Statistical Machine Translation Part 2

Te Rutherford; Based on Philipp Koehn's slides

CS114 Spring 2015

How to train a translation model (from $f \rightarrow e$)

- Train for both $e \rightarrow f$ and $f \rightarrow e$
  - Train IBM Model 1 (Why?)
  - Train Higher-order IBM Model of your choice
  - Get the best many-to-one word alignment for each sentence pair

- Use heuristics to get many-to-many word alignment.

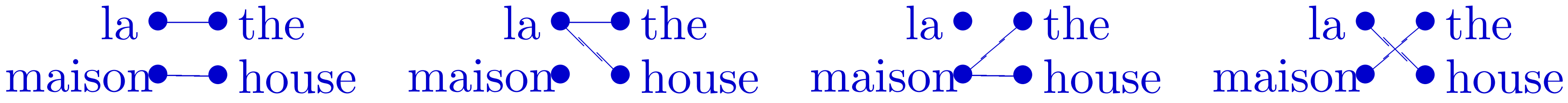
- Extract phrases and score them.

IBM Model 1 and EM

- **Probabilities**

$$p(\text{the}|\text{la}) = 0.7 \qquad p(\text{house}|\text{la}) = 0.05$$
$$p(\text{the}|\text{maison}) = 0.1 \qquad p(\text{house}|\text{maison}) = 0.8$$

- **Alignments**



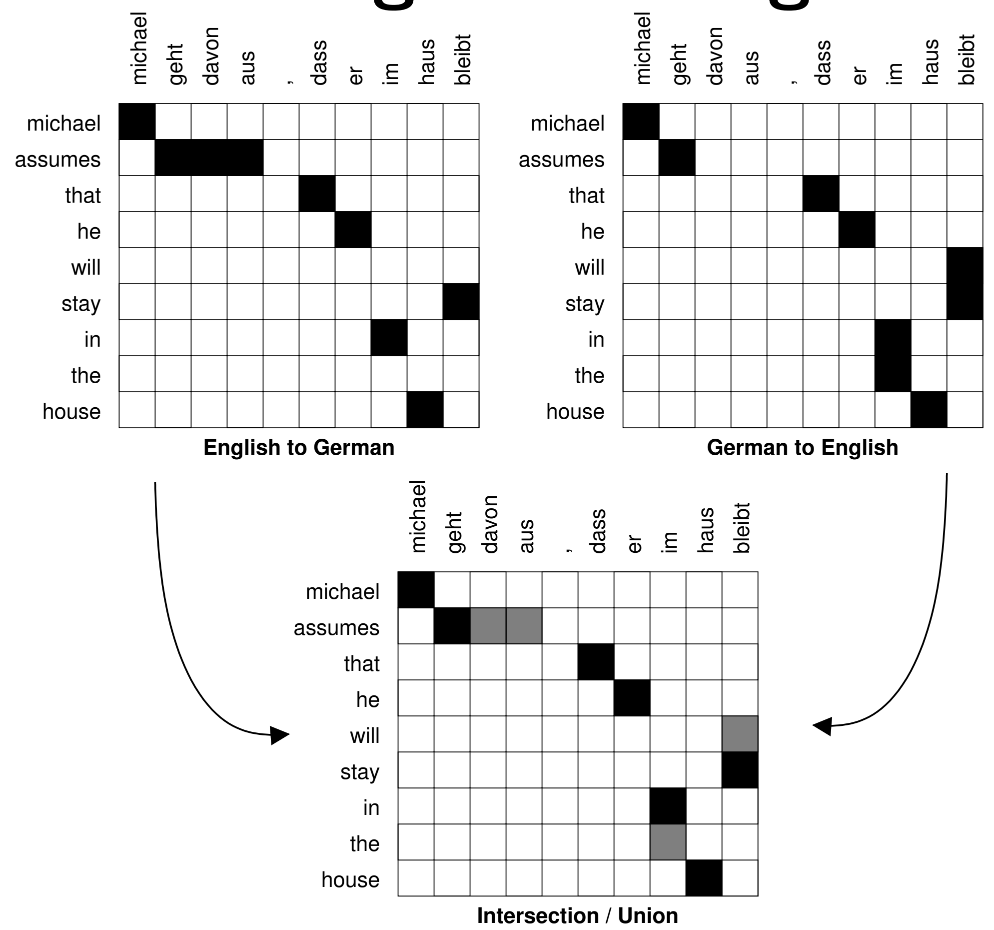
$$p(\mathbf{e}, a|\mathbf{f}) = 0.56 \qquad p(\mathbf{e}, a|\mathbf{f}) = 0.035 \qquad p(\mathbf{e}, a|\mathbf{f}) = 0.08 \qquad p(\mathbf{e}, a|\mathbf{f}) = 0.005$$

