

Statistical NLP Spring 2010



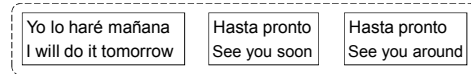
Lecture 17: Word / Phrase MT

Dan Klein – UC Berkeley

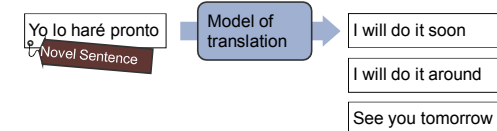
Corpus-Based MT

Modeling correspondences between languages

Sentence-aligned parallel corpus:



Machine translation system:

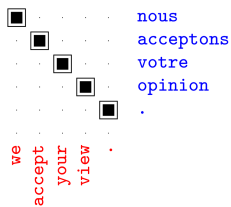


Unsupervised Word Alignment

- Input: a *bitext*: pairs of translated sentences

nous acceptons votre opinion .
we accept your view .

- Output: *alignments*: pairs of translated words
 - When words have unique sources, can represent as a (forward) alignment function a from French to English positions



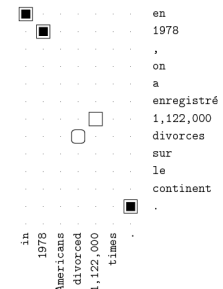
Alignment Error Rate

- Alignment Error Rate

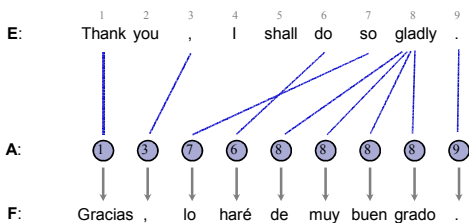
- = Sure
- = Possible
- = Predicted

$$AER(A, S, P) = \left(1 - \frac{|A \cap S| + |A \cap P|}{|A| + |S|}\right)$$

$$= \left(1 - \frac{3 + 3}{3 + 4}\right) = \frac{1}{7}$$



IBM Models 1/2

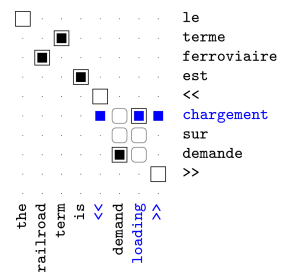


Model Parameters

Emissions: $P(F_1 = \text{Gracias} | E_{A1} = \text{Thank})$ Transitions: $P(A_2 = 3)$

Problems with Model 1

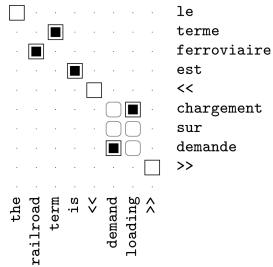
- There's a reason they designed models 2-5!
- Problems: alignments jump around, align everything to rare words
- Experimental setup:
 - Training data: 1.1M sentences of French-English text, Canadian Hansards
 - Evaluation metric: alignment error Rate (AER)
 - Evaluation data: 447 hand-aligned sentences



Intersected Model 1

- Post-intersection: standard practice to train models in each direction then intersect their predictions [Och and Ney, 03]
- Second model is basically a filter on the first
 - Precision jumps, recall drops
 - End up not guessing hard alignments

Model	P/R	AER
Model 1 E→F	82/58	30.6
Model 1 F→E	85/58	28.7
Model 1 AND	96/46	34.8

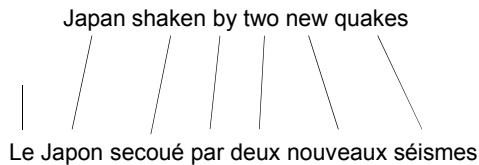


Joint Training?

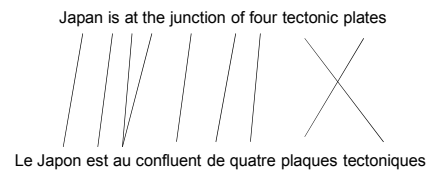
- Overall:
 - Similar high precision to post-intersection
 - But recall is much higher
 - More confident about positing non-null alignments

Model	P/R	AER
Model 1 E→F	82/58	30.6
Model 1 F→E	85/58	28.7
Model 1 AND	96/46	34.8
Model 1 INT	93/69	19.5

Monotonic Translation



Local Order Change



IBM Model 2

- Alignments tend to the diagonal (broadly at least)

$$P(f, a|e) = \prod_j P(a_j = i|j, I, J) P(f_j|e_i)$$

$$P(\text{dist} = i - j \frac{I}{J})$$

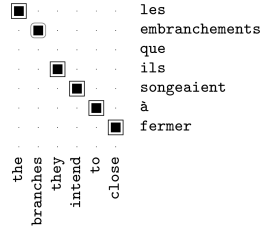
$$\frac{1}{Z} e^{-\alpha(i-j \frac{I}{J})}$$

