Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.776523

AUTOMATIC INFORMATION EXTRACTION FROM DIFFERENT SOURCES USING USER-DEFINED TEMPLATE

Srisudha Garugu
Research Scholar, Computer Science and System Engineering, Andhra University College of Engineering (A), Andhra University, Visakhapatnam-530003, India

D. Lalitha Bhaskari
Professor, Computer Science and System Engineering, Coordinator, IQAC, Andhra University College of Engineering (A), Andhra University, Visakhapatnam-530003, India

Abstract-Information Extraction deals with the automated extraction of knowledge from unstructured sources. This field unfolded new avenues for querying, organizing, and analyzing information by drawing clean linguistics of structured information and also the abundance of unstructured data. This study aimed to differentiate the assorted styles of informational text structure within the text information. Classification of informational text structure during a given text is a vital space of analysis for locating data within the text content. Several previous studies outlined a collection of classes of informational text structures that identify respective signal words. This paper proposes automatic extraction of text informational structure from various sources with extracting techniques like extractive summarization, abstractive summarization, name entity recognition, event extraction, and question answering which is useful to the reader to know the information exactly. It makes effective for the various steps inclined in information extraction adapting to dynamic data, integrating with existing entities, and handling uncertainty in the existing process.

Keywords-Text structure, Signal word, Extractive Summarization, Abstractive Summarization, Name Entity Recognition, Event Extraction, Question Answering

1.Introduction

Information Extraction parses through unstructured information and extracts key data into additional editable structured information formats. Extracting structure from noisy, unstructured sources is a daunting task that has occupied the true research community for over two decades. With its roots in the natural language processing (NLP) community, today's topic of structure extraction is involved in a variety of communities like machine learning, information retrieval, database, web, and document analysis. Early extraction tasks focused on identifying named entities such as personal and company names and relationships between them from natural language text.

Almost Extracting structured data from different types of documents is an important problem with the potential to automate many real-world business workflows such as purchase Orders, payment, recruitment, payroll, and Incident Reports. Almost different types of documents are used in our daily workflow. Text classification is of great importance, given the high volume

of text available today. M.Chaudhuri, B.B., (2000)Unlimited quantity of text data is around us, a method of organizing this high quantity of information is to classify it into a collection of knowledge kind classes. Belkin,N.J.,et al.,(1993) by classifying the text information users will find the desired information in their search for knowledge and information. The idea is to spot that data go serves the supposed purpose and the way to spot it.
structured information to be extracted from unstructured texts shown in fig.1. Preprocessing of text has to be done than finding and classifying concepts after that connecting the concepts for unifying which gets rid of the noise thereby enriching the knowledge base.

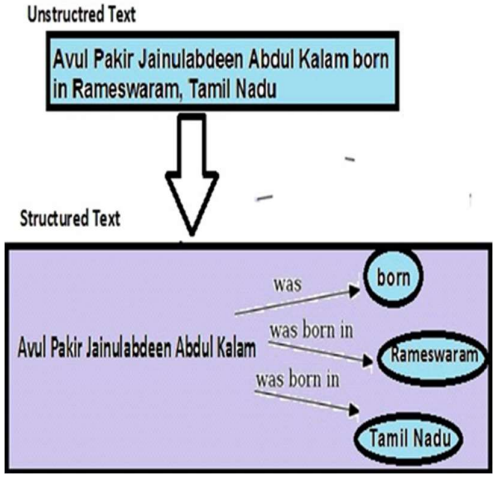

Fig.1. Unstructured to Structured data

Applications

Information extraction can be applied to a wide variety of text sources while categorizing according to the type of application as scientific, personal, enterprise, or Web-oriented.

1. Scientific Applications: For relevant papers suggestion or automated references discovery. The rise in bio-informatics has broadened the scope of earlier extractions from named entities to biological objects such as proteins and genes.

2. Personal Information Management: Personal information management (PIM) systems seek to organize personal data like documents, emails, projects, and people in a structured inter-linked format. The success of such systems will depend on being able to automatically extract structure from existing predominantly file-based unstructured sources.

