COMP 790.139 (Fall 2017)
Natural Language Processing

Machine Translation 2; Guest Task; Coding-HW2 Discussion

THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL

Mohit Bansal

(various slides adapted/borrowed from courses by Dan Klein, JurafskyMartin-SLP3, Manning/Socher, others)

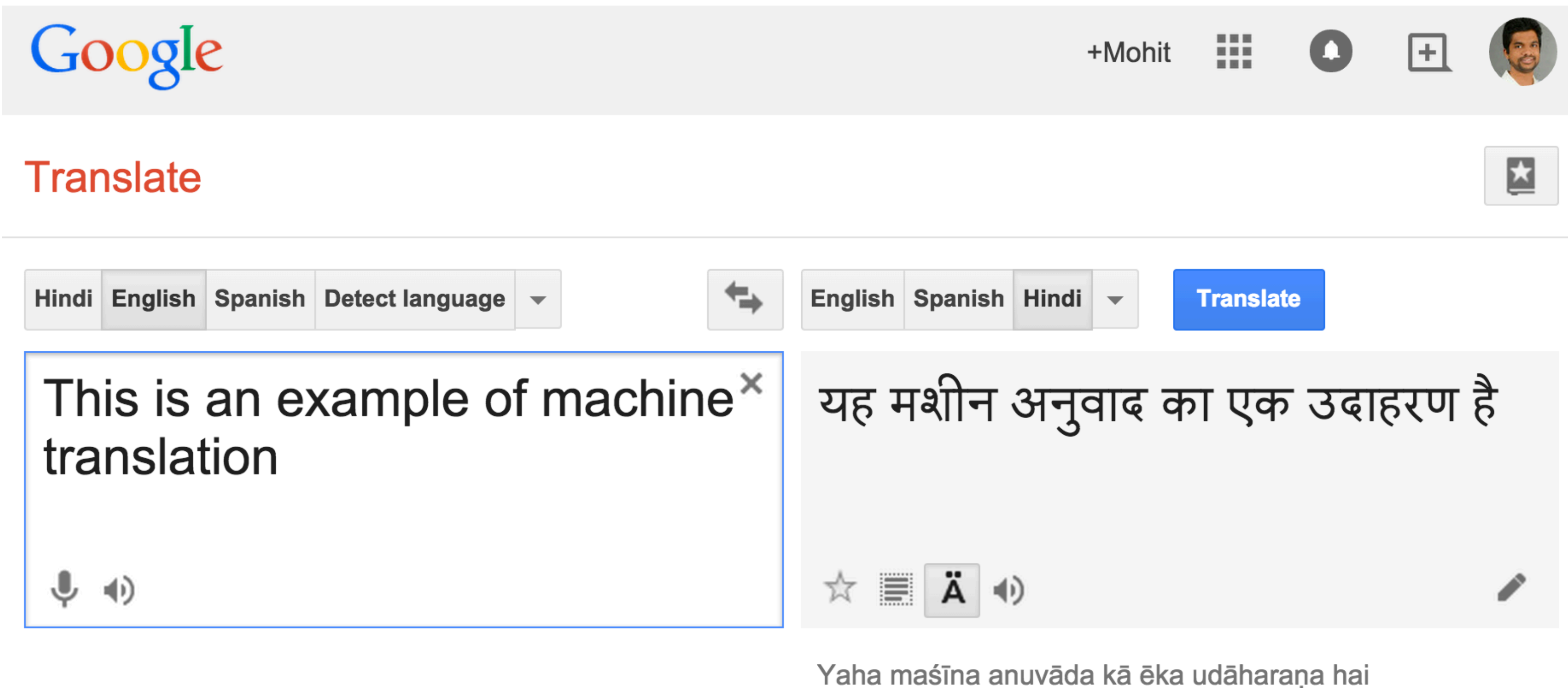
Machine Translation 2

Announcements

- Robotics+ML talk today from 11am-12pm by Dr. Animesh Garg (Stanford) on “Towards Generalizable Imitation in Robotics”!
- Come back to class at 12.05pm when TA Yixin will present HW2 (on entailment classification) and we will formally release the HW today/ tomorrow.

Machine Translation

- Useful for tons of companies, online traffic, and our international communication!



Statistical Machine Translation

- Source language f (e.g., French)
- Target language e (e.g., English)
- We want the best target (English) translation given the source (French) input sentence, hence the probabilistic formulation is:

$$\hat{e} = \mathrm{argmax}_e p(e|f)$$

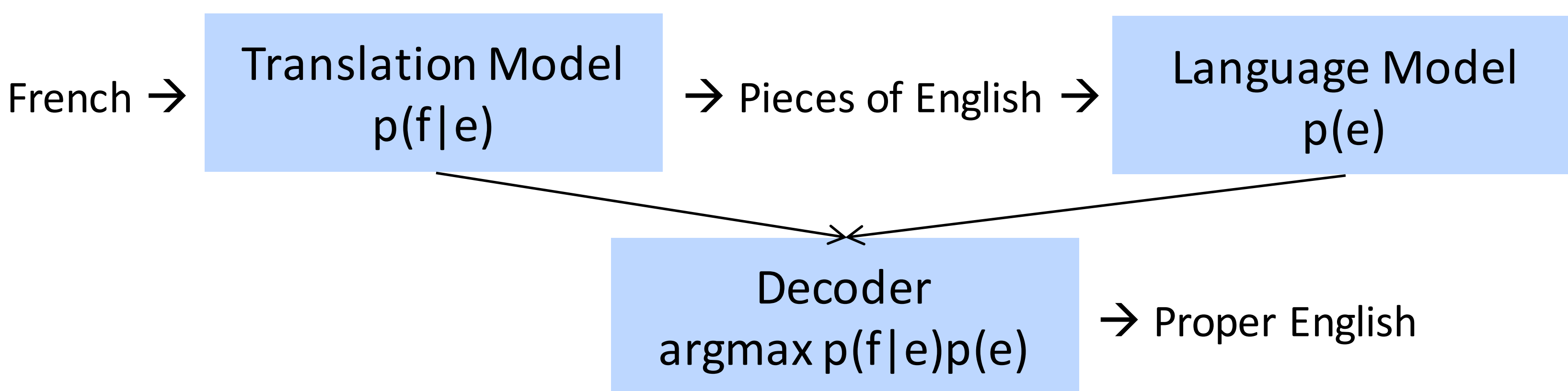
- Using Bayes rule, we get the following (since p(f) in the denominator is independent of the argmax over e):

$$\hat{e} = \mathrm{argmax}_e p(e|f) = \mathrm{argmax}_e p(f|e)p(e)$$

[Richard Socher CS224d]

Statistical Machine Translation

- The first part is known as the ‘Translation Model’ p(f|e) and is trained on parallel corpora of {f,e} sentence pairs, e.g., from EuroParl or Canadian parliament proceedings in multiple languages
- The second part p(e) is the ‘Language Model’ and can be trained on tons more monolingual data, which is much easier to find!