- Other schemes for biasing alignments towards the diagonal:
 - Relative vs absolute alignment
 - Asymmetric distances
 - Learning a full multinomial over distances

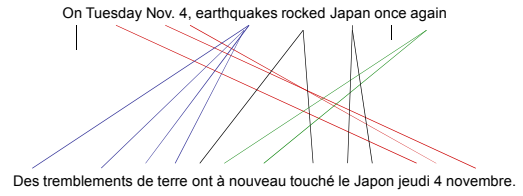
EM for Models 1/2

- Model parameters:
 - Translation probabilities (1+2) $P(f_j|e_i)$
 - Distortion parameters (2 only) $P(a_j = i|j, I, J)$
 - Start with $P(f_j|e_i)$ uniform, including $P(f_j|null)$
 - For each sentence:
 - For each French position j
 - Calculate posterior over English positions
- $$P(a_j = i|f, e) = \frac{P(a_j = i|j, I, J) P(f_j|e_i)}{\sum_{i'} P(a_j = i'|j, I, J) P(f_j|e_{i'})}$$
- (or just use best single alignment)
 - Increment count of word f_i with word e_i by these amounts
 - Also re-estimate distortion probabilities for model 2
- Iterate until convergence

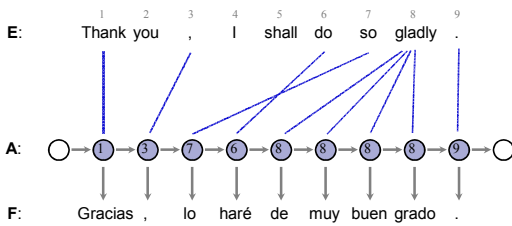
Example: Model 2 Helps



Phrase Movement



The HMM Model



Model Parameters
 Emissions: $P(F_1 = \text{Gracias} | E_{A_1} = \text{Thank})$ Transitions: $P(A_2 = 3 | A_1 = 1)$

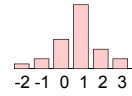
The HMM Model

- Model 2 preferred global monotonicity
- We want local monotonicity:
 - Most jumps are small
- HMM model (Vogel 96)

f	$t(f e)$
nationale	0.469
national	0.418
nationaux	0.054
nationales	0.029

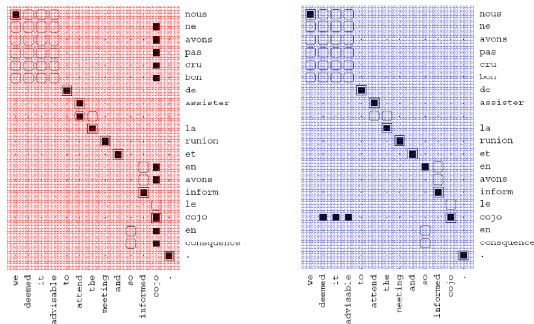
$$P(f, a|e) = \prod_j P(a_j | a_{j-1}) P(f_j | e_i)$$

$$P(a_j - a_{j-1})$$



- Re-estimate using the forward-backward algorithm
- Handling nulls requires some care
- What are we still missing?

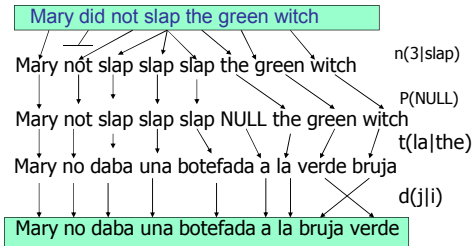
HMM Examples



AER for HMMs

Model	AER
Model 1 INT	19.5
HMM E→F	11.4
HMM F→E	10.8
HMM AND	7.1
HMM INT	4.7
GIZA M4 AND	6.9

IBM Models 3/4/5



[from Al-Onaizan and Knight, 1998]

Examples: Translation and Fertility

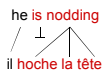
f	t(f e)	ϕ	n(ϕe)
le	0.497	1	0.746
la	0.207	0	0.254
les	0.155		
l'	0.086		
ce	0.018		
cette	0.011		

f	t(f e)	ϕ	n(ϕe)
ne	0.497	2	0.735
pas	0.442	0	0.154
non	0.029	1	0.107
rien	0.011		

farmers

f	t(f e)	ϕ	n(ϕe)
agriculteurs	0.442	2	0.731
les	0.418	1	0.228
cultivateurs	0.046	0	0.039
producteurs	0.021		

