General Overview on Text Summarization

Osman Onur KUZUCU
CE Hacettepe University, n21131025@cs.hacettepe.edu.tr

Prof. Dr. Ebru AKÇAPINAR SEZER
CE Hacettepe University, ebru@hacettepe.edu.tr

***Abstract* - In recent years, with the development of technology, many sources of information have emerged. Therefore text summarization has become popular to find correct information. In this paper, we have aimed to focus on text summarization. In our perspective, we have tried to explain why people use text summarization, text summarization types, methods of text summarization.**

Introduction

With the development of the Internet, people are stuck between a lot of information and many documents on the Internet while doing research. Due to this situation, developing a technology that can automatically summarize texts is needed. With this technology, people use this technology to get information that is needed easily.

Text summarization creates summaries with important sentences and includes all important relevant information from the original document. In recent years, different approaches have been built. Some researchers focus on only news summarization. Some researchers focus on specific fields such as biology, physics, computer science etc.

Text summarization is a difficult area. Sometimes people read more than one to understand important parts of the documents. If the reader does not have enough information in a specific domain, the reader can not understand easily. In some cases, the reader must understand multiple documents. This case is very difficult and time consuming.

In text summarization history, research began in the 1950s. Luhn[1] worked on a summarization technique that is based on word and phrase frequency. In these years, this research was very important for this domain. After this research, a lot of research has been done. In recent years with development of the technology, researchers have focused on neural networks, machine learning integration with text summarization.

In general text summarization techniques are seperated 2 parts. One technique is extractive text summarization. The other technique is abstractive text summarization. In extractive text summarization techniques, important information is gathered from original text sentences. After gathering the important information, sentences are selected based on containing important information from the original text . In abstractive text summarization, important information is gathered from the original text. After gathering important information, sentences are generated based on gathered information.

In this paper, our focus point is extractive text summarization. In addition, abstractive text summarization is explained. Some methods are explained with some articles. After the introduction, extractive text summarization is explained in Section 2. In Section 3, abstractive text summarization is explained. In Section 4, the history of text summarization is explained. In section 5, the conclusion of this paper is presented.

Section 2
Extractive text summarization

In this section, extractive text summarization methodology and methods are explained with articles. First term frequency based extractive text summarization methods are explained. You can see the Figure 1.

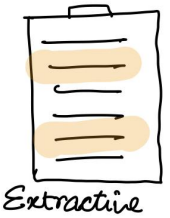

Figure 1:Extractive Text Summarization Figure

1- Term Frequency Method

Tf(term frequency), Idf(inverse document frequency) and Tf-idf(term frequency-inverse document frequency) are some statistical parameters that are used to calculate whether the word or term is important or not. Words and terms are ordered using their importance that is calculated with statistics. Every sentence in the original document is ordered by an important word/phrase. Therefore it is a problem sometimes. For example, very long sentences that contain more than one important word/term, the sentence is ordered of

high importance. However in summary,very long sentences do not want to be selected

In “Word Sequence Models for Single Text Summarization” research paper, researchers used some preprocessing techniques that are eliminating stop-words, stemming. Researchers used different term selection, term weighting methods. Then they used Garcia, an unsupervised algorithm to understand the similarity between sentences. After that, they used the k-means algorithm to cluster sentences.You can see detailed information about this journal in Appendix 1.

In “An Approach to Summarizing Bengali News Documents”, researcher studied news document şn Bengali language. Researcher used tf-idf, sentence position value and sentence length value. Researcher compared the proposed method with 3 methods. One of these methods is called Lead that summary text with the first 100 words. The second method used tf-idf and sentence position value. The last method used tf-idf and sentence length value. Proposed method from the researcher performs better than other methods. You can see detailed information in Appendix 1.

In “MWI-Sum: A Multilingual Summarizer Based on Frequent Weighted Itemsets” researchers wanted to solve multi-document and multilingual text summarization problems. Researchers wanted to build a method that is used for multi-document and multilingual text summarization. reachers used preprocessing(stemming, removing stop-word). The preprocessing process was language based. Researcher used Natural Language Toolkit for the preprocessing process. Researchers used tf-df(term frequency-document frequency) to weight items. In every document, researchers filtered k-top sentences based on tf-df. Researchers mined frequent weighted items from documents. Then, researchers selected sentences that are most relevant to generate a summary. Researchers tested this method with different languages and different datasets. In general test results, this method performs very well. You can see detailed information in Appendix 1.

2-Cluster Based Method

In documents, different ideas and different sections are included. Therefore, different ideas and sections are separated to generate a summary. Different ideas and sections are assigned to different clusters. In every cluster, k sentences are combined to generate a summary. You can see Figure 2.

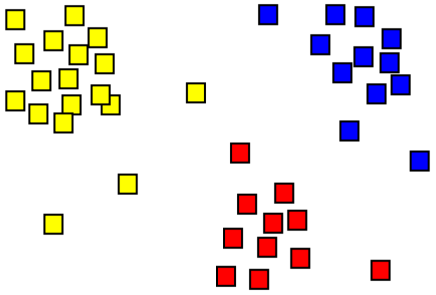

Figure 2: Cluster Figure Example

In “Automatic text summarization based on sentences clustering and extraction”, researchers wanted to solve information overload problems. Researchers wanted to separate sentences into clusters using the similarity value of each sentence. Researchers used word form similarity, word order similarity and word semantic similarity to calculate similarity value for every two sentence pairs in source documents. Researchers calculated cluster numbers using term number in document, term number in sentence and number of the sentences in document. Then, sentences are clustered by the k-means algorithm. Proposed method selected to generate summary from clusters. Proposed method performs better than the MMr and WAA method on the DUC2003 dataset. You can see detailed information in Appendix 1.

