Name – Alok Bhawankar

Panel - D

Roll No. -PD 09

Subject – CC & NLP

School of Computer Engineering and Technology

1. What is “Word Sense Disambiguation” with respect to NLP? Explain in detail.

Word sense disambiguation, in natural language processing (NLP), may be defined as the ability to determine which meaning of word is activated using word in a particular context. Lexical ambiguity, syntactic or semantic, is one of the very first problem that any NLP system faces. Part-of-speech (POS) taggers with high level of accuracy can solve Word’s syntactic ambiguity. On the other hand, the problem of resolving semantic ambiguity is called WSD (word sense disambiguation). Resolving semantic ambiguity is harder than resolving syntactic ambiguity.

For example, consider the two examples of the distinct sense that exist for the word “bass”

- I can hear bass sound.
- He likes to eat grilled bass.

The occurrence of the word **bass** clearly denotes the distinct meaning. In first sentence, it means frequency and in second, it means fish. Hence, if it would be disambiguated by WSD then the correct meaning to the above sentences can be assigned as follows −

- I can hear bass/frequency sound.
- He likes to eat grilled bass/fish.

Different approaches to WSD

- Dictionary and Knowledge based methods:

  These rely primarily on dictionaries, thesauri, and lexical knowledge bases, without using any corpus evidence.

- Supervised methods:

  These make use of sense-annotated corpora for training.

- Semi-supervised methods:

  These methods require very small amount of annotated text and large amount of plain unannotated text.

- Unsupervised methods:

  This approach works directly from raw unannotated corpora.

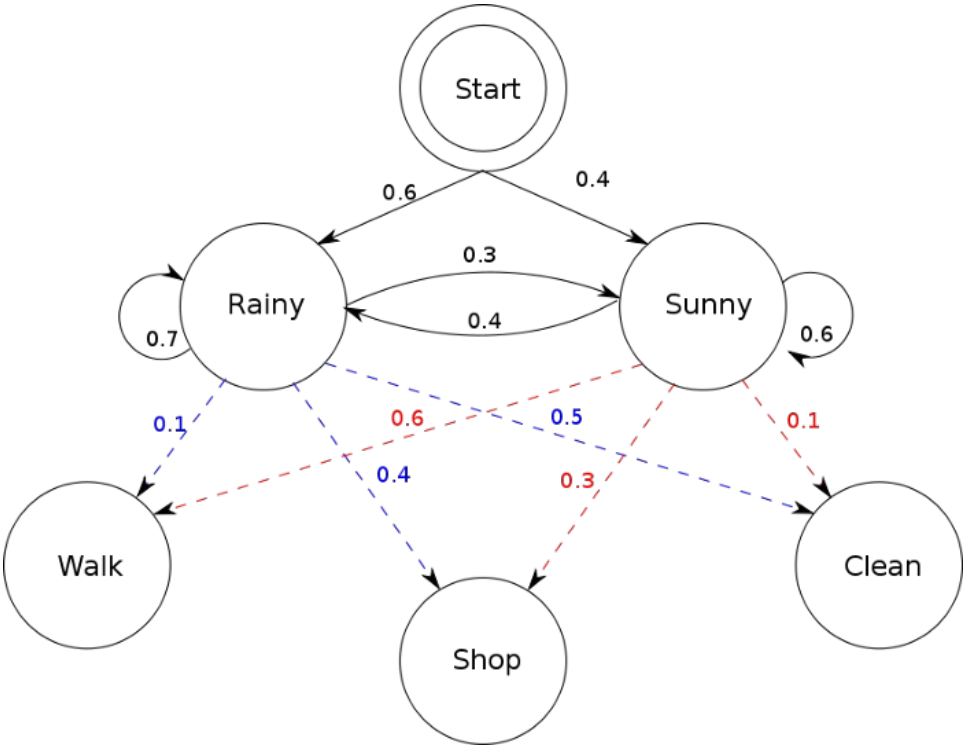
2. **Explain Hidden Markov Model in detail. Give one real life application where HMM can be used.**

   Hidden Markov Model (HMM) is a special type of Bayesian network. Hidden Markov Model (HMM) is a statistical Markov model in which the system being modeled is assumed to be a Markov process with unobserved (i.e. *hidden*) states.

   Hidden Markov models are especially known for their application in reinforcement learning and temporal pattern recognition such as speech, handwriting, gesture recognition, part-of-speech tagging, musical score following, partial discharges and bioinformatics.

   Terminology in HMM

   The term hidden refers to the first order Markov process behind the observation. Observation refers to the data we know and can observe. Markov process is shown by the interaction between “Rainy” and “Sunny” in the below diagram and each of these are **HIDDEN STATES**.



