ISBN: 978-1-948012-17-1

Ethics and Information Technology (ETIT)

DOI: <http://doi.org/10.26480/etit.02.2020.142.146>

CrossMark

A HYBRID APPROACH FOR TEXT SUMMARIZATION

Shabbir Sidhpurwala, Saiyam Jain, Sushma Verma*

Department of Computer Science and Engineering, Medi-Caps University, AB Rd, Pigdambar, Rau, Indore-453331, India

*Corresponding Author Email: sushma.verma@medicaps.ac.in

This is an open access article distributed under the Creative Commons Attribution License CC BY 4.0, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

ARTICLE DETAILS

Article History:

Received 25 October 2020
Accepted 26 November 2020
Available online 03 December 2020

ABSTRACT

In the era of information and communication technology, the data plays a vital role. The majority of these data can be reached online that empowers the one anytime and anywhere. The boundless access, cooperation and distribution of data have therefore uplifted the enduring of 'Data Glut'. This paper focuses on helping the clients with instructive requisite of data by avoiding its examining or surfing time. A system has been proposed that can provide data by making the assessing precise. Automatic text summarization is the way to consolidate any unique report into shorter format to make a littler or smaller form of it. Despite of the numerous systems proposed, there are no 'silver slugs' to accomplish the superlative outcomes as of human created outlines. Therefore, the necessity of advancement in the field of text summarization is well chief. The proposed system implements four different techniques of summarization in hybrid form. All the four methodologies are combined using a merge algorithm that summarizes all small and multi datasets. The method proposed produces a more effective and accurate text summarization when compared to the existing methods.

KEYWORDS

Fuzzylogic, Latent semantic analysis, LexRank, TexRank.

1. INTRODUCTION

As indicated by www.worldwidewebsize.com, the indexed Web contains 5.65 billion pages (11 may, 2020). The enormous expansion in the speed, volume and assortment of data available on the web has given a solid rebirth of enthusiasm for Web Information Retrieval (WIR). A definitive test of WIR is to give improved frameworks that recover the most significant data accessible on the web to fulfil a client's data need. With the change of Web into a standard choice help and proposal device, handling the test of "data over-burden" on the Web has gotten progressively imperative. The Web IR investigate is normally sorted out in errands with explicit objectives to be accomplished. Without any difficulty in accessibility of Internet and expanded utilization of savvy gadgets like cell phone, personal computers, tablets the entrance and capacity of information has heightened quickly, bringing about need of devices which can improve the client's efficiency and experience. A basic keyword search on the Internet brings about many outcomes in under a second, some of which are not in any case applicable to the clients' inquiry and finding the appropriate data is a troublesome and tedious errand.

Summarization has been distinguished as a successful WIR task, which encourages clients to find the correct data at the ideal time encouraging convenient choices. Human's way to summarize can be one-sided, setting subordinate and may differ with human perception. Subsequently, reasonable procedures and devices are expected to coerce relevant and basic areas with the goal that basic data in the form of summary is procured; giving a machine produced synopsis liberated from predisposition. The thought is to make littler, consolidated renditions from the bounteous data that is accessible, protecting the centre idea and importance with the end goal that it addresses the issues of the client. Automatic text summarization strategies can be utilized for just removing

the keywords, regardless of what the aim, type or context of synopsis created is, the essential target is to help client's educational pursuits with diminished surfing time and furthermore improve the record ordering effectiveness simultaneously. It helps to fasten the searching procedure. The significance of an article to our intrigued theme can be deducted by the significant catchphrases of it. Removing watchwords and the utilizing text summarization calculation will improve the inquiry procedure.

A text summarization strategy has been proposed which is hybrid of four strategies namely: Fuzzy logic (Ghalehtaki et al., 2016), Latent Semantic Analysis (Aggarwal et al., 2012), LexRank and TextRank (Cambria et al., 2014), each having their own upsides and downsides. United, they can bring about progressive and sure outcomes. Fuzzy logic has showed up as a ground-breaking hypothetical structure for examining human thinking and its application, it has been investigated inside the space of text summarisation in the past few years. One key part of text summarization is exact distinguishing proof of keywords from the given printed content. In this undertaking, another strategy dependent on Fuzzy logic has been proposed utilizing two diagram-based strategies named as TextRank and LexRank and one Semantic based procedure named as Latent semantic analysis (LSA).

