Word Alignment

MTM'2014, Trento, Sept. 8

Mark Fishel
Uni. Zürich / Uni. Tartu

TASK:

En: One pizza of the day, please!

It: Una pizza del giorno, per favore!

$\times 10^6$

TASK:

LV: Lūdzu jusu dienas pic!

ET: Üks päevapizza, palun!

WHY?

* mainly : to learn to align unseen sentences
  - generalization, learning
* to extract phraze pairs (PBMT), gappy phraze pairs (HPBMT), etc.
* generate dictionaries from corpora
* decipher lost/unknown languages

ETC.

FORMAL DEF-S:

*Parallel corpus:

$$D = \{ \langle \bar{f}^{(1)}, \bar{e}^{(1)} \rangle, \ldots, \langle \bar{f}^{(N)}, \bar{e}^{(N)} \rangle \}$$

$$\bar{f}^{(n)} = f_1^{(n)}, \ldots, f_{I^{(n)}}^{(n)}$$

$$\bar{e}^{(n)} = e_1^{(n)}, \ldots, e_{J^{(n)}}^{(n)}$$

E.g:

EN	IT
the car	la auto
my car	la mia auto
my house	la mia casa

$N = 3$ $\quad f_1^{(2)} = \text{"la"} \quad I_2 = 3$

$\bar{e}^{(1)} = (\text{"the"}, \text{"car"})$

*Alignment function:

$\bar{a}^{(n)}: a_i^{(n)} \in [1..J_n]$, $|\bar{a}^{(n)}| = I_n$

$f_i^{(n)}$ aligns to $e_{a_i^{(n)}}^{(n)}$, or

f_i aligns to e_{a_i}

E.g:

$\bar{e}$: the car

$\bar{f}$: la auto

$\bar{a} = (0, 1)$

$a_1 = 0$

LEARNING, GENERALIZATION?

* let's define a set of tunable parameters:

$\Theta : \Theta_{e,f} = p(f/e)$ ← probability of f translating into e

...WHY PROBABILITIES?

* and express the overall corpus likelihood using them:

$$p(D;\Theta) = \prod_{n=1}^{N} p(\bar{f}^{(n)}, \bar{a}^{(n)} | \bar{e}^{(n)}; \Theta) = \dots$$

* and then find such $\hat{\Theta}$ that gives the highest data likelihood:

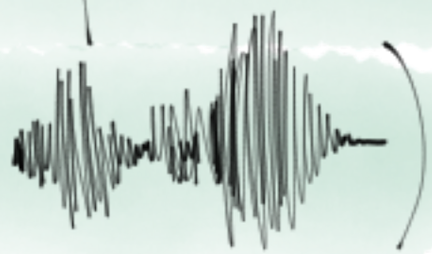
$$\hat{\Theta} = \underset{\Theta}{\operatorname{argmax}}\, p(D;\Theta)$$

MLE

WHY PROBABILITIES?

TO USE THEM!

We want:

* translation: p("una pizza del..." | "ow pi...")
* tagging: p(DT NN VB RB | "the lion sleeps tonight")
* ASR: p("wreck a nice beach" | )

...

We can:

p(HEADS/TAILS)
p(tired | was beginning to get)
p(casa | home)

...

how? independence assumptions

CORPUS LIKELIHOOD

* assuming sentence pairs indep. of each other:

$$p(D) = \cancel{p(N)} \cdot \prod_{n=1}^{N} p(\bar{f}^{(n)}, \bar{a}^{(n)} \mid \bar{e}^{(n)})$$

* assuming lexical dependences only:

$$p(\bar{f}, \bar{a} \mid \bar{e}) = \cancel{p(I \mid \bar{e})} \cdot \prod_{i=1}^{I} \cancel{p(a_i \mid \ldots)} \cdot p(f_i \mid e_{a_i})$$

$\theta_{f_i, e_{a_i}}$

* let's assume we are trying to find the parameters based on present (annotated) alignments: constant $p(a_i \mid \ldots)$

OPTIMIZATION

$a > 0$

$b > 0$

$\forall a \forall b.\ a < b \Leftrightarrow$

$\log(a) < \log(b)$

$$\hat{\theta} = \arg\max_{\theta} \log p(D) =$$

$$= \arg\max_{\theta} \log \prod_n \prod_i p\left(f_i^{(n)} \mid e_{a_i^{(n)}}^{(n)}\right) =$$

