The CNN-Corpus: A Large Textual Corpus for Single-Document Extractive Summarization

Rafael Dueire Lins
rdl.ufpe@gmail.com
UFPE/UFRPE, Recife, Brazil

Hilario Oliveira, Luciano Cabral
Jamilson Batista, Bruno Tenorio
UFPE, Recife, Brazil

Rafael Ferreira, Rinaldo Lima
Gabriel de França Pereira e Silva
UFRPE, Recife, Brazil

Steven J. Simske
Colorado State University
Fort Collins, CO, USA

ABSTRACT

This paper details the features and the methodology adopted in the construction of the CNN-corpus, a test corpus for single document extractive text summarization of news articles. The current version of the CNN-corpus encompasses 3,000 texts in English, and each of them has an abstractive and an extractive summary. The corpus allows quantitative and qualitative assessments of extractive summarization strategies.

CCS CONCEPTS

• **Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability.

KEYWORDS

Single-document Summarization, Corpus, Extractive Summarization, Multi-language Summarization

ACM Reference Format:

Rafael Dueire Lins, Hilario Oliveira, Luciano Cabral, Jamilson Batista, Bruno Tenorio, Rafael Ferreira, Rinaldo Lima, Gabriel de França Pereira e Silva, and Steven J. Simske. 2019. The CNN-Corpus: A Large Textual Corpus for Single-Document Extractive Summarization. In *ACM Symposium on Document Engineering 2019 (DocEng '19), September 23–26, 2019, Berlin, Germany*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3342558.3345388

1 INTRODUCTION

A summary is a short version of a text document, which should provide enough information about its content for the reader to decide on his interest in reading the entire document [15]. Since the advent of the Internet, the amount of information is growing at a rate unprecedented in human history. Readers cannot keep track of the large and quickly expanding volume of data in any area of knowledge. Thus, automatically creating good quality summaries has become an issue of paramount importance.

Automatic Text Summarization (ATS) could be defined as the computer process of creating a condensed version (summary) from a single document (single-document) or a collection of documents (multi-document), keeping only the most relevant information [22]. Automatic summarization methods for text have been an active object of research in Information and Computational Sciences since the works of Luhn in 1958 [20] and Edmundson in 1969 [3]. Nonetheless, the fundamental challenges in the area remain unsolved. The main goal in this area is to develop a method that automatically creates a summary with a similar quality to one created by a qualified human reader. This means that the summary must contain important information on the content and be written in a coherent style. Such a summary is usually referred to as *abstractive summary*. Although automatic methods for generating abstractive summaries are the final goal in this research area, simpler models of summaries, such as, *extractive* [28] [11] and, more recently, *semi-extractive summaries* [10] have been heavily studied. An *extractive* or *cut-and-paste* summary is created by selecting sentences *verbatim* from the text. Such a summary not necessarily offers coherent information and reading fluency as sentences are extracted from various positions of the original text. The selected sentences may either not be self-contained or have redundant or irrelevant information. A *semi-extractive summary* is created starting from an extractive summary, and then the coherence of the final text is increased by removing redundancies, mapping co-references, etc.

An important open problem is the development of a methodology for evaluating the quality of the generated summaries [23] [27] taking into account several aspects, such as the semantic equivalence, coherence, and cohesion of the generated and the original texts. Indeed, having corpora that allow quantifying the results of automatic text summarization, reducing the subjectivity of the assessment, is central to increase the validity of the results obtained in this area.

The lack of appropriate corpora for text summarization is justified by the exhaustive manual effort required in the annotation task. For each document, a human must read the entire content at least once and manually create a summary (gold standard). Several issues may appear at this point. First, there is more than one acceptable summary for any given content, due to summarizer subjectivity, the informational user needs, the wideness and richness of the vocabulary used, etc. Second, when one needs to scale up the corpus, a large team of experts in the subject must be set in place and work for many man-hours. Even if one can assemble such a team, they will soon start to experience fatigue, which may affect

DocEng '19, September 23–26, 2019, Berlin, Germany
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-6887-2/19/09…$15.00
https://doi.org/10.1145/3342558.3345388

the quality of the summaries generated. Therefore, those issues lead to a trade-off between the size of the corpus and its quality. In fact, [21] states that "*given the cost and tediousness of the annotation process, it is very unlikely that we will ever manually annotate for textual importance sufficiently large corpora*".

One of the possible requirements is that the corpus should be large enough so that any experimental results using it can be considered statistically relevant. However, what is the appropriate minimum size for it, except for following the popular knowledge which states that *the larger, the better*, remains an open question. The quality requirements are even harder to specify. First, one needs the original texts to be clearly written by native speakers, using standard and grammatically correct language, reporting on non-controversial subjects, with a straightforward outlook. Such high-standard documents should be summarized down to a specified number of sentences by more than one expert on the subject, following a uniform methodology. Second, the minimal set of features used for each document in a summarization corpus is usually the full original content and one human-made extractive summary, which is a summary consisting of sentences extracted as written in the original content. However, for an appropriate test corpus, a set of additional metadata, title, authors, primary source, categories, keywords, published and collected date is needed. Additional summaries, possibly abstractive, with different views and compression rate may also be part of the test corpus.

Once a test corpus for text summarization is built, several problems can be addressed, and the research can focus on the development of new summarization methods and their evaluation, instead of going through the same process of constructing a corpus, wasting a lot of research time. Furthermore, there is no reason to re-implement the methods proposed in the literature to assess the quality of a new one, if the same corpus is used in the experiments. This allows better measuring the progress of the field.

Currently, most of the corpora in the literature for single-document summarization have less than one thousand documents, and many of them do not contain a broad diversity of subjects. For example, the corpora of the Document Understanding Conference [1] (DUC) competitions of 2001 (DUC 2001) and 2002 (DUC 2002), which are the most widely used for assessing single-document systems [9], have 308 and 533 documents, respectively. Another important aspect of the current single-document corpora is that most of them have only abstractive summaries. This type of reference summary is important, because if new summarization approaches can generate similar summaries, this indicates that such methods are getting closer to the summaries generated by humans. However, the current summary evaluation measures have a significant limitation, as most of them only consider the lexical similarity between texts. Thus, elements such as paraphrasing, which are commonly used in the production of abstractive summaries, are not taken into consideration during the evaluation process. Therefore, the comparison between extractive and abstractive summaries further demonstrates these constraints, which may lead to incorrect or at least inaccurate conclusions.