3. Enterprise Applications: Any customer-oriented enterprise which collects many forms of unstructured data from customer interaction which rises to many interesting extraction problems like identification of product names, product attributes from customer emails, linking of customer emails to a specific transaction in a sales database, extraction of merchant name, addresses from sales invoices, the extraction of repair records from insurance claim forms, the extraction of customer moods from phone conversation transcripts, and the extraction of product attribute value pairs from textual product descriptions. News tracking a classical application of information extraction which raised early research in the NLP community is

tracking specific event types from news sources. Classified ads and other listings such as restaurant lists are another domain with an implicit structure that when exposed can be invaluable for querying. Many researchers have specifically targeted such as record-oriented data in their extraction research.

4. Web Oriented: Opinion database which relates Innumerable websites storing unmoderated opinions about a range of topics, including products, books, movies, people, and music. Many of the opinions are in free text form hidden behind Blogs, newsgroup posts, review sites, and so on. The value of these reviews can be greatly enhanced if organized along structured fields. Many citation databases on the web have been created through elaborate structure extraction steps from sources like Citeseer, Cora, and Google Scholar.

Information comes in many shapes and sizes. One interesting and important form is structured data which contains entities and relationships for example company-location specific data could help companies understand which region they should invest their resources into. Taking the 'Berlin' company as an exemplar: given a specific location, one can ask for the locations where Berlin work. we would like to discover which companies do business in that location.

Information extraction uses several tasks including 1) Lample et al., (2016) Suggested Identifying entities like people, companies, and dates mentioned within a document (i.e., Named Entity Recognition). 2) Liu et al., (2013) Assigning unique identities associated with the entity's in a text (e.g., entity IDs in a knowledge base), and 3) Raghunathan et al., (2010) Finding all aspects refer t the same entity (e.g., named entities, nouns and pronouns). this means that we must identify all entities in the text and pinpoint the relationships between them.4) Mintz et al., (2009) this process of identifying co-reference of different entities require us to detect semantic relationships between entities. We are minimizing the ambiguity in meaning by mostly labeling things that are established as instances of particular meaning and to make sense of semantic relationships like ownership and marriage i.e., Relation Extraction.

A named entity is a specific noun phrase that indicates the type of people such as Organizations, people, data, etc. Ratinov and Roth,(2009) proposed Named Entity Recognition (NER) and Zelenko et al.,(2003) relation extraction (RE) are two important subtasks of Information Extraction (IE).Li and Ji, (2014) and Gupta et al., (2016) have shown the benefits of solving these two tasks together, both in terms of efficiency and accuracy. Chan and Roth et al. (2011) suggested that compared to the pipeline approach, a model that extracts named entities (NEs) and relationships together can capture dependencies between entities and relationships. For example, Miwa and Sasaki (2014) used handcrafted syntactic features such as the shortest path between two words in the syntactic tree to extract syntactic information using encoders from pre-trained syntactic parsers. The authors Miwa and Sasaki (2014) examine decoding search strategies for filling tables based on historical predictions. In addition, they examined six strategies for determining the order for filling table cells. History-based prediction is also an obstacle to parallelizing label decoding.

2. Proposed System

The architecture of a proposed system is given in Fig.2:

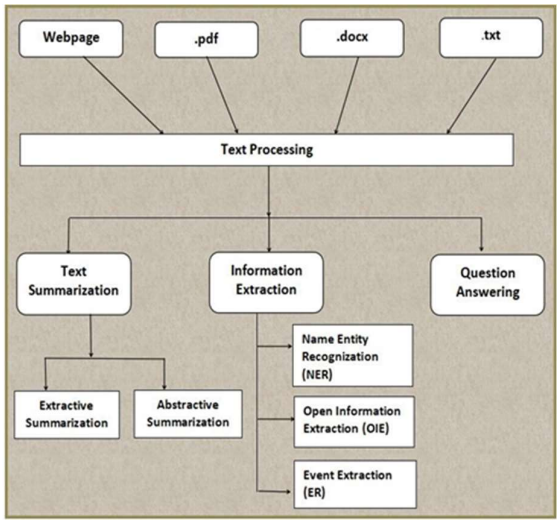

Fig.2. Proposed system architecture

The proposed system can input any kind of text that may be from the webpage, any Text file (.txt), Document (.doc), Portable Document Format (.pdf), or text paragraph shown in Fig.3, and performs automatic Summarization, Named Entity recognition(NER), Relationship Extraction (RE), and Open Information Extraction(OIE), and also provides Question answering where the system automatically answers the questions posed by humans in a natural language.

