A Survey on evaluation of summarization methods

Liana Ermakova, Jean-Valère Cossu, Josiane Mothe

A Survey on Evaluation of Summarization Methods

Liana Ermakova

HCTI – EA 4249, Université de Bretagne Occidentale, Brest, France

Jean Valère Cossu

My Local Influence, Aubagne, France

Josiane Mothe

IRIT, UMR5505 CNRS, ESPE, Univ. de Toulouse, France

Abstract

The increasing volume of textual information on any topic requires its compression to allow humans to digest it. This implies detecting the most important information and condensing it. These challenges have led to new developments in the area of Natural Language Processing (NLP) and Information Retrieval (IR) such as narrative summarization and evaluation methodologies for narrative extraction. Despite some progress over recent years with several solutions for information extraction and text summarization, the problems of generating consistent narrative summaries and evaluating them are still unresolved. With regard to evaluation, manual assessment is expensive, subjective and not applicable in real time or to large collections. Moreover, it does not provide re-usable benchmarks. Nevertheless, commonly used metrics for summary evaluation still imply substantial human effort since they require a comparison of candidate summaries with a set of reference summaries. The contributions of this paper are three-fold. First, we provide a comprehensive overview of existing metrics for summary evaluation. We discuss several limitations of existing frameworks for summary evaluation. Second, we introduce an automatic framework for the evaluation of metrics that does not require any human annotation. Finally, we evaluate the existing assessment metrics on a Wikipedia data set and a collection of scientific articles using this framework. Our findings show that the majority of existing metrics based on vocabulary overlap are not suitable for assessment based on comparison with a full text and we discuss this outcome.

Keywords: automatic summarization, text compression, evaluation campaigns, assessment metrics, extraction, extractive summarization, ROUGE

Highlights

1. Manual assessment is not re-usable
2. Re-use of the gold standard by non-participants is often problematic
3. Overlap-based metrics are not suitable for full text comparison-based evaluation
4. GRAD exceeds word-based metrics to distinguish between generated and human written summaries
5. Overlap metrics and GRAD can identify native abstracts among ones from different texts

Email address: liana.ermakova@univ-brest.fr (Liana Ermakova)

6. Existing metrics, except GEM, have relative values and so are not interpretable
7. The majority of the metrics are normalized, but in practice, their values tend to 0

1 Introduction

In recent decades, data explosion has obliged people to manage the constantly growing amount of information. Thus, more data has been created in the past two years than in the entire history of human existence [1]. The increasing volume of textual information in its various forms such as news articles, comments or posts on social networks poses new challenges for those who aim to understand the storyline of an event [2]. Despite recent significant progress in the domains of information extraction and text mining, the problem of constructing consistent narratives has not yet been solved. New trends in the area of Natural Language Processing (NLP) and Information Retrieval (IR) have emerged such as narrative summarization, story evolution and shift detection, and evaluation methodologies for narrative extraction. They aim to find methods for detecting the most important information and condensing it for users.

Summarization is, by far, the most concrete and most recognized kind of text condensation [3, 4]. A summary is a "*condensed version of a source document having a recognizable genre and a very specific purpose: to give the reader an exact and concise idea of the contents of the source*" [5].

Despite some progress over the last years in information extraction and text summarization, two major challenges remain: the problem of generating consistent narrative summaries and consequently how to evaluate summarization methods and obtained summaries. In order to compare the algorithms for automatic summarization, it is crucial to have a reliable metric for summary quality evaluation.

Summaries can be evaluated considering different facets, but primarily in terms of informativeness and readability [6, 7]. In this paper, we focus mainly on informativeness assessment since this aspect is the most automatized in contrast to readability that is still usually evaluated manually. Manual assessment is expensive, subjective and not applicable to real time scenarios or large collections. Commonly-used metrics for summary evaluation involve substantial human effort in any case. Indeed, they require the comparison of the candidate summaries with a set of reference summaries, although there have been some attempts to use the full text as a reference [8, 9, 10].

In this paper we focus on evaluation methodologies for narrative extraction. Our first contribution is to propose a comprehensive overview of existing metrics and data sets for summary evaluation. An assessment metric allows us to compare different approaches. However, in turn, the assessment metric should be meticulously chosen. Hence, a second contribution of this paper is to introduce an automatic framework to evaluate assessment metrics that does not require any human annotation. Finally, as a third contribution, we evaluate the main existing metrics from the literature on two different data sets: a Wikipedia data set and a collection of scientific articles. Our experiments show that the majority of existing metrics based on vocabulary overlap are not suitable to assess the quality of produced summaries when compared to the original full text.

The rest of the paper is organized as follows. Firstly, in Section 2, we provide a description of the most important international evaluation programs and collections for text summarization. Then, in Section 3 we present an overview of existing assessment metrics. The framework for the evaluation of assessment metrics

is described in Section 4. Section 5 discusses the main challenges in the domain of summary evaluation revealed in the previous section and concludes this paper.

2 Evaluation Programs and Collections

An acute need for automatic text compression has led to the emergence of a number of evaluation programs sponsored by research and governmental institutions. These programs include reference data and metrics.

Shortly after the Internet explosion in the late 90s, the first annual summary evaluation campaign DUC (Document Understanding Conference) appeared. The first evaluation programs were general-purpose, while more recent ones cover specific domains, e.g. biomedical literature, search engine snippets, or are targeting specific applications, e.g. tweet contextualization within the Initiative for the Evaluation of XML Retrieval (INEX).

These assessment programs provide evaluation frameworks containing text collections, evaluation metrics, and gold standards (also called reference summaries or ground truth). In some cases, the evaluation frameworks are automatic (not requiring a gold standard) or semi-automatic (exploiting ground truth) and, thus, can be re-used by non-participants to the evaluation campaign to assess their approaches.

In this section, we discuss the major international evaluation campaigns in detail, namely Document Understanding Conference, Text Analysis Conference, Tweet Contextualization, Snippet retrieval, Sentence Ordering Assessment, and Stream Summarization.

2.1 Document Understanding Conference (DUC)

In 2000-2007, the Document Understanding Conferences (DUC)[1] was a continuing evaluation campaign in the area of text summarization. DUC was an emerging response to the requirement of an evaluation framework for summarization systems within campaigns such as TREC (Text Retrieval Conferences) organized by NIST (National Institute of Standards and Technology), TIDES (Translingual Information Detection Extraction and Summarization) organized by DARPA (Defense Advanced Research Projects Agency), and the ARDA (Advanced Research and Development Activity) Advanced Question & Answering Program. DUC started out as a workshop in the TIDES program and evolved into a text summarization evaluation forum. Sponsored by ARDA, DUC was organized by NIST with the objective of further progress in automatic summarization by allowing researchers to participate in large-scale experiments.

No data were distributed in 2000. In 2001, single- and multi-document summaries (around 10 documents per set) were evaluated by their coverage and readability by 10 human judges. In 2002, NIST provided 60 reference sets, 30 for training and 30 for testing. Each collection contained documents, per-document summaries, and multi-document summaries, with sets defined by different types of criteria.

In 2003, 3 collections were used:

- 30 TREC document clusters of 10 documents on average (AP 1998-2000, New York Times 1998-2000, Xinhua News Agency 1996-2000);

[1] http://duc.nist.gov/

- 30 TDT (Topic Detection and Tracking) document clusters of 10 documents on average (TDT topics/events/timespans and a subset of corresponding documents);
- 30 TREC Novelty track document clusters of 22 documents on average (Financial Times of London 1991-1994, Federal Register 1994, FBIS 1996, Los Angeles Times 1989-1990).

4 tasks were proposed:

1. very short summaries (10 words, no specific format other than linear) manually evaluated by coverage and usefulness;
2. short summaries focused by events (100 words) manually evaluated by quality and length-adjusted coverage;
3. short summaries focused by viewpoints (100 words) manually evaluated by quality and length-adjusted coverage;
4. short summaries in response to a question (100 words) manually evaluated by quality, length-adjusted coverage, and responsiveness.

In 2004 DUC organized 5 tasks:

1. very short single-document summaries;
2. short multi-document summaries focused by TDT events;
3. very short cross-lingual single-document summaries;
4. short cross-lingual multi-document summaries focused by TDT events;
5. short summaries focused by questions.

The summaries in tasks 1-4 were evaluated by ROUGE metric [11]. The summaries of the task 5 were manually evaluated intrinsically for quality and coverage and by their responsiveness to the question.

In 2006-2007, the main task was real-world complex question answering, namely, given a topic and a set of 25 relevant documents, to produce a 250-word summary answering the question in the topic statement [12]. There were 45 topics in the test data. The documents for summarization were taken from the AQUAINT corpus, made of news articles from the Associated Press and New York Times (1998-2000) and Xinhua News Agency (1996-2000). Reference summaries were written by 4 different NIST assessors. In 2005, the task was also answering complex questions, but the documents were chosen from Financial Times of London and Los Angeles Times [13].

The update task was introduced in 2007 and aimed to produce 100-words multi-document update summaries of news articles under the assumption that the user has already read earlier articles. This task was kept in TAC campaign.

The DUC campaign saw large improvements in participants' automatic systems with regards to annotators' proposals. However, there were limitations to participants' summaries since they were topic-dependent and extractive. The DUC data set contains: documents, manually created summaries, automatic baselines, summaries submitted by the participants, evaluation results, additional supporting data and software. The data are distributed according to the Agreement Concerning Dissemination of DUC Results and the User

Agreements. The DUC-2000 data is not available. DUC 2001-2005 collections are distributed under the TREC and TIPSTER user agreements. DUC 2006-2007 data are under the AQUAINT agreement.

In 2008, DUC became a Summarization track in the Text Analysis Conference (TAC).

2.2 Summarization track in the Text Analysis Conference (TAC)

The Text Analysis Conference (TAC)[2] was a series of evaluation campaigns organized to promote research in Natural Language Processing (NLP) by providing a large test collection, standard evaluation framework, and a forum to share results. Being a multi-task workshop, TAC was made of tracks each focusing on a particular NLP problem. TAC was interested in end-user tasks, as well as evaluation methods within the context of end-user tasks.

The Summarization Track evolved during the TAC campaigns [6]. In 2008, Summarization Track included:

- Update Task. The goal was to write a short summary of a set of news articles, assuming that the user is already familiar with a given set of earlier articles. The summaries were evaluated for readability and content by the Pyramid Method (see Section 3 for details). Documents were issued from the AQUAINT-2 collection.
- Opinion Pilot. This task aimed at writing short coherent summaries of opinions from blogs. Summaries should be based either on the text snippet output created by Question Answering (QA) systems as a response to questions from the TAC QA Track, or on the associated documents. Informativeness evaluation was based on the Nugget Pyramid Method. Documents were taken from the Blog06 collection [14].

The Update Summarization task continued in 2009.

In 2010, the summarization task focused on Guided Summarization: the goal was to write a short summary of a set of newswire articles for a given topic from predefined categories. A summary had to include all aspects found for a category, e.g. "*Accidents and Natural Disasters: what happened; date; location; reasons for accident/disaster; casualties; damages; rescue efforts/countermeasures*". Summaries were evaluated for readability, informativeness, and overall responsiveness (see Section 3) .

