Machine Translation

Dan Klein
UC Berkeley

Many slides from John DeNero and Philip Koehn

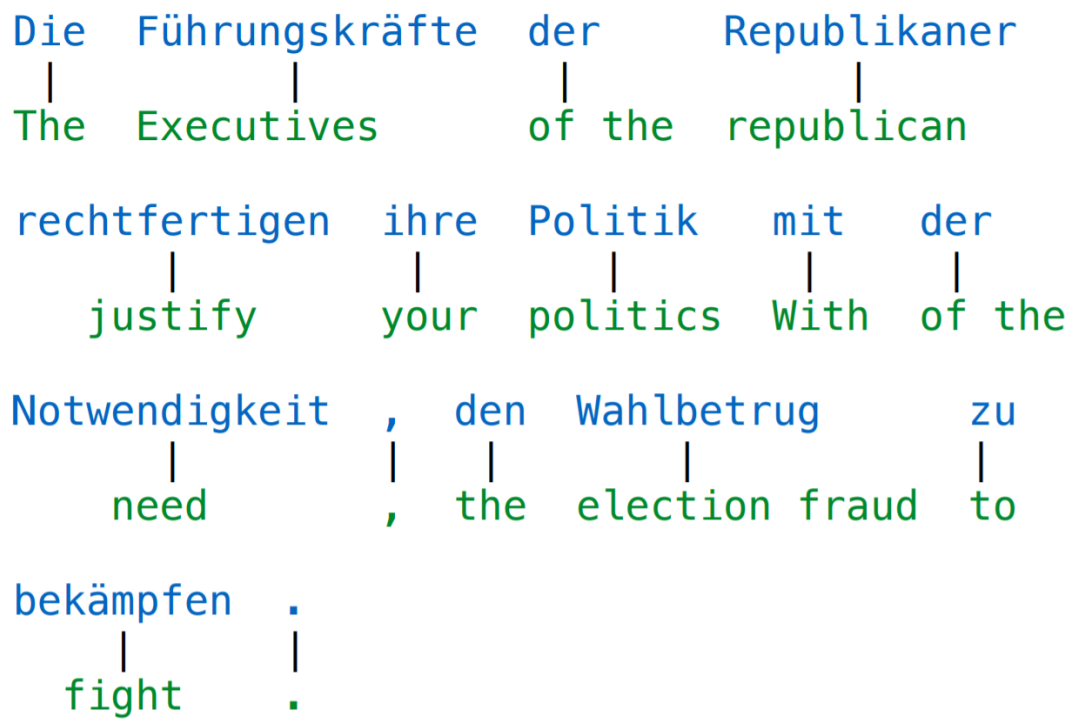
Translation Task

- Text is both the input and the output.
- Input and output have roughly the same information content.
- Output is more predictable than a language modeling task.
- Lots of naturally occurring examples.

Translation Examples

English-German News Test 2013 (a standard dev set)

Republican leaders justified their policy by the need to combat electoral fraud.



Variety in Translations?

Human-generated reference translation

A small planet, whose is as big as could destroy a middle sized city, passed by the earth with a distance of 463 thousand kilometers. This was not found in advance. The astronomists got to know this incident 4 days later. This small planet is 50m in diameter. The astonomists are hard to find it for it comes from the direction of sun.

A commercial system from 2002

A volume enough to destroy a medium city small planet is big, flit earth within 463,000 kilometres of close however were not in advance discovered, astronomer just knew this matter after four days. This small planet diameter is about 50 metre, from the direction at sun, therefore astronomer very hard to discovers it.

Google Translate, 2020

From https://catalog.ldc.upenn.edu/LDC2003T17

An asteroid that was large enough to destroy a medium-

Evaluation

BLEU Score

BLEU score: geometric mean of 1-, 2-, 3-, and 4-gram precision vs. a reference, multiplied by brevity penalty (harshly penalizes translations shorter than the reference).

$$\text{Matched}_i = \sum_{t_i} \min \left\{ C_h(t_i), \max_j C_j(t_i) \right\}$$

If "of the" appears twice in hypothesis h but only at most once in a reference, then only the first is "correct"

$$P_i = \frac{\text{Matched}_i}{H_i}$$

"Clipped" precision of n-gram tokens

$$B = \exp \left\{ \min \left(0, \frac{n - L}{n} \right) \right\}$$

Brevity penalty only matters if the hypothesis **corpus** is shorter than the sum of (shortest) references.

