

CS395T: Structured Models for NLP

Lecture 12: Machine Translation



Greg Durrett

Adapted from Dan Klein – UC Berkeley

Machine Translation



Administrivia

Project 2 due one week from today!

P1 test set results: top 3

Su Wang: 84.03 F1 (86.10 P / 82.05 R)

Larger window size and Wikipedia gazetteer

Prateek Shrishail Kolhar: 82.32 F1 (82.61 P / 82.07 R)

Conjunctions of words, POS, and shapes in neighborhood
Very fast vectorized implementation (15s per epoch)

Yasumasa Onoe: 78.55 F1 (78.27 P / 78.83 R)

Used transition probabilities from HMM, character
5-grams and other feature tuning



Machine Translation: Examples

Atlanta, preso il killer del palazzo di Giustizia

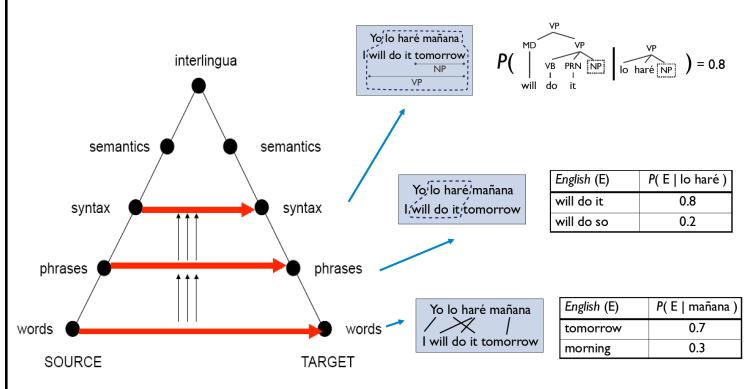
ATLANTA - La grande paura che per 26 ore ha attanagliato Atlanta è finita: Brian Nichols, l'uomo che aveva ucciso tre persone a palazzo di Giustizia e che **ha poi ucciso un agente di dogana**, s'è consegnato alla polizia, dopo avere cercato rifugio nell'alloggio di una donna in un complesso d'appartamenti alla periferia della città. Per tutto il giorno, il centro della città, sede della **Coca Cola** e dei Giochi 1996, cuore di una popolosa **area metropolitana**, era rimasto paralizzato.

Atlanta, taken the killer of the palace of Justice

ATLANTA - The great fear that for 26 hours has gripped Atlanta is ended: Brian Nichols, the man who had killed three persons to palace of Justice and that **a customs agent has then killed**, s'is delivered to the police, after to have tried shelter in the lodging of one woman in a complex of apartments to the periphery of the city. For all the day, the center of the city, center of the **Coke Strains** and of Giochi 1996, heart of one popolosa **metropolitan area**, was remained paralyzed.



Levels of Transfer



Word-Level MT: Examples

la politique de la haine .

politics of hate .

the policy of the hatred .

(Foreign Original)

(Reference Translation)

(IBM4+N-grams+Stack)

nous avons signé le protocole .

we did sign the memorandum of agreement .

we have signed the protocol .

(Foreign Original)

(Reference Translation)

(IBM4+N-grams+Stack)

où était le plan solide ?

but where was the solid plan ?

where was the economic base ?

(Foreign Original)

(Reference Translation)

(IBM4+N-grams+Stack)



Phrasal MT: Examples

Le président américain Barack Obama doit annoncer lundi de nouvelles mesures en faveur des constructeurs automobile. General motors et Chrysler avaient déjà bénéficié fin 2008 d'un prêt d'urgence cumulé de 17,4 milliards de dollars, et ont soumis en février au Trésor un plan de restructuration basé sur un total de 22 milliards de dollars d'aides publiques supplémentaires.

U.S. President Barack Obama to announce Monday new measures to help automakers. General Motors and Chrysler had already received late in 2008 a cumulative emergency loan of 17.4 billion dollars, and submitted to the Treasury in February in a restructuring plan based on a total of 22 billion dollars in additional aid .

