J., Rankel, P. A., & O'Leary, D. P. (2011). CLASSY 2011 at TAC: Guided and multi-lingual summaries and evaluation metrics. In *Proceedings of the 2011 text analysis conference (TAC 2011)*.

Conroy, J. M., Schlesinger, J. D., Rankel, P. A., & O'Leary, D. P. (2010). Guiding CLASSY toward more responsive summaries. In *Proceedings of the 2010 text analysis conference (TAC 2010)*.

Dalianis, H., & Hassel, M. (2001). *Development of a Swedish corpus for evaluating summarizers and other IR-tools*. Technical report TRITA-NAP0112, IPLab-188, NADA, KTH.

Dang, H. T. (2005). Overview of DUC 2005. In *Proceedings of the document understanding conference (DUC)*.

Dang, H. T. (2006). Overview of DUC 2006. In *Proceedings of the document understanding conference (DUC)*.

Donaway, R. L., Drummey, K. W., & Mather, L. A. (2000). A comparison of rankings produced by summarization evaluation measures. In *Proceedings of NAACL-ANLP 2000 workshop on automatic summarization* (pp. 69–78).

Dong, Z., & Dong, Q. (2003). HowNet—A hybrid language and knowledge resource. In *Proceedings of natural language processing and knowledge engineering conference* (pp. 820–24).

Edmundson, H. P. (1969). New methods in automatic extracting. *Journal of the ACM (JACM)*, *16*(2), 264–85.

El-Haj, M., Kruschwitz, U., & Fox, C. (2010). Using mechanical turk to create a corpus of arabic summaries. In *Proceedings of the seventh conference on international language resources and evaluation, Valletta, Malta*.

Elhadad, N., Kan, M. Y., Klavans, J. L., & McKeown, K. R. (2005). Customization in a unified framework for summarizing medical literature. *Artificial Intelligence in Medicine*, *33*(2), 179–198. doi:10.1016/j.artmed.2004.07.018.

Ellouze, S., Jaoua, M., & Belguith, L. H. (2016). Automatic evaluation of a summary's linguistic quality. In *Natural language processing and information systems—21st international conference on applications of natural language to information systems, NLDB 2016, Salford, UK*, June 22–24, 2016, Proceedings (pp. 392–400).

Ellouze, S., Jaoua, M., & Hadrich Belguith, L. (2017). Machine learning approach to evaluate multilingual summaries. In *Proceedings of the MultiLing 2017 workshop on summarization and summary evaluation across source types and genres*. Association for Computational Linguistics (pp. 47–54).

Feng, D., Besana, S., & Zajac, R. (2009). Acquiring high quality non-expert knowledge from on-demand workforce. In *Proceedings of the 2009 workshop on the people's web meets NLP: Collaboratively constructed semantic resources*. People's Web '09. Association for Computational Linguistics, Morristown, NJ, USA (pp. 51–6). http://portal.acm.org/citation.cfm?id=1699765.1699773.

Field, D., Pulman, S., Van Labeke, N., Whitelock, D., & Richardson, J. (2013). Did I really mean that? Applying automatic summarisation techniques to formative feedback. In *Proceedings of the international conference recent advances in natural language processing RANLP 2013*. INCOMA Ltd. Shoumen, BULGARIA, Hissar, Bulgaria (pp. 277–84). http://www.aclweb.org/anthology/R13-1036.

Fiori, A. (2014). Innovative document summarization techniques: Revolutionizing knowledge understanding: Revolutionizing knowledge understanding. In *Advances in data mining and database management: IGI Global.*

Fiszman, M., Demner-Fushman, D., Kilicoglu, H., & Rindflesch, T. C. (2009). Automatic summarization of medline citations for evidence-based medical treatment: A topic-oriented evaluation. *Journal of Biomedical Informatics*, *42*(5), 801–813. doi:10.1016/j.jbi.2008.10.002.