The 2011 Summarization Track was composed of three tasks:

- Guided Summarization. In contrast to 2010, redundancy in the update summaries was evaluated by the Pyramid score over initial summaries. Test documents were taken from the TAC 2010 KBP Source Data, rather than AQUAINT and AQUAINT-2. Source documents for summarization came from the newswire subset of the TAC 2010 Knowledge Base Population (KBP) Source Data. The collection consists of articles from the New York Times, the Associated Press, and the Xinhua News Agency newswires in 2007-2008.

[2] https://tac.nist.gov/

- Automatically Evaluating Summaries Of Peers (AESOP). In addition to metrics summary informativeness (Pyramid, Responsiveness), AESOP targeted readability. Correlations at the summary level (within each topic) was reported.

- Multiling Pilot Task. This task aimed at promoting the use of multi-lingual algorithms for summarization, e.g. transforming an algorithm or a set of resources from a mono-lingual to a multi-lingual version.

Participants of the Guided Summarization Task had to write 100-word summaries of a set of 10 news articles for a given topic from a predefined category:

- accidents and natural disasters;
- attacks;
- health and safety;
- endangered resources;
- investigations and trials.

Summaries should cover all given important aspects for a category. The data set was composed of 44 topics. A topic had ID, category, title, and 20 relevant documents divided into 2 sets: Set A and Set B. All the documents in Set A chronologically preceded the documents in Set B. No topic narrative were provided, but the category and its aspects expressing the information need of a reader.

An "update" Guided Summarization Task was aimed at written a 100-word "update" summary of a subsequent 10 news articles for the topic, under the assumption that the user has already read the earlier articles. Thus, the goal of the "update" Guided Summarization Task was to recognize new (non-redundant) information in the second set of documents on the same topic.

TAC AESOP task was introduced in 2009. The AESOP task focused on metrics for content evaluation, such as overall responsiveness and Pyramid scores. Within the AESOP task, a collection of automatic evaluation tools was built. AESOP was running until 2011. In 2009 and 2010, AESOP task was focused on developing automatic metrics for content summary evaluation, more precisely measuring the average quality of summarization systems. In 2011, participating metrics should also be applicable on the level of individual summaries. Besides content evaluation, in 2011 the readability measuring should be also considered. Participants obtained the test data from the TAC Guided Summarization task, the human-authored and automatic summaries from that task, namely:

- topic statements;
- two sets of 10 documents for each topic;
- four human written reference summaries for each document set;
- summaries evaluated in the TAC Guided Summarization task;
- list of topic categories and target aspects.

The output of participants automatic metrics was compared against manual metrics: (1) the Pyramid score for content evaluation; (2) Overall Readability for linguistic quality evaluation; and (3) Overall Responsiveness, which is a combined measure for content and linguistic quality evaluation. Pearson's, Spearman's, and Kendall's correlation coefficients between scores assigned by each automatic metric and the three manual metrics were computed. The assumption under the evaluation framework was that a good automatic metric should make the same significant difference between automatic summarizing systems as the manual metrics, but should not provide a ranking to two summarizing systems contradicting the manual metric. Thus, each run was evaluated by (1) the correlation with the manual metric and (2) its discriminative power compared with the manual metric.

TAC Summarization Track was not running in 2012 and 2013. In 2014, TAC restricted the summarization task with only biomedical literature in the Biomedical Summarization track. Nowadays, this task is extremely important since the integration of big data into health care could save around $1000 a year per person according to recent studies [1].

The TAC test data are distributed by the Linguistic Data Consortium (LDC)[3], an open consortium of universities, libraries, corporations and government research laboratories.

The majority of campaigns are focused on evaluation of summarization methods, while AESOP track aims at evaluating assessment metrics by measuring the correlation with the manual metrics and the discriminative power compared with the manual metric. In contrast to that, our framework is automatic and does not require human assigned scores.

2.3 Issue of Streams Summarization

The idea of summarizing streams of information related to a given event is not a recent research issue. However, the rise of social medias and user generated content like micro-blogs gives a new opportunity to tackle the lack of framework.

It recently attracted several research teams in Europe used to focus on automatic summarization of Events or in TREC where it has its own track `http://trecrts.github.io/` [15].

2.3.1 From scheduled events

Indeed, to deal with the overflow of less relevant information, especially in the real-time news streams when human reference is hardly available within a short period (sports events, awards ceremonies or during elections early results phase), some researchers tried to use social media extraction to build so-called summaries.

For instance, the work presented by [16] can be extended to any similar event where relatedness of social media extracts can be compared to official reports. These works as well as [17] intended to be representative but their evaluation was event-detection centered since their purpose was to find the most relevant tweet with regards to the given event. By the way, they not intended to generate a story from the events.

[3] https://www.ldc.upenn.edu

The main drawback of these data sets is that although they are reusable, it may be difficult to compare with existing work since there is no given reference and each researcher can manually select her/his evaluation criterion (e.g. events detection and coverage). Moreover, it could be difficult to replicate such experiments since the context evolved through time.

2.3.2 Real-Time Summarization

From 2016 to 2018, TREC Real-Time Summarization (RTS) focused on different data set and considered, users with multiple information needs. Systems had to help users to keep up to date on their topics of interest. RTS can be considered as an extended version of real-time filtering task in the TREC 2015 Microblog and Temporal Summarization tracks which were on their last edition while both were running from 2010 and 2013.

Two scenarios had been proposed within RTS:

- the first one focuses on push notifications (immediate selection of relevant information, for instance). For instance: notify me when an accident involving autonomous vehicle occurs;
- while the second one was based on email (daily recap of what happened yesterday). For instance, build a digest of all tweets published yesterday related to Apple products bug fixes.

TREC proposed experiments and evaluation using crowdsourced assessments. However, the assessment process evolved through years going through push and fully controlled system to semi- blind, pull-based process [18]. The topics considered for each edition are still available as well as the "Evaluation Broker" API's code. Except tweets from the evaluation periods that have been removed by their authors, this framework could be re-usable.

2.4 INEX/CLEF Tweet Contextualization and CLEF Microblog Contextualization

The rise of user generated content and short texts like micro-blogs or tweets has led to new trends in summarization. Recently, the idea of contextualizing this content to improve tweet understanding was introduced into challenges such as INEX/CLEF Tweet Contextualization (TC) and CLEF Microblog Contextualization [7].

Contextualization Track has grown as an evaluation forum at the intersection of text summarization and information retrieval. TC aims at automatically summarizing large text resources to contextualize (i.e. help to understand) short text passages (e.g. tweets).

Several systems, such as Linguamatics [19], proposed to automatically retrieve the wide range of vocabulary used in tweets, including topic tags, and use linguistic processing to collect and summarize the thousands of ways people have of saying the same thing. Recently, Meij et al. (2012) mapped a tweet into a set of Wikipedia articles, but instead of a summary they provided users with a set of related links [20]. SanJuan et al. took another step by introducing Tweet Contextualization (TC) Track [21] .

In 2011, the Question Answering Track aimed to evaluate TC in terms of (1) relevance of the retrieved information to tweets, and (2) readability of the presented results [21]. In 2012, this track was renamed Tweet Contextualization.

Table 1: Test collections (INEX/CLEF Tweet Contextualization 2011-2014)

	INEX 2011	INEX 2012	INEX 2013	INEX 2014
Corpus	XML dump of English Wikipedia			
	April 2011	November 2011	November 2012	November 2012
Queries	132 tweets (tweet = title + 1-st snt of a NYT article)	1000 tweets from informative accounts	598 tweets from informative accounts	240 topics from RepLab 2013 (tweet + entity + category)
Evaluation (informativeness/ readability)	50 tweets / 53 tweets	50 tweets / 18 tweets	50 tweets / 10 tweets with the largest text references	50 tweets/12 summaries per run
Gold standards	New York Times articles Pool of relevant passages	Pool of relevant passages	Prior set of relevant pages Pool selection of submitted passages All relevant texts+10 random tweets	Pool of relevant sentences Pool of noun phrases

In 2012, answers needed to be a concatenation of textual passages from an external textual resource providing background information that helped to understand the tweet content. Providing an answer to these kinds of task (retrieval and concatenation of relevant textual passages), thus, implies using text summarization.

Associated test collections are described in the table 1.

In 2011, the query data set included 132 tweets. Each tweet consisted of the identifier (*id*), the title (*title*), and the first sentence (*txt*) of a New York Times article published in July 2011.

For each tweet, participants had to provide a summary of up to 500 words in TREC format that is contextualizing the tweet, i.e. answering the question "what is this tweet about?". The summary had to contain as much relevant information as possible and not include irrelevant or redundant passages.

The summary had to be made solely of extracts from the XML dump of English Wikipedia articles (April 2011): 3,217,015 non-empty pages in total. All notes, history and bibliographic references were removed. Thus, a page was composed of a title (*title*), an abstract (*a*) and sections (*s*). Each section had a header (*h*)). Each abstract and all sections contained paragraphs (*p*) and entities (*t*) referring to other pages.

The summaries submitted by participants were compared with each other, with the baseline summary made of sentences (BaselineSum) and with the key terms (BaselineMWT). The baseline system was based on the Indri index with stemming and without stop words (language model). Part of speech tagging was performed by TreeTagger. The summarization algorithm was TermWatch [21].

In 2012, the text corpus was presented by an updated Wikipedia dump from November 2011. The query set was dramatically changed. It consisted of approximately 1000 real tweets written in English collected from informative accounts such as @CNN, @TennisTweets, @PeopleMag, @science etc. However, the task remained the same: to provide a summary up to 500 words in the TREC format. In 2013 there were 598 tweets in English to be contextualized from the Wikipedia dump of November 2012.

For all test collections, 50 tweets[4] were selected to evaluate the informativeness of the summaries [21]. For each of those topics, all submitted passages were merged into a pool. Passages were sorted in alphabetic

[4]Although this amount can be considered as limited, the INEX organizers shown than competitors ranking is not affected by variating the amount of tweets selected for evaluation.

order and therefore each passage was judged whether it was relevant independently from others. Submitted summaries were compared with the corresponding pools of relevant passages. In 2011 summaries were also evaluated according to the overlap with the original New York Times articles. In 2013 the informativeness was estimated as the overlap of a summary with 3 pools of relevant passages [22]:

- prior set (PRIOR) of relevant pages selected by organizers (40 tweets, 380 passages);
- pool selection (POOL) of the most relevant passages (1,760) from participant submissions for 45 selected tweets;
- all relevant texts (ALL) merged together with extra passages from a random pool of 10 tweets (70 tweets, 2,378 relevant passages).

In 2014 there were 240 tweets in English collected by the organizers of CLEF RepLab 2013. In 2014 participants should provide a context to tweets from the perspective of the related entities. Tweets were at least 80 characters long and did not contain URLs. A tweet had the following annotation types: the category (4 distinct), an entity name from the Wikipedia (64 distinct) and a manual topic label (235 distinct). The context had to explain the relationship between a tweet and an entity. As in previous years it should be a summary extracted from a Wikipedia dump.

In 2014, 2 gold standards (1/5 of the topics/tweets) were used:

- pool of relevant sentences per topic/tweet (SENT);
- pool of noun phrases (NOUN) extracted from these sentences together with the corresponding Wikipedia entry.