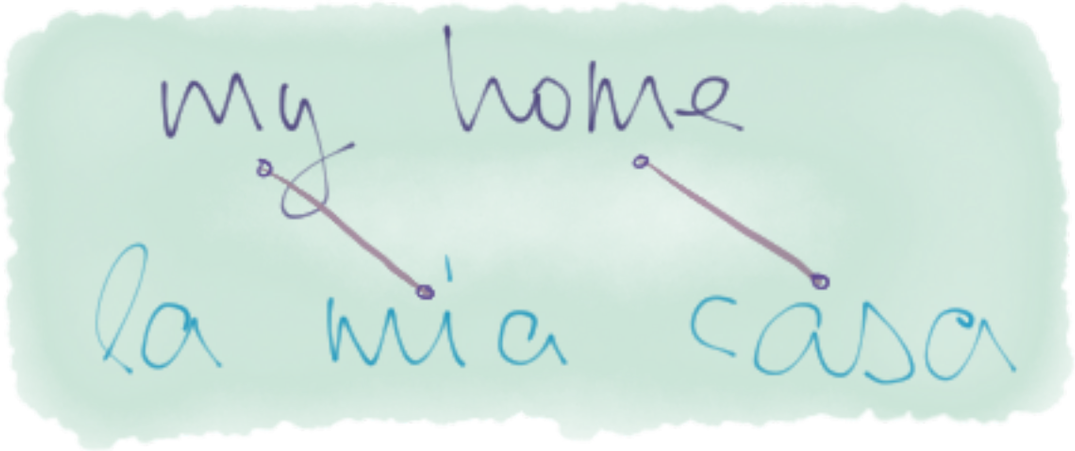
$$= \arg\max_{\theta} \sum_n \sum_i \log p\left(f_i^{(n)} \mid e_{a_i^{(n)}}^{(n)}\right) =$$

$$= \arg\max_{\theta} \sum_{(e,f)} \underline{\text{count}(\langle e,f \rangle)} \cdot \log p(f|e)$$

↖ in the whole corpus

$$\hat{\theta}_{e,f} = \frac{\text{count}(\langle e,f \rangle)}{\sum_{f'} \text{count}(\langle e,f' \rangle)}$$

AND WITHOUT ANNOTATION?



$$p(\text{casa} \mid \text{home}) = \frac{\text{count}(\text{"home" aligned to "casa"})}{\text{count}(\text{"home" total})}$$

$$p(\text{casa} \mid \text{home}) = ?$$

* no given alignment
* J^I (exponential) nr. of possible al-s.
* could we give preference to some of the alignments?

— in order to sum over all possible alignments

ALIGNMENT LIKELIHOOD

$$p(\bar{a}\,|\,\bar{f},\bar{e}) = \frac{p(\bar{f},\bar{a}\,|\,\bar{e})}{p(\bar{f}\,|\,\bar{e})} = \frac{p(\bar{f},\bar{a}\,|\,\bar{e})}{\sum_{\bar{a}'} p(\bar{f},\bar{a}'\,|\,\bar{e})}$$

$$\sum_{\bar{a}} \prod_{i=1}^{I} p(f_i\,|\,e_{a_i}) = \sum_{a_1=1}^{J}\sum_{a_2=1}^{J}\cdots\sum_{a_I=1}^{J} \prod_{i=1}^{I} p(f_i\,|\,e_{a_i}) =$$

nasty

$$\sum_{a_1=1}^{2}\sum_{a_2=1}^{2}\prod_{i=1}^{2} p(f_i|e_{a_i}) = p(f_1|e_1)\cdot p(f_2|e_1) + p(f_1|e_1)\cdot p(f_2|e_2) + p(f_1|e_2)\cdot p(f_2|e_1) + p(f_1|e_2)\cdot p(f_2|e_2) =$$

$$= p(f_1|e_1)\cdot\left(p(f_2|e_1)+p(f_2|e_2)\right) + p(f_1|e_2)\cdot\left(p(f_2|e_1)+p(f_2|e_2)\right) = \left(p(f_1|e_1)+p(f_1|e_2)\right)\cdot\left(p(f_2|e_1)+p(f_2|e_2)\right) =$$

$$= \prod_{i=1}^{2}\sum_{a_i=1}^{2} p(f_i|e_{a_i}) = \prod_{i=1}^{2}\sum_{j=1}^{2} p(f_i|e_j)$$

$$= \prod_{i=1}^{I}\sum_{j=1}^{J} p(f_i\,|\,e_j)$$

, DO . . .

