Maximum Entropy Markov Models

Nikos Karampatziakis

September 19th 2006

Background

- Preliminary CRF work. Published in 2000.
- Authors: McCallum, Freitag and Pereira.
- Key concepts: Maximum entropy / “Overlapping” features.
- MEMMs as (non deterministic) probabilistic finite automata: We have to estimate a probability distribution for transitions from a state to other states given an input.

Limitations of HMMs

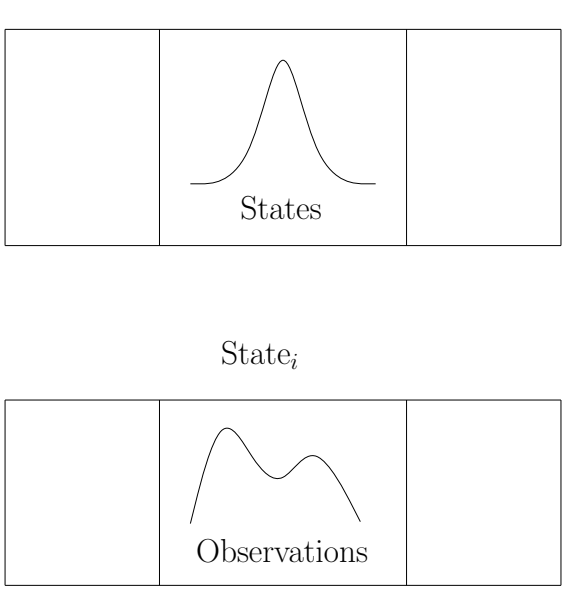
"US official questions regulatory scrutiny of Apple"

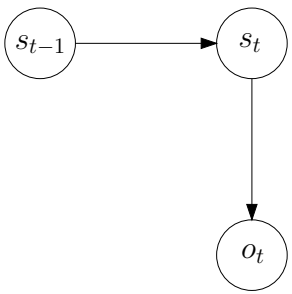
- Problem 1: HMMs only use word identity.
- Cannot use richer representations. Apple is capitalized.
- MEMM Solution: Use more descriptive features (b_0: Is-capitalized, b_1: Is-in-plural, b_2: Has-wordnet-antonym, b_3: Is-"the" etc)
- Real valued features can also be handled.
- Here features are pairs $< b, s >$: b is feature of observation and s is destination state e.g. <Is-capitalized, Company>
- Feature function:
  $$f_{<b,s>}(o_t, s_t) = \begin{cases} 1 & \text{if } b(o_t) \text{ is true and } s = s_t \\ 0 & \text{otherwise} \end{cases}$$
  e.g. $f_{<\text{Is-capitalized,Company}>}(\text{"Apple"}, \text{Company}) = 1$.

HMMs vs. MEMMs (I)