This paper details the features and development methodology of the CNN-corpus for single-document extractive text summarization of news articles. The starting point for such a corpus were articles from the CNN website[12]. Besides meeting the high linguistic standards listed above needed for the documents in a test corpus, each of the texts selected has a good quality abstractive summary written by their original authors, the *highlights*, which served as the basis for generating the *gold standard*, an extractive summary for each text. The gold standard may serve as the reference for qualitative and quantitative assessment of automatic summarization techniques. The selection of the articles from the CNN website, besides having the highlights (not all pages have them), met the restriction of being self-contained in such a way that the understanding of the text should not depend on images, graphical elements, videos, and others, that may be present or referenced in the web page of the news article. The current version of the CNN-corpus encompasses 3,000 texts written in English. Several tests were performed to assess the performance of the building process and the quality of the resulting CNN-corpus in English.

2 RELATED WORKS

This section presents a brief overview of the available text summarization corpora and their development methodology, which are summarized in Table 1. The first and second columns are the corpus names and year of its announcement, respectively. The *Quality* column of Table 1 indicates the number of experts involved in the generation of each of the summaries. An interrogation mark is used whenever that information is not available.

The other columns in Table 1 are:

- **Building Process**: the building process feature indicates if the corpus was created manually, with some computer assistance or completely automatic.
- **Type**: the summary type indicates whether the corpus consists of abstracts, extracts, and/or highlight summaries;
- **Mode**: the summary mode indicates whether a single or several documents (i.e. multi-document) were used to create each summary;
- **Size** the size informs the number of pairs summary/text and, in parenthesis, the number of multi-document summaries in the corpus;
- **Language**: the languages present in each corpus are listed in the last column.

This work focuses on corpora with documents written in English.

In a manual summary building process, each summary is generated by a single person or a team of people, who needs to read each text entirely before starting to write the summary. In general, some guidelines are used for instructing the humans on basic rules for the summarization task to provide some standardization. The main advantage of such an approach is yielding good quality final abstractive summaries. On the other hand, such a process leads to a high development cost and subjectivity. Such factors, in general, are the main reasons large test corpora for extractive or abstractive summarization do not exist.

Several corpora were built employing only one expert per summary to increase the size of the corpus and decrease its generation cost. This decision may affect the quality of the corpus because humans are prone to subjectivity and mistakes, which can be mitigated

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

[2] 1 www.cnn.com

Table 1: Overview of text summarization corpora

Name	Year	Building Process	Quality	Type	Mode	Size	Language
SUMMAC	1998	Manual	High, authors	Abstract	Single	183	English
Jing	1999	Manual	?	Abstract	Single	300	English
DUC	2001	Manual	High, 10 analysts	Abstract	Both	60 sets of 10 documents	English
	2002	Manual	High, 10 analysts	Abstract, Extract	Both	60 sets of 10 documents	English
	2003	Manual	High	Abstract	Both	900	English
	2004	Manual	High	Abstract	Both	1,250	English, Arabic
Microsoft	2007	Manual	High, authors	Highlight	Single	1,365	English
New York Times	2007	Manual	Mid, 1 librarian	Abstract	Single	650,000	English
Hasler	2003	Computer-assisted	Mid, 1 expert	Extract	Single	163	English
CNN (HP-UFPE)	2012	Computer-assisted	High, 3 experts	Extract	Single	400	English
	2018	**Semi-automatic**	**High, 5 experts**	**Extract, Highlight**	**Single**	**3,000**	**English**
Kupiec	1995	Semi-automatic	?	Extract	Single	188	English
Teufel	1997	Semi-automatic	?	Extract	Single	202	English
Marcu	1999	Automatic	Low	Extract	Single	6,942	English

by using more experts per summary. Therefore, employing only one human per summary is not enough to guarantee a high level of quality. As a consequence of the simple central limit theorem, any scientific research that uses such summaries as a gold standard may have the validity of their experimental results affected as well. Nevertheless, several corpora were produced using this approach.

The SUMMAC corpus[3] consists of 183 documents in English collected from the Computation and Language collection[4]. The documents are scientific papers, which appeared in Association for Computational Linguistics (ACL) sponsored conferences, and they have an abstract written by the authors of the paper.

Jing and Mckeown (1999) [13] built a corpus of 300 news articles written in English collected various news portals, having a wide variety of subjects. The documents have 1,642 sentences and were summarized by humans, with abstracts ranging from 2 to 21 sentences. The authors analyzed the summaries generated using the proposed decomposition approach and found that 315 (19%) of sentences present the summaries did not have a mapping with no original sentence of documents, 686 (42%) of the sentences had a direct mapping with an original sentence in the documents, 592 (36%) of the sentences were mapped to 2 or 3 sentences of the documents, and only 49 (3%) of the sentences were mapped for four or more sentences. This result demonstrated that the vast majority of sentences (78%) written by human summarizers during the construction of the abstracts had little or no changes regarding the original sentences in the documents.

The Document Understanding Conference [25, 26] promoted annual competitions of text summarization between 2001 and 2004. Annually, a committee constructed a corpus consisting of several sets of documents each containing on average 10 documents in English, and once in Arabic (2004), gathered from TDT, TIPSTER, TREC, and AQUAINT collections. In general, each document has abstracts and extracts produced manually based on guidelines. The DUC 2002 test set was developed by a team of ten NIST information analysts who followed the construction strategy detailed in [25]. The DUC 2001 developers produced a set of 60 documents (30 for training and 30 for tests), while DUC 2002 also generated a set of 60 documents all for tests. The DUC 2005-2007 tasks [25] were question-focused summarization tasks using multiple documents to provide the answer. Hence, they are out of scope and were not included in Table 1. Assuming that the same strict development methodology described in [26] was followed in the other DUC data sets, one may assign a high quality standard to all instances of the DUC data sets.

The Microsoft corpus [30] consists of 1,365 documents in English gathered from the CNN website. Each document was extracted by hand, where up to 50 documents were collected per day. The documents were hand-collected on consecutive days during the month of February 2007. Each document includes the title, date, story highlights and article text.

The New York Times corpus[5] is currently, to the best of the authors' knowledge, the largest corpus for single-document summarization found in the literature for news article written in English. The corpus has over 1.8 million articles written and published by The New York Times in the years 1987 to 2007. In an analysis carried out on this corpus by the authors of this paper, it was found that only about 650,000 documents (about 35 %) have summaries. Most summaries are a single sentence in length and have between 1 to 10 words. With such short summaries, it is impossible to provide the central information of the news; such abstracts are more similar to a title than a real summary. During such analysis, it was found that the ratio between sentences or words of the original documents regarding abstracts is very low, i.e., the compression rate used when creating the abstract was too high. Currently, the corpus is provided by the Linguistic Data Consortium[6]. At present, the NWT corpus has no quality standards to be useful in any kind of summarization research, unfortunately.