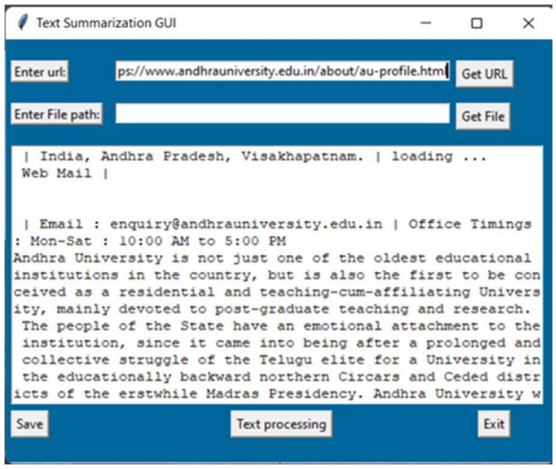

Fig.3. Inputting the Text

- **Automatic Summarization**

Automatic Summarization can be performed in two ways i.e., Extractive and Abstractive. Extractive is a deductive automatic summarization method, which forms a summary by extracting the existing keywords in the document; Abstractive is a generative automatic summarization method, which forms an abstract by establishing abstract semantic representations and using natural language generation techniques [11].

- **Extractive Summarization**

The Extractive summarization model has been built using the LexRank algorithm. LexRank algorithm

[12] is a graph-based Lexical Centrality as Salience in Text Summarization. LexRank is a new approach for computing sentence importance based on the concept of eigenvector centrality in a graph representation of sentences. This approach model works similar to Google's PageRank algorithm to find top-ranked sentences from the document and represents them as a graph. This TF-IDF formulation is then used as a measurement for the similarity between sentences by using it in this idf-modified-cosine formula.

$$\text{idf-modified-cosine}(x,y) = \frac{\sum_{w \in x,y} tf_{w,x} tf_{w,y} (idf_w)^2}{\sqrt{\sum_{x_i \in x} (tf_{x_i,x} idf_{x_i})^2} \times \sqrt{\sum_{y_i \in y} (tf_{y_i,y} idf_{y_i})^2}}$$

The above formula will measure the distance between two sentences x and y. the more similar two sentences, the closer they are to each other. The LexRank algorithm measures the importance of sentences in the graph by considering their relative importance to their neighboring sentences, where a positive value will increase the importance of a sentence's neighbor, while a negative value will decrease the importance value of a sentence's neighbor.

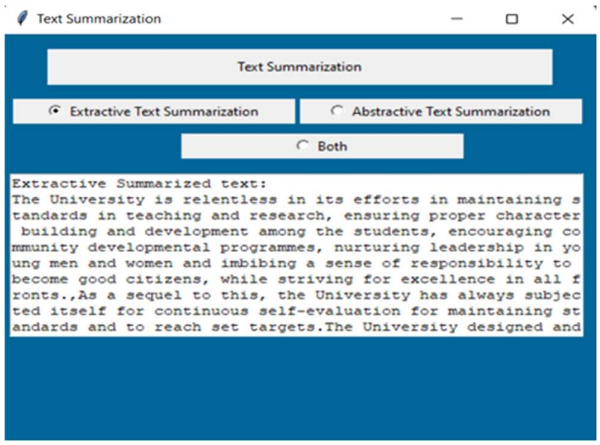

Fig.4. Extractive Text Summarization

- **Abstractive Summarization**

The Abstractive summarization model has been built using T5Ttransformer.T5 (Text-To-Text Transfer Transformer) [13] is a transformer model that is trained in an end-to-end manner with text as input and modified text as output. T5 is an encoder-decoder model [14] and converts all NLP problems into a text-to-text format. It is trained using teacher forcing. This means that for training, the application had taken an input sequence and a corresponding target sequence. These input sequences are given to the model using input_ids. The target sequence is moved to the right, i.e., the sentences are embedded with a start-sequence token and given to the decoder using the decoder_input_ids. In teacher-forcing style, corresponding labels of the target sequence are appended by the EOS token and the PAD token is used as the start-sequence token. T5 can perform training / fine-tuning both in a supervised and unsupervised fashion.

