A Hybrid Approach to Generation of Missing Abstracts in Biomedical Literature

Suchet K Chachra
suchet.chachra@gmail.com

Asma Ben Abacha
asma.benabacha@nih.gov

Sonya Shooshan
sonya@nlm.nih.gov

Laritza Rodriguez
laritza.rodriguez@nih.gov

Dina Demner-Fushman
ddemner@mail.nih.gov

U.S. National Library of Medicine, Bethesda, MD, USA

Abstract

Readers usually rely on abstracts to identify relevant medical information from scientific articles. Abstracts are also essential to advanced information retrieval methods. More than 50 thousand scientific publications in PubMed Central lack author-generated abstracts, and the relevancy judgements for these papers have to be based on their titles alone. In this paper, we propose a hybrid summarization technique that aims to select the most pertinent sentences from articles to generate an extractive summary in lieu of a missing abstract. We combine i) health outcome detection, ii) keyphrase extraction, and iii) textual entailment recognition between sentences. We evaluate our hybrid approach and analyze the improvements of multi-factor summarization over techniques that rely on a single method, using a collection of 295 manually generated reference summaries. The obtained results show that the hybrid approach outperforms the baseline techniques with an improvement of 13% in recall and 4% in F1 score.

1 Introduction

PubMed Central[1] (PMC) is a repository of biomedical and life sciences journals supported by the U.S. National Library of Medicine (NLM). PMC provides access to the abstracts as well as the full-text content of biomedical articles. The open-access subset of PubMed Central contains over one million biomedical articles as of Fall 2015, and is widely used as a public resource to discover, read and build upon its vast portfolio of biomedical knowledge. Given the abundance and variety of information available within PMC, many user queries return a multitude of results, which makes it more difficult to identify relevant data. The amount of potentially relevant results often increases further due to information retrieval techniques such as query expansion using synonyms.

Article abstracts are usually considered to be entry points into the full-text. Abstracts often contain key health-outcome data and clinical findings that help identifying relevant data when narrowing down the number of returned results to a select few. The abstracts, however, are missing in 50,000 articles available within the open access subset of PMC. Therefore, for this large set of articles, the only way for the users to judge the relevancy of an article is either through the article title, which is not always reliable, or through the full-text of the paper, which can be time-consuming.

In this paper, we describe a novel hybrid approach that builds upon textual entailment, keyphrase extraction and health outcome detection to generate surrogate abstracts for biomedical articles where none are available. Using a set of 295 documents and manually generated extractive summaries that we make publicly available with this publication, we also analyze how this approach compares to baseline methods relying on a single technique.

2 Related Work

Single and multi-document text summarization of biomedical articles received much attention over the years. Lloret *et al.* developed COMPENDIUM, a text summarization system for generating abstractive and extractive summaries for individual biomedical papers (Lloret et al., 2013). They observed that

[1] http://www.ncbi.nlm.nih.gov/pmc

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 1093–1100, Osaka, Japan, December 11-17 2016.

extractive methods are as effective as abstractive summarization or text generation. Kim *et al.* proposed a sub-topic or theme detection method for multi-document clustering and topical summarization of citation data (Kim et al., 2015).

Previous work shows that Recognizing Textual Entailment (RTE) can provide effective information for text summarization. RTE is the task of recognizing an inference relation between two sentences expressing the fact that the meaning of one sentence is entailed by the other (Androutsopoulos and Malakasiotis, 2010; Dagan et al., 2013). In particular, Entailment-based minimum vertex cover method (Gupta et al., 2014) is an RTE method for single document summarization using graph-based algorithms. Textual entailment and logic segmentation based methods also improved performance for single document summarization (Tatar et al., 2008).

Keyword identification methods were also used in single and multiple document summarization and document clustering (Frigui and Nasraoui, 2004; Hammouda et al., 2005), as well as summary generation based on the salience of sentences (Erkan and Radev, 2004). A more detailed survey of summarization methods is presented in (Nenkova and McKeown, 2012).

In this paper, we propose a novel hybrid approach that combines both textual entailment and keyword extraction for the construction of relevant extractive summaries. We particularly show that such combination yields more comprehensive and informative summaries for a variety of documents. Another contribution of our work is a manually created collection of summaries for 295 documents that have no author-generated abstracts. The summaries created by two experts are released with this paper.