In “A Novel Approach for Research Paper Abstract Summarization Using Cluster Based Sentence Extraction”, Researchers want to summarize research paper abstracts. Researchers extracted terms that contain at least one seed and occur more than 3 times. Researchers used fixed length summary numbers to calculate the number of clusters. Fixed length summary number is taken from the user that generated the summary. Then, the k-means algorithm clusters the sentences. Researchers used 2 search strategies. One of the strategies is local search. In the local search strategy there are three methods. First method centroid sentences were selected using minimum distance between vectors. Second method tf-idf is calculated for abstract. Last method uses term frequency in the abstract. The other search strategy is the global search strategy. In this strategy, term frequency and term length are used. Between these methods, Tf-idf ranking, term frequency ranking and global search produced similar results. However, the centroid sentence method has lower performance.You can see detailed information in Appendix 1.

3- Neural Network Based Method

Neural networks are an important part of computer science. Neural networks are modeled on the human brain. Neural networks are trained from scratch. After training, neural networks act like an expert on the

domain. Therefore, neural networks can be used in text summarization. In this part, some journals are explained that contain neural networks. You can see Figure 3

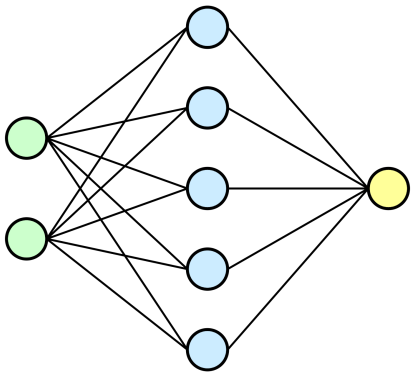

Figure 3: Neural Network Figure Example

In "Automatic Text Summarization with Neural Networks", researchers wanted to build a neural network model that decides which sentence should be in news' summary. Researchers wanted to solve a problem that when users want to get information, they are faced with a lot of relevant documents. Users cannot get the information easily. Therefore, researchers want to solve this problem.
In this journal, researchers extracted features from every sentence. The features are paragraph title, paragraph location, sentence location, first sentence, sentence length, number of thematic words and number of the title words. After extraction, the neural network is trained with a dataset. Then, some features are eliminated that are not very important. Three neural networks are built by researchers. First neural network(N1) has seven input neurons, 12 hidden neurons and one output neuron. Second neural network(N2) has twenty three input neurons, thirty five hidden neurons and one output neuron. Third neural network(N3) has fifty four input neurons, seventy hidden neurons and one output neuron. The difference between neural networks is based on the features vector of the sentences. N3 performs better than N2, N2 performs better than N1. You can see detailed information in Appendix 1.

In "Extractive Summarization using Continuous Vector Space Model" , Researchers built feed-forward neural networks for text summarization. In the first case(CW_RAE_{COS}), researchers vectorize every sentence in the document using Collobert & Weston. Recursive Auto-encoder is used for word embeddings. Cosine similarity is used. In the second case(CW_RAE_{EUC}), researchers vectorize every sentence in the document using Collobert & Weston. Recursive Auto-encoder is used for word embeddings. Euclidean distance is used for similarity measurement. In the third case(CW_Add_{COS}), researchers vectorize every sentence in the document using Collobert & Weston. Vector addition is used for word embeddings. Cosine similarity is used. In the fourth case(CW_Add_{EUC}), researchers vectorize every sentence in the document using Collobert & Weston. Vector addition is used for word embeddings. Euclidean distance is used for similarity measurement. In the fifth case($W2V\_Add_{COS}$), researchers vectorize every sentence in the document using Word2Vec. Vector addition is used for word embeddings. Cosine similarity is used. In the sixth case($W2V\_Add_{EUC}$), researchers vectorize every sentence in the document using Word2Vec. Vector addition is used for word embeddings. Euclidean distance is used for similarity measurement. In the recall result, CW_Add_{COS} got a better result. In the precision and f-measure scores, CW_Add_{EUC} has gotten better results. Collobert & Weston vectorization is better than Word2Vec.

In "A Novel Evolutionary Connectionist Text Summarizer", researchers explain the problems of the previous systems. Problems include difficulty in selecting architecture, excessive training time required, lack of knowledge facilities. Researchers have built system architecture which is displayed in Figure 4.
Researchers extract seven features. Features are parahpagh title, paragraph location, sentence location, first sentence of the paragraph, sentence length, number of thematic words, number of the title words. After extraction of the features from every sentence, a recurrent neural network is trained. Recurrent neural network is tested with 50 different articles from the internet summarized by human.The overall accuracy is %94.

