Probabilistic Graphical Models:

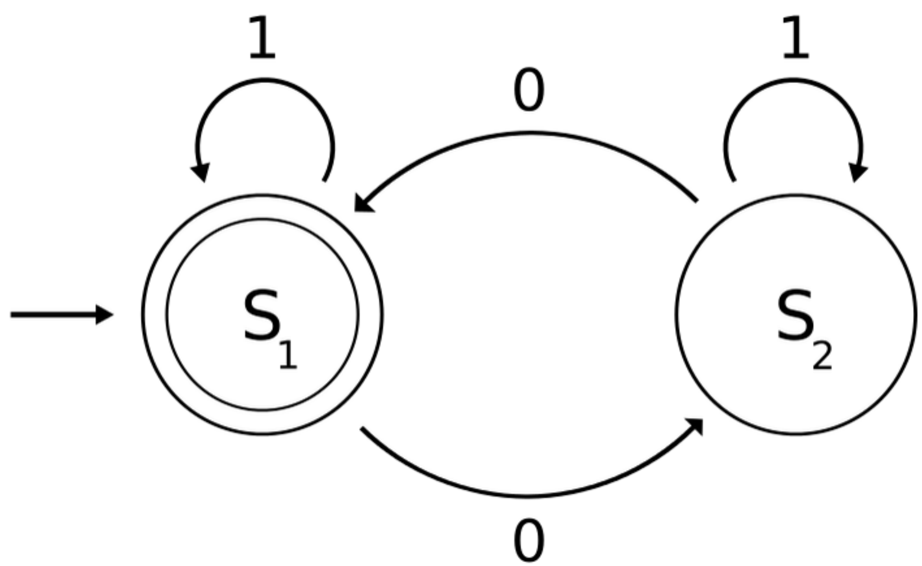
Hidden Markov Models & beyond

Juan Miguel Cejuela

PhD Student in CS & Bioinformatics @ TUM

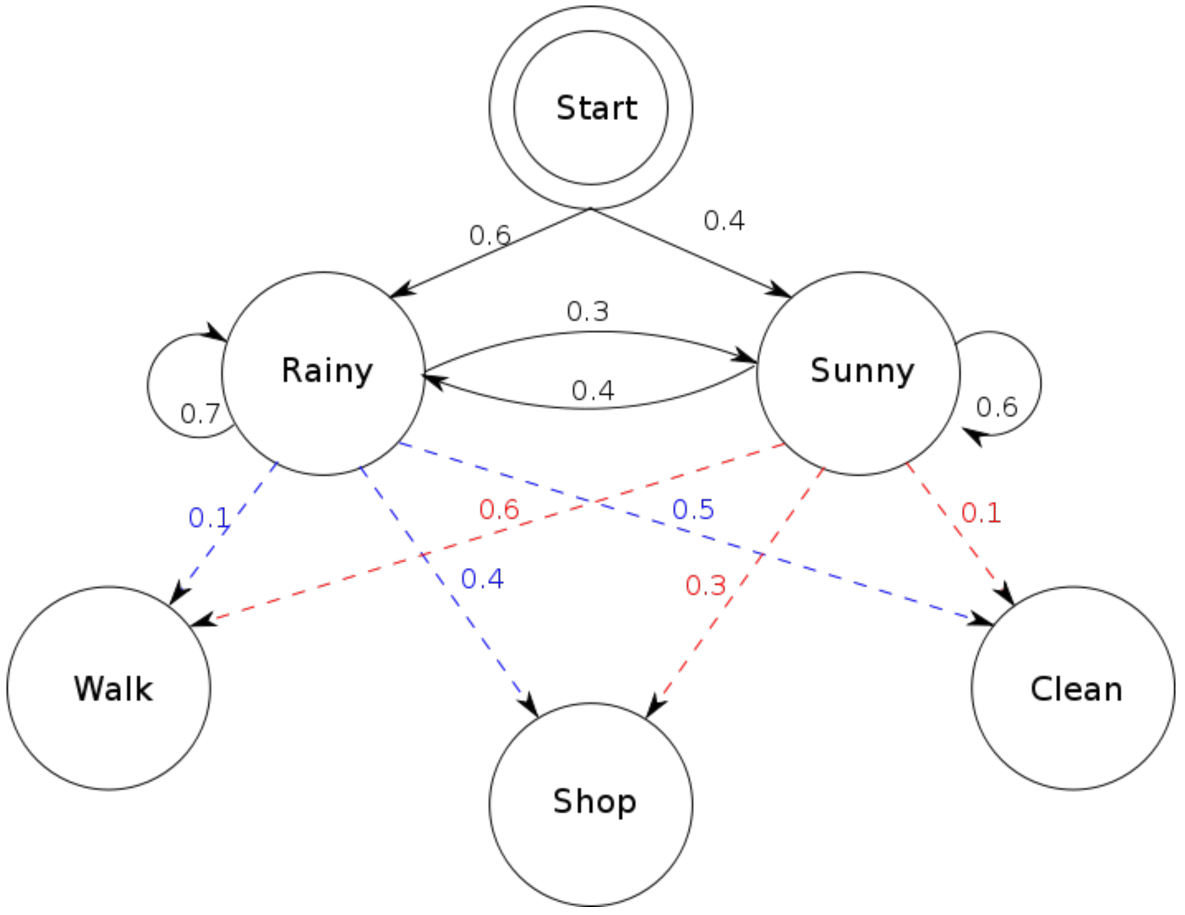
tagtog

(Probabilistic) Graphical Models



Wikipedia, Finite-state machine

Hidden Markov Models (HMM)

