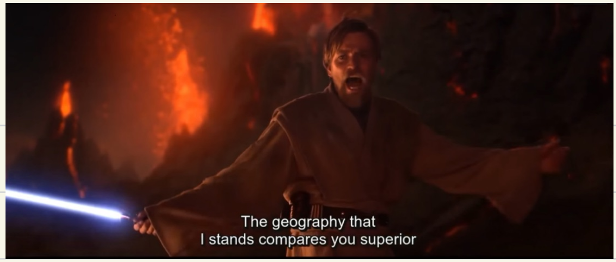
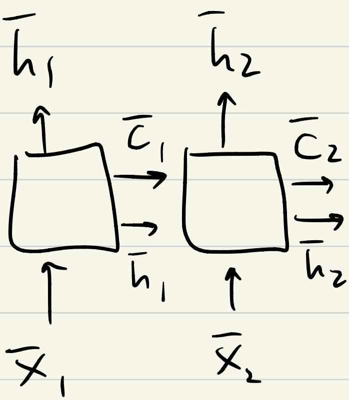


CS 378 Lecture 17
Machine Translation

Today
- Phrase-based MT
- Word alignment

Announcements
- Midterm
- FP
- A3
- A4

Recap LSTMs



involves gated connections

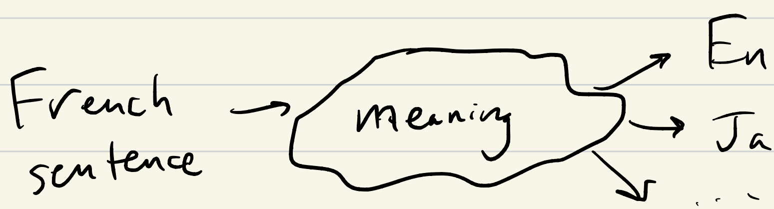
$$\bar{h}_i = \bar{f} \odot \bar{h}_{i-1}$$

Machine Translation

Today: phrase-based (pre-2015)
Next time: neural w/ seq2seq models (post-2015), following on how we used RNNs for LM

Input: $\bar{s}$ source sentence
Output: $\bar{t}$ target language

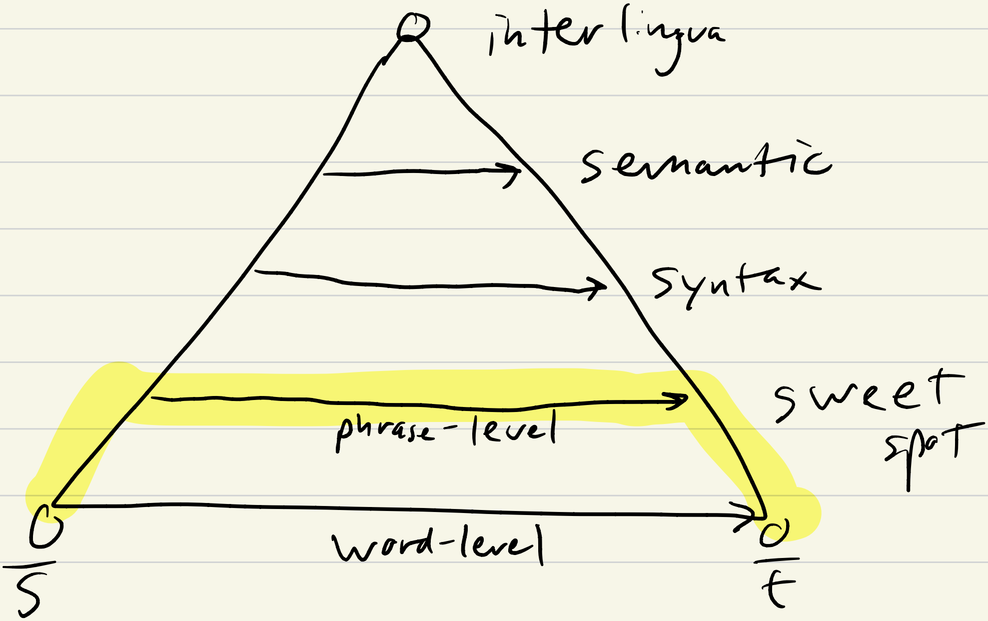
Data: bitext. Set of $(\bar{s}, \bar{t})$ pairs



We don't know how to do this!

interlingua

Bernard Vauquois (1968)

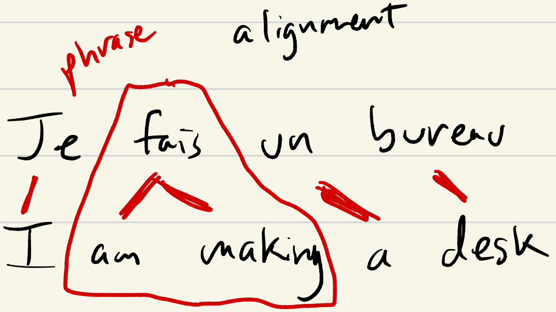


Phrase-based MT

Bitext $\xrightarrow{\text{word alignment}}$ aligned bitext → phrase table

phrase table ↓ decoder

LM ↙ decoder



Decoder: searches over the space of phrase-by-phrase translations to find one that scores best (including an LM score)

Je fais un bureau

use an LM to figure out which candidate is grammatical English

Word Alignment (focus of today)

Input: bitext $(\bar{s}, \bar{t})$ pairs

Output: one-to-many alignments from $\bar{s}$ to $\bar{t}$

placeholder ↓

$\bar{s}$ = Je vais le faire NULL

$\bar{a}$ = 1 2 2 4 4 3

$\bar{t}$ = I am going to do it

$a_2 = 2$
$a_3 = 2$
...