Gambhir, M., & Gupta, V. (2017). Recent automatic text summarization techniques: A survey. *Artificial Intelligence Review*, *47*(1), 1–66. doi:10.1007/s10462-016-9475-9.

Giannakopoulos, G., Conroy, J., Kubina, J., Rankel, P. A., Lloret, E., Steinberger, J., Litvak, M., & Favre, B. (2017). Multiling 2017 overview. In *Proceedings of the MultiLing 2017 workshop on summarization and summary evaluation across source types and genres*. Association for Computational Linguistics, Valencia, Spain (pp. 1–6). http://www.aclweb.org/anthology/W17-1001.

Giannakopoulos, G., & Karkaletsis, V. (2011a). AutoSummENG and MeMoG in evaluating guided summaries. In *Proceedings of the 2011 text analysis conference (TAC 2011).*

Giannakopoulos, G., & Karkaletsis, V. (2011b). Autosummeng and memog in evaluating guided summaries. In *Proceedings of the text analysis conference (TAC 2011), Gaithersburg, Maryland, USA*.

Giannakopoulos, G., & Karkaletsis, V. (2013). Together we stand npower-ed. In *Proceedings of CICLing 2013, Karlovasi, Samos, Greece*.

Giannakopoulos, G., Karkaletsis, V., Vouros, G., & Stamatopoulos, P. (2008). Summarization system evaluation revisited: N-gram graphs. *ACM Transactions on Speech and Language Processing*, *5*(3), 1–39.

Gillick, D., & Liu, Y. (2010). Non-expert evaluation of summarization systems is Risky. In *Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon's mechanical turk*.

Grosz, B. J., Weinstein, S., & Joshi, A. K. (1995). Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, *21*(2), 203–25.

Hand, T. (1997). A proposal for task-based evaluation of text summarization systems. In *Proceedings of the association for computational linguistics conference, Madrid, Spain* (pp. 31–38).

Harnly, A., Nenkova, A., Passonneau, R. J., & Rambow, O. (2015). Automatation of summary evaluation by the pyramid method. In *Proceedings of the international conference recent advances in natural language processing (RANLP), Borovets, Bulgaria* (pp. 226–232).

Hasler, L. (2008). Centering theory for evaluation of coherence in computer-aided summaries. In *Proceedings of the sixth international conference on language resources and evaluation.*

Hasler, L., Orăsan, C., & Mitkov, R. (2003). Building better corpora for summarization. In *Proceedings of corpus linguistics 2003, Lancaster, UK* (pp. 309–19).

Hassel, M. (2004). *Evaluation of automatic text summarization: A practical implementation.*

He, T., Chen, J., Ma, L., Gui, Z., Li, F., Shao, W., & Wang, Q. (2008). ROUGE-C: A fully automated evaluation method for multi-document summarization, Granular Computing, 2008. GrC 2008. In *IEEE international conference on* (pp. 269–74).

Hong, K., Conroy, J., Favre, B., Kulesza, A., Lin, H., & Nenkova, A. (2014). A repository of state of the art and competitive baseline summaries for generic news summarization. In N. C. C. Chair, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the ninth international conference on language resources and evaluation (LREC'14)*. European Language Resources Association (ELRA), Reykjavik, Iceland.

Hovy, E. (2005). *The Oxford handbook of computational linguistics*. Oxford University Press, Ch. Text Summarization (pp. 583–98).

Hovy, E., Lin, C.-Y., Zhou, L., & Fukumoto, J. (2006). Automated summarization evaluation with basic elements. In *Proceedings of the 5th international conference on language resources and evaluation*.

Jimeno-Yepes, A. J., Plaza, L., Mork, J. G., Aronson, A. R., & Díaz, A. (2013). MeSH indexing based on automatically generated summaries. *BMC Bioinformatics*, *14*, 208.

Jing, H., Barzilay, R., McKeown, K. & Elhadad, M. (1998). Summarization evaluation methods: Experiments and analysis. In *AAAI symposium on intelligent summarization* (pp. 51–9).