HMMs

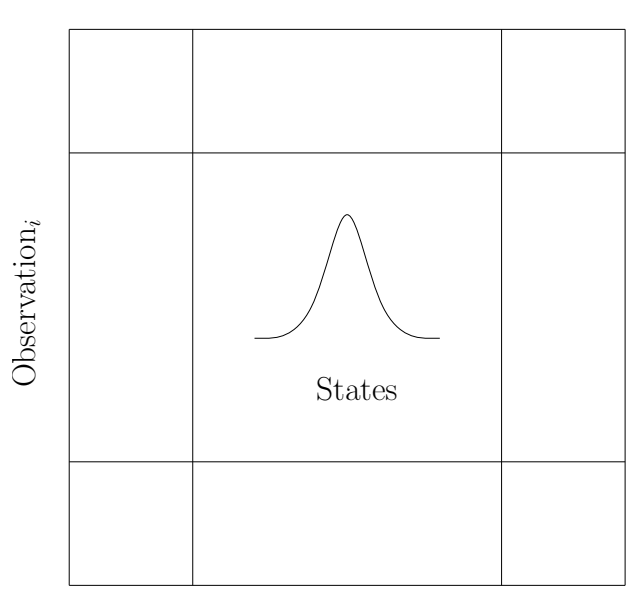
$P(s|s')$, $P(o|s)$

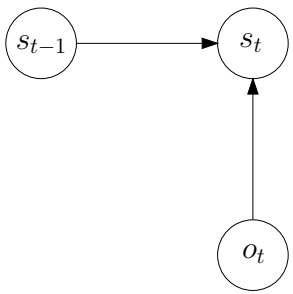




MEMMs

$P(s|s', o) \Leftrightarrow |S|$ distributions: $P_{s'}(s|o)$





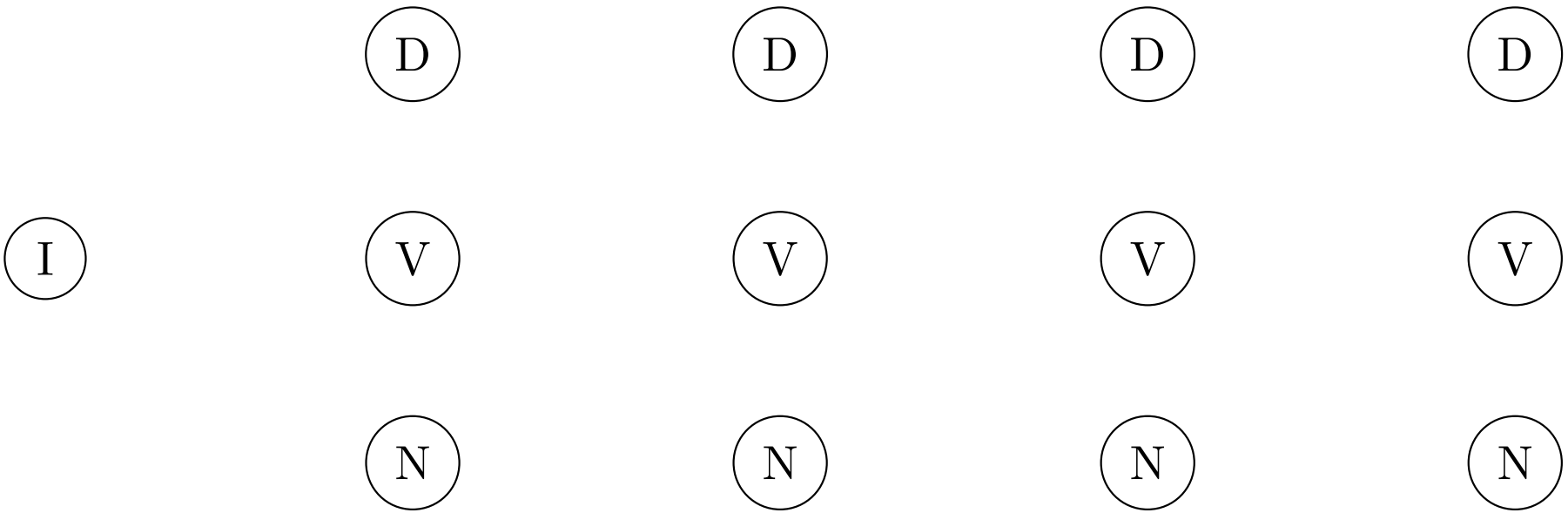
HMMs vs. MEMMs (II)

HMMs	MEMMs
$\alpha_t(s)$ the probability of producing $o_1, \ldots, o_t$ and being in s at time t.	$\alpha_t(s)$ the probability of being in s at time t given $o_1, \ldots, o_t$.
$\alpha_{t+1}(s) = \sum_{s' \in S} \alpha_t(s') P(s\|s') P(o_{t+1}\|s)$	$\alpha_{t+1}(s) = \sum_{s' \in S} \alpha_t(s') P_{s'}(s\|o_{t+1})$
$\delta_t(s)$ the probability of the best path for producing $o_1, \ldots, o_t$ and being in s at time t.	$\delta_t(s)$ the probability of the best path that reaches s at time t given $o_1, \ldots, o_t$.
$\delta_{t+1}(s) = \max_{s' \in S} \delta_t(s') P(s\|s') P(o_{t+1}\|s)$	$\delta_{t+1}(s) = \max_{s' \in S} \delta_t(s') P_{s'}(s\|o_{t+1})$

Viterbi in MEMMs

"Matt saw the cat"