The pool of relevant passages was constructed through the summaries submitted by the participants of the track. Only these passages were judged as relevant or not. All other passages are considered as irrelevant. Thus, a brand new system which is able to catch relevant but not judged passages can be potentially underscored due to this bias. The same remark can be applied to traditional ad-hoc information retrieval campaigns since measures like MAP (mean average precision) consider all non-judged results as irrelevant (see for example TREC Robust data[5]).

2.5 INEX Snippet Retrieval Task

Search engines return to a user an immense volume of results that it is impossible to read. Therefore, to define whether a web page is relevant to a query without clicking on a link, a search engine provides a user with snippets. Snippets are small text passages appearing under every search result.

For the Snippet Retrieval Track 2012, the data collection consisted of the dump of the Wikipedia of October 2008 annotated with YAGO [23] and 35 topics. Participants should provide 20 snippets per topic limited to 180 characters [24]. In 2013, the Snippet Retrieval track was using the same document collection as the Tweet Contextualisation track, based on a dump of the English Wikipedia from November 2012. The

[5] https://trec.nist.gov/data/robust.html

set of topics was the same as in 2012. The DTD for the submission format was as follows. The topic title, description, and narrative (intent) provided the idea of the user information need.

The main drawback of this data set is that it is not re-usable after the track since the entire evaluation was manual.

2.6 Collections for Sentence Ordering Assessment

Automatic text generation systems, particularly multi-document extractive summarization, face sentence ordering problem [25, 22]. While evaluating informativeness of automatically produced content without involving manual process interests numerous research teams, the evaluation of readability is still mainly an expert work. However, there exist frameworks that allow the assessment of sentence ordering, i.e. readability, since sentence ordering is the only way to influence the readability of an extractive summary.

Two corpora are widely used for sentence ordering assessment:

- airplane **Accidents** from the National Transportation Safety Board,
- articles about **Earthquakes** from the North American News Corpus [26, 27, 28].

Each of these corpora has 100 original texts and for each document 20 permutations (2,000 in total).

3 Evaluation Measures

In this section, we present an overview of the metrics used to evaluate summaries that have been generated automatically. When applicable, we mention the evaluation program or collection from Section 2 the metric has been used for. Of course, performing an exhaustive survey seems hardly possible due to the number of related works. However, we consider a wide range of studies and adopt a high-level approach, focusing on the relevance of the measures depending on the context in which the summaries are generated. We organize them in a typology with regard to the evaluation use case and resources available.

3.1 Informativeness Evaluation

3.1.1 Questionnaire-based Metrics

Summaries may be evaluated according to compression rate, i.e. proportion of summary length over full text length, or retention rate, i.e. proportion of the retained information [29].

A good summary should have low compression rate and high retention rate. Compression rate is well-defined and can be easily computed while retention rate estimation is more problematic since it involves less formalized concepts. To measure the retained information, one assessor team develops a set of questions based on the input texts, while another group answers these questions reading only summaries [30]. An assessor may be asked to evaluate the importance of each sentence/passage and the obtained annotation of importance allows summaries with a predefined compression rate to be generated [31] or serve as expert extracts which may be used as a gold standard.

In [32], the Responsiveness metric is proposed. It was designed for query-focused summarization and shows how well a summary satisfies the user's information need expressed by the query. The Responsiveness metric was applied at DUC. The expert nature of this metric makes it impossible for further re-use.

At INEX Snippet Retrieval Track 2012-2013 [22], the relevance of the documents was judged independently from the relevance of the snippets. It was done in order to determine the effectiveness of a snippet to provide sufficient information about the corresponding document. Thus, assessors should evaluate results in two ways: (a) relevance evaluation of documents; (b) relevance evaluation of snippets.

Assessors had to go through the snippets, and decide whether the underlying document seemed relevant to the topic by reading only the snippet. They put 1 if it seemed to be relevant and 0 otherwise. After that, assessors had to read all of the documents and judge their relevance. Then, snippet-based relevance judgments were compared with the document-based relevance judgments (ground truth), i.e. a good snippet should be judged the same as the corresponding document. Then, these judgments were integrated by the following measures [22]:

- Mean prediction accuracy (MPA) — the average percentage of results the assessor correctly assessed:

$$MPA = \frac{TP+TN}{TP+FN+TN+FP} \tag{1}$$

  where TP refers to true positive, TN to true negative, FN to false negative, and FP to false positive.

- Mean normalized prediction accuracy (MNPA) is the average of the relevant results correctly assessed and the irrelevant results correctly assessed:

$$MNPA = 0.5 \times \frac{TP}{TP+FN} + 0.5 \times \frac{TN}{TN+FP} \tag{2}$$

- Recall is the average percentage of relevant documents correctly assessed:

$$R = \frac{TP}{TP+FN} \tag{3}$$

- Negative recall (NR) is the average percentage of irrelevant documents correctly assessed:

$$NR = \frac{TN}{TN+FP} \tag{4}$$

- Positive agreement (PA) is the conditional probability of agreement between snippet assessor and document assessor, given that one of the two is judged relevant:

$$PA = 2 \times \frac{TP}{2*TP+FP+FN} \tag{5}$$

- Negative agreement (NA) is the conditional probability of agreement between snippet assessor and document assessor, given that one of the two judged irrelevant:

$$NA = 2 \times \frac{TN}{2*TN+FP+FN} \tag{6}$$

- Geometric mean (GM) of recall and negative recall:

$$GM = \sqrt{R \times NR} \tag{7}$$

TREC RTS organizers considered in-situ and bach evaluation metrics Precision and Utilty [33].

$$Precision = \times \frac{relevant}{relevant + redundant + notrelevant} \tag{8}$$

and

$$Utility = relevant - redundant - notrelevant \tag{9}$$

However, the interleaving of outputs from several systems was responsible for introducing the redundancy and may disadvantaged some systems.

3.1.2 Overlap-based Measures

Reference summaries allow the metrics commonly used in information retrieval to be calculated: recall (R) and precision (P) over the number of terms/sentences appearing in reference and candidate summaries [29]:

$$R = \frac{|S \cap C|}{|S|} \tag{10}$$

$$P = \frac{|S \cap C|}{|C|} \tag{11}$$

where S is a set of terms/sentences in the set of reference summaries, C is a set of terms/sentences in the candidate summary.

Recall and precision may be integrated into the $F - measure$ [34]:

$$F = \frac{(\beta^2 + 1) \times R \times P}{\beta^2 \times P + R} \tag{12}$$

$$\beta^2 = \frac{1 - \alpha}{\alpha}, \alpha \in [0, 1] \tag{13}$$

The F-measure is widely used in ad-hoc information retrieval but it is less useful in summary evaluation since a search engine result is potentially infinite while a summary is limited. Besides, the sentence-based measures of information retrieval types cannot be applied to abstract assessment since abstracting implies the reformulation of initial sentences.

Similarity between reference and candidate summaries may be estimated as cosine, dice or Jaccard coefficient, as well as the number of shared n-grams or longest common subsequence etc.

The cosine similarity between a summary and a reference hereinafter referred to as **COS** is estimated as follows:

$$cos(C, S) = \frac{S\dot{C}}{\sqrt{\sum_{i=1}^{n} S_i^2} \sqrt{\sum_{i=1}^{n} C_i^2}} \tag{14}$$

$$cos(S, C) \in [0, 1] \tag{15}$$

where C is the summary under evaluation and S is the reference text, S_i and C_i are TF-IDF of the $i-$th term in the vector representation S and C respectively. In our experiments, we used IDFs learned on the entire Wikipedia collection (see Section 4).

One of the most efficient metrics of summary evaluation is ROUGE (Recall-Oriented Understudy for Gisting Evaluation) used at the Document Understanding Conference (DUC) [34]. ROUGE is also based

on comparison with a set of reference summaries. Proposed by Chin-Yew Lin in 2004, ROUGE aims to evaluate summaries but it is also used to assess the quality of machine translation. Several variants of the ROUGE metrics exist, e.g. $ROUGE_N$ (n-grams recall):

$$ROUGE_N(S,C) = \frac{|S \cap C|}{|S|} \tag{16}$$

where S is a multi-set of n-grams in the set of reference summaries, C is a a multi-set of n-grams in the candidate summary, N refers to the n-gram length. Thus, the main difference of $ROUGE_N$ from the recall is the use of multi-set instead of a simple set. Another dissimilarity is n-gram application. $ROUGE_N$ implies that a summary gets a higher score as it contains more n-grams co-occurring with reference summaries. This metric is called further as **ROUGE**.

$ROUGE_{N\_multi}$ computes pairwise n-gram recall with each reference summary S_i and takes the maximal value:

$$ROUGE_{N\_multi} = \arg\max_i ROUGE_N(S_i,C) \tag{17}$$

Another method $ROUGE_L$ is based on the search for the longest common substring (LCS) shared by two sentences:

$$ROUGE_L = F_{lcs} = \frac{(\beta^2+1) \times R_{lcs} \times P_{lcs}}{\beta^2 \times P_{lcs} + R_{lcs}} \tag{18}$$

$$R_{lcs} = \frac{|LCS(X,Y)|}{|X|} \tag{19}$$

$$P_{lcs} = \frac{|LCS(X,Y)|}{|Y|} \tag{20}$$

where $LCS(X,Y)$ is the longest common substring of the sentences X and Y. If there is no shared subsequence $ROUGE_L = 0$. $ROUGE_L$ includes the longest common n-gram and there is no need to compute its length in advance. $ROUGE_L$ allows to compare the sentence structure but only with respect to the longest shared part. For the whole texts $ROUGE_L$ can be estimated by the formulas:

$$ROUGE_L = F_{lcs} = \frac{(\beta^2+1) \times R_{lcs} \times P_{lcs}}{\beta^2 \times P_{lcs} + R_{lcs}} \tag{21}$$

$$R_{lcs} = \frac{\sum_{Si \in S} |LCS_\cup(S_i,C)|}{|S|} \tag{22}$$

$$P_{lcs} = \frac{\sum_{Si \in S} |LCS_\cup(S_i,C)|}{|C|} \tag{23}$$

where $|LCS_\cup(S_i,C)|$ is the LCS score of the union of the longest common subsequences between each reference sentence s_i and candidate summary C. For example, let $s_i = w_1w_2w_3w_4w_5$, and $C = c_1c_2$, $c_1 = w_1w_2w_6w_7w_8$, $c_2 = w_1w_3w_8w_9w_5$, then LCS for s_i, c_1 is w_1w_2, and for s_i, c_2 LCS is $w_1w_3w_5$. The union is $w_1w_2w_3w_5$. $|LCS_\cup(s_i,C)| = 4/5$.