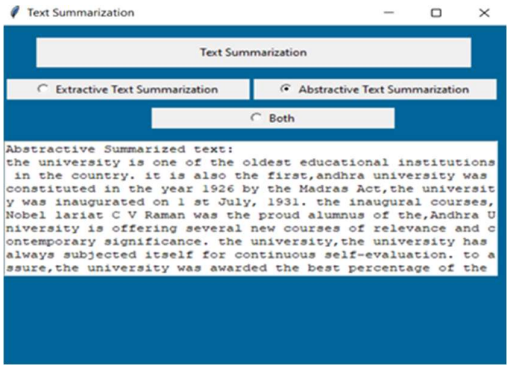

Fig.5. Abstractive Text Summarization

Information Extraction

In natural language processing, open information extraction (OIE)[15] is the task of generating a structured, machine-readable representation of the information in text, usually in the form of triples or n-ary propositions. Sentences are the combination of words that belong to eight different Parts of Speech (POS) in the English language: noun, pronoun, verb, adjective, adverb, preposition, conjunction, and intersection. The POS specifies the functions of the word in the context of meaning in a given sentence. The Part Of Speech tagging is the process of tagging each word with the part of speech it belongs to and that would carry a lot of significance when it comes to understanding the meaning of a sentence. And it can be used for the extraction of meaningful information from text [16].

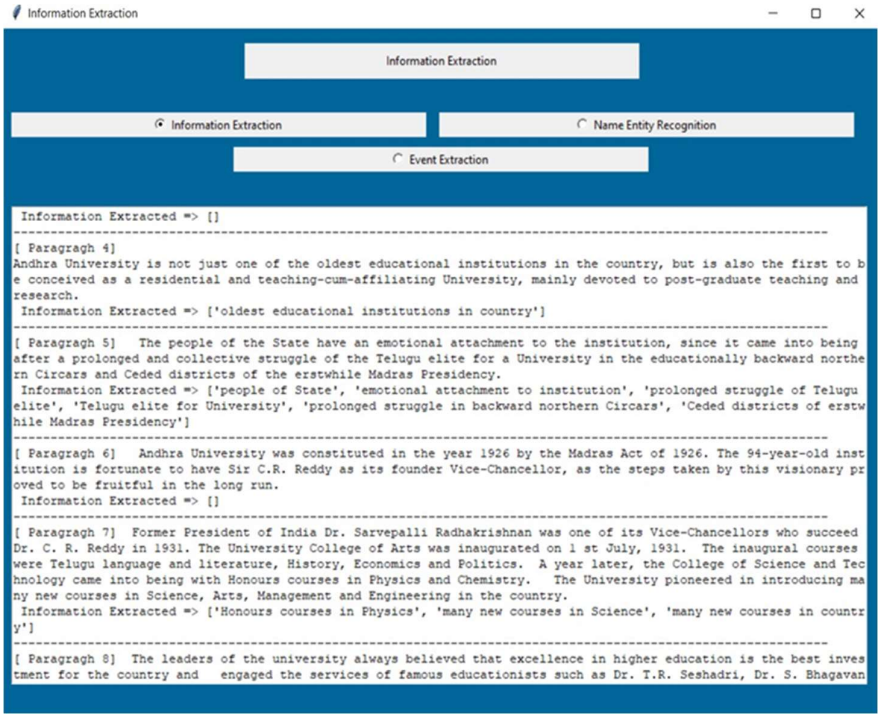

Fig. 6. Information Extraction

Name Entity Recognition

A named entity is an object from the real world that is assigned by a name – for example, a person, a country, a product, or a book title. This approach can recognize various types of named entities in a document, by asking the system to predict. This application provides a good accurate statistical entity recognition system, that assigns labels to contiguous spans of tokens. The trained pipelines can be able to identify a variety of named and numeric entities, including companies, locations, organizations, and products.