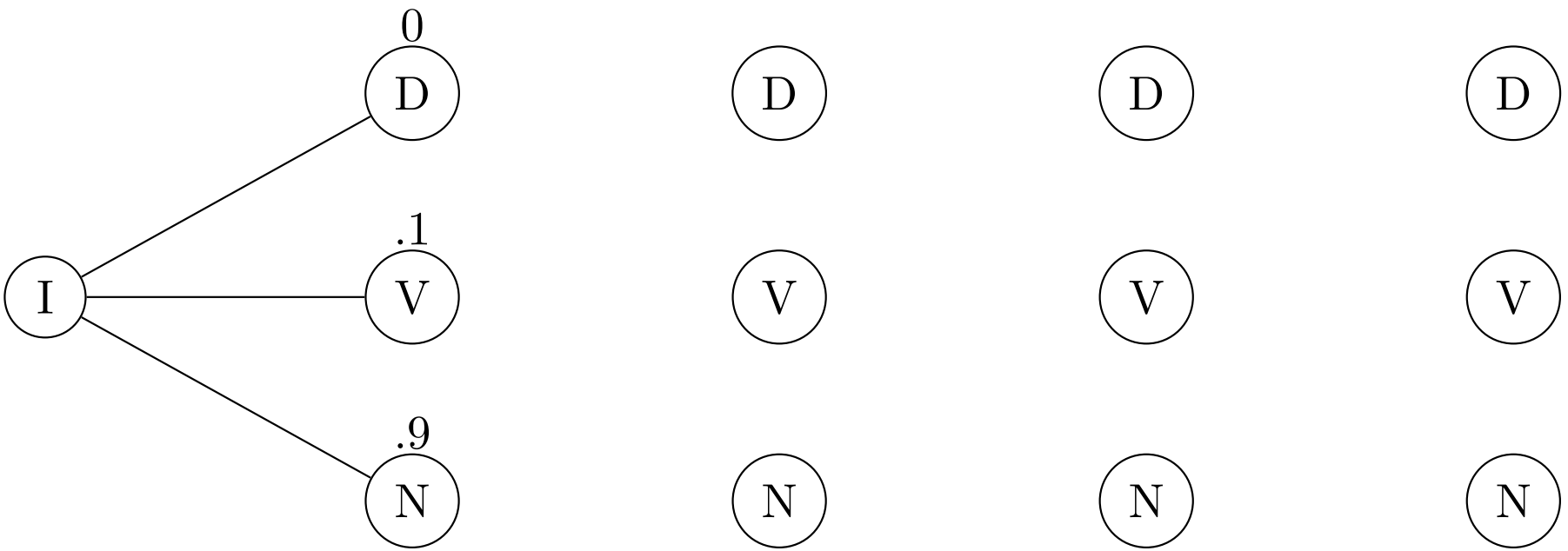
	I or N	V	D
Matt	$p_N = .9, p_V = .1$	$p_N = .8, p_V = .2$	$p_N = .9, p_V = .1$
saw	$p_N = .2, p_V = .8$	$p_N = .7, p_V = .3$	$p_N = 1$
the	$p_D = 1$	$p_D = 1$	$p_D = 1$
cat	$p_N = .9, p_V = .1$	$p_N = .95, p_V = .05$	$p_N = 1$



Viterbi in MEMMs

"*Matt* saw the cat"

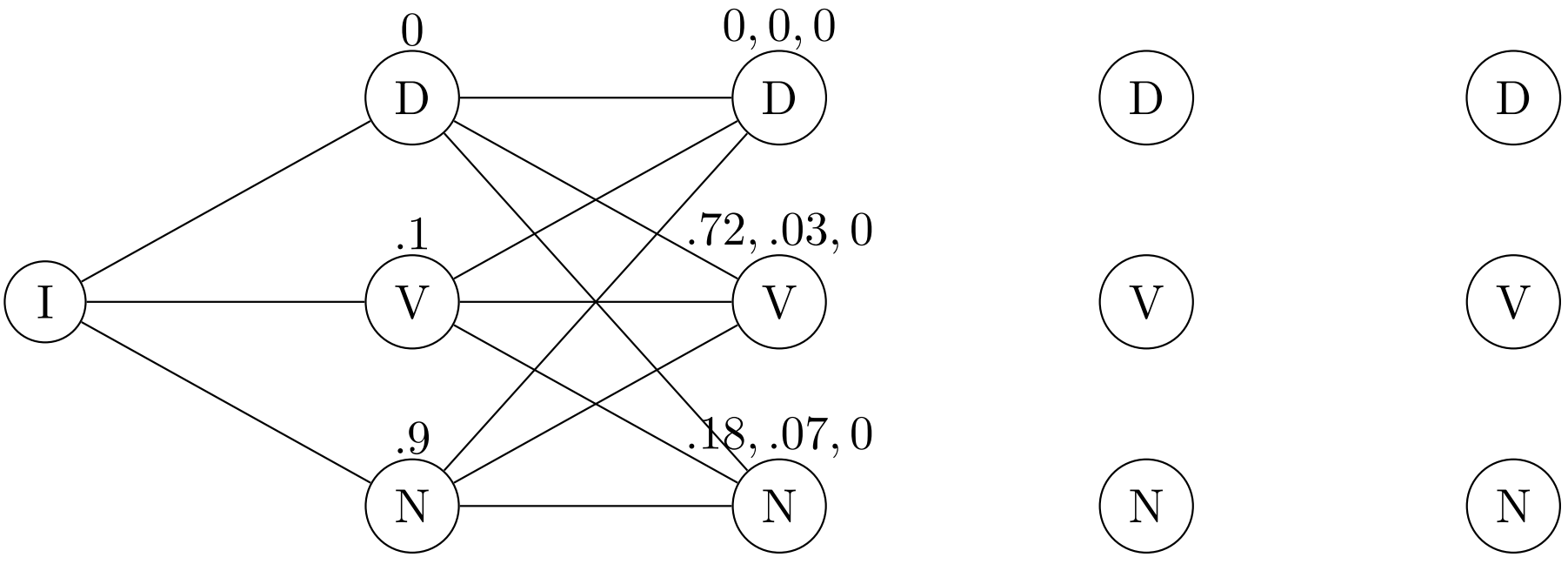
	I or N	V	D
Matt	$p_N = .9, p_V = .1$	$p_N = .8, p_V = .2$	$p_N = .9, p_V = .1$
saw	$p_N = .2, p_V = .8$	$p_N = .7, p_V = .3$	$p_N = 1$
the	$p_D = 1$	$p_D = 1$	$p_D = 1$
cat	$p_N = .9, p_V = .1$	$p_N = .95, p_V = .05$	$p_N = 1$