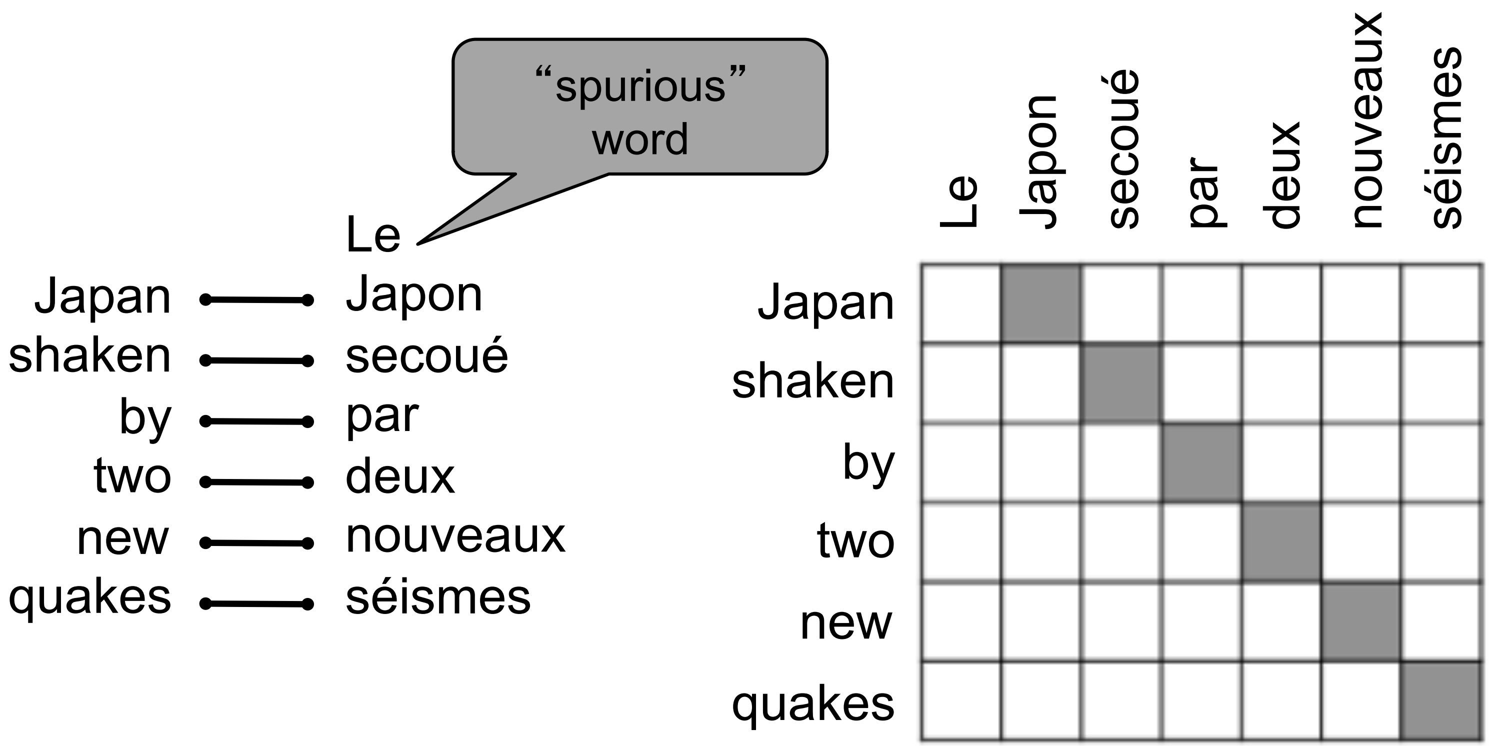


[Richard Socher CS224d]

Statistical Machine Translation

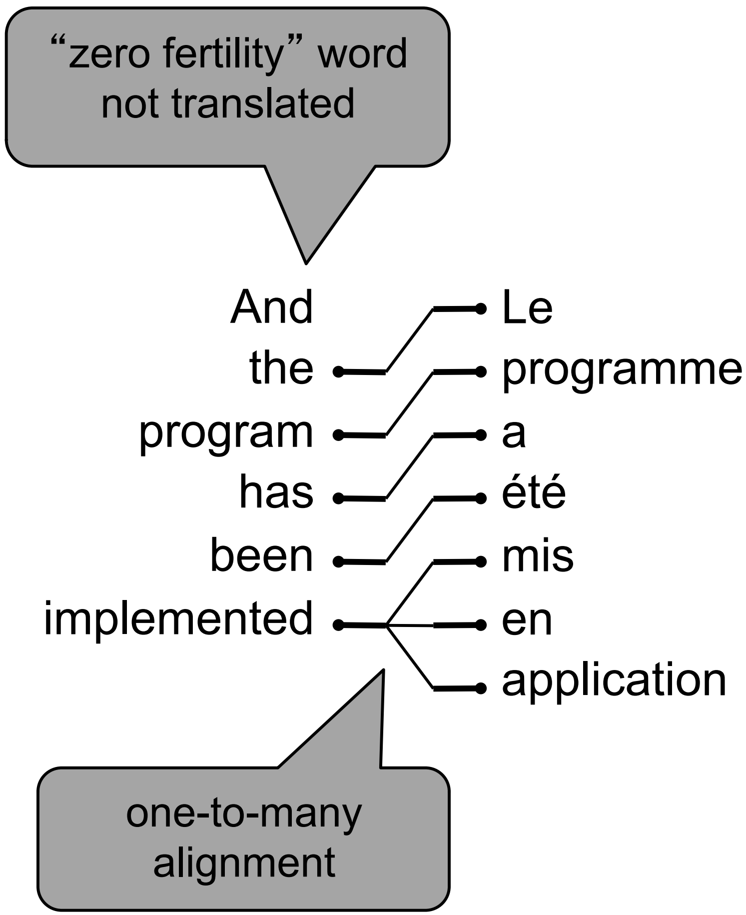
- First step in traditional machine translation is to find alignments or translational matchings between the two sentences, i.e., predict which words/phrases in French align to which words/phrases in English.

