Applications

Machine Translation

Machine Translation (MT) is the automated process of translating one natural language to another. Machine translation, an integral part of Natural Language Processing where translation is done from source language to target language preserving the meaning of the sentence.

There are few challenging aspects of MT:

1) The wide variety of languages, alphabets and grammar;

2) The task to translate a sequence to a sequence is harder for a system than working with numbers only

3) There is no one correct answer Eg

- Different types of Machine Translation

1 Statistical Machine Translation (SMT)

- SMT functions by referring to statistical models that are based on the analysis of large volumes of bilingual text

- It work towards or aims to determine the correspondence between a word from the target language and a word from the source language.

- Google translate is an example!

- SMT is good for basic translation. But its disadvantage is that it does not factor in context, which means translations can often be erroneous. We can also say that it dont expect high quality translations

2. Neural Machine Translation (NMT)

- It is a new approach that makes machines learn to translate through one large neural network (multiple processing devices modeled on the brain)

- The approach has become increasingly popular among MT researchers and developers, as trained NMT systems have begun to show better translation performance in many language pairs compared to the

phrase based statistical approach.

3. Rule based Machine Translation (RBMT)

- It translates on the basis of grammatical rules
- To generate the translated sentence RBMT conducts a grammatical analysis of the source language and the target language.
- It requires proof reading, and its heavy dependence on lexicons means that efficiency is achieved after a long period of time

4. Hybrid Machine Translation (HMT)

- It is a mix of RBMT and SMT. It takes upp, the translation memory making it far more effective in terms of quality
- HMT also has its disadvantages:
  1. It needs heavy editing
  2. There is a requirement of Human Translators.

Information Retrieval (IR)
or
Write a Short Note on Information Retrieval (IR)

Sol: Information Retrieval is one of the most challenging problems of Natural Language

- IR is defined as a software program that deals with the organization, Storage, retrieval and evaluation of information from document repositories particularly textual information.

- IR system assists users in finding the information and it informs the existence and location of documents that might consist of the required information.

- The documents that satisfy the user's requirment are called relevant documents. Example: Google, Yahoo, Altavista etc.

- Traditionally, the IR system techniques are based on keyword.

- They use lists of keywords to describe the content of information but they do not say reveal semantic relationships between keywords nor consider the meaning of words and phrases.

=> Basic IR system involves following stages:

1. Indexing the collection of documents.
2. Transforming the query in the same way as the document content is represented
3. Comparing the description of each document with that of the query.
4. Listing the results in order of relevancy

- In general all IR Systems consists of mainly two processes as

a) Indexing: Indexing is the process of selecting terms to represent a text which involves tokenization of string, removing frequent words and stemming.

b) Matching: It is the process of computing a measure of similarity between two text representations. Relevance of a document is computed based on parameters like term frequency and inverse document frequency.

Question Answers System

Question Answering (QA) System is a task of automatically answering to the questions asked in natural language using either a pre structured database or a collection of natural language documents

- It presents only the requested information instead of searching full documents like search engine.

- The basic idea behind the QA System is that the users just have to ask the question and the system will retrieve the most appropriate and correct answer for the question

- Example
  Q. "What is the birth place of Shree Krishna"?
  A. Mathura.

- Question answering System helps users to find the precise answers to the question articulated in natural language.

- Question answering system provides explicit, concise and accurate answer to user questions rather than providing a set of relevant documents or web pages as

answers as most of the information retrieval system does.

- Question answering system basically consists of three parts as- Question processing - answer retrieval - answer generation.
- QAS has become part of daily life of users.
- over a period of time many personal assistance software like Siri, Cortana, Google Now, Alexa are developed which provide precise and accurate answer to user's questions.
- Datasets for Q A systems are
  1. Stanford Question Answer Dataset
  2. Wiki QA dataset
  3. TREC-QA
  4. News-QA

=> Question Answering system challenges

- Lexical Gap: In natural Language, the same meaning can be expressed in different ways. Because a question can usually only be answered if every refered concept is identified, bridging this gap significantly increases the proportion of questions that can be answered by a system.