The proposed approach is compared with Fuzzy logic, Ranking calculation, chart base calculation and Sentence, utilizing ROUGE measures. The outcome shows that the proposition gives better f-measure with 95% more than previously mentioned strategies examination of this exploration is practiced utilizing DUC-2002 (Witte et al., 2005), a standard dataset for archive rundown. Hence, this examination presents a semantic diagram-based strategy for MIDAS (multiple intelligence development assessment scale), Fuzzy guideline, and other chart base methodology (TextRank and LexRank), which endeavours to conquer the burdens of

Quick Response Code	This paper was presented at	Access this article online
International Conference on Contemporary Issues in Computing (ICCIC-2020) - Virtual IETE Sector V, Salt Lake, Kolkata From 25th-26th July 2020	Website: www.intelcomp-design.com	DOI: 10.26480/etit.02.2020.142.146

existing chart based and other approaches of machine learning. The commitments of this exploration are featured as given:

- Introduction of merging algorithm,
- Use of Fuzzy logic,
- Use of semantic graph approach,
- Use of graph positioning algorithm.

We have validated our proposed work with the relative execution of the existing work. The proposed technique is utilized in creating hybrid model of extractive outline type. The strategies are assessed on opiosis informational collection utilizing 'ROUGE-1' and 'time to remove the keywords or watchwords' (Witte et al., 2005).

The rest of the paper is organized as follows: Section 2 covers the related work with their limitations. The proposed methodologies elaborated in section 3. Section 4 comes up with the result of the proposed approach and section concludes the paper along with the future scope.

2. RELATED WORK

2.1 Related work on Fuzzy standards

A study built up a book outline strategy dependent on Fuzzy logic and WordNet. They proposed two methods; by breaking down the equivalent words of the words in the sentences and the entire content. The strategy was designed with nine info factors and four yields (Kyoomarsi et al., 2010). The outcomes with ROUGE and the DUC 2003 dataset accomplished 0.60 in exactness study proposed a methodology hybridizing Fuzzy logic, GA and Semantic Role Annotation. According to ROUGE-1 measure, the proposed technique accomplished 0.47 in f-measure (Binwahlan et al., 2010).

A recent study proposed a book outline strategy dependent on Cellular Automata. The technique additionally utilizes genetic algorithm and Fuzzy logic (Kyoomarsi et al., 2007). Three summarization techniques are inspected: (i) content rundown dependent on the genetic algorithm strategy, (ii) Text summarization dependent on the Fuzzy logic technique (Ghalehtaki et al., 2016), and (iii) Text summarization dependent on the category algorithm (CA) strategy. Every technique utilized is using eight highlights to compute sentence score, and afterward the best sentences in the positioning are removed from each summarizer. The writers utilized an example of 17 English scientific articles to assess their techniques (16 archives for preparing and 1 for testing). The Fuzzy logic accomplished 0.46 in accuracy, while the techniques utilizing CA and GA accomplished 0.31 and 0.59 in exactness, individually. We adapt reinforcement learning to our formulation of extractive summarization to rank sentences for summary generation (Narayan et al., 2018).

Hannah et al portray a book summarization Fuzzy logic utilizing seven qualities. The characteristics were arranged into three sets Fuzzy (lower, centre and higher) with trapezoidal membership functions. The Fuzzy strategy classified sentences into three types namely: not significant, medium significant or major importance. The outcomes with ROUGE and the DUC 2002 dataset demonstrated a presentation 0.48 in f-measure.

In a work studied the performance of two *text summarization* techniques. One of them utilized the Fuzzy logic while the other used the Artificial Neural Network (ANN) system. The creators selected preliminaries legitimate writings for the analyses and continued with a non-automated assessment. The neural system summarization technique accomplished 0.42 in f-measures, while 0.46 in Fuzzy logic. Also represented a summarizer framework utilizing separate measures with Fuzzy logic to create the synopsis and Conditional Random Field (CRF) for classifying the summary segments. The framework was tried with the contribution of 30 Legal Judgements. In resultant analysis, the presentation was 0.26 in f-measure (Megala et al., 2014).