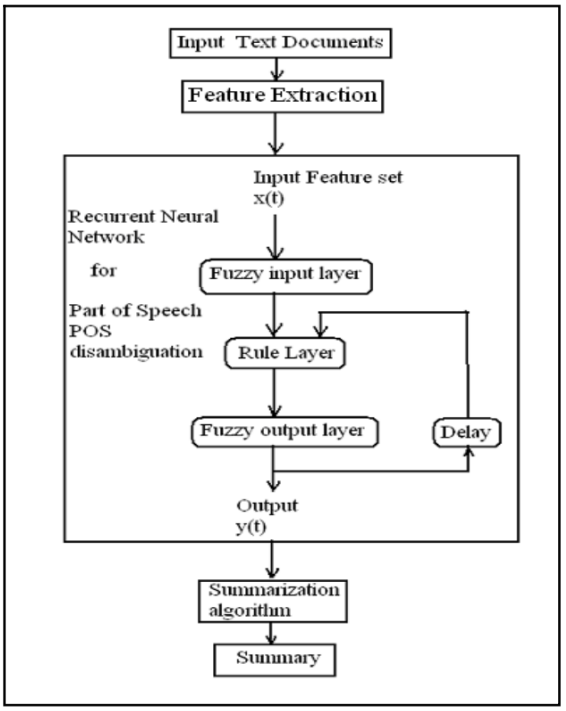

Figure 4: System architecture

4- Fuzzy Logic Method

Fuzzy logic is an interesting approach that provides more than one true or false statement. In other words,

fuzzy logic does not have specific true or false. In boolean logic, there are one true and one false statement. You can understand this by looking at Figure 2 clearly. In Figure 5 2 logics are displayed. Boolean logic has 2 statements that are true and false. In Figure 2, fuzzy logic has 3 statements. The statements weights are different from each other and boolean logis's statements. In fuzzy logic, "if-then" rule is used. With this rule, we can extract features from sentences.

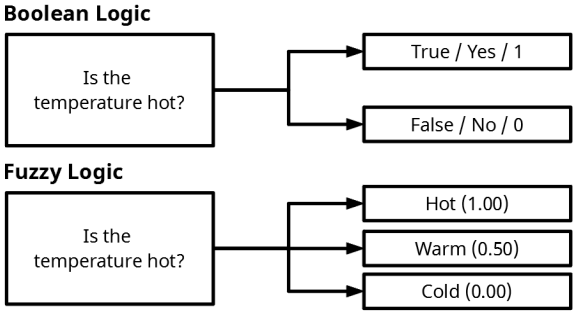

Figure 5: Fuzzy logic and boolean logic comparison[10]

In "Improving Performance of Text Summarization", Researchers wanted to increase performance. Researchers used two methods in this paper. The first method is the fuzzy logic method. Feature extraction is maden by researchers using fuzzy logic. Features are title word, sentence length, sentence position, numerical data, thematic words, sentence similarity, term weight and proper noun. In fuzzy logic methodology, features are weighted differently.
The second method is latent semantic analysis. Latent semantic analysis compares word usage with semantic similarity. This method is from the statistics domain. Firstly, the input matrix is created from the document. Secondly, sentences extracted from the features matrix are calculated from the input matrix. Lastly, the sentence extracted features matrix is ordered .
Two methods are combined. After this combination, this algorithm increases the text summarization performance. This algorithm is compared with only fuzzy logic methods. Between two algorithms, there is %5 percent performance increment. You can see detailed information in Appendix 1.

5- Graph Based Method

Graph is a data structure used in computer science that has vertices and edges. Two or more vertices are connected with edges. Edges display relationships between vertices. Therefore, usage of graphs can increase the text summarizer performance. Graph is built from document. Every sentence is a vertex. Edge between nodes displays relationships. edge weighting is an important parameter that displays important relations. You can see Figure 6.

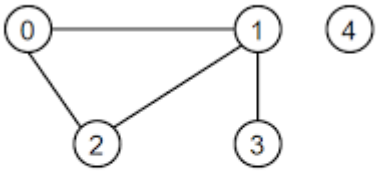

Figure 6: Graph Figure Example

In "Graph-based Ranking Algorithms for Sentence Extraction, Applied to Text Summarization", researcher investigates graph based text summarization algorithms. In this paper, HITS, Positional Power Function, PageRank, undirected graphs and weighted graphs are explained. Researcher uses the TextRank algorithm to generate graphs from text. Every vertex is a sentence. Edge is the relation between two or more sentences based on similarity. researcher used the DUC 2002 dataset. After generating graphs from the DUC 2002 dataset, researcher compares the summarization methods that are $HITS_A$, $HITS_H$, POS_P,POS_W and PageRank with different graphs that are undirected, directed forward and directed backward. $HITS_A$ and PageRank perform best by using directed backward graphs. If overall performance is important, undirected graphs perform better than others in overall performance.

6-Latent Semantic Analysis Based Method

Latent semantic analysis(LSA) is a statistical analysis algorithm that is used to understand word usage in sentences and which word is used commonly in the text. LSA makes a term-sentence matrix from a document. Then, Singular Value Decomposition is built. Singular Value Decomposition is used to understand the relationship between sentences and words.