Normalized pairwise comparison $LCS_{MEAD}(S,C)$ [35] is similar to $ROUGE_L$ when $\beta = 1$ [34], but LCS_{MEAD} takes the maximal value of LCS, while $ROUGE_L$ deals with the union of LCS [36].

$$LCS_{MEAD} = \frac{(\beta^2+1) \times R_{MEAD} \times P_{MEAD}}{\beta^2 \times P_{MEAD} + R_{MEAD}} \tag{24}$$

$$R_{MEAD} = \frac{\sum_{s_i \in S} \max_{s_j \in C} LCS(s_i, s_j)}{m} \tag{25}$$

$$P_{MEAD} = \frac{\sum_{s_j \in C} \max_{s_i \in S} LCS(s_i, s_j)}{n} \tag{26}$$

One of the serious shortcomings of LCS is the fact that it does not consider the distance between words. Let consider an example. Let $S = ABCDEFG$ be a reference, $C1 = AZBYCZD$ a candidate, $C2 = ABCDXYZ$ another candidate. The longest common substring is ABC and therefore, $ROUGE_L$ attributes the same score to both candidates. However, the second candidate is better. Weighted LCS, called WLCS, takes into account the length of consecutive matches by using dynamic programming approach memorizing the length of the current consecutive matches ending at words s_i and c_j. For more details about the algorithm see [11]. The weighting function f is a function of consecutive matches at the table position (i, j) and should satisfy the following constraint: $(\forall x, y \in N) : f(x+y) > f(x) + f(y)$. That is to say consecutive matches should have higher score than non-consecutive ones. f may be the linear $f(k) = \alpha k - \beta, \alpha > 0, \beta > 0$, polynomial or quadratic function $f(k) = k^2$. Let f^{-1} be the inverse function of f, for example, $f = k^2$, $f^{-1} = k^{1/2}$. In this case F-measure is estimated as follows:

$$F_{wlcs} = \frac{(\beta^2 + 1) \times R_{wlcs} \times P_{wlcs}}{\beta^2 \times P_{wlcs} + R_{wlcs}} \tag{27}$$

$$R_{wlcs} = f^{-1}\left(\frac{WLCS(X,Y)}{f(m)}\right) \tag{28}$$

$$P_{wlcs} = f^{-1}\left(\frac{WLCS(X,Y)}{f(n)}\right) \tag{29}$$

LCS based algorithms are a special case of edit distance [37].

The metric ROUGE-S is based on the counting of shared bi-grams the elements of which may be separated by arbitrary number of other words. The distance may be limited by d_{skip}. Sometimes uni-gram smoothing is applied (ROUGE-SU).

To compute ROUGE-S the following formulas are applied:

$$F_{skip2} = \frac{(\beta^2 + 1) \times R_{skip2} \times P_{skip2}}{\beta^2 \times P_{skip2} + R_{skip2}} \tag{30}$$

$$R_{skip2} = \frac{SKIP2(X,Y)}{C(m,2)} \tag{31}$$

$$P_{skip2} = \frac{SKIP2(X,Y)}{C(n,2)} \tag{32}$$

where $C(n,k)$ is the binomial coefficient $\binom{n}{k}$, and $SKIP2(X,Y)$ is the number of common bi-grams with arbitrary distance in the texts X and Y respectively. The distance may be limited by d_{skip}. Sometimes uni-gram smoothing is applied (ROUGE-SU).

The experiments conducted by the organizers of Automatically Evaluating Summaries of Peers (AESOP) task within Text Analysis Conference (TAC) showed that the metrics proposed by TAC participants,

such as ROUGE-BE, DemokritosGR2, catholicasc1, and CLASSY1 significantly outperform ROUGE-2 which is the best from all ROUGE variants [32].

The metric BLEU (BiLingual Evaluation Understudy) commonly used for machine translation evaluation is also suitable for assessment of any generated text [34]. As ROUGE, BLEU also estimates as the number of shared n-grams. In contrast to ROUGE which is recall oriented, BLEU uses a modified form of precision to compare a candidate against multiple references [38]:

$$p^{BLEU} = \frac{|S \cap C|}{|C|} \tag{33}$$

where S is a multi-set of n-grams in the reference summary, C is a a multi-set of n-grams in the candidate summary. Note that intersection of multi-sets captures the notion of clipped counts, i.e. upper bounding of the total count of each candidate word by its maximum reference count, introduced in [38].

Thus, BLEU measures how much n-grams in the candidate summary appear in the reference summaries, while ROUGE estimates how much n-grams in the reference summaries appear in the candidate one. Precision-based metrics have tendency to prefer short texts making them less appropriate to summary evaluation. However, a good summary should provide as much important information as possible. Besides, one can imagine an extreme case when a summarization system provides only one word which may be relevant and a precision-based metric would attribute the maximal score to it. Therefore, brevity penalty was introduced to penalizes texts shorter than the length of a reference [11]:

$$BP = \begin{cases} 1 & \text{if } |C| > |S|_{avg} \\ \exp^{1-\frac{|S|_{avg}}{|C|}} & \text{if } |C| \leq |S|_{avg} \end{cases} \tag{34}$$

Thus, the final score of BLEU is calculated as [38]:

$$BLEU = BP \exp^{\sum_{n=1}^{N} w_n \log p_n^{BLEU}} \tag{35}$$

or

$$\log BLEU = \min\left(1 - \frac{|S|_{avg}}{|C|}, 0\right) + \sum_{n=1}^{N} w_n \log p_n^{BLEU} \tag{36}$$

where N is the maximal n-gram order, w_n is n-gram weight.

Further improvement of BLEU was done by US National Institute of Standards and Technology within DUC conference. They introduced the metric NIST, which does not only calculates n-gram precision, but also weights the informativeness of a particular n-gram, for example by IDF [11].

In contrast to BLEU, the METEOR (Metric for Evaluation of Translation with Explicit ORdering) metric is able to treat spelling variants, WordNet synsets and paraphrase tables and distinguishes function and content words [39]. METEOR is calculated based on the F-measure between recall and precision and a penalty p for fragmentation:

$$p = 0.5\left(\frac{|chunks|}{|S \cap C|}\right)^3 \tag{37}$$

$$METEOR = F_{mean}(1-p) \tag{38}$$

where $|chunks|$ is the total number of chunks, i.e. a set of uni-grams that are adjacent in the candidate and in the reference. Basic Elements (BE) which can be considered as paraphrases where proposed by

Tratz and Hovy [36]. A BE is a syntactic unit up to 3 words with associated tags such as named entities and parts of speech. BE can deal lemmas, synonyms, hyponyms and hyperonyms, identical prepositional phrases, spelling variants, nominalization and denominalization (derivation in WordNet), transformations like prenominal noun - prepositional phrase, noun swapping for IS-A type rules, pronoun transformations, pertainym adjective transformation.

One of the derivatives of the edit distance is Word Error Rate (WER) [40]:

$$WER = \frac{Sub + Del + Ins}{|S|} \tag{39}$$

where Sub is the number of substitutions, Del is the number of deletions, Ins is the number of insertions, and $|S|$ is the number of words in the reference. In fact, WER is the length normalized edit distance.

The organizers of INEX Tweet Contextualization Track 2011-2014 evaluated extractive summaries by comparing them with the pool of passages judged as relevant. As the distance they used the Kullback-Leibler divergence or simple log difference [41]. They state that the Kullback-Leibler divergence is very sensitive to smoothing in case of small number of relevant passages in contrast to the absolute log-diff between frequencies [41].

Until 2011, the informativeness was evaluated by the Kullback-Leibler (KL) divergence [42]:

$$KL = \sum_{w \in S \cup C} p_C(w) \log \frac{p_C(w)}{p_S(w)} \tag{40}$$

where $p_C(w)$ and $p_S(w)$ refer to the probability to see w in C and S respectively. Dirichlet smoothing was applied.

In 2011 the informativeness was estimated as the log difference:

$$Div = \sum_{w \in S} \left| \log \left(\frac{|S(w)|}{|S|} + 1 \right) - \log \left(\frac{|C(w)|}{500} + 1 \right) \right| \tag{41}$$

where S is the set of terms in the pool of relevant passages, $|S(w)|$ is the frequency of a term w in the pool, $|S|$ is the total number of terms in the pool, $|C(w)|$ is the frequency of a term w in a summary, $|C|$ is the total number of terms in a summary. A term may refer to a uni-gram, a bi-gram (two consecutive lemmas in the same sentence) or a bi-gram allowing a gap up to two lemmas between its component (with 2-gap). The lower values of Div corresponds to higher matching of tokens in a pool and a summary. The evaluation was carried out by FRESA[43] package which includes a special lemmatizer.

Since 2012 the informativeness was evaluated by the following formula:

$$Dis = \sum_{w \in S} \frac{|S(w)|}{|S|} \times \left(1 - \frac{\min(\log P, \log Q)}{\max(\log P, \log Q)} \right) \tag{42}$$

where P and Q are computed as:

$$P = \frac{|S(w)|}{|S|} + 1 \tag{43}$$

$$Q = \frac{|C(w)|}{|C|} + 1 \tag{44}$$

Since $\frac{|C(w)|}{|C|} \in (0,1]$ and $\frac{|S(w)|}{|S|} \in (0,1]$, $P > 1$ and $Q > 1$. Therefore, $\max(\log P, \log Q) > 1$. The logarithm allows dealing with highly frequent words. The evaluation toolkit was based on Porter stemmer. The lower

values of *Dis* correspond to the higher informativeness. The complement of this dissimilarity measure $1 - Dis$ has similar properties than usual information retrieval Interpolate Precision measures. Thus, we used this complement as a competing metric hereafter referred to as **INEX**.

[44] introduced a trivergent model that outperformed the divergence score. In [45], the authors proposed to use the similarity within semantic representation such as LSA, LDA, Word2Vec and Doc2Vec. However, ROUGE-1 outperformed all these metrics. In [46], ROUGE metric was modified by word embeddings but this variant showed lower results than the standard one.

3.1.3 Other Informativeness Metrics

A Pyramid score is based on the number of repetitions of Summary Content Units (SCU) in the gold-standards [47]. SCUs are information units of variable length inside a sentence labeled by experts in their own words. SCUs are weighted according to the number of summaries they occur in and, thus, are organized into a pyramid. The final Pyramid score is estimated as follows:

$$Pyramid = \frac{\sum_{i=1}^{n} i \times D_i}{\sum_{i=j+1}^{n} i \times |T_i| + j \times \left(X - \sum_{i=j+1}^{n} |T_i|\right)} \tag{45}$$

$$j = \max_i \left(\sum_{t=i}^{n} |T_t| \geq X \right) \tag{46}$$

where n is the number of tiers, $|T_i|$ denotes the number of SCUs at the level T_i, D_i is the number of SCUs in the summary appearing in T_i, X is summary size in SCUs. Pyramid score was used at the DUC conference. The main drawback of the Pyramid score is that it is based on manual assessment not only of the reference summaries, but also of the candidate ones, and therefore it can not be re-used for evaluation of new candidates.

GRAD (GRAph Distance) metric aims to estimate how well summary terms are connected to full text terms and is based on the assumption that a good summary is made of the terms that refer to the central vertices in the semantic graph, i.e. the terms that are connected to the maximal number of other terms in a full text [10]. According to this metric, the score of a summary is estimated as a normalized inverted sum of distances from every term in the text to its closest term appearing in the summary S:

$$score(S) = \frac{1}{|S| \sum_{v_i} \min_{v_j \in V \cap S} d(v_j, v_i)} \tag{47}$$

where $d(v_j, v_i)$ is the shortest path between v_i and v_j. To calculate minimal distances from every term in the text to its closest term from the summary, a modified Dijkstra's algorithm is used [48].

The metric GEM (GEnerosity Measure) attributes an absolute score [0,1] to a summary [9]. GEM relies on the importance of the different sections of a scientific paper. The GEM score is calculated as the sum of the weights of the section classes retrieved both in a summary and a full text normalized over the total sum of weights of section classes in a full text. For each sentence, the section class in a summary is assigned according to the class of the sentence from the full text with the maximal cosine similarity. As GRAD, GEM fails to distinguish native abstracts from other human summaries.