Kabadjov, M., Steinberger, J., Barker, E., Kruschwitz, U., & Poesio, M. (2015). Onforums: The shared task on online forum summarisation at multiling'15. In *Proceedings of the 7th forum for information retrieval evaluation, FIRE '15. ACM, New York, NY, USA* (pp. 21–26). doi:10.1145/2838706.2838709.

Katragadda, R. (2010). GEMS: Generative modeling for evaluation of summaries. In *Proceedings of the 11th international conference on computational linguistics and intelligent text processing* (pp. 724–35).

Khan, A., Salim, N., & Kumar, Y. J. (2015). A framework for multi-document abstractive summarization based on semantic role labelling. *Applied Soft Computing*, *30*, 737–747.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., & et al. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the acl on interactive poster and demonstration sessions*. Association for Computational Linguistics (pp. 177–80).

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

Labeke, N. V., Whitelock, D., Field, D., Pulman, S., & Richardson, J. (2013a). What is my essay really saying? Using extractive summarization to motivate reflection and redrafting. In *Proceedings of the workshops at the 16th international conference on artificial intelligence in education AIED 2013, Memphis, USA, July 9–13. Vol. 1009 of CEUR workshop proceedings*. CEUR-WS.org.

Labeke, N. V., Whitelock, D., Field, D., Pulman, S., & Richardson, J. T. E. (2013b). OpenEssayist: extractive summarisation and formative assessment of free-text essays. In *1st international workshop on discourse-centric learning analytics. A pre conference workshop at LAK13*. http://oro.open.ac.uk/37548/.

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

Lapata, M., & Barzilay, R. (2005b). Automatic evaluation of text coherence: Models and representations. In *Proceedings of the 19th international joint conference on artificial intelligence*. IJCAI'05. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (pp. 1085–1090). http://dl.acm.org/citation.cfm?id=1642293.1642467.

Lin, C.-Y. (2001). *Summary evaluation environment*. http://www.isi.edu/~cyl/SEE.

Lin, C.-Y. (2004a). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop*. Association for Computational Linguistics, Barcelona, Spain (pp. 74–81).

Lin, C.-Y. (2004b). ROUGE: A package for automatic evaluation of summaries. In *Proceedings of association of computational linguistics text summarization workshop* (pp. 74–81).

Lin, C.-Y., & Hovy, E. (2002). Manual and automatic evaluation of summaries. In *Proceedings of the workshop on automatic summarization post conference workshop of ACL-02 (DUC 2002)*.

Lin, C.-Y., & Hovy, E. (2003). Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 conference of the North American chapter of the association for* *computational linguistics on human language technology—Volume 1. NAACL '03*. Association for Computational Linguistics, Stroudsburg, PA, USA (pp. 71–78). doi:10.3115/1073445.1073465

Lin, Z., Liu, C., Ng, H. T., & Kan, M.-Y. (2012). Combining coherence models and machine translation evaluation metrics for summarization evaluation. In *Proceedings of the 50th annual meeting of the association for computational linguistics: Long papers—Volume 1*. Association for Computational Linguistics (pp. 1006–1014).

Liseth, A. (2004). *En evaluering av NorSum en automatisk tekstsammenfatter for norsk*. Hovedfagsoppgave. Technical report: Universitetet i Bergen, Seksjon for lingvistiske fag.

Liu, F., & Liu, Y. (2008a). Correlation between rouge and human evaluation of extractive meeting summaries. In *Proceedings of the 46th annual meeting of the association for computational linguistics on human language technologies: Short papers. HLT-Short '08*. Association for Computational Linguistics, Stroudsburg, PA, USA (pp. 201–204). http://dl.acm.org/citation.cfm?id=1557690.1557747

Liu, F., & Liu, Y. (2008b). Correlation between ROUGE and human evaluation of extractive meeting summaries. In *Proceedings of the 46th annual meeting of the association of computational linguistics: Human language technologies, short papers* (pp. 201–4).