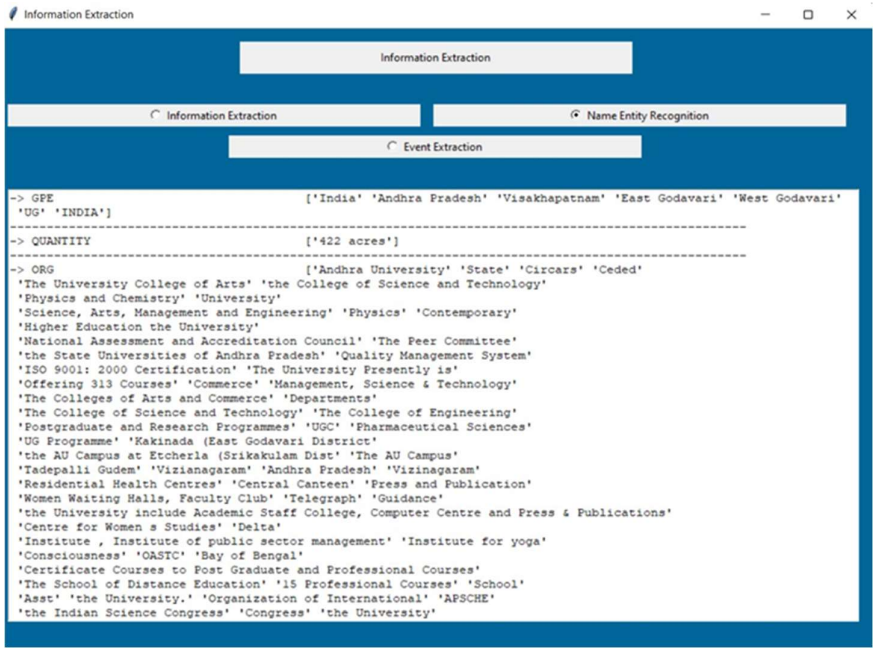

Fig.7.Name Entity Recognition

Event Extraction

Event Extraction is the process of gathering knowledge about periodical incidents found in texts, automatically identifying information about what happened and when it happened. The following approach will perform pre-training of words and build an embedding model which considers a sentence's vector as the average between its token vectors. each sentence will have a respective 300th-dimensional array.

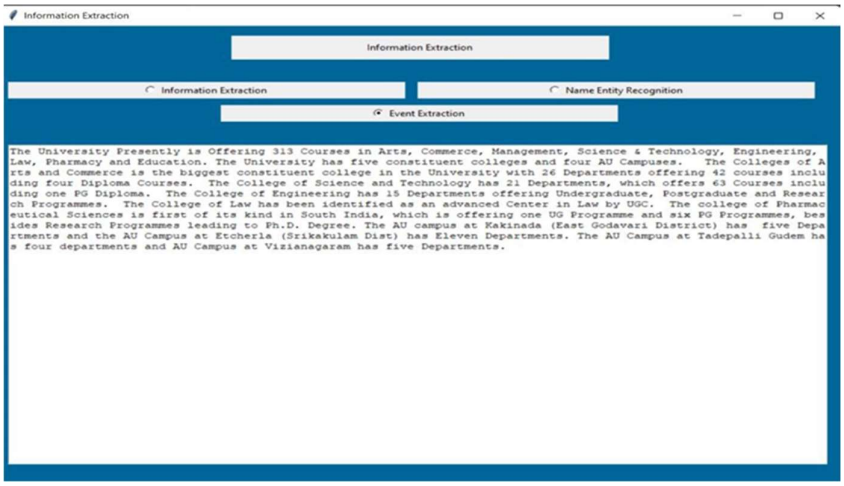

Fig.8. Event Extraction

There are many sentences in the document and it should pick the one that best represents the event itself. To achieve that it will divide the vectors into a cluster and, for each cluster, it will choose the one closest to the cluster center and display it as an event.

Question Answering

Question Answering System is built using the BERT technique, which stands for Bidirectional Encoder Representations from Transformers, BERT takes two parameters, the input question, and passage as a single packed sequence. The input embedding model is the combination of the token embedding and the segment embedding.

- **Token embedding:** Tokens are added to the input word tokens of the question. A 'CLS' token is added at the beginning of the question, and A 'SEP' token is inserted at the end of both the question and the paragraph.
- **Segment embedding:** Sentences in the text will be identified and added to each token. This allows the model to differentiate between sentences.

Every token in the text will be fed its final embedding into the start token classifier. In the start token classifier, a single set of weights will be applied to each word. After applying the dot product between the output embedding and the 'start' weights, the system would activate the softmax activation to produce a probability distribution over all of the words. The word with the highest probability of being the start token will be chosen and corresponding sentences will be considered as answer.

