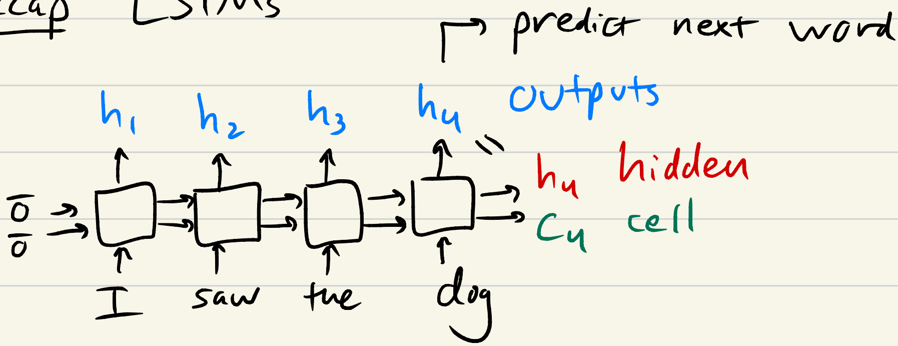
# Announcements

- A4 due in 1 week


Star Wars The Third Gathers: The Backstroke of the West
(subtitles machine translated from Chinese)

# Recap LSTMs




# Machine Translation

Phrase-based MT: 2000 – 2015

Neural MT: 2015 – present

This lecture: PBMT and word alignment

Thursday: Finish PBMT, start NMT: seq2seq models

→ leads to attention: critical idea in modern deep learning w/roots in word alignment

Goal: understand core concepts of PBMT, understand word alignment (A5)

Learn to translate sentence in one lang (source) to another lang (target)

French → English

<u>Data</u> bitext, parallel sent pairs

| | |
|---|---|
| Je fais un bureau | I make a desk |
| Je fais une soupe | I'm making a soup |
| Qu'est-ce que tu fais? | What are you doing? |
| Quelle erreur! | What a mistake! |

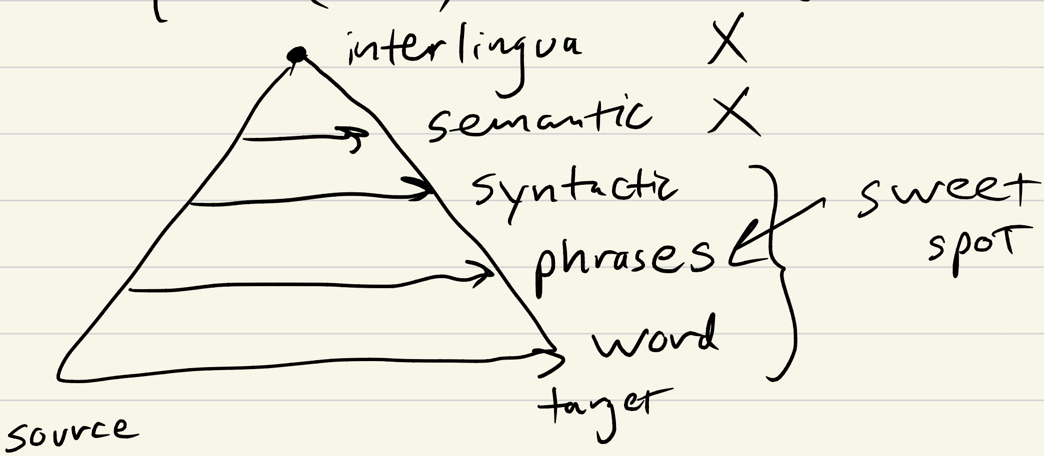
you make a mistake

1. What does "tu fais une erreur" mean?
2. How do you know what fais means?

- Many-to-many translations
- Cooccurrence → alignment
- Phrasal translations

French → meaning → En sentence
→ Ja sentence
⋮

Vauquois (1968) levels of transfer




① How to get chunks/phrases?
② How to translate sentences? next class

## Word alignment

Input: bitext

Output: Je fais un bureau

$a_1=1$ $a_2=2$ $a_3=2$ $a_4=3$ $a_5=4$

I am making a desk

1 2 2 3 4

Source = French, target = English

Alignment: family of models

$P(\bar{a}, \bar{f} \mid \bar{s})$

$\bar{f}$ : Eng words

$\bar{s}$ : Fr words

$\bar{a}$ : alignments

$a_i = j$ : the $i$th En word aligns to $j$th Fr word

$a_i$ takes one value

one-to-many alignment

$a_i$ are not necessarily monotonic

Fais-tu un bureau?

