Other uses of HMMs: Topic extraction

- More fine-grained IR: a text can be about multiple topics, some words of it generated by each
- Train classifiers for individual topics
- Commonly use hierarchical classification, so smaller classifiers can be trained for individual decisions

The federal government and many of the nation's leading telephone companies agreed today to a reshuffling of phone rates that would reduce the bills of tens of millions of less-affluent consumers.

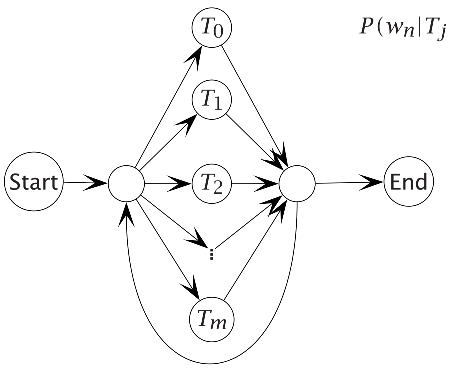
→

telecommunications
U.S. government
FCC
social equity

Generative HMM for topic extraction

- Choose a set of topics $T = \{T_1, \ldots, T_m\}$
- Put a distribution over how many topics are in the set – disfavoring a lot of topics
- For each word in the story, choose a topic according to $P(T_j|T)$
- Choose a word according to $P(w_n|T_j)$
- Proceed to decode by finding out the use of which set of topics would give highest probability to the observed story [often restrict when topic can change. . .]
- BBN: 4600 topics, 76% of first choices in annotated labels (vs. 45% for IR-style tf.idf)

Generative HMM for topic extraction



$P(w_n|T_j)$

Other uses of HMMs: Extracting facts

- *The problem with IR:* You search for "soldiers attacking rebels" and the top matches are:
  - Hutu rebels attacked soldiers in one of Bujumbura's suburbs (Google 2000/10/03)
  - Sudanese rebels say they have killed or wounded more than 300 government soldiers (Hotbot 2000/10/03)
  - [Altavista: a Confederate soldier's memorial!]
- We need to be able to match relationships like:
  - attack(soldiers, rebels)
- Models that see sentence structure, like HMMs or dependency parsers, let us capture these relations, though we still need to deal with synonymy and polysemy

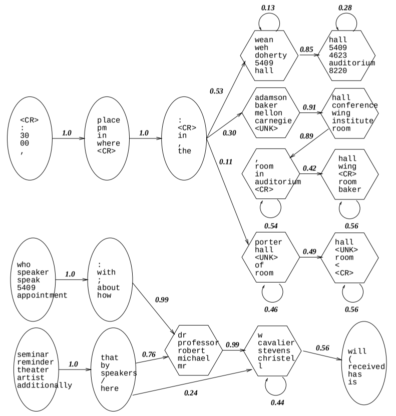
Other uses of HMMs: Information Extraction (Frietag and McCallum 1999)

- IE: extracting instance of a relation from text snippets
- States correspond to fields one wishes to extract, token sequences in the context that are good for identifying the fields to be extracted, and a background "noise" state
- Estimation is from tagged data (perhaps supplemented by EM reestimation over a bigger training set)
- The Viterbi algorithm is used to tag new text
- Things tagged as fields to be extracted are returned

Information Extraction (Frietag and McCallum 1999)

- State topology is set by hand. Not fully connected
- Use simpler and more complex models, but generally:
  - Background state
  - Preceding context state(s)
  - Target state(s)
  - Following context state(s)
- Preceding context states connect only to target state, etc.

Information extraction (2)



Information Extraction (Frietag and McCallum 1999)

- Each HMM is for only one field type (e.g., “speaker”)
- Use different HMMs for each field (*bad:* no real notion of multi-slot structure)
- Semi-supervised training: target words (generated only by target states) are marked
- Shrinkage/deleted interpolation is used to generalize parameter estimates to give more robustness in the face of data sparseness
- Some other work has done multi-field extraction over more structured data (Borkar et al. 2001)

Information Extraction (Frietag and McCallum 1999)

- Tested on seminar announcements and corporate acquisitions data sets
- Performance is generally equal to or better than that of other information extraction methods
- Though probably more suited to semi-structured text with clear semantic sorts, than strongly NLP-oriented problems
- HMMs tend to be especially good for robustness and high recall

HMMs in Computational Biology

- Determining coding/non-coding regions of DNA
- Determining protein-binding sites of DNA
- Aligning and analyzing protein families
- Aligning DNA regions
- Unlike human languages, observation alphabets are normally small, and so clever smoothing/shrinkage methods are less necessary

Identifying G-protein-coupled receptors (Baldi and Chauvin 1994)

- Long but very variable length chains (200–1200) of amino acids
- HMM trained on known examples, with states equal to average length
- Effectively HMM is computing a kind of probabilistic ‘edit distance’, measuring changes, repeats, and deletions
- Model can effectively discriminate GPCR sequences from random or non-GPCR sequences based on calculated log-likelihood of sequence

Gene finding in DNA sequences (Krogh, Mian and Haussler 1994)

- Finds protein-coding genes in *E. coli* DNA.
- HMM states model codons, intergenic regions (incl. palindromic sequences, etc.)
- Allows for insertions and deletions
- HMM parameters estimated from 1 million nucleotides of annotated DNA
- On independent data, model finds exact location of about 80% of known *E. coli* genes, and approximate location of a further 10%.
- It also potentially finds several new genes. . . .