In a computer-assisted building process, the summary is created by a human with the aid of a computer program that can automate some summarization tasks, such as, spell checking and suggesting synonyms. This approach, compared to the manual one, reduces the cost, time and subjectivity while improving the quality of the final summary. There is the additional cost of developing the software platform and training users, however. Such an approach is cost-effective only if the final corpus size is large enough. In fact, only a few corpora use this approach, which was used in the corpus described here.

Reference [12] reports on the development of a corpus consisting of 163 documents in English. It was built with texts drawn from the Reuters corpus and from popular science texts of the British

[3] http://www.itl.nist.gov/iaui/894.02/related_projects/tipster_summac/

[4] http://arxiv.org/list/cmp-lg/recent

[5] https://catalog.ldc.upenn.edu/LDC2008T19

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

National Corpus (BNC). In such a corpus, 113 of the texts were annotated by only one person and the remaining by two or three experts. An annotation tool was used to facilitate the task of marking the 6,584 sentences as essential or important, and a guideline was used for maintaining the consistency of the annotation task. A guideline was developed to help annotators during the process of annotation of the sentences. The sentences marked as *essential* were considered as more relevant than the ones marked as *important*. Such a distinction was made to enable the creation of two summaries for each text. The *short* summaries had only sentences marked as essential, while the *complete* summaries included the sentences marked as important. Besides the annotation of the sentences, the annotators could mark fragments of the sentences they judge irrelevant. Such type of labeling is important in evaluating summarization systems that include sentence compression stages.

In a semi-automatic building process, the summary is produced by a computer program, and the humans read and validate it. If mistakes are found, then the humans can and should correct them. In this case, the software may process several documents in a short time because it automatically performs all the tasks required to build a summary. The human effort is, thereby, significantly reduced requiring just reading of the result and, only when necessary, correcting the summary. This approach is cost-effective only if the software platform produces good-quality summaries. The main advantage of this approach is that it is feasible to build a large corpus while maintaining a high quality of the summaries because they are validated and corrected by humans. Although the disadvantages are greatly reduced, it still requires human intervention and, thus, there is still cost, time and subjectivity associated with it. There is also the additional cost of developing a new software platform and training its users. There were several attempts to build a corpus using this approach, which are presented as follows.

In 1995, Kupiec and colleagues [14] built a corpus consisting of 188 scientific and technical documents in English and their extracts. Their strategy was, in the first pass, to automatically match the sentences from the abstract with the ones in the original text and using them as the starting point for the manual assignment in the second pass. They reported an alignment in the first pass of 79% of the sentences. A guideline was used to instruct the experts, but no information about the number of experts per summary was provided.

Teufel and Moens (1997) [31] took a similar approach to the one described [14] but achieved a much lower percentage of alignment (31.7%), using a corpus of 202 articles on computational linguistics.

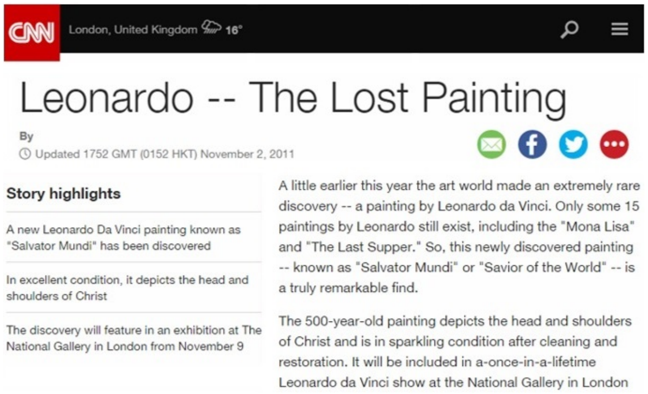
In an automatic building process, the summary is generated automatically by computer software, and there is no human intervention. Therefore, a large volume of documents may be summarized in a short time, greatly reducing the cost, time and subjectivity. It is still an open problem to construct an algorithm that creates a summary with the same quality of a human summary, however. Hence, any corpus produced using this approach has no guarantee of quality and it may not be suitable for most applications, such as research in automatic summarization. Nonetheless, Marcu (1999) [21] proposed creating a corpus automatically. He applied his algorithm on the Ziff-Davis corpus, which consists of newspaper articles announcing computer products, to create a text summarization corpus of 6,942 documents in English. Although it is one of the largest corpora found, there is no way to know which if a summary is incorrect or has poor quality, except if humans check each of them. Therefore, there is no guarantee that the corpus is suitable for scientific research.

The initial ideas behind the CNN corpus were outlined in reference [19]. This paper describes the methodology followed in not only in growing the number of documents, but also in keeping the high quality of the summaries produced.

3 BUILDING THE HP-UFPE CNN-CORPUS

This section describes the basic methodology in the development of CNN-corpus. As the name of the corpus suggests, the starting point is the set of news articles harvested from the CNN website[7].

The CNN news articles are high-quality, grammatically correct texts, report on subjects of general interest, and use standard vocabulary. Besides those fundamental features, there is one particularity of many of such texts that make them especially valuable for automatic text summarization: the *highlights*. The highlights can be seen as abstractive summaries provided by the authors of the original texts. The data collected from the CNN website were the original text, the story highlights and several other metadata such as the name of the authors, title, subject classification, date, and others. Figure 1 shows an example of a news article from the CNN website, in which one can find the title of the article in the center in boldface font, the highlights on the left side, the date of publication, and part of the text content on the right side. As will be further detailed, the story highlights play a fundamental role in the development of the CNN-corpus.



Leonardo -- The Lost Painting

By

Updated 1752 GMT (0152 HKT) November 2, 2011

Story highlights

A new Leonardo Da Vinci painting known as "Salvator Mundi" has been discovered

In excellent condition, it depicts the head and shoulders of Christ

The discovery will feature in an exhibition at The National Gallery in London from November 9

A little earlier this year the art world made an extremely rare discovery -- a painting by Leonardo da Vinci. Only some 15 paintings by Leonardo still exist, including the "Mona Lisa" and "The Last Supper." So, this newly discovered painting -- known as "Salvator Mundi" or "Savior of the World" -- is a truly remarkable find.

The 500-year-old painting depicts the head and shoulders of Christ and is in sparkling condition after cleaning and restoration. It will be included in a-once-in-a-lifetime Leonardo da Vinci show at the National Gallery in London

Figure 1: Example of a news article collected from the CNN's portal available in http://goo.gl/A8cLMP.