Metrics



MT: Evaluation

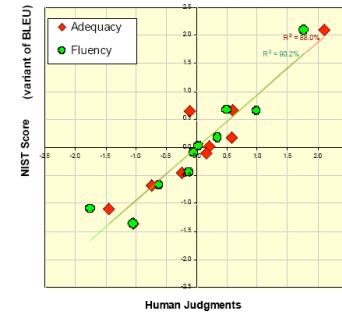
- Human evaluations: subject measures, fluency/ adequacy
- Automatic measures: n-gram match to references
 - NIST measure: n-gram recall (worked poorly)
 - BLEU: n-gram precision (no one really likes it, but everyone uses it)
 - Lots more: TER, HTER, METEOR, ...
- BLEU:
 - P1 = unigram precision
 - P2, P3, P4 = bi-, tri-, 4-gram precision
 - Weighted geometric mean of P1-4
 - Brevity penalty (why?)
 - Somewhat hard to game...
 - Magnitude only meaningful on same language, corpus, number of references, probably only within system types...

Reference (human) translation:
The U.S. Island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport.

Machine translation:
The American [?] international airport and its office all received one calls self the sand Arab riot business [?] and so on electronic mail, which sends out ; The threat will be able after public place and so on the airport to start the biochemistry attack, [?] highly alerts after the maintenance.



Automatic Metrics Work (?)



slide from G. Doddington (NIST)

Systems Overview



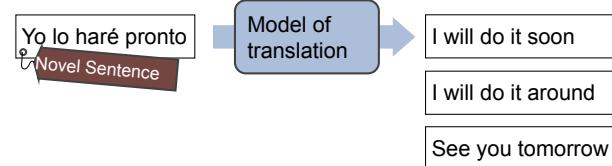
Corpus-Based MT

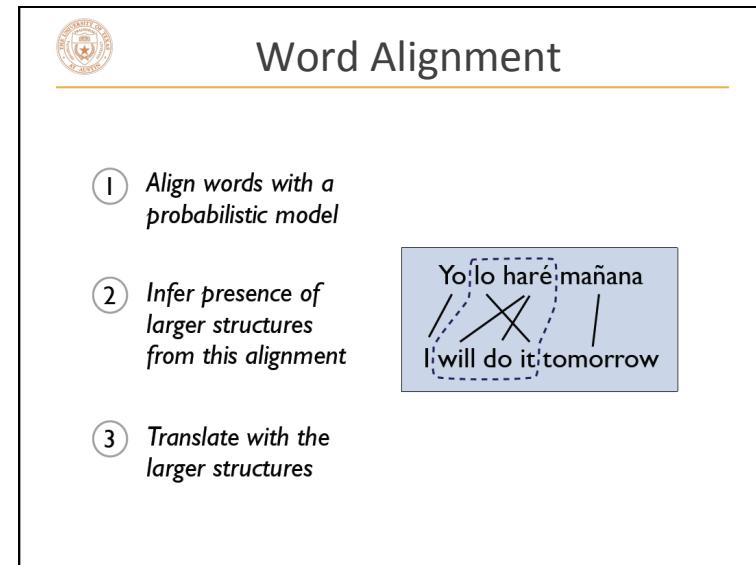
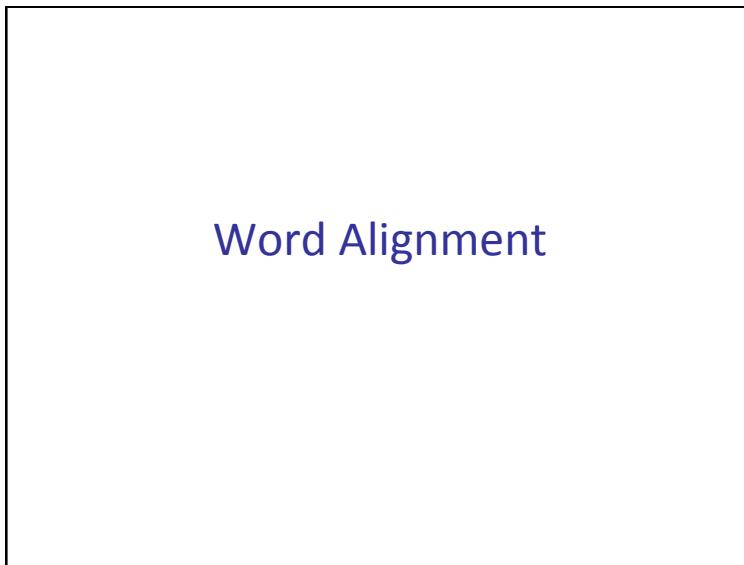
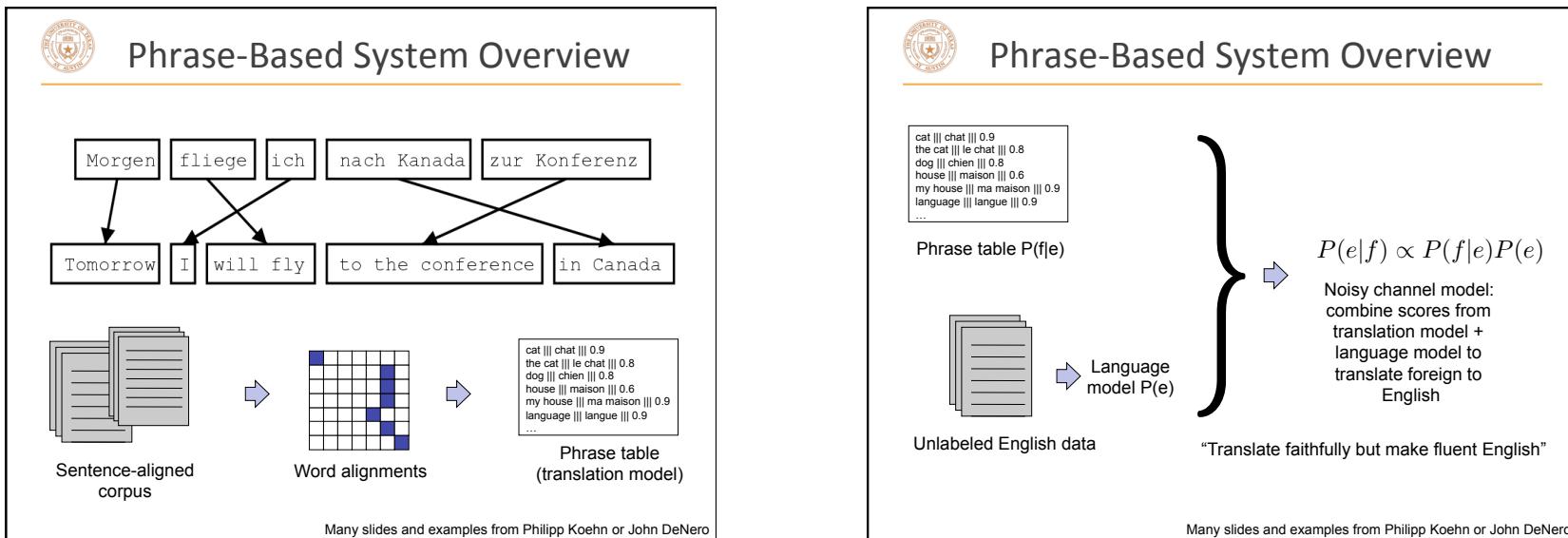
Modeling correspondences between languages