Example: Idioms



nodding

f	t(f e)	ϕ	n(ϕe)
signe	0.164	4	0.342
la	0.123	3	0.293
tête	0.097	2	0.167
oui	0.086	1	0.163
fait	0.073	0	0.023
que	0.073		
hoche	0.054		
hocher	0.048		
faire	0.030		
me	0.024		
approuve	0.019		
qui	0.019		
un	0.012		
faites	0.011		

Example: Morphology

should

f	t(f e)	ϕ	n(ϕe)
devrait	0.330	1	0.649
devraient	0.123	0	0.336
devrions	0.109	2	0.014
faudrait	0.073		
faut	0.058		
doit	0.058		
aurait	0.041		
doivent	0.024		
devons	0.017		
devrais	0.013		

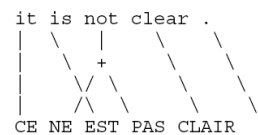
Some Results

- [Och and Ney 03]

Model	Training scheme	0.5K	8K	128K	1.47M
Dice		50.9	43.4	39.6	38.9
Dice+C		46.3	37.6	35.0	34.0
Model 1	1^5	40.6	33.6	28.6	25.9
Model 2	$1^5 2^5$	46.7	29.3	22.0	19.5
HMM	$1^5 H^5$	26.3	23.3	15.0	10.8
Model 3	$1^5 2^5 3^5$	43.6	27.5	20.5	18.0
	$1^5 H^2 3^3$	27.5	22.5	16.6	13.2
Model 4	$1^5 2^5 3^4 3^3$	41.7	25.1	17.3	14.1
	$1^5 H^2 3^3 4^3$	26.1	20.2	13.1	9.4
	$1^5 H^2 4^3$	26.3	21.8	13.3	9.3
Model 5	$1^5 H^2 4^2 5^3$	26.5	21.5	13.7	9.6
	$1^5 H^2 3^3 4^2 5^3$	26.5	20.4	13.4	9.4
Model 6	$1^5 H^2 4^2 6^3$	26.0	21.6	12.8	8.8
	$1^5 H^2 3^3 4^2 6^3$	25.9	20.3	12.5	8.7

Decoding

- In these word-to-word models
 - Finding best alignments is easy
 - Finding translations is hard (why?)



Bag "Generation" (Decoding)

Exact reconstruction (24 of 38)

Please give me your response as soon as possible.
 ⇒ Please give me your response as soon as possible.

Reconstruction preserving meaning (8 of 38)

Now let me mention some of the disadvantages.
 ⇒ Let me mention some of the disadvantages now.

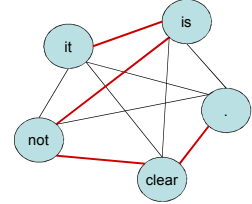
Garbage reconstruction (6 of 38)

In our organization research has two missions.
 ⇒ In our missions research organization has two.

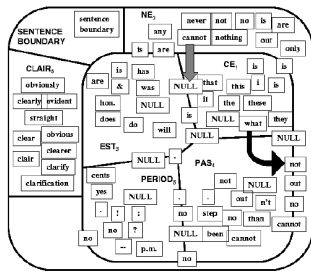
Bag Generation as a TSP

Imagine bag generation with a bigram LM

- Words are nodes
- Edge weights are $P(w|w')$
- Valid sentences are Hamiltonian paths
- Not the best news for word-based MT!



IBM Decoding as a TSP



Greedy Decoding

NULL well heard , it talks a great victory .
 bien entendu , il parle de une belle victoire .
 translateTwoWords(2,understood,0,about)

NULL well understood , it talks about a great victory .
 bien entendu , il parle de une belle victoire .
 translateOneWord(4,he)

NULL well understood , he talks about a great victory .
 bien entendu , il parle de une belle victoire .
 translateTwoWords(1,quite,2,naturally)

NULL quite naturally , he talks about a great victory .
 bien entendu , il parle de une belle victoire .