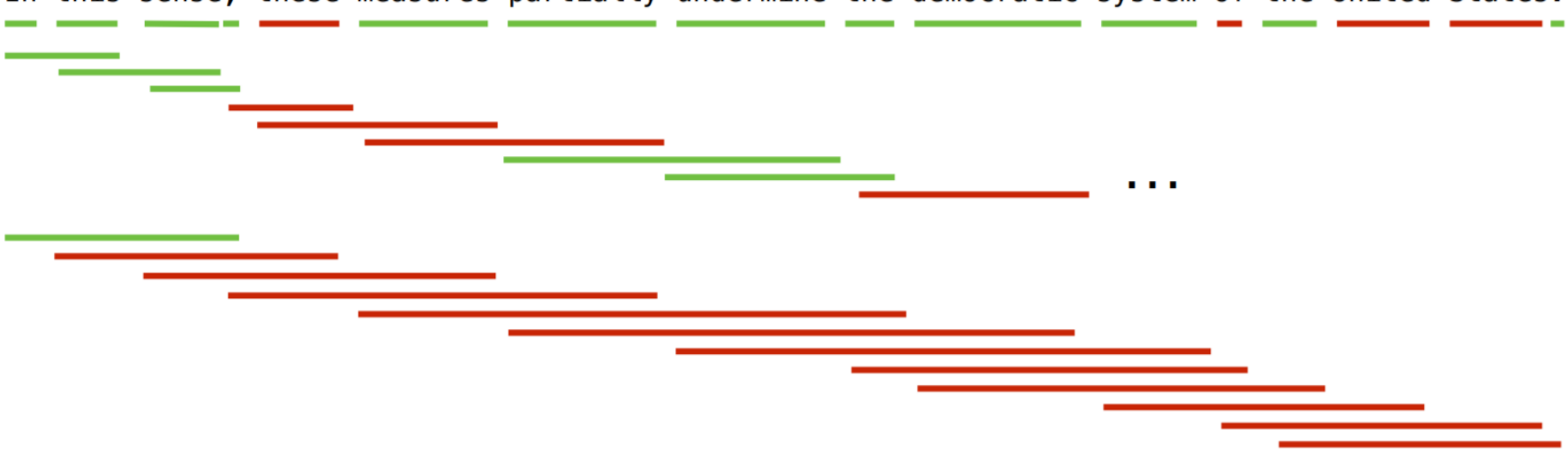
$$\text{BLUE} = B \left(\prod_{i=1}^{4} P_i \right)^{\frac{1}{4}}$$

BLEU is a mean of clipped precisions, scaled down by the brevity penalty.

Evaluation with BLEU

In this sense, the measures will partially undermine the American democratic system.

In this sense, these measures partially undermine the democratic system of the United States.



BLEU = 26.52, 75.0/40.0/21.4/7.7 (BP=1.000, ratio=1.143, hyp_len=16, ref_len=14)

(Papineni et al., 2002) BLEU: a method for automatic evaluation of machine translation.

Corpus BLEU Correlations with Average Human Judgments

These are ecological correlations over multiple segments; segment-level BLEU scores are noisy.

Commercial machine translation providers seem to all perform human evaluations of some sort.

(Ma et al., 2019) Results of the WMT19 Metrics Shared Task: Segment-Level and Strong MT Systems Pose Big Challenges