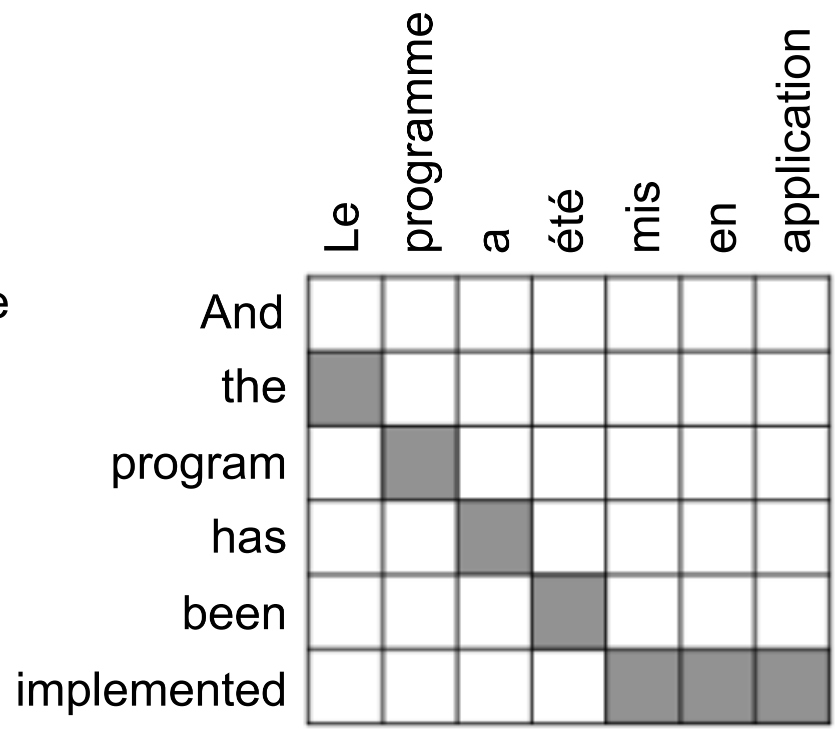
- Challenging problem: e.g., some words may not have any alignments:



Statistical Machine Translation

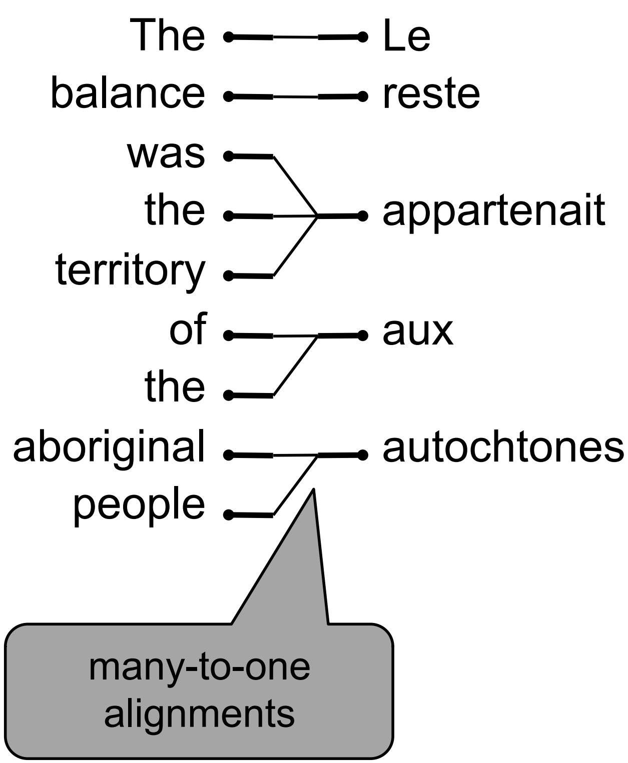
- One word in the source sentence might align to several words in the target sentence:

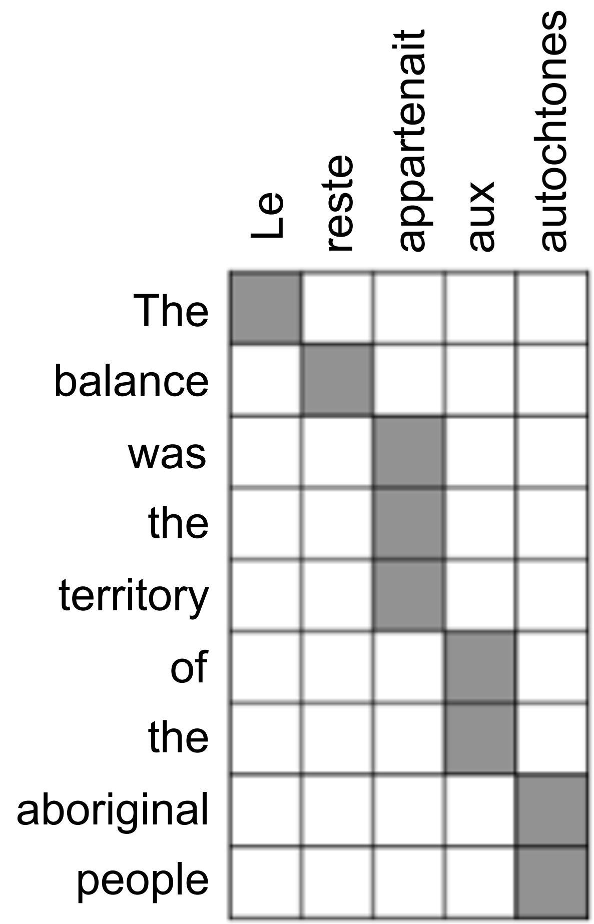




Statistical Machine Translation

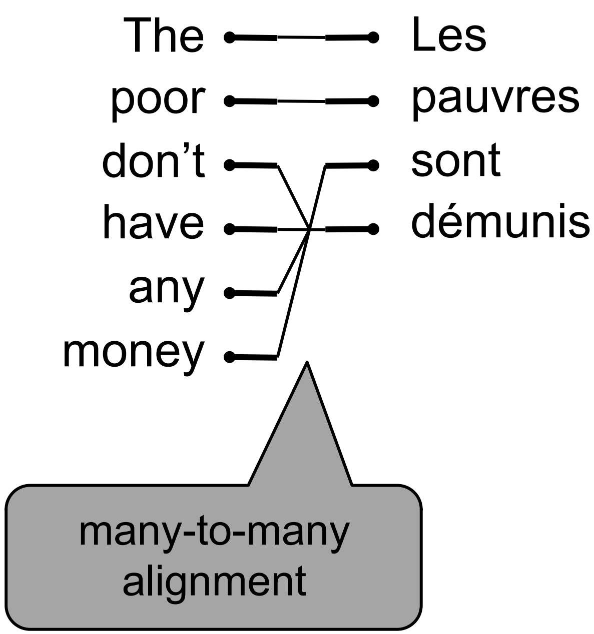
- Many words in the source sentence might align to a single word in the target sentence:

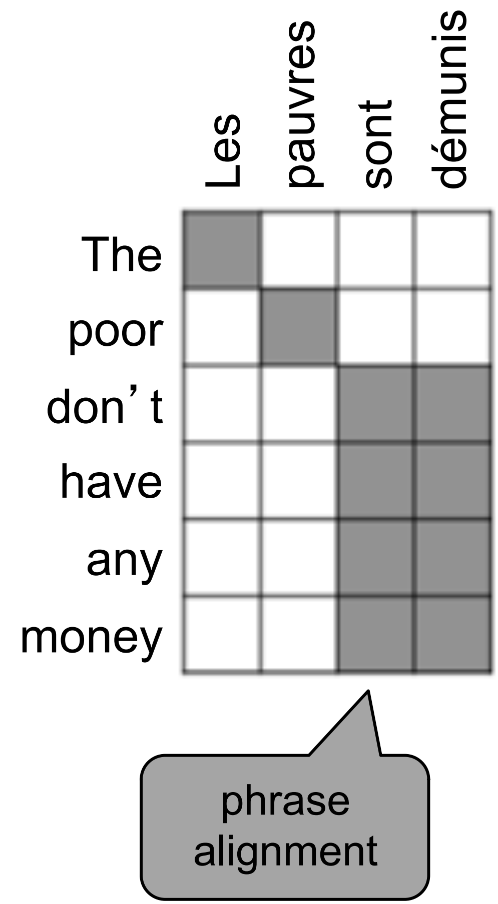




Statistical Machine Translation

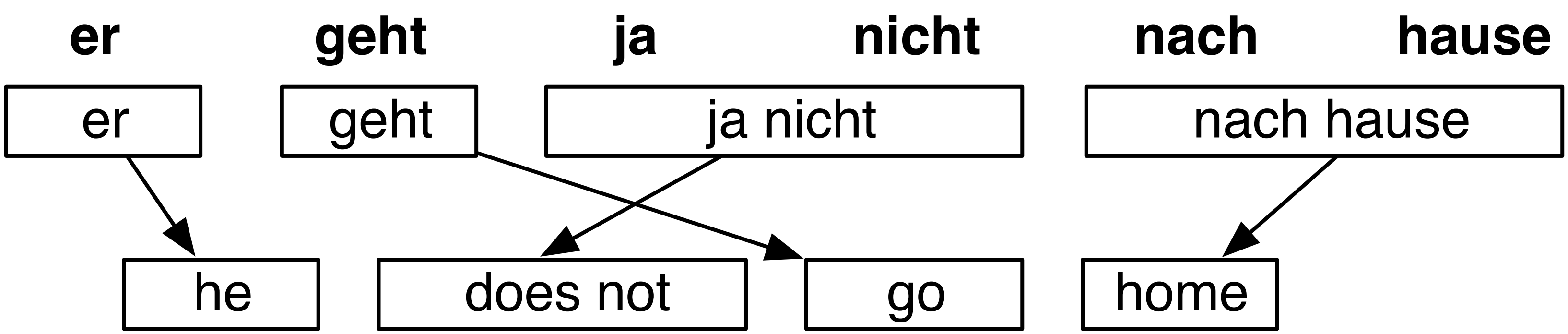
- And finally, many words in the source sentence might align to many words in the target sentence:





Statistical Machine Translation

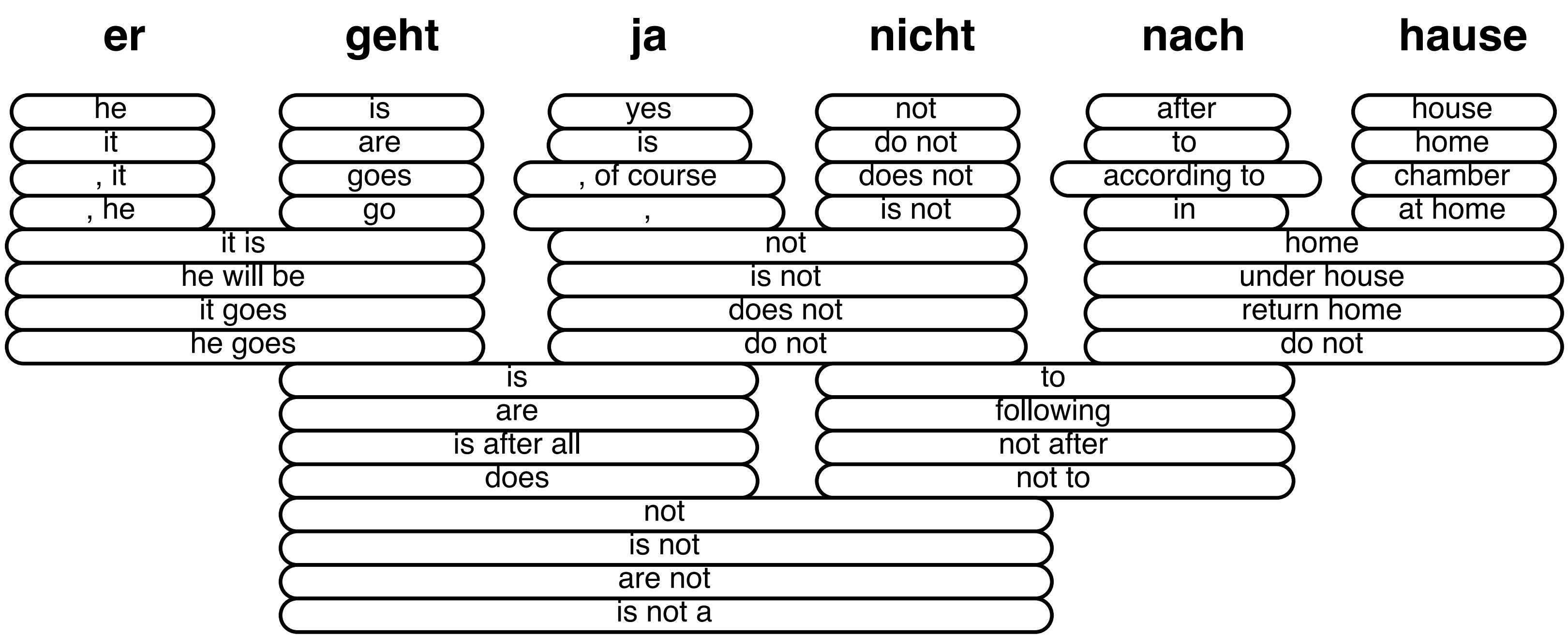
- After learning the word and phrase alignments, the model also needs to figure out the reordering, esp. important in language pairs with very different orders!