3.1 The First Version on the CNN-corpus

The first version of the HP-UFPE CNN-corpus was developed to meet the need to assess automatic extractive text summarization algorithms and was used in several research works [19] [5] [6] [7] [2]. It consisted of 400 news articles in English with highlights, which were manually collected from the CNN website during a period of one week by a team of four experts. A software tool

[7] http://edition.cnn.com

developed by the group assisted them in the task, which at first filtered out any advertisement, graphical elements such as images, videos, and others, from the web page, producing a plain text version of the document. The text sentences were then numbered leaving one line per sentence. The software interface provided two windows for the person who would do the mapping of the each sentence in the story highlights onto one or two sentences extracted from the text *verbatim* for each of the 400 articles. Before such mapping, the tool asks the user if there was any error in the numbered text (e.g., broken sentences, invalid characters). If any error is detected, then the document is discarded. Otherwise, a set of candidate sentences for the summary is generated. The candidate sentences are chosen by a linear combination of the Levenshtein distance [16] from the story highlights to the original sentences of the document and the number of votes of several summarization algorithms. The top-ranked sentences became strong candidates for the best match of the sentence in the story highlight onto the original text. Then, the expert is asked whether those sentences are representative of the content. If yes, then the software places a checkbox to the left of each candidate sentence to the expert to choose the number of the sentences from the original text that best represents the essence of the content of the sentence in the highlights. Otherwise, the software tool opens a text box for the expert to manually list the numbers of the sentences for the extractive summary by reading the original text. After the appropriate sentences were chosen, the software stores them and loads the next article. Eventually, a different person, the *reviewer*, would check the mappings of each of the three experts and may make the final decision, in case of divergence. If there was the need to include more sentences in the extractive summary to better match the content of the sentences in the highlights, the reviewer would manually include them.

As one may observe, this process is computer-assisted since the software does not propose a summary. Instead, it only presents a list of candidate sentences, in which the experts have to refine to appropriately create an extractive summary, the *gold standard*, which allows the quantitative evaluation of automatic summarization algorithms by comparing the matches between the results of the sentences chosen by the summarization methods and the ones in the gold standard. Another observation is that, although four experts mapped each article, sentences may be chosen by only one person. For keeping the quality in the sentence selection process, no expert worked more than 2 hours per day in the process, during the four weeks elapsed to map the 400 articles. No statistics of the building process was recorded. Although the first version of the corpus was successfully used to assess several automatic summarization algorithms, the developing team considered it unsatisfactory in measures of size and diversity of subjects for the purpose of automatic text summarization.

3.2 The Second Version of the CNN-corpus

Following the same ideas adopted in construction of the first version of the CNN-corpus [19], the original text highlights are taken as reference to select the sentences in the original text to be included *verbatim* in the extractive summary that is considered the gold standard. The articles selected in this version are different from those of the first version. A set of rules was adopted during the process to ensure as much as possible that the information in the highlights is also contained in the extractive summaries generated.

The construction of the CNN-corpus, which was performed in four steps: Gathering, Selection, Documents Mapping, and Divergence resolution. In the first step, the collection of documents from the CNN news portal is performed. To make the text selection process more efficient a *web-crawler* was developed, which gathered 18,337 articles from April 2014 to Jun 2014. For each news article, the crawler extracted the full text, the highlights, and the following metadata: title, published date, keywords, author, and category. For some articles, the metadata was not available. The crawler automatically discarded the articles that mentioned videos or figures in the text, as well as Internet links, so that the texts selected were self-contained. Articles with less than three sentences in the story highlights were also discarded because the compression rate of the summary would be too high. Furthermore, for each article, the text was segmented into paragraphs and sentences using the natural languages processing tasks of tokenization and sentence splitting methods implemented in the Stanford Core Natural Language Processing[8]. Each article, story highlight, paragraph and sentence was numbered using a unique identifier. All the documents collected are preprocessed and stored in the Extensible Markup Language (XML) format, to facilitate the recovery of structured information associated with each news article. At this point, 10,000 documents were selected to pass onto the next processing phase.

After the gathering process, aiming to reduce the human effort in the creation of the extractive summaries, all documents collected were processed using the sentence semantic similarity measure proposed by [8]. The similarity measure adopted was applied comparing each sentence in the highlights with each of the original sentences of the document, using a three-layer sentence representation:

1. the lexical layer, which consists of the lexical analysis and the preprocessing steps, such as, removing stop words and stemming;
2. the syntactic layer performs the syntactic analysis and;
3. the semantic layer annotates the entities using the semantic role label technique.

Each layer represents different aspects of both sentences and it is used by the algorithm for estimating their similarity. The final similarity score, which is between 0 and 1, is the mean of the lexical, syntactic and semantic measure values. The sentence similarity algorithm performs an important role in the entire process, because if it succeeds, then less effort is made by the people that have to do the tasks that follow in the generation of the gold standard. To illustrate it, Table 2 presents the top-4 sentences with the highest degree of similarity of the article presented in Table 3. In this case, the topmost sentence s_3 is indeed the most suitable mapping for the story highlight h_1.

During the Selection step, the documents are assigned to one of two classes depending on the degree of probable difficulty in matching the sentences in the highlights onto the original text. This filter discards the documents for which the generation of the gold standard summary would be less straightforward, giving a wider margin to subjectivity. Two thresholds T_1 and T_2 are used. Their

[8] http://nlp.stanford.edu/software/corenlp/

Table 2: Top-4 items of the similarity list for the first sentence in the highlights of the article presented in Table 3.

First sentence in the highlights:
A new Leonardo Da Vinci painting known as "Salvator Mundi" has been discovered.

Original text sentences	Degrees of Similarity			
	Lexical	Syntactic	Semantic	Final (v_{l_j})
3) So, this newly discovered painting – known as "Salvator Mundi" or "Savior of the World" – is a truly remarkable find.	0.555	0.652	0.705	0.638
12) Will another painting like this ever be discovered and is the adjective "priceless", accurate in this instance?	0.324	0.391	0.622	0.446
5) It will be included in a-once-in-a-lifetime Leonardo da Vinci show at the National Gallery in London from November 9th – the first time "Salvator Mundi" will have shared space with other Leonardos for many centuries.	0.560	0.452	0.286	0.433
7) In a half hour special program, Nick Glass travels to New York to meet Robert Simon, an Old Masters dealer, who is acting on behalf of the owners of "Salvator Mundi", and gets to see the painting first hand.	0.513	0.409	0.364	0.429

values are 0.6 and 0.5, respectively, which selected 5,797 (31.6%) documents. Other values of thresholds were either too restrictive or too permissive. For instance, for $T_1 = T_2 = 0.7$ only 6.8% of the texts were selected, while for $T_1 = T_2 = 0.8$ only 1.3% were selected. For values less than $T_1 = T_2 = 0.6$, an experiment was performed with the team and showed that the story highlights were difficult to map onto the sentences of the text, wasting too much time. At the end of this step, 4,800 documents were selected and used in the mapping step.