Are you making a desk?

Unsupervised learning problem

- We do not see labeled $\bar{a}$ anywhere in our data

## IBM Model 1 (1993)

- Canadian Hansards Fr – En

$\bar{a} = (a_1, \ldots, a_n)$ $\bar{t} = (t_1, \ldots, t_n)$

$\bar{s} = (s_1, \ldots, s_m, \text{NULL})$ allow unaligned words

$$P(\bar{t}, \bar{a} \mid \bar{s}) = P(\bar{a})\, P(\bar{t} \mid \bar{a}, \bar{s})$$

$$= \prod_{i=1}^{n} P(a_i) \underbrace{P(t_i \mid s_{a_i})}_{\text{"emission" probs (next page)}}$$

$P(a_i)$: dist over $\{1, \ldots, m+1\}$

uniform $\frac{1}{m+1}$ prob for each

no parameters

<u>M1</u>: $\prod_{i=1}^{n} P(a_i) P(t_i \mid s_{a_i})$

Model params: emissions in an HMM

$|V_s| \times |V_t|$ matrix of probs

includes NULL

$P(\text{target word} \mid \text{src word})$

$P(t_i \mid s_{a_i})$: look up the prob of word $t_i$ conditioned on word $s_{a_i}$ in the matrix

$a_i$ is a "switch" that tells $t_i$ what to condition on

<u>Ex</u> params

$$P(t \mid s)$$

| | 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 | 0.3 |

$Je_1 \quad NULL_2$

$I \quad a_1 = \{1, 2\}$

$$P(a_1 = 1, t = \text{"I"} \mid s = \text{"Je"})$$

$$= P(a_1 = 1)\, P(I \mid Je)$$

$$= 0.5 \cdot 0.8$$

<u>Inference</u>: $P(\bar{a} \mid \bar{s}, \bar{t})$ posterior <u>over</u> $\bar{a}$

From HMMs: $\bar{a} \approx \bar{y}$, $\bar{t} \approx \bar{x}$

$\bar{s}$ just hangs out

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

constant

doesn't depend on $\bar{a}$

constant

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

prop. to

$\text{Je}_1 \quad \text{NULL}_2$ $\qquad$ $\underset{\text{I}}{=} \quad \underset{\text{Je}}{=}$

$\text{I} \qquad a_1 = \{1, 2\}$ $\qquad P(t_1 | s_1)$

$$P(a_1 | \bar{S}, \bar{F}) \propto \begin{cases} a_1 = 1: & 0.8 \\ a_1 = 2: & P(t_1 | s_2) \text{ ``NULL''} \\ & 0.4 \end{cases}$$

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

---

J' $\quad$ aime $\quad$ NULL

I $\quad$ like

What is $P(a_1 | \bar{S}, \bar{F})$ ?

| $a_1 = 1$ | $a_1 = 2$ | $a_1 = 3$ |
|---|---|---|
| $P(\text{I} \mid \text{J'})$ | $P(\text{I} \mid \text{aime})$ | $P(\text{I} \mid \text{NULL})$ |
| 0.8 | 0 | 0.4 |
| 2/3 | 0 | 1/3 |

# Learning

No labeled $\bar{a}$

If we had labeled $\bar{a}$:

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

$(\bar{s}^{(i)}, \bar{t}^{(i)}, \bar{a}^{(i)})$ labeled aligned bitext

Problem: no $\bar{a}$ labeled

Instead: $$\max \sum_{i=1}^{D} \log \underbrace{\sum_{\bar{a}} P(\bar{a}, \bar{t}^{(i)} \mid \bar{s}^{(i)})}_{P(\bar{t}^{(i)} \mid \bar{s}^{(i)})}$$

log marginal probability

Expectation Maximization

For Model 1: $$\sum_{i=1}^{D} \log \underbrace{\prod_{j} \sum_{a_j} P(a_j, \bar{t}_j^{(i)} \mid \bar{s}^{(i)})}_{O(nm)}$$

because of independence assumptions