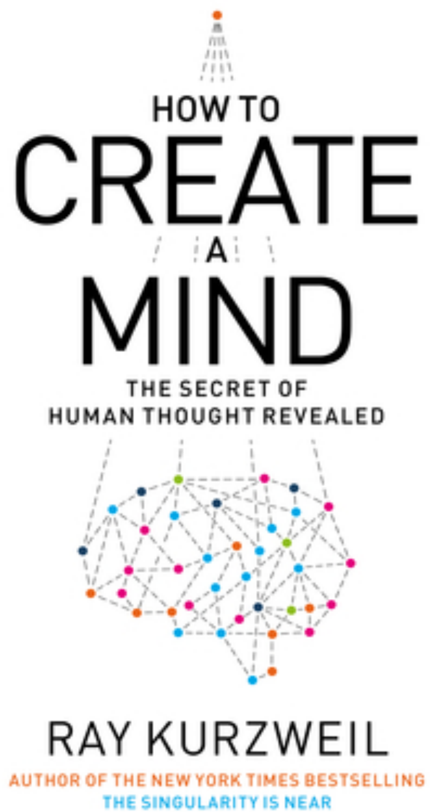


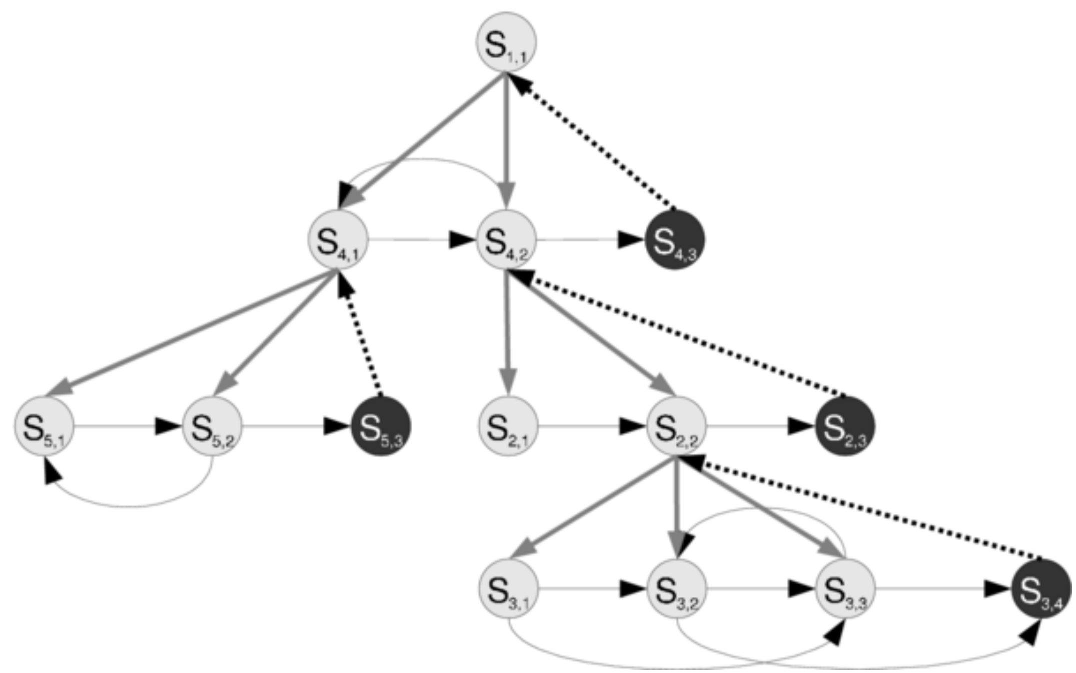
History & Applications

- ~1960s, L. E. Baum & Ruslan L. Stratonovich
- Rabiner 1989 – A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition

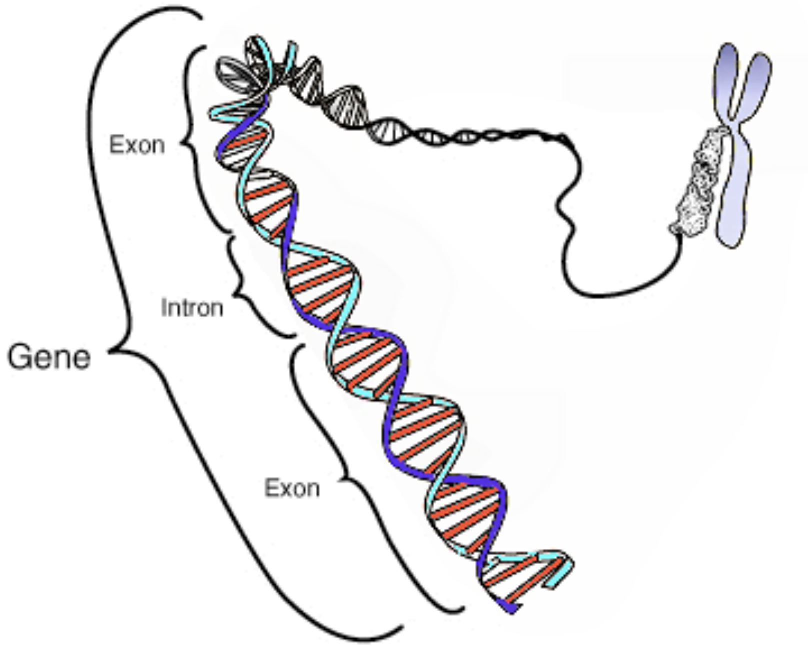
- Sequence labeling
- Speech Recognition
- Gene Finding
- Machine Translation
- POS Tagging
- Protein Sequence Alignment
- …

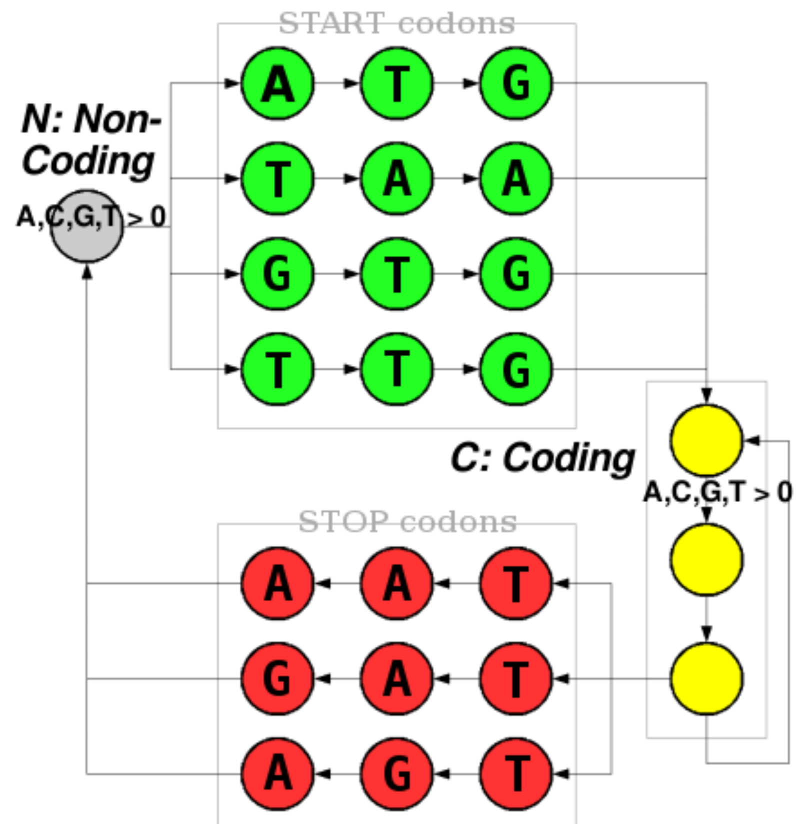
Hierarchical Hidden Markov Models (HHMM)



