According to the IEEE Article Sharing and Posting Policies, the uploaded full-text on our server is the accepted paper. The final version of the publication is available at

https://doi.org/10.1109/EIT.2018.8500186.

© 2018 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

Towards Extractive Text Summarization using Multidimensional Knowledge Representation

Johannes Zenkert, André Klahold and Madjid Fathi

University of Siegen
Department of Electrical Engineering and Computer Science
Institute of Knowledge Based Systems and Knowledge Management
Germany
johannes.zenkert@uni-siegen.de, andre.klahold@uni-siegen.de, fathi@informatik.uni-siegen.de

***Abstract*—Multidimensional knowledge representation (MKR) is the result of integrative text mining. Analysis results from individual text mining methods such as named entity recognition, sentiment analysis, or topic detection are represented as dimensions in a knowledge base to support knowledge discovery, visualization or complex computer-aided writing tasks. Extractive text summarization is a content-oriented task which uses available information from text to shorten its length in order to summarize it. In this regard, a MKR knowledge base provides a structure which is applicable as an innovative selection instrument for text summarization. This paper introduces cross-dimensional text summarization based on dimensional selection and filtering of results retrieved from MKR knowledge base.**

***Keywords*—Text Summarization, Integrative Text Mining, Knowledge Representation**

I. Introduction

In the task of text summarization, authors typically select information from several relevant sources, extract the most important core information and then write a topic-specific text or abstract. This approach is carried out manually by authors, reporters and journalists, day by day. The demand for text summarization is very high since the information available in the internet is typically higher than needed and surpasses the possibilities of human capacity to summarize a text or a even a collection of texts in short time and a meaningful way.

For computers, however, text summarization is also a difficult task because it has a high complexity in dealing with natural language, identification of relationships between textual information and especially the recognition of the importance of individual issues in the big picture. This is the reason, why there is currently no general, domain-independent solution for the (automatic) text summarization task. Approaches and methods for text summarization are typically adapted to the respective context, require a special vocabulary or use an ontology in order to understand complex relationships.

Integrative text mining is an approach which combines results of individual natural language processing (NLP) and text mining methods in a knowledge base to achieve a multi-perspective analysis of the input text. The insights gained typically reflect the results of methods such as named entity recognition (NER), sentiment analysis (SA) or topic detection (TD) applied on the text. In the MKR framework proposed by the authors [1], [2], the results are stored in the knowledge base to facilitate knowledge discovery processes, provide contextual (semantic) information in adaptive visualization and support knowledge discovery or complex computer-aided writing tasks. In the MKR approach, texts are analyzed with the methods mentioned and the results are stored as dimensions. For example, the semantic relationships between the extracted entities from a text and their entity-level sentiment, given the information about multi-topics from each document, are stored in the knowledge base. Accordingly, in a large collection of documents, related documents can be retrieved based on their entity information given a topic in a respective sentiment or even emotion classification. In this paper, we use such dimension information as a selection and filtering instrument for extractive text summarization to take advantage of a combination of the individual analysis results from mentioned methods. In addition, we use semantically related information and its importance to aggregate large volumes of documents by identifying relevant parts of documents for the text summary via MKR.

Section 2 discusses related work to put the paper into the overall context of text summarization. In Section 3, the approach of dimensional text summarization is introduced. A case study and experimental results are described in Section 4. In the conclusion, the results of the paper are summarized and future work is mentioned.

II. Related Work

Several surveys have been published related to text summarization over the last decades with focus on information extraction, summarization methods and systems [3], [4], [5], [6], [7]. Text summarization concepts are generally divided into two related approaches - the *abstractive* and the *extractive text summarization*. While abstractive text summarization focuses on rewriting text based on core information, extractive text summarization identifies the most important parts from a text, a document or a collection of documents to use them in the summarization process.