ALIGNMENT LIKELIHOOD

$$p(\bar{a}|\bar{e},\bar{f}) = \frac{p(\bar{f},\bar{a}|\bar{e})}{\sum_{\bar{a}'} p(\bar{f},\bar{a}|\bar{e})} = \frac{\prod_{i=1}^{I} p(f_i|e_{a_i})}{\prod_{i=1}^{I}\sum_{j=1}^{J} p(f_i|e_j)} = \prod_{i=1}^{I} \frac{p(f_i|e_{a_i})}{\sum_{j=1}^{J} p(f_i|e_j)}$$

$$p(D) = \prod_{n=1}^{N} \sum_{\bar{a}} p(\bar{f}^{(n)},\bar{a}|\bar{e}^{(n)}) = \prod_{n=1}^{N} \sum_{\bar{a}} p(\cancel{I}|\bar{e}) \prod_{i=1}^{I} p(a_i|\ldots)\cdot p(f_i|e_{a_i})$$

$$\hat{\theta} = \arg\max_{\theta} p(D;\theta)$$

$$\theta_{e,f} = \frac{\text{exp.count}(\langle e,f\rangle)}{\sum_{f'} \text{exp.count}(\langle e,f'\rangle)}$$

sum over $p(a_i|\ldots)$ which depends on θ !!!

WHY DID THE CHICKEN CROSS THE ROAD ?

* we need parameters to get $p(\bar{a}|...)$
* and we need alignments to get the parameters

= no analytic solution like before

WHAT DO WE DO ?

EXPECTATION-MAXIMIZATION

Iteration!

1. Start with some uneducated guess of the param. values
   * e.g. random
   * or uniform
2. Collect exp. counts by computing $p(a_i|\ldots)$ over the whole corpus
3. Re-estimate the parameters using the exp. counts
4. Repeat steps 2 & 3 ad nauseam

EM DETAILS

Step 2: collecting expected counts:

for $n = 1..N$, $i = 1..I^{(n)}$, $j = 1..J^{(n)}$:

$$p(a_i^{(n)} = j \mid ..) = \frac{p(f_i \mid e_j)}{\sum_{k=1}^{J^{(n)}} p(f_i \mid e_k)}$$

$$\text{exp.count}(f_i^{(n)}, e_j^{(n)}) \mathrel{+}= p(a_i^{(n)} = j \mid ..)$$

$$\text{exp.count}(*, e_j^{(n)}) \mathrel{+}= p(a_i^{(n)} = j \mid ..)$$

Step 3: re-estimating parameters:

for f in foreign words, e in english words:

$$\theta_{e,f} = \frac{\text{exp.count}(f, e)}{\text{exp.count}(*, e)}$$

D:

the car

la auto

my car

la mia auto

my house

la mia casa

Expected counts:

	∅	the	my	car	house
la					
mia					
auto					
casa					
*					

$\hat{\Theta}$:

	∅	the	my	car	house
la					
mia					
auto					
casa					

$N = 3$

$f_2^{(2)} = \text{'mia'}$

$a_2^{(2)} =$ ———

D:

the car

la auto

my car

la mia auto

my house

la mia casa

Expected counts:

	∅	the	my	car	house
la					
mia					
auto					
casa					
*					

$\hat{\Theta}$:

	∅	the	my	car	house
la	0.25	0.25	0.25	0.25	0.25
mia	0.25	0.25	0.25	0.25	0.25
auto	0.25	0.25	0.25	0.25	0.25
casa	0.25	0.25	0.25	0.25	0.25

$\Sigma = 1$

Step 1: Initialize parameters

D:

the car

la auto

my car

la mia auto

my house

la mia casa

Expected counts:

	∅	the	my	car	house
la	1.167	0.500	0.667	0.833	0.333
mia	0.667	0.000	0.667	0.333	0.333
auto	0.833	0.500	0.333	0.833	0.000
casa	0.333	0.000	0.333	0.000	0.333
*	3.000	1.000	2.000	2.000	1.000

$\hat{\Theta}$:

	∅	the	my	car	house
la	~~0.25~~	~~0.25~~	~~0.25~~	~~0.25~~	0.25
mia	~~0.25~~	~~0.25~~	~~0.25~~	~~0.25~~	0.25
auto	0.25	[illegible]	[illegible]	[illegible]	0.25
casa	0.25	0.25	[illegible]	[illegible]	0.25

