

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.

Die	Führungskräfte	der	Republikaner
The	Executives	of the	republican
rechtfertigen	ihre	Politik	mit der
justify	your	politics	With of the
Notwendigkeit	,	den	Wahlbetrug zu
need	,	the	election fraud to
bekämpfen	.		
fight	.		

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 astronomists 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.

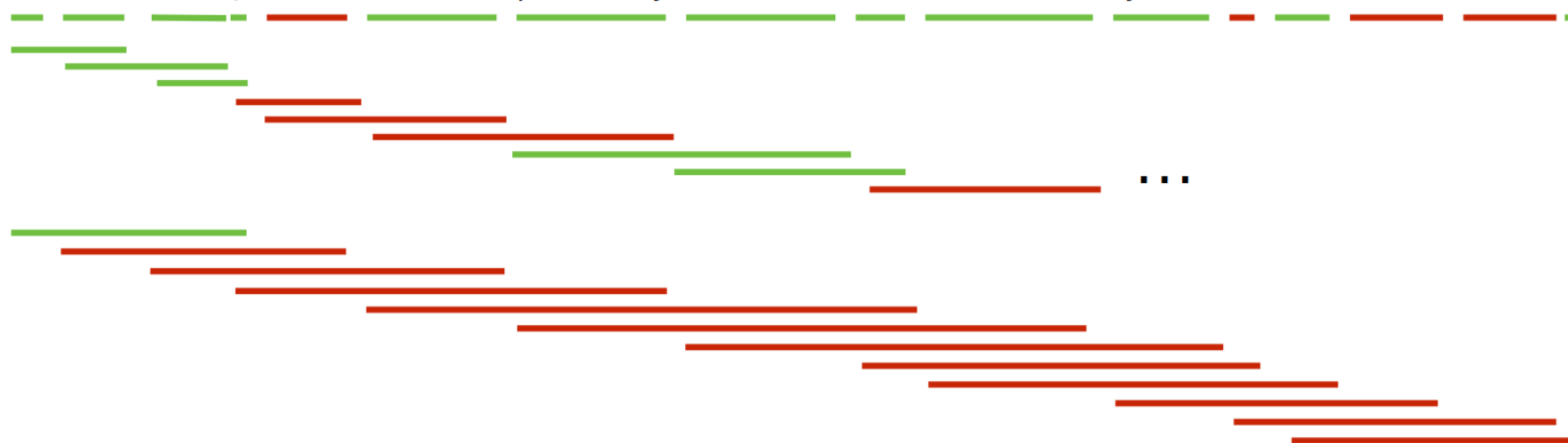
$$\text{BLUE} = B \left(\prod_{i=1}^4 P_i \right)^{\frac{1}{4}}$$