A previous research portray a multi document summarization dependent on news segments utilizing Fuzzy cross documents relations. The proposed methodology has three stages: (i) part sentence extraction, (ii) Cross document Structure Theory (CST) model which depicts the semantic relations between literary units, (iii) sentence scoring utilizing Fuzzy reasoning. The CST connection is identified utilizing a Genetic-Case Base Reasoning (CBR) model with five diverse features for each sentence pair taken from a report group. The Fuzzy reasoning model is utilized for sentence scoring (Kumar et al., 2014). The most elevated positioned sentences are chosen until the ideal synopsis length is met. The creators utilized the dataset acquired from CST Bank for the preparation and testing, the outcomes with ROUGE-1 and DUC 2002 dataset accomplished 0.33 in f-measure. A study proposed a method in which they are mapping

a sentence selected by an extractive model into the summary by training a sentence compression model; the quality of a selected sentence was evaluated using the model with respect to the summary. However, the method was used for both single and a multi document and the optimized techniques are applied in an optimized way (Zhang et al., 2018). The method has the limitations i.e. for further improvement in the result Fuzzy logic can be used. The method's pre-processing data sets remains same. In 2019, proposed an algorithm that greedily selected the sentences which can maximize the ROUGE scores as the oracle sentences. In the method, they assigned label 1 to sentences selected in the oracle summary and 0 otherwise. The model has the advantage that it used a trained masked language named BERT for extractive summarization. It also used trigram blocking that reduced the redundancies in the summary. But, the method worked for only single document file and unlike abstractive summarization, the rewriting operation was impossible (Liu, 2019).

2.2 Related work on Semantic Graph Approach:

In this segment, from the outset we exhibit the past methodologies presented for MIDAS, at that point we examine the diagram-based methodologies proposed for multi-record extractive summarization and single archive abstractive rundown. A couple of analysts have strived to make abstractive summaries utilizing an assortment of techniques, which can be organized into two classes: Syntactic or Linguistic and Semantic methodology. All phonetic based methodologies proposed for abstractive synopsis use syntactic parser to speak to and examine the content. The striking weakness of these investigations is the non-presence of semantic representation of source record content. It is critical to speak to the content semantically, as significant semantic content investigation is carried out in abstractive summarization. Be that as it may, distinctive semantic methodologies have been researched for MDAS and are shown as follows.

GISTEXTER is multi document summarization (MDS) framework which utilizes the format-based technique to produce abstractive summaries from various news reports (Khan et al., 2015). However, the extraction rules or etymological examples were physically created, which require more exertion and time. A Fuzzy ontology approach is introduced for summarization of Chinese news, which models dubious data to portray the area information in a superior manner (Lee et al., 2003). In this methodology, the Chinese word reference and space metaphysics are characterized accurately by a human master, which is a monotonous errand. The system introduced in creates abstractive summary from the semantic model, speaking to a multimodal record. The information spoke to by ideas of philosophy is used to fabricate the semantic model. The shortcoming of this model is that it relies upon human master to develop space cosmology, and doesn't make a difference to different areas. The approach introduced by delivers elegantly composed and compact abstractive outlines from the gatherings of news stories on comparative theme. The impediment of the philosophy was that age examples and data extraction rules were physically composed, which requires exertion and time.

As of now, graph-based techniques are in highlights which are effectively attempted for MDS. These models utilize PageRank calculation and its varieties to dole out positions to sentences or paragraphs. Authors in introduced a network model dependent on chart, which expect that hubs which are connected to a few different hubs are most presumably to convey critical data.

3. PROPOSED METHODOLOGY

In various fields and informational index, the text summarizer performs diversely for single informational set. At mostone uses LexRank and don't utilize this in multi dataset. In light of the fact that it is not giving improved arrangement, therefore for we utilize Fuzzy logic or latent semantic analysis. In contrast of the available program we give an efficient method for this. An alternate methodology is used to give an outcome which depends on nature of data collected.

Our proposed work introduces four different techniques of summarization. Each technique produces its own results. Combination of all the four techniques resulted in extraordinary summarization results. Out of four techniques some are good for small datasets and some are good for multi-datasets, therefore united they produce effective results. Each technique generates different value of precision, recall, and f-measure for different datasets.

Initially our dataset which we needed to summarize are setting off to all the chosen techniques (TextRank, LexRank, Latent semantic analysis and

Fuzzy logic). Each technique produces its own summary by its own logic. After generating the summary from all the techniques, it needs to be merged in order to generate a single effective summary. Merging algorithm is developed to merge all the summaries.