The third step, the experts check the ranking of the automatic mapping process between each of the sentences in the highlights and the ones in the original text. A computational tool was developed to assist in such a mapping step. To create a high-quality extractive summary, each document was mapped by at least two annotators, and in case of divergence, a third annotator performed the review process.

In the Mapping step, the annotators were assisted by software to map the sentences of the highlights onto the original text. The first interface used presents:

(1) The title of the document;
(2) The highlights of the text in analysis;
(3) The most similar sentence pointed to by the similarity algorithm;
(4) The degree of similarity assigned by the algorithm; and
(5) The navigation options.

If the annotator does not agree with the indicated sentence or he thinks that it is necessary to map more than one sentence for a sentence in the highlight, he can load all the sentences of the document and then perform the mapping process. In addition to these actions, the annotator can also:

(1) Visualize the full text of the document;
(2) Skip the highlight in mapping; or
(3) Mark the document as an Error.

This last option was included for the cases in which the annotator cannot identify all of the information of the highlights in the document sentences, or if he identifies errors in the structure of the document which may have been caused during the collection step. In addition to assisting in the mapping process, the tool also recorded the time required to perform the mapping of each highlight, and which annotator was responsible for the process. This information is important to verify the degree of agreement among the annotators, the quality of the work of each annotator, and an estimate of the effort/time spent by each of them in this process.

A team of six people was organized for the mapping task. All of them were proficient, but non-native, English speakers. Thus, by the distribution policies, 3 pairs of annotators were used, which alternated every week, and each one of the pairs received a batch of the same 100 articles per week. After 16 weeks of work, 4,800 documents were sent to the annotators. At the end of the mapping process, 4,299 documents were mapped and 501 were discarded from the process due to the presence of errors in information extracted. The presence of Hypertext Markup Language (HTML) tags or incomplete information were among the common errors.

Divergent Mapping Resolution. Of the 4,299 documents mapped, 2,702 (62.85%) showed no divergence between the two annotators, while 1,597 (37.15%) had at least one divergence in the mapping process. A document was considered divergent in the following situations:

- there is a divergence in at least one of sentences mapped from the highlights onto the set of the sentences candidate to become the gold standard; and
- only one document was marked as an error, while the other was mapped as normally.

The documents with divergence were sent to a third annotator for a new mapping process by. As in the first mapping step, the third annotator was randomly chosen. Thus, each of the six annotators could perform the divergence resolution, obeying the restriction that he had not done the mapping of this document in the first round. Another graphical interface was used to assist in such a task. In such an interface, the third annotator can check the set of sentences mapped for every highlight by each annotator, or if the document was marked as an error. Thus, the third annotator can clearly see the points of divergence, and he can agree with

Table 3: "Leonardo – The Lost Painting" http://edition.cnn.com/2011/11/02/living/lost-leonardo-synopsis/.

Highlights	Gold standard
A new Leonardo Da Vinci painting known as "Salvator Mundi" has been discovered.	**1)** A little earlier this year the art world made an extremely rare discovery – a painting by Leonardo da Vinci. **3)** So, this newly discovered painting – known as "Salvator Mundi" or "Savior of the World" – is a truly remarkable find.
In excellent condition, it depicts the head and shoulders of Christ.	**4)** The 500-year-old painting depicts the head and shoulders of Christ and is in sparkling condition after cleaning and restoration.
The discovery will feature in an exhibition at The National Gallery in London from November 9.	**5)** It will be included in a-once-in-a-lifetime Leonardo da Vinci show at the National Gallery in London from November 9th – the first time "Salvator Mundi" will have shared space with other Leonardos for many centuries.

Original and full content

1) A little earlier this year the art world made an extremely rare discovery – a painting by Leonardo da Vinci.
2) Only some 15 paintings by Leonardo still exist, including the "Mona Lisa" and The Last Supper."
3) So, this newly discovered painting – known as "Salvator Mundi" or "Savior of the World" – is a truly remarkable find.
4) The 500-year-old painting depicts the head and shoulders of Christ and is in sparkling condition after cleaning and restoration.
5) It will be included in a-once-in-a-lifetime Leonardo da Vinci show at the National Gallery in London from November 9th – the first time "Salvator Mundi" will have shared space with other Leonardos for many centuries.
6) CNN has been given rare access to this newly discovered work.
7) In a half hour special program, Nick Glass travels to New York to meet Robert Simon, an Old Masters dealer, who is acting on behalf of the owners of "Salvator Mundi," and gets to see the painting first hand.
8) During his time in New York, Glass also talks with restorer Dianne Modestini, who brought the painting back to its original state.
9) Intrigued, Glass heads to Florence to learn a little more about the life and work of da Vinci, the ultimate Renaissance man.
10) Further interviews with world renowned da Vinci experts shed more light on the man, his paintings and his contributions to science and medicine.
11) Finally, Glass returns to the National Gallery where the "Salvator Mundi" will be on show to the public until February 2012.
12) Will another painting like this ever be discovered and is the adjective "priceless," accurate in this instance?
13) This program will attempt to answer these questions, while telling the story of the most talked about piece of art unveiled this century.

one of the first two annotators, perform his mapping, or mark the document as an error.

At the end of the Mapping and Divergence Mapping Resolution steps, there were 3,384 documents mapped without divergence and agreed on by three experts. For those mapped documents, there was a final manual inspection to ensure the quality of extractive summaries generated. After this re-checking phase, 384 documents were removed because of problems such as the presence of broken coding symbols; sentences mapped generated much disagreement among all annotators; or because the ratio between the sentences in the extractive summary and the sentences of the full document content was very low. The resulting corpus has 3,000 documents in English with the original text, highlight and extractive summaries and the following metadata: title, published date, keywords, URL, author, and category.

Each document of the corpus is stored in an XML file, which is segmented in several sections. One of the sections is *summaries* that contains the highlight summary in subsection *highlights*, and the extractive summary in subsection *gold_standard*. Another section is *article* that contains the text of the article segmented in paragraphs and sentences with identifiers.

