Research Projects on NLP

Nitin Arvind Shelke

Question Answering

- **Question answering** is one of the most prevalent research problems in NLP.
- Some of its applications are chatbots, information retrieval, dialog systems, among others.
- It serves as a powerful tool to automatically answer questions asked by humans in natural language, with the help of either a pre-structured database or a collection of natural language documents.
- **Models:** Models like LSTM, BiLSTM, BiDAF, BERT, and XLNet can be used for question-answering projects.
- **Dataset:** Stanford Question Answering Dataset (SQuAD), Conversational Question Answering systems (CoQA), etc.

Text Classification

- **Text Classification** or Text Categorization is the technique of categorizing and analyzing text into some specific groups. This technique supports a comparative evaluation of the impact of linguistic information concerning approaches based on word matching.
- **Models**: LSTM, BiLSTM, BERT, XLNet, and RoBERTa can be used for text classification.
- **Dataset**: Amazon Reviews dataset, IMDB dataset, SMS Spam Collection, etc.

Text Summarization

- **Text summarization** is one of the most efficient methods to interpret text information.
- Text summarization methods can be mainly categorized into two parts – extractive summarization and abstractive summarization.
- In extractive summarization, the process involves selecting sentences of high rank from any document based on word and sentence features and fusing them to generate a summary.
- On the other hand, an abstractive summarization is mainly used to understand the main concepts in any given document and then express those concepts in any natural language.

Text Summarization (Cont.)

- Models: LSTM, BiLSTM, BERTSumExt, BERTSumAbs, and UniLM (s2s-ft) can be used for text summarization.

- Dataset: BBC News Summary, Large-Scale Chinese Short Text Summarization Dataset, 20 Newsgroups dataset etc.

Sentiment Analysis

- **Sentiment Analysis** is the technique of understanding human sentiments implied in a text, and helps classify emotions using text analysis methods.
- This technique has witnessed significant traction due to the growth of social media platforms like Facebook, Instagram, and more. Some of the applications of this technique are market research, brand monitoring, customer service, among others.
- **Models:** Models like LSTM, BiLSTM, Dependency Parser, BERT, and RoBERTa can be used for sentiment analysis.
- **Dataset:** Stanford Sentiment Treebank, Multi-Domain Sentiment Dataset, Sentiment140, etc.

Sentence Similarity

- **Sentence similarity** portrays an important part in text-related research and applications in areas such as text mining and dialogue systems. This technique has proven to be one of the best to improve retrieval effectiveness, where titles are used to represent documents in the named page finding task.
- **Models:** LSTM, BiLSTM, BERT, GloVe, etc. can be used for sentence similarity projects.

- **Dataset:** Paraphrase Adversaries from Word Scrambling (PAWS)

Speech Recognition

- Speech Recognition is the technique used in identifying spoken words or phrases and translating them into machine language.
- Speech recognition has gained attention in recent years with the dramatic improvements in acoustic modeling yielded by deep feedforward networks.
- **Models:** LSTM, BiLSTM, BERT, RoBERTa, etc. can be used for speech recognition projects.
- **Dataset:** Google AudioSet, LibriSpeech ASR corpus, etc.

Neural Machine Translation

- Neural machine translation is one of the most popular approaches in NLP research. The neural machine translation aims at building a single neural network that can be jointly tuned to maximize translation performance.
- **Models:** LSTM, BiLSTM, BERT, RNN Encoder-Decoder, etc.
- **Dataset:** English-Persian parallel corpus, Japanese-English Bilingual Corpus, etc.