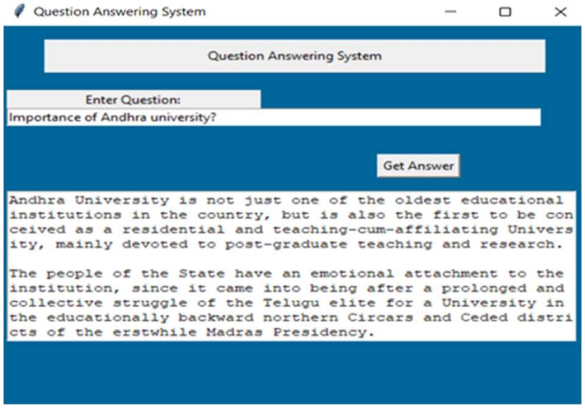

Fig.9.Question Answering system

3. Performance Analysis

We have used various texts(ranges from 2000 to 20,000 approx. words) to perform relation extraction and calculated the time of execution (in seconds) for the features of the application.

- **Text Processing Time**

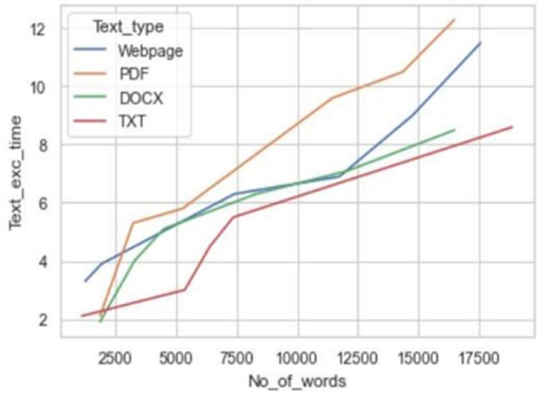

Fig.10. Line graph for Text Processing time

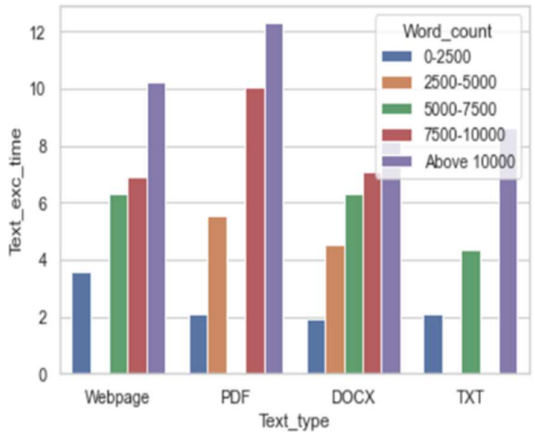

Fig.11. Bar graph for Text Processing time

The above graphs interpret that the Text processing time of portable document format(.pdf) is more and the text file(.txt) processing time is less compared to other text file formats.

- **Extractive and Abstractive summarization**

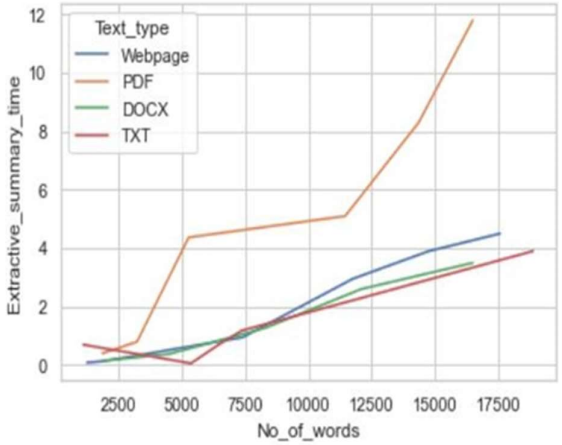

Fig.12. Line graph for the Execution time of Extractive Summarization

The following graphs (Figures 12&13) interpret the various text format's extractive and abstractive summarization execution time concern word count and The portable document format(.pdf) is taking more time to execute compared to other text formats.

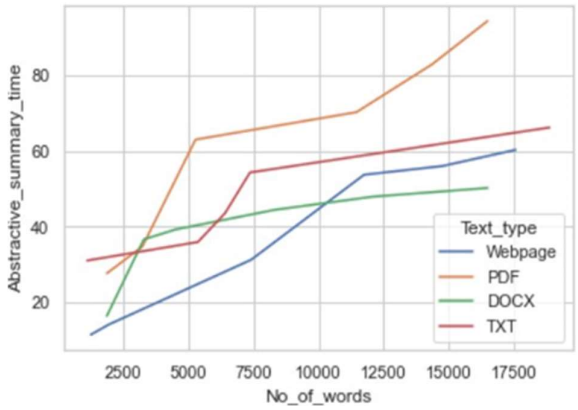

Fig.13. Line graph for the Execution time of Abstractive Summarization

Information Extraction

The following graphs (Figures 14&15) interpret the various text format's Information extraction execution time with respect to word count. The portable document format(.pdf) is taking more time to execute compared to other text formats.