In "Using Latent Semantic Analysis in Text Summarization and Summary Evaluation", researchers wanted to build a new LSA based approach to text summarization. Proposed model builds on a content based method. That method is to calculate the similarities between the full text and the summary. In similarity calculation, cosine similarity, main topic similarity and term significance similarity are used. Proposed method uses Singular Value Decomposition of term-sentences matrix. Using this matrix, the proposed model finds the important term in the document. Proposed model is compared with the LSA summarizer from Gong and Lui, a random summarizer based on ten extracts, positional heuristic, mutual reinforcement heuristic and tf-idf based summarizer. Between all these methods, the proposed model performs better than other summarizers on three different similarity methods.

7-Machine Learning Based Method

Model is trained with a dataset to classify whether that sentence should be in summary or not. In the model, Naive-bayes, Decision Trees, Hidden Markov Model, and Log-linear algorithms can be used.

In “Using Machine Learning for Medical Document Summarization”, researchers wanted to build a text summarizer on the medical domain. Firstly, researchers have maden a preprocessing of the document. Bagging method is used. Bagging method is a multiple decision tree that improves the performance of the summarization compared with the decision tree. Then, researchers built a corpun focus on the medical domain. Features are extracted from documents. Features are centroid value, similarity to first sentence, sentence position, medical phrases score, cue phrase position, acronym score and sentence length. After training with extracted sentence features, the sentence ranking algorithm has run. Sentences are marked “summary worthy”, “moderately summary worthy” and “summary unworthy”. Sentence selection process takes the number of the sentence(“summary worthy”) to make a summary. If there are not enough sentences to generate a summary, sentences are selected from “moderately summary worthy” until there are enough sentences to generate a summary. Before selection sentences, idf modified cosine similarity is calculated for every sentence. Proposed method performs better than MEAD and Baseline-lead.

8-Query Based Summarization

Users can limit the summary topics. In the query-based summarization, users can determine the limit of the summary with the query, model summary and the document based on the query. For example, users want to summarize text based on tf-idf and terms’ frequency must be more than 5 times. Generally query-based text summarization performs very well and gives accurate results.

In “A System for Query-Specific Document Summarization”, researchers built a structure-based query-based summarization algorithm. In this algorithm, researchers have focused on keyboard queries that are the most effective methods to find information. Firstly, algorithms apply preprocessing. Then, documents are parsed into text fragments using delimiters. Nodes contain text fragments. Edges are built between nodes that have a relationship. Based on the relationship, edges are weighted.There are many relationship models between nodes that contain text fragments. Then, a document graph is built using the document. Document graphs are labeled, weighted. Algorithm gets a set of keywords. It generates a summary that contains all keywords with minimum loss. Proposed approach performs better than Multi-result Enumeration Algorithm, Top-1 Enumeration Algorithm, Multi-Result Expanding Search Algorithm, Top-1 Expanding Search Algorithm.

SECTION 3
ABSTRACTIVE TEXT SUMMARIZATION

Abstractive text summarization generates a summary from a document. In summary, all sentences are constructed from scratch. In this methodology, all sentences are generated uniquely. All sentences must be different from the source document.

Figure 7: Abstractive Text Summarization Figure Example

1-Structure Based Method

In this method, most important information is gathered using phrases, verbs, nouns and structure.

1.1 Tree Based Method

Only one dependency tree is built from source text. Text documents are displayed using the dependency tree easily. The parser and dependency between words affect the performance of the summarization.

In “Dependency Tree Based Sentence Compression”, Researchers have explained detailed information about the proposed method. Researchers aimed to generate a summary using a tree-based model and sentence compression. In this journal, researchers proposed a method that is a novel unsupervised dependency-based approach. This approach contains three sentence comparison steps. Tree transformation, tree compression and tree linearization are the steps. In tree transformation, a tree is generated from a sentence. You can see in Figure 8. In tree compression, a dependency is removed on the tree or not.This part is the optimization part of the tree. In tree linearization, a tree linearizes to study. Researchers used an English and a German corpus. Proposed method performs well when Stanford parser is used to separate the corpus. When the proposed method runs with RASP, that is another parser performs well but not as well as with the Stanford parser. In conclusion, the proposed method(dependency-based) performs better than GOLD and Clarke&Lapata’s method. In the German corpus, the proposed method did not achieve a better result than Clarke&Lapata’s method

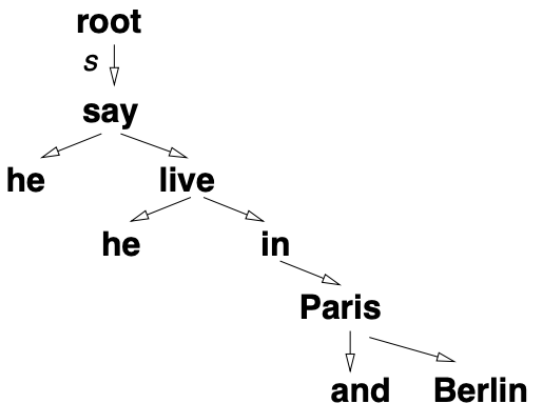

Figure 8: "He said that he lived in Paris and Berlin" transform a tree.

1.2 Template Based Method

This summarization technique uses templates. Using templates to summarize required highly detailed semantic information and relevant contents. In this technique, pattern is very important.