A. Abstractive text summarization

In abstractive text summarization, a summarization system generally tries to understand all textual information given in documents or collection of documents. After a general knowledge has been gained or trained through machine learning, an abstract or text summary is created using pieces of content rephrased through trained vocabulary or frequent sequences of n-grams, words or even only characters. Particularly in the field of natural text generation (NLG), this approach requires a profound understanding and good modelling through machine learning and related disciplines. The abstractive text summarization task is more difficult than the extractive summarization since human readers will recognize errors in the text generated by machines and subconsciously evaluate the texts for writing style and its readability. Those difficulties make automatic abstractive text summarization a very complex and non-trivial NLP/NLG task. Through the advances in deep learning, abstractive text summarization gained more attention. Research in this area uses sequence-to-sequence Recurrent Neural Networks (RNN) for the task of abstractive summarization [8].

B. Extractive text summarization

Extractive text summarization typically uses features created from the content of the text such as term frequency [9], inverse document frequency [10], contained named entities or word co-occurrences. Furthermore, the title similarity [10] and the utilization of cue words have been proposed in the literature [11], [12]. Sentences are weighted with terms which are similar to the words in the headline [10], [11]. This method assumes that the title or subtitle is the shortest possible form of the text summary. Moreover, the layout of an article has also been considered as relevant for summarization. Therefore, the sentence location [10], [11], [12], [13] and the font styles [14] have been proposed as useful text summarization indicators.

C. Hybrid approach

The combination of extractive and abstractive text summarization has been shown in a two-stage extractive-abstractive framework by Liu et al. [15]. Extractive summarization has been used to identify relevant information and a neural abstractive model has been applied to generate Wikipedia lead sections and full articles with promising results [15].

III. Dimensional Text Summarization

The dimensional text summarization is based on the concept of MKR in knowledge bases [1], [2]. Since all potentially required information is already computed by individual text mining methods and stored in a JSON format (Listing 1), the proposed extractive text summarization algorithm benefits from dimensional selection or filtering of content and can be executed specifically on content which matches the conditions of the user request.

```
{
    "_id" : ObjectId(...),
    "created" : ISODate(...),
    "_documentId" : ObjectId(...),
    "language" : "en-EN",
    "sentimentRelation" : [
        {
            "_id" : ObjectId(...),
            "language" : "en-EN",
            "nGram" : "powerful",
            "value" : 0.639,
            "created" : ISODate(...)
        },
        {...}
    ],
    "documentSentimentValue" : 0.0458,
    "documentSentimentEmotion" : {
        "Anger" : 0.10,
        "Anticipation" : 0.10,
        "Disgust" : 0.20,
        "Fear" : 0.000,
        "Joy" : 0.70,
        "Sadness" : 0.00,
        "Surprise" : 0.100,
        "Trust" : 0.00
    },
    "sentimentEntityRelation" : [
        {
            "_entityId" : ObjectId(...),
            "_sentimentNGramId" : ObjectId(...),
            "language" : "en-EN",
            "IsShifted" : false,
            "Shifter" : [ ],
            "IsIntensified" : true,
            "Intensifier" : [
                "most"
            ]
        },
        {...}
    ],
    "topicRelation" : [
        {
                "language" : "en-EN",
                "topic" : "society",
                "subtopic" : "politics"
        },
        {...}
    ],
    "entityRelation" : [
        {
            "_id" : ObjectId(...),
            "language" : "en-EN",
            "entityName" : "Angela Merkel",
            "entityType" : "Person",
            "created" : ISODate(...)
        },
        {...}
    ]
}
```

Listing 1
JSON Format of MKR (Example Analysis Result of one Document [2])

In the following, the proposed algorithm is explained with a focus on entity related text summarization. For example, if a user wants to summarize all documents, in which German chancellor Angela Merkel is mentioned in politics topic, normally all documents need to be scanned if they contain information about the entity and match to requested politics topic. Also documents which contain only the term "Merkel" won't be found if the search is similar to a full-text search. In the dimensionally structured knowledge base, this information is available after the document has been analyzed once since it stores the results of NER and TD applied on the document in corresponding $entityRelation$ and $topicRelation$ in the final MKR output. In this way, the search is backward-oriented and

starts with the filtered analysis results in order to process less documents and correct content. Especially for very complex multi-dimensional search queries, the dimensionally arranged structure has a great advantage, since the analysis results can be used for text summarization applying selection or filtering operations.

