CS 378 Lecture 18

Today

- Finish IBM Model 1
- Phrase-based MT
- Syntactic MT (briefly)
- Neural MT and seq2seq models

Recap IBM Model 1

$\bar{s} =$ Je fais un bureau NULL

$\bar{a} =$ $a_1=1$ $a_2=2$ $a_3=2$ $a_4=3$ $a_5=4$

$\bar{t} =$ I am making a desk

$$P(\bar{t}, \bar{a} \mid \bar{s}) = \prod_{i=1}^{n} P(a_i)\, P(t_i \mid s_{a_i})$$

NULL

$P(a_i) =$ uniform dist over $\{1, 2, \ldots, m+1\}$

$P(t_i \mid s_{a_i})$ $\qquad$ $P(I \mid Je) = 0.8$

...

Announcements

- Custom FP proposals returned w/comments
- Optimal lecture Tuesday
- A4
- A5

Inference in Model 1

What we care about: $P(\bar{a} \mid \bar{t}, \bar{s})$

$\left(\underset{\bar{a}}{\operatorname{argmax}}\, P(\bar{a} \mid \bar{t}, \bar{s}) \right)$

$$P(\bar{a} \mid \bar{t}, \bar{s}) = \frac{P(\bar{a}, \bar{t} \mid \bar{s})}{\cancel{P(\bar{t} \mid \bar{s})}} = \prod_{i=1}^{n} \underbrace{\cancel{P(a_i)}}_{\frac{1}{m+1}} P(t_i \mid s_{a_i})$$

constant

$$P(\bar{a} \mid \bar{t}, \bar{s}) \text{ proportional to } \prod_{i=1}^{n} P(t_i \mid s_{a_i})$$

$$P(a_i \mid \bar{t}, \bar{s}) \text{ proportional to } P(t_i \mid s_{a_i})$$

(model params)

$P(t|s)$

	I	like	eat
Je	0.8	0.1	0.1
J'	0.8	0.1	0.1
mange	0	0	1.0
aime	0	1.0	0
NULL	0.4	0.3	0.4

Example

$\bar{s} =$ Je NULL

a_1

$\bar{t} =$ I

$$P(a_1 | \bar{t}, \bar{s}) \text{ prop. to } \begin{cases} a_1 = 1 \ (\text{Je}) & P(I | Je) = 0.8 \\ a_1 = 2 \ (\text{NULL}) & P(I | NULL) = 0.4 \end{cases}$$

$$P(a_1 | \bar{t}, \bar{s}) = \begin{cases} a_1 = 1 & 2/3 \\ a_1 = 2 & 1/3 \end{cases}$$

Ex 2   J' aime NULL
I   like

$$P(a_1 \mid \bar{s}, \bar{t}) = \begin{cases} 0.8 & J' \\ 0 & \text{aime} \\ 0.4 & \text{NULL} \end{cases} \Rightarrow \begin{matrix} 2/3 \\ 0 \\ 1/3 \end{matrix}$$

$$P(a_2 \mid \bar{s}, \bar{t}) = \begin{cases} 0.1 & J' \\ 1.0 & \text{aime} \\ 0.3 & \text{NULL} \end{cases} \Rightarrow \begin{matrix} 1/14 \\ 10/14 \\ 3/14 \end{matrix}$$

More complex models : HMM alignment model

$$P(a_i \mid a_{i-1})$$

IBM Models 2-4

Learning   Unsupervised learning: $\bar{s}, \bar{t}$
no $\bar{a}$

EM (Expectation Maximization)

maximize

$$\sum_{i=1}^{D} \log \overbrace{\sum_{\bar{a}} P(\bar{a}, \bar{t}^{(i)} | \bar{s}^{(i)})}^{P(\bar{t}^{(i)} | \bar{s}^{(i)})}$$