Viterbi in MEMMs

"Matt *saw* the cat"

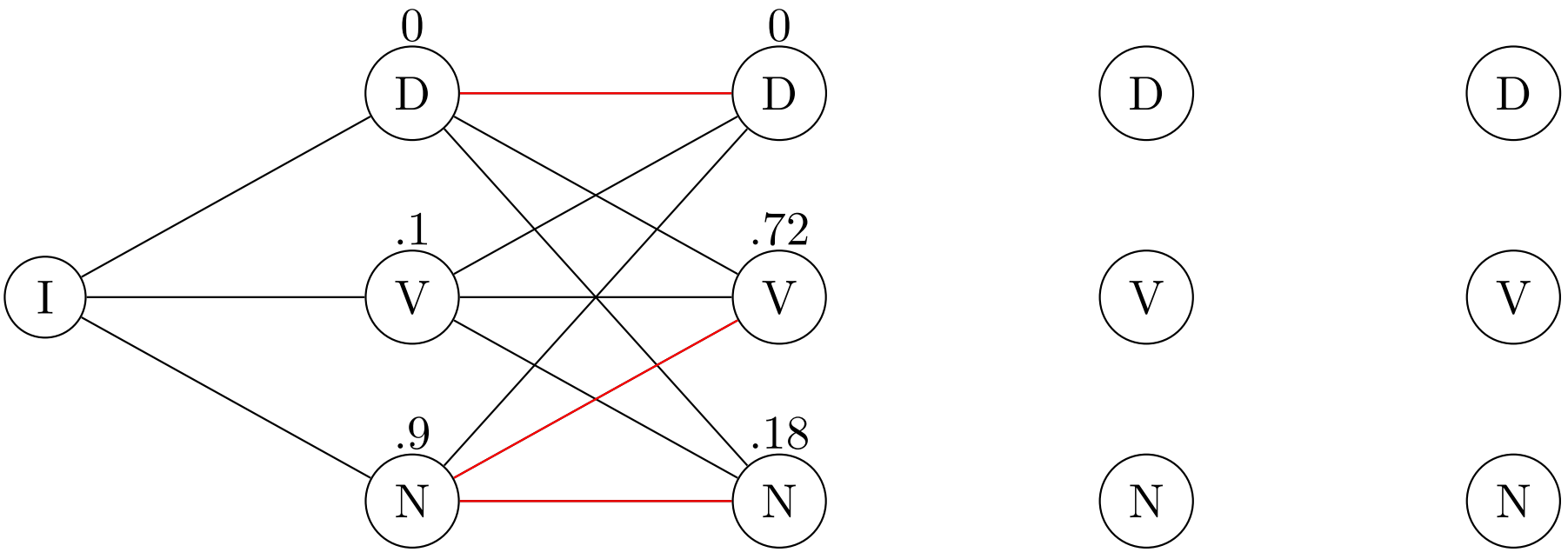
	I or N	V	D
Matt	$p_N = .9, p_V = .1$	$p_N = .8, p_V = .2$	$p_N = .9, p_V = .1$
saw	$p_N = .2, p_V = .8$	$p_N = .7, p_V = .3$	$p_N = 1$
the	$p_D = 1$	$p_D = 1$	$p_D = 1$
cat	$p_N = .9, p_V = .1$	$p_N = .95, p_V = .05$	$p_N = 1$



Viterbi in MEMMs

"Matt *saw* the cat"