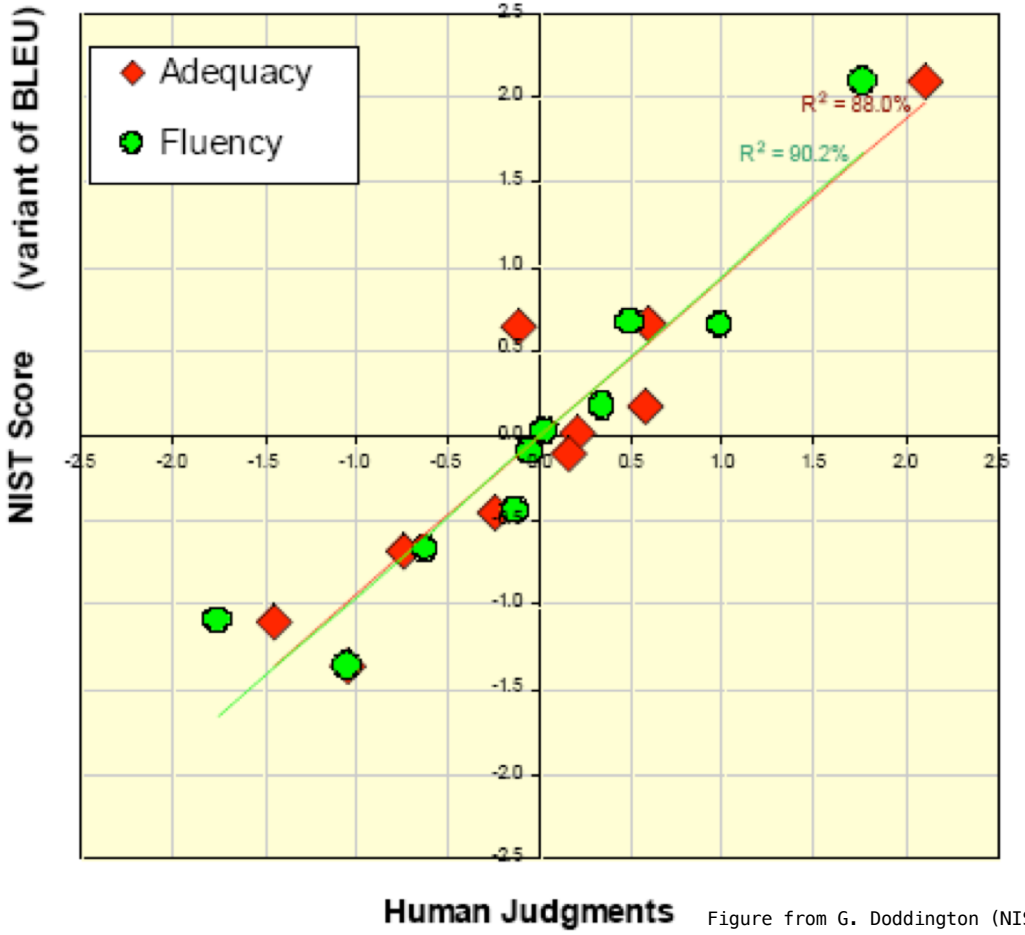


Figure from G. Doddington (NIST)

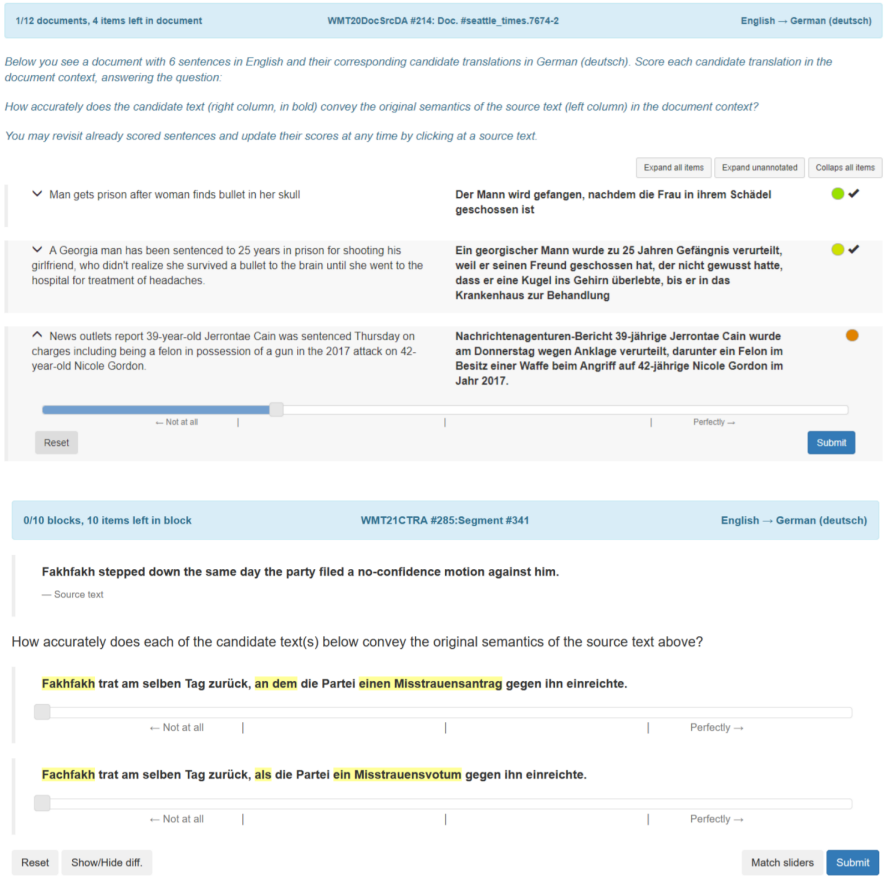
Human Evaluations

Direct assessment: adequacy & fluency

- Monolingual: Ask humans to compare machine translation to a human-generated reference. (Easier to source annotators)
- Bilingual: Ask humans to compare machine translation to the source sentence that was translated. (Compares to human quality)
- Annotators can assess segments (sentences) or whole documents.
- Segments can be assessed with or without document context.