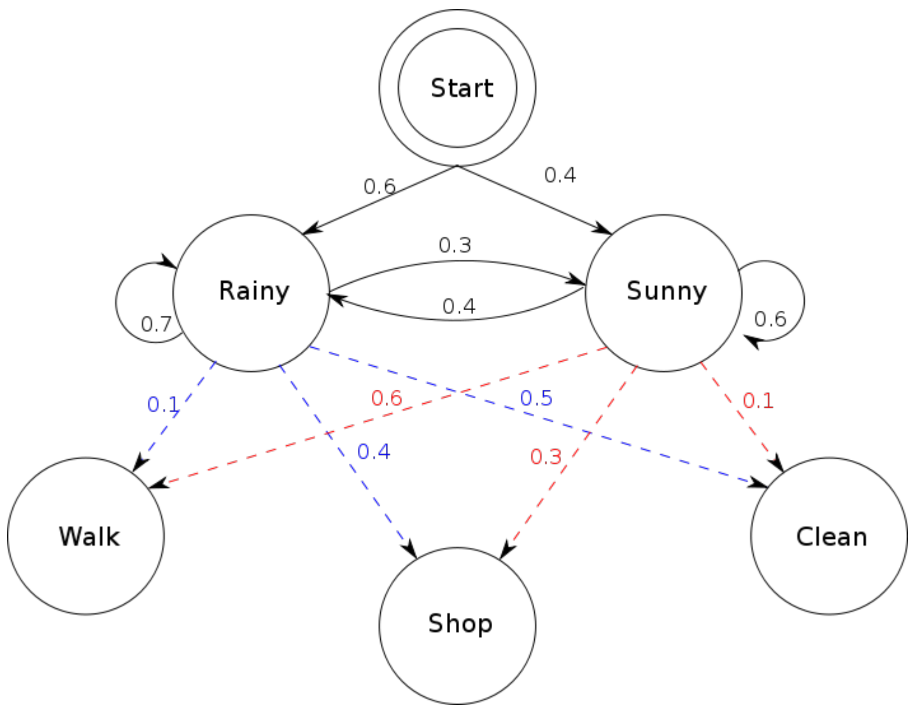


Application Case: Gene Finding





HMM: Definition (I)



- Directed
- Generative
- Markov process

Wikipedia, HMM

HMM: Definition (II)

$S = \{S_1, S_2, \ldots S_N\}$
$O = \{O_1, O_2, \ldots O_M\}$
$\pi = \{\pi_i\}$
$A = \{a_{ij}\}$
$B = \{b_j(k)\}$

$\lambda = \{\pi, A, B\}$

```
states = ('Rainy', 'Sunny')

observations = ('walk', 'shop', 'clean')

start_probability = {'Rainy': 0.6, 'Sunny': 0.4}

transition_probability = {
   'Rainy' : {'Rainy': 0.7, 'Sunny': 0.3},
   'Sunny' : {'Rainy': 0.4, 'Sunny': 0.6},
   }

emission_probability = {
   'Rainy' : {'walk': 0.1, 'shop': 0.4, 'clean': 0.5},
   'Sunny' : {'walk': 0.6, 'shop': 0.3, 'clean': 0.1},
   }
```

Example code from Wikipedia

HMM: 3 problems

1. Scoring, $P(X/\lambda)$
2. Decoding, $\text{argmax}_Y P(Y/X, \lambda)$
3. Training, $\text{argmax}_\lambda \prod P(X_k/\lambda)$

Viterbi: decoding (predict state sequence)

- Find *Y*, path of states that best explains the given observation sequence *X*

- Viterbi: path with highest probability

$$\delta_t(i) = \max_{q_1, q_2, \ldots, q_{t-1}} P[q_1\ q_2\ \ldots\ q_t = S_i, O_1\ O_2\ \ldots\ O_t | \lambda]$$

Initialisation

$$\delta_1(i) = \pi_i b_i(O_1), \quad 1 \le i \le N$$
$$\psi_1(i) = 0$$

Recursion

$$\delta_t(j) = \max_{1 \le i \le N} [\delta_{t-1}(i) a_{ij}] b_j(O_t), \quad 2 \le t \le T$$
$$1 \le j \le N$$

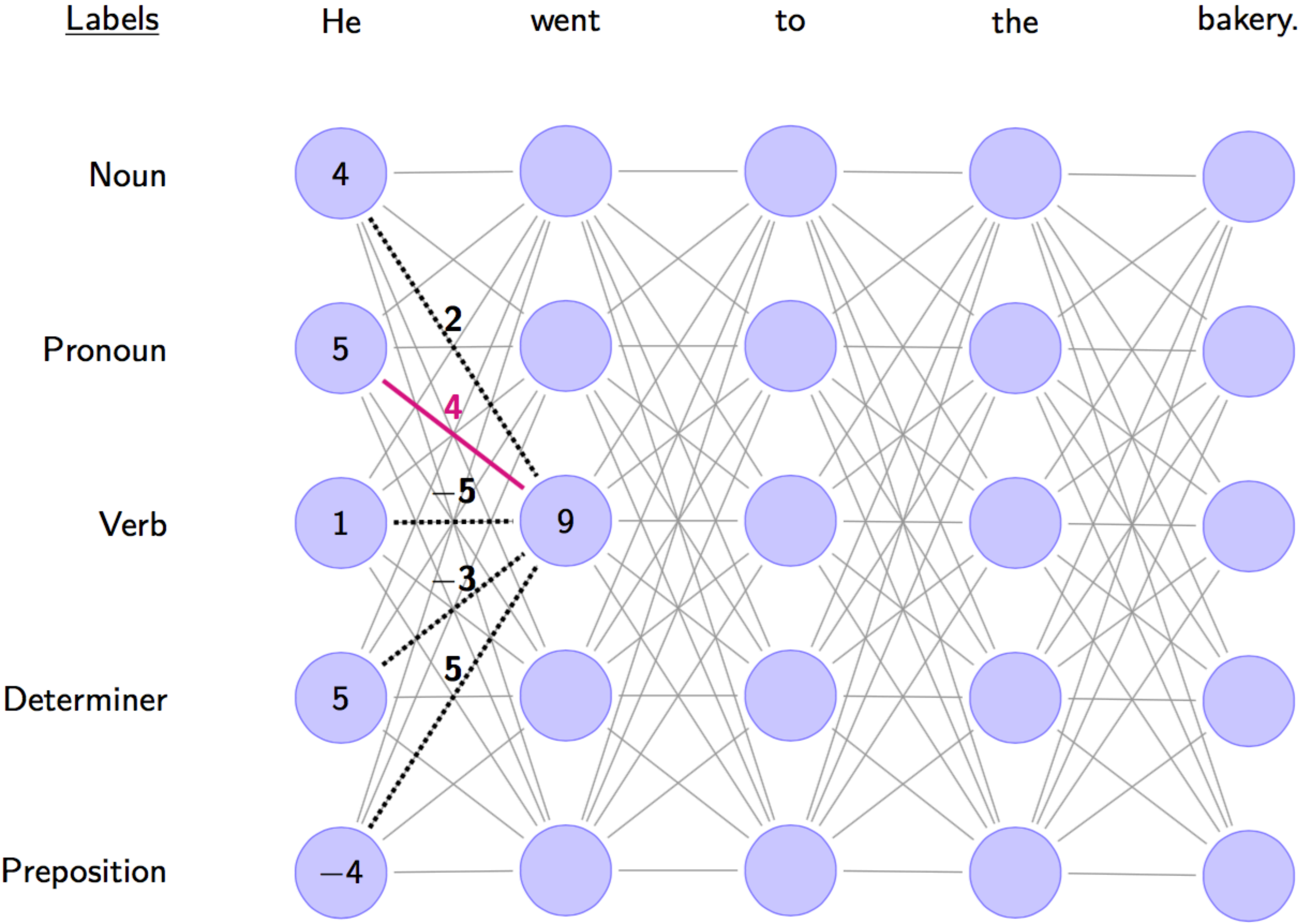
$$\psi_t(j) = \operatorname*{arg\,max}_{1 \le i \le N} [\delta_{t-1}(i) a_{ij}], \quad 2 \le t \le T$$
$$1 \le j \le N$$