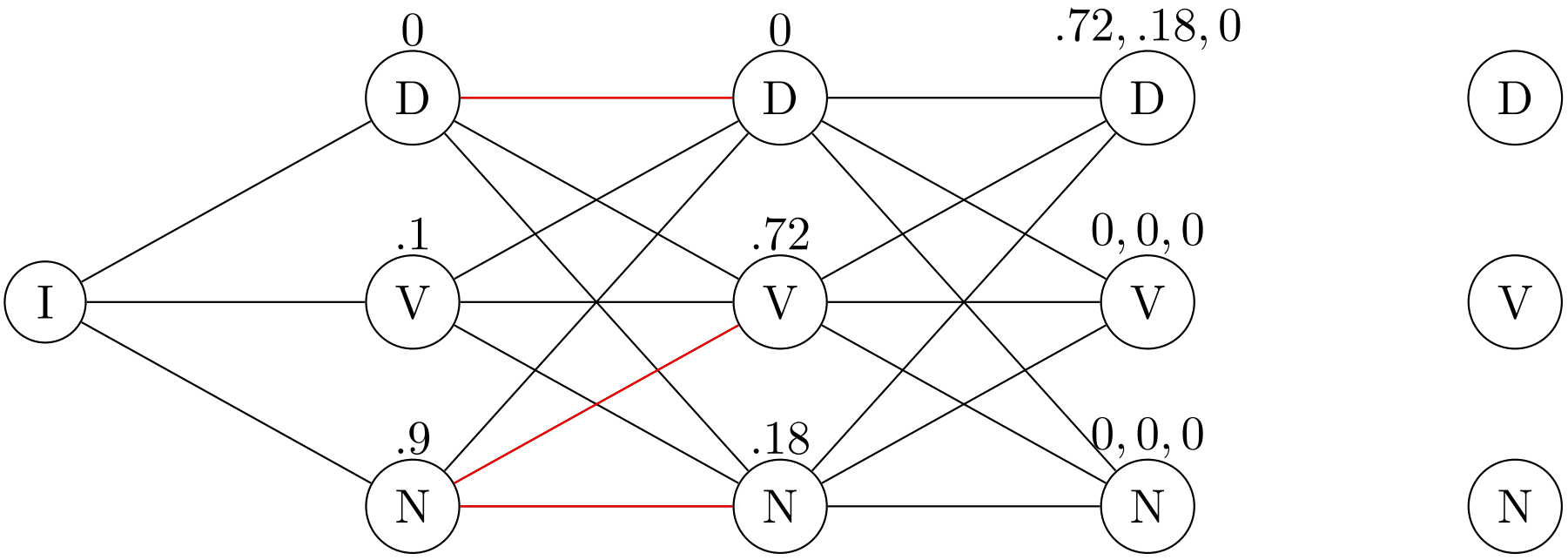
	I or N	V	D
Matt	$p_N = .9, p_V = .1$	$p_N = .8, p_V = .2$	$p_N = .9, p_V = .1$
saw	$p_N = .2, p_V = .8$	$p_N = .7, p_V = .3$	$p_N = 1$
the	$p_D = 1$	$p_D = 1$	$p_D = 1$
cat	$p_N = .9, p_V = .1$	$p_N = .95, p_V = .05$	$p_N = 1$



Viterbi in MEMMs

"Matt saw *the* cat"

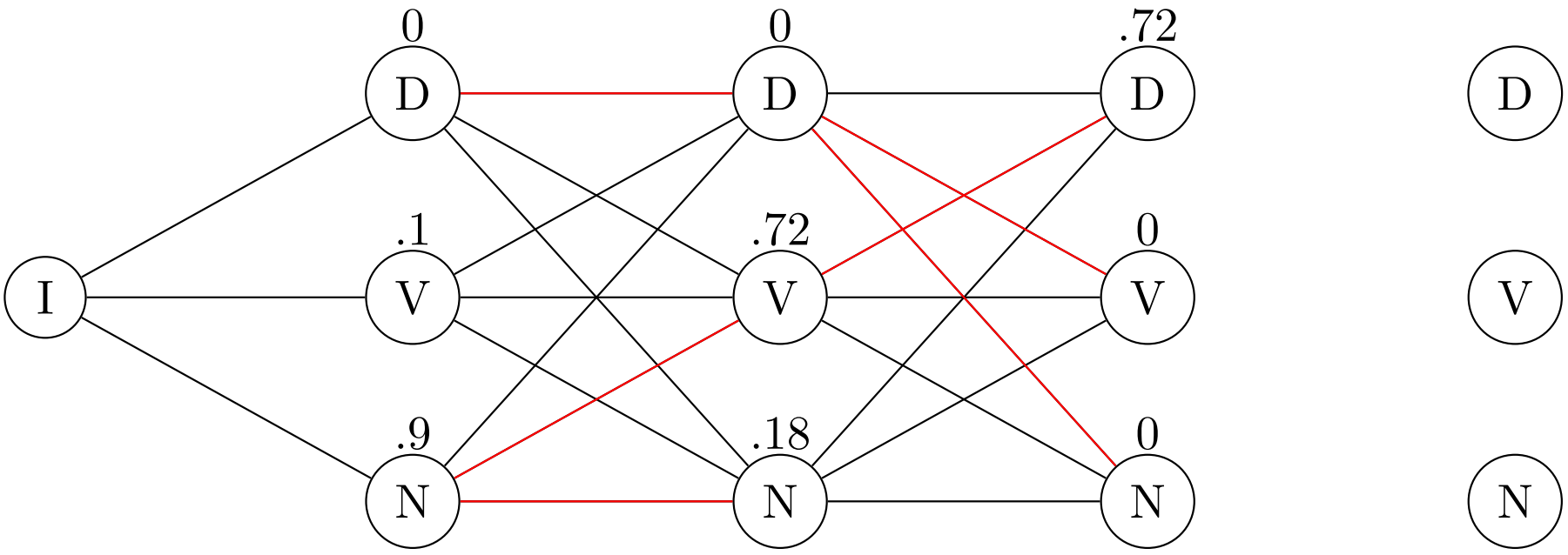
	I or N	V	D
Matt	$p_N = .9, p_V = .1$	$p_N = .8, p_V = .2$	$p_N = .9, p_V = .1$
saw	$p_N = .2, p_V = .8$	$p_N = .7, p_V = .3$	$p_N = 1$
the	$p_D = 1$	$p_D = 1$	$p_D = 1$
cat	$p_N = .9, p_V = .1$	$p_N = .95, p_V = .05$	$p_N = 1$