The algorithm illustrated in Fig. 1 is explained as follows:

1) The knowledge base (KB) is queried based on different filters for MKR analysis results. The analysis results represent the combinations of analysis dimensions *NER, SA, TD*.
   a) A $time$ filter removes content which doesn't fit into user request (if $time$ specified).
   b) Filtering or selection of dimensions $entity$, $sentiment$ and $topic$ (relations in MKR).
2) The analysis results from the subset are retrieved from KB.
3) While the list of $MKR_{i...n}$ is not empty, perform following actions:
   a) Load the MKR_i related document D_i.
   b) Perform *NLP operations* (e.g. paragraph splitting, sentence splitting, tokenizing, part-of-speech tagging) in order to pre-process D_i.
   c) For each sentence $S_{j...m}$ from D_i:
      i) If S_j contains the requested named entity e ($entityRelation$), S_j is kept, otherwise it is skipped.
      ii) If the user specified *entity-to-entity* relationship, S_j is further checked if it contains all entities $e_{i...n}$. If yes, S_j is kept, if no, S_j is skipped.
      iii) If the user specified a *sentiment polarity* $p = \{p \in \mathbb{R} \mid -1 \geq p \leq 1\}$ in *range* or *level*, S_j is checked if it matches requested p ($sentimentEntityRelation$). If yes, S_j is kept, if no, S_j is skipped.
      iv) S_j is appended on the output text T. A *reference* R_i to the document D_i (metadata) is created in a *reference section* at the end of the text (if R_i is not yet in *reference section*).
      v) If S_m is reached, remove MKR_i from list and select the next analysis result MKR_{i+1} and its document D_{i+1}. The algorithm continues at step 3.
4) The algorithm exits if n is reached in the MKR list and therefore, all related documents D_n have been analyzed and all related information has been extracted for summarization.

IV. Experimental Result

For demonstration, we extracted a small collection of 81 news articles crawled from Reuters News[1] between October 2017 and November 2017 in the detected $topic$ area of "society" via MKR. The $entity$ "Angela Merkel" has been selected as dimensional filter ($entity$ is listed in the $entityRelation$ of MKR). The proposed algorithm has extracted the text given in Table I. In a second run, the $entity$ combination of "Angela Merkel" and "SPD" has been chosen as dimensional entity filter with an overall positive sentence-level $sentiment$. The summarization results are given in Table II.

As this example shows, the result is a merged text from various sources (webpages) which contain the identified sentences. Therefore, the abstraction is linguistically weak, but it is very well tailored to the requested search, as each sentence actually contains information related to the input (in this case, the $entity$ "Angela Merkel"). Table II further illustrates how the dimensional filtering influenced the summarization given in Table I. Here, only information is given related to the combination of $entities$ "Angela Merkel" and her opponent party "SPD" (Germany's Social Democrats) with overall positive sentence-level $sentiment$.

A. Limitations

The identification of relevant information is particularly important for the extraction of textual content from documents. This process can be algorithmically modelled in a similar way as human authors would do in their search for information. However, parts from the content of identified sources have to be put together meaningfully in order to create a coherent text, especially stylistically, but also chronologically.

The outlined approach has so far left these two problems untreated. Nevertheless, the approach is suitable for Wikipedia-like chronological summarization, since the algorithm can process documents one after the other using time stamp information. Thus, according to the user input, a summary from a chronological sequence of referenced articles is created, which can even be further adapted or modified in different criteria (entities, sentiment, topic) via the MKR approach.

Here we identify further potentials for abstractive text summarization since the produced text could be used as an input or training dataset to rewrite or rephrase selected sentences.

V. Conclusion

In this paper, we have presented an approach for extractive text summarization using the multidimensional knowledge representation (MKR) framework developed by the authors in previous research [1], [2]. The dimensional selection and filtering operations are applied on analysis results as an instrument to identify (entity, sentiment, and topic) related documents for extractive text summarization. In an example, the presented algorithm selects relevant sentences based on entity, topic and sentiment relationships and creates a chronologically ordered text summary. In future work, we will use a character length criterion on the output text to further modify the summary and select the best matching sentences according to available MKR information in sentence weighting. Furthermore, we are using the approach to select relevant training data for a long short-term memory (LSTM) recurrent neural network (RNN) in order to rephrase or rewrite the text, similar as a human author would do in the next step of text summarization.