Statistical Machine Translation

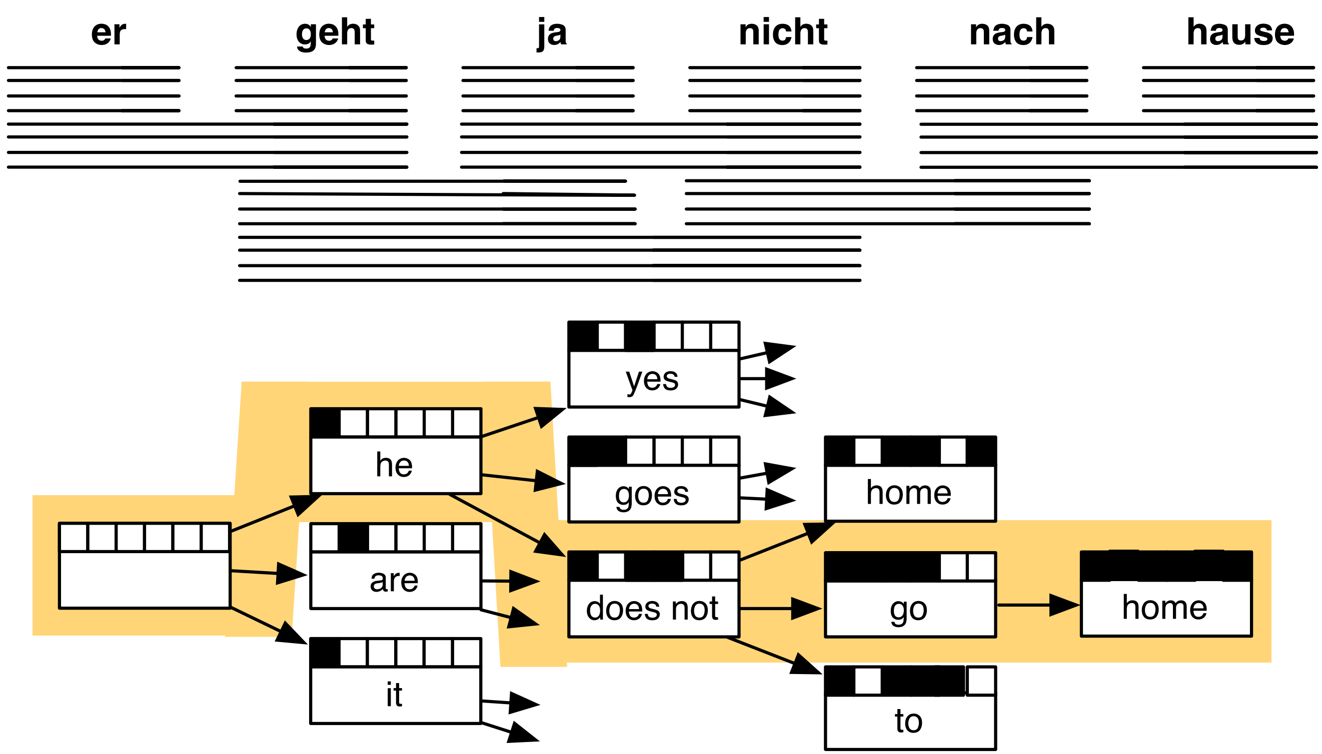
- After many steps, you get the large ‘phrase table’. Each phrase in the source language can have many possible translations in the target language, and hence the search space can be combinatorially large!

Translation Options

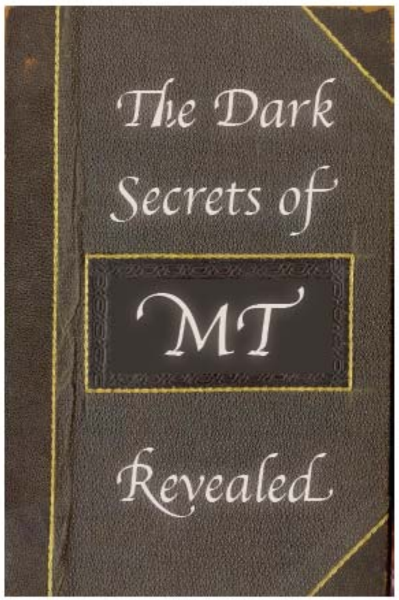


Statistical Machine Translation

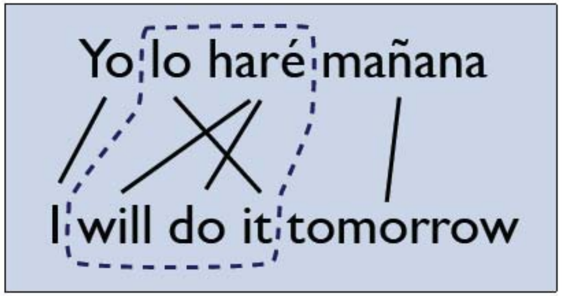
- Finally, you decode this hard search problem to find the best translation, e.g., using beam search on the several combinatorial paths through this phrase table (and also include the language model p(e) to rerank)



Alignment Model Details



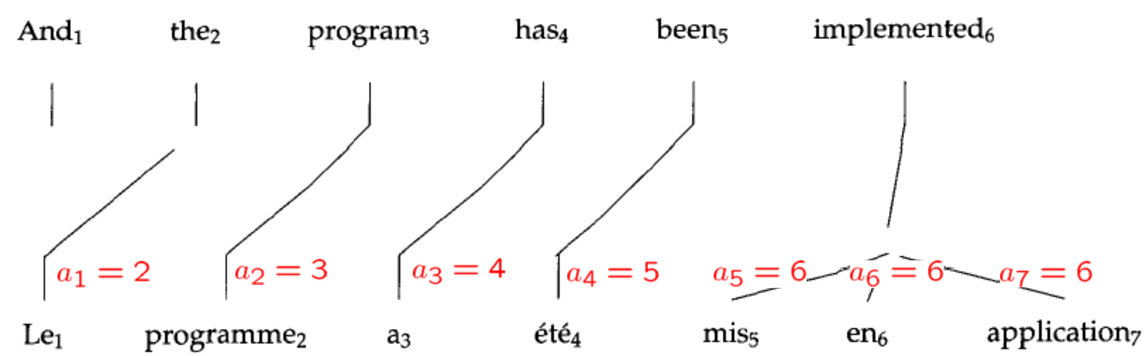
1. *Align words with a probabilistic model*
2. *Infer presence of larger structures from this alignment*
3. *Translate with the larger structures*