Lloret, E., Llorens, H., Moreda, P., Saquete, E., & Palomar, M. (2011). Text summarization contribution to semantic question answering: New approaches for finding answers on the web. *International Journal of Intelligent Systems*, *26*(12), 1125–52.

Lloret, E., & Palomar, M. (2012). Text summarisation in progress: A literature review. *Artificial Intelligence Review*, *37*(1), 1–41. doi:10.1007/s10462-011-9216-z.

Lloret, E., Plaza, L., & Aker, A. (2013). Analyzing the capabilities of crowdsourcing services for text summarization. *Language Resources and Evaluation*, *47*(2), 337–69. doi:10.1007/s10579-012-9198-8.

Louis, A., & Nenkova, A. (2008). Automatic summary evaluation without human models. In *Proceedings of the text analysing conference, (TAC 2008)*.

Louis, A., & Nenkova, A. (2009a). Automatically evaluating content selection in summarization without human models. In *Proceedings of the 2009 conference on empirical methods in natural language processing: Volume 1*. Association for Computational Linguistics (pp. 306–314).

Louis, A., & Nenkova, A. (2009b). Predicting summary quality using limited human input. In *Proceedings of the 2009 text analysis conference (TAC 2009)*.

Mani, I. (2001). *Automatic summarization* (Vol. 3). Amsterdam: John Benjamins Publishing Company.

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*. Association for Computational Linguistics (pp. 77–85).

Marcu, D. (1997). From discourse structures to text summaries. In *Proceedings of the ACL. Vol. 97* (pp. 82–88).

Martschat, S., & Markert, K. (2017). Improving rouge for timeline summarization. In *Proceedings of the 15th conference of the european chapter of the association for computational linguistics: Volume 2, short papers*. Association for Computational Linguistics (pp. 285–290).

Mason, W., & Watts, D. J. (2010). Financial incentives and the "performance of crowds". *ACM SigKDD Explorations Newsletter*, *11*, 100–8.

McKeown, K., Barzilay, R., Evans, D., Hatzivassiloglou, V., Kan, M. Y., Schiffman, B., & Teufel, S. (2001). Columbia multi-document summarisation: Approach and evaluation. In *Proceedings of the DUC 2001*.

McKeown, K., Passonneau, R., Elson, D., Nenkova, A., & Hirschberg, J. (2005). Do summaries help? A task-based evaluation of multi-document summarization. In *28th annual ACM SIGIR conference on research and development in information retrieval, ACM, Salvador, Brazil* (pp. 210–17).

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

Nenkova, A., & McKeown, K. (2011). Automatic summarization. *Foundations and Trends in Information Retrieval*, *5*(2–3), 103–233. doi:10.1561/1500000015.

Nenkova, A., & Passonneau, R. (2004). Evaluating content selection in summarization: The pyramid method. In *HLT-NAACL 2004: Main Proceedings* (pp. 145–52). Association for Computational Linguistics, Boston, Massachusetts, USA.

Nenkova, A., Passonneau, R., & McKeown, K. (2007). The pyramid method: Incorporating human content selection variation in summarization evaluation. *ACM Transactions on Speech and Language Processing (TSLP)*, *4*(2).

Ng, J.-P., & Abrecht, V. (2015a). Better summarization evaluation with word embeddings for rouge. In *Proceedings of the 2015 conference on empirical methods in natural language processing*. Association for Computational Linguistics, Lisbon, Portugal (pp. 1925–1930).

Ng, J.-P., & Abrecht, V. (2015b). Better summarization evaluation with word embeddings for rouge. In *Proceedings of the 2015 conference on empirical methods in natural language processing*. Association for Computational Linguistics, Lisbon, Portugal (pp. 1925–1930). http://aclweb.org/anthology/D15-1222.

Ono, K., Sumita, K., & Miike, S. (1994). Abstract generation based on rhetorical structure extraction. In *Proceedings of the 15th conference on Computational linguistics—Volume 1*. Association for Computational Linguistics (pp. 344–48).