[1] https://www.reuters.com/

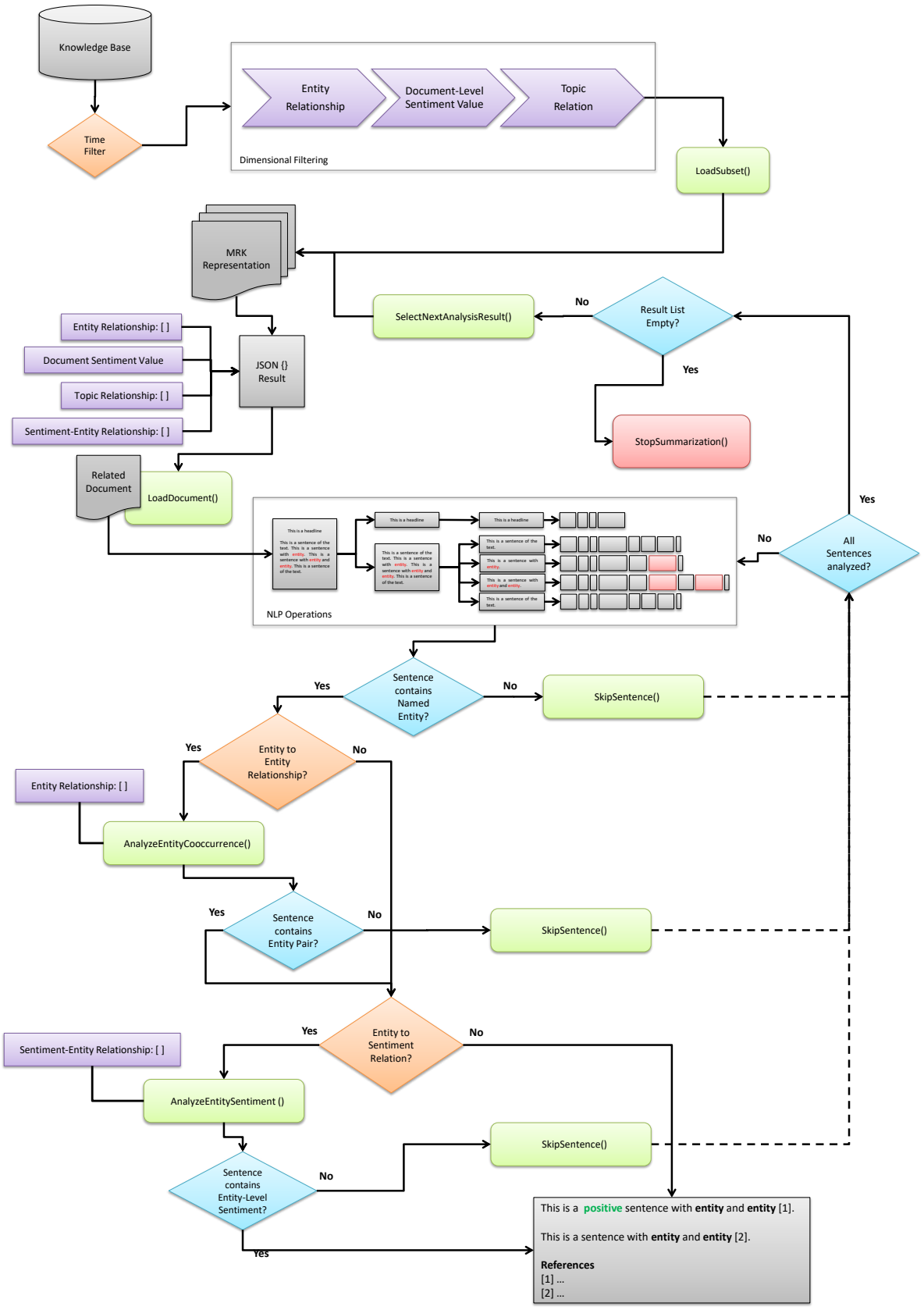

Fig. 1. Dimensional Text Summarization Algorithm in MKR knowledge base

TABLE I

DIMENSIONAL TEXT SUMMARY RELATED TO NAMED ENTITY "ANGELA MERKEL" (VIOLET) WITH TOPIC FILTER "SOCIETY" FROM REUTERS.COM NEWS CORPUS BETWEEN 08.10.17 AND 27.11.17.