Automatic Coreference Resolution. During the construction process of the CNN corpus it was noticed that some reference summaries presented open coreferences; that is, some sentences had pronouns that were not connected to any entity (e.g. noun) of the summary. This fact occurred because the summary construction process was extractive and based on the highlights, in which the annotators selected the sentences most similar to the highlight to compose the summary. To correct such a problem an automatic method of anaphora resolution (AES) described in [1] was applied. The method listed 1,519 extractive summaries that had at least one free (unbound) pronoun. For each of those summaries the method indicated to the annotators the corresponding entity with which to replace the non-connected pronoun.

The quality of the result of the AES method for correction of the extractive reference summaries obtained was made through the Amazon Mechanical Turk (AMT) platform [9]. The evaluators were responsible for assessing if the automatic treatment of anaphora was correct. The evaluation was done through surveys, each survey is represented by a Human Intelligence Task (HIT) in the AMT. Only native English speakers with minimum secondary education, residing in the U.K., U.S.A., Canada, and Australia were allowed to

[9] https://requester.mturk.com/

enroll in such a task. The following example of a text[10] from the survey was applied:

> **She (Arianna Huffington)** describes herself as a "sleep evangelist," has nap rooms in her offices at the AOL headquarters in New York and tries to start every day with meditation. Huffington, 62, founded Huffington Post in 2005, and two years ago sold it to AOL for $315 million.

The evaluators had to answer the following question:

(1) *Does the mention in parentheses correspond to the respective pronoun? (Yes or No)*

The AES algorithm was able to correctly replace 86% of the total of evaluated summaries, which demonstrates a good indicator that the method can be used for treatments of the open coreferences in the extractive reference summaries. The remaining open coreferences were corrected by the people of the CNN-corpus development team.

3.3 Some Features of the Summaries

Several statistic measures and features of the second version of the CNN-Corpus are presented in Table 4.

The position of the sentences in the extractive summaries generated is also analyzed. The total number of sentences of a document is divided into thirds, representing the regions of the beginning, middle, and end of the document. The 10,754 sentences present in the extractive summaries are distributed as follows: 6,399 sentences (59.50%) are located at the beginning of documents, 2,624 sentences in the middle, and 1,731 sentences at the end of the document. This corroborates several studies in the literature [5] [24] that demonstrated that sentences at the beginning of the documents are more likely to be included in the summaries.

4 SOME DEVELOPMENT STATISTICS

This section evaluates some of the steps performed during the construction of the CNN-corpus, allowing one to better understand the challenges faced on its development. Section 4.1 shows the experimental results of the assessment of the sentence similarity measure (SSM) proposed by Ferreira *et al.* (2014), used to select documents and point out the most similar sentences for each highlight. Section 4.2 and Section 4.3 presents the statistics of the time taken and the agreement among the annotators responsible for the mapping process, respectively.

4.1 Evaluating the SSM

This first experiment assessed the effectiveness of the SSM algorithm adopted [4]. As each highlight was mapped for one or more sentences of the document, and this process was reviewed by humans, it is possible to assess the accuracy of the similarity algorithm in estimating the most similar sentences correctly for each highlight. The experimental results are presented in Table 5. The first column indicates how many sentences from the original text were mapped onto a single sentence of the highlights. It is possible to note that the vast majority of sentences in the highlights (94.72%) were mapped onto one sentence of the original text. The maximum number of mapped sentences for a single sentence from the highlights was four (observed only six times). Accuracy measurements are also provided in Table 5. These are based on the agreement between the human annotators and the algorithm for spotting the key sentences in the summary. As more sentences are mapped onto a single sentence of the highlights, the accuracy of the similarity algorithm decreased. This behavior was due to two factors:

- Given the challenge in the task of identifying when information in a single highlight is fragmented in several judgments of the text; and
- Because the similarity measure proposed by Ferreira *et al.* (2014) was not adapted to deal with such particular cases.

The results of this experiment showed that in 95.89% of the sentences in the highlights, i.e., 10,235 of the 10,674 highlights mapped, the human annotator needed only to read the five more similar sentences indicated by the similarity algorithm to find the set of most suitable phrases in the original text to be selected. These results demonstrate that the inclusion of the selection step impacted positively on the construction process of the CNN-corpus, reducing the manual effort spent for the creation of extractive summaries.

4.2 Time Assessment

Here, one finds some account of the time taken for an expert to validate and fix the mapping of a sentence in story highlights onto the original text in the formation of the gold standard, using the platform developed. The mapping took an average of 141.0 seconds, including the effort of two annotators and one judge. The mean time is 42.6 seconds per sentence in the story highlights, including two annotators and one judge. The mean mapping and divergent times are 18.7 and 18.6 seconds per sentence of the highlights.

The overall and mapping times increased slightly with the number of sentences in the text. This may be so because there are few mappings that are not easy and the expert has to scan other sentences of the text to appropriately perform the task. On the other hand, the divergent time appears to be constant, which sounds reasonable because the annotators have to read only those sentences selected by the annotators to take a decision.

The efficient and user-friendly interfaces developed in this project minimized the human effort and decreased the time elapsed in the tasks. One annotator took on average 13.4 seconds to map a sentence of story highlight, while another one required 29.0 seconds. The mapping and divergent times are 20.7 and 4.8 seconds per story highlight, in the mean. The overall and mapping times oscillate heavily with the number of sentences in the text, not enabling the detection of a pattern. The mean divergent time appears to be almost constant as well.

4.3 Annotators' Agreement Level

The degree of agreement of the mappings between the experts may also be used as an indicator of the quality, because if at least two annotators agreed on a mapping, then it was assumed that there is a higher probability that the summary was adequately mapped.

There were six experts involved in this task. The total number of mapped highlights is twice the number reported in the earlier Section, *Mapping Story Highlights onto the Original Text*, because each story highlight was mapped by two annotators according

[10] http://edition.cnn.com/2013/03/07/business/arianna-huffington-leading-women/

Table 4: Overview statistics of the CNN-corpus.

Categories	Articles	Avg.Sentence/ Summary	Avg.Words/ Story High.	Avg.Sentences/ Text	Avg.Words/ Sentences	Avg. Sentences/ Gold. Summary	Avg.Words/ Gold. Sentence
Business	161	3.3	14.1	30.8	21.6	3.4	25.5
Health	290	3.3	11.7	47.0	18.6	3.4	23.0
Justice	224	3.7	11.9	35.6	20.2	3.5	25.6
Living	98	3.6	12.9	53.3	19.3	3.7	26.9
Opinion	192	3.8	13.5	43.8	20.7	3.9	26.1
Politics	195	3.5	12.2	37.8	21.7	3.5	26.7
Showbiz	241	3.5	11.6	28.8	19.0	3.5	23.2
Sport	148	3.7	11.6	31.3	20.9	3.6	27.0
Technology	132	3.4	12.2	39.1	19.0	3.4	25.4
Travel	171	3.3	12.6	55.4	17.7	3.5	24.7
US	160	3.6	11.9	39.7	18.8	3.6	23.6
World	988	3.7	12.2	35.6	20.6	3.7	25.2
Total/Average	3,000	3.6	12.3	38.4	19.9	3.6	25.0

Table 5: Results of the complexity of the mapping process of the highlights onto the original text in the CNN-corpus.

