ACS Introduction to NLP

Lecture 2: Part of Speech (POS) Tagging

UNIVERSITY OF CAMBRIDGE

Stephen Clark

Natural Language and Information Processing (NLIP) Group

sc609@cam.ac.uk

The POS Tagging Problem

England|NNP ’s|POS fencers|NNS won|VBD gold|NN on|IN day|NN 4|CD in|IN Delhi|NNP with|IN a|DT medal|JJ -winning|JJ performance|NN .|.

This|DT is|VBZ Dr.|NNP Black|NNP ’s|POS second|JJ gold|NN of|IN the|DT Games|NNP .|.

- Problem is difficult because of ambiguity

Part-of-Speech (POS) Tagging

- Task: given a set of POS tags and a sentence, assign a POS tag to each word
- What knowledge is required and where does it come from?
  - tag dictionary plus contextual statistical models
  - dictionary and probabilities are obtained from labelled data
- What's the algorithm for assigning the tags?
  - the Viterbi algorithm for labelled sequences

Probabilistic Formulation

$$y^* = \arg\max_{y \in Y} P(y|x)$$

where $x = (x_1, \ldots, x_n)$ is a sentence and $y = (y_1, \ldots, y_n) \in Y$ is a possible tag sequence for x

- Two problems:
  - where do the probabilities come from? (age-old question in statistical approaches to AI)
  - how do we find the arg max?
- Problem 1 is the problem of *model estimation*
- Problem 2 is the *search problem*

Statistical Methods in NLP (and AI)

- In 1990 less than 5% of papers at an ACL conference used statistical methods
- Now it's more like 95%
- How did this *paradigm change* come about?

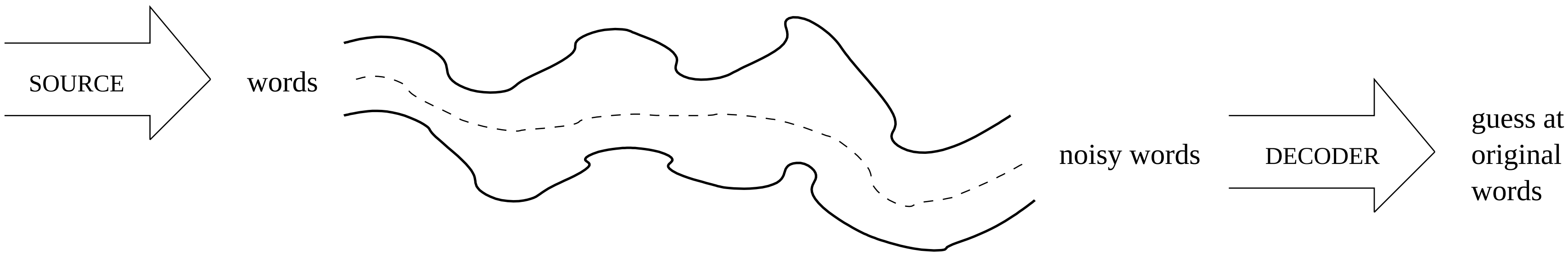
Our Statistical NLP Hero

- Fred Jelinek (1932 - 2010)

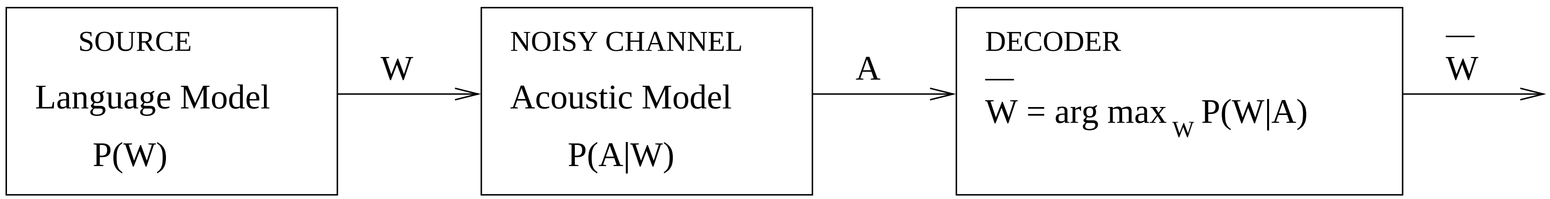
An Historical Aside

- Speech recognition
  - originally used a rule-based approach based on linguistic expertise
  - work in the 70s at IBM showed that a *data-driven* approach worked much better
- Statistical MT
  - IBM applied similar statistical models to translation in the early 90s
  - initially a lot of scepticism and resistance, but now the dominant approach (and used by Google)