Over, P., & Liggett, W. (2002). Introduction to DUC: An intrinsic evaluation of generic news text summarization systems. In *Proceedings of DUC 2002*.

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. 190–98).

Owczarzak, K., Conroy, J. M., Dang, H. T., & Nenkova, A. (2012a). An assessment of the accuracy of automatic evaluation in summarization. In *Proceedings of workshop on evaluation metrics and system comparison for automatic summarization*. Association for Computational Linguistics, Stroudsburg, PA, USA (pp. 1–9).

Owczarzak, K., Conroy, J. M., Dang, H. T., & Nenkova, A. (2012b). An assessment of the accuracy of automatic evaluation in summarization. In *Proceedings of workshop on evaluation metrics and system comparison for automatic summarization*. Association for Computational Linguistics, Stroudsburg, PA, USA (pp. 1–9). http://dl.acm.org/citation.cfm?id=2391258.2391259.

Owczarzak, K., & Dang, H. T. (2011). Overview of the TAC 2011 summarization track: Guided task and AESOP task. In *Proceedings of the text analysis conference (TAC)*.

Paice, C. D. (1990). Constructing literature abstracts by computer: Techniques and prospects. *Information Processing & Management*, *26*(1), 171–86.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (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).

Passonneau, R. J., Chen, E., Guo, W., & Perin, D. (2013). Automated pyramid scoring of summaries using distributional semantics. In *Proceedings of the 51st annual meeting of the association for computational linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Sofia, Bulgaria (pp. 143–147).

Passonneau, R. J., Nenkova, A., McKeown, K., & Sigelman, S. (2005). Applying the pyramid method in DUC 2005. In *Proceedings of the document understanding conference (DUC 05), Vancouver, BC, Canada*.

Perea-Ortega, J. M., Lloret, E., Ureña López, A., & Palomar, M. (2013). Application of text summarization techniques to the geographical information retrieval task. *Expert Systems with Applications*, *40*(8), 2966–74. doi:10.1016/j.eswa.2012.12.012.

Pitler, E., Louis, A., & Nenkova, A. (2010). Automatic evaluation of linguistic quality in multi-document summarization. In *Proceedings of the 48th annual meeting of the association for computational linguistics. ACL '10*. Association for Computational Linguistics, Stroudsburg, PA, USA (pp. 544–554). http://dl.acm.org/citation.cfm?id=1858681.1858737.

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. 186–95).

Plaza, L. (2014). Comparing different knowledge sources for the automatic summarization of biomedical literature. *Journal of Biomedical Informatics*, *52*, 319–328, special Section: Methods in clinical research informatics. http://www.sciencedirect.com/science/article/pii/S1532046414001610.

Plaza, L., Stevenson, M., & Díaz, A. (2010). Improving summarization of biomedical documents using word sense disambiguation. In *Proceedings of the 2010 workshop on biomedical natural language processing. BioNLP '10*. Association for Computational Linguistics, Stroudsburg, PA, USA (pp. 55–63). http://dl.acm.org/citation.cfm?id=1869961.1869968.

Radev, D. R. (2001). Experiments in single and multidocument summarization using mead. In *First document understanding conference (DUC 2001)*.

Radev, D. R., & Tam, D. (2003). Summarization evaluation using relative utility. In *CIKM '03: Proceedings of the 12th international conference on information and knowledge management* (pp. 508–11).

Rankel, P., Conroy, J. M., Slud, E. V., & O'Leary, D. P. (2011). Ranking human and machine summarization systems. In *Proceedings of the conference on empirical methods in natural language processing. EMNLP '11*. Association for Computational Linguistics, Stroudsburg, PA, USA (pp. 467–473).

Rankel, P. A., Conroy, J. M., Dang, H. T., & Nenkova, A. (2013). A decade of automatic content evaluation of news summaries: Reassessing the state of the art. In *Proceedings of the 51st annual meeting of the association for computational linguistics, ACL 2013, 4–9 August 2013, Sofia, Bulgaria, Volume 2: Short Papers* (pp. 131–136).