3.1 Merge Algorithm

The ultimate goal of the merging algorithm is to generate a single effective summary by combining all the summaries generated by the specific techniques. The framework of merging algorithm is designed in such a way that it will produce the most efficient result. Presently we discuss the mechanism of merging algorithm and how it works. Merging algorithms are used for consolidating and finding the best yielded summary which is created from various techniques. This algorithm combines all the summaries generated by the respective techniques and then united present a single, efficient and most accurate summary. Merging algorithm also helps in finding the best technique among them by comparing the resulted summary (summary generated by merging algorithm) with the individual summary (individual summary includes the summary of TextRank using LexRank, Latent Semantic Analysis and Fuzzy Logic). After comparing, individual vector score is given to each of the technique according to its performance. The technique with maximum vector score tells us that it produces the most similar results as the result is produced by merging algorithm.

3.2 Proposed Framework

Existing works use either graph-based techniques or semantic techniques or on the Fuzzy logic for text summarization. Graph based technique consists of TextRank and LexRank where the semantic techniques consist of Latent semantic analysis (Cambria et al., 2014). Graph based technique extract the keyword using structure of the article. Our hybrid model comprises of four parts: TextRank, LexRank, LSA, and Fuzzy Logic. In this model, every text summarization strategy is utilized to remove the keywords. Every technique has positioned the keyword dependent on its significance (repeat, centrality, thing and so on.), and these keywords with their scores are then given as input to the last keyword extractor. In this stage, the keywords occurred is the conclusive outcome of all the four strategies taken into definite keyword list. At that point the rest of the keywords of the considerable number of techniques are organized in decreasing order of the scores. From this list the top keywords are chosen in the last yield. Keyword extraction should be done according to the need of the keywords. Usually 1% of the complete keywords (aside from stop words) are sufficient to speak to the central idea of the document.

At first the document goes through a pre-handling stage to retrieve the necessary information. For example, In the previous line keywords like: is, the, a, to, just won't be the keywords speaking the central idea of the document. Along these lines, we first expel the undesirable stuffs from the record and afterward additionally include few words which semantically mean the equivalent in order to improve the semantic comprehension. For instance, the keywords like improving, upgrading are regularly mean the equivalent. Therefore, we can supplant one with the other, so that the actual occurring of the word could be distinguished.

In TextRank the keywords or sentences of the documents are represented as the vertices of the graph and the edges connect the vertices which has some sort of relationship. For example, in case of sentences the common words, the position of two sentences, for keywords nouns, adjectives etc. can be used for defining the relationship between the vertices.

The numbers of common tokens like noun, verbs which are keywords to the text define the similarity of sentences or also called as overlapping nature.

LexRank computes the significance of a sentence or keywords by finding their eigenvector centrality in the graph representation of the sentences.

3.3 Architectural View

According to the proposed approach, a document needs to be chosen from opinois dataset. Followed by its pre-processing to expel the undesirable words and changing over rest of the words to its stem(root). For keyword extraction the four techniques form the archive (process the document) in its own specific manner. Afterward, the keywords extricated from all the strategies are combined and orchestrated in plunging request of their scores.

At long last, the top 1% keywords are extricated from the document which shows the overview of the framework proposed in this research.

Figure 1: Framework of hybrid approach of text summarization

3.4 Proposed Algorithm

Each of the four techniques has been implemented individually; at that point their output is given as the input to the algorithm to extract the progressively viable keywords. It needs to be done very effectively to get the central idea of the document.

```

Initialize four variables
- let Fuzzy_sum = []
- let lsa_sum = []
- let Text_rank = []
- let lex_rank = []
- let total_line
// Initialize one more variable to store text sentence which is separated by "."
-       Let text=this.faqform.value.text_split(".")
-       total_line = text_line
// Now we take data from all the four algorithm and then store in the array
[]
-let Fuzzyrank = await this.post()
let lsatextrank = await this.post()
let lexrank = await this.post()
// Initialize two more variable
let sentence = {}
let array []
// Concatenate all the summary and store in the array
Arr = arr.concat(lex,Fuzzy,text,lsa)
// We calculate the important highlighted keywords from each and every
sentence by giving rank to them on the basis of their vector score.
For(let S=0; s<arr.length; S++)
{
    Let element:strings = arr[S]
let fullstop = element.substring(element.length - 1, element.length)
if (fullstop == ".")
{
    element = element.substring(Q.element.length - 1)
}
If(type of element == "string")
{
    If(sentence[element])
    {
        ++sentence[element]
    }
    Else
    {
        sentence[element] = 1
    }
}
}
// Now we had given rank to every line/sentence, let us concatenate and
sort in a order so that we can present the summary
For (let P=Q; f<object.key(sentence).length; f++)
{
    Const. key = object.keys(sentence)[f];
    Summary.push{
    Text : key,
    Value : sentence[key]
  