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

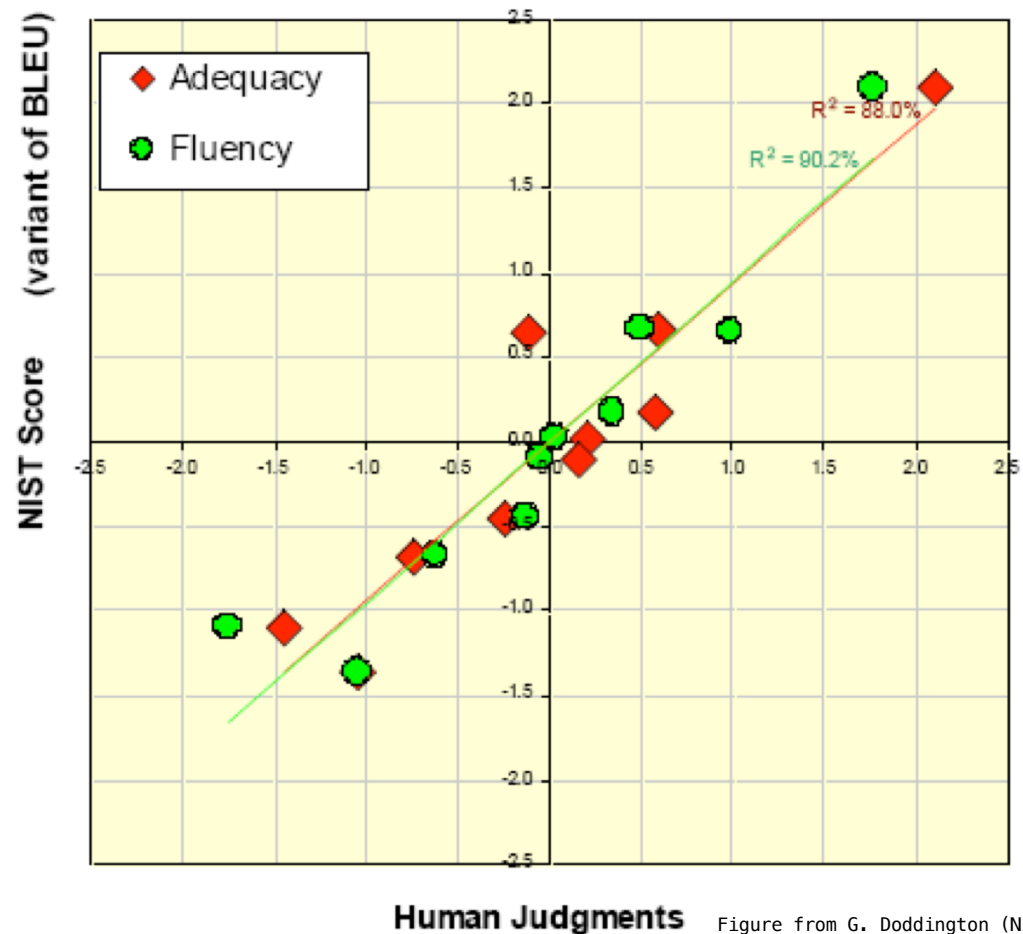


Figure from G. Doddington (NIST)

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

1/12 documents, 4 items left in document WMT20DocSrcDA #214: Doc. #seattle_times.7674-2 English → German (deutsch)

Below you see a document with 6 sentences in English and their corresponding candidate translations in German (deutsch). Score each candidate translation in the document context, answering the question:

How accurately does the candidate text (right column, in bold) convey the original semantics of the source text (left column) in the document context?

You may revisit already scored sentences and update their scores at any time by clicking at a source text.

Expand all items Expand unannotated Collapse all items

Man gets prison after woman finds bullet in her skull	Der Mann wird gefangen, nachdem die Frau in ihrem Schädel geschossen ist	100% ✓
A Georgia man has been sentenced to 25 years in prison for shooting his girlfriend, who didn't realize she survived a bullet to the brain until she went to the hospital for treatment of headaches.	Ein georgischer Mann wurde zu 25 Jahren Gefängnis verurteilt, weil er seinen Freund geschossen hat, der nicht gewusst hatte, dass er eine Kugel ins Gehirn überlebte, bis er in das Krankenhaus zur Behandlung	100% ✓
News outlets report 39-year-old Jerrontae Cain was sentenced Thursday on charges including being a felon in possession of a gun in the 2017 attack on 42-year-old Nicole Gordon.	Nachrichtenagenturen-Bericht 39-jährige Jerrontae Cain wurde am Donnerstag wegen Anklage verurteilt, darunter ein Felon im Besitz einer Waffe beim Angriff auf 42-jährige Nicole Gordon im Jahr 2017.	0% ✗

Reset Submit

0/10 blocks, 10 items left in block WMT21CTRA #285:Segment #341 English → German (deutsch)

Fakhfakh stepped down the same day the party filed a no-confidence motion against him.

— Source text

How accurately does each of the candidate text(s) below convey the original semantics of the source text above?

Fakhfakh trat am selben Tag zurück, an dem die Partei einen Misstrauensantrag gegen ihn einreichte.

Not at all Perfectly

Fakhfakh trat am selben Tag zurück, als die Partei ein Misstrauensvotum gegen ihn einreichte.