#Sentences Mapped (S)	#Highlights (%)	#Hits Top-S sentences	Accuracy (%)	Hits #Top-5 sentences	Accuracy (%)
1	10,110 (94.72)	9,214	91.14	9,979	98.70
2	513 (4.81)	191	37.23	248	48.34
3	45 (0.42)	7	15.56	7	15.56
4	6 (0.05)	1	16.67	1	16.67

to the quality policy adopted. The overall annotator agreement rate is 89.2%, which means that 10.8% were discarded. The level of agreement between the six annotators is similar, ranging from 85.3% to 91.6%. After the judge decided on the divergent cases, the agreement level was also similar, ranging from 90.7% to 95.5%. For instance, the expert B annotated 3,772 story highlights, of which 89.9% of them were agreed by another expert (e.g. A, C, D, E or F) and 95.5% of them were agreed by another expert in the mapping task or by a judge in the divergent task. Specifically for the divergent cases, which correspond to 10.7% of the total number of mapped highlights (i.e. 10,674), the referees agreed with one of the annotators in 62.6% of the cases. For the rest of the divergent highlights (37.4%), their documents were discarded from the final corpus. This means that the resulting corpus is formed by articles that are many levels easier than most articles in the original CNN news articles, besides being self-contained (there are no external web links) and text-only (there are no internal references to figures, tables, videos, etc.).

5 CONCLUSIONS AND FURTHER WORK

This paper presents the CNN-corpus, possibly the largest corpus for assessing algorithms for the extractive single-document automatic summarization, the result of a team of more than eight people, over eight years that consumed several man-years of work. The corpus has as a starting point some of the CNN news articles in English, which address subjects of general interest and follow very high vocabulary and grammatical standards. Besides that, each text has an abstractive summary associated with the *highlights*, which served as the key point to select the sentences in the original text. Such sentences form the *gold standard*, an extractive summary for each of the chosen texts. The gold standards may be used as a reference for making quantitative assessments in extractive summarization algorithms.

The semi-automatic methodology fully described here used in the development of the CNN-corpus may be considered as an indicator of its quality. It may also be followed to further enlarge it, either in the number of texts in English, or in including other languages. The same methodology may also be followed in the development of other test corpora for other areas of knowledge. The corpus building process attempts to minimize the human intervention by automatically choosing documents for which the mapping of the sentences in the highlights onto the text may be done in a simple way. The efficient and user-friendly interfaces developed minimized the human effort, decreased the time elapsed in the mapping, and lowered the chances of human errors.

Strict quality policies were enforced: every extractive summary must be agreed on by, at least, two independent experts. Such a rule decreases the probability of a summary being affected by human subjectivity and mistakes. The current version of the CNN-corpus encompasses 3,000 documents in English.

The authors of this paper and some other associates also developed the CNN-corpus in Spanish [18] following the methodology described here. Such a task was even more challenging because most good quality text analytics tools were developed for the English language and are not available for other languages. An intermediate Spanish into English translation step was introduced, to translate each sentence of the original text keeping the original ordering, making possible the use of such tools to better select the candidate sentences to the extractive summary, prior to human checking.

As presented here, the original tagging of CNN-articles encompass only twelve categories that overlap in many aspects, and are too general sometimes. The category "world" is responsible for almost one-third of the articles in the current version of the CNN-Corpus. Another research line that are being followed by the authors of this paper is in the development of a better and more detailed ontology for news articles. Such a new ontology encompasses ten categories on the top level, and each of those are further refined in two sub-level deep categories. All the CNN-documents are thus being semi-automatically retagged based on the three-level deep ontology developed. Such a huge research effort is fundamental for analyzing automatic document classification techniques and the effect of summarization on document classification [29]. Besides the automatic analysis of document subject, research is also being developed to analyze the time-span the document describes, its geographic placing, etc.

The CNN-corpus is currently being used in a large number of research initiatives ranging from the analysis and resolution of dangling coreference, improving extractive summarization techniques, automatically generating abstractive summaries from extractive ones. It was also recently used in the DocEng'19 Competition on Extractive Text Summarization [17].

The CNN-corpus, with the original texts, their highlights, gold-standard summaries, and all its annotated versions will be made freely available for research purposes, under request to the authors.

6 ACKNOWLEDGMENTS

The authors are partly supported by a R&D project between HP Brazil R&D and UFPE originated from tax exemption (IPI - Law n^{o} 8.248, of 1991 and later updates). The authors are also grateful to PROPESQ-IFPE and CNPq by supporting this research.

REFERENCES

[1] Jamilson Batista, Rafael Dueire Lins, Rinaldo Lima, Steven J. Simske, and Marcelo Riss. 2016. Towards Cohesive Extractive Summarization Through Anaphoric Expression Resolution. In *Proceedings of the 2016 ACM Symposium on Document Engineering (DocEng '16)*. ACM, New York, NY, USA, 201–204. https://doi.org/10.1145/2960811.2967159

[2] Luciano de Souza Cabral, Rafael Dueire Lins, Rafael Fe Mello, Fred Freitas, Bruno Ávila, Steven Simske, and Marcelo Riss. 2014. A platform for language independent summarization. In *Proceedings of the 2014 ACM symposium on Document engineering - DocEng '14*. ACM Press, New York, New York, USA, 203–206. https://doi.org/10.1145/2644866.2644890

[3] H. P. Edmundson. 1969. New methods in automatic extracting. *Journal of the ACM (JACM)* 16, 2 (1969), 264–285.

[4] Rafael Ferreira, Luciano de Souza Cabral, Frederico Freitas, Rafael Dueire Lins, Gabriel de França Silva, Steven J. Simske, and Luciano Favaro. 2014. A multi-document summarization system based on statistics and linguistic treatment. *Expert Systems with Applications* 41 (2014), 5780–5787.

[5] Rafael Ferreira, Luciano de S. Cabral, Rafael D. Lins, Gabriel de F. P. e Silva, Frederico L. G. Freitas, George D. da C. Cavalcanti, Rinaldo J. de Lima, Steven J. Simske, and Luciano Favaro. 2013. Assessing sentence scoring techniques for extractive text summarization. *Expert Systems with Applications* 40 (2013), 5755–5764.