Sentence-aligned parallel corpus:

Yo lo haré mañana	Hasta pronto	Hasta pronto
I will do it tomorrow	See you soon	See you around

Machine translation system:





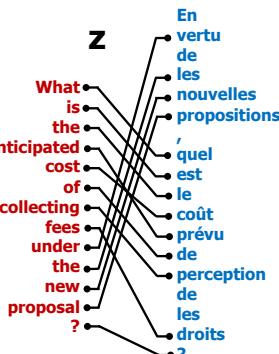


Word Alignment

X

What is the anticipated cost of collecting fees under the new proposal?

En vertu des nouvelles propositions, quel est le coût prévu de perception des droits?

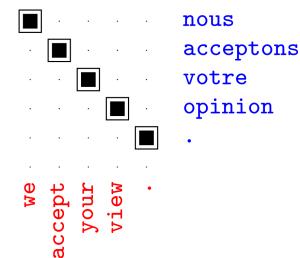


Unsupervised Word Alignment

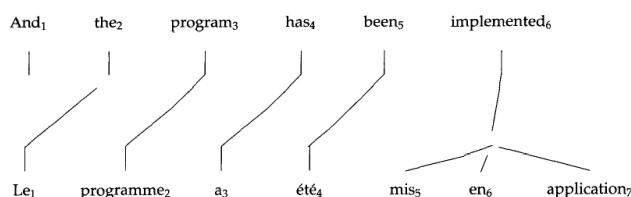
- Input: a *bilingual text*: pairs of translated sentences

nous acceptons votre opinion .
we accept your view .

- Output: *alignments*: pairs of translated words
- Not always one-to-one!



