Statistical approaches to NLP

Basic Terminology

Roberto Basili

Department of Computer Science, System and Production
University of Roma, *Tor Vergata*
Via Della Ricerca Scientifica s.n.c., 00133, Roma, ITALY

e-mail: `basili@info.uniroma2.it`

Contenuti:

- Probabilita'
- Probabililta' Condizionata
- Statistica Bayesiana
- IT Elementare

Probability and language models :

- Language Models: predict next word on basis of knowledge or assumptions

- Probability theory developed as a way of reasoning about uncertainty and (especially) games of chance

- Strategy: conceptualise language as a game of chance, then use probability theory

Statistical Thinking :

- Language Models: predict the next word assumptions
- Applications: act on limited information
- Linguistics 1 : prefer economical theories
- Linguistics 2 : explain natural language phenomena
- Linguistics 1 : explain language acquisition

Events, Trials and Outcomes :

- We sometimes say that an event (e.g. heads), is the outcome of a trial (e.g Tossing a coin)
- Now imagine a series of trials (repeatedly tossing a coin)
- Plausible that the outcome of trial $n + 1$ will be unaffected by that of trial n

Random variables :

- Random variables formalise the idea of a trial (e.g. W_n for the n-th word in a string)
- Random variables represent what you know about the trial before you have seen its outcome
- After the trial, you have an outcome (e.g. W_n=*dog* or $W_n = w_k$)

Notation for probabilities :

- $P(X = x_i)$, a probability (i.e. number) for x_i
- $P(x_i)$ an abbreviation for the above
- $P(X)$, or either of above to mean the function that assigns a value to each x_i
- $|X = x_i|$ for the number of times x_i occurs
- Can define $P(X = x_i)$ as $\frac{|X=x_i|}{\sum_j |X=x_j|}$ if large number of trials

Conditional probabilities:

- Conan-Doyle does not play dice:
  Holmes follows every use of *Sherlock*

- The formal statement of this fact is that the conditional probability of the n-th word being *Holmes* if the $(n-1)$-th is *Sherlock* appears to be 1 for ConanDoyle stories

- $P(W_n = holmes | W_{n-1} = sherlock) = 1$

Joint events:

- Joint event of the $(n-1)$-th word being "*Sherlock*" and the n-th "*Holmes*":
  $P(W_{n-1} = sherlock, W_n = holmes)$

- Know identity of the next word when we have seen the *Sherlock*, so
  $P(W_n = holmes, W_{n-1} = sherlock) = P(W_{n-1} = sherlock)$

- $P(W_n = holmes|W_{n-1} = sherlock) = 1$

Decomposing joint events:

In general, for any pair of words w_k , w'_k , we will have :

- $P(W_n = w'_k, W_{n+1} = w_k) =$
  $P(W_n = w_k)P(W_{n+1} = w'_k | W_n = w_k)$

  which is usually written more compactly in a form like:

- $P(w_n^k, w_{n+1}^{k'}) = P(w_n^k)P(w_{n+1}^{k'} | w_n^k)$

Pitfalls:

- $P(w_n^k, w_{n+1}^{k'}) \neq P(w_{n+1}^{k'}, w_n^k)$ order different

- $P(w_n^k | w_{n+1}^{k'}) \neq P(w_{n+1}^k | w_n^{k'})$ order same, but given is different

Just because *Holmes* is the only word that follows *Sherlock*, it need not be that *Holmes* is always preceded by *Sherlock*, e.g. *Mr. Holmes*

Bayes' theorem:

- $P(w_n^k, w_{n+1}^{k'}) = P(w_n^k)P(w_{n+1}^{k'}|w_n^k) = P(w_{n+1}^{k'})P(w_n^k|w_{n+1}^{k'})$

  Divide through by $P(w_n^k)$

- $P(w_{n+1}^{k'}|w_n^k) = P(w_{n+1}^{k'})\frac{P(w_n^k|w_{n+1}^{k'})}{P(w_n^k)}$

- Which is an instance of Bayes' theorem

  $P(A|B) = P(A)\frac{P(B|A)}{P(B)}$

Medical diagnosis :

- The doctors problem: $P(S,C)$ vs $P(S,P)$

- Causal Information: $P(S|C) = P(S|P) = 1$

- Base Rates: $P(P) = 10^{-6}$, $P(P) = 0.25$, $P(S) = 0.33$

- Wants: $P(C|S)$ and $P(P|S)$

Medical diagnosis: Bayes' rule applied :

- $P(P|S) = \frac{P(P)P(S|P))}{P(S)}$

- In this case $(10^{-6} \cdot 1)/0.33 = 3 \cdot 10^{-6}$

- In an epidemic the prior might change dramatically, affecting the outcome

- The prior dominates the posterior

Bayesian inference:

- Posterior $= \alpha$ Prior $\times$ Likelihood

- $P(L|X) = \alpha \quad P(L) \times P(X|L)$

- Grammar inference $=$
  Prior beliefs about possible grammars $\times$ Language model