Not at all Perfectly

Reset Show/Hide diff. Match sliders Submit

(Akbari 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 translation 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 Gujarati: all systems are outperformed by the human translator; |
| × English to Chinese: all systems are outperformed by the human translator; | × English to Kazakh: all systems are outperformed by the human translator; |
| × English to Czech: all systems are outperformed by the human translator; | × English to Lithuanian: all systems are outperformed by the human translator; |
| × English to Finnish: all systems are outperformed by the human translator; | ✓ English to Russian: Facebook-FAIR is tied with human performance. |
| ✓ English to German: Facebook-FAIR achieves super-human translation performance; several systems are tied with human performance; | |

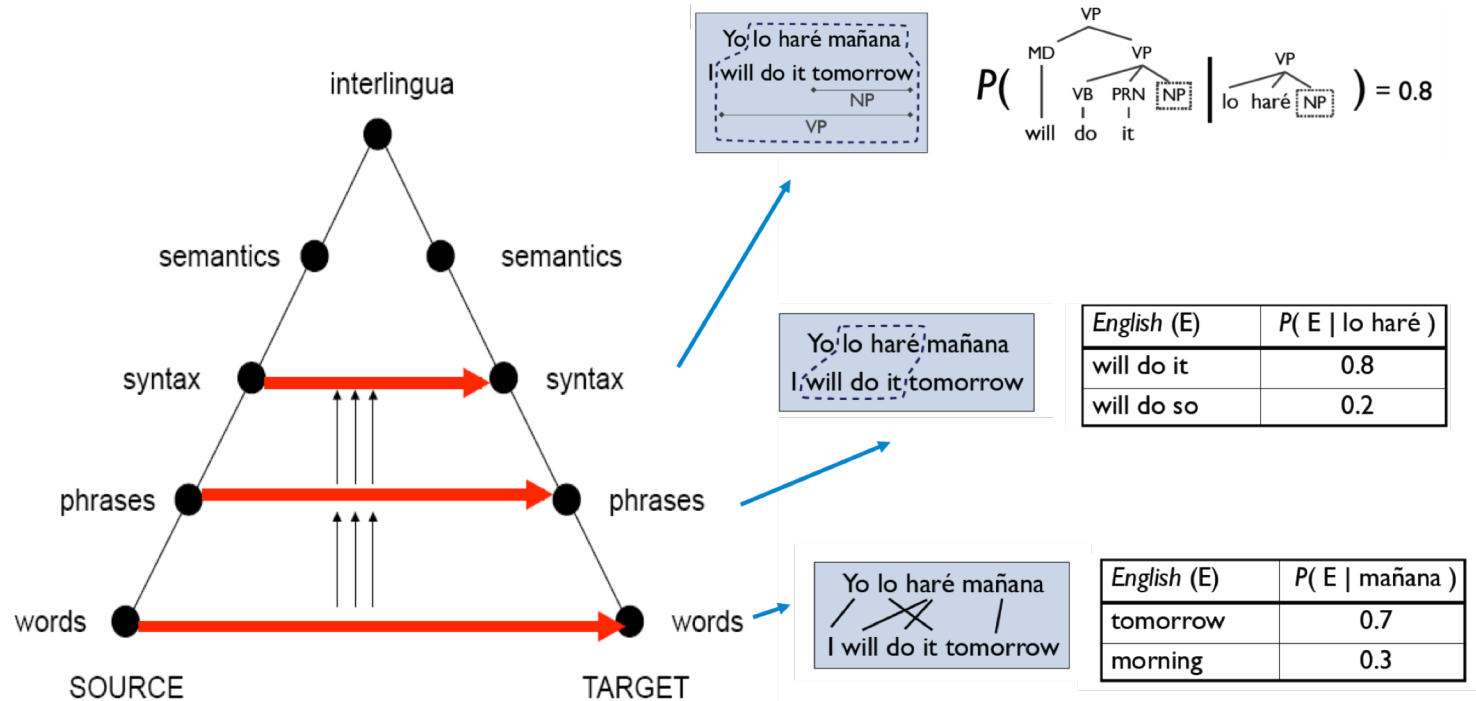
Statistical Machine Translation (1990 - 2015)



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)



Data-Driven Machine Translation

Target language corpus gives examples of well-formed sentences

I will get to it later

See you later

He will do it

Parallel corpus gives translation examples

I will do it gladly

Yo lo haré de muy buen grado

You will see later

Después lo veras

Machine translation system:

Source language

Yo lo haré después

NOVEL SENTENCE

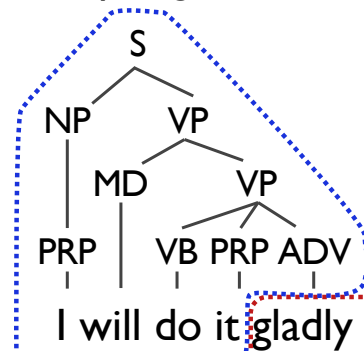
Model of
translation

Target language

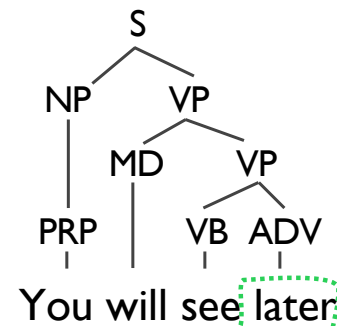
I will do it later

Stitching Together Fragments

Parallel corpus gives translation examples

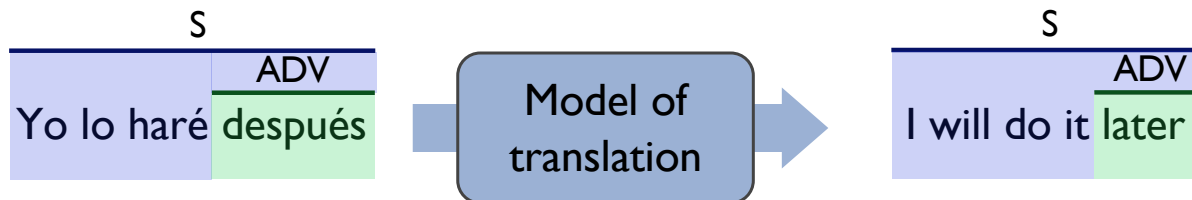


Yo lo haré de muy buen grado



Después lo veras

Machine translation system:



Evolution of the Noisy Channel Model

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

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

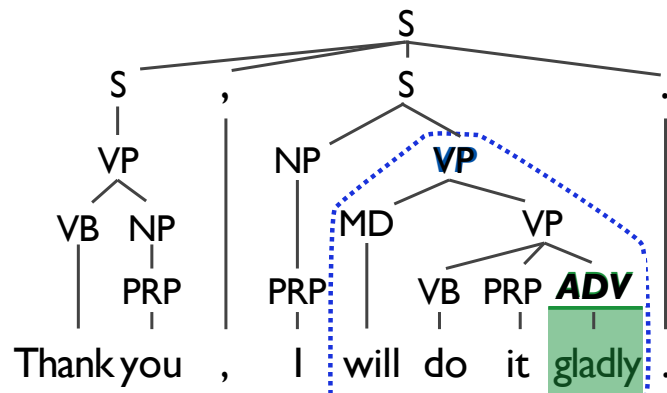
$$P(e|f) \propto \exp \left\{ \sum_i w_i \cdot f_i(e, f) \right\}$$

Chosen to minimize loss

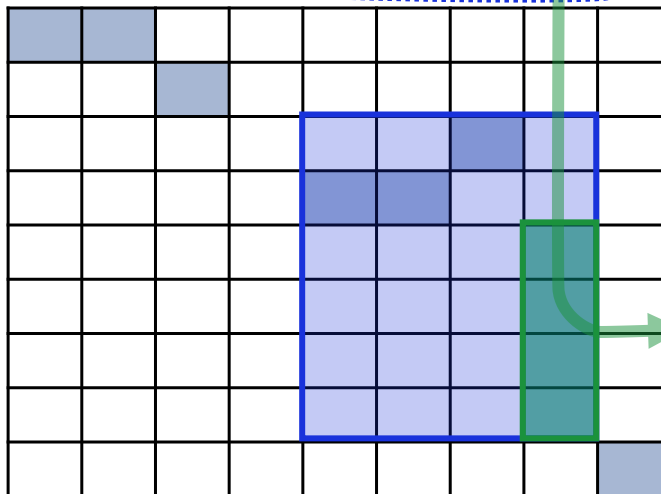
E.g., $\log P(e)$

Word Alignment and Phrase Extraction

Extracting Translation Rules



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



Gracias

,

lo

haré

de

muy

buen

grado

.