Stack Decoding

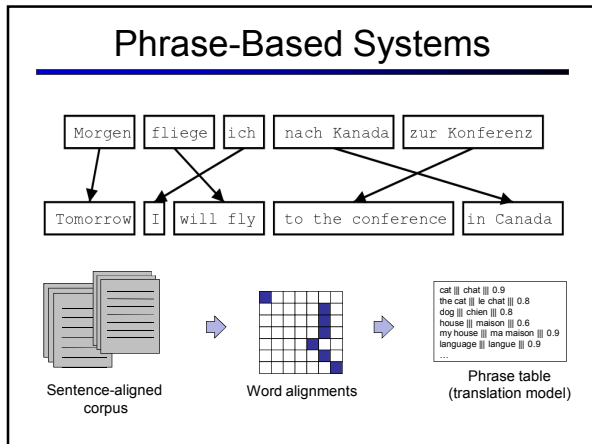
- Stack decoding:
 - Beam search
 - Usually A^* estimates for completion cost
 - One stack per candidate sentence length
- Other methods:
 - Dynamic programming decoders possible if we make assumptions about the set of allowable permutations

sent length	decoder type	time (sec/sent)	search errors	translation errors (semantic and/or syntactic)	NE	PME	DSE	FSE	HSE	CE
6	IP	47.50	0	57	44	57	0	0	0	0
6	stack	0.79	5	58	43	53	1	0	0	4
6	greedy	0.07	18	60	38	45	5	2	1	10
8	IP	495.00	0	75	29	74	0	0	0	0
8	stack	5.67	20	75	24	57	1	2	2	15
8	greedy	2.66	43	75	20	38	4	5	1	33

Stack Decoding

- Stack decoding:
 - Beam search
 - Usually A^* estimates for completion cost
 - One stack per candidate sentence length
- Other methods:
 - Dynamic programming decoders possible if we make assumptions about the set of allowable permutations

sent length	decoder type	time (sec/sent)	search errors	translation errors (semantic and/or syntactic)	NE	PME	DSE	FSE	HSE	CE
6	IP	47.50	0	57	44	57	0	0	0	0
6	stack	0.79	5	58	43	53	1	0	0	4
6	greedy	0.07	18	60	38	45	5	2	1	10
8	IP	495.00	0	75	29	74	0	0	0	0
8	stack	5.67	20	75	24	57	1	2	2	15
8	greedy	2.66	43	75	20	38	4	5	1	33

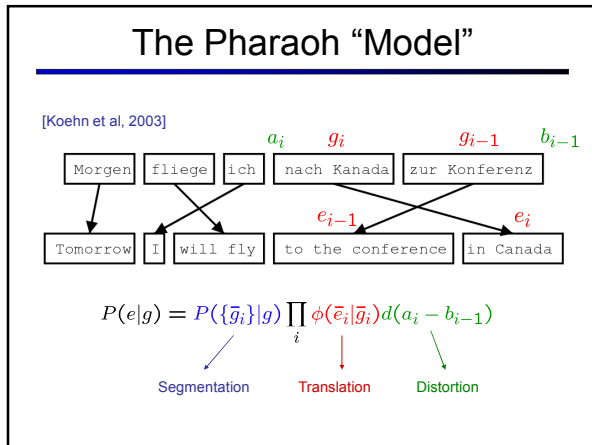


Phrase-Based Decoding

这 7人 中包括 来自 法国 和 俄罗斯 的 宇航 员 .

the	7	people	including	by	some	and	the	russian	the	astronauts	international	astronautical	.	.	
a	7	people	included	by	france	and	the	russian	the	astronauts	international	astronautical	of	rapporteur .	
this	7	out	including	the	from	the	franch	and	the	russian	the	frith	.	.	
these	7	among	including	from		the	franch	and	of	the	ruissian	of	space	members .	
that	7	person	including	from	the	of	france	and	aid	to	russian	of	the	airspce	members .
7	include					from	the	of	france	and	russian	of	the	astronauts	the
7	members	include				from	france	and	and	russian	of	the	astronauts	who	.
7	populations	include				those	from	france	and	russian	in	the	astronauts	who	.
7	experts	included				come	from	france	and	russia	in	the	astronautical	personnel	.
7	phylum	including	those	from		france	and	russia	in	a	space	astronaut	member	.	.
						including	representatives	from	france	and	the	russian	astronaut		.
						include	come	from	france	and	russia	by	cosmonauts		.
						include	representatives	from	franch	and	russia	cosmonauts			.
						include	come	from	france	and	russia	cosmonauts			.
						includes	coming	from	franch	and	russia	cosmonaut			.
									franch	and	russian	's	astronaut	member	.
									franch	and	russia	's	astronauts		.
									franch	and	russia	's	astronauts	special	rapporteur
									and	russia	astronauts			rapporteur	.
									and	russia				rapporteur .	.
									and	russia				rapporteur .	.
									or	russia	's				.

Decoder design is important: [Koehn et al. 03]



The Pharaoh "Model"

$$P(f|e) = P(\{\bar{e}_i\}|e) \prod_i \phi(\bar{f}_i|\bar{e}_i) d(a_i - b_{i-1})$$

$\frac{1}{K}$

$\frac{\text{count}(\bar{f}_i, \bar{e}_i)}{\text{count}(\bar{e}_i)}$

$\alpha^{|a_i - b_{i-1}|}$

Where do we get these counts?