IBM Model 1

- Alignments: a hidden vector called an alignment specifies which English source is responsible for each French target word.
- The first, simplest IBM model treated alignment probabilities as roughly uniform:

$$a = a_1 \dots a_J$$



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

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

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

[Brown et al., 1993]

IBM Model 2 (Distortion)

- The next more advanced model captures the notion of 'distortion', i.e., how far from the diagonal is the alignment

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

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

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

- Other approaches for biasing alignment towards diagonal include relative vs absolute alignment, asymmetric distances, and learning a full multinomial over distances

[Brown et al., 1993]

IBM Models 1/2 EM Training

- Model Parameters:
  - Translational Probabilities: $P(f_j|e_i)$
  - Distortion Probabilities: $P(a_j = i|j, I, J)$

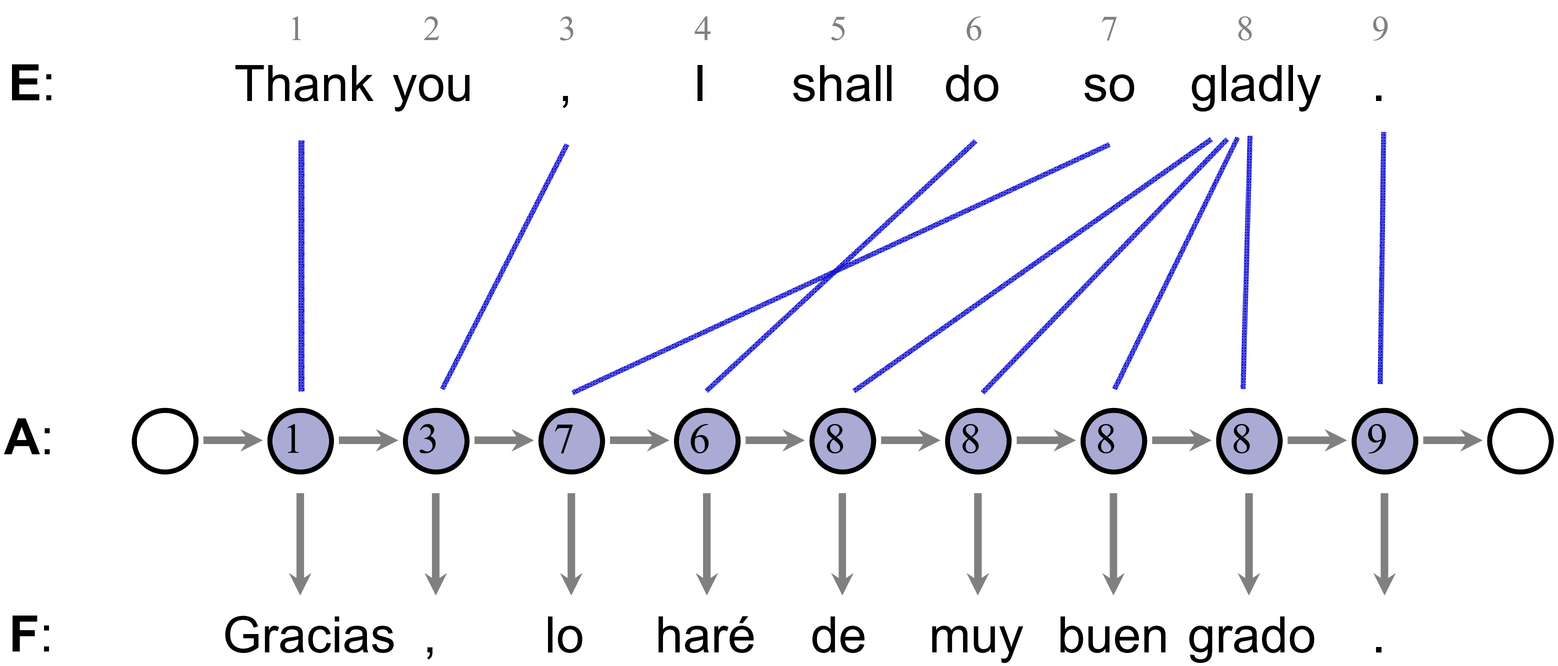
- Start with uniform P(f_j | e_i) parameters, including P(f_j | null)
- For each sentence in training corpus:
  - For each French position *j*:
    - Calculate posterior over English positions using:

$$P(a_j = i|f, e) = \frac{P(a_j = i|j, I, J)P(f_j|e_i)}{\sum_{i'} P(a_j = i'|j, I, J)P(f_j|e'_i)}$$

    - Increment count of word f_j with word e_i by these amounts
    - Similarly re-estimate distortion probabilities for Model2
- Iterate until convergence

[Brown et al., 1993]