ADV

VP

will do it ADV

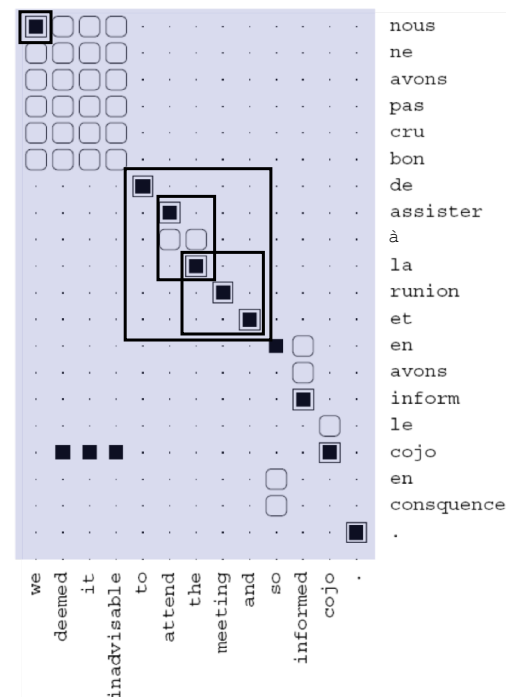
LO HARÉ

ADV

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.

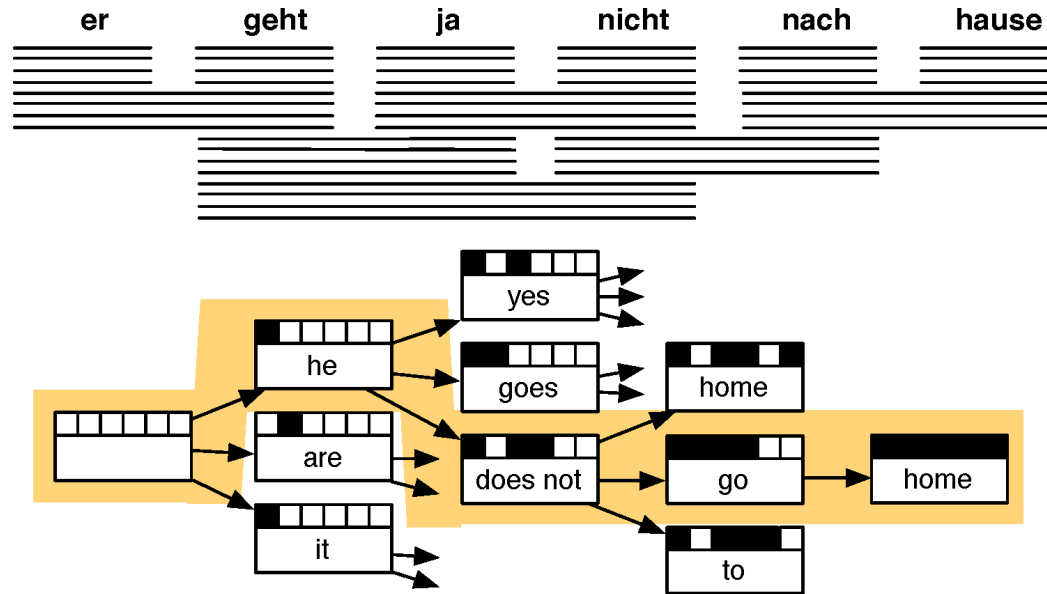


Translation Options

er	geht	ja	nicht	nach	hause
he	is	yes	not	after	house
it	are	is	do not	to	home
, it	goes	, of course	does not	according to	chamber
, he	go	,	is not	in	at home
it is		not		home	
he will be		is not		under house	
it goes		does not		return home	
he goes		do not		do not	
	is		to		
	are		following		
	is after all		not after		
	does		not to		
	not				
	is not				
	are not				
	is not a				

- 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
	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			