Step 2: collect exp. counts

(iteration 1)

D:

the car

la auto

my car

la mia auto

my house

la mia casa

Expected counts:

	∅	the	my	car	house
la	1.167	0.500	0.667	0.833	0.333
mia	0.667	0.000	0.667	0.333	0.333
auto	0.833	0.500	0.333	0.833	0.000
casa	0.333	0.000	0.333	0.000	0.333
*	3.000	1.000	2.000	2.000	1.000

$\hat{\theta}$:

	∅	the	my	car	house
la	0.389	0.500	0.333	0.417	0.333
mia	0.222	0.000	0.333	0.167	0.333
auto	0.278	0.500	0.167	0.417	0.000
casa	0.111	0.000	0.167	0.000	0.333

Step 3: re-estimate param. distribs

(iteration 1)

D:

the car

la auto

my car

la mia auto

my house

la mia casa

Expected counts:

	∅	the	my	car	house
la	1.559	0.500	0.800	0.917	0.333
mia	0.588	0.000	0.800	0.167	0.333
auto	0.729	0.500	0.200	0.917	0.000
casa	0.154	0.000	0.200	0.000	0.333
*	3.000	1.000	2.000	2.000	1.000

$\hat{\theta}$:

	∅	the	my	car	house
la	0.389	0.500	0.333	0.417	0.333
mia	0.222	0.000	0.333	0.167	0.333
auto	0.278	0.500	0.167	0.417	0.000
casa	0.111	0.000	0.167	0.000	0.333

Step 2: collect exp. counts (from new params)

(Iteration 2)

D:

the car

la auto

my car

la mia auto

my house

la mia casa

Expected counts:

	∅	the	my	car	house
la	1.559	0.500	0.800	0.917	0.333
mia	0.588	0.000	0.800	0.167	0.333
auto	0.729	0.500	0.200	0.917	0.000
casa	0.154	0.000	0.200	0.000	0.333
*	3.000	1.000	2.000	2.000	1.000

$\hat{\Theta}$:

	∅	the	my	car	house
la	0.520	0.500	0.400	0.458	0.333
mia	0.186	0.000	0.400	0.083	0.333
auto	0.243	0.500	0.100	0.458	0.000
casa	0.051	0.000	0.100	0.000	0.333

Step 3: re-estimate parameters (again)

(Iteration 2)

QUESTIONS

* when/how do we stop?

* we have the parameters, how to find the alignments?

* why does that iterative process work?

* are lex. deps enough?

WHY WORKS IT?

ADVANCED MODELS

* what we just did = "IBM model 1"
* model 2 adds $p(i|j,I,J)$
* model 3 adds fertility & NULL-probabilities

• • •

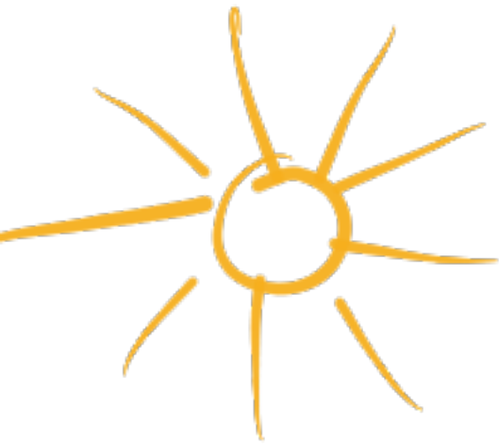
* HMM-based alignment promotes "parallel" consecutive alignments
* agreement-driven alignment promotes agreement for parallel $p(f|e) \times p(e|f)$
* FastAlign is fast

TO CONCLUDE

* word alignment can be done on parallel corpora
* it works by tracking word pair co-occurrences
* it is an essential component in many state-of-the-art SMT paradigms
* and it's FUN

FURTHER INFO

* the SMT book, P. Koehn 2009
* "the mathematics of statistical machine translation: parameter estimation" Brown, Della Pietra x2, Mercer 1993
* "HMM-based word alignment in statistical translation", Vogel, Ney, Tillmann 1996
* "Alignment by agreement" Liang, Taskar, Klein 2006
* "A simple, fast and effective reparameterization of IBM model 2", Dyer, Chahuneau, Smith 2013



THANK YOU!