1-to-Many Alignments



Evaluating Models

- How do we measure quality of a word-to-word model?
 - Method 1: use in an end-to-end translation system
 - Slow development cycle
 - Misleading if your MT system was “tuned” for certain aspects of bad alignments
 - Method 2: measure quality of the alignments produced
 - Easy to measure
 - Hard to know what the gold alignments should be
 - Often does not correlate well with translation quality



Alignment Error Rate

- Alignment Error Rate

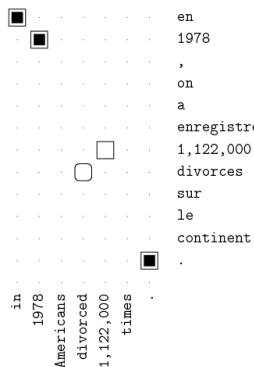
\square = Sure

\square = Possible

\blacksquare = Predicted

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

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

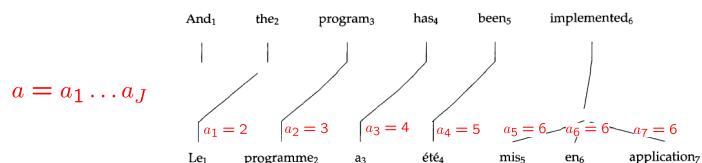


IBM Model 1



IBM Model 1 (Brown 93)

- Alignments: a hidden vector called an *alignment* specifies which English source (or a special *null token*) is responsible for each French target word.



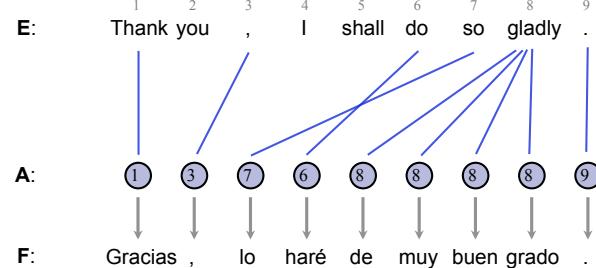
$$P(f, a|e) = \prod_j P(a_j = i) P(f_j|e_i)$$

$$= \prod_j \frac{1}{I+1} P(f_j|e_i)$$

$$P(f|e) = \sum_a P(f, a|e)$$



IBM Model 1



Model Parameters

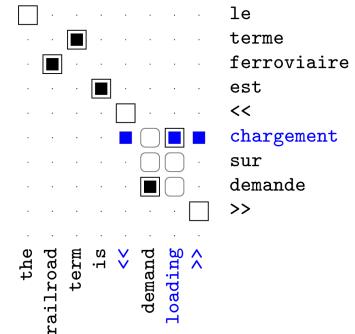
$P(A_1 = 1) = 1/10$, nothing to learn

$P(F_1 = \text{Gracias} | A_1 = 1) = P(\text{Gracias} | \text{Thank})$ <- learn these translation probs



EM for Model 1

- Model 1 Parameters:
Translation probabilities $P(f_j|e_i)$
 - Start with $P(f_j|e_i)$ uniform, including $P(f_j|null)$
 - For each sentence, for each foreign position j :
 - Calculate posterior over English positions
- $$P(a_j = i|\mathbf{f}, \mathbf{e}) = \frac{P(f_j|e_i)}{\sum_{i'} P(f_j|e'_i)}$$
- Increment count of word f_j with word e_i by these amounts
 - Do for whole corpus, re-estimate $P(\mathbf{f}|\mathbf{e})$ with M-step



Problems with Model 1

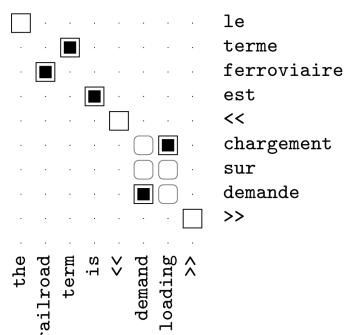
- There's a reason they designed models 2-5!
- Problems: alignments jump around, align everything to rare words
- Experimental setup:
 - Training data: 1.1M sentences of French-English text, Canadian Hansards
 - Evaluation metric: alignment error Rate (AER)
 - Evaluation data: 447 hand-aligned sentences



