Experiment No.1
Study various applications of NLP and Formulate the Problem Statement for Mini Project based on chosen real world NLP applications
Date of Performance:
Date of Submission:

Aim: Study various applications of NLP and Formulate the Problem Statement for Mini Project based on chosen real world NLP applications.

Objective: Understand the different applications of NLP and their techniques by reading and critiquing IEEE/ACM/Springer papers.

Theory:

1. **Machine Translation**

Machine translation is a process of converting the text from one language to the other automatically without or minimal human intervention.

2. **Text Summarization**

Condensing a lengthy text into a manageable length while maintaining the essential informational components and the meaning of the content is known as summarization. Since manually summarising material requires a lot of time and is generally difficult, automating the process is becoming more and more popular, which is a major driving force behind academic research.

Text summarization has significant uses in a variety of NLP-related activities, including text classification, question answering, summarising legal texts, summarising news, and creating headlines. Additionally, these systems can incorporate the creation of summaries as a middle step, which aids in shortening the text.

The quantity of text data from many sources has multiplied in the big data era. This substantial body of writing is a priceless repository of data and expertise that must be skillfully condensed in order to be of any use. A thorough investigation of NLP for automatic text summarization has been necessitated by the increase in the availability of documents. Automatic text summarising is the process of creating a succinct, fluid summary without the assistance of a human while maintaining the original text's meaning.

3. Sentiment Analysis

Sentiment analysis, often known as opinion mining, is a technique used in natural language processing (NLP) to determine the emotional undertone of a document. This is a common method used by organisations to identify and group ideas regarding a certain good, service, or concept. Text is mined for sentiment and subjective information using data mining, machine learning, and artificial intelligence (AI).

Opinion mining can extract the subject, opinion holder, and polarity (or the degree of positivity and negative) from text in addition to identifying sentiment. Additionally, other scopes, including document, paragraph, sentence, and sub-sentence levels, can be used for sentiment analysis.

Businesses must comprehend people's emotions since consumers can now communicate their views and feelings more freely than ever before. Brands are able to listen carefully to their customers and customise their products and services to match their demands by automatically evaluating customer input, from survey replies to social media chats.

4. Information Retrieval

A software programme that deals with the organisation, storage, retrieval, and evaluation of information from document repositories, particularly textual information, is known as information retrieval (IR). The system helps users locate the data they need, but it does not clearly return the questions' answers. It provides information about the presence and placement of papers that may contain the necessary data. Relevant documents are those that meet the needs of the user. Only relevant documents will be pulled up by the ideal IR system.

5. Question Answering System (QAS)

Building systems that automatically respond to questions presented by humans in natural language is the focus of the computer science topic of question answering (QA), which falls under the umbrella of information retrieval and natural language processing (NLP).

Text Summarization

Abstract

Text Summarization is the process of creating a summary of a certain document that contains the most important information of the original one, the purpose of it is to get a summary of the main points of the document. Abstractive summarization of multi-documents aims to generate a concentrated version of the document while keeping the main information. Due to the massive amount of data these days, the importance of summarization arose. Finally, this paper collects the most recent and relevant research in the field of the text summarization to study and analysis for future research. It will be significant by giving a new direction to who are interested in this domain in the future. There is a huge amount of data surfacing digitally, therefore the importance of developing a punctuate procedure to shorten long texts immediately while keeping the main idea of it is necessary. Summarization also helps shorten the time needed for reading, fasten the search for information and help to get the most amount of information on one topic .The central object of computerized text summarization is decreasing the reference text into a smaller version maintaining its knowledge alongside with its meaning. Several descriptions of text summarization are provided, for example explained the report as text that is generated from one or more documents that communicate relevant knowledge in the first text, and that is no higher than half of the primary text(s) and usually significantly more limited than that.

Methodology:

1) Extractive Summarization:

- In extractive summarization, the goal is to select and extract the most important sentences or phrases from the original text to create a summary. It doesn't involve rephrasing or generating new sentences.
- Common techniques for extractive summarization include:
- TextRank: It's based on PageRank and uses graph-based algorithms to identify important sentences based on their connections to other sentences.
- TF-IDF (Term Frequency-Inverse Document Frequency): It calculates the importance of words in sentences based on their frequency and uniqueness.
- Machine Learning Models: Supervised models like Support Vector Machines or deep learning models (e.g., LSTM, BERT) can be used to classify sentences as important or not.