Rankel, P. A., Conroy, J. M., & Schlesinger, J. D. (2012). Better metrics to automatically predict the quality of a text summary. *Algorithms*, *5*(4), 398. http://www.mdpi.com/1999-4893/5/4/398.

Reeve, L. H., Han, H., & Brooks, A. D. (2007). The use of domain-specific concepts in biomedical text summarization. *Information Processing & Management*, *43*(6), 1765–1776, text summarization. http://www.sciencedirect.com/science/article/pii/S030645730700074X.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of international joint conferences on artificial intelligence (IJCAI), Montreal, Canada* (pp. 448–53).

Saggion, H., & Lapalme, G. (2000). Selective analysis for automatic abstracting: Evaluating indicativeness and acceptability. In *Proceedings of content-based multimedia information access* (pp. 747–64).

Saggion, H., & Szasz, S. (2012). The CONCISUS corpus of event summaries. In N. Calzolari, K. Choukri, T. Declerck, M. U. Dogan, B. Maegaard, J. Mariani, J. Odijk & S. Piperidis (Eds.), *LREC*. European Language Resources Association (ELRA) (pp. 2031–37).

Saggion, H., Teufel, S., Radev, D., & Lam, W. (2002). Meta-evaluation of summaries in a cross-lingual environment using content-based metrics. In *Proceedings of the 19th international conference on Computational linguistics* (pp. 1–7).

Saggion, H., Torres-Moreno, J., da Cunha, I., SanJuan, E., & Velázquez-Morales, P. (2010). Multilingual summarization evaluation without human models. In *COLING 2010, 23rd international conference on computational linguistics, posters volume, 23–27 August 2010, Beijing, China* (pp. 1059–1067).

Salton, G., Singhal, A., Mitra, M., & Buckley, C. (1997). Automatic text structuring and summarization. *Information Processing & Management.*, *33*, 193–207.

Schlesinger, J. D., O'Leary, D. P., & Conroy, J. M. (2008). *Arabic/English multi-document summarization with CLASSY—The past and the future, Springer, Berlin* (pp. 568–581). 10.1007/978-3-540-78135-6_49.

Schluter, N. (2017). The limits of automatic summarisation according to rouge. In *Proceedings of the 15th conference of the european chapter of the association for computational linguistics: Volume 2, short papers*. Association for Computational Linguistics (pp. 41–45).

Sjöbergh, J. (2007). Older versions of the ROUGEeval summarization evaluation system were easier to fool. *Information Processing & Management*, *43*(6), 1500–5.

Smith, C., Danielsson, H., & Jönsson, A. (2012). A more cohesive summarizer. In *COLING 2012, 24th international conference on computational linguistics, proceedings of the conference: Posters, 8–15 December 2012, Mumbai, India* (pp. 1161–1170).

Spärck Jones, K. (2007). Automatic summarising: The state of the art. *Information Processing & Management*, *43*(6), 1449–1481. doi:10.1016/j.ipm.2007.03.009.

Sparck Jones, K., & Galliers, J. (1996). *Evaluating natural language processing systems (an analysis and review)*. In Lecture Notes in Computer Science, Springer.

Steinberger, J., Kabadjov, M., Pouliquen, B., Steinberger, R., & Poesio, M. (2009). WB-JRC-UT's participation in TAC 2009: Update summarization and AESOP tasks. In *Proceedings of the 2009 text analysis conference (TAC 2009)*.

Stewart, J. (2008). *Genre oriented summarization*. Ph.D. thesis. https://www.lti.cs.cmu.edu/sites/default/files/research/thesis/2008/jade_goldstein_stewart_genre_oriented_summarization.pdf

Tang, J., & Sanderson, M. (2010). Evaluation and user preference study on spatial diversity. In *Proceedings of the 32nd European conference on information retrieval (ECIR)*.

Teufel, S. (2001). Task-based evaluation of summary quality: Describing relationships between scientific papers. In *Workshop automatic summarization, NAACL* (pp. 12–21).