Each word in $\bar{t}$ aligns to one word in $\bar{s}$

Define a vector $\bar{a}$

a_i = index in $\bar{s}$ that word t_i aligns to

Alignment models: place distribution over $p(\bar{t}, \bar{a} \mid \bar{s})$

generative model of $\bar{t}, \bar{a}$

$\bar{t} \approx$ words in an HMM

$\bar{a} \approx$ tags

IBM Model 1 (1993)

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

$\bar{s} = (s_1, \ldots, s_m, \text{NULL})$ $\quad m$ source words

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

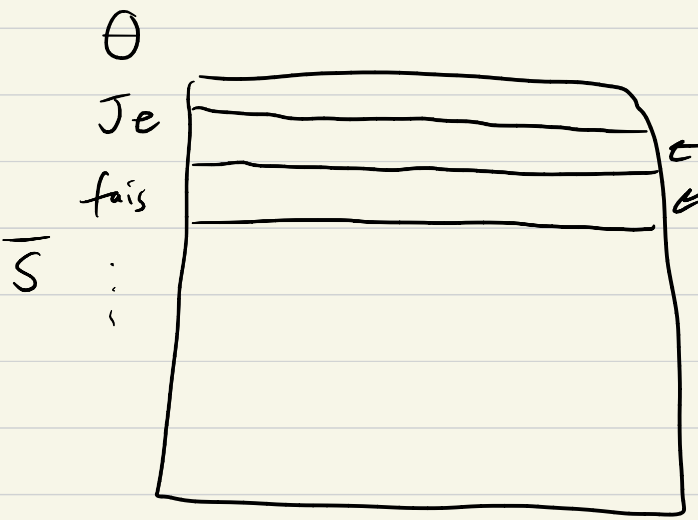
Generative process: for each target word i, pick a source index a_i

$P(a_i) = \frac{1}{m+1}$ uniform over these options

Generate t_i conditioned on s_{a_i}

a_ith source word

Model params: translation dictionary



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

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

like emissions in HMM!

Each a_i is like a switch, tells you what row of θ to use

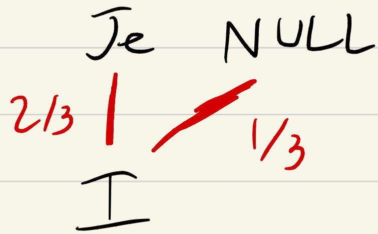
Je fais

$a_1 = 1 : P(t_1 \mid \text{Je})$

$a_1 = 2 : P(t_1 \mid \text{fais})$

$T = \{I, \text{like}, \text{eat}\}$ $\quad S = \{\text{Je}, \text{J'}, \text{mange}, \text{aime}, \text{NULL}\}$

	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



Ex 1

$\bar{S} =$ Je NULL

$\bar{T} =$ I

$P(a_1 \mid \bar{T}, \bar{S})$?

prop. to $\begin{cases} P(I \mid Je) & a_1 = 1 \\ P(I \mid NULL) & a_1 = 2 \end{cases}$

$= \begin{cases} 0.8 \Rightarrow & 2/3 \\ 0.4 & 1/3 \end{cases}$

What do we want?

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

HMM: $P(\bar{x}, \bar{y}) \Rightarrow P(\bar{y} \mid \bar{x})$ posterior

model (tagging)

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

$$\downarrow$$

$$\underset{\bar{a}}{\operatorname{argmax}}\ P(\bar{a} \mid \bar{t}, \bar{s})$$ multiply top/bottom

$$p(\bar{t} \mid \bar{s})$$

$$P(\bar{a} \mid \bar{t}, \bar{s}) = \frac{P(\bar{a}, \bar{t} \mid \bar{s})}{P(\bar{t} \mid \bar{s})}$$ constant w.r.t. $\bar{a}$

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

constant $\frac{1}{m+1}$

$$= \underset{\bar{a}}{\operatorname{argmax}} \prod_{i=1}^{n} \cancel{P(a_i)}\, P(t_i \mid s_{a_i})$$

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

$$\begin{array}{cccc} & 1 & 2 & 3 \\ \underline{\text{Ex 2}} & \text{J'} & \text{aime} & \text{NULL} & \bar{s} \\ a_1 & \underline{\text{I}} & \text{like} & & \bar{t} \end{array}$$

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

argmax $a_1 \overset{?}{=} 1$

$P(a_2 \mid \bar{s}, \bar{t})$

argmax $a_2 \overset{?}{=} 2$

Bitext Je I

J' aime I like

Je fais un bureau I am making...

$\Rightarrow$ alignments + phrases

Learning Hard!

Unsupervised: no examples of labeled $\bar{a}$.

Expectation Maximization:

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

$(\bar{s}^{(i)}, \bar{t}^{(i)})_{i=1}^{D}$

Phrase-based MT

$(\bar{s}, \bar{t})$ pairs $\Rightarrow$ learn an aligner

$\Rightarrow$ align our data

phrase extraction: aligned sents

$\Rightarrow$ phrase translation options

Phrase table: huge!

	separator		score (from alignment)
Je fais	|||	I make	0.9
Je fais	|||	I am making	0.6
J'aime	|||	I like	0.7

$\downarrow$ +LM

decoder

<u>Takeaways</u> Idea of alignment

Learn a model to get a distribution over source words given a target word