Termination

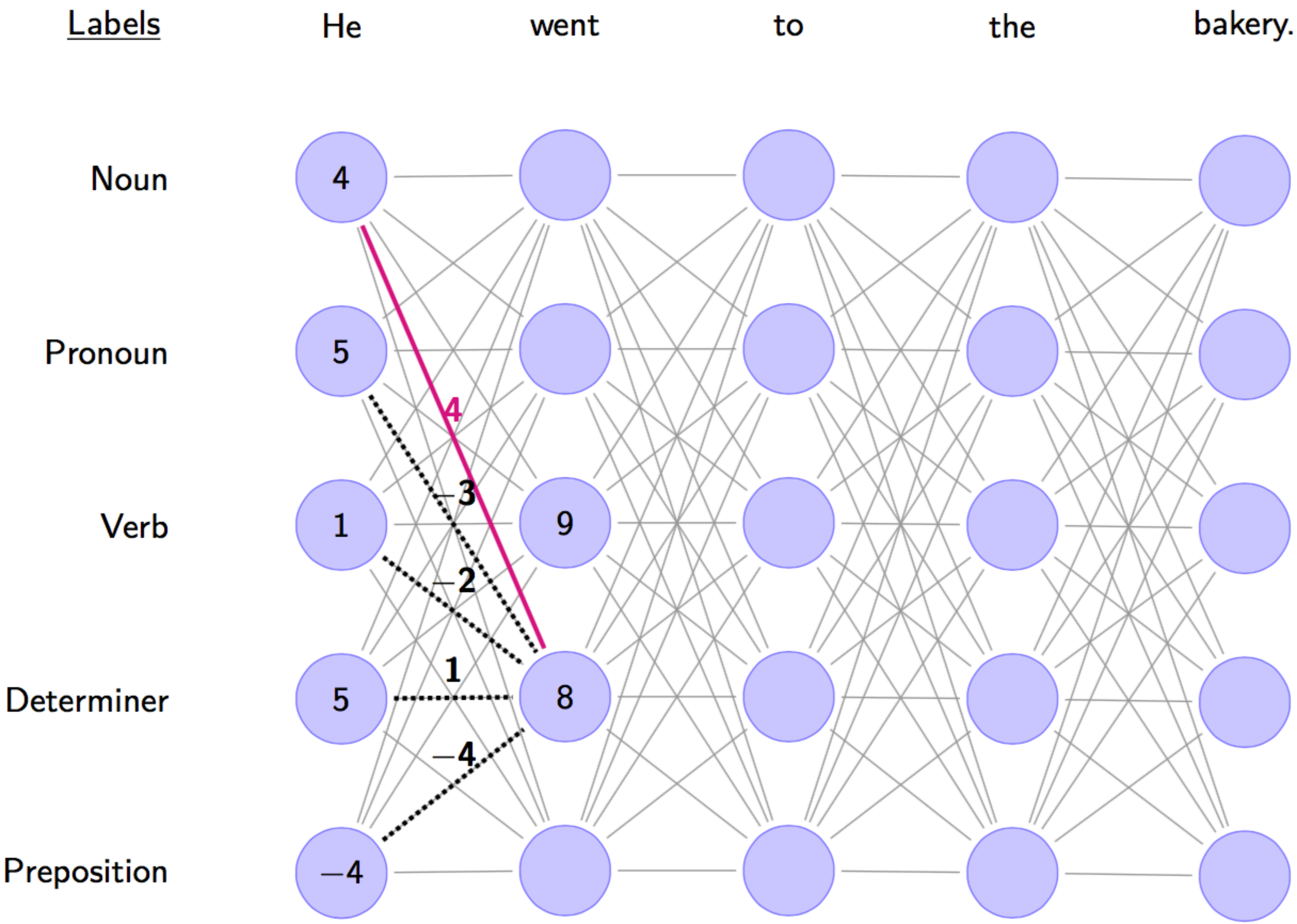
$$P^* = \max_{1 \le i \le N} [\delta_T(i)]$$
$$q_T^* = \operatorname*{arg\,max}_{1 \le i \le N} [\delta_T(i)]$$