OBSERVATIONS are known data and refers to “Walk”, “Shop”, and “Clean” in the above diagram. In machine learning sense, observation is our training data, and the number of hidden states is our hyper parameter for our model. **State transition probabilities** are the arrows pointing to each hidden state. **Observation probability matrix** are the blue and red arrows pointing to each observation from each hidden state. **Initial state distribution** gets the model going by starting at a hidden state.

Now with the HMM what are some key problems to solve?

i. **Problem 1**, Given a known model what is the likelihood of Observational sequence happening?
ii. **Problem 2**, Given a known model and sequence Observational sequence, what is the optimal hidden state sequence? This will be useful if we want to know if the weather is “Rainy” or “Sunny”
iii. **Problem 3**, Given sequence O and number of hidden states, what is the optimal model which maximizes the probability of Observational sequence?

Hidden Markov Models (HMMs) are used for facial expression recognition because they perform well in the spatio-temporal domain and are analogous to human performance (e.g., for speech and gesture recognition). We use the Facial Action Coding System (FACS) to identify facial action.

3. Present a case study of sentiment analysis in any of the following areas:

a. Market research analysis- (Reference- http://wps.fep.up.pt/wps/wp489.pdf)

Sentiment analysis is a data mining technique that systematically evaluates textual content using machine learning techniques. As a research method in marketing, sentiment analysis presents an efficient and effective evaluation of consumer opinions in real time. It allows data collection and analysis from a very large sample without hindrances, obstructions and time delays. Through sentiment analysis, marketers collect rich data on attitudes and opinion in real time, without compromising reliability, validity and generalizability. Sentiment analysis provides an opportunity for marketers to collect data on customers in their natural cyber environment, without the presence of the researcher being felt. Therefore, this method eliminates the problem of people reacting differently when they know their responses are being collected.

Market research is regarded as a systematic approach that involves data collection and data analysis on any relevant marketing related issues. In marketing, research is used for a variety of purposes including obtaining insights into customer attitudes and beliefs, measuring customer satisfaction, ascertaining the effectiveness of advertising, etc. Some larger companies have their own market research departments whilst smaller companies usually outsource the function to research specialists. Research is carried out in two basic ways: qualitative and quantitative.

In a qualitative approach, the researcher makes knowledge claims based primarily on a constructivist perspective (i.e. the multiple meanings of individual experiences, meanings socially and historically constructed, with an intent of developing a theory or pattern) or advocacy/participatory perspectives (i.e. political, issue-oriented, collaborative or change oriented) or both. Qualitative research involves finding out what people think, and how they feel - or at any rate, what they say they think and how they say they feel. This kind of information is subjective. It involves feelings and impressions, rather than numbers.

On the other hand, quantitative research focuses on measuring an objective fact. Key to conducting quantitative research is definition of variables of interest and to a large extent a sense of detachment in the data collection by the researcher. Quantitative research analyses data using statistics and relies on large samples to make generalized statements. Marketers also gather feedback on attitudes and opinions as they occur without having to invest in lengthy and costly market research activities.

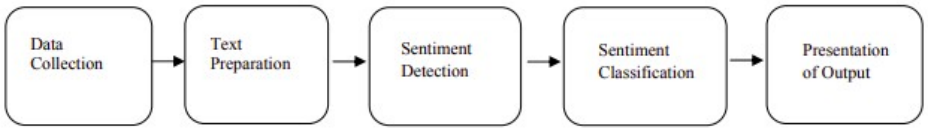


Figure 1: Sentiment analysis process

Researchers have found ways to avoid the use of manual annotation by utilizing existing online textual content generated from sites such as Epinion, Amazon, Rotten Tomatoes, Twitter, Facebook. Several sentiment search engines exist where users run typical queries on any topic of interest and generate text results. Usually the results are coded and categorized into two or three polar categories. Some examples currently available are:

i. Twitrratr – www.twitrratr.com
ii. Sentiment 140 - http://www.sentiment140.com
iii. Tweetfeel – www.tweetfeel.com
iv. Opinmind – www.opinmind.com
v. Social Mention – www.socialmention.com

Sentiment search engines make sentiment analysis quite easy. But, the online reviews on sites like Amazon and Epinion have been found to be skewed towards the positive which raises questions on validity and reliability of sentiment classification. However, Pang and Lee (2008) admit that although the content might be skewed, the validity of the process is acclaimed. Another tool in sentiment analysis is word lists or annotated databases which categorize words based on their emotions for example - attractive (positive valance) or aversive (negative valance). Some examples include: ANEW, General Inquirer and LIWC. Other tools include sentiment analysis programs that are specifically designed to categorize short textural documents. One example is sentistrength.