Ranking assessment:

- Raters are presented with 2 or more translations.
- A human-generated reference may be provided, along with the source.
- "In a pairwise ranking experiment, human raters assessing adequacy and fluency show a



(Laubli et al., 2018) Has Machine Translation Achieved Human Parity? A Case for Document-level Evaluation

(Akhbardeh et al., 2021) Findings of the 2021 Conference on Machine Translation

Translationese and Evaluation

Translated text can: (Baker et al., 1993; Graham et al., 2019)

- be more explicit than the original source
- be less ambiguous
- be simplified (lexically, syntactically, and stylistically)
- display a preference for conventional grammaticality
- avoid repetition
- exaggerate target language features
- display features of the source language

"If we consider only original source text (i.e. not translated from another language, or translationese), then we find evidence showing that human parity has not been achieved."
(Toral et al., 2018)

(Baker et al., 1993) Corpus linguistics and transla- tion studies: Implications and applications.
(Graham et al., 2019) Translationese in Machine Translation Evaluation.
(Toral et al, 2018) Attaining the Unattainable? Reassessing Claims of Human Parity in Neural Machine Translation

How are We Doing? Example: WMT 2019 Evaluation

2019 segment-in-context direct assessment (Barrault et al, 2019):

✓ German to English: many systems are tied with human performance;

× English to Chinese: all systems are outperformed by the human translator;

× English to Czech: all systems are outperformed by the human translator;

× English to Finnish: all systems are outperformed by the human translator;

✓ English to German: Facebook-FAIR achieves super-human translation performance; several systems are tied with human performance;

× English to Gujarati: all systems are outperformed by the human translator;

× English to Kazakh: all systems are outperformed by the human translator;

× English to Lithuanian: all systems are outperformed by the human translator;

✓ English to Russian: Facebook-FAIR is tied with human performance.

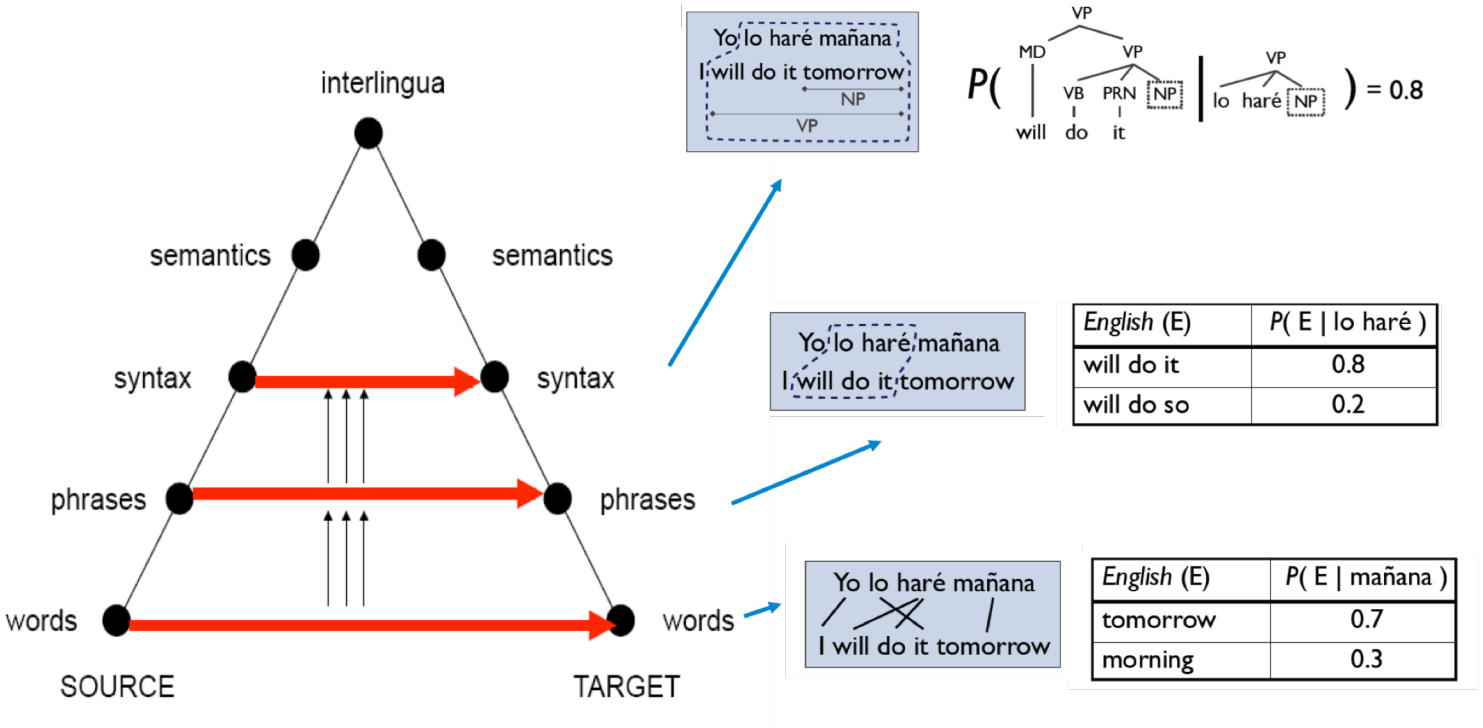
(Barrault et al, 2019) Findings of the 2019 Conference on Machine Translation (WMT19)

Statistical Machine Translation
(1990 - 2015)

When I look at an article in Russian, I say: “This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.”