- Ambiguity: It is the phenomenon of the same phrase having different meanings. This can be structural and syntactic (like 'flying planes') or lexical and semantic ('like' 'bank'). The same string accidentally refers to different concepts (as in money bank Vs. river bank) and polysemy, where the same string refers to different but related concepts (as in bank as a company vs bank as a building)

- Multilingualism: Knowledge on the Web is expressed in various languages. While ~~RDS~~ RDF resources can be described in multiple languages at once using language tags, there is not a single language that is always used in web documents.

- Additionally: Users have different native languages. A Q&A system is expected to recognize a language and get the results on the go!

Text Categorization System

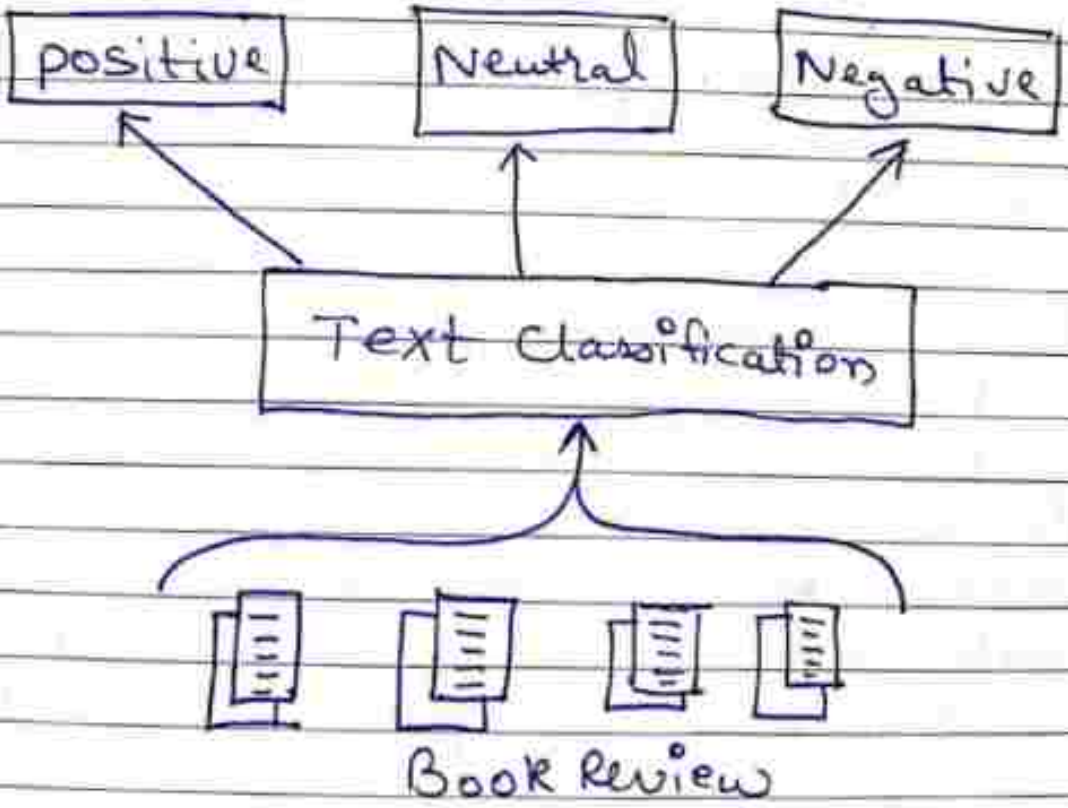
Need of text categorization System

- Rapid development of Information technology has led to massive growth in text data found on internet
- Data mining field worked mostly on English text document.
- Nowadays, millions of documents are present in indian regional languages like Telugu, Tamil, Hindi, Punjabi, Bengali, Urdu, Marathi.
- TO classify such documents manually is an expensive and time consuming task.
- Automatic classification can help in better management and retrieval of these text documents.
- Also their accuracy and time efficiency is much better than manual text classification.

- Text categorization (also known as text classification or topic spotting) is the task of automatically sorting a set of documents into categories (clusters)

o Uses of Text categorization

- Filtering of content
- Spam filtering Identification of document content
- Survey coding.