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



HMM Model: Local Monotonicity



Monotonic Translation

Japan shaken by two new quakes

 Le Japon secoué par deux nouveaux séismes



Local Order Change

Japan is at the junction of four tectonic plates

 Le Japon est au confluent de quatre plaques tectoniques



The HMM Model

- 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

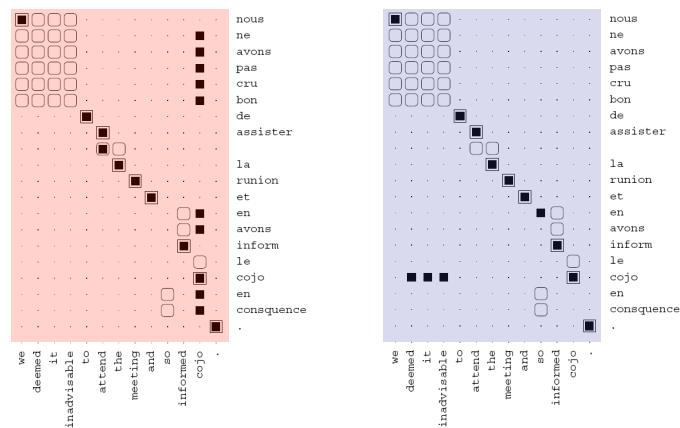
$$P(f, a|e) = \prod_j P(a_j|a_{j-1}) P(f_j|e_i)$$

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

- Re-estimate using the forward-backward algorithm



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

Language Modeling



Phrase-Based System Overview

```
cat ||| chat ||| 0.9  
the cat ||| le chat ||| 0.8  
dog ||| chien ||| 0.8  
house ||| maison ||| 0.6  
my house ||| ma maison ||| 0.9  
language ||| langue ||| 0.9  
...
```

Phrase table $P(f|e)$



Language model $P(e)$

Unlabeled English data

$$P(e|f) \propto P(f|e)P(e)$$

Noisy channel model:
combine scores from
translation model +
language model to
translate foreign to
English

“Translate faithfully but make fluent English”

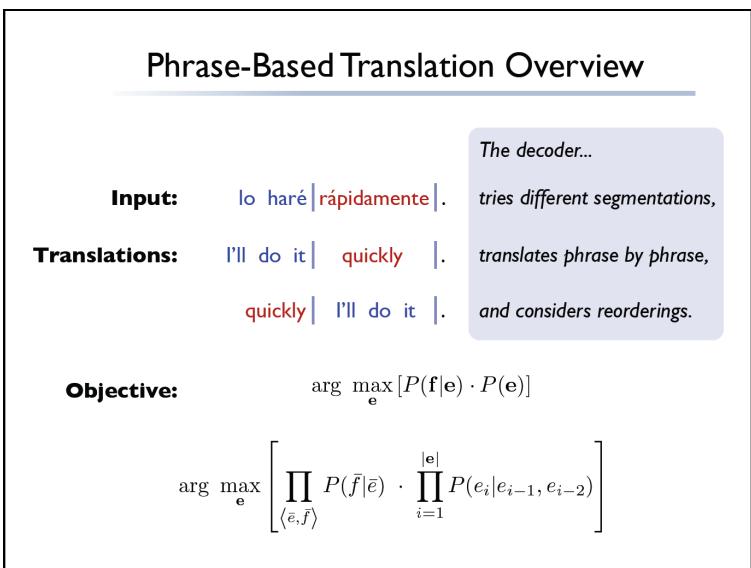
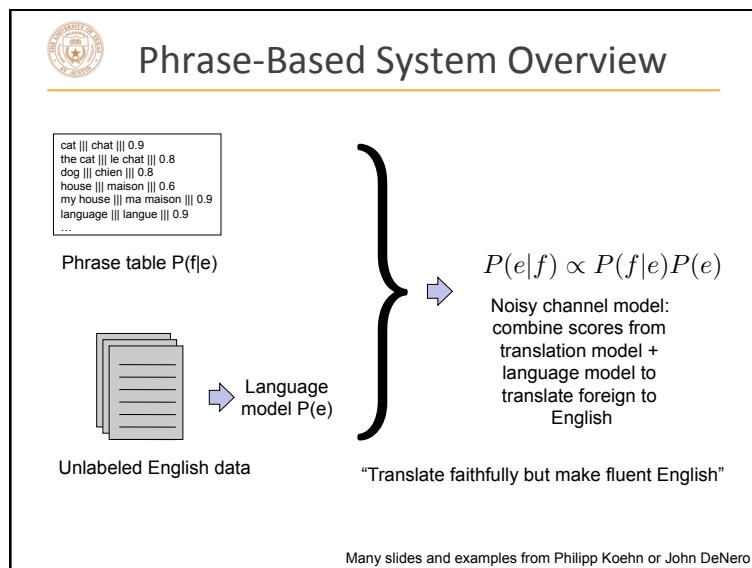
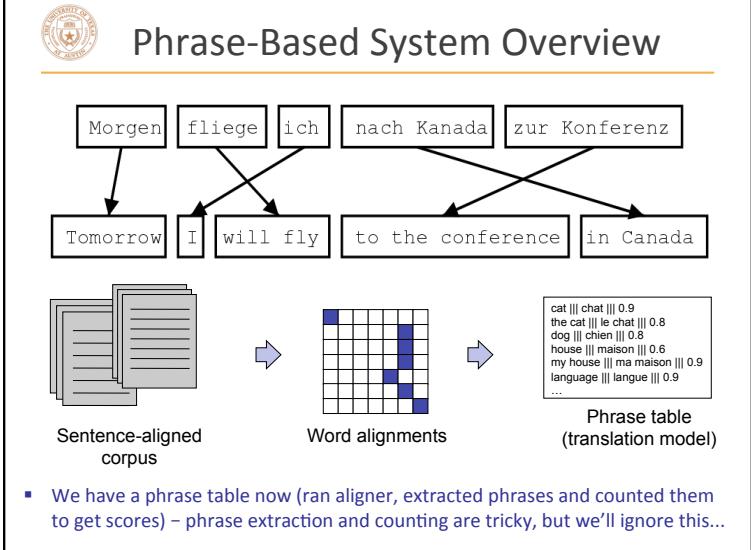
Many slides and examples from Philipp Koehn or John DeNero