$$p(a|\mathbf{e}, \mathbf{f}) = 0.824 \qquad p(a|\mathbf{e}, \mathbf{f}) = 0.052 \qquad p(a|\mathbf{e}, \mathbf{f}) = 0.118 \qquad p(a|\mathbf{e}, \mathbf{f}) = 0.007$$

- **Counts**

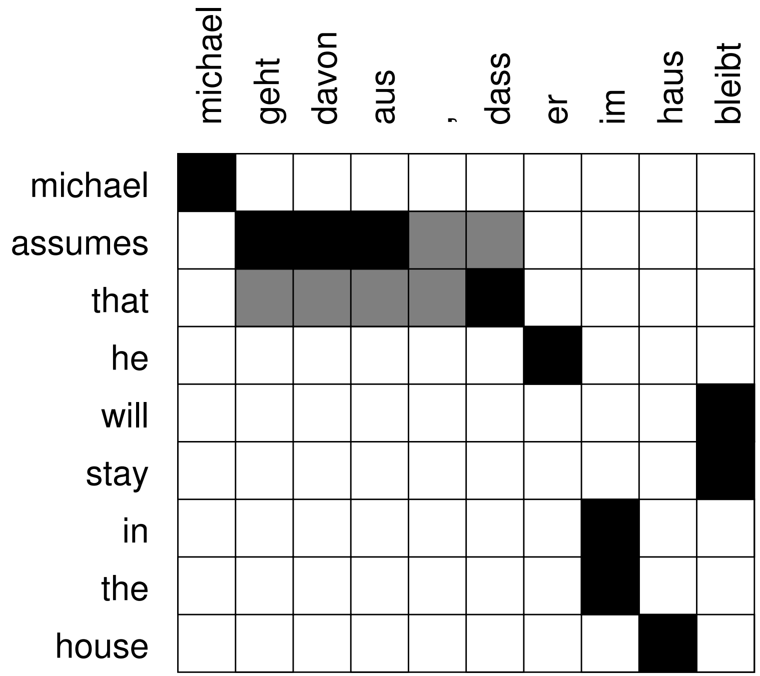
$$c(\text{the}|\text{la}) = 0.824 + 0.052 = \frac{0.7}{0.7+0.1}$$
$$c(\text{house}|\text{la}) = 0.052 + 0.007 = \frac{0.05}{0.05+0.8}$$
$$c(\text{the}|\text{maison}) = 0.118 + 0.007 = ?$$
$$c(\text{house}|\text{maison}) = 0.824 + 0.118 = ?$$

Symmetrizing Word Alignments



- Intersection of GIZA++ bidirectional alignments
- Grow additional alignment points [Och and Ney, CompLing2003]

Extracting Phrase Pairs



extract phrase pair consistent with word alignment:

assumes that / geht davon aus , dass

Real Example

- Phrase translations for den Vorschlag learned from the Europarl corpus:

English	$\phi(\bar{e}\|\bar{f})$	English	$\phi(\bar{e}\|\bar{f})$
the proposal	0.6227	the suggestions	0.0114
's proposal	0.1068	the proposed	0.0114
a proposal	0.0341	the motion	0.0091
the idea	0.0250	the idea of	0.0091
this proposal	0.0227	the proposal ,	0.0068
proposal	0.0205	its proposal	0.0068
of the proposal	0.0159	it	0.0068
the proposals	0.0159	...	...