Warren Weaver (1949)

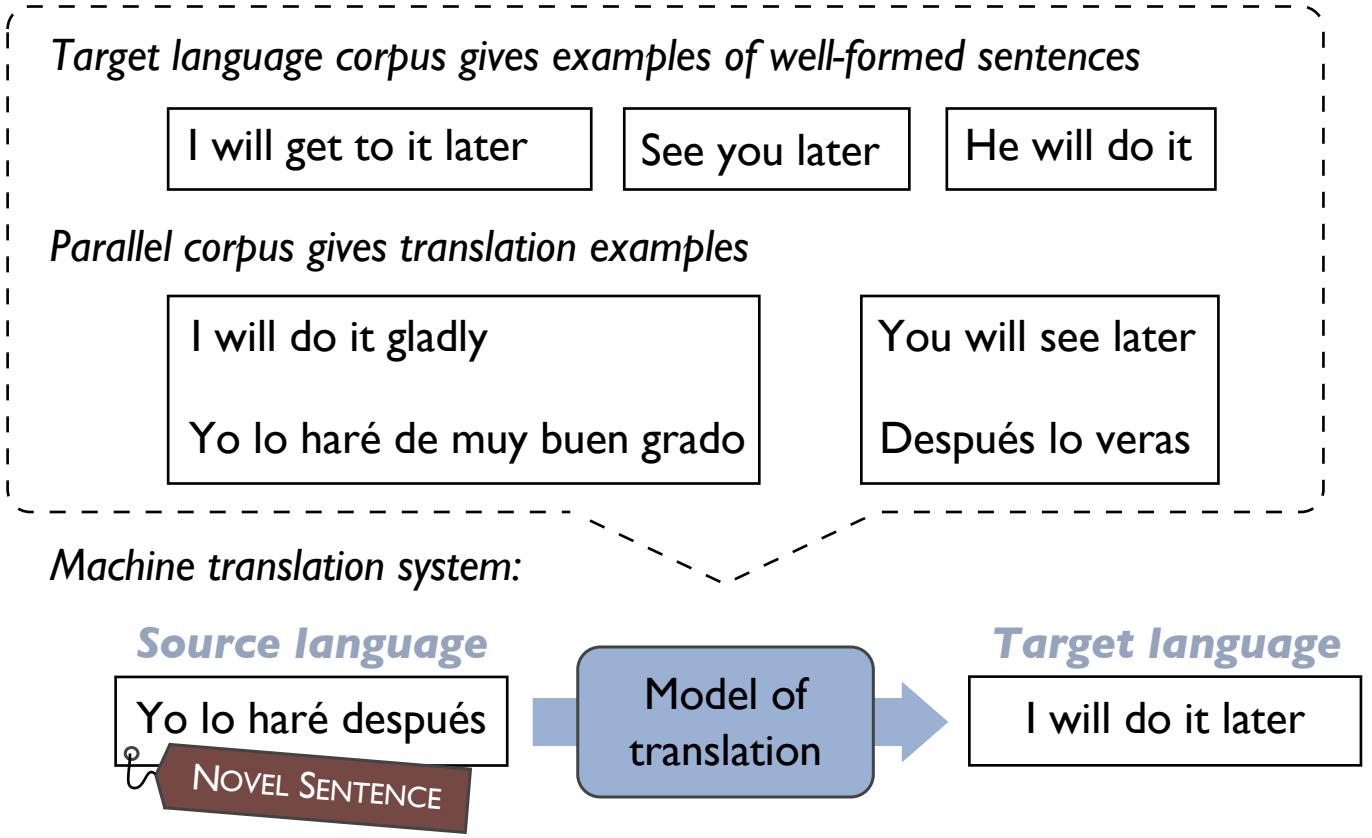
Levels of Transfer: Vauquois Triangle (1968)