3.2 Readability Evaluation

Readability, coherence, conciseness, content, grammar, recall, pithiness etc. are usually assessed manually [34, 31] since often these parameters are not numerically expressed [29].

The first metrics of readability were proposed outside the scope of automatic summarization. For example, Gunning fog index estimates the years of formal education a person needs to understand the text on the first reading and is computed based on the average sentence length and percentage of complex words [49]. The Flesch–Kincaid readability tests is designed to indicate how difficult a passage in English is to understand based on word length and sentence length [50]. Despite the easiness of these two metrics, they are not reliable [51].

Language models were also used for estimate the readability [51, 52, 53]. Machine learning techniques were proposed in [54, 55]. Feng et al. proposed to use discourse, language modelling, parsing, POS, and quantitative features [55]. However, all these metrics are not applicable to the readability evaluation of extractive summaries since summary phrases are taken from source documents as they are.

Traditional methods of readability evaluation are based on familiarity of terms [56, 57, 58] or their length [59] and syntax complexity [60]. Another set of methods is based on syntax analysis [61, 62, 63]. Syntactical methods may be combined with statistics (e.g. sentence length, the depth of a parse tree, omission of personal verb, rate of prepositional phrases, noun and verb groups etc.) [64]. The latter methods are suitable only for the readability evaluation of a particular sentence and therefore they cannot be used for extracts assessment. Researches also proposed to use language models for prediction of readability difficulties, i.e. predicting word difficulty based on the language model [60, 65]. Usually assessors assign score to the readability of text in some range [66].

Syntactical errors, unresolved anaphora, redundant information and coherence influence readability and therefore the score may depend on the number of these mistakes [41]. At INEX Tweet Contextualization Track 2011-2014, assessors were not asked to evaluate the relevance of the summaries. There were two metrics:

- Relaxed metric: a passage was considered valid if it was not marked as trash,
- Strict metric: a passage was considered valid if it did not have any problems mentioned above.

In 2011, the readability of summaries was estimated as the number of words (up to 500) in valid passages [21]. Since 2012, the score of a summary was the average normalized number of words in valid passages [22]. Sentence ordering was not judged by conference organizers, however several authors affirm that it is quite important for text understanding [66].

Different non-parametric rank correlation coefficients (e.g. Kendall, Spearman or Pearson coefficients) may be used to find the dependence [67]. However, as shown in [68], Kendall coefficient is the most suitable for sentence ordering assessment.

Current methods to evaluate readability are based on the familiarity of terms and syntax complexity [52]. Word complexity may be estimated by humans [69, 57, 58] or according to its length [59]. Researches also propose to use language models [52, 65]. Usually assessors assign a score to the readability of a text in some range [66]. BLEU with high-order n-grams can be used for readability evaluation. Edit distance

may be applied as well for readability evaluation to assess word order or sentence order. These metrics are semi-automatic because they require a gold standard. Another set of methods is based on syntax analysis which may be combined with statistics (e.g. sentence length, depth of a parse tree, omission of personal verbs, rate of prepositional phrases, noun and verb groups) [61, 62, 63, 64], but they remain suitable only for the readability evaluation of a particular sentence and, therefore, cannot be used for assessing extracts.

As shown, evaluating informativeness of automatically produced contents without involving manual process interest numerous research teams. However, the evaluation of readability is still mainly manual which leads to the need of designing self-sufficient metric that could measure sentence ordering hence readability. Ermakova [70] proposes an automatic approach for sentence order assessment where the similarity between adjacent sentences is used as a measure of text coherence. However, it assigns equal scores to initial and inverse sentence order due to the symmetric similarity measure. In contrast, the topic-comment based method proposed in [71] can deal with this problem.

4 Limitations of the Existing Assessment Metrics

4.1 Framework to Evaluate Assessment Metrics

An assessment metric allows comparing various methods in order to choose the most appropriate one. However, this assessment metric should also be carefully selected. In the previous section we described the main metrics used for summary evaluation. In this section we present the experimental results that revealed some limitations of existing metrics.

Traditionally, the quality of assessment metrics is evaluated as a correlation between expert results and candidate metrics (e.g. Kendall, Spearman, or Pearson coefficients) [34]. A good metric should give low score to summaries which have low score according to human judgment and high score otherwise.

In the case where a reference summary is compared to other references, re-sampling methods are often used, e.g. jackknifing (use of the subsets of the reference summaries, each of which is missing one reference) or bootstrapping (random replacement of points in the data set). This is important, since comparing a reference to itself leads to the maximal score, but ignoring it results into different number of references. In case of re-sampling, the final score is the mean of all computed values [36].

In [8], Louis and Nenkova suggested an automatic approach for summary evaluation without a gold standard. Instead of a set of reference summaries, a full text is used. The authors suggested to estimate summary scores by Kullback-Leibler divergence, Jensen Shannon divergence, and cosine similarity measure. Although these metrics have some correlation with traditional gold-standard based ROUGE scores, ROUGE-1 demonstrated better results.

In this section, we will compare the main overlap-based metrics ROUGE and INEX as defined in 42. We took into account only automatic metrics that is why we did not consider the Pyramid score nor Responsiveness. ROUGE and INEX metrics seems to be the most widely used automatic metrics. They were applied during several years in different evaluation campaigns. We also compared the results with a simple cosine similarity. We will show that they have similar drawbacks since they are based on vocabulary overlap. Moreover, we compared the results with those of less classical metrics GRAD and GEM since they are not overlap based.

We show the limitations of the use of the full text as a gold standard for vocabulary overlap based metrics. In contrast to [8], besides cosine similarity measure, we also compared INEX and ROUGE metrics that use the full text as a gold standard.

The evaluation framework we defined is two folds: on the one hand, we assume a human written summary should be considered as better than a automatically generated one, on the other hand we consider that the measure should assign higher score to the summary coming from the text under consideration than to the summaries of other texts.

The intuition underlying the first part of the evaluation is that a good assessment metric should assign a high score to a good summary and a low score to a bad one. In contrast to [34], rather than calculating the correlation between the scores assigned to summaries by assessors and metrics, we propose to compare the percentage of times when a good summary of a text is scored lower than a worse one of the same text (accuracy).

This approach requires at least two summaries for every article. The data sets we used contain one human provided summary and another two summaries we generated automatically using poor methods described below. We assume that human written summaries are better than the generated ones and a good metric should reflect that. It is important to notice that we deliberately chose very simple methods for automatic summary creation to ensure that human written abstracts are much better than the generated ones. This evaluation framework does not require explicit human assigned scores and may be performed on very large collections. However, rank correlation coefficients are not applicable for these data since there is no ground truth beyond a pair of summaries and, thus, it is impossible to compare summaries of different full texts.

The second part of the evaluation aims at evaluating the capacity of an automatic measure to identify the appropriate abstract among human provided summaries of different texts (the text under consideration and other texts). We performed pair-wise comparison between each of 31 documents and abstracts coming from different texts using the automatic assessment metrics. A good metric should assign higher score to the abstract initially associated to the document by its authors not the abstract from another document.

4.2 Data Sets

We conducted experiments on two data sets with existing abstracts: scientific and Wikipedia articles.

4.2.1 Scientific Articles.

For our experiments we used the data from the ISTEX (Excellence Initiative of Scientific and Technical Information) platform[6] that contains collections of scientific literature in all disciplines covering journal archives, digital books, databases, texts corpora etc. from the following publishers: Elsevier, Wiley, Springer, Oxford University Press, British Medical Journal, IOP Publishing, Nature, Royal Society of Chemistry, De Gruyter, Ecco Press, Emerald, Brill, Early English Books Online. We collected articles that contain authors abstracts, editorial or/and web summaries. We selected 4,234 full texts in TXT format with corresponding summaries provided as XML descriptions. We call this data set **ISTEX**.

[6] http://www.istex.fr/

We also selected 4,535 articles with abstracts about environmental sciences from ISTEX (referred as **Environmental**). ISTEX documents are labeled by Web of Science categories. So, we retrieved documents labeled *Environmental Studies* or *Environmental Science*.

4.2.2 Wikipedia.

The second data set we used was a cleaned recent English Wikipedia XML dump created by the INEX organizers for Tweet Contextualization Track [41]. All notes, history and bibliographic references were removed. Thus, a page was composed of a title (*title*), an abstract (*a*) and sections (*s*). A section had a header (*h*). Abstract and sections contained paragraphs (*p*) and entities (*t*) referring to other pages. We selected the 100,000 longest articles. However, only 43,611 articles had both abstracts and not-empty sections and were kept for experiments. This data set is referred as **Wikipedia**.

4.3 Methods for Summary Generation

Generated summaries have low readability because of abrupt topic changes and unresolved anaphoras. However from informativeness standpoint we assume they will be competitive with human provided summaries. Since our goal is mainly to evaluate evaluation metrics, we deliberately applied very simple methods to generate summaries since these simple methods should produce in average summaries of lower quality than summaries created by using more sophisticated methods. We believe that simple methods for summary generation are sufficient since (1) even in this case vocabulary-based metrics fail and (2) there is no need to check manually if they are less good than the human written ones. The two methods are presented below.

4.3.1 Random Extractive Summary.

The first method we used is extractive summarization that randomly selects the sentences from the full texts while the total number of words does not exceed a predefined threshold. To avoid bias because of the different summary sizes, the size of the generated summaries was set to the average size of human provided ones (200). For evaluation purpose, we generated a summary for each document only once. The random bias, however, should be vanished due to the average score over a high number of documents. We call this method **RandSum**.

4.3.2 Cosine Similarity Based Summary.

The second approach we applied for summary generation is based on the cosine similarity measure between bag-of-words representations of a candidate sentence C and a full texts A:

$$cos(C,A) = \frac{\sum_{i=1}^{n} C_i A_i}{\sqrt{\sum_{i=1}^{n} C_i^2}\sqrt{\sum_{i=1}^{n} A_i^2}} \tag{48}$$

A_i and C_i refer to term TF-IDF from the corresponding texts. The sentences with the highest scores are selected until the total number of words in a summary is less than a predefined threshold. This method is later denoted by **CosSum**.

4.4 Competing Metrics

In our experiments we compared three vocabulary-based metrics (INEX, ROUGE-N, and COS) with a graph-based measure GRAD.

Since vocabulary overlap based metrics have very low performance when a full text is used as a gold standard, we also compared results with a random score assigned to a summary in the range $[0, 1]$ (hereafter called **RAND**).

Additional details regarding metrics comparison including Jensen–Shannon and Kullback-Leibler divergence [8] can be found in [72].

4.5 Revealed Limitations of the Use of the Full Text as a Gold Standard

Table 2 provides the evaluation results of competing metrics as the percentage of times when a human provided abstract is scored higher/lower/equally than/to a generated summary. We assume that human written summaries are better than the generated ones. Thus, we use the accuracy as a performance metric; here, the accuracy means the percentage of times when human provided summary has higher score than the generated ones. The higher accuracy is, the better is the metric. We performed Student's t-test to check whether the results are significantly different from the random ones at the level $p < 0.05$. $^{+}/^{-}$ indicates the accuracy of the metrics that is significantly higher/lower than the accuracy of RAND.