Noisy Channel Model



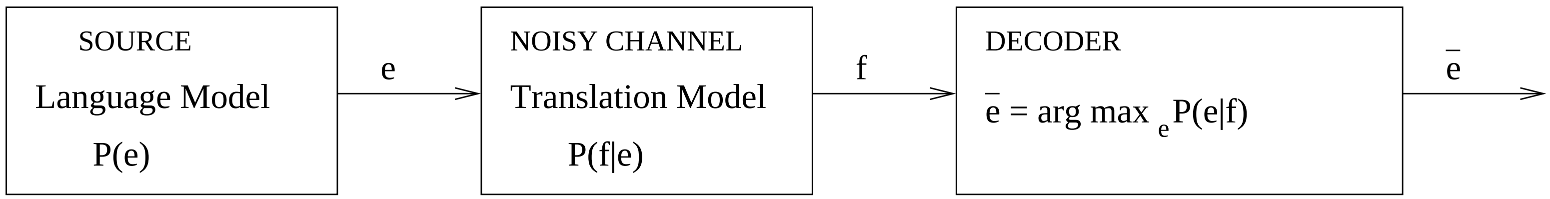
Speech Recognition as a Noisy Channel



- Speaker has word sequence W
- W is articulated as acoustic sequence A
- This process introduces noise:
  - variation in pronunciation
  - acoustic variation due to microphone etc.
- Bayes theorem gives us:

$$
\begin{aligned}
\overline{W} &= \arg\max_W P(W|A) \\
&= \arg\max_W \underbrace{P(A|W)}_{likelihood} \underbrace{P(W)}_{prior}
\end{aligned}
$$

Machine Translation as a Noisy Channel



- Translating French sentence (f) to English sentence (e)
- French speaker has English sentence in mind ($P(e)$)
- English sentence comes out as French via the noisy channel ($P(f|e)$)

- Can use the same mathematics of the noisy channel to model the POS tagging problem
- Breaking the problem into two parts makes the modelling easier
  - can focus on tag transition and word probabilities separately
  - allows convenient independence assumptions to be made

$$\begin{aligned}\overline{T} &= \arg\max_T P(T|W) \\ &= \arg\max_T P(W|T)P(T)\end{aligned}$$

Hidden Markov Models (HMMs)

- $P(T|W) = \frac{P(W|T)P(T)}{P(W)}$ (Bayes theorem)
- $\arg\max_T P(T|W) = \arg\max_T P(W|T)P(T)$ (W is constant)
- Using Chain Rule and (Markov) independence assumptions:

$$
\begin{aligned}
P(W|T) &= P(w_1, \ldots, w_n | t_1, \ldots, t_n) \\
&= P(w_1|t_1, \ldots, t_n)P(w_2|w_1, t_1, \ldots, t_n)P(w_3|w_2, w_1, t_1, \ldots, t_n) \\
&= P(w_n|w_{n-1}, \ldots, w_1, t_1, \ldots, t_n) \\
&\approx \prod_{i=1}^{n} P(w_i|t_i)
\end{aligned}
$$

$$
\begin{aligned}
P(T) &= P(t_1, \ldots, t_n) \\
&= P(t_1)P(t_2|t_1)P(t_3|t_2, t_1) \ldots P(t_n|t_{n-1}, \ldots, t_1) \\
&\approx \prod_{i=1}^{n} P(t_i|t_{i-1})
\end{aligned}
$$

- A tagger which conditions on the previous tag is called a **bigram** tagger
- Trigram taggers are typically used (condition on previous 2 tags)
- HMM taggers use a **generative** model, so-called because the tags *and* words can be thought of as being generated according to some stochastic process
- More sophisticated **discriminative** models (e.g. CRFs) can condition on more aspects of the context, e.g. suffix information

Parameter Estimation

- Two sets of parameters:
  - $P(t_i|t_{i-1})$ tag transition probabilities
  - $P(w_i|t_i)$ word emission probabilities
- Note *not* $P(t_i|w_i)$ (reversed because of use of Bayes theorem)
  - one of the original papers on stochastic POS tagging reportedly got this wrong
- Estimation based on counting from manually labelled corpora
  - so we have a *supervised* machine learning approach
- For this problem, simple counting (relative frequency) method gives *maximum likelihood* estimates

Relative Frequency Estimation

- $$\hat{P}(t_i|t_{i-1}) = \frac{f(t_{i-1},t_i)}{f(t_{i-1})}$$
  - where $f(t_{i-1}, t_i)$ is the number of times t_i follows t_{i-1} in the training data; and $f(t_{i-1})$ is the number of times t_{i-1} appears in the data
- $$\hat{P}(w_i|t_i) = \frac{f(w_i,t_i)}{f(t_i)}$$
  - where $f(w_i, t_i)$ is the number of times w_i has tag t_i in the training data
- It turns out that for an HMM the intuitive relative frequency estimates are the estimates which *maximise the probability of the training data*
- What if the numerator (or denominator) is zero?

Search

- Why is there a search problem?
  - there are an exponential number of tag sequences for a sentence (exponential in the length)
  - finding the highest scoring sequence of tags is complicated by the n-th order Markov assumption ($n>0$)
- More on this next time

Other Models for POS Tagging

- Generative models suffer from the need for restrictive independence assumptions
  - how would you modify the generative process to account for the fact that a word ending in *ing* is likely to be VBG?
- *Discriminative models*, e.g. Conditional Random Fields, are similar to HMMs but model the conditional probability $P(T|W)$ *directly*, rather than via Bayes and a generative story

Reading for Today’s Lecture

- Jurafsky and Martin, Speech and Language Processing, Chapter on Word Classes and Part of Speech Tagging
- Manning and Schutze, Foundations of Statistical Natural Language Processing, Chapter on Part of Speech Tagging and also Mathematical Foundations
- Historical: A statistical approach to machine translation, Peter Brown et al., 1990