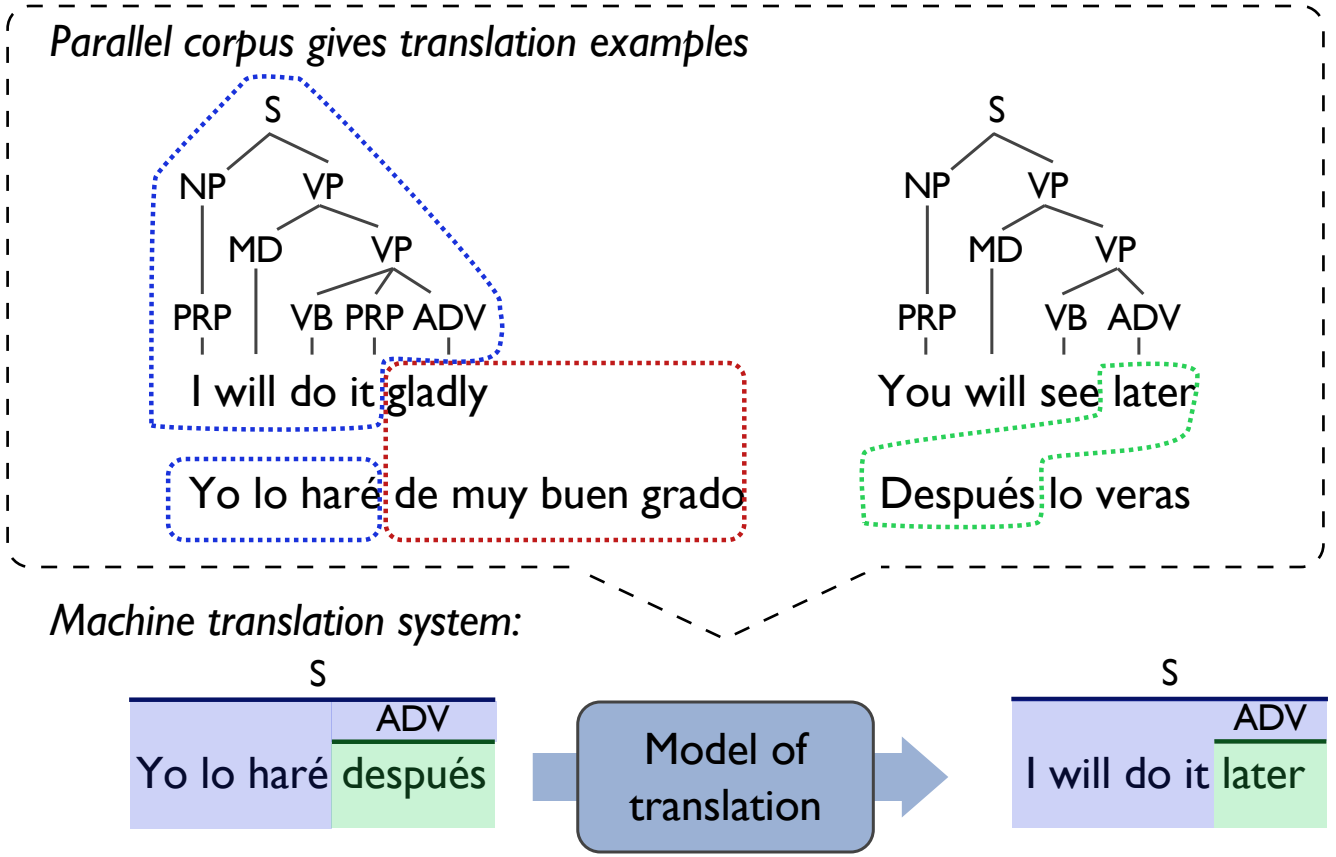
English (E)	*P*(E | lo haré)
will do it	0.8
will do so	0.2