In "Generating Single and Multi-Document Summaries with GISTexter", Researcher built a system that is named GISTexter. GISTexter has 2 summarization models. The first model focuses on a single document summarization based template. You can see an example template in Figure 9. In the first model, firstly sentence extraction is applied using single-document decomposition. Secondly, unnecessary information is filtered out. Lastly, summary reduction is applied. The second model is a multi-document summarizer. The multi-document summarizer model processes the topic of the document. After that, the information extraction system runs to summarize. With information extraction, all information is gathered from multi-documents. GISTexter uses a template to summarize. GISTexter performs best on single document summarization in the DUC 2002 dataset. GISTexter performs second best on multi-document summarization in the DUC 2002 dataset.

```
TEMPLATE
Doc_NR:    CNN19980301.1000.0329
Event:     <Natural_Disaster–CNN19980301.1000.0329–1>
Comment:   Prototypical
<Natural_Disaster–CNN19980301.1000.0329–1> :=
      Disaster:            last week's TORNADOES
      Amount Damage:   $100 million
      Number Dead:       40
                        / four of the victims
                        / a husband, wife, their daughter and her fiancee
      Location:          Florida
                       / central Florida
      Date:              last week
```

Figure 9: Template Example

1.3 Ontology Based Method

With domain specific information, ontology based text summarization performs well. The main performance bottleneck is domain specific information in ontology.

In "Ontology Based Text Document Summarization System Using Concept Terms", Researchers wanted to build a text summarizer based on ontology. Proposed method's structure is mentioned in Figure 10. In the preprocessing state, stop words are filtered out. Stemming is applied. In the clustering state, concept extraction is applied using tf, idf or tf-idf. The K-means algorithm is applied after this state. Then, ontology is generated from the text. In the summarization state, concept matching, concept depth and concept count are applied. The ontology-based summarization methods prove the accuracy of the summary.

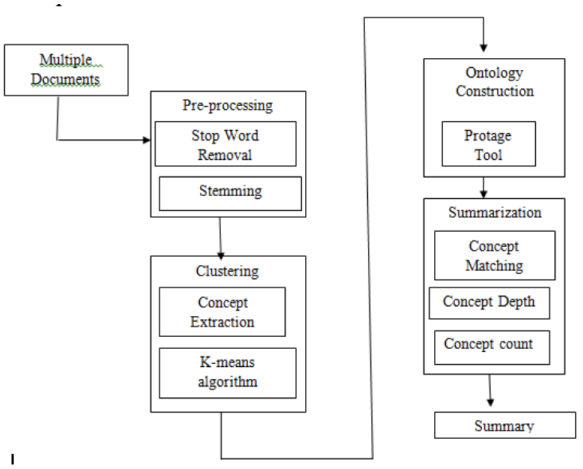

Figure 10: Proposed Method's Structure

1.4 Lead and Body Phrase Method

This model is built on phrases. Usage of the phrases affects the performance of the summarization. In the document, sentences must be informative, long enough and able to be rephrased.

In "Hybrid Text Summarization Method based on the TF Method and the LEAD Method", Researchers wanted to build a system that uses phrases. The proposed method is a combination of the TF-based summarization and LEAD-based summarization. The proposed method uses different parameters from TF, Head-TF, Hyb-LEAD and LEAD methods.The proposed method performs best in these methods on the NTCIR-2 TSC dataset.

1.5 Rule Based Method

Generating summary sentences uses the design pattern of the documents. This method generates a high density summary. This method's rule is designed by users.

In "Fully Abstractive Approach to Guided Summarization", researchers wanted to build rule-based text summarization methods. Guided summarization has 5 main categories. In every category, specific questions must be answered. Proposed method contains information extraction methodology. In information extraction, algorithms must follow some rules. These rules are changed for every category. Content selection is a progressive process after information extraction. Then, the SimpleNLG realizer is used to generate a summary. Proposed method generates a high density-summary. However, the proposed method does not increase summary's coverage.

2-Semantic Based Method

These methods focus on linguistic information. Nouns, verbs and phrases are identified by using these methods.

2.1 Multimodal Semantic Method

In this method, summary score is calculated from concepts that are in the documents

In "Towards a Framework for Abstractive Summarization of Multimodal Documents", a researcher wanted to build a framework that generates summaries from unified semantic models. The proposed model has steps. The first step is building the semantic model using Sparser and SIGHT. You can see an example semantic model in Figure 11. Boxes represent an individual concept in the document. Line connection between boxes represents the relationship between boxes. The second step is content rating. In this step, information density is used to rate content. In the last step, contents are selected with ratings to generate a summary. Proposed method can produce summaries from multimodal documents by covering all the documents. In the journals, the proposed method's result is not displayed.

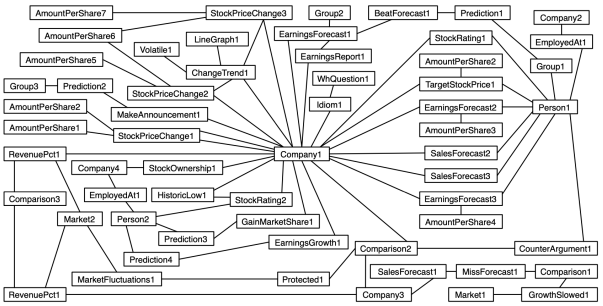

Figure 11: Example semantic model

2.2 Information Item Based Method

In this method, summary is less redundant than other methods. The item can be the smallest information word in the document.