Teufel, S., & van Halteren, H. (2004). Evaluating information content by factoid analysis: Human annotation and stability. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 419–26).

Tombros, A., & Sanderson, M. (1998). Advantages of query biased summaries in information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval, ACM, New York, NY, USA* (pp. 2–10).

Torres-Moreno, J. (2011). Résumé automatique de documents. Recherche d'information et web. Hermes Science Publications. https://books.google.es/books?id=9HeLsuRFRJMC.

Torres-Moreno, J. (2014). Automatic Text Summarization. Cognitive science and knowledge management series. Wiley. https://books.google.es/books?id=aPHsBQAAQBAJ.

Torres-Moreno, J., Saggion, H., da Cunha, I., SanJuan, E., & Velázquez-Morales, P. (2010a). Summary evaluation with and without references. *Polibits: Research Journal on Computer Science and Computer Engineering with Applications*, *42*, 13–19.

Torres-Moreno, J., Saggion, H., da Cunha, I., Velázquez-Morales, P., & SanJuan, E. (2010b). Evaluation automatique de résumés avec et sans références. In *TALN'10, Montréal, Canada*.

Tratz, S., & Hovy, E. (2008). Summarization evaluation using transformed basic elements. In *Proceedings of the 1st text analysis conference*.

Turchi, M., Steinberger, J., Kabadjov, M., & Steinberger, R. (2010). Using parallel corpora for multilingual (multi-document) summarisation evaluation. In *Multilingual and multimodal information access evaluation. Vol. 6360 of Lecture Notes in Computer Science* (pp. 52–63).

Ulrich, J., Murray, G., & Carenini, G. (2008). A publicly available annotated corpus for supervised email summarization. In *AAAI08 EMAIL Workshop, AAAI, Chicago, USA*.

Vadlapudi, R., & Katragadda, R. (2010a). Quantitative evaluation of grammaticality of summaries. In *Proceedings of the 11th international conference on computational linguistics and intelligent text processing, CICLing 2010, Iasi, Romania* (pp. 736–47).

Vadlapudi, R., & Katragadda, R. (2010b). On automated evaluation of readability of summaries: Capturing grammaticality, focus, structure and coherence. In *Proceedings of the NAACL HLT 2010 student research workshop. HLT-SRWS '10*. Association for Computational Linguistics, Stroudsburg, PA, USA (pp. 7–12). http://dl.acm.org/citation.cfm?id=1858146.1858148.

Vadlapudi, R., & Katragadda, R. (2010c). 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).

Van Dijk, T. (1972). *Some aspects of text grammars. A study in theoretical linguistics and poetics*. Paris, Mouton: The Hague.

Voorhees, E. (2003). Overview of the TREC 2003 question answering track. In *Proceedings of the twelfth text retrieval conference (TREC)*.

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. 82–7).

Wang, X., Evanini, K., & Zechner, K. (2013). Coherence modeling for the automated assessment of spontaneous spoken responses. In *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: Human language technologies*. Association for Computational Linguistics, Atlanta, Georgia (pp. 814–819). http://www.aclweb.org/anthology/N13-1101.

Wu, M., Wilkinson, R., & Paris, C. (2004). An evaluation on query-biased summarisation for the question answering task. In *Proceedings of the Australasian language technology workshop 2004, Sydney, Australia* (pp. 32–8). http://www.aclweb.org/anthology/U/U04/U04-1005.

Yin, W., & Schütze, H. (2015). Discriminative phrase embedding for paraphrase identification. In *Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics: Human language technologies*. Association for Computational Linguistics, Denver, Colorado (pp. 1368–1373).

Zhou, L., Lin, C.-Y., Munteanu, D. S., & 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. 447–54).

Zhu, X., & Cimino, J. J. (2013). Clinicians' evaluation of computer-assisted medication summarization of electronic medical records. *Computers in Biology and Medicine*, *59*, 221–231.