English (E)	*P*(E | mañana)
tomorrow	0.7
morning	0.3

Data-Driven Machine Translation



Stitching Together Fragments



Evolution of the Noisy Channel Model

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

$$P(e|f) \propto P(f|e)^{\phi_{\text{tm}}} \cdot P(e)^{\phi_{\text{lm}}}$$

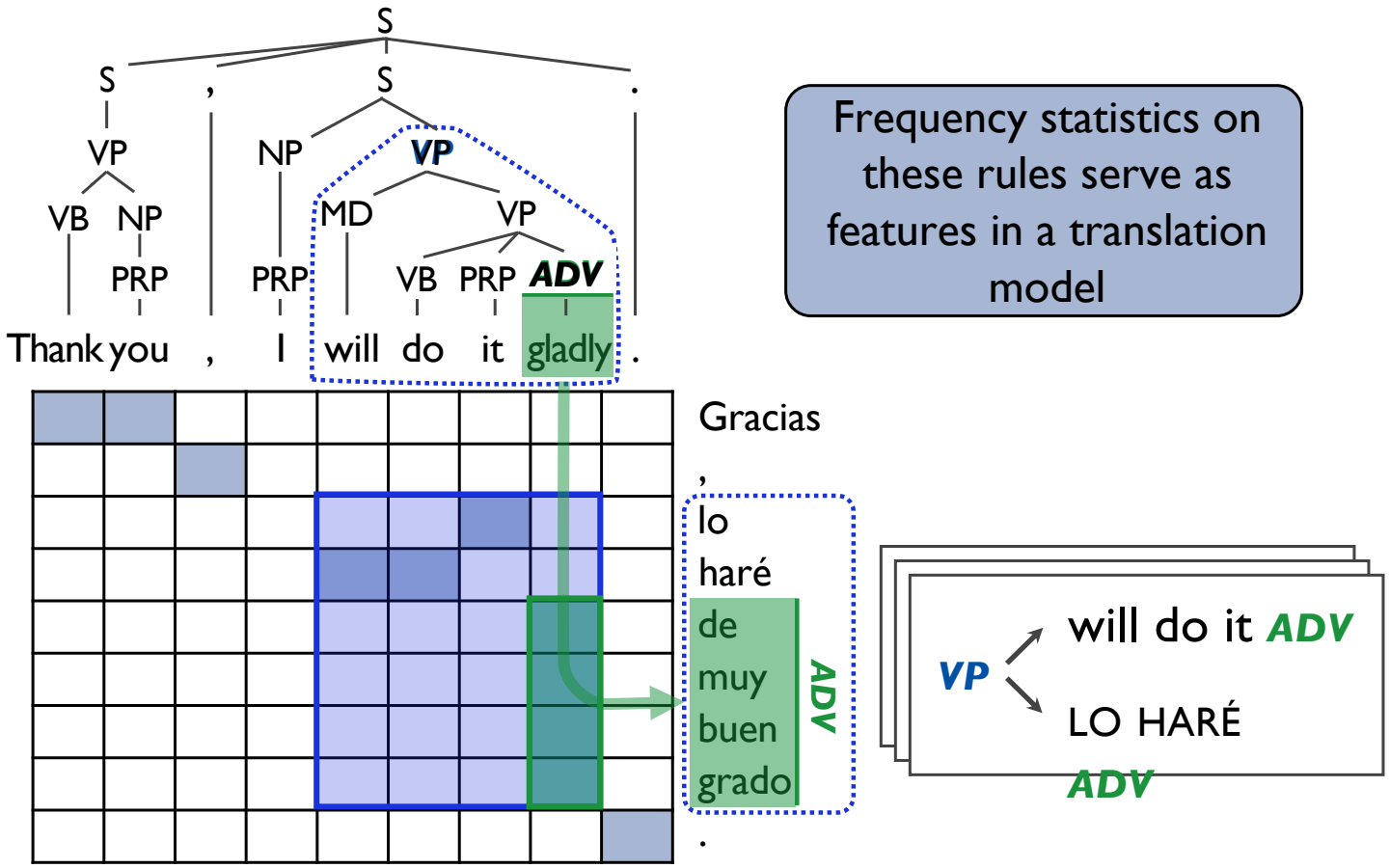
$$P(e|f) \propto \exp\left\{\sum_i w_i \cdot f_i(e, f)\right\}$$

Chosen to minimize loss

E.g., log P(e)

Word Alignment and Phrase Extraction

Extracting Translation Rules



Frequency statistics on these rules serve as features in a translation model

Counting Aligned Phrases

d’assister à la reunion et ||| to attend the meeting and

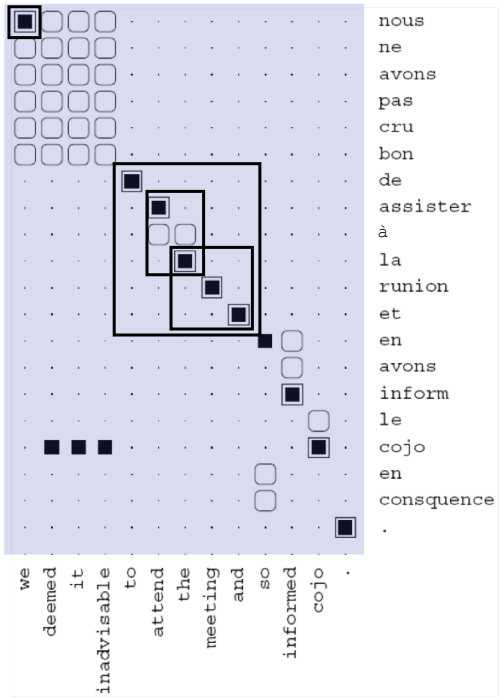
assister à la reunion ||| attend the meeting