In "Framework for Abstractive Summarization using Text-to-Text Generation", researchers wanted to build a framework that is used to generate summaries from abstract representations of the source documents. In the framework, "INIT" is used to get the smallest element of the information from text or sentences. After using "INIT", getting elements of the information, elements are selected using frequency-based or query-driven or guided summarization methods. After selecting the elements from the documents, SimpleNLG is used to generate a summary of the proposed model to achieve a satisfactory result on the TAC 2010 dataset. .

2.3 Semantic Graph Based Method

In this method, documents are represented with graphs. Nodes contain nouns and verbs, edges are the product of the relationship between nodes. This method generates grammatically correct summaries.

In "Semantic Graph Reduction Approach for Abstractive Text Summarization", researchers wanted to build a method to summarize text using semantic graph reduction. In the first phase of the proposed method is rich semantic graph creation. In this phase, semantic graphs are produced from text. You can see the first phase structure in Figure 12. After the first phase, semantic graphs are reduced by WordNet. The third phase is generating the summary. In this phase, some rules are considered like a fuzzy logic if-then technique. Proposed method achieves a fifty percent reduction of the original text.

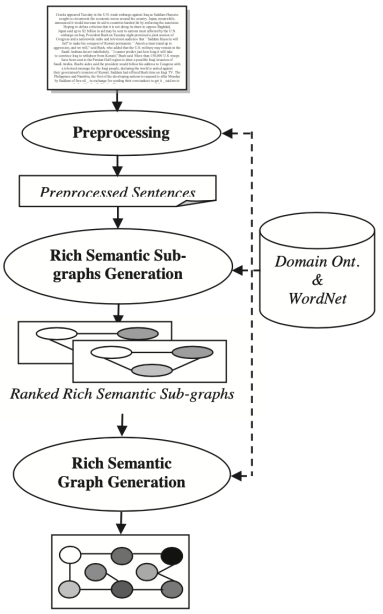

Figure 12: First Phase Structure

Section 4
history of text summarization

The first publication was published by Luhn. Luhn focused on statistical methods. Luhn worked on extractive text summarization. After increasing usage of term frequency-inverse document frequency, research numbers have increased rapidly. While developing the tf-idf method, pattern based text summarization is developed. After patter-based text summarization became popular, some researchers combined it with Latent Semantic Analysis. Early 2019, neural network technology was developing rapidly, researchers have studied neural network and machine learning on text summarization.

Section 4
Conclusion

In this paper, we want to explain text summarization with detailed information. Extractive text summarization and its methods are explained clearly using example journals. Also, abstractive text summarization and its methods are introduced detailly using example journals. History of the text summarization is discoursed.

References

[1] https://en.wikipedia.org/wiki/Automatic_summarization#History

[2] https://www.researchgate.net/publication/224384763_Word_Sequence_Models_for_Single_Text_Summarization

[3] https://www.researchgate.net/publication/254462618_An_approach_to_summarizing_Bengali_news_documents

[4] https://dl.acm.org/doi/10.1145/2809786#sec-ref

[5] https://www.researchgate.net/publication/224588010_Automatic_text_summarization_based_on_sentences_clustering_and_extraction

[6] https://www.researchgate.net/publication/220902555_A_novel_approach_for_research_paper_abstracts_summarization_using_cluster_based_sentence_extraction

[7] https://www.researchgate.net/publication/4096321_Automatic_text_summarization_with_neural_networks

[8] https://www.researchgate.net/publication/280730254_Extractive_Summarization_using_Continuous_Vector_Space_Models

[9] https://arxiv.org/pdf/1912.03035.pdf

[10] https://medium.com/@siosond/introduction-to-fuzzy-logic-3664c610d98c

[11] https://medium.com/@prasasthy.sanal/brief-history-of-text-summarization-9d1b3787a707#:~:text=It%20was%20during%201980s%20and,Automatic%20summarization%20also%20gained%20interest.

[12]https://www.researchgate.net/publication/224599254_A_novel_evolutionary_connectionist_text_summarizer_ECTS

[13]https://www.researchgate.net/publication/275366875_Improving_Performance_of_Text_Summarization

[14]https://www.researchgate.net/publication/228996372_Graph-based_Ranking_Algorithms_for_Sentence_Extraction_Applied_to_Text_Summarization

[15]https://www.researchgate.net/publication/220017752_Using_Latent_Semantic_Analysis_in_Text_Summarization_and_Summary_Evaluation

[16]https://www.earticle.net/Article/A147691

[17]https://dl.acm.org/doi/10.1145/1183614.1183703

[18]https://www.semanticscholar.org/paper/Dependency-Tree-Based-Sentence-Compression-Filippova-Strube/6d42e405d9b7270079eb33e228141de9d57bded7

[19]https://www.researchgate.net/publication/246302615_Generating_single_and_multi-document_summaries_with_GISTEXTER

[20]http://researchgate.net/publication/283227649_Ontology_based_text_document_summarization_system_using_concept_terms

[21]https://www.researchgate.net/publication/2373730_Hybrid_Text_Summarization_Method_based_on_the_TF_Method_and_the_LEAD_Method

[22]https://www.researchgate.net/publication/262237231_Fully_abstractive_approach_to_guided_summarization