Backtracking

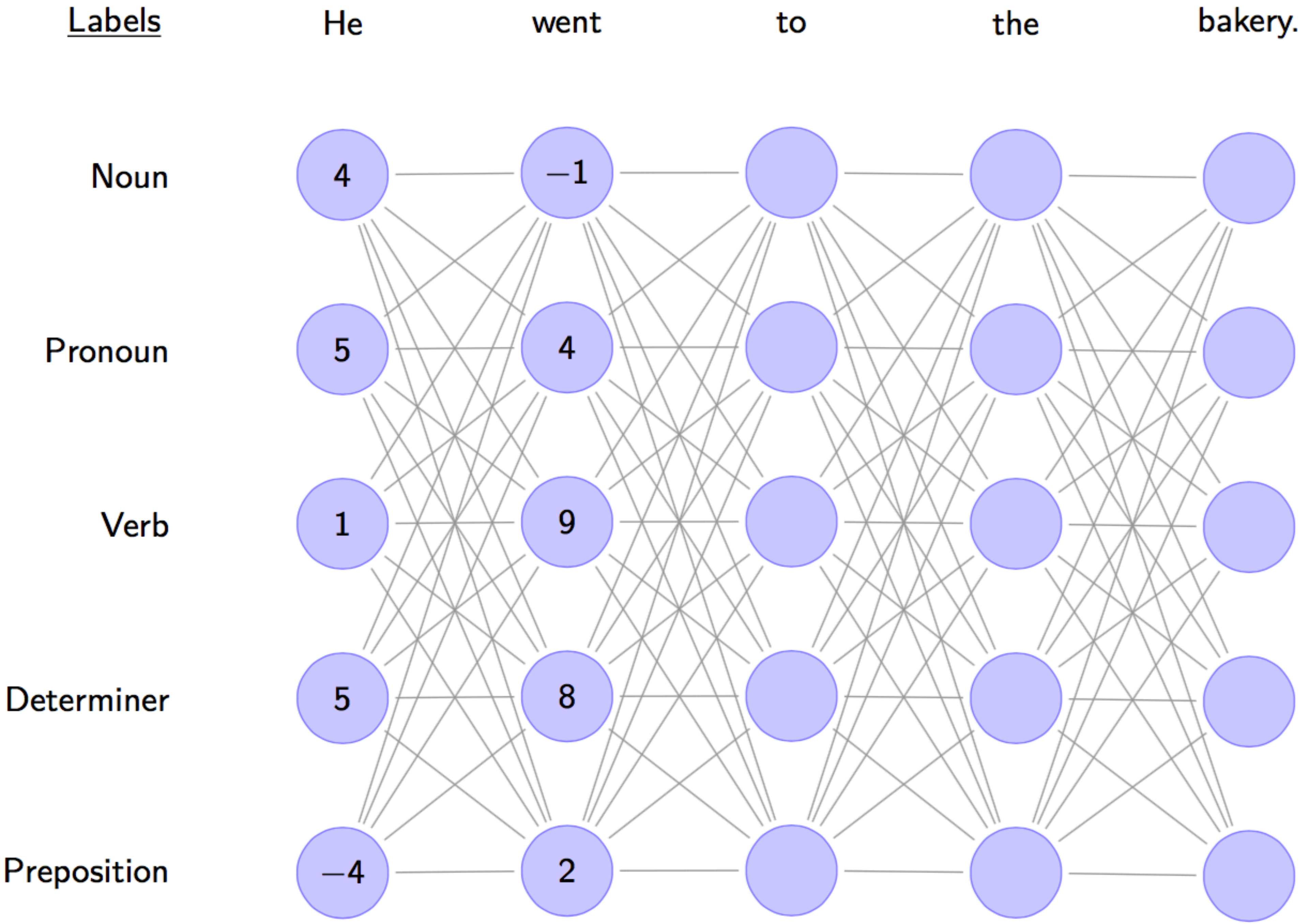
$$q_t^* = \psi_{t+1}(q_{t+1}^*), \quad t = T-1,\ T-2,\ \dots\ ,1$$



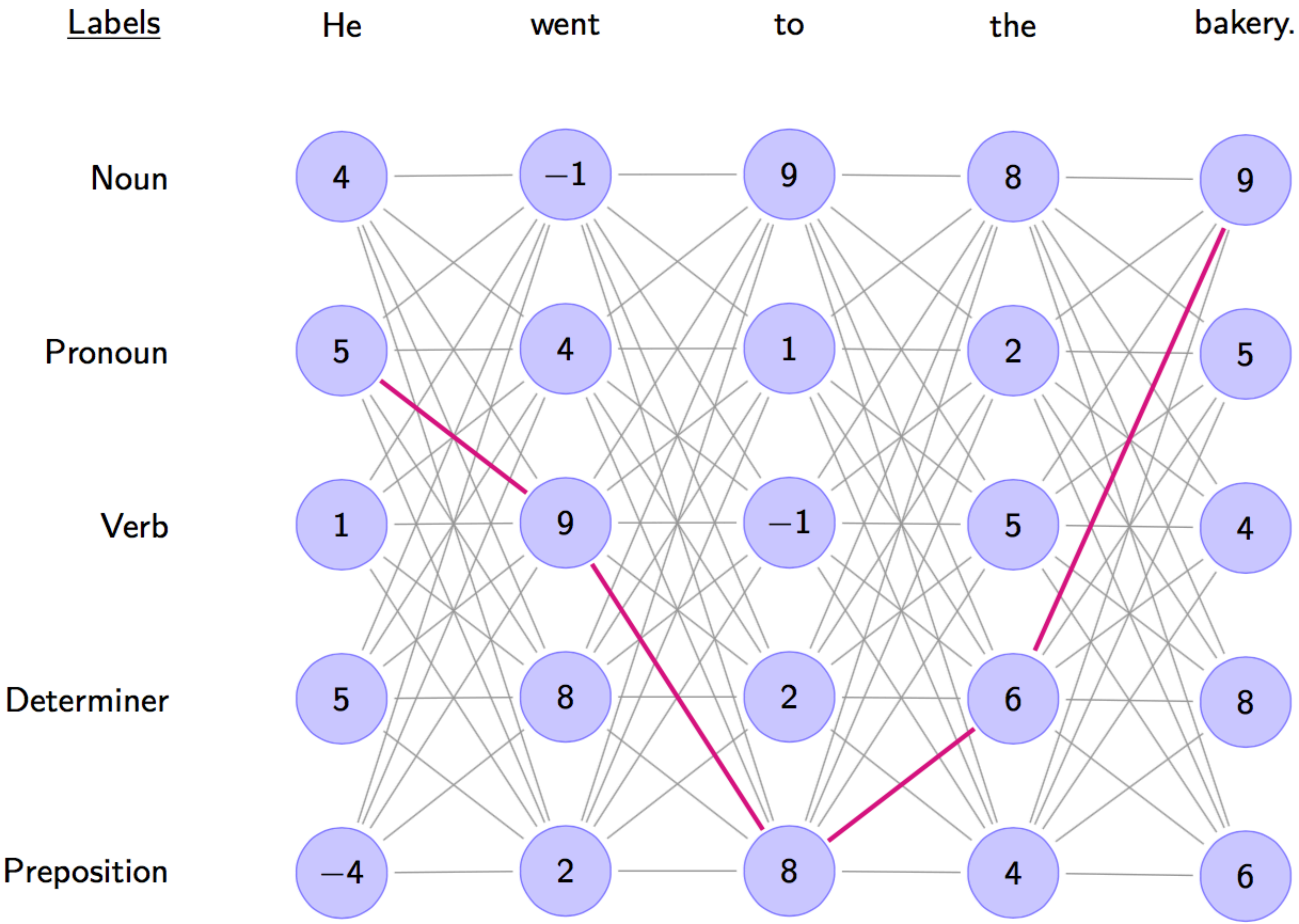
Labels
He
went
to
the
bakery.
Noun
Pronoun
Verb
Determiner
Preposition
4
5
1
5
−4
9
2
4
−5
−3
5