la reunion et ||| the meeting and

nous ||| we

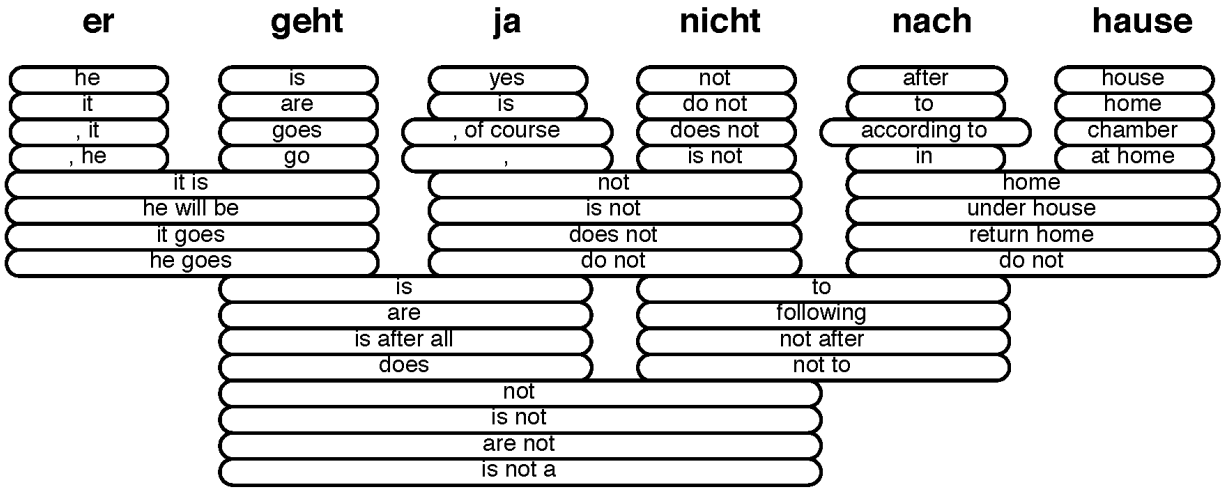
…

- Relative frequencies are the most important features in a phrase-based or syntax-based model.
- Scoring a phrase under a lexical model is the second most important feature.
- Estimation does not involve choosing among segmentations of a sentence into phrases.



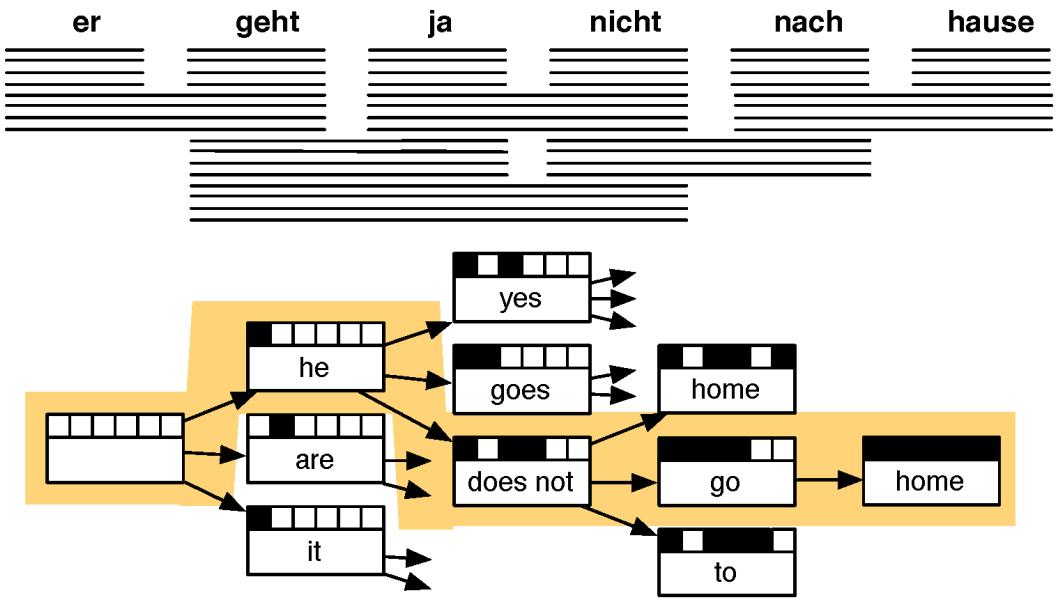
Slide by Greg Durrett

Translation Options



- Many translation options to choose from
  - in Europarl phrase table: 2727 matching phrase pairs for this sentence
  - by pruning to the top 20 per phrase, 202 translation options remain

Decoding: Find Best Path



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**	the astronauts		,
it	7 people included		by france		and the	the russian		international astronautical	of rapporteur .	
this	7 out	including the	**from**	the french	and the russian		the fifth		.	
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 .	
	7 include		from the	of france and		russian		**astronauts**		. the
	7 numbers include		**from france**		and russian		of astronauts who			. ”
	7 populations include		those from france		and russian			astronauts .		
	7 deportees included		come from	**france**	**and russia**		in	astronautical	personnel	;
	7 philtrum	including those from		**france and**		**russia**	a space		**member**	
		including representatives from		france and the		**russia**		astronaut		
		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				