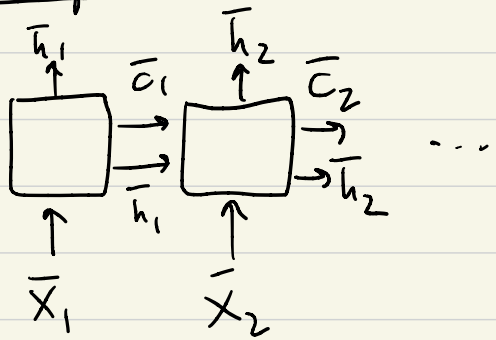
CS 378 Lecture 17

Today
- Finish LSTMs and language modeling
- Visualizing LSTMs
- Machine translation
- Word alignment

Recap LSTMs



See Olah blog for architecture

Main difference from Elman networks: gates

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

$\odot$: elementwise product

$\bar{f}$: vector $\in [0,1]^d$

Announcements

- Midterm
- Custom FP proposals
- A4

Machine Translation

- Learn to translate sentences from lang A into lang B

Input: source sentence $\overline{s}$
Output: target sent $\overline{t}$

Data: bitext. Set of sentences + their translations in the other lang.

Given a bitext, how do we learn to translate?

Je fais un bureau | I make a desk

Qu'est-ce que tu fais? | What are you doing?

faire (Fr) = to make
to do

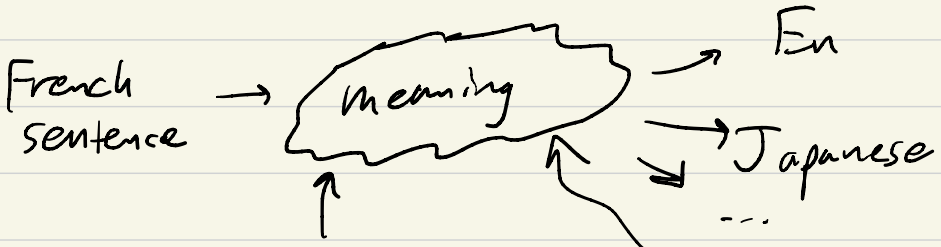
--

① Word mapping is not one-to-one
② Some phrases need to translate as a unit

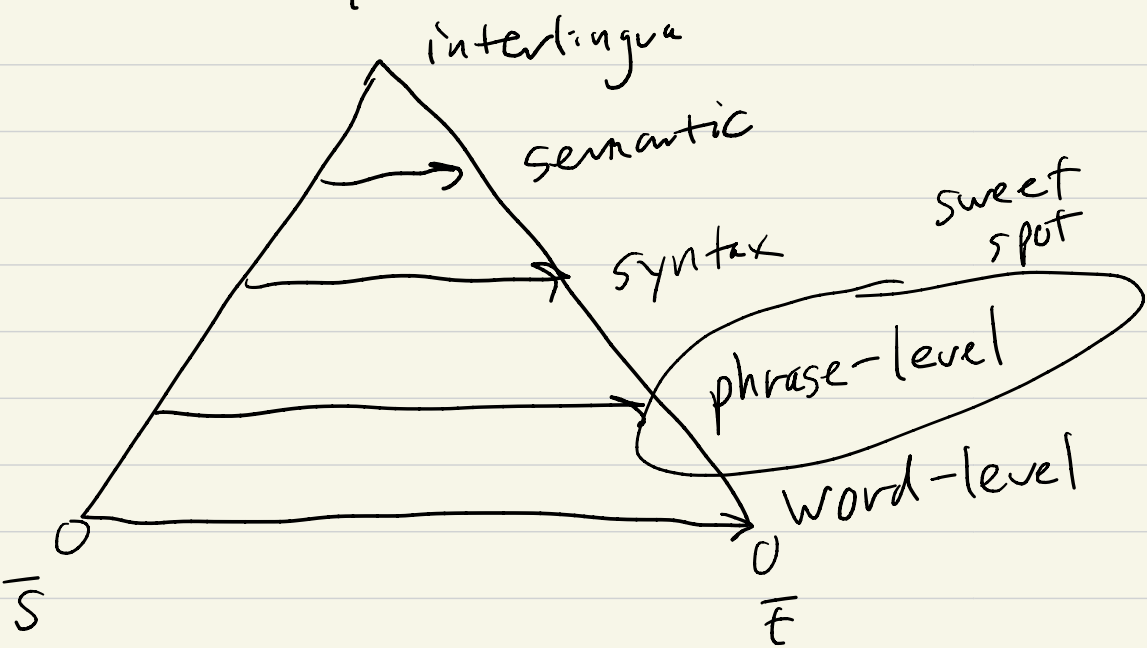
Step 1 Word alignment: discover word-level mapping between sentences in the bitext

Unsupervised learning problem

How to do MT?



