Human Communication 1
Lecture 9

Grammar-based language models

- Rules for
  - parsing
  - computing meanings
- Language-controlled calculator
  - many rules
  - limited coverage
  - what is not covered by the grammar is not “understood”

Statistical Language Models (SLM)

- A Statistical Language Model (SLM) captures regularities of a natural language
- Main idea: estimate the probability distribution of various linguistic units, such as words, sentences and whole documents
- Purpose: improve the performance of natural language applications

Statistical Language Model

- An SLM
  - assigns a probability to a sequence of *m* words or *m* word groups
  - by means of a probability distribution (e.g. a normal distribution also called bell curve)
  - extracts information about structure or content from these probabilities

Example: document retrieval

- Situation: large database of documents
- Problem: how to retrieve relevant documents?
- Solution: instead of parsing the content of all documents and trying to interpret them, compute the similarities between documents

Other areas of applications

- Question answering, e.g. automated booking systems
- Speech recognition (capture the properties of a language in order to predict the next word in a speech sequence)
- Machine translation
- Parsing

n-grams – one kind of SLM

- n-gram models go back to an experiment by Claude Shannon
- Shannon asked: Given a sequence of letters (for example, the sequence “for ex”), what is the likelihood of the next letter?

Shannon’s result

- We cannot consider all possible cases (obviously)
- So, we look at a subset (training data)
- From the training data it is possible to
  - derive a probability distribution for the next letter,
  - given a history of size *n*:
    - a = 0.4,
    - b = 0.00001,
    - c = 0, …
  - where the sum of probabilities of all possible next letters is 1.0

n-gram models

- n-grams are a generalisation of Shannon's result
- Used in statistical natural language processing
  - Speech recognition: phonemes and sequences of phonemes are modelled using n-gram distribution (Hidden Markov Model)
  - Parsing: words are modelled such that each n-gram is composed of n words
  - Language recognition: sequences of letters are modelled for different languages

Unigrams, bigrams trigrams

- n-gram models look at letter or word sequences, e.g.
  - unigram → "of"
  - bigram → "of the"
  - trigram → "of the house"

Trigram examples

1. Sequence of words: "The dog smelled like a skunk"
2. Sequence of characters: "good morning"

Trigrams for "the dog smelled like a skunk"

- # the dog
- the dog smelled
- dog smelled like
- smelled like a
- like a skunk
- a skunk #

Trigrams for “good morning”

- go
- goo
- ood
- od
- dm
- mo
- mor

⋮

Computing probabilities

Bigram

$$P(I,saw,the,red,house) =$$
$$P(I|<s>)*P(saw|I)*P(the|saw)*P(red|the)*P(house|red)$$

Trigram

$$P(I,saw,the,red,house) =$$
$$P(I|<s><s>)*P(saw|<s>I)*P(the|I,saw)*$$
$$P(red|saw,the)*P(house|the,red)$$

Learning grammars

- Up to now we wrote grammars by hand
- SLMs enable the computer to learn a grammar from a corpus of language data
- Many different approaches but resulting grammar is not necessarily a grammar of the kind we had up to now

Function words vs. content words

- Function words have mainly a grammatical function, content words mainly carry meaning
- Function words are more frequent
- Function words: closed class; content words: open class
- For example, the bigram “of the” is very frequent but carries almost no meaning

Frequency and meaning

In English text the most commonly occurring bigrams and trigrams convey little meaning. Manning & Schutze (2003b) analysed roughly 14 million words from New York Times newswire articles and the 10 most common bigrams found are: "of the", "in the", "to the", "on the", "for the", "and the", "that the", "at the", "to be" and "in a" – not very useful at all. In fact, of the 20 most common bigrams found "New York" (15th on the list) was the only phrase that could be considered to convey useful meaning. Our analysis of text from about 10,000 English language web pages also returned similar results.

More Effective Web Search Using Bigrams and Trigrams, D. Johnson et al, 2006

Latent Semantic Analysis (LSA)

- n-grams are useful on the lower levels of linguistic analysis
- Difficult to use for computing meanings, e.g. finding documents with similar content
- A popular technique for this task is Latent Semantic Analysis (LSA)

LSA

- LSA computes a matrix between documents and the terms/concepts

	doc1	doc2	doc3
house	20	10	2
price	3	4	1
dog	7	88	9
tree	12	22	6

LSA

- Based on this matrix, different computations are possible
  - Comparing documents (data clustering, document classification)
  - Find similar documents across languages (cross language retrieval)
  - Find relations between terms (synonymy, polysemy)
  - Given a query of terms, find matching documents (information retrieval).

Criticism of SLMs

- Criticism of SLMs coming from linguistics is mainly based on
  - SLMs do not explicitly capture the competence–performance distinction introduced by Noam Chomsky: they do not model linguistic knowledge as such
  - SLMs do not capture long range dependencies (e.g. wh-movement)