Labels
He
went
to
the
bakery.
Noun
Pronoun
Verb
Determiner
Preposition
4
5
1
5
−4
9
8
4
−3
−2
1
−4



Labels
He
went
to
the
bakery.
Noun
Pronoun
Verb
Determiner
Preposition
4
−1
5
4
1
9
5
8
−4
2


Labels
He
went
to
the
bakery.
Noun
Pronoun
Verb
Determiner
Preposition
4
−1
9
8
9
5
4
1
2
5
1
9
−1
5
4
5
8
2
6
8
−4
2
8
4
6

Training

- $\text{argmax}_{\lambda} \; \prod \; P(X_k/\lambda)$
- Baum-Welch (forward & backward) ~ EM
- Supervised& semi-supervised learning

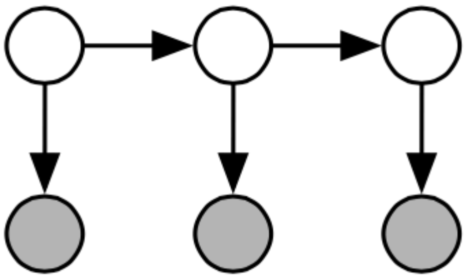
Implementations

- www.ghmm.org (C & Python)

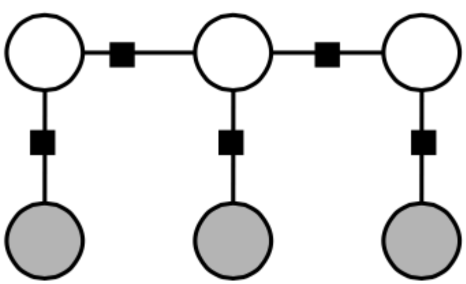
- github.com/jmcejuela/CL-HMM

log space!

Generative vs Discriminative



HMMs



Linear-chain CRFs

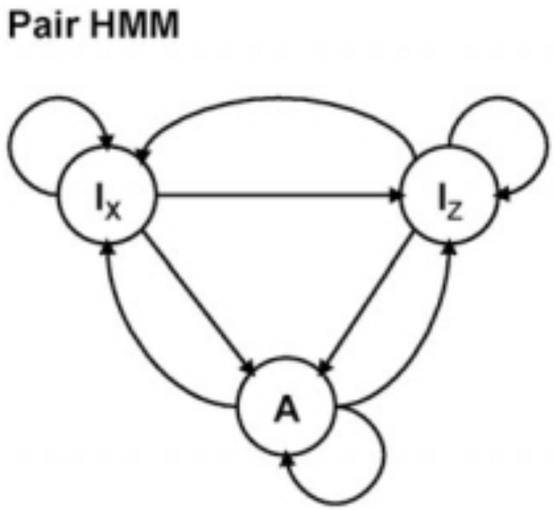
$$P(X,Y) \text{ vs } P(Y|X)$$

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp\left\{\sum_{k=1}^{K} \lambda_k f_k(y_t, y_{t-1}, \mathbf{x}_t)\right\}$$

Pair Hidden Markov Models (PHMM)

- **L**eft & **R**ight observation alphabets
- $B = \{b_j(l, r)\}$
- argmax$_Y$ $P(Y|X, Z, \lambda)$
- forward & backward trickier in log space

(Pairwise) Sequence Alignment



I_x: insertion in **x** (seq 1)

I_z: insertion in **z** (seq 2)