Viterbi in MEMMs

"Matt saw *the* cat"

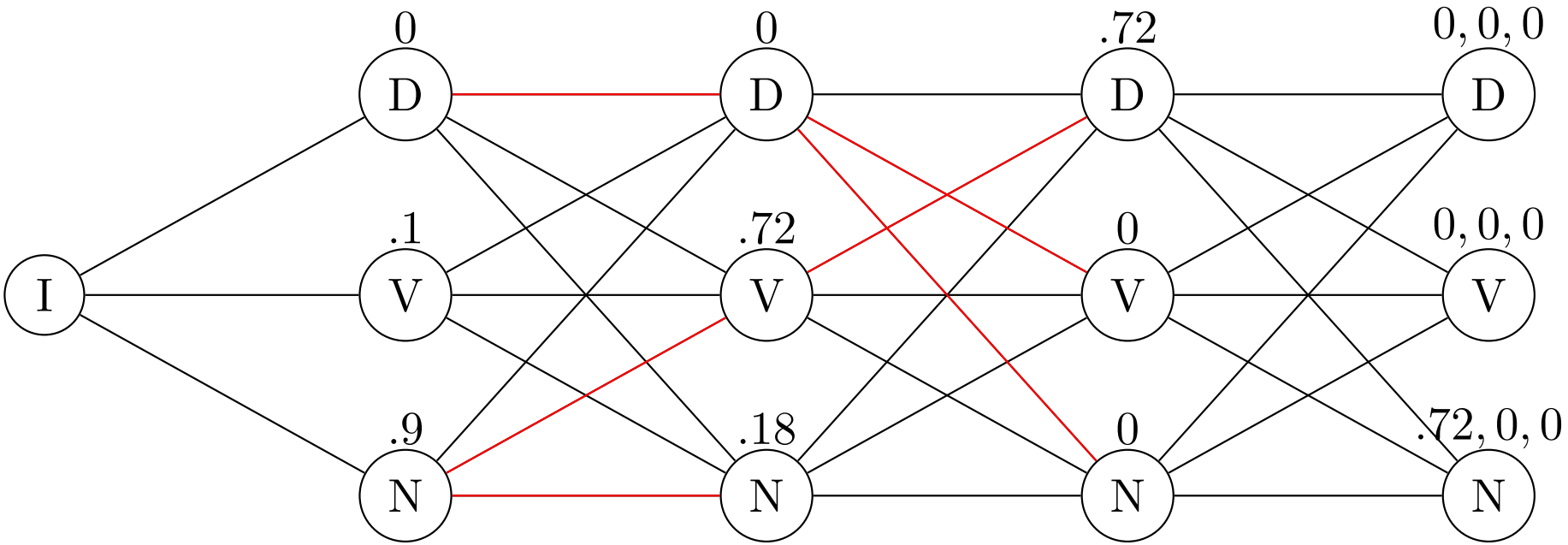
	I or N	V	D
Matt	$p_N = .9, p_V = .1$	$p_N = .8, p_V = .2$	$p_N = .9, p_V = .1$
saw	$p_N = .2, p_V = .8$	$p_N = .7, p_V = .3$	$p_N = 1$
the	$p_D = 1$	$p_D = 1$	$p_D = 1$
cat	$p_N = .9, p_V = .1$	$p_N = .95, p_V = .05$	$p_N = 1$



Viterbi in MEMMs

"Matt saw the *cat*"