N-gram Language Modeling

- Could give several lectures on this!
- Estimate $P(w_n | w_{n-k}, w_{n-k+1}, \dots, w_{n-1})$
- Generative model: read off counts and normalize
 - $P(\text{fox} | \text{the quick brown}) = 0.9$, etc.
- Very complex distributions, need to smooth
 - Interpolate with lower-order models
 - Lots of complex techniques

Phrase-Based MT





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 the russian the the astronauts of rapporteur .										
it	7 people included	by france	and the	the russian	the	the astronauts	of rapporteur .			
this	7 out	including the from	the french and the russian	the fifth	international astronomical					
these	7 among	including from	the french and of the russian	of space	members					
that	7 persons	including from the	of france and to russian	of the aerospace	members					
includes	including from the	of france and to the russian	astronauts							
7 members include	from the	and russian	of astronauts who							
7 populations include	those from france	and russian	astronauts							
7 deportees included	come from	france and russia	in astronomical	personnel						
7 philtum	including those from	france and russia	a space	member						
include	came from	france and russia	by cosmonauts							
include representatives from	french and russia		cosmonauts							
include	came from france	and russia's	cosmonauts							
includes	coming from	french and russia's	cosmonaut							
		french and russian	's	astronavigation	member					
		french	and russia	astronauts						
			and russia's		special rapporteur					
			, and russia		rapporteur					
			, and russia		rapporteur					
			, and russia							
			or russia's							

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



Phrase-Based Decoding

Maria	no	dio	una	bofetada	a	la	bruja	verde
Mary	not	give	a	slap	to	the	witch	green
	did not				by			
	no							
	did not give							
							slap	the witch

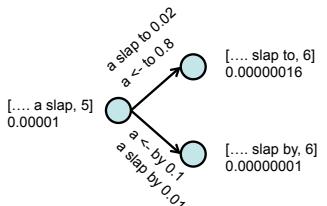


Monotonic Word Translation

Maria	no	dio	una	bofetada	a	la	bruja	verde
Mary	not	give	a	slap	to	the	witch	green

- Cost is $LM * TM$
- It's an HMM?
- State includes
 - Exposed English
 - Position in foreign
- Dynamic program loop?

```
for (fPosition in 1...!f)
  for (eContext in allEContexts)
    for (eOption in translations[fPosition])
      score = scores[fPosition-1][eContext] * LM(eContext+eOption) * TM(eOption, fWord[fPosition])
      scores[fPosition][eContext[2]+eOption] = max(score)
```



Beam Decoding

- For real MT models, this kind of dynamic program is a disaster (why?)
- Standard solution is beam search: for each position, keep track of only the best k hypotheses

```
for (fPosition in 1...!f)
  for (eContext in bestEContexts[fPosition])
    for (eOption in translations[fPosition])
      score = scores[fPosition-1][eContext] * LM(eContext+eOption) * TM(eOption, fWord[fPosition])
      bestEContexts.maybeAdd(eContext[2]+eOption, score)
```