```



```

}
Summary.sort(($Q,$1) => {
    Return $Q.value - $1 value
})
// Now we join summary with "." And final output

For(let s=0; s < summary.length; s++)
{
    Const sum summary[s];
    If(s < summary.length)
    {
        Final summary +=sum.text+"."
    }
}
this.summary = final summary

```

3.5 Programming Aspect

Initially, feature selection needs to be done. The selection procedure of feature extraction includes eight highlights for deciding the noticeable quality score of a sentence. These are 'title word, sentence length, sentence position, numerical information, thematic words, sentence to sentence comparability, term weight and proper noun'. The feature scores are then given as the contribution to the Fuzzifier. It utilizes the trapezoidal membership function and represents the scores in s Fuzzy values. These are then given to the inference engine which works on the basis of Fuzzy set and then the status of the sentence is extracted. De-Fuzzifier is the last step which provides a prominent score of each individual sentence with the help of triangular membership function. These sentences with their score are then given as input to the merger subsystem of the proposed technique.

The remaining three strategies TextRank, LexRank and LSA were executed in a Linux domain using bash and python. The bash contents were composed to go about as drivers for mechanization of the undertaking. The calculations were executed with python. The generated output file is then passed through the merger system.

The final summary done by an algorithm which took four files generated by each method as input files containing the sentence and their scores determining their importance in the document. Firstly, the content of all four files are merged in one file, then the sentences which were occurring in more than one input file are kept only once to remove the redundancy. The sentences which were coming in more than one algorithm as the important sentence are first taken into the output file and are removed from the merged document. The rest of the sentences in the merged document are then arranged in descending order of their scores. If a document is of 1000 sentences and we want summary of only 20 lines that is only 2% of the actual document. We will select the top 20 sentences from the arranged list, only if none of the sentence were occurring in more than one method. If some sentences were coming as important in more than one method, say 3 sentences were found in summary of more than one method, and then from final list only 17 sentences will be selected.

4. RESULT

The evolution of Web from the 'Web of Documents' to the 'Web of People' (Social Web) to the 'Web of Data' (Semantic Web) has multiplied the volume, velocity and variety of information available online in different ways. Among them the motive is to find useful, relevant, trending, and interesting information from it. Text summarization has emerged as one of the highlighted techniques to handle this problem. The incredible ability of Fuzzy inference systems to make logical assessment in an ambiguous and uncertain environment has made it a trending choice for practical applications such as text summarization, that involve imprecision and uncertainty.

In our work, we have reviewed various studies on research papers based on text summarization from 2017 to 2019, published in conference proceedings and journals of high reputation nationally and internationally. The purpose is to evaluate the progress made so far and identify the trends and gaps in studies to ascertain the future scope of research within the domain. The following key-points were observed:

Fuzzy logic is an attractive choice to achieve optimal summaries due to its resemblance to human reasoning. The goal of text summarization is to achieve superlative results comparable to human generated summaries and mapping the Fuzzy logic inference mechanism gives the desired brainpower.

It is promising to see the change in perspective from ordinary summarization strategies to contemporary, novel insight-based techniques. About half of the investigations have been done with hybrid models of Fuzzy logic with improvement strategies, trailed by equivalent number of studies done on hybrids of Fuzzy logic with statistical and semantic systems individually. Combinational hybrids using at least two or more procedures with Fuzzy logic have also been accounted but the circle is an open research issue to accomplish improved summary results. Hybrid models taking semantics into consideration and generating the better result comparative to other techniques.

The model that is proposed in this paper is applied on random text which we get from internet, like we use news article, scientific article or paper, research paper and short story. Some of the screens attached here, on which the method is applied. The corresponding outputs are also attached near below.

Figure 2: Our Test data

We have used the article on big bang theory as a test data set, reference of this www.livescience.com/65700-big-bang-theory.html, title are "The Big Bang Theory: How the Universe Began" by Charlie wood June 12, 2019, shown in Figure 2.