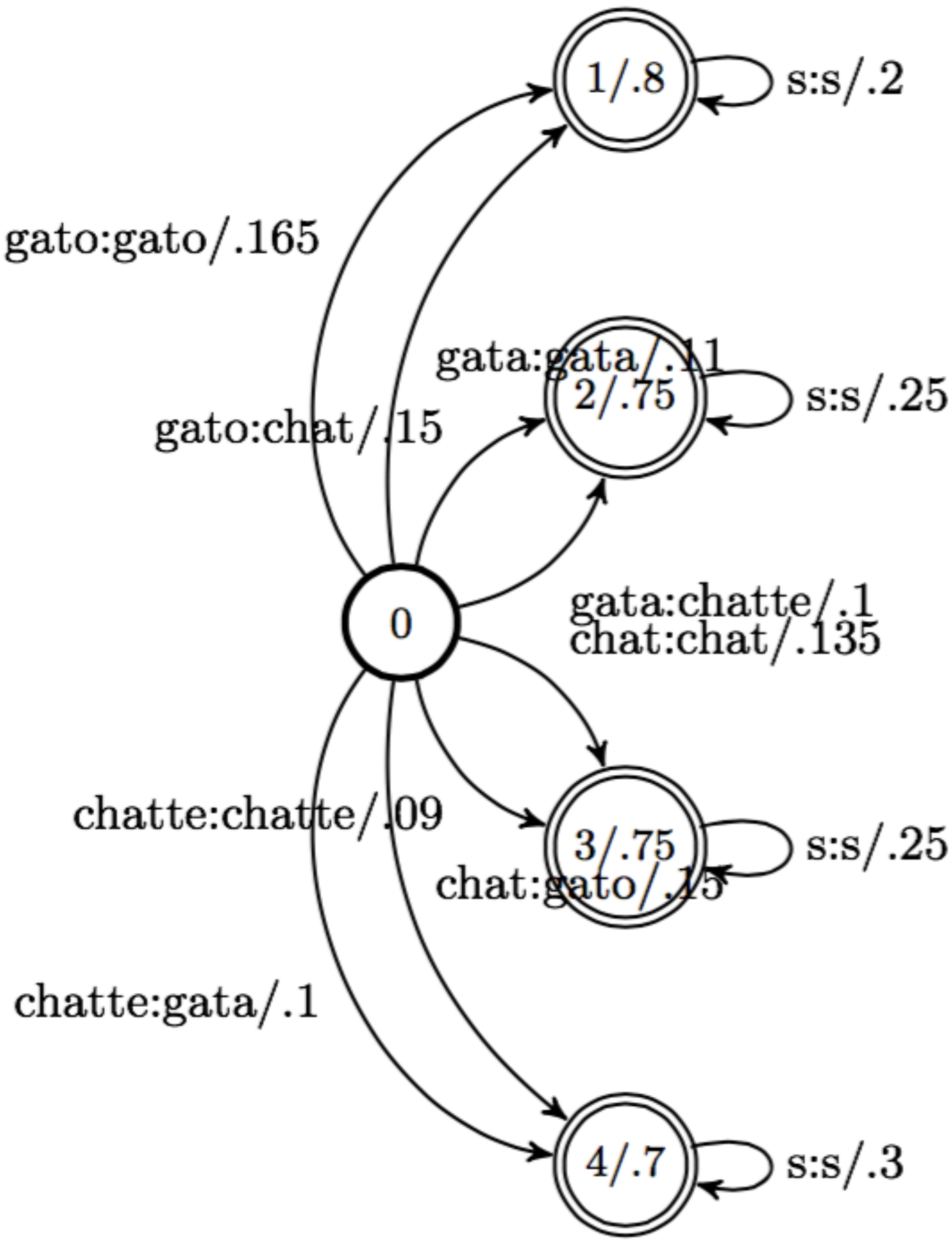
A: aligned symbols in **x** and **z**

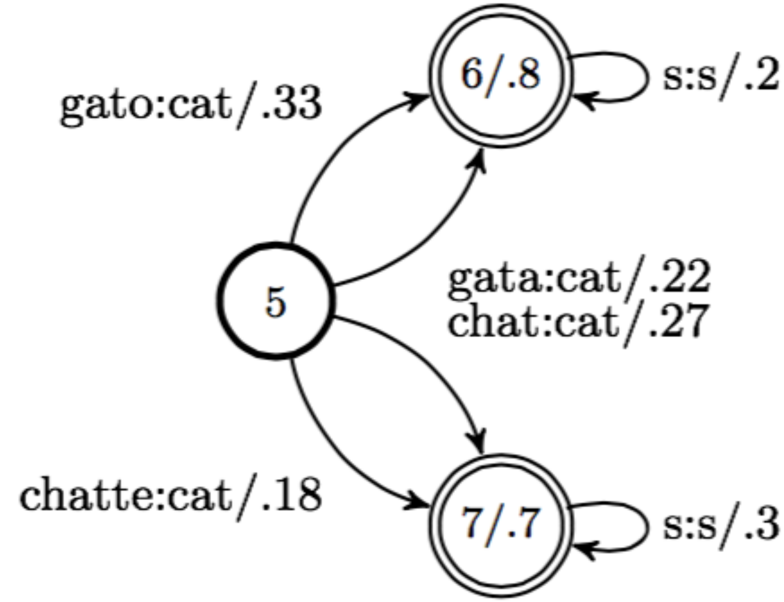
x (seq 1) :	T	T	C	C	G	–	–
z (seq 2) :	–	–	C	C	G	T	T
y (states) :	I_x	I_x	A	A	A	I_z	I_z

Levenshtein Distance, Needleman-Wunsch algorithm

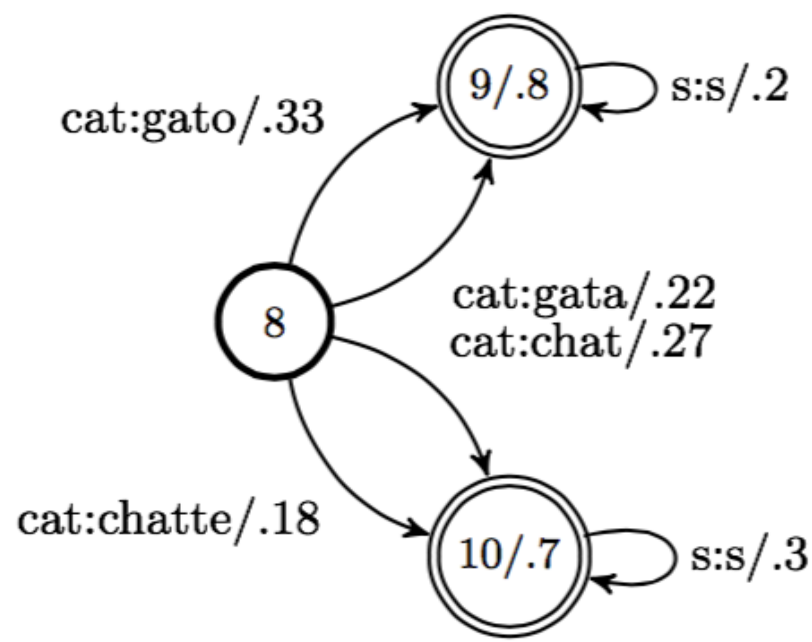
Machine Translation

- PHMMs as generalization of:
- Weighted Finite State Transducers


0
1/.8
s:s/.2
gato:gato/.165
2/.75
gata:gata/.11
s:s/.25
gato:chat/.15
gata:chatte/.1
chat:chat/.135
3/.75
s:s/.25
chatte:chatte/.09
chat:gato/.15
chatte:gata/.1
4/.7
s:s/.3