- Applications of text categorization

- Email message filtering
- News & event tracked and filtered by topics
- Web pages organized into categories hierarchy

- Text Classification can be achieved through three main Approaches

o Rule based approach: ~~Trust app~~
- Use of handcrafted linguistic rules to classify text is been made here.
- A way do group text is to creat list of words related to a certain column and then judge the text based on occurances of these words.
- for example ~~word~~ 'fur', 'feathers', 'claws' and 'Scales' could help Zoologist identify texts talking about animals online.
- But this approach requir a lot of domain knowledge to be extensive, take a lot of time to compile and are difficult to scale.

o Machine learning approach: Machine learning is used to train models on large scale of text data to predict categories of new text. For the training of models, we need to transform text data into numerical data. This is feature extraction.

Important feature extraction techniques include bag of words and n-grams. There are many useful ML algos which can be used for text classification.

Naive Bayse classifiers, SVM are the most popular or famous once.

- Hybrid Approach:

They are a combination of ML Approach and Rule-based approach. They make use of these two to model a classifier that can be fine-tuned in certain scenarios.

⇒ Language detection, NLP, Topic detection, Semantic analysis are some of the most common examples and use cases for Automatic text classification.

Text Summarization

- I dont want a full report, just give me a summery of the results. I have often found myself in this situation - both in college as well as my professional life. We prepare a comprehensive report and the teacher/supervisor only has time to read the summary.

- Summarization means to reduce the size of the document without changing its meaning.

- Automatic text summarization is the task of producing a concise and fluent summary while preserving key information content and overall meaning.

- A good summary should cover the most vital information of the original document or a cluster of documents. While being coherent, non redundant and grammatically readable

Text Summarization

- Abstractive Summarization
- Extractive Summarization

- Extractive based Summarization

- The extractive text Summarization technique involves pulling keyphrases from the source document and combining them to make a summery.

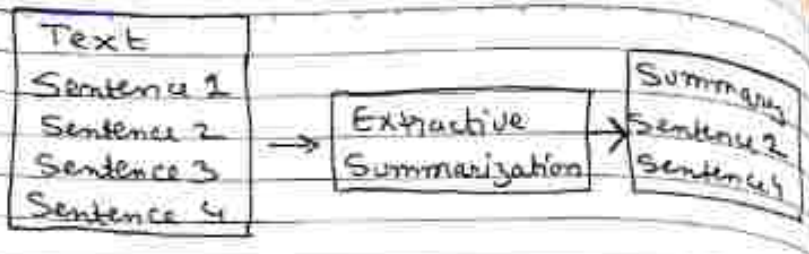
- The Extraction is made according to the defined metric without making any changes to the texts.

- Source text: Joseph and Marya rode on a donkey to attend the annual event in Jerusalem. In the city Mary gave birth to a child named ~~Jesus~~ Zeus

- ~~Extra~~ Extractive Summary: Joseph and Marya attend event Jerusalem. Marya birth Zeus.

- Source text: Joseph and Marya rode on a donkey to attend the annual event in Jerusalem. In the city Mary gave birth to a child named Zeus

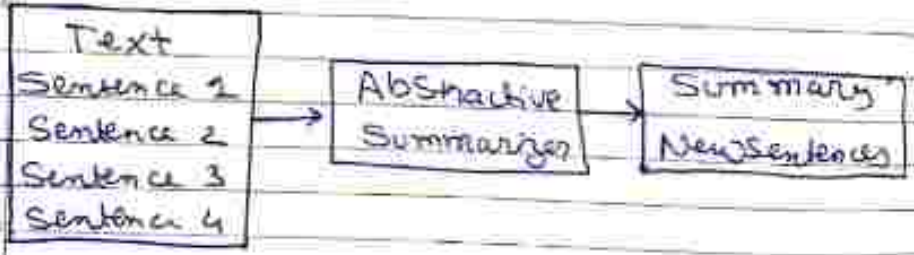
- Extractive Summary: Joseph and Marya attend event Jerusalem. Marya birth Zeus.



- Abstraction based Summarization

- The abstraction technique entails paraphrasing and shortening parts of the source document

- The abstractive text summarization algorithms create new phrases and sentences that relay the most useful information from the original text. Just like humans do. Therefore abstraction performs better than extraction.



- Applications of text summarization
  - can be used as a preliminary stage for information retrieval tasks
  - simplifies text categorization
  - widely used due to information overload problem where information searched is very large and where there is a need of meaningful summary and saves time.