3 Methods

In this section, we first describe three baseline methods for abstract generation, where the abstract is generated by combining the top five scoring sentences according to each method, in the order in which they appear in the original text. Second, we describe our method for the recognition of entailment relationships between sentences. Thereafter, we present our hybrid approach that aggregates the best-performing baseline methods and exploits textual entailment relationships in the article full-text to enrich the set of selected sentences.

3.1 Summary Generation based on Health Outcome Identifier (HO)

We used an existing Health Outcome identifier, previously shown to perform well on extracting health outcomes, also called bottom-line, from PubMed abstracts (Demner-Fushman et al., 2006). The health outcome detector employs an ensemble of rule-based, Naïve Bayes, n-gram based, position based, document-length based and semantic classifiers to compute the likelihood scores for each sentence in the article to contain a health outcome. The rule-based classifier analyzes each sentence for existence of cue phrases such as "significantly greater" and "dropout rate". The Naïve Bayes classifier generates a likelihood score based on a bag of words representation of the sentence. The n-gram based classifier looks for uni- and bi-grams that provide a high information gain measure and are strong positive predictors of outcomes such as "superior" and "especially useful". Positional and document length classifiers factor the position of a given sentence in the supplied text and the length of each sentence to provide probability estimates for containing health outcomes. The semantic classifier uses the results of a biomedical concept-extractor that detects presence of biomedical concepts belonging to outcome-related semantic types, such as diseases, symptoms, and medications, within the sentence and concept discovery information from previous sentences to generate a likelihood score. Finally, the probability scores from each classifier are combined to compute the final score $S(x)$ for each sentence x:

$$S(x) = \sum_{k=1}^{n} \alpha_k P_k(x)$$

Where P_1(x), ..., P_n(x) are the probability scores from various classifiers and (α_1, ..., α_n) are the coefficients or weights used to add the likelihood scores.

3.2 Summary Generation based on Keyphrase Extraction

A *keyword* is "a single word that is highly relevant" and a *keyphrase* is "a sequence of two or more words that is considered highly relevant". Our two remaining baseline methods for summary generation are inspired by (Luhn, 1958; Edmundson, 1969) and identify salient sentences to be selected based on detection of keywords or multi-word keyphrases. More precisely, the task is to identify "key sentences" within a given text, defined as the sentences that contain more keyphrases or keywords compared to others. Each method uses a different algorithm for extracting keyphrases from a given text. The extracted keyphrases are then normalized before being used to generate a score for each sentence, using the frequency of contained keywords. The two methods are described below.

3.2.1 Keyphrase Extraction with KEA

The Keyphrase Extraction Algorithm (KEA), developed by (Witten et al., 1999), uses a Naïve Bayes classifier to identify key phrases within text and a discretization scheme developed by (Fayyad and Irani, 1993) based on Minimum Descriptor Length Principle. The algorithm first splits the input text into phrase boundaries, based on punctuation and word boundaries, to look for sequences of words of length up to three to be used as candidate phrases for further examination. Candidate phrases that end in a stopword or occur only once in the text are dropped. Next, the Naïve Bayes model is used on each candidate phrase with feature values t (for $TF \times IDF$) and d (for $distance$) to compute probabilities $P[yes]$ and $P[no]$ that candidate phrase is a keyphrase in the document. The overall probability that the candidate phrase is a keyphrase is then calculated as:

$$p = \frac{P[yes]}{(P[yes] + P[no])}$$

Finally, candidate phrases are ranked according to the above value and the top n keywords are returned, where n is the number of requested keywords. In our experiments, KEA was restricted to output no more than 15 keyphrases per document.

3.2.2 Keyphrase Extraction with Microsoft Text Analytics (MSTA)

The Microsoft Azure Machine Learning suite provides access to Text Analytics web services, which is based on Microsoft Office's sophisticated Natural Language Processing toolkit. MS Text Analytics was used in our experiments to extract keyphrases from the full-text article. For our task, parts of article text were broken into chunks of successive sentences up to 1000 characters long to support the web-service requirement of maximum text length per individual request.

3.2.3 Keyphrase Normalization and Sentence Ranking

Before sentence ranking, a normalization step is performed to remove selected keywords that also occur as complete words within other keyphrases. Acronyms with all uppercase characters are always selected and not filtered during the normalization step. An example of the keyword normalization step is shown below:

$$microCT\ scans,\ microCT \longrightarrow microCT\ scans$$

$$Italian\ Purine\ Clubs,\ Italian\ Purine,\ Purine\ Clubs \longrightarrow Italian\ Purine\ Clubs$$

Once the keyphrase normalization is complete, article sentences are ranked in the order of keyphrase frequency, counting multiple occurrences of the same keyphrase as one. The keyphrase frequency, i.e. the number of keyphrases contained in each sentence is later used to identify target areas of the article text that are relatively more informative compared to others.