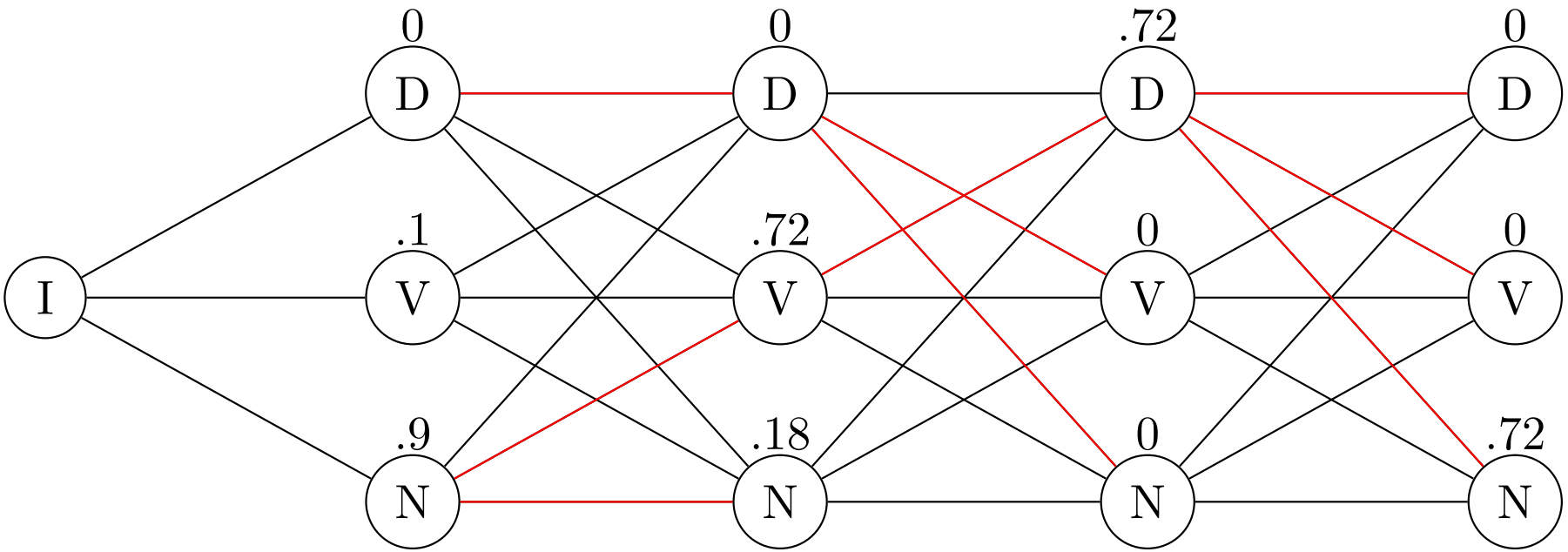
	I or N	V	D
Matt	$p_N = .9, p_V = .1$	$p_N = .8, p_V = .2$	$p_N = .9, p_V = .1$
saw	$p_N = .2, p_V = .8$	$p_N = .7, p_V = .3$	$p_N = 1$
the	$p_D = 1$	$p_D = 1$	$p_D = 1$
cat	$p_N = .9, p_V = .1$	$p_N = .95, p_V = .05$	$p_N = 1$



Viterbi in MEMMs

"Matt saw the *cat*"

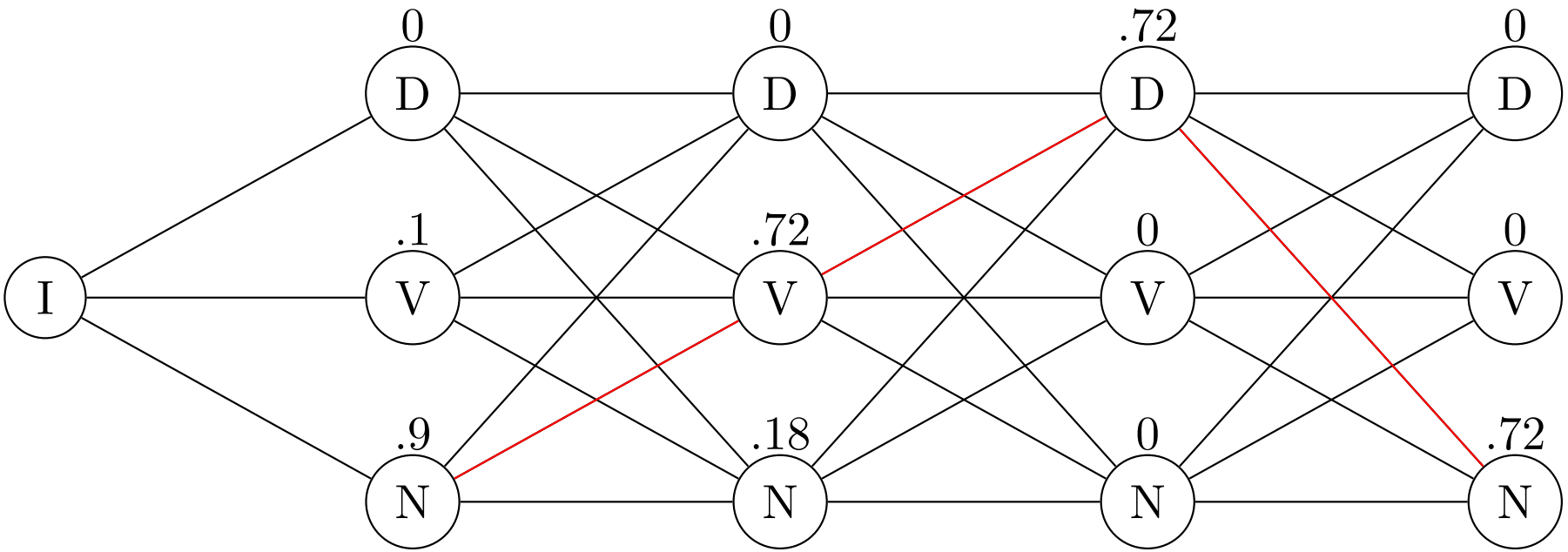
	I or N	V	D
Matt	$p_N = .9, p_V = .1$	$p_N = .8, p_V = .2$	$p_N = .9, p_V = .1$
saw	$p_N = .2, p_V = .8$	$p_N = .7, p_V = .3$	$p_N = 1$
the	$p_D = 1$	$p_D = 1$	$p_D = 1$
cat	$p_N = .9, p_V = .1$	$p_N = .95, p_V = .05$	$p_N = 1$