5
6/.8
s:s/.2
gato:cat/.33
gata:cat/.22
chat:cat/.27
chatte:cat/.18
7/.7
s:s/.3

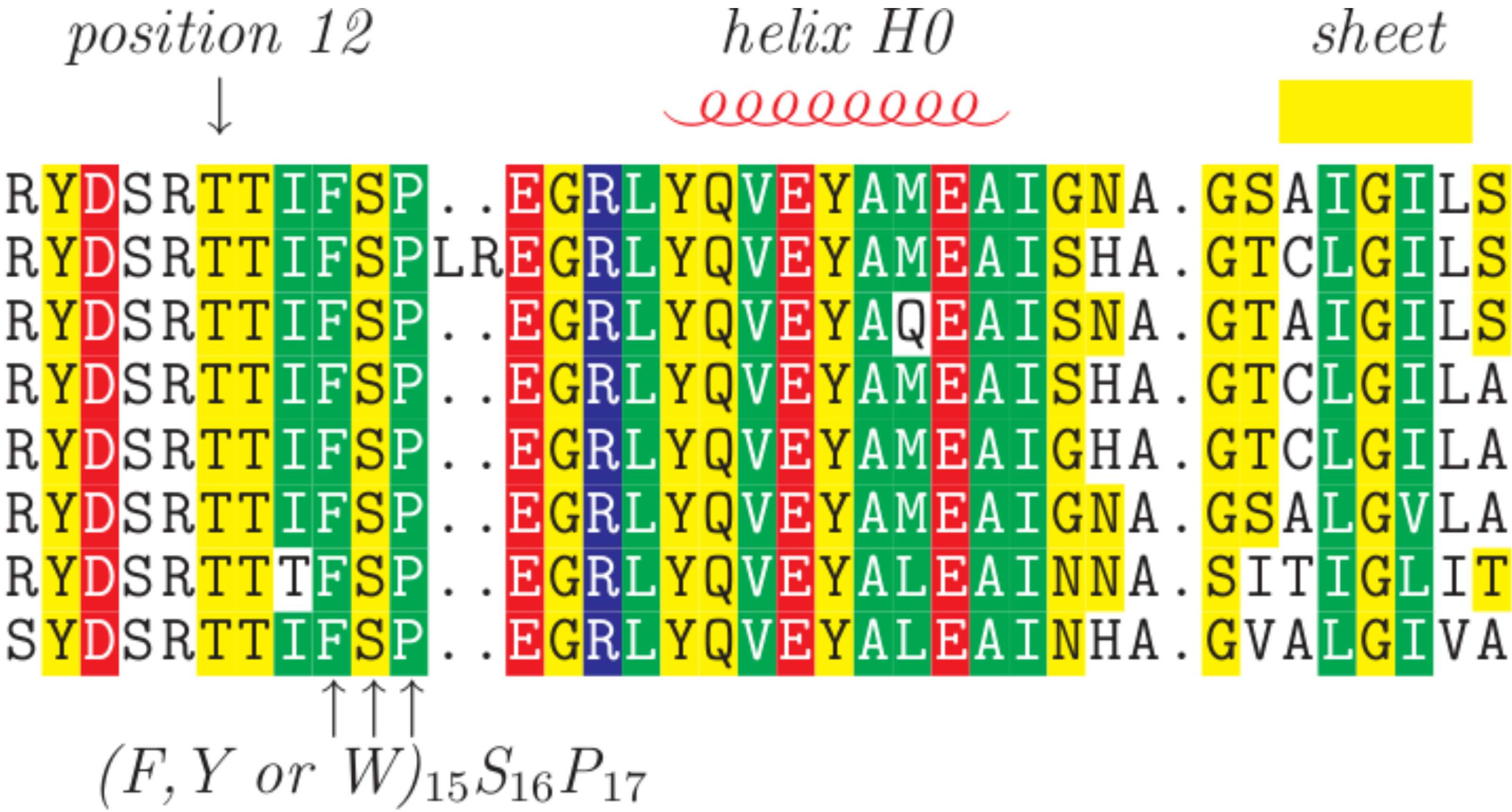
(b)


8
9/.8
s:s/.2
cat:gato/.33
cat:gata/.22
cat:chat/.27
cat:chatte/.18
10/.7
s:s/.3

WFST Implementation

- OpenFst Library (C++)
  http://openfst.org/twiki/bin/view/FST/WebHome

Profile-HMMs & Multiple Sequence Alignment



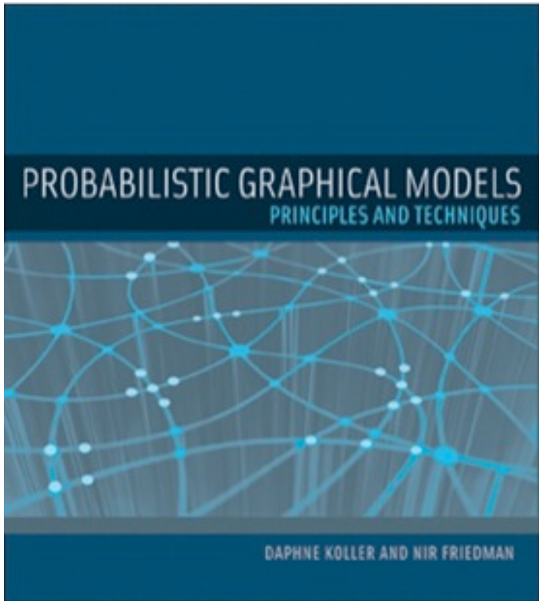
More Complexity

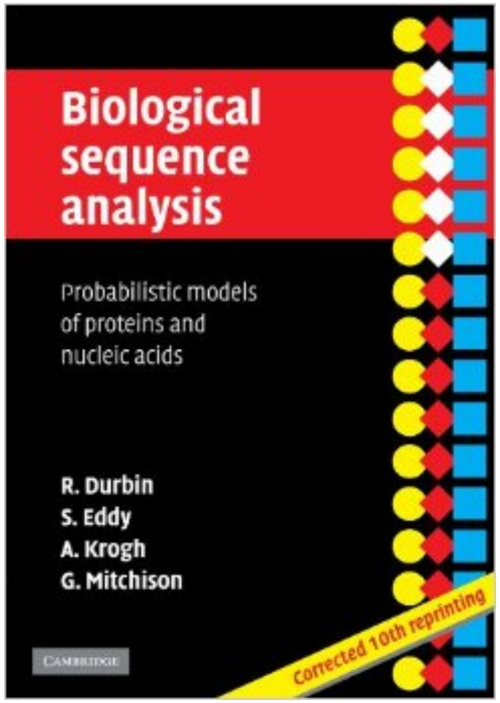
- General Hidden Markov Models (GHMM) (linear vs fully connected models)
- Other orders: 1^{st}, 2^{nd}, 3^{rd},…
- Discrete vs Continuous observations (speech recognition)
- Explicit State Duration Density

More…

Rabiner 1989

coursera.org/course/pgm by Daphne Koller





Thanks!

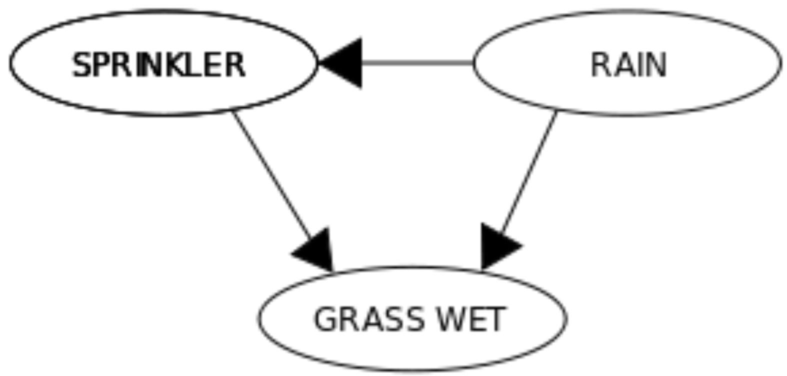
Juanmi @ tagtog.net

Conditional Random Fields (CRF)

Weighted Finite State Transducers (WFST)

Bayesian Networks

	SPRINKLER	
RAIN	T	F
F	0.4	0.6
T	0.01	0.99



RAIN	
T	F
0.2	0.8

		GRASS WET	
SPRINKLER	RAIN	T	F
F	F	0.0	1.0
F	T	0.8	0.2
T	F	0.9	0.1
T	T	0.99	0.01