ROUGE and COS accuracy is less than 25% for all data sets. INEX has a high accuracy on ISTEX data but very poor accuracy on other collections. In all cases, the difference with the random baseline is significant. The table testifies that GRAD significantly outperforms the baseline assigning score randomly as well as the state-of-the-art vocabulary overlap based metrics (INEX, ROUGE).

Since the used methods for summary generation are the extractive ones, the metrics based on the overlap between terms should have a low performance. ROUGE and COS are in average twice as worse as the random baseline on both test collections. Although, the INEX measure showed quite competitive results on the ISTEX data set, on the Wikipedia and environmental science articles its results are much lower than the results obtained by RAND baseline. Thus, we can conclude that full texts can not serve as a gold standard especially in case of extractive summaries since in this case the measures are mainly reduced to the ratio of the size of a summary over the one of a full text.

One of the possible explanation why GRAD significantly outperforms the vocabulary overlap based metrics is that other measures are dealing with pure term frequencies regardless of their context. Under the circumstances, a term may be frequent in a particular context or a specific document part but not very informative for the entire text. In contrast, the GRAD metric takes into account how strongly a term is connected to all other terms in the text.

The major conclusion we can draw is that vocabulary overlap based metrics should be not used with full texts as references.

4.6 Ability to Find the Native Summary among Summaries of Different Texts

The second experiment aimed at evaluating the capacity of metrics to identify the appropriate abstract among human written summaries and follows the second part of the framework we suggested. We performed pair-wise comparison between each of 31 documents and abstracts coming from different texts

Table 2: Percentage of times when a human written abstract (H) is scored higher/lower/equally than/to a generated summary (S). The best values of $H > S$ are in bold. $^{+}/^{-}$ indicates accuracy significantly higher/lower than the accuracy of RAND.

Data	Method	RandSum			CosSum		
		$H > S$	$H < S$	$H = S$	$H > S$	$H < S$	$H = S$
ISTEX	COS	16.60%$^{-}$	83.40%	0.00%	16.33%$^{-}$	83.66%	0.01%
	ROUGE	14.49%$^{-}$	85.31%	0.19%	17.61%$^{-}$	82.13%	0.27%
	INEX	61.39%$^{+}$	38.61%	0.00%	62.04%$^{+}$	37.88%	0.08%
	RAND	49.67%	50.33%	0.00%	50.01%	49.99%	0.00%
	GRAD	**62.97%**$^{+}$	35.54%	1.48%	**79.63%**$^{+}$	19.24%	1.13%
Environmental	COS	3.2%$^{-}$	96.2%	0.6%	0.4%$^{-}$	99.0%	0.6%
	ROUGE	8.0%$^{-}$	91.2%	0.8%	14.9%$^{-}$	74.2%	0.9%
	INEX	5.8%$^{-}$	93.5%	0.7%	6.2%$^{-}$	93.1%	0.7%
	RAND	48.9%	51.1%	0.0%	49.3%	50.7%	0.00%
	GRAD	**89.8%**$^{+}$	10.2%	0.0%	**84.2%**$^{+}$	15.8%	0.0%
Wikipedia	COS	23.13%$^{-}$	76.85%	0.02%	17.17%$^{-}$	82.80%	0.02%
	ROUGE	14.48%$^{-}$	85.31%	0.21%	22.21%$^{-}$	77.37%	0.43%
	INEX	14.08%$^{-}$	85.75%	0.17%	12.66%$^{-}$	87.17%	0.17%
	RAND	49.49%	50.51%	0.00%	49.00%	51.00%	0.00%
	GRAD	**71.60%**$^{+}$	21.80%	6.60%	**92.91%**$^{+}$	1.84%	5.25%

($31 \times 31 = 961$ comparisons, 496 different text pairs). These documents on environmental sciences were selected from the ISTEX database. We used the same documents as in [9].

In Figures 1 to 4 we present the results of the subtraction of the score assigned to original abstracts and and the score of abstracts coming from the other texts when considering these 961 cross-text comparisons and various metrics. In these figures, if an abstract coming from another text is preferred by a metric over the native abstract, the point is below the zero line.

It is evident from the figure 1 that in most cases ROUGE is able to find the appropriate abstract from a set of abstracts summarizing different texts. This fact does not conflict with the conclusion from the previous section. It is important to notice that ROUGE is based on vocabulary overlap. A human-authored abstract can be paraphrased with regard to the source text, i.e. it may have different words leading lower ROUGE score. An extractive summary is coming from the source text, i.e. summary words are a subset of words from the full text leading higher ROUGE score. In contrast, when comparing human-written abstracts for different texts, the vocabulary overlap of the native abstract is higher than the one of the foreign summary because it presents another text.

In [10], it was reported that despite GRAD significantly outperforming vocabulary overlap based metrics, it fails to distinguish native abstracts from other human written summaries. Only in 15% of cases the score assigned to the native abstract was higher than the scores assigned to foreign abstracts. However, our new experiments demonstrated that GRAD is able to distinguish native abstracts from abstracts coming from other texts (see Figure 3). The difference is that here we do not normalize over the summary length

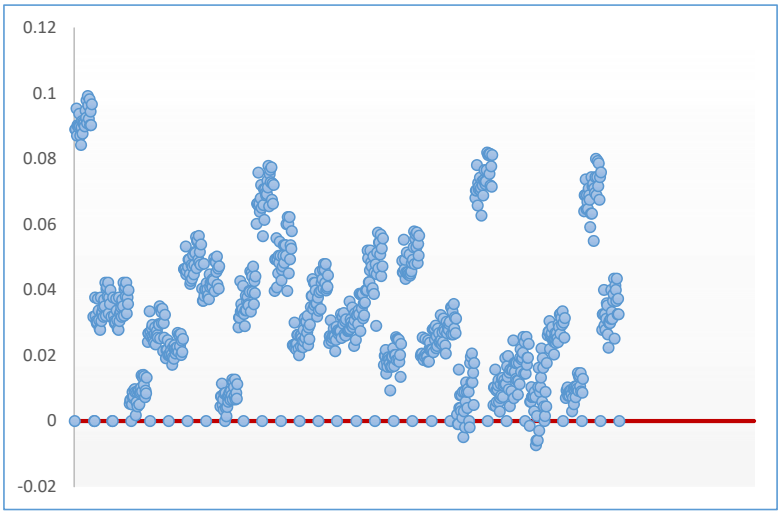

Figure 1: The difference between the ROUGE scores of native abstracts and abstracts associated with the other texts

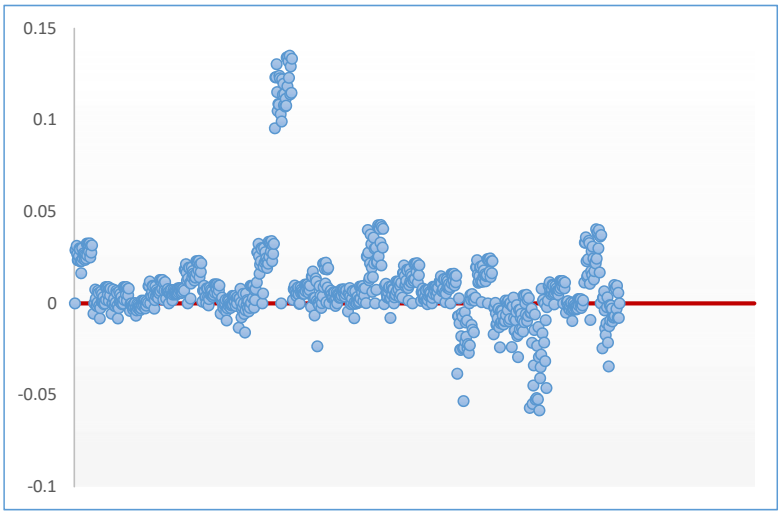

Figure 2: The difference between the INEX scores of native abstracts and abstracts associated with other the texts

and IDF is considered.

In contrast, GEM produced a lot of errors (see Figure 4). This inability to find the right abstract is due to an implementation shortcoming of GEM since for each sentence in a summary, the section class is assigned according to the class of the sentence from the full text with the maximal cosine similarity. The section class would be assigned in any case, even if no section match at all the sentence under consideration (the first section class would be assigned). GEM was not designed to search for right abstract. Thus, in contrast to traditional metrics, GEM compares overlap scores of sentences within a particular abstract rather than different summaries. Assuming that the summary is the native one, GEM aims to detect the presence of sections from a predefined list based on a set of rules (see [9] for more details). Vocabulary overlap-based metrics did not faced this issue: the trend is completely opposite (see Figures 1 and 2).

Figures 1 and 3 show that ROUGE and GRAD perform equally well in their ability to assign a higher

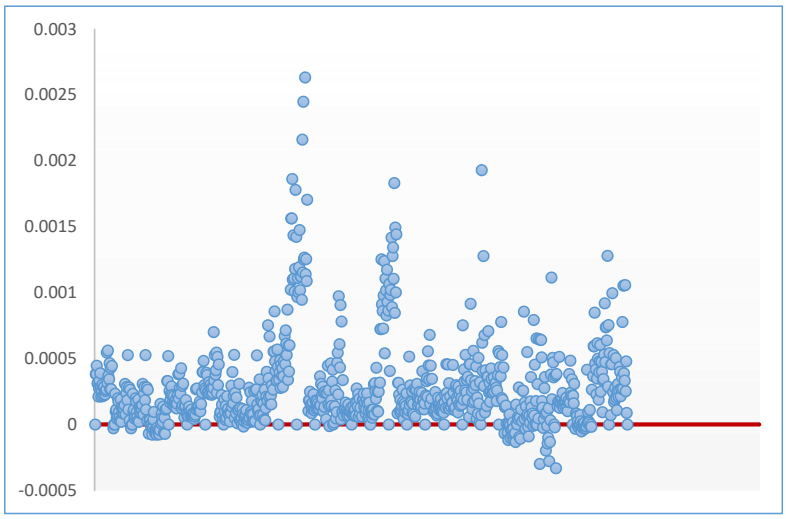

Figure 3: The difference between the GRAD scores of native abstracts and abstracts associated with the other texts

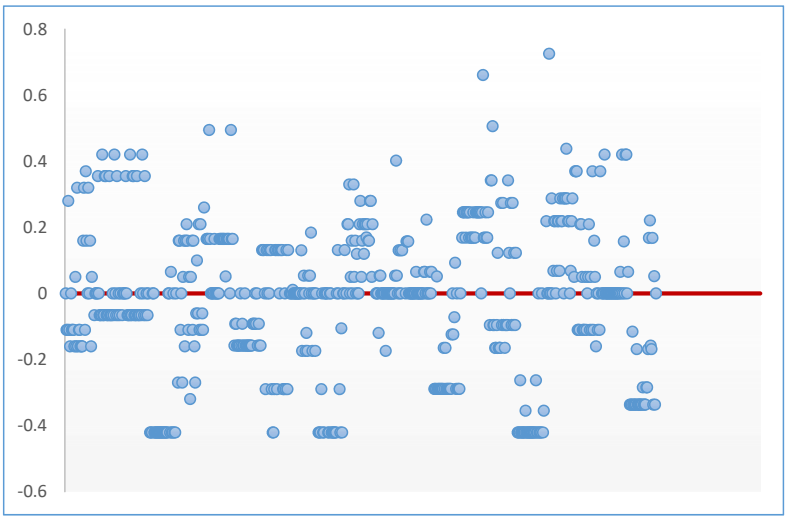

Figure 4: The difference between the GEM scores of native abstracts and abstracts associated with the other texts

score to abstracts originating from the target document than abstracts originating from other documents.