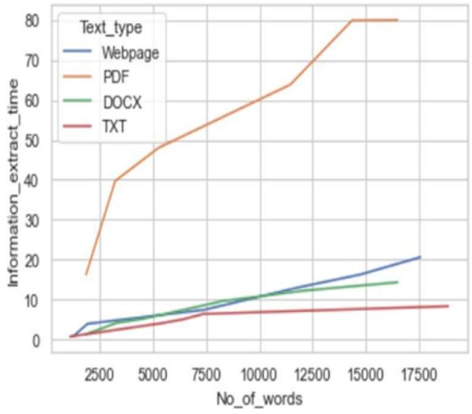

Fig.14. Line graph for the Execution time of Information Extraction

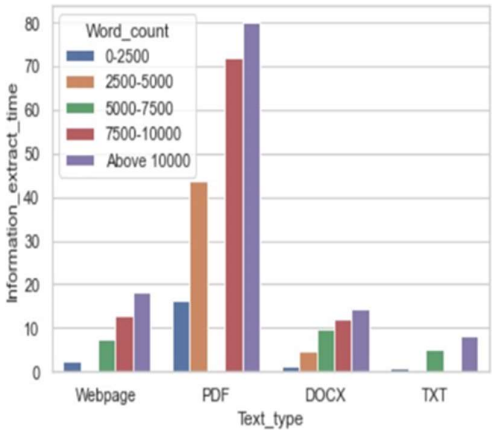

Fig.15. Bar graph for the Execution time of Information Extraction

4. Related Work

Summarization is the process of reducing large amounts of information into short summaries or summaries containing only the most important information. Summarizing is common in everyday communication, but it is also a specialized skill for journalists and scientific writers. The automatic summarization feature is necessary for Internet users who want to take advantage of available information without being overwhelmed. Therefore, the primary use of summarization is to summarize a set of documents returned by an information retrieval system. There are many other possible uses for the summarization technique like document retrieval, automatic generation of comparison tables, just-in-time
knowledge acquisition, Finding answers to specific questions, and biographical profiling.

Text summarization is the technique for generating a concise and precise summary of voluminous texts while focusing on the sections that convey useful information, and without losing the overall meaning. In extraction-based summarization, a subset of words that represent the most important points is pulled from a piece of text and combined to make a summary i.e the main information from a source text. In abstraction-based summarization, advanced deep learning techniques are applied to paraphrase and shorten the original document, just like humans do. abstraction performs better at text summarization, developing its algorithms requires complicated deep learning techniques and sophisticated language modeling.

The Unsupervised approach can also be divided into two categories single-document summaries and multi-document summaries. Chu and Liu (2019) propose an end-to-end unsupervised abstract neural summarization model. Their model consists of an LSTM-based auto encoder that learns representations of each input text and his summarization engine that learns to generate summaries based on the representations encoded by the auto encoder. consists of two parts. Naeem et al. (2018) proposed an unsupervised multi-document summarization system consisting of word graph-based (Filippova, 2010) sentence fusion and integer linear programming (ILP)-based sentence ranking. They first apply a hierarchical agglomerative clustering that perfectly links all sentences in the document and the distance between two sentences is based on their continuous representations.

Sundheim and Riloff et al. (1996) suggested that event extraction is a long-studied and difficult task in the field of information. Its goal is to extract structured information. Takeshi Sakaki and Yutaka Matsuo (2012) develop a system that uses social media to extract real-time driving information andprovide drivers with important events such as traffic congestion and weather forecasts. It is advantageous in areas where intelligent transportation systems (ITS) are poor. Semi-supervised and remote monitoring methods can produce high-quality training data. Amir Polan, Ben Weisse, et al. (2021) explore new open-domain event detection methods by improving a pre-trained language model GPT-2 to automatically generate new training data. In particular, a novel teacher-student architecture is employed to maintain consistency between original and generated data.