3.3 Hybrid Approach Using Baseline Methods and Textual Entailment

Our hybrid approach uses the output of both Health Outcome identification method and KEA keyphrase extraction method to identify a set of candidate sentences, C, as an initial summary. The next factor in the hybrid approach is based on inference relations obtained by recognizing textual entailment between article sentences.

3.3.1 Recognizing Textual Entailment

Textual entailment between two sentences of the same article is recognized using a feature-based classifier. We use a set of similarity measures as learning features. We will refer to it as SimSet in the remainder of the paper.

For a sentence pair (S_1, S_2), three features are computed after stopword removal and word stemming. The first feature is the word overlap between S_1 and S_2. The second feature is the Dice coefficient based on the number of common bigrams. The third feature is the maximum similarity value between five similarity measures: Levenshtein distance, Dice coefficient, Jaccard similarity, Cosine and Word Overlap.

We trained our RTE classifier on the SNLI corpus (Bowman et al., 2015) which contains 570K sentence pairs annotated with three labels: entailment, contradiction and neutral. The authors showed that the size of this corpus allows lexicalized classifiers to outperform some existing sophisticated entailment models. Also, the tested RNN models (a plain RNN and an LSTM RNN) and the feature-rich/lexicalized model show similar performance when trained on the full corpus.

In the scope of our study, we converted the contradiction and neutral labels to the same non-entailment class. Table 1 presents the results of our classifier using the SVM and Logistic Regression algorithms. We apply our RTE method to all possible sentence pairs in each article of our collection.

Classifier	**Accuracy**
SimSet (SVM)	75.86
SimSet (Logistic Regression)	75.64
Lexicalized classifier (Bowman et al., 2015)	75.00

Table 1: 2-class test accuracy on the SNLI corpus for recognizing textual entailment.

3.3.2 Improving Summaries Using Textual Entailment Graph Traversal

The extracted entailment relations are used to generate one or more directed graphs. The vertices V in these directed graphs are the article sentences for which entailment is detected, and the edges E represent directional entailment relations.

The next step involves iterating over each candidate sentence, $C_i \in C \cap V$, involved in at least one entailment relation and selected by the baseline systems. The sub-graph starting at node C_i is then traversed to check if there exists a sentence $V_j \neq Ci$ directly entailed by C_i, or indirectly entailed through a descendant of C_i such that $f(V_j) > f(C_i)$, for a given function f. If such a sentence is found, and has not been previously selected during similar optimization, then V_j is recorded to replace C_i in the final summary.

After the above iteration has been performed for each sentence in $C \cap V$, any unexplored graphs, formed by vertices $V_{rem} \subseteq V - C$ and disjoint from the candidate sentences obtained in earlier steps, are explored beginning from the source node to select additional sentences (one sentence for each disjoint entailment graph). This helps in the enrichment of the final summary by selecting vital hypothesis chains missed by the baseline systems. This allows addressing scenarios where the entailment relations discovered in the article involve other sentences that were not previously selected by the baseline systems, i.e. $C \cap V = \varnothing$. For our various experiments, function f was designed to prefer: i) shorter sentences (Hybrid MinLength), ii) longer, more informative sentences (Hybrid MaxLength) and iii) sentences with higher scores from baseline systems (Hybrid MaxScore).

4 Experiments & Discussion

4.1 Evaluation Dataset

Our experiments were conducted on 295 articles (the evaluation dataset) taken from the open-access subset of PMC. We picked a predetermined number of articles at random from 16 different article types, a classification provided by PMC for each article (e.g., research article, patient's case description or review articles) to ensure a diverse evaluation collection. Table 2 shows the breakup of our evaluation

dataset by article type. We note here that the articles selected for evaluation did not contain author-generated abstracts, so we had to manually generate a reference set of extractive summaries ("Golden Summaries") to be used in the evaluation of the baseline and hybrid system generated summaries.