Example from David Chiang



Phrase Translation

Maria	no	dio	una	bofetada	a	la	bruja	verde
Mary	not	give	a	slap	to	the	witch	green
	did not			a slap	by		green	witch
	no		slap		to the			
	did not give				to			
					the			
			slap			the	witch	

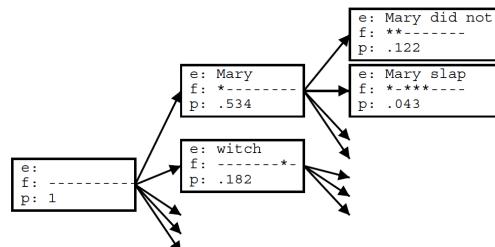
- If monotonic, almost an HMM; technically a semi-HMM

```
for (fPosition in 1...|f|)
    for (lastPosition < fPosition)
        for (eContext in eContexts)
            for (eOption in translations[fPosition])
                ... combine hypothesis for (lastPosition ending in eContext) with eOption
```

- If distortion... now what?



Non-Monotonic Phrasal MT



Pruning: Beams + Forward Costs

Maria	no	dio	una	bofetada	a	la	bruja	verde
	-----			-----				

Two hypotheses are shown with their forward costs:

- $e: \text{Mary did not}$, $f: *----$, $p: 0.154$
- $e: \text{the}$, $f: -----*$, $p: 0.354$

Annotations below the hypotheses:

- "better partial translation" next to the first hypothesis
- "covers easier part --> lower cost" next to the second hypothesis

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



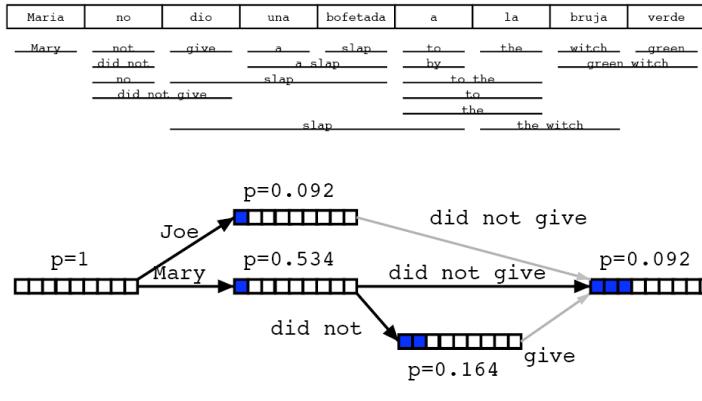
The Pharaoh Decoder

Maria	no	dio	una	bofetada	a	la	bruja	verde
Mary	not	give	a	slap	to	the	witch	green
	did not			a slap	by		green	witch
	no		slap		to the			
	did not give				to			
					the			
			slap			the	witch	

Maria	no	dio	una	bofetada	a	la	bruja	verde
Mary	did not	slap	the	green				



Hypothesis Lattices



Syntactic Models

Translating with Tree Transducers

Input
lo haré de muy buen grado .

Grammar

Translating with Tree Transducers

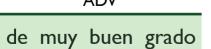
Input
lo haré de muy buen grado .

Grammar

ADV → < de muy buen grado ; gladly >

Translating with Tree Transducers

Input

lo haré  ADV de muy buen grado .

Output

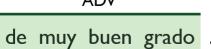
ADV
|
gladly

Grammar

ADV → < de muy buen grado ; gladly >

Translating with Tree Transducers

Input

lo haré  ADV de muy buen grado .

ADV
|
gladly

Grammar

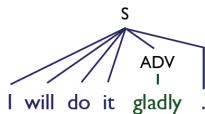
$s \rightarrow \langle \text{lo haré ADV . ; I will do it ADV .} \rangle$
 ADV → < de muy buen grado ; gladly >

Translating with Tree Transducers

Input

S  ADV  de muy buen grado .

Output

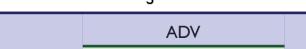


Grammar

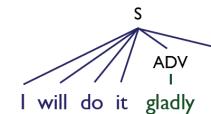
$s \rightarrow \langle \text{lo haré ADV . ; I will do it ADV .} \rangle$
 ADV → < de muy bien grado ; gladly >

Translating with Tree Transducers

Input

S  ADV  de muy bien grado .

Output

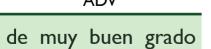


Grammar

$s \rightarrow \langle \text{lo haré ADV . ; I will do it ADV .} \rangle$
 ADV → < de muy bien grado ; gladly >

Translating with Tree Transducers

Input

lo haré  ADV
de muy buen grado .