HMM Model

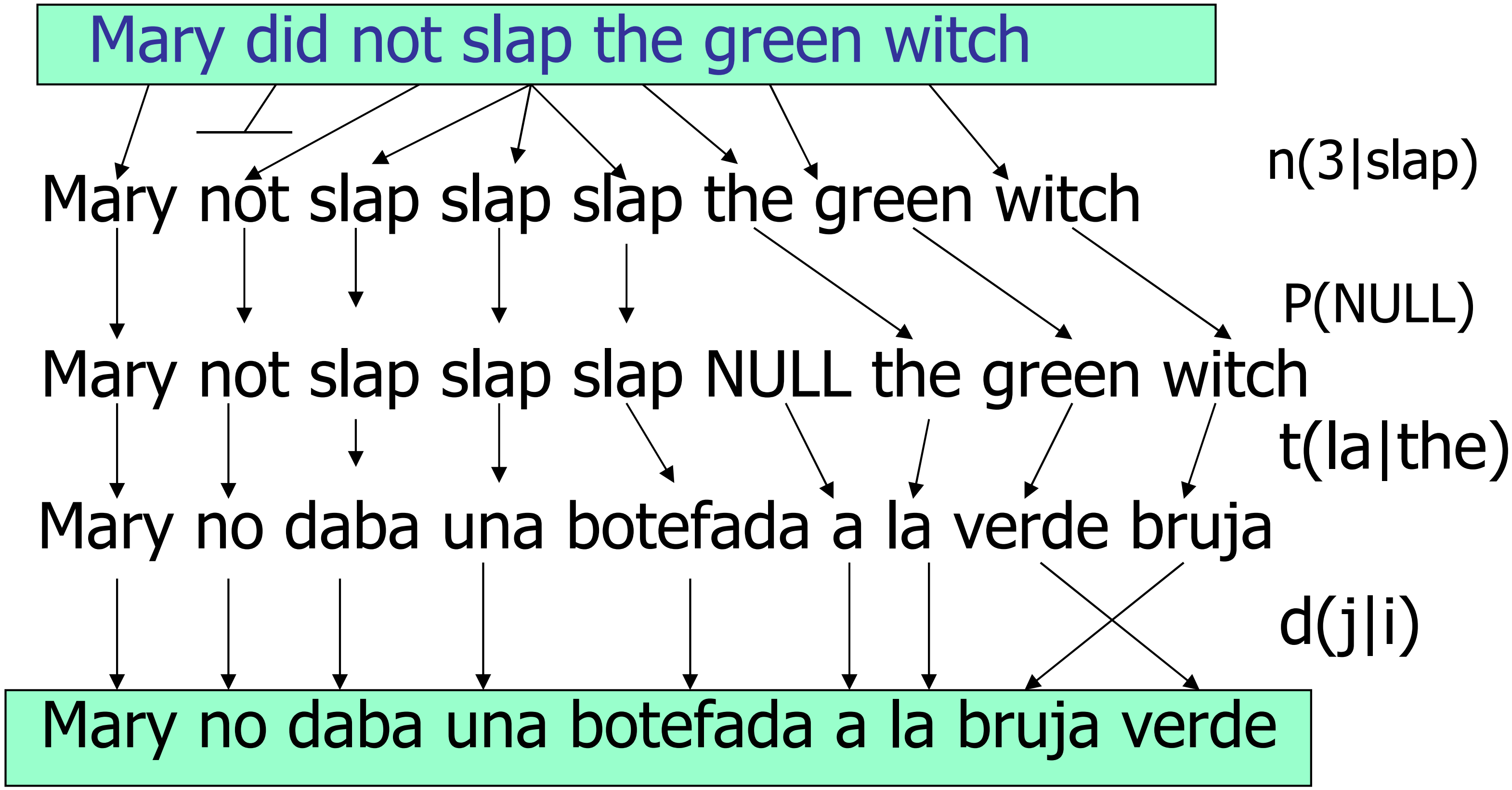


Model Parameters

Emissions: P(F_1 = Gracias | E_{A_1} = Thank) *Transitions*: P(A_2 = 3 | A_1 = 1)

[Vogel et al., 1996]

IBM Models 3/4/5 (Fertility)



[Vogel et al., 1996]

IBM Models 3/4/5 (Fertility)

the

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

not

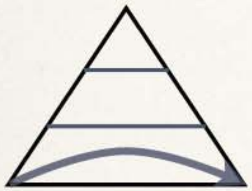
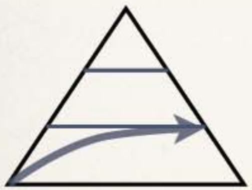
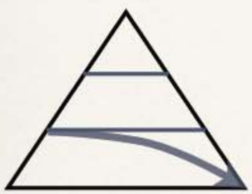
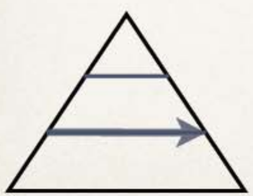
f	$t(f \mid e)$	ϕ	$n(\phi \mid e)$
ne	0.497	2	0.735
pas	0.442	0	0.154
non	0.029	1	0.107
rien	0.011		

farmers

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

[Vogel et al., 1996]

Syntactic Machine Translation

	string-to-string	ITG (Wu 1997)	Hiero (Chiang 2005)
	string-to-tree	Yamada & Knight 2001	Galley et al 2004/2006
	tree-to-string		Huang et al 2006 Y Liu et al 2006
	tree-to-tree	DOT (Poutsma 2000) Eisner 2003	Stat-XFER (Lavie et al 2008) M Zhang et al. 2008 Y Liu et al., 2009

Hiero

$$S \rightarrow \langle S_{\boxed{1}} X_{\boxed{2}}, S_{\boxed{1}} X_{\boxed{2}} \rangle$$

$$S \rightarrow \langle X_{\boxed{1}}, X_{\boxed{1}} \rangle$$

$$X \rightarrow \langle \text{yu } X_{\boxed{1}} \text{ you } X_{\boxed{2}}, \text{have } X_{\boxed{2}} \text{ with } X_{\boxed{1}} \rangle$$

$$X \rightarrow \langle X_{\boxed{1}} \text{ de } X_{\boxed{2}}, \text{the } X_{\boxed{2}} \text{ that } X_{\boxed{1}} \rangle$$

$$X \rightarrow \langle X_{\boxed{1}} \text{ zhiyi, one of } X_{\boxed{1}} \rangle$$