Article Type	Count
Extended Abstracts	69
Research Articles	48
Review Articles	48
Case Reports	30
Editorials	15
Book Reviews	10
Brief Reports	10
Discussions	10
Letters	10
Meeting Reports	10
News	10
Introduction	5
Obituary	5
Oration	5
Product Reviews	5
Replies	5

Table 2: Article Type counts in the Evaluation Dataset.

4.2 Manual Extraction of Reference Summaries

Two experts, a clinician trained in medical informatics and a medical librarian, were asked to extract "golden summaries" from the articles in the evaluation dataset. Articles of various types were uniformly distributed between the human evaluators. The task was to identify and select key article sentences from the article text. The preferable length in number of sentences for reference summaries was set at 10 sentences, but the system allowed human evaluators to override this limit and adjust it to the minimal length needed to capture all key points of an article. It is important to specify here that manually extracting and compiling reference summaries is highly laborious and required the experts to read the supplied articles, hence being more time-consuming than using author supplied abstracts for evaluations. Figure 1 shows our sentence selection interface for reference summary extraction.

The manually extracted reference summaries are available for download at `https://archive.nlm.nih.gov/ridem/infobot_docs/reference-summaries{.zip,.tar.gz}`

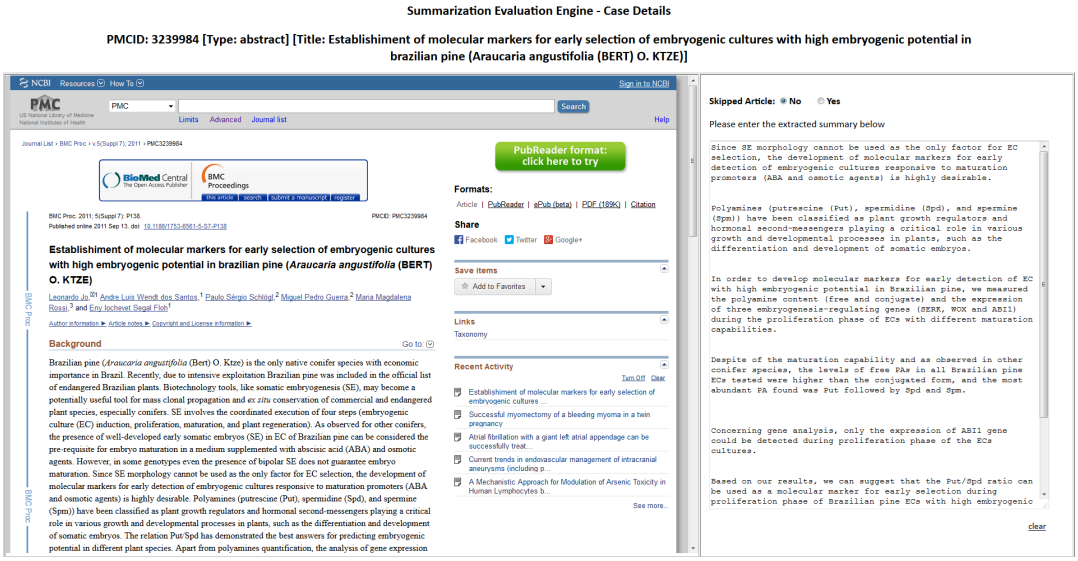
4.3 Judging Baseline System Summaries for Content and Coverage

An additional evaluation task for the human evaluators was to judge the summaries generated by the three baseline systems for content coverage and potential usefulness by rating the baseline summaries on a scale of 1-5 (1=Not at all, 5=Perfect) for the below criteria:

- Is the summary informative?
- Does the summary reflect the most important issues?
- Does the summary capture the bottom-line?

For this task, the generated summaries from three baseline systems were presented to the human evaluators unlabeled and in random order. Table 3 shows for each baseline system the number of articles where the evaluators judged the baseline system summary as acceptable or better (Score, $S_{criterion}$, greater than or equal to 3) for each of the above mentioned criteria. Figure 2 shows our interface for recording such judgements.