  - lexical variation (proposal vs suggestions)
  - morphological variation (proposal vs proposals)
  - included function words (the, a, ...)
  - noise (it)

Scoring Phrase Translations

- Phrase pair extraction: collect all phrase pairs from the data
- Phrase pair scoring: assign probabilities to phrase translations
- Score by relative frequency:

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

Today

- Decoding – how to translate using the models.
- Evaluation – how to figure out if we have improved.

Decoding

- We have a mathematical model for translation

$$p(\mathbf{e}|\mathbf{f})$$

- Task of decoding: find the translation $\mathbf{e}_{\text{best}}$ with highest probability

$$\mathbf{e}_{\text{best}} = \text{argmax}_{\mathbf{e}}\, p(\mathbf{e}|\mathbf{f})$$

- Two types of error
  - the most probable translation is bad $\rightarrow$ fix the model
  - search does not find the most probably translation $\rightarrow$ fix the search

- Decoding is evaluated by search error, not quality of translations (although these are often correlated)

Computing Translation Probability

- Probabilistic model for phrase-based translation:

$$\mathbf{e}_{\text{best}} = \text{argmax}_{\mathbf{e}} \prod_{i=1}^{I} \phi(\bar{f}_i|\bar{e}_i)\ d(start_i - end_{i-1} - 1)\ p_{\text{LM}}(\mathbf{e})$$

- Score is computed incrementally for each partial hypothesis

- Components

Phrase translation Picking phrase $\bar{f}_i$ to be translated as a phrase $\bar{e}_i$
$\rightarrow$ look up score $\phi(\bar{f}_i|\bar{e}_i)$ from phrase translation table

Reordering Previous phrase ended in end_{i-1}, current phrase starts at $start_i$
$\rightarrow$ compute $d(start_i - end_{i-1} - 1)$

Language model For n-gram model, need to keep track of last $n-1$ words
$\rightarrow$ compute score $p_{\text{LM}}(w_i|w_{i-(n-1)}, ..., w_{i-1})$ for added words w_i