$$X \rightarrow \langle \text{Aozhou, Australia} \rangle$$

$$X \rightarrow \langle \text{shi, is} \rangle$$

$$X \rightarrow \langle \text{shaoshu guojia, few countries} \rangle$$

$$X \rightarrow \langle \text{bangjiao, diplomatic relations} \rangle$$

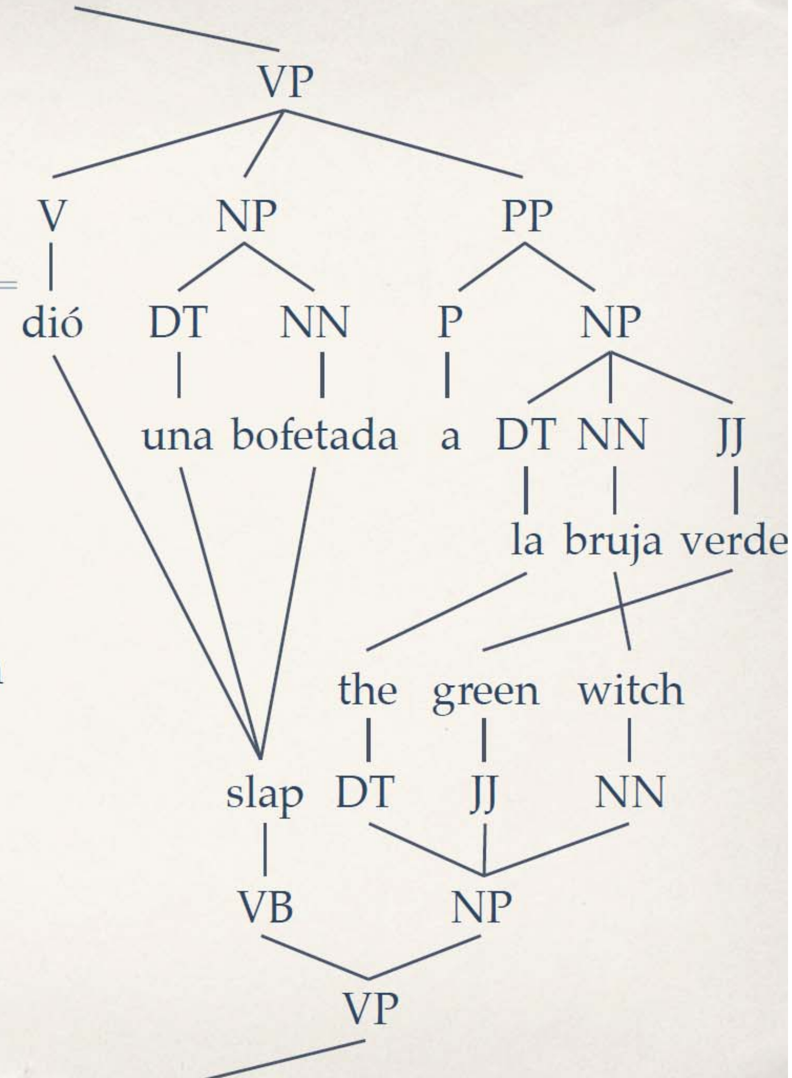
$$X \rightarrow \langle \text{Bei Han, North Korea} \rangle$$

From [Chiang et al, 2005]

Synchronous Tree-Substitution Grammars

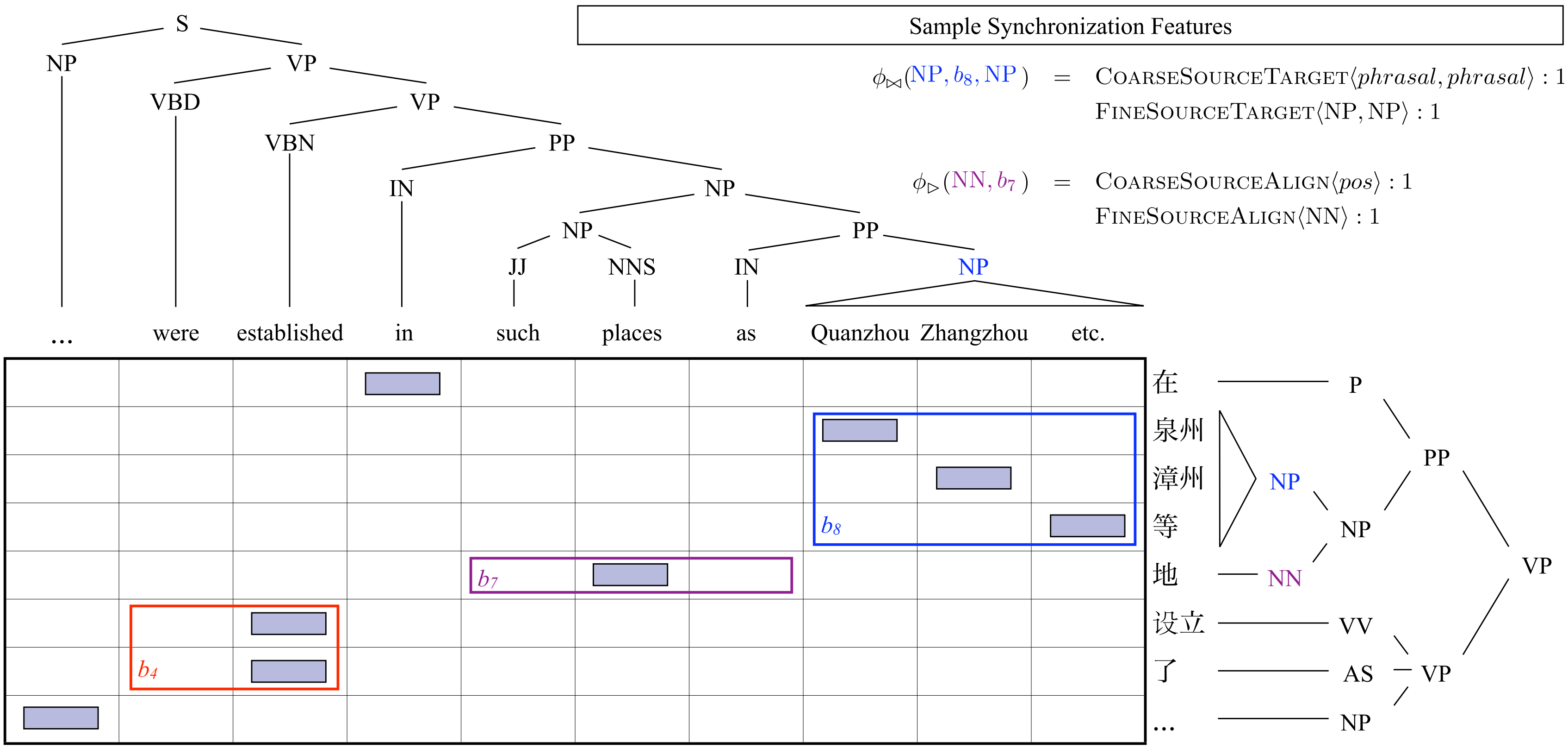
STSG extraction

1. Phrases
   - respect word alignments
   - are syntactic constituents on *both* sides
2. Phrase pairs form rules
3. Subtract phrases to form rules



[Shieber, 2004; Graehl et al., 2008]

Joint Parsing and Alignment



[Burkett et al., 2011]

Guest Task by Dr. Animesh Garg (Stanford):

“Towards Generalizable Imitation in Robotics”

(11am-12pm)

Coding-HW2 Presentation by TA Yixin Nie:

“Sequence-to-Label Learning for Entailment Recognition”

(12.10pm-12.40pm)