Viterbi in MEMMs

"Matt saw the cat"

	I or N	V	D
Matt	$p_N = .9, p_V = .1$	$p_N = .8, p_V = .2$	$p_N = .9, p_V = .1$
saw	$p_N = .2, p_V = .8$	$p_N = .7, p_V = .3$	$p_N = 1$
the	$p_D = 1$	$p_D = 1$	$p_D = 1$
cat	$p_N = .9, p_V = .1$	$p_N = .95, p_V = .05$	$p_N = 1$



Maximum Entropy

- Problem 2: HMMs are trained to maximize the likelihood of the training set. Generative, joint distribution.
- But they solve conditional problems (observations are given).
- MEMM Solution: Maximum Entropy (duh).
- Idea: Use the least biased hypothesis, subject to what is known.
- Constraints: The expectation E_i of feature i in the learned distribution should be the same as its mean F_i on the training set. For every state s' and feature i:

$$F_i = \frac{1}{n_{s'}} \sum_{\substack{k=1 \\ s_k = s'}}^{n} f_i(o_k, s')$$

$$E_i = \frac{1}{n_{s'}} \sum_{\substack{k=1 \\ s_k = s'}}^{n} \sum_{s \in S} P_{s'}(s|o_k) f_i(o_k, s)$$

More on MEMMs

- It turns out that the maximum entropy distribution is unique and has an exponential form:

$$P_{s'}(s|o) = \frac{1}{Z(o,s')} \exp\left(\sum_{i \in \text{features}} \lambda_i f_i(o,s)\right)$$

- We can estimate λ_i with Generalized Iterative Scaling.
- Adding a feature $x : f_x(o,s) = C - \sum_i f_i(o,s)$ does not affect the solution.
- Compute F_i. Set $\lambda_i^{(0)} = 0$.
- Compute current expectation $E_i^{(j)}$ of feature i from model.
- $\lambda_i^{(j+1)} = \lambda_i^{(j)} + \frac{1}{C} \log\left(\frac{F_i}{E_i^{(j)}}\right)$

Extensions

- We can train even when the labels are not known using EM.
- E step: determine most probable state sequence and compute F_i. M step: GIS.
- We can reduce the number of parameters to estimate by moving the previous state in the features: "Subject-is-female", "Previous-was-question", "Is-verb-and-no-noun-yet".
- We can even add features regarding actions in a reinforcement learning setting: "Slow-vehicle-encountered-and-steer-left".
- We can mitigate data sparseness problems by simplifying the model:

$$P(s|s', o) = P(s|s')\frac{1}{Z(o, s')}\exp\left(\sum_i \lambda_i f_i(o, s)\right)$$

Experiments

- Task: Label each line in a FAQ with either “H”, “Q”, “A”, “T”.
- Train on one part of a multi part FAQ; test on the rest.
- Formatting consistent across parts but can be different accross FAQs.
- Metrics: Co-occurence agreement probability, Segmentation Precision, Segmentation Recall.
- Models: Maxent Stateless, Token HMM, Feature HMM, MEMM.
- Conclusions: Features help both maximum entropy and maximum likelihood models, states help maxent models, MEMMs outperform other models most notably in Segmentation Precision.