Bernard Vauquois (1968)



Phrase-based MT

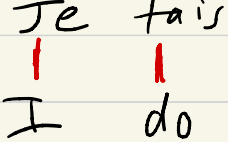
(this lecture + part of next)

Then: neural MT

Bitext

↓ word alignment

Aligned bitext



↓ extract phrases

Phrase table

Je fais ||| I do

↓

Decoder { Phrase T. / LM

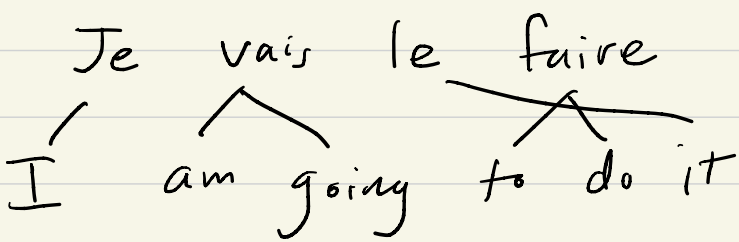
this is a phrase translation pair

produce translation w/ high LM prob using high-scoring phrases

Word Alignment

Input: bitext, source sents $\bar{S}$
target sents $\bar{T}$

Output:



Alignments: one-to-many

Each word in $\bar{T}$ aligns to 1 word in $\bar{S}$

Define $\bar{a}$ as follows:

a_i = index in the source that the ith target word aligns to

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

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

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

placeholder for unaligned words

Alignment models: distribution

$p(\bar{t}, \bar{a} | \bar{s})$ generative model of $\bar{t}, \bar{a}$

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

$\bar{a}$ is like the tags

IBM Model 1 (1993)

n target words

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

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

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

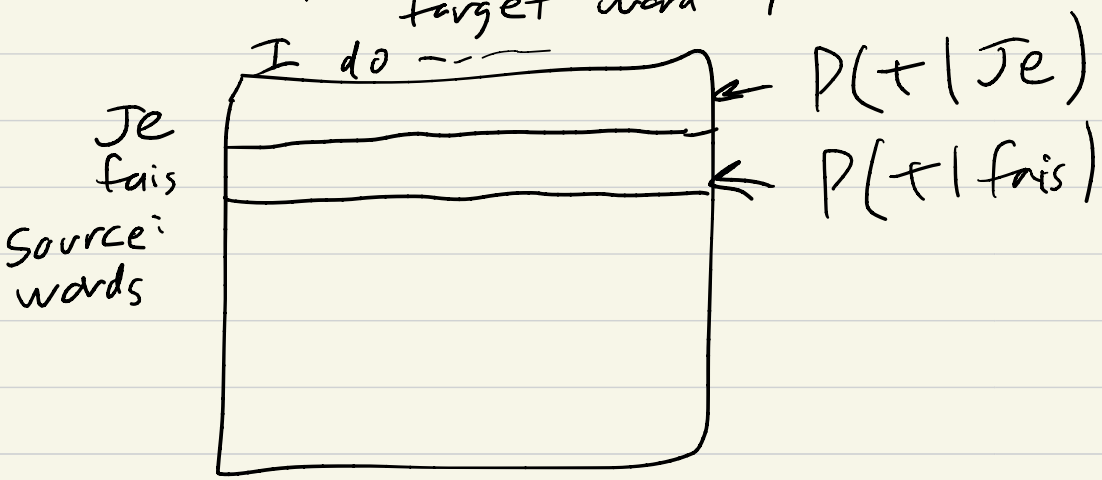
Generative process: for each target word, pick a source index to align to $P(a_i) =$ uniform dist over $\frac{1}{m+1}$ options

Generate t_i conditioned on s_{a_i}

the a_ith source word

Model params: dictionary

target word



a_i is like a switch. Tells you which source word to condition on to generate the given target

$a_1 = 1$ vs. $a_1 = 2$

$P(I|Je)$ very high prob

$P(I|fais)$ very low prob