2. Abstractive Summarization:

- Abstractive summarization involves paraphrasing and generating new sentences to convey the main ideas of the text. It often requires more advanced natural language processing techniques.
- Common techniques for abstractive summarization include:
- Seq2Seq models: These are neural network architectures that use Recurrent Neural Networks (RNNs) or Transformer models to generate summaries by predicting words or phrases.
- Pointer-Generator Networks: These models can copy words directly from the source text, allowing for more fluent and coherent summaries.
- Reinforcement Learning: Some models are fine-tuned using reinforcement learning to produce more readable and coherent summaries.

3. Hybrid Approaches:

- Hybrid approaches combine elements of both extractive and abstractive summarization to create more accurate and readable summaries. For example, they might extract important sentences and then use abstractive techniques to rephrase and combine them.

4. Evaluation:
- It's important to evaluate the quality of generated summaries. Common evaluation metrics include ROUGE (Recall-Oriented Understudy for Gisting Evaluation), BLEU (Bilingual Evaluation Understudy), and METEOR (Metric for Evaluation of Translation with Explicit ORdering).

5. Preprocessing:
- Text preprocessing is essential, including tasks like sentence and word tokenization, stop-word removal, and stemming or lemmatization.

6. Data:
- You'll need a dataset for training and testing your summarization model. Depending on the task, you might use news articles, scientific papers, or any other text source.

7. Fine-Tuning:
- If using pre-trained models like BERT or GPT, you may need to fine-tune them on your specific summarization task to achieve better results.

8. Post-Processing:
- After generating a summary, post-processing techniques can be applied to ensure coherence, remove redundancy, and improve readability.

Technical Terms:

1. Abstractive Summarization: A text summarization approach that generates a summary by paraphrasing and rephrasing the content rather than directly extracting sentences or phrases from the source text.

2. Extractive Summarization: A summarization technique that selects and extracts important sentences or phrases from the source text to create a summary.

3. ROUGE (Recall-Oriented Understudy for Gisting Evaluation): A metric commonly used to evaluate the quality of machine-generated summaries by measuring the overlap of n-grams (sequences of n words) between the generated summary and reference summaries.

4. BLEU (Bilingual Evaluation Understudy): A metric used for evaluating the quality of machine-generated text, including summaries, by comparing the overlap of n-grams in the generated text with reference text.

5. METEOR (Metric for Evaluation of Translation with Explicit ORdering): An evaluation metric that takes into account both precision and recall of n-grams and includes stemming and synonym matching.

6. TF-IDF (Term Frequency-Inverse Document Frequency): A numerical statistic that evaluates the importance of words in a document within a collection of documents. It is often used in extractive summarization to determine the significance of words.

7. PageRank: An algorithm used in graph-based extractive summarization methods like TextRank to assess the importance of sentences or phrases based on their connections to other sentences in a text.

8. Seq2Seq (Sequence-to-Sequence): A neural network architecture commonly used for abstractive summarization. It consists of an encoder-decoder framework for converting a sequence of words into another sequence, such as generating a summary.

9. Pointer-Generator Networks: A type of neural network architecture used in abstractive summarization that combines elements of extraction and generation, allowing the model to copy words directly from the source text when necessary.

10. Supervised Learning: A machine learning approach where the model is trained on labeled data, i.e., data with known summaries or labels, to predict or classify text based on the training examples.

11. Reinforcement Learning: A machine learning approach used for fine-tuning summarization models. It involves rewarding or penalizing the model based on the quality of generated summaries, often using reinforcement signals.

12. Fine-Tuning: The process of taking a pre-trained language model (e.g., BERT, GPT) and training it on a specific summarization task to adapt it to the task's requirements.

13. Preprocessing: The initial step in text summarization, which includes tasks such as text tokenization (splitting text into words or sentences), stop-word removal, and stemming or lemmatization.

14. Post-Processing: The step that follows the generation of a summary, involving actions like removing redundancy, ensuring coherence, and improving the readability of the generated summary.

15. Coverage: A concept in abstractive summarization referring to how well the generated summary covers the important content from the source text.