Announcements

- A4 due Tues
- A5 out Tues, due the following Tues

Recap Word alignment w/ IBM Model 1

$\bar{s}$ = Je fais un bureau NULL $(s_1, \ldots, s_m, \text{NULL})$

$\bar{a}$ = 1 [1, 2, 2, 3, 4]

$\bar{t}$ = I am making a desk $(t_1, \ldots, t_n)$

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

$P(a_i)$: constant $\frac{1}{m+1}$

$P(t_i \mid s_{a_i})$: parameters $P(t \mid s)$ matrix $|V_s| \times |V_t|$ "emissions"

Inference:

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

Learning: $\max \sum_{j=1}^{D} \log P(\bar{t}^{(j)}, \bar{a}^{(j)} \mid \bar{s}^{(j)})$ no labeled $\bar{a}$

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

Expectation Maximization: repeat E+M:

E-step: compute posteriors $P(a_i | \bar{s}^{(j)}, \bar{t}^{(j)})$ on each train ex

"labeling the data"

M-step: count + normalize using posteriors as labels (soft labels)

Je NULL
I

Je fais NULL
I do

I-Je ≥ 50% prob ⇒ will lead to high prob of P(I|Je) after M-step

---

Today: ① HMM alignment
② Phrase extraction
③ Phrase-based decoding
④ Syntactic MT
⑤ seq2seq models / NMT (next time)

HMM Alignment (Vogel 1996)

Same as Model 1, except: ~~$P(\bar{a}) = \prod_{i=1}^{n} \frac{1}{m+1}$~~

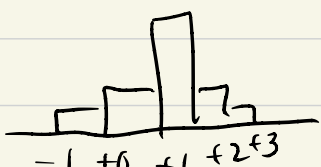
$P(a_i)$

HMM:

$$P(\bar{a}) = \prod_{i=1}^{n} P(a_i \mid a_{i-1})$$

Categorical $(a_i - a_{i-1})$

$a_i = 1 \ldots 20$



NN → VBZ

14 → 15 ???

Phrase Extraction ① Run alignment model in both directions (Fr → En, En → Fr)

② Intersect them

③ Find aligned phrases by drawing "boxes"

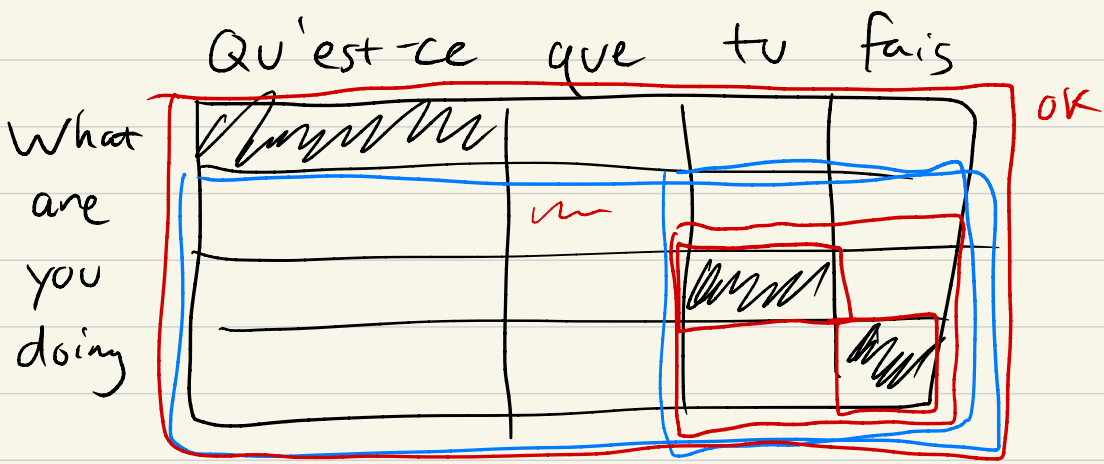
④ Count phrases, build phrase table

Qu'est-ce que tu fais NULL

What are you doing

What are you doing

Qu'est-ce que tu fais



tu - you

fais - doing

tu fais = you doing

not a phrase

Requirements

- "Corners" of rectangle are aligned
- Words in phrase cannot align outside rectangle

Phrase-based decoding

One thing we could do:

$$\max_{\bar{t}} \sum_{\bar{a}} P(\bar{a}, \bar{t} \mid \bar{s})$$ using Model 1/ HMM

Word level translation doesn't work well

New modeling paradigm

Noisy channel (Shannon / Weaver 50s)

I see source $\bar{s}$

Model is that there's an underlying "signal" $\bar{t}$, $\bar{s}$ is emitted from $\bar{t}$

$$\underbrace{P(\bar{t} \mid \bar{s})}_{\text{likely } \bar{t}} \propto \overbrace{\underbrace{P(\bar{s} \mid \bar{t})}_{\text{emission of } \bar{s}}}^{\text{TM}} \overbrace{P(\bar{t})}^{\text{LM}}$$ ← prior over $\bar{t}$

View translation in "reverse" direction

TM is based on phrases