The summarized text after applying the test data is kept below in Figure 3.

Figure 3: Result

The exploration in this work includes a Fuzzy logic-based hybrid model taking two graphs-based technique and one semantic based strategy together to produce a redesigned summary. It performs better than the entire Fuzzy logic-based model. The proposed system can outflank the current strategies, when contrasted and the outcomes given by the real examinations.

5. CONCLUSION AND FUTURE SCOPE

THE INVESTIGATION IS OBSERVATIONAL IN NATURE and can be improved by giving different weights to each of the technique among the four techniques. The weight given to the strategies can influence the output of the proposed technique. Features highlighted in the feature extraction step can likewise be given diverse weightage to make the summary creation process like people. As human don't accept each attribute as same, similarly, the title keywords are given more significance than the noun. Along these lines, the various weight allotted to each feature can bring more efficient and accurate outcomes.

We can additionally broaden this work by:

- Assigning various weights to every technique utilized.

- Assigning various weights to the highlights removed during the time spent in summarization produced utilizing Fuzzy logic.
- Incorporate some more Nature Language based procedures to produce abstractive summary.
- The proposed method can be implemented using standard data sets.

ACKNOWLEDGEMENT

We thank Pramod Kumar of the Indian Institute of Technology (Indian School of Mines), Dhanbad for his assistance with the appropriate ideas in scripting this paper.

REFERENCES

- Aggarwal, Charu C., and Zhai C.X., eds. , 2012. Mining text data. Springer Science & Business Media, 2012.
- Binwahlan, Salem M, Salim N, and Suanmali L. 2010. Fuzzy swarm diversity hybrid model for text summarization. Information processing & management, 46(5), 571-588.
- Cambria, Erik, and White B. 2014. Jumping NLP curves: A review of natural language processing research. IEEE Computational intelligence magazine 9(2), 48-57.
- Cheng J, Lapata M. 2016. Neural Summarization by Extracting Sentences and Words ACL2016 conference paper with appendix.
- Chopra S, Auli M, Rush A.M. 2016. Abstractive Sentence Summarization with Attentive Recurrent Neural Networks. San Diego, California, c 2016 Association for Computational Linguistics.
- Ghalehtaki A, Razieh, Khotanlou H, and Esmaeilpour M. 2016. Fuzzy evolutionary cellular learning automata model for text summarization. Swarm and Evolutionary Computation. 30, 11-26.
- Khan, Atif, Salim N, and Kumar Y.J. 2015. A framework for multi-document abstractive summarization based on semantic role labelling. Applied Soft Computing, 30, 737-747.
- Kumar Y.J, Salim N, Abuobieda A, and Albaham A.T. 2014. Multi document summarization based on news components using Fuzzy cross-document relations. Applied Soft Computing 21, 265-279.
- Kyoomarsi, Farshad, Khosravi H, Eslami E, and Davoudi M. 2010. Extraction-based text summarization using Fuzzy analysis. Iranian Journal of Fuzzy Systems 7(3), 15-32.
- Kyoomarsi, Farshad, Torkestani J.A, Dehkordy P.K, and Peyravi M.H. , 2007. Using chemical cellular automata in simulation of chemical materials. In 5th ACIS International Conference on Software Engineering Research, Management & Applications (SERA 2007), 769-773. IEEE.
- Lee, Shing C, Chen Y.J, and Jian Z.W. 2003. Ontology-based Fuzzy event extraction agent for Chinese e-news summarization. Expert Systems with Applications. 25(3), 431-447.
- Liu Y. 2019. Fine-tune BERT for Extractive Summarization Cornell University.
- Megala, Santhana S., Kavitha A., and Marimuthu A. 2014. Enriching text summarization using Fuzzy logic. International Journal of Computer Science and Information Technologies, 5(1), 863-867.
- Narayan S, Cohen S.B, Lapata M. 2018. Ranking Sentences for Extractive Summarization with Reinforcement Learning. NAACL 2018.
- Witte, René, Krestel R, and Bergler S. 2005. ERSS 2005: Coreference-based summarization reloaded. In DUC 2005 Document Understanding Workshop, Canada.
- Zhang X, Lapata M, Wei F, Ming Z. 2018. Neural Latent Extractive Document Summarization to appear in EMNLP 2018.