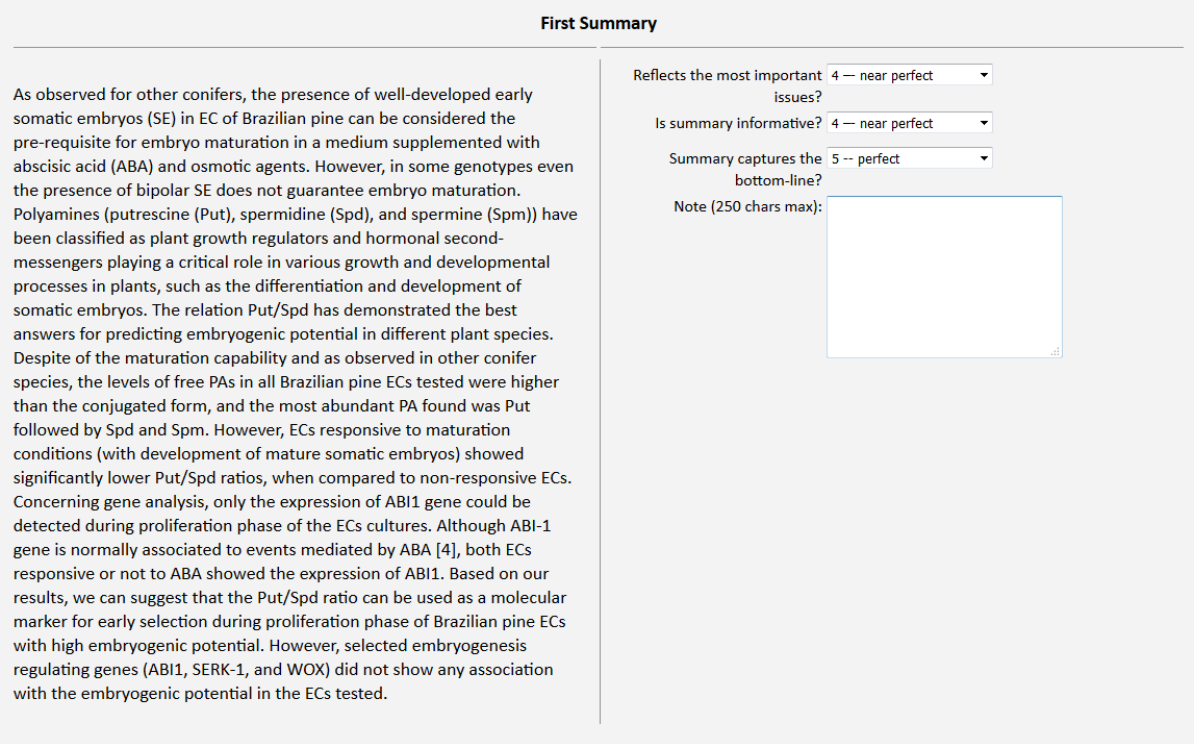
Figure 1: Interface for reference summary sentence selection.



System	Informative ($S_{info} \geq 3$)	Overall Rating ($S_{imp.issues} \geq 3$)	Bottom Line ($S_{bottom} \geq 3$)
HO	242	208	228
KEA	258	228	181
MSTA	244	212	155

Table 3: Manual evaluation of the automatically generated baseline summaries.

Figure 2: Interface for the evaluation of unlabeled baseline summaries.



Based on the preliminary results shown above and the Rouge-2 scores provided in the next section, we decided to base our hybrid approach on the HO and KEA baseline systems and to further improve the combined summaries using Textual Entailment relations recognized in the article text.

4.4 Rouge-2 Evaluation of Generated Baseline and Hybrid Summaries

In addition to manual evaluation of the baseline summaries, we compared them with the hybrid summaries as whole paragraphs to human extracted "golden summaries" using the recall based evaluation metric ROUGE-2 (with stopword removal) for automatic overlap measurement. Table 4 and Table 6 present the Rouge-2 results for the baseline systems and hybrid systems respectively.

System	R (%)	P (%)	F (%)
HO	27.97	27.96	27.13
KEA	28.42	29.44	28.03
MSTA	24.90	20.71	21.99

Table 4: ROUGE-2 evaluation results for summaries generated by baseline systems.

We also tested for, and observed a low overlap between baseline summaries generated by HO and KEA-based systems, which indicates that a hybrid approach could be more comprehensive and likely to outperform individual systems in terms of content recall and coverage. We also observed that in 46 of 130 cases in which an entailment relation was present, our feature-based RTE classifier was able to correctly identify at least one relation involving a sentence that was also selected as prominent by human evaluators. Table 5 presents the summary overlap between baseline systems and inter-annotator agreement.

System	R (%)	P (%)	F (%)
HO Vs. KEA	21.85	19.69	20.21
Inter-Annotator	46.33	42.99	43.47

Table 5: ROUGE-2 results for HO and KEA summary overlap and inter-annotator agreement.