[23]https://www.researchgate.net/publication/220873703_Towards_a_Framework_for_Abstractive_Summarization_of_Multimodal_Documents

[24]https://www.researchgate.net/publication/262282952_Framework_for_abstractive_summarization_using_text-to-text_generation

[25]https://www.researchgate.net/publication/235326941_Semantic_graph_reduction_approach_for_abstractive_Text_Summarization

Appendix 1

ID	Journal Name	Full Reference	Publication Year	Total Citation	Investigated Problem	Proposed solution	Experimental Setup	Achieved Result	Performance Evaluator
1	Word Sequence Models for Single Text Summarization	28 references in this link	2009	45	-Gathered information that is needed easily from different sources -Detection of the most relevant information in source document	-Term selection (n-gram) -Term weighting(Bool,TF,IDF, TF/IDF) -Sentence Selection (Garcia, k-means)	on DUC 2002 dataset comparison between different n-gram, term weighting, and MFS methods	- best result: IDF, 3-gram -second best result: Bool, MFS -third best result: Bool,1-gram	Recall, precision and f-measure are compared between different approaches.
2	An Approach to Summarizing Bengali News Documents	23 references in this link	2012	28	-Reducing information overload problem on the internet.	-Preprocessing(removing stop-words, stemming) -Ranking Sentences (TF-IDF, sentence positional value, sentence length) -Sentence selection	-Randomly selected 10 document-summary pairs from research corpus.	-Proposed method is better than Lead method, only tf-idf and position used method and only tf-idf and length parameter used method.	Recall, precision and f-measure are compared between different methods.
3	MWI-Sum: A Multilingual Summarizer Based on Frequent Weighted Itemsets	69 references in this link	2015	35	-low performance in multi-document and multilingual text summarization. -reduce information overload.	-Preprocessing(removing stop-words, stemming) -Ranking Sentences (TF-DF) -Sentence selection (k-top)	-TAC 2011 Multilingual Benchmark collections(different languages) -DUC 2004 English-written Benchmark Collections -News Collections with different cases from Google	-Proposed method is better than CIST, CLASSY, ItemSum, JRC, AMTS, LIF, OTS, ICSIsumm,	Recall, precision, f-measure and Avg. Similarity are compared between different methods.

							News	UBSummarizer	
4	Automatic text summarization based on sentences clustering and extraction	20 references in this link	2009	68	-Information overload problem	-Preprocessing(removing stop-words) -Similarity calculation between every two sentences -Separate sentences to clustors(k-means) -Sentence selection from clustors	on DUC 2003 Dataset	Proposed method performs better than compared with MMR and WAA	Recall, precision and f-measure are compared between MMR and WAA methods.
5	A Novel Approach for Research Paper Abstract Summarization Using Cluster Based Sentence Extraction	8 references in this link	2011	5	-Users want to get information from the internet or other sources. But users are faced with a lot of documents. Researchers want to solve this problem with summarizing multi documents' abstract.	-Term extraction(at least one seed, occurs at least 3 times in the abstract.) -Sentence Clustering(fixed length summary) -Sentence selection(local and global search strategies)	IEEE International Conferences on Data Mining dataset.	Tf-idf, tf and global search strategies perform similarly. But, centroid sentences perform worse than other methods.	Recall, precision and f-measure are compared between local and global search strategies.
6	Automatic Text Summarization with Neural Networks	17 references in this link	2004	71	-When users want to get information, they are faced with a lot of relevant documents. Users cannot get the information easily. Therefore, researchers want to solve this problem.	-Feature extraction(paragraph title, paragraph location, sentence location, first sentence, sentence length, number of thematic words and number of the title words) -build 3 different network	-Dataset is generated by a human reader.	-N1: %93 accuracy -N2: %96 accuracy -N3: %99 accuracy	Accuracy results of the neural networks are compared.

						-train/test network			
8	Extractive Summarization using Continuous Vector Space Model	34 references in this link	2014	134	-Reducing information overload problem on the internet.	-Built feed-forwards neural networks -Using different vectorizers(Collobert & Weston and Word2Vec) -Using different word embeddings(Recursive Auto-encoder and Vector addition) -Using different similarity measurement(Cosine similarity and Euclidean distance similarity)	-The Opinosis dataset is used. Dataset contains 51 topics. Each topic has at least 50 sentences. In some topics, they have 571 sentences.	CW_Add_{COS} and CW_Add_{EUC} have gotten much better results than the original Lin-Bilmes method which is published with the dataset.	Recall, precision and f-measure are compared between different methods and the original method which is published with the dataset.
9	A Novel Evolutionary Connectionist Text Summarizer	28 references in this link	2009	5	-Reducing information overload problem on the internet. -increase previous text summarization techniques	-Build a recurrent neural network structure. -Extracting features from all sentences in the documents -Train recurrent neural network with dataset.	-Dataset is generated by a human reader.	-This structure with connectionist model is performed very well. This algorithm is very powerful for text summarization.	Accuracy results of the recurrent neural networks are interpreted.
10	Improving Performance of Text Summarization	21 references in this link	2015	81	-increase text summarization performance	-Extract 8 features using fuzzy logic algorithm -LSA is used to build sentence feature matrix -Combine these two methods to find the	-Ten datasets are used for performance comparison.	Proposed method performs better than using only a fuzzy logic methodology in text summarizer.	Recall, precision and f-measure are compared between two algorithms on ten datasets.