[6] Rafael Ferreira, Fred Freitas, Luciano De Souza Cabral, Rafael Dueire Lins, Rinaldo Lima, Gabriel Franca, Steven J. Simske, and Luciano Favaro. 2013. A Four Dimension Graph Model for Automatic Text Summarization. In *Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2013 IEEE/WIC/ACM International Joint Conferences on*, Vol. 1. 389–396. https://doi.org/10.1109/WI-IAT.2013.55

[7] Rafael Ferreira, Fred Freitas, Luciano De Souza Cabral, Rafael Dueire Lins, Rinaldo Lima, Gabriel Franca, Steven J. Simske, and Luciano Favaro. 2014. A Context Based Text Summarization System. In *Document Analysis Systems (DAS), 2014 11th IAPR International Workshop on*. 66–70. https://doi.org/10.1109/DAS.2014.19

[8] Rafael Ferreira, Rafael Dueire Lins, Fred Freitas, Bruno Avila, Steven J Simske, and Marcelo Riss. 2014. A New Sentence Similarity Method based on a Three-Layer Sentence Representation. In *2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*. IEEE Computer Society, 110–117. https://doi.org/10.1109/WI-IAT.2014.23

[9] Mahak Gambhir and Vishal Gupta. 2016. Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review* (2016), 1–66.

[10] Pierre-Etienne Genest. 2013. *Génération de résumés par abstraction*. Thesis, in French. Université de Montréal Génération.

[11] Vishal Gupta and Gurpreet Singh Lehal. 2010. A Survey of Text Summarization Extractive Techniques. *Journal of Emerging Technologies in Web Intelligence* 2, 3 (Aug. 2010), 258–268.

[12] Laura Hasler, Constantin Orasan, and Ruslan Mitkov. 2003. Building better corpora for summarisation. In *Proceedings of Corpus Linguistics 2003*. 309–319.

[13] Hongyan Jing and Kathleen R. McKeown. 1999. The decomposition of human-written summary sentences. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '99*. ACM Press, New York, New York, USA, 129–136.

[14] Julian Kupiec, Jan Pedersen, and Francine Chen. 1995. A trainable document summarizer. *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '95* (1995), 68–73.

[15] F. W. Lancaster. 2003. *Indexing and Abstracting in Theory and Practice* (3 ed.). Library Association, London.

[16] VI Levenshtein. 1966. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady* 10 (1966), 707.

[17] Rafael Dueire Lins, Rafael Ferreira, and Steven J. Simske. 2019. DocEng'19 Competition on Extractive Text Summarization. In *Proceedings of the 2019 ACM Symposium on Document Engineering (DocEng '19)*. ACM, New York, NY, USA, 216–217. https://doi.org/10.1145/3342558.3351874

[18] Rafael Dueire Lins, Hilario Oliveira, Bruno Tenorio, Jamilson Batista, Rafael Ferreira, Gabriel Pereira e Silva, Rinaldo Lima, Diego Salcedo, and Steven J. Simske. 2019. The CNN-Corpus in Spanish: a Large Corpus for Extractive Text Summarization in the Spanish Language. In *Proceedings of the 2019 ACM Symposium on Document Engineering (DocEng '19)*. ACM, New York, NY, USA, 211–215. https://doi.org/10.1145/3342558.3345423

[19] Rafael Dueire Lins, Steven J Simske, Luciano de Souza Cabral, Gabriel de França Pereira e Silva, Rinaldo Lima, Rafael Ferreira de Mello, and Luciano Favaro. 2012. A multi-tool scheme for summarizing textual documents. In *11st IADIS international conference WWW and INTERNET 2012*. Madrid, Spain, 1–8.

[20] H. P. Luhn. 1958. The automatic creation of literature abstracts. *IBM Journal of Research and Development* 2, 2 (1958), 159–165.

[21] Daniel Marcu. 1999. The automatic construction of large-scale corpora for summarization research. In *22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '99*. ACM Press, New York, New York, USA, 137–144. https://doi.org/10.1145/312624.312668

[22] Ani Nenkova and Kathleen McKeown. 2012. A Survey of Text Summarization Techniques. In *Mining Text Data*, Charu C. Aggarwal and ChengXiang Zhai (Eds.). Springer US, Boston, MA, 43–76. https://doi.org/10.1007/978-1-4614-3223-4_3

[23] C Orasan. 2002. Building annotated resources for automatic text summarisation. In *Third International Conference on Language Resources and Evaluation (LREC)*. Las Palmas de Gran Canaria, 1780–1786.

[24] You Ouyang, Wenjie Li, Qin Lu, and Renxian Zhang. 2010. A Study on Position Information in Document Summarization. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters (COLING '10)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 919–927. http://dl.acm.org/citation.cfm?id=1944566.1944672

[25] Paul Over, Hoa Dang, and Donna Harman. 2007. DUC in Context. *Information Process and Management* 43, 6 (Nov. 2007), 1506–1520. https://doi.org/10.1016/j.ipm.2007.01.019

[26] P. Over and W. Liggett. 2002. Introduction to DUC: An Intrinsic Evaluation of Generic News Text Summarization Systems. http://www-nlpir.nist.gov/projects/duc/pubs/2002slides/overview.02.pdf, last visited on 18th March 2019. (2002).

[27] DR Radev, Simone Teufel, and Horacio Saggion. 2003. Evaluation challenges in large-scale document summarization. In *41st Annual Meeting on Association for Computational Linguistics*, Vol. 1. Association for Computational Linguistics.

[28] J. E. Rush, R. Salvador, and A. Zamora. 1971. Automatic abstracting and indexing. II. Production of indicative abstracts by application of contextual inference and syntactic coherence criteria. *Journal of the American Society for Information Science* 22, 4 (July 1971), 260–274.

[29] Steven J. Simske and Rafael Dueire Lins. 2018. Automatic Text Summarization and Classification. In *Proceedings of the ACM Symposium on Document Engineering 2018, DocEng 2018, Halifax, NS, Canada, August 28-31, 2018*. 1:1–1:2. https://doi.org/10.1145/3209280.3232791

[30] Krysta M Svore, L Vanderwende, and CJC Burges. 2007. Enhancing Single-Document Summarization by Combining RankNet and Third-Party Sources.. In *EMNLP-CoNLL*. Prague, Czech Republic, 448–457.

[31] Simone Teufel and Marc Moens. 1997. Sentence Extraction as a Classification Task. In *Proceedings of the ACL'97/EACL'97 Workshop pn Intelligent Scallable Text Summarization*. Madrid, 58–59.