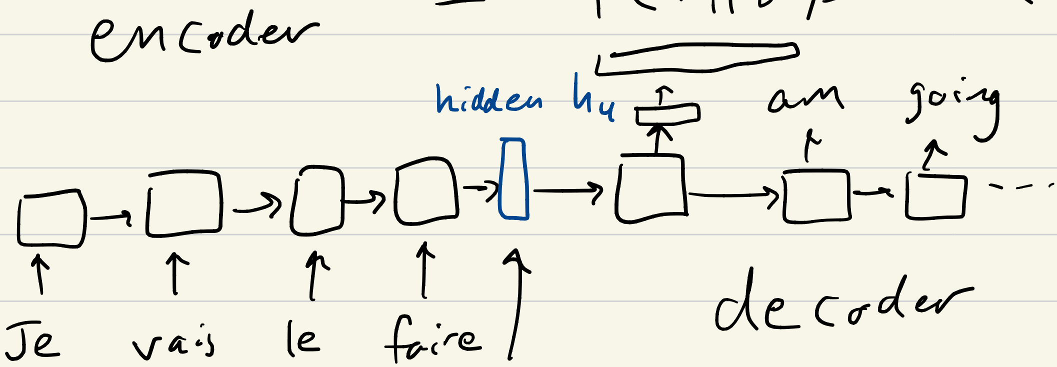
log marginal likelihood

SLIDES

Seq 2 seq models

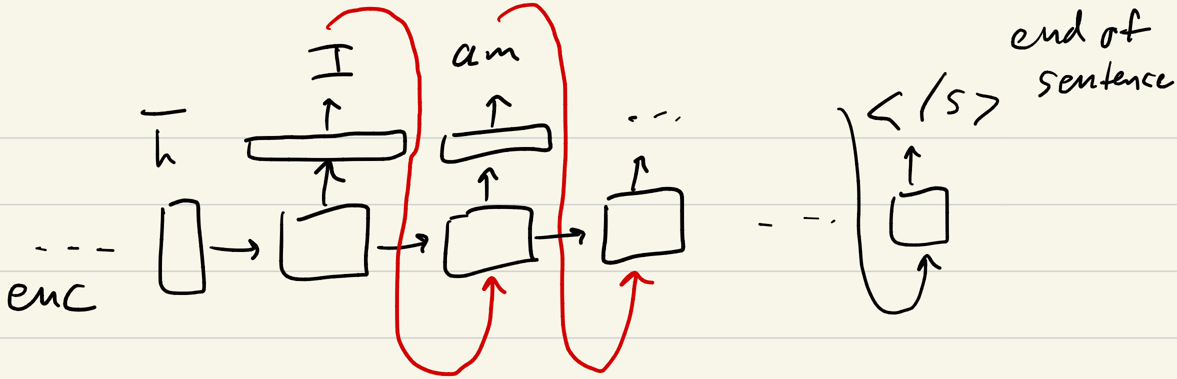
Key idea: encode source sent w/RNN,
"decode" target w/another RNN

$$I \leftarrow P(t_1 | \bar{s}) = \text{softmax}(W\bar{h}_i)$$



representation of the French sentence

$$P(\bar{t} | \bar{s}) = P(t_1 | \bar{s}) P(t_2 | \bar{s}, t_1) \cdots$$



Feed output t_i into cell input for t_{i+1}

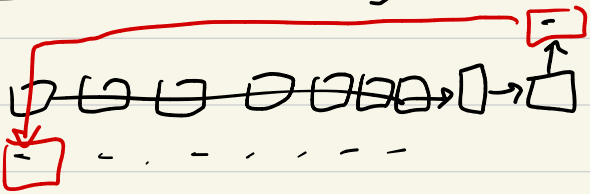
Why? Model capture word-word dependencies more easily

Training Given sequence pairs $(\bar{s}, \bar{t})$

$$loss = \sum \log P(t_i^* \mid \bar{s}, t^*_{1,\dots,i-1})$$

Like in language modeling, assume everything up until now matches our reference/gold

Problem: Long-range dependencies



What we want is a way to look back at the input more easily

Solution: attention