System	R (%)	P (%)	F (%)
Hybrid MinLength	38.82	27.88	31.73
Hybrid MaxLength	41.76	27.41	32.18
Hybrid MaxScore	39.87	27.71	32.88

Table 6: ROUGE-2 evaluation results for summaries generated hybrid systems.

4.5 Discussion

Using a hybrid approach to abstract generation significantly improved the recall while still providing similar precision values, despite the fact that hybrid summaries are generally longer compared to baseline systems. More generally, our experiments show that combining multiple single-factor techniques like keyword extraction, health outcome detection and utilizing semantic relations in text using textual entailment works well for different kinds of articles, and is more likely to outperform traditional baseline approaches for text summarization.

5 Conclusion

We presented a new hybrid approach combining textual entailment and keyword extraction for the summarization of biomedical articles. Our results show that such combination yields substantial improvement in recall while maintaining the precision at the same level. In future work, we plan to incorporate named entity recognition and use the extracted named entities as additional keywords to improve precision and to apply a similar approach to multi-document summarization.

References

Ion Androutsopoulos and Prodromos Malakasiotis. 2010. A survey of paraphrasing and textual entailment methods. *J. Artif. Int. Res.*, 38(1):135–187, May.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Ido Dagan, Dan Roth, Mark Sammons, and Fabio Massimo Zanzotto. 2013. *Recognizing Textual Entailment: Models and Applications*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Günes Erkan and Dragomir R. Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Intell. Res. (JAIR)*, 22:457–479.

Usama M. Fayyad and Keki B. Irani. 1993. Multi-interval discretization of continuous-valued attributes for classification learning. In *Proceedings of the 13th International Joint Conference on Artificial Intelligence. Chambéry, France, August 28 - September 3, 1993*, pages 1022–1029.

Hichem Frigui and Olfa Nasraoui. 2004. Unsupervised learning of prototypes and attribute weights. *Pattern Recognition*, 37(3):567–581.

Anand Gupta, Manpreet Kaur, Shachar Mirkin, Adarsh Singh, and Aseem Goyal. 2014. Text summarization through entailment-based minimum vertex cover. In *Proceedings of the Third Joint Conference on Lexical and Computational Semantics (*SEM 2014)*, pages 75–80, Dublin, Ireland, August. Association for Computational Linguistics and Dublin City University.

Khaled M. Hammouda, Diego N. Matute, and Mohamed S. Kamel. 2005. Corephrase: Keyphrase extraction for document clustering. In *Machine Learning and Data Mining in Pattern Recognition, 4th International Conference, MLDM 2005, Leipzig, Germany, July 9-11, 2005, Proceedings*, pages 265–274.

Sun Kim, Lana Yeganova, and W. John Wilbur. 2015. Summarizing topical contents from pubmed documents using a thematic analysis. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 805–810.

Elena Lloret, María Teresa Romá-Ferri, and Manuel Palomar. 2013. COMPENDIUM: A text summarization system for generating abstracts of research papers. *Data Knowl. Eng.*, 88:164–175.

Hans Peter Luhn. 1958. The automatic creation of literature abstracts. In *IBM Journal of research and development 2(2)*, pages159-165.

Harold P. Edmundson. 1969. New Methods in Automatic Extracting. In *Journal of the ACM 16 (2)*, pages 264-285.

Ani Nenkova and Kathleen McKeown. 2012. A survey of text summarization techniques. In Charu C. Aggarwal and ChengXiang Zhai, editors, *Mining Text Data*, pages 43–76. Springer.

Dina Demner-Fushman, Barbara Few, Susan E. Hauser, and George Thoma. 2006. Automatically identifying health outcome information in MEDLINE records.. In *Journal of the American Medical Informatics Association 13(1)*, pages 52-60.

Doina Tatar, Andreea Diana Mihis, and Dana Lupsa. 2008. Text entailment for logical segmentation and summarization. In *Natural Language and Information Systems, 13th International Conference on Applications of Natural Language to Information Systems, NLDB 2008, London, UK, June 24-27, 2008, Proceedings*, pages 233–244.

Ian H. Witten, Gordon W. Paynter, Eibe Frank, Carl Gutwin, and Craig G. Nevill-Manning. 1999. KEA: practical automatic keyphrase extraction. In *Proceedings of the Fourth ACM conference on Digital Libraries, August 11-14, 1999, Berkeley, CA, USA*, pages 254–255.