Thus, we can conclude that vocabulary overlap-based metrics and GRAD are able to find the native abstracts, while GEM can not capture this difference.

5 Discussion and Conclusion

In this paper, we discuss a wide range of evaluation frameworks for automatic summarization, including data sets and metrics used, as well as main competitions in the field, for the last 18 years. Manual assessment is time and money consuming and, therefore, not applicable to real time scenarios or large collections. The expert nature of metrics, such as responsiveness, retention rate, metrics used in INEX Snippet Retrieval track, makes it impossible for re-use, since manual assessment requires expert intervention for each summary under evaluation.

Table 3: Descriptive statistics on Wikipedia data

	INEX	ROUGE	COS	GRAD
Mean	0.0671	0.0666	0.2788	0.0002
Standard error	0.0004	0.0003	0.0004	0.0000
Median	0.0297	0.0501	0.2825	0.0001
Standard deviation	0.0838	0.0603	0.0860	0.0050
Variance	0.0070	0.0036	0.0074	0.0000
Kurstosis	1.8581	10.6040	-0.1688	16482.4494
Skewness	1.5985	2.3678	-0.1224	120.1888
Minimum	0.0000	0.0000	0.0000	0.0000
Maximum	0.5255	0.8750	0.6173	0.7757

Moreover, on average, assessment agreement is about 70% due to the fact that judges may have different opinions about summary quality and evaluation metrics [29]. Inter-rater agreement is traditionally measured by Cohen's kappa [73] or Krippendorff's alpha in the case of an arbitrary number of coders [74]. This average 70% agreement between judges is also a big issue for expert metrics since it means that for an objective evaluation at least, all summaries of the same text should be evaluated by the same annotator. Otherwise, there is a possibility that different methods are treated in different ways by different annotators. In other words, one judge can be very strict, another very generous.

In contrast to manual assessment, gold-standard based evaluation is less subjective and theoretically re-usable for new methods.

Traditional evaluation frameworks (data sets and metrics) have several drawbacks. Sometimes the re-use of the gold standard by non-participants is problematic since often the ground truth is constructed only from the results initially submitted by the participants, i.e. some relevant results may not occur in a gold standard.

The use of the full text rather than a set of reference summaries for summary evaluation provides poor results [8] since traditional overlap-based metrics are mainly reduced to the ratio of the size of a summary to the size of a full text since they are designed to be compared with summaries created by humans.

To provide evidence, we proposed a completely automatic framework for the evaluation of metrics to assess automatically produced summaries which does not require any human annotation. We conducted experiments on Wikipedia data set and a collection of scientific articles from the ISTEX database. The obtained results show that metrics based on vocabulary overlap are not suitable for measuring the quality of a summary with regard to a full text. However, the GRAD measure significantly outperformed overlap-based baselines on both test collections in distinguishing human written abstracts from generated summaries of poor quality.

The second problem with the full text as a gold standard is that every completely automatic metric without hidden data can be optimized, i.e. transformed into an automatic method that maximizes this metric. This optimization does not necessarily correlate with the quality of the produced summary.

All existing metrics, except GEM, have relative values allowing candidate summaries to be ranked, but they are not applicable for comparison of an isolated summary with the full text nor for comparison of metric scores for summaries of different documents. Thus, they enable us to answer the question *Which summary is better?* but not *Is this summary a good one?*

Another problem with the existing metrics (except GEM) is their output values. Theoretically, the majority of metrics are normalized, but in practice, the values tend to be quite small, usually ROUGE score is less than 0.2 (see Table 3). This means that the isolated values produced by these metrics cannot be interpreted, i.e. it is impossible to say, for example, whether the score of ROUGE-1 equal to 0.17 is good or not.

Structured abstracts tend to be informative [75]. One of the metrics taking into account document structure is BM25F [76], a field-based extension of Okapi's BM25 from the information retrieval. However, it is not suitable for summary scoring since it also gives a relative score allowing search result ranking. GEM considers the document structure but its major problem is that it is unable to distinguish summaries coming from different texts. This is an implementation drawback of the metric GEM explained in the previous section.

The analysis of the existing evaluation metrics shows the following future perspectives in the area:

- absolute values of metrics would allow us to extend their application, for example, to the domain of education or to the automatic evaluation of scientific abstract quality and even automatic reviewing;
- the interpretability of metrics remains problematic (i.e. how to interpret the score of an individual summary and answer the question *Is this summary a good one?*);
- vocabulary overlap-based metrics should be not used with full texts as references. Metrics not based on vocabulary overlap should be applied in the case where a full text is used as a reference.

Acknowledgements

The authors would like to thanks Marianne Noël, Frédérique Bordignon, Nicolas Turenne, Anton Firsov, and Université Paris Descartes, Frontières du vivant.

References

[1] B. Marr, Big Data: 20 Mind-Boggling Facts Everyone Must Read.
URL https://www.forbes.com/sites/bernardmarr/2015/09/30/big-data-20-mind-boggling-facts-everyone-must-read/

[2] A. Jorge, R. Campos, A. Jatowt, S. Nunes, C. Rocha, J. P. Cordeiro, A. Pasquali, V. Mangaravite, ECIR 2018: Text2story Workshop - Narrative Extraction from Texts, ACM SIGIR Forum 52 (1) (2018) 150–152. doi:10.1145/3274784.3274801.
URL http://dl.acm.org/citation.cfm?doid=3274784.3274801

[3] ANSI, American National Standard for Writing Abstracts, Tech. rep., American National Standards Institute, Inc., New York, NY, (ANSI Z39.14.1979) (1979).

[4] J. Torres-Moreno, Resume automatique de documents : une approche statistique, Hermes-Lavoisier, 2011.

[5] H. Saggion, G. Lapalme, Generating Indicative-Informative Summaries with SumUM, Association for Computational Linguistics 28 (4) (2002) 497–526.

[6] K. Owczarzak, H. T. Dang, Overview of the TAC 2011 summarization track: Guided task and AESOP task, in: Proceedings of the Text Analysis Conference (TAC 2011), Gaithersburg, Maryland, USA, November, 2011.

[7] P. Bellot, V. Moriceau, J. Mothe, E. SanJuan, X. Tannier, INEX tweet contextualization task: Evaluation, results and lesson learned, Inf. Process. Manage. 52 (5) (2016) 801–819. doi:10.1016/j.ipm.2016.03.002.
URL https://doi.org/10.1016/j.ipm.2016.03.002

[8] A. Louis, A. Nenkova, Automatically assessing machine summary content without a gold standard, Comput. Linguist. 39 (2) (2013) 267–300. doi:10.1162/COLI_a_00123.
URL http://dx.doi.org/10.1162/COLI_a_00123

[9] L. Ermakova, F. Bordignon, N. Turenne, M. Noel, Is the Abstract a Mere Teaser? Evaluating Generosity of Article Abstracts in the Environmental Sciences, Frontiers in Research Metrics and Analytics 3. doi:10.3389/frma.2018.00016.
URL https://www.frontiersin.org/articles/10.3389/frma.2018.00016/full

[10] L. Ermakova, A. Firsov, GRAD: A Metric for Evaluating Summaries, in: Conférence francophone en Recherche d'Information et Applications (CORIA-2018), 2018.

[11] C.-Y. Lin, E. Hovy, 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, Association for Computational Linguistics, 2003, pp. 71–78.

[12] H. T. Dang, Overview of duc 2006, in: In Proceedings of HLT-NAACL 2006, 2006.

[13] H. T. Dang, Duc 2005: Evaluation of question-focused summarization systems, in: Proceedings of the Workshop on Task-Focused Summarization and Question Answering, SumQA '06, Association for Computational Linguistics, Stroudsburg, PA, USA, 2006, pp. 48–55.
URL http://dl.acm.org/citation.cfm?id=1654679.1654689

[14] C. Macdonald, I. Ounis, The trec blogs06 collection: Creating and analysing a blog test collection, Department of Computer Science, University of Glasgow Tech Report TR-2006-224 1 (2006) 3–1.

[15] J. Allan, D. Harman, E. Kanoulas, D. Li, C. Van Gysel, E. Vorhees, Trec 2017 common core track overview, TREC, 2017.

[16] A. Zubiaga, D. Spina, E. Amigó, J. Gonzalo, Towards real-time summarization of scheduled events from twitter streams, in: Proceedings of the 23rd ACM Conference on Hypertext and Social Media, HT '12, 2012, pp. 319–320. doi:10.1145/2309996.2310053.

[17] C. Shen, F. Liu, F. Weng, T. Li, A participant-based approach for event summarization using twitter streams, in: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2013, pp. 1152–1162.

[18] J. Lin, A. Roegiest, L. Tan, R. McCreadie, E. Voorhees, F. Diaz, Overview of the trec 2016 real-time summarization track, in: Proceedings of the 25th text retrieval conference, TREC, Vol. 16, 2016.

[19] Text mining software, text analytics software, big data management, knowledge discovery – linguamatics.

[20] E. Meij, W. Weerkamp, M. de Rijke, Adding semantics to microblog posts, in: Proceedings of the fifth ACM international conference on Web search and data mining, WSDM '12, ACM, New York, NY, USA, 2012, p. 563–572. doi:10.1145/2124295.2124364.
URL http://doi.acm.org/10.1145/2124295.2124364

[21] E. SanJuan, V. Moriceau, X. Tannier, P. Bellot, J. Mothe, Overview of the inex 2011 question answering track (qainex), in: S. Geva, J. Kamps, R. Schenkel (Eds.), Focused Retrieval of Content and Structure, Vol. 7424 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2012, pp. 188–206.
URL http://dx.doi.org/10.1007/978--3--642--35734--3_17

[22] P. Bellot, V. Moriceau, J. Mothe, E. SanJuan, X. Tannier, Overview of INEX 2013, in: Information Access Evaluation. Multilinguality, Multimodality, and Visualization, Vol. 8138 of LNCS, 2013, pp. 269–281. doi:10.1007/978-3-642-40802-1_27.

[23] R. Schenkel, F. M. Suchanek, G. Kasneci, YAWN: A Semantically Annotated Wikipedia XML Corpus, in: BTW, 2007, pp. 277–291.

[24] M. Trappett, S. Geva, A. Trotman, F. Scholer, M. Sanderson, Overview of the INEX 2011 Snippet Retrieval Track, in: S. Geva, J. Kamps, R. Schenkel (Eds.), Focused Retrieval of Content and Structure, Vol. 7424 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2012, pp. 283–294.
URL http://dx.doi.org/10.1007/978-3-642-35734-3_27

[25] D. Bollegala, N. Okazaki, M. Ishizuka, A bottom-up approach to sentence ordering for multi-document summarization, Information processing & management 46 (1) (2010) 89–109.

[26] Z. Lin, H. T. Ng, M.-Y. Kan, Automatically evaluating text coherence using discourse relations, in: Proc. of the 49th Annual Meeting of the ACL: Human Language Technologies - vol. 1, ACL, Stroudsburg, PA, USA, 2011, pp. 997–1006.