German Chancellor **Angela Merkel**'s Christian Democrats (CDU) have agreed on the divisive issue of a refugee cap with her conservative Bavarian allies, two conservative sources told Reuters, removing a hurdle to coalition talks with other parties.[a]

German Chancellor **Angela Merkel** attends a Lower Saxony's Christian Democratic Union's (CDU) regional election campaign rally in Stade, Germany October 13, 2017. The latest opinion poll put the SPD on 34.5 percent in Lower Saxony, giving it a 1.5 point lead over **Merkel**'s Christian Democrats (CDU) - who had been 12 points ahead at the start of the campaign in August. A deal brokered last weekend between **Merkel**'s CDU and its conservative Bavarian sister party, the CSU, to cap the number of immigrants is likely to be hard for the Greens to swallow. **Merkel**'s reverse in September left her with no viable option other than a "Jamaica" coalition, so named because the three parties' colors correspond with the black, yellow and green of Jamaica's flag.[b]

German Chancellor **Angela Merkel** attends the first plenary session of German lower house of Parliament, Bundestag, after a general election in Berlin, Germany, October 24, 2017.[c]

Two veteran allies of Chancellor **Angela Merkel** appealed to Germany's parties on Tuesday to strike compromise to form a stable government that could drag Europe's biggest economy out of a political impasse. It has also cast some doubt over whether **Merkel**, Europe's most powerful leader after 12 years in office, will serve a fourth term after her conservatives bled support to the far-right in a Sept. 24 election, though still won the most seats.[d]

Members of Germany's Social Democrats (SPD) will likely approve a renewed coalition with Chancellor **Angela Merkel**'s conservatives if party leaders present a convincing proposal, a member of the party's executive leadership said on Saturday. German Chancellor **Angela Merkel** speaks with Social Democratic Party (SPD) leader Martin Schulz as they attend a meeting of the Bundestag in Berlin, Germany, November 21, 2017. Schulz said party leaders agreed to talks out of a sense of responsibility to Germany and Europe after **Merkel**'s attempt to form a government with two smaller parties collapsed on Sunday.[e]

The Alternative for Germany (AfD) far-right party asked members of the public to share pictures showing extra security measures at their local markets and post them on social media in protest against Chancellor **Angela Merkel**'s decision in 2015 to open Germany's borders to more than a million asylum seekers. The AfD blames **Merkel**'s immigration policy for what it says is a rise in crime and Islamist attacks.[f]

[a] http://www.reuters.com/article/us-germany-politics-conservatives/merkel-bavaria-allies-agree-on-migrant-policy-sources-idUSKBN1CD0S1 (08.10.2017 17:38)
[b] http://www.reuters.com/article/us-germany-election-lower-saxony/state-vote-unlikely-to-give-merkel-boost-in-german-coalition-talks-idUSKBN1CJ0T5 (14.10.2017 22:14)
[c] http://www.reuters.com/article/us-germany-politics/merkel-to-try-to-kickstart-german-coalition-talks-media-idUSKBN1CY0M3 (29.10.2017 15:24)
[d] https://www.reuters.com/article/us-germany-politics/german-political-grandees-press-parties-to-compromise-for-stability-idUSKBN1DL0TC (21.11.2017 12:49)
[e] https://www.reuters.com/article/us-germany-politics/momentum-grows-for-another-grand-coalition-in-germany-idUSKBN1DP01O (25.11.2017 02:07)
[f] https://www.reuters.com/article/us-christmas-season-germany/germanys-christmas-markets-open-under-tight-security-a-year-after-attack-idUSKBN1DR2HM (27.11.2017 19:25)

TABLE II

DIMENSIONAL TEXT SUMMARY RELATED TO NAMED ENTITIES "ANGELA MERKEL" AND "SPD" (VIOLET) WITH TOPIC FILTER "SOCIETY" AND OVERALL POSITIVE SENTENCE-LEVEL SENTIMENT (GREEN) FROM REUTERS.COM NEWS CORPUS BETWEEN 08.10.17 AND 27.11.17.