						important sentences to make summary			
11	Graph-based Ranking Algorithms for Sentence Extraction, Applied to Text Summarization	11 references in this link	2004	332	-compare graph based text summarization method and enhance the performance	-Using TextRank to generate undirected, directed forward and directed backward graphs from text. - Run $HITS_A$, $HITS_H$, POS_P,POS_W and PageRank algorithms to compare the summarization.	-DUC 2002 dataset is used to run comparisons.	-$HITS_A$ and PageRank perform best by using directed backward graphs. -Undirected graph perform better than other in overall performance.	-Accuracy results are compared..
12	Using Latent Semantic Analysis in Text Summarization and Summary Evaluation	9 references in this link	2004	134	-increase the performance of the text summarizers.	-proposed method is built on similarity calculation between full text and its summary. -build Single value decomposition matrix to understand important term in text.	-Reuters collections are used for testing proposed model.	-The proposed method performs better than the LSA summarizer from Gong and Lui, a random summarizer based on ten extracts, positional heuristic, mutual reinforcement heuristic and tf-idf based summarizer.	-Accuracy results are compared.
13	Using Machine Learning for Medical Document Summarization	101 references in this link	2006	46	-Information overload problem -Improve performance of text summarizers.	-Preprocessing -Bagging(multi decision tree) -Feature extractions -idf based cosine	-total document number is 75. -50 documents for training, 25 documents for testing	The proposed method performs better than MEAD and	Precision and recall results are compared.

						similarity.		baseline-lead.	

14	A System for Query-Specific Document Summarization	47 references in this link	2006	79	-improve performance of the query-based text summarization technique.	-Preprocessing -Document graph is built from a document using some features. -Sentence selection	-DUC 2005 dataset is used.	-Proposed method performs better than Multi-result Enumeration Algorithm, Top-1 Enumeration Algorithm, Multi-Result Expanding Search Algorithm and Top-1 Expanding Search Algorithm.	-Accuracy results are compared.
15	Dependency Tree Based Sentence Compression	21 reference in this link	2008	120	-Information overload problem -Improve text summarizer performance	-Tree transformation -Tree compression -Tree linearization -Tree generated from text	-English Compression Corpus and TüBa-D/Z are used.	-The proposed method(dependency-based) performs better than GOLD and Clarke&Lapata's method.	-F-measure and compression rate are compared.
16	Generating Single and Multi-Document Summaries with	12 references in this link	2002	49	-Information overload problem -Improve text summarizer	-Single document summarization -Multi-document summarization	-DUC 2002 dataset is used.	-GISTexter performs best on single document	Precision, recall are compared. s

	GISTexter				performance	-Sentence extraction -Filter out non-important information		summarization in the DUC 2002 dataset. -GISTexter performs second best on multi-document summarization in the DUC 2002 dataset.	
17	Ontology Based Text Document Summarization System Using Concept Terms	4 references in this link	2015	8	-Documents are difficult to understand without reading the entire document entirely.--This is time consuming.	-Preprocessing(stop words, stemming) -Clustering state(tf, idf, tf-idf) -K-means algorithm to cluster -Concept matching -Concept depth -Concept count to generate summary.	-Not stated.	-The proposed methods prove the accuracy of the summary.	Accuracy is compared but not stated in the document.
18	Hybrid Text Summarization Method based on the TF Method and the LEAD Method	0 references in this link	2001	7	-Information overload problem -Improve text summarizer performance	-TF based summarization -LEAD based summarization -Combine these two methods.	-NTCIR-2 TSC dataset is used.	The proposed method performs best.	F-measure is compared.
19	Fully Abstractive Approach to Guided Summarization	20 references in this link	2012	72	-Information overload problem -Improve text summarizer performance	-Guided summarization -Information extraction -Content selection based on rules	-Comparison with summary made by human reader.	-The proposed method does not increase summary's coverage.	-Comparison was made, faced with not good results

						-SimpleNLG			
20	Towards a Framework for Abstractive Summarization of Multimodal Documents	198 references in this link	2011	32	-translation into morphologically rich languages -Information overload problem -Improve text summarizer performance	-Building the semantic model using parsers -Content rating -selecting contents	-190K Chinese-english from LDC2003E14 corpus. - NIST'06 and NIST'08 are used to test the model.	Proposed method can produce summaries from multimodal documents by covering all the documents.	-Results are not displayed.
21	Framework for Abstractive Summarization using Text-to-Text Generation	22 references in this link	2011	130	-Information overload problem -Improve text summarizer performance	-"INIT" is used to get the smallest element of the information from text or sentences. -Getting elements of the information -Element selection using frequency-based or query-driven or guided summarization methods	-Not stated	-Proposed model achieves a satisfactory result.	-Linguistic quality and pyramid score are compared.
22	Semantic Graph Reduction Approach for Abstractive Text Summarization	23 references in this link	2012	97	-Information overload problem -cannot manage the amount of information is available	-Rich semantic graphs are created -semantic graphs are produced from text by WordNet. -Generating summary using fuzzy logic IF-THEN algorithm.	-WordNet is used.	-Proposed method achieves a fifty percent reduction of the original text.	-Results are not stated.