[27] M. Elsner, J. L. Austerweil, E. Charniak, A Unified Local and Global Model for Discourse Coherence., in: HLT-NAACL, 2007, pp. 436–443.

[28] A. Louis, A. Nenkova, A coherence model based on syntactic patterns, in: Proc. of EMNLP-CoNLL '12, ACL, Stroudsburg, PA, USA, 2012, pp. 1157–1168.

[29] S. Gholamrezazadeh, M. A. Salehi, B. Gholamzadeh, A comprehensive survey on text summarization systems, Computer Science and its Applications (2009) 1–6.

[30] Y. Seki, Automatic summarization focusing on document genre and text structure, ACM SIGIR Forum 39 (1) (2005) 65–67.

[31] H. Saggion, D. Radev, S. Teufel, W. Lam, S. M. Strassel, Developing Infrastructure for the Evaluation of Single and Multi-document Summarization Systems in a Cross-lingual Environment, LREC (2002) 747–754.

[32] K. Owczarzak, J. M. Conroy, H. T. Dang, A. Nenkova, 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, 2012, pp. 1–9.
URL http://dl.acm.org/citation.cfm?id=2391258.2391259

[33] J. Lin, S. Mohammed, R. Sequiera, L. Tan, N. Ghelani, M. Abualsaud, R. Mc-Creadie, D. Milajevs, E. Voorhees, Overview of the trec 2017 real-time summarization track (notebook draft), in: Pre-Proceedings of the 26th Text REtrieval Conference, TREC, 2017.

[34] C.-Y. Lin, ROUGE: a package for automatic evaluation of summaries, Text Summarization Branches Out: Proc. of the ACL–04 Workshop (2004) 74–81.

[35] D. Radev, S. Teufel, H. Saggion, W. Lam, J. Blitzer, A. Elebi, H. Qi, E. Drabek, D. Liu, Evaluation of text summarization in a cross-lingual information retrieval framework, Tech. rep., Center for Language and Speech Processing, Johns Hopkins University, Baltimore (2002).

[36] E. Hovy, S. Tratz, Summarization evaluation using transformed basic elements, Proceedings TAC 2008.

[37] S. Bangalore, O. Rambow, S. Whittaker, Evaluation metrics for generation, Proceedings of the first international conference on (2000) 1–8.

[38] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, BLEU: a method for automatic evaluation of machine translation, in: Proceedings of the 40th annual meeting on association for computational linguistics, Association for Computational Linguistics, 2002, pp. 311–318.

[39] M. Denkowski, A. Lavie, Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems, Proc. of the EMNLP 2011 Workshop on Statistical Machine Translation (2011) 85–91.

[40] D. Klakow, J. Peters, Testing the correlation of word error rate and perplexity, Speech Communication 38 (1) (2002) 19 – 28. doi:https://doi.org/10.1016/S0167-6393(01)00041-3.
URL http://www.sciencedirect.com/science/article/pii/S0167639301000413

[41] P. Bellot, A. Doucet, S. Geva, S. Gurajada, J. Kamps, G. Kazai, M. Koolen, A. Mishra, V. Moriceau, J. Mothe, M. Preminger, E. SanJuan, R. Schenkel, X. Tannier, M. Theobald, M. Trappett, Q. Wang, Overview of inex, in: Information Access Evaluation. Multilinguality, Multimodality, and Visualization - 4th International Conference of the CLEF Initiative, CLEF 2013, Valencia, Spain, September 23-26, 2013. Proceedings, 2013, pp. 269–281.

[42] E. SanJuan, P. Bellot, V. Moriceau, X. Tannier, Overview of the inex 2010 question answering track (qa@inex), in: S. Geva, J. Kamps, R. Schenkel, A. Trotman (Eds.), Comparative Evaluation of Focused Retrieval, Vol. 6932 of Lecture Notes in Computer Science, Springer Berlin / Heidelberg, 2011, pp. 269–281.

[43] H. Saggion, J.-M. Torres-Moreno, I. da Cunha, E. SanJuan, Multilingual summarization evaluation without human models, in: 23rd Int. Conf. on Computational Linguistics, COLING '10, ACL, Beijing, China, 2010, pp. 1059–1067.

[44] L. A. Cabrera-Diego, J.-M. Torres-Moreno, B. Durette, Evaluating Multiple Summaries Without Human Models: A First Experiment with a Trivergent Model, Springer International Publishing, Cham, 2016, pp. 91–101. doi:10.1007/978-3-319-41754-7_8.
URL http://dx.doi.org/10.1007/978-3-319-41754-7_8

[45] M. Campr, K. Ježek, Comparing Semantic Models for Evaluating Automatic Document Summarization, Springer International Publishing, Cham, 2015, pp. 252–260. doi:10.1007/978-3-319-24033-6_29.
URL http://dx.doi.org/10.1007/978-3-319-24033-6_29

[46] J.-P. Ng, V. Abrecht, 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, 2015, pp. 1925–1930.
URL http://aclweb.org/anthology/D15-1222

[47] A. Nenkova, R. Passonneau, K. McKeown, The pyramid method: Incorporating human content selection variation in summarization evaluation, ACM Trans. Speech Lang. Process. 4 (2). doi:10.1145/1233912.1233913.
URL http://doi.acm.org/10.1145/1233912.1233913

[48] E. W. Dijkstra, A note on two problems in connexion with graphs, Numerische Mathematik 1 (1) (1959) 269–271. doi:10.1007/BF01386390.
URL http://dx.doi.org/10.1007/BF01386390

[49] R. Gunning, The technique of clear writing, McGraw-Hill, 1968.
URL https://books.google.fr/books?id=vJZpAAAAMAAJ

[50] R. Flesch, A new readability yardstick., Journal of Applied Psychology 32 (3) (1948) p221 – 233.
URL http://libezproxy.open.ac.uk/login?url=http://search.ebscohost.com.

libezproxy.open.ac.uk/login.aspx?direct=true&db=pdh&AN=apl-32-3-221&site=ehost-live&scope=site

[51] L. Si, J. Callan, A statistical model for scientific readability, in: Proceedings of the Tenth International Conference on Information and Knowledge Management, CIKM '01, ACM, New York, NY, USA, 2001, pp. 574–576. doi:10.1145/502585.502695.
URL http://doi.acm.org/10.1145/502585.502695

[52] K. Collins-Thompson, J. Callan, A language modeling approach to predicting reading difficulty, Proceedings of HLT/NAACL 4.

[53] M. Heilman, K. Collins-Thompson, M. Eskenazi, An analysis of statistical models and features for reading difficulty prediction, in: Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications, EANL '08, Association for Computational Linguistics, Stroudsburg, PA, USA, 2008, pp. 71–79.
URL http://dl.acm.org/citation.cfm?id=1631836.1631845

[54] S. E. Petersen, M. Ostendorf, A machine learning approach to reading level assessment, Comput. Speech Lang. 23 (1) (2009) 89–106. doi:10.1016/j.csl.2008.04.003.
URL http://dx.doi.org/10.1016/j.csl.2008.04.003

[55] L. Feng, M. Jansche, M. Huenerfauth, N. Elhadad, A comparison of features for automatic readability assessment, in: Proceedings of the 23rd International Conference on Computational Linguistics: Posters, COLING '10, Association for Computational Linguistics, Stroudsburg, PA, USA, 2010, pp. 276–284.
URL http://dl.acm.org/citation.cfm?id=1944566.1944598

[56] A. J. Stenner, I. Horablin, D. R. Smith, M. Smith, The Lexile Framework. Durham, NC: Metametrics.

[57] J. S. Chall, E. Dale, Readability revisited: The new Dale–Chall readability, MA: Brookline Books, Cambridge, 1995.

[58] E. Fry, A readability formula for short passages, Journal of Reading 8 (1990) 594–597, 33.

[59] J. Tavernier, P. Bellot, Combining relevance and readability for INEX 2011 question–answering track (2011) 185–195.

[60] K. Collins-Thompson, J. Callan, A Language Modeling Approach to Predicting Reading Difficulty, Proceedings of HLT/NAACL 4.

[61] A. Mutton, M. Dras, S. Wan, R. Dale, Gleu: Automatic evaluation of sentence-level fluency, ACL–07 (2007) 344–351.

[62] S. Wan, R. Dale, M. Dras, Searching for grammaticality: Propagating dependencies in the viterbi algorithm, Proc. of the Tenth European Workshop on Natural Language Generation.

[63] S. Zwarts, M. Dras, Choosing the right translation: A syntactically informed classification approach, Proc. of the 22nd International Conference on Computational Linguistics (2008) 1153–1160.

[64] J. Chae, A. Nenkova, Predicting the fluency of text with shallow structural features: case studies of machine translation and human–written text, Proc. of the 12th Conference of the European Chapter of the ACL (2009) 139–147.

[65] L. Si, J. Callan, A statistical model for scientific readability, Proceedings of the tenth international conference on Information and knowledge management (2001) 574–576.

[66] R. Barzilay, N. Elhadad, K. R. McKeown, Inferring strategies for sentence ordering in multidocument news summarization, Journal of Artificial Intelligence Research (2002) 35–5517.

[67] G. Lebanon, J. Lafferty, Cranking: Combining rankings using conditional probability models on permutations, Machine Learning: Proceedings of the Nineteenth International Conference (2002) 363–370.

[68] M. Lapata, Probabilistic text structuring: Experiments with sentence ordering, Proc. of ACL (2003) 542–552.

[69] A. Stenner, I. Horabin, D. R. Smith, M. Smith, The lexile framework, Durham, NC: MetaMetrics.

[70] L. Ermakova, Automatic Sentence Ordering Assessment Based on Similarity, in: Proc. of EVIA 2016, Tokyo, Japan, 07/06/2016, NII, 2016.

[71] L. Ermakova, J. Mothe, A. Firsov, A Metric for Sentence Ordering Assessment Based on Topic-Comment Structure (short paper), in: ACM SIGIR Special Interest Group on Information Retrieval (SIGIR), Tokyo, Japan, 07/08/2017-11/08/2017, 2017, selection rate 30

[72] R. Deveaud, V. Moriceau, J. Mothe, E. SanJuan, Informativeness for adhoc ir evaluation: A measure that prevents assessing individual documents, in: European Conference on Information Retrieval, Springer, 2016, pp. 818–823.

[73] J. Carletta, Assessing agreement on classification tasks: The kappa statistic, Computational Linguistics 22 (2) (1996) 249–254.

[74] K. Krippendorff, Content Analysis: An Introduction to Its Methodology, Sage, 2004.
URL https://books.google.co.za/books?id=q657o3M3C8cC

[75] P. Fontelo, A. Gavino, R. F. Sarmiento, Comparing data accuracy between structured abstracts and full-text journal articles: implications in their use for informing clinical decisions., Evidence-based medicine 18 (6) (2013) 207–11. doi:10.1136/eb-2013-101272.
URL http://www.researchgate.net/publication/240308203_Comparing_data_accuracy_between_structured_abstracts_and_full-text_journal_articles_implications_in_their_use_for_informing_clinical_decisions

[76] S. Robertson, H. Zaragoza, M. Taylor, Simple BM25 Extension to Multiple Weighted Fields, in: Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management, CIKM '04, ACM, New York, NY, USA, 2004, pp. 42–49. doi:10.1145/1031171.1031181.
URL http://doi.acm.org/10.1145/1031171.1031181