The latest opinion poll put the **SPD** on 34.5 percent in Lower Saxony, giving it a 1.5 point **lead** over **Merkel**'s Christian Democrats (CDU) - who had been 12 points ahead at the start of the campaign in August. The north German state of Lower Saxony holds an election on Sunday that looks likely to hand the **Social Democrats (SPD)** a narrow **victory**, and **deprive** Chancellor **Angela Merkel**'s conservatives of a **boost** in **looming** national coalition talks.[a]

Members of Germany's **Social Democrats (SPD)** will likely **approve** a **renewed** coalition with Chancellor **Angela Merkel**'s conservatives if party leaders present a **convincing** proposal, a member of the party's executive leadership said on Saturday. German Chancellor **Angela Merkel** speaks with **Social Democratic Party (SPD)** leader Martin Schulz as they **attend** a meeting of the Bundestag in Berlin, Germany, November 21, 2017.[b]

[a] http://www.reuters.com/article/us-germany-election-lower-saxony/state-vote-unlikely-to-give-merkel-boost-in-german-coalition-talks-idUSKBN1CJ0T5 (14.10.2017 22:14)
[b] https://www.reuters.com/article/us-germany-politics/momentum-grows-for-another-grand-coalition-in-germany-idUSKBN1DP01O (25.11.2017 02:07)

REFERENCES

[1] J. Zenkert and M. Fathi, "Multidimensional knowledge representation of text analytics results in knowledge bases," in *Electro Information Technology (EIT), 2016 IEEE International Conference on*. IEEE, 2016, pp. 0541–0546.

[2] J. Zenkert, A. Klahold, and M. Fathi, "Knowledge discovery in multidimensional knowledge representation framework - an integrative approach for the visualization of text analytics results," *Iran Journal of Computer Science*, pp. 1–18, 2018.

[3] K. Zechner, "A literature survey on information extraction and text summarization," *Computational Linguistics Program*, vol. 22, 1997.

[4] E. Lloret, "Text summarization: an overview," *Paper supported by the Spanish Government under the project TEXT-MESS (TIN2006-15265-C06-01)*, 2008.

[5] V. Gupta and G. S. Lehal, "A survey of text summarization extractive techniques," *Journal of emerging technologies in web intelligence*, vol. 2, no. 3, pp. 258–268, 2010.

[6] A. Nenkova and K. McKeown, "A survey of text summarization techniques," in *Mining text data*. Springer, 2012, pp. 43–76.

[7] M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E. D. Trippe, J. B. Gutierrez, and K. Kochut, "Text summarization techniques: A brief survey," 2017.

[8] R. Nallapati, B. Zhou, C. Gulcehre, B. Xiang *et al.*, "Abstractive text summarization using sequence-to-sequence rnns and beyond," *arXiv preprint arXiv:1602.06023*, 2016.

[9] H. P. Luhn, "The automatic creation of literature abstracts," *IBM Journal of research and development*, vol. 2, no. 2, pp. 159–165, 1958.

[10] C. Nobata, S. Sekine, M. Murata, K. Uchimoto, M. Utiyama, and H. Isahara, "Sentence extraction system assembling multiple evidence." in *NTCIR*, 2001.

[11] H. P. Edmundson, "New methods in automatic extracting," *Journal of the ACM (JACM)*, vol. 16, no. 2, pp. 264–285, 1969.

[12] M. A. Fattah and F. Ren, "Ga, mr, ffnn, pnn and gmm based models for automatic text summarization," *Computer Speech & Language*, vol. 23, no. 1, pp. 126–144, 2009.

[13] P. B. Baxendale, "Machine-made index for technical literaturean experiment," *IBM Journal of Research and Development*, vol. 2, no. 4, pp. 354–361, 1958.

[14] J. Kupiec, J. Pedersen, and F. Chen, "A trainable document summarizer," in *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1995, pp. 68–73.

[15] P. J. Liu, M. A. Saleh, E. Pot, B. Goodrich, R. Sepassi, L. Kaiser, and N. Shazeer, "Generating wikipedia by summarizing long sequences," 2018.