Translation Process

- Task: translate this sentence from German into English

er	**geht**	**ja**	**nicht**	**nach**	**hause**

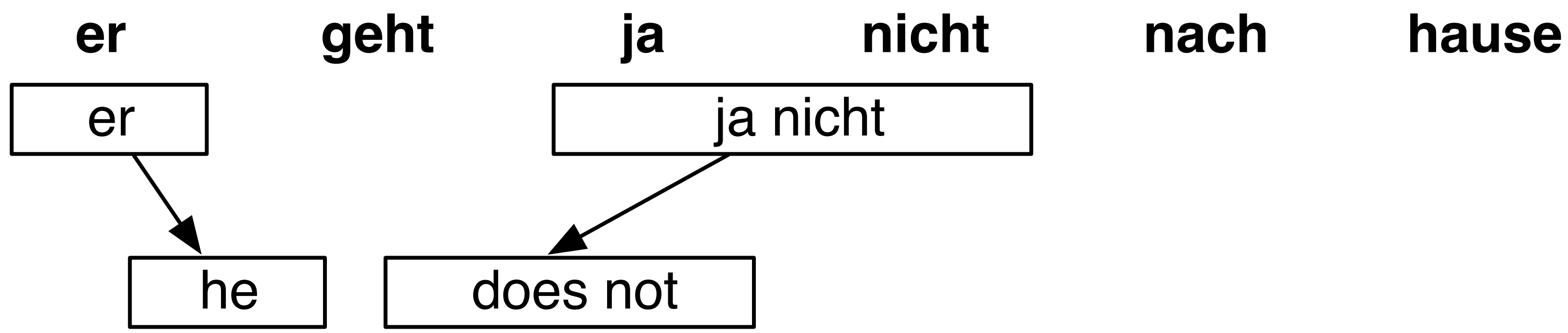
Translation Process

- Task: translate this sentence from German into English

- Pick phrase in input, translate

Translation Process

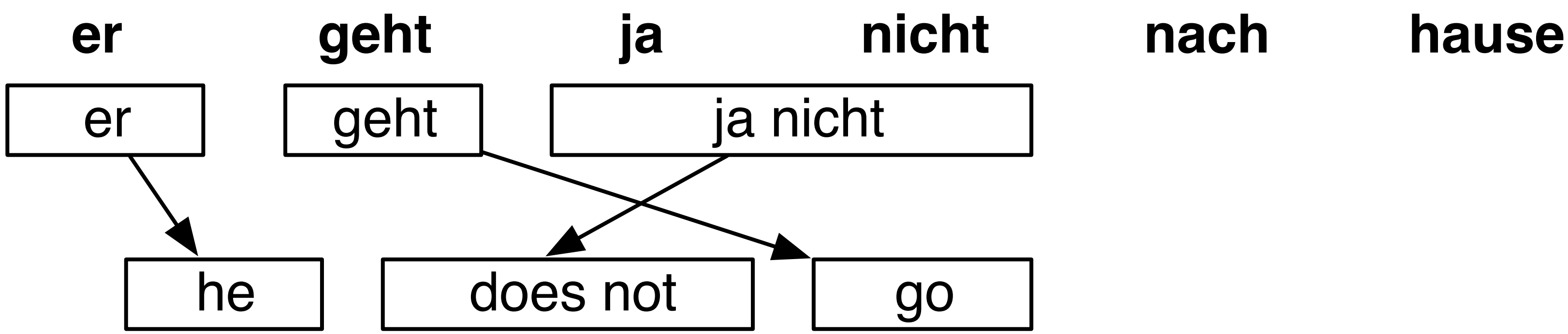
- Task: translate this sentence from German into English



- Pick phrase in input, translate
  - it is allowed to pick words out of sequence reordering
  - phrases may have multiple words: many-to-many translation

Translation Process

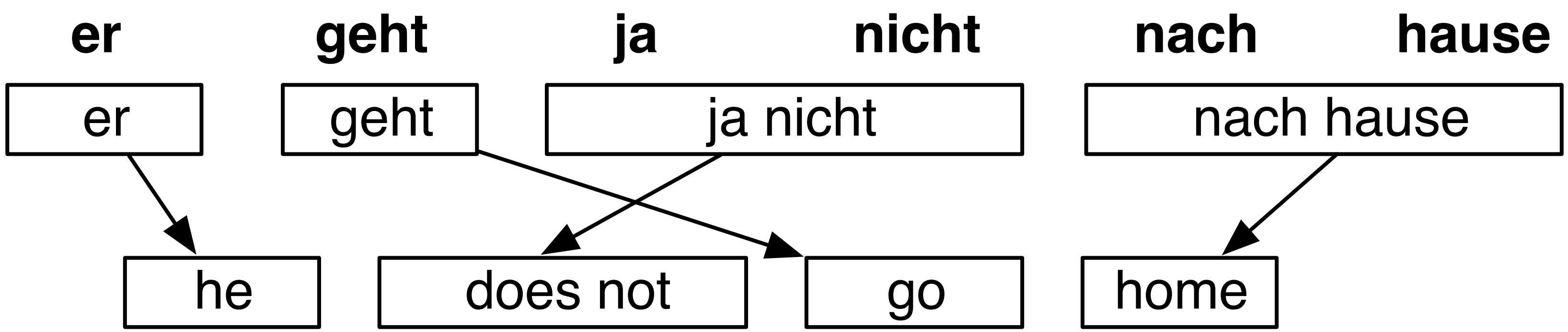
- Task: translate this sentence from German into English



- Pick phrase in input, translate

