Week4

Semantic Text Similarity

Applications

- Grouping similar words into semantic concepts
- As a building block in NLU like Textual entailment or paraphrasing

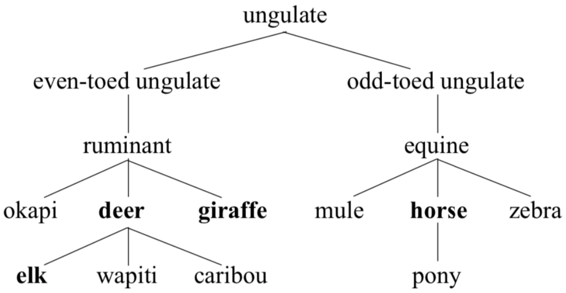
WordNet

- WordNet organizes information in a hierarchy
- Many similarity measures use the hierarchy in some way
- Verbs, nouns, adjectives all have separate hierarchies
- Find the shortest path between two concepts

Path similarity

- Similarity measure inversely related to path distance
    - PathSim(deerk,elk) = 1/(path+1)=1/2

- PathSim(deer, giraffe) = 1/3
- PathSim(deer, horse) = 1/7



Lowest common subsumer (LCS)

- Find the closest ancestor to both concepts
  - LCS(deer, elk) = deer
  - LCS(deer, giraffe) = ruminant
  - LCS(deer, horse) = ungulate

Lin similarity

- Similarity measure based on the information contained in the LCS of the two concepts.
- LinSim(u, v) = 2 x log P(LCS(u,v)) / (log P(u) + log P(v))
- P(u) is given by the information content learnt over a large corpus.

Collocations and Distributional Similarity

- Two words that frequently appears in similar contexts are more likely to be semantically related
- Words before, after, within a small window
- Parts of speech of words before, after, in a small window
- Specific syntactic relation to the target word
- Words in the same sentence, same document, …
- How frequent are these?
  - Not similar if two words don’t occur together often
- Also important to see how frequent are individual words.
  - ‘the’ is very frequent, so high chances it co-occurs often with every other word

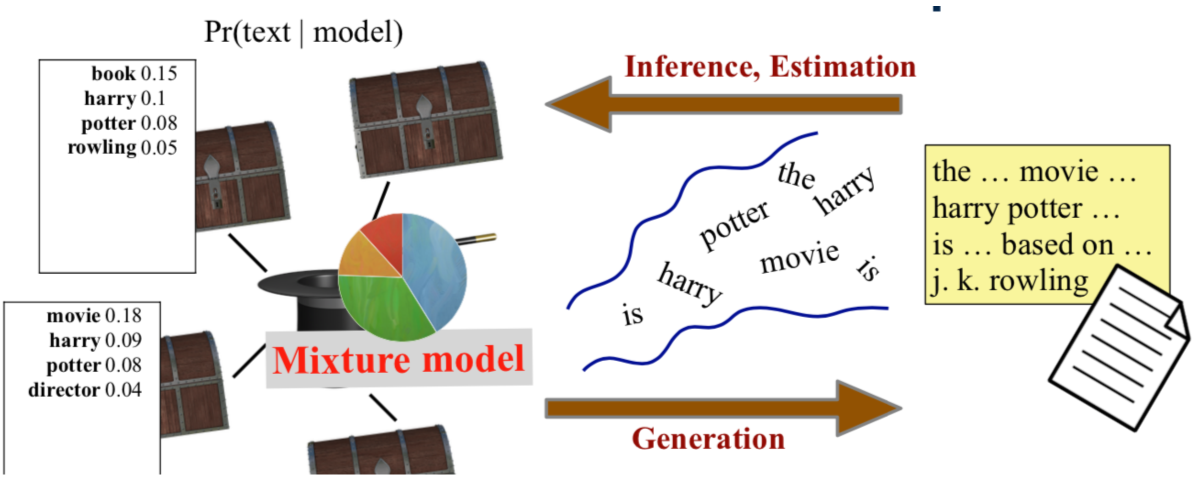
- So to see if a word is important or related to other or just very common —>
    - **Pointwise Mutual Information**: $PMI(w,c) = log\frac{P(w,c)}{P(w)P(c)}$

Topic modelling

- A course-level analysis of what's in a text collection.
- Topic : the subject (theme) of a discourse.
- Topics are represented as a word distribution (Each word in a document has a probability of belonging to a set of topics).
- A document is assumed to be a mixture of topics.
- Essentially, a text clustering problem
    - Documents and words clustered simultaneously
- Known:
    - The text collection or corpus
    - Number of topics
- Unknown:
    - The topics
    - Topic distribution for each document

Generative models and LDA (Latent Dirichlet Allocation)

- Using a corpus of words, for each topic, a document is generated. Then the process is reversed and each word is assigned a Pr of belonging to that topic.



- Generative model for a document d
  - Choose length of document d
  - Choose a mixture of topics for document d.
  - Use a topic’s multinomial distribution to output words to fill that topic’s quota

- In practice:
  - Choose how many topics → Finding or even guessing the number of topics is hard.
  - Interpreting topics
    - Topics are just word distributions.
    - Making sense of words / generating labels is subjective.
  - Preprocess text:
    - Tokenize, normalize (lowercase)
    - Stop word removal
    - Stemming
  - Convert tokenized documents to a document - term matrix
  - Build LDA models on the doc-term matrix

Information extraction

Goal: Identify and extract fields of interest from free text

Named entity recognition

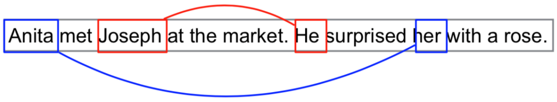
- Named entities: Noun phrases that are of specific type and refer to specific individuals, places, organizations, …
- Named Entity Recognition: Technique(s) to identify all mentions of pre-defined named entities in text
    - Identify the mention/phrase: *Boundary classification* (a task on itself).
    - Identify the type: *Tagging / classification*.

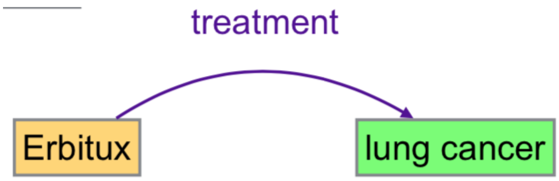
The approach to the task depends on the kind of entities that need to be identified (for simple extractions, regular expressions may be very successful.

Person, Organization, Location/GPE, Other

Typically there are 4 classes: PER, ORG, LOC/GPE, OTHER/OUTSIDE any other class.

- Co-reference Resolution: Disambiguate mentions and group mentions together
- Relation extraction: Identify relationships between named entities


Anita met Joseph at the market. He surprised her with a rose.


treatment
Erbitux
lung cancer