Question and answering (QA) is one kind of the typical application of information retrieval for document repositories such as the World Wide Web and local repositories. The system should be able to retrieve answers to questions asked in natural language. QA is considered a more complex Natural Language Processing (NLP) technique than other types of information retrieval such as document retrieval. retrieval, such as B. Document Search. Srihari and Li, (2000) say that it is sometimes seen as the next step beyond search engines. There was a tendency to view non-QA NLP tasks as QA tasks (McCann et al., 2018). Information Extraction (IE) (Friday, 1998) is an emerging research area of information processing that can be used as a data source for QA. So it's an important part of QA. Natural Language Processing (NLP) (Laender et al., 2002), Ontologies (Srihari et al., 2008), HTML Structure (Arocena and Mendelzon, 1998), Web Search (Zhao et al., 2005), Semantic Patterns (Kim and MoNovan, 1993) or rappers (Liu et al., 1999). These technologies are already widely used, but none are suitable for all web data.

5. Conclusion and future work

Text is the largest repository of human knowledge. A deep understanding of board language is a difficult task. Knowledge about the language, knowledge about the world, and a way to combine knowledge resources to extract information in a meaningful manner is achieved by taking different sources so that relations can be known in a well-defined manner. In the future, we can enhance the system to known general knowledge (or relations) using images, audio, and video and at the same time it will be useful for extracting relations from Scientific Applications, Personal Information Management, Enterprise Applications, Web Oriented applications like citation which gives better performance.

References

[1] Mitra, M., Chaudhuri, B.B., 2000. “Information retrieval from documents”: A survey. Inf. Retr. 2, 141–163. URL:https://doi.org/10.1023/A:1009950525500,doi:10.1023/A:1009950525500.

[2] Belkin,N.J.,etal.,1993.Interactionwithtexts: “Informationretrievalasinformationseekingbehavior”.Information retrieval93,55–66.

[3] Takeshi Sakaki, Yutaka Matsuo, Tadashi Yanagihara, Naiwala P. Chandrasiri, and Kazunari Nawa. “Real-time event extraction for driving information from social sensors”. In 2012 IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems (CYBER), pages 221–226. IEEE, 2012.

[4] Qi Li and Heng Ji. 2014. “Incremental joint extraction of entity mentions and relations”. In Proceedings of the52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (ACL), pages 402–412.

[5] Lev Ratinov and Dan Roth. 2009. “Design challenges and misconceptions in named entity recognition”. In Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009), pages 147–155.

[6] Makoto Miwa and Yutaka Sasaki. 2014. “Modeling joint entity and relation extraction with table representation”.In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing(EMNLP), pages 1858–1869.

[7] James Pustejovsky, Patrick Hanks, Roser Sauri, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, and Lisa Ferro. The timebank corpus. In Corpus linguistics, volume 2003 of Corpus linguistics, page 40. Lancaster, UK., 2003.

[8] G. Arocena, A. Mendelzon, WebOQL:" Restructuring Documents, Databases and Webs" , Proc. of 14th. Intl. Conf. on Data Engineering (ICDE 98), Florida, 1998.

[9] Rohini Srihari and Wei Li. 2000. “A Question Answering System Supported by Information Extraction”. In Sixth Applied Natural Language Processing Conference, pages 166–172, Seattle, Washington, USA.

[10] Katja Filippova, 2010. “Multi-Sentence Compression: Finding Shortest Paths in Word Graphs”. In Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pages 322–330, Beijing, China. Coling 2010 Organizing Committee.

[11] Xiangdong You, “Automatic Summarization and Keyword Extraction from Web Page or Text File”. In Proceedings of the 2019 IEEE 2nd International Conference on Computer and Communication Engineering Technology-CCET.

[12] Vishnu Preethi K,and Vijaya MS. “Text Summarizers for Education News Articles”. In Proceedings of the International Journal of Engineering Science Invention (IJESI).

[13] https://towardsdatascience.com/simpleabstractive-text-summarization-with-pretrained-t5-text-to-text-transfer-transfor mer-10f6d602c426.html.Accessed 18July 2022.

[14] https://huggingface.co/docs/transformers/model_doc/t5.html.Accessed 18July 2022.

[15] https://en.wikipedia.org/wiki/Open_information_extraction.html.Accessed 18July 2022.

[16] https://medium.com/analytics-vidhya/question-answering-system-with-bert.html.Accessed 18July 2022.