Phrase Weights

How the MT community estimates $P(\bar{f}|\bar{e})$

Parallel training sentences

provide phrase pair counts.

Gracias ,lo haré de muy buen grado .
Thank you , I shall do so gladly .

▶

lo haré ↔ I shall do so
44 times in the corpus

All phrase pairs are counted,

and counts are normalized.

Gracias ,lo haré de muy buen grado
Thank you , I shall do so gladly .

▶

$P(\bar{f}|\bar{e}) = \frac{\text{count}(\bar{f}, \bar{e})}{\text{count}(\bar{e})}$

Counting Phrase Pairs

Input:
Gracias ,lo haré de muy buen grado .
Thank you , I shall do so gladly .

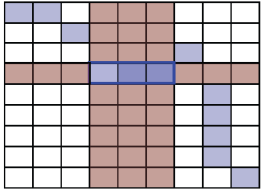
▶

Gloss
Thanks
,
that
do [first; future]
of
very
good
degree

First, we learn word alignments,

▶

then we infer aligned phrases.

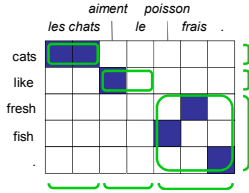


▶

Thank you , I shall do so gladly .

Phrase Scoring

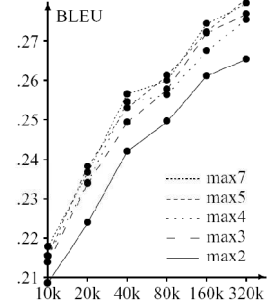
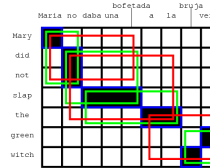
$$\phi_{\text{trans}}(\bar{e}_j | \bar{f}_i) = \frac{c(\bar{f}_i, \bar{e}_j)}{c(\bar{f}_i)}$$



- Learning weights has been tried, several times:
 - [Marcu and Wong, 02]
 - [DeNero et al, 06]
 - ... and others
- Seems not to work well, for a variety of partially understood reasons
- Main issue: big chunks get all the weight, obvious priors don't help
 - Though, [DeNero et al 08]

Phrase Size

- Phrases do help
 - But they don't need to be long
 - Why should this be?



Lexical Weighting

$$\phi(\bar{f}_i | \bar{e}_i) = \frac{\text{count}(\bar{f}_i, \bar{e}_i)}{\text{count}(\bar{e}_i)} p_w(\bar{f}_i | \bar{e}_i)$$

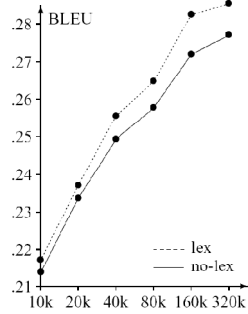
f1 f2 f3
 NULL -- -- ##
 e1 ## -- --
 e2 -- ## --
 e3 -- ## --

$$p_w(\bar{f} | \bar{e}, a) = p_w(f_1 f_2 f_3 | e_1 e_2 e_3, a)$$

$$= w(f_1 | e_1)$$

$$\times \frac{1}{2} (w(f_2 | e_2) + w(f_2 | e_3))$$

$$\times w(f_3 | \text{NULL})$$

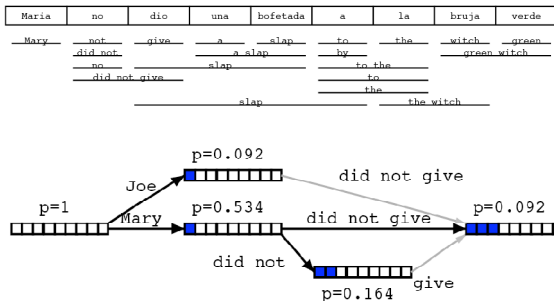


The Pharaoh Decoder



- Probabilities at each step include LM and TM

Hypothesis Lattices



Pruning

Maria no dio una bofetada a la bruja verde

e: Mary did not
 f: **-----
 p: 0.154

better
 partial
 translation

e: the
 f: -----*---
 p: 0.354

covers
 easier part
 --> lower cost

- Problem: easy partial analyses are cheaper
 - Solution 1: use beams per foreign subset
 - Solution 2: estimate forward costs (A*-like)

WSD?

- Remember when we discussed WSD?
 - Word-based MT systems rarely have a WSD step
 - Why not?