NLP - MODULE 6 - CHAPTER 6

APPLICATIONS

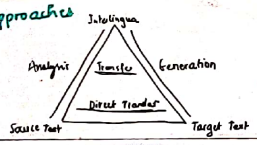
Machine Translation

- refers to the process where computer s/w translates the source language into target language without human intervention
- task to translate text from source language to its counterpart in target language
- Challenges:
  - large variety of languages, alphabet & grammar
  - task to translate a sequence to sequence is harder for computer than working with numbers only.
  - no one correct answer

Types of Machine Translation

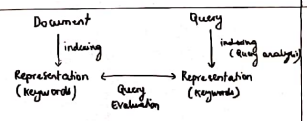
1) Statistical Machine Translation (SMT)
2) Rule Based Machine Translation (RBMT)
3) Hybrid Machine Translation (HMT)
4) Neural Machine Translation (NMT)

Approaches



Information Retrieval (IR)

- finding material (usually docs) of an unstructured nature (usually text) that satisfy an information need from within large collection (usually stored on computer)



Type of IR Models

First Dimention: Mathematical basis

- Set Theoretic Model
  - Standard boolean model
  - Extended boolean model
  - Fuzzy retrieval
- Algebric Model
  - Vector space model
  - Generalized vector space model
  - Enhanced topic based vector space model
  - Extended boolean model
  - Latent Semantic indexing
- Probabilistic Models
  - Binary Independence model
  - Probabilistic relevance model
  - Uncertain inference
  - Divergence from randomness model
  - Language models
  - Latent dirichlet allocation

Second Dimentional: Properties of model

- Models without term-interdependencies treat different terms/words as independent.
- Models with immanent term interdependencies allow a representation of interdependencies between terms
- Models with transcedent term interdependencies allow a representation of interdependencies between terms but they do not allege how the interdependency between two terms is defined

Classification of IR models

1) Classical IR model
2) Non classical IR model
3) Alternative IR model

Design features of IR system

1) Inverted Index
2) Stopword Elimination
3) Stemming
4) Boolean model

Question Answer System

- branch of learning of IR & NLP which focuses on building system that automatically answer posed by users in natural language
- Process:
  1) Query processing
  2) Document retrieval
  3) Passage retrieval
  4) Answer extraction

Categorization

- also known as text tagging
- process of categorizing text into organised groups
- Approaches:
  1) Rule Based Approach
  2) ML Approach
  3) Hybrid Approach

Summarization

- reductive transformation of source text into summary text by extraction or generation

Types of Summarization

1) Extractive Summarization
2) Abstractive Summarization

Sentiment Analysis

- process of detecting positive or negative sentiment in text.

Types of Sentiment Analysis

1) Rule Based
2) Automatic
3) Hybrid

Application of Sentiment Analysis

1) Social media monitoring
2) Brand monitoring
3) Voice of customer
4) Customer service
5) Market Research

Named entity Recognition

- technique that automatically identifies named entities in a text & classifies them into predetermined categories
- entities can be names of people, organization, locations, quantities, monetary values, percentage & more

Types of Named Entity Recognition

1) Entity Name Types
2) Numerical Expression
3) Time Expression

Challenges in NER

1) morphologically rich
2) no capitalization feature
3) Ambiguity
4) Spell Variation
5) Less resource
6) Lack of easy availability of annotated data

RBMT

1) Consistency between versions
2) provides good out of domain quality & is by nature predictable
3) Knows grammatical rules
4) High performance & robustness
5) Lack of fluency
6) Hard to handle exception to rules
7) High development & customization costs

SMT

1) Inconsistency between versions
2) provides good quality when large & qualified corpora are available
3) Doesn't know grammar
4) High CPU & disk space requirement
5) Good fluency
6) Good to catching exception to rules
7) Rapid & cost effective development cost provided the required corpus exists.

Data Retrieval

1) Determine which document of a collection contain the keyword in user query
2) All objects which satisfy clearly defined conditions are retrieved
3) A single erroneous object means total failure
4) Data has a well defined structure & semantics

Information Retrieval

1) Retrieve information about subject rather than data which satisfies a given query.
2) IR system somehow "interpret" the content of document in a collection and rank them according to a degree of relevance to user query
3) Retrieval objects might inaccurate & small errors are ignored
4) Data is natural language text which is not always well structured & could be semantically ambiguous.