Output

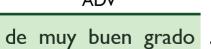
ADV
|
gladly

Grammar

$S \rightarrow \langle \text{lo haré} \text{ ADV} . ; \text{I will do it ADV} \rangle$
 $\text{ADV} \rightarrow \langle \text{de muy buen grado} ; \text{gladly} \rangle$

Translating with Tree Transducers

Input

lo haré  ADV
de muy buen grado .

Output

ADV
|
gladly

Grammar

$VP \rightarrow \langle \text{lo haré} \text{ ADV} ; \text{will do it ADV} \rangle$
 $S \rightarrow \langle \text{lo haré} \text{ ADV} . ; \text{I will do it ADV} \rangle$
 $\text{ADV} \rightarrow \langle \text{de muy buen grado} ; \text{gladly} \rangle$

Translating with Tree Transducers

Input

VP
ADV
lo haré de muy buen grado .

Output

VP
|
will do it
ADV
|
gladly

Grammar

$VP \rightarrow \langle \text{lo haré} \text{ ADV} ; \text{will do it ADV} \rangle$
 $S \rightarrow \langle \text{lo haré} \text{ ADV} . ; \text{I will do it ADV} \rangle$
 $\text{ADV} \rightarrow \langle \text{de muy buen grado} ; \text{gladly} \rangle$

Translating with Tree Transducers

Input

VP
ADV
lo haré de muy buen grado .

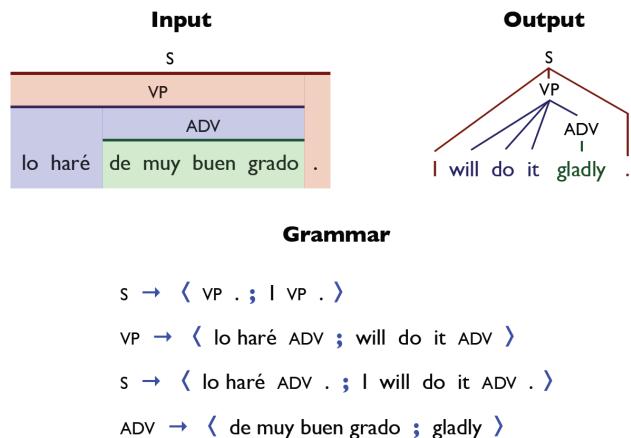
Output

VP
|
will do it
ADV
|
gladly

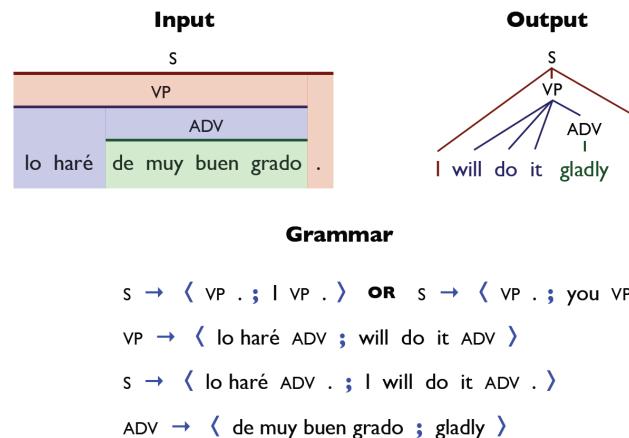
Grammar

$S \rightarrow \langle VP . ; I VP . \rangle$
 $VP \rightarrow \langle \text{lo haré} \text{ ADV} ; \text{will do it ADV} \rangle$
 $S \rightarrow \langle \text{lo haré} \text{ ADV} . ; \text{I will do it ADV} \rangle$
 $\text{ADV} \rightarrow \langle \text{de muy buen grado} ; \text{gladly} \rangle$

Translating with Tree Transducers



Translating with Tree Transducers



Syntactic Translation

- Lots of complexity: large phrase tables, errors introduced by parsers, parses don't agree, inference is harder, ...
- Good for some languages (Japanese->English), but generally more trouble than it's worth
- Easier method: syntactic “pre-reordering”



MT: Takeaways

- Word alignments: unsupervised process for finding word-level correspondences. Turn these into phrase level correspondences -> phrase table
- Language model: estimate n-gram model on a very large corpus
- Translation process: use beam search to find the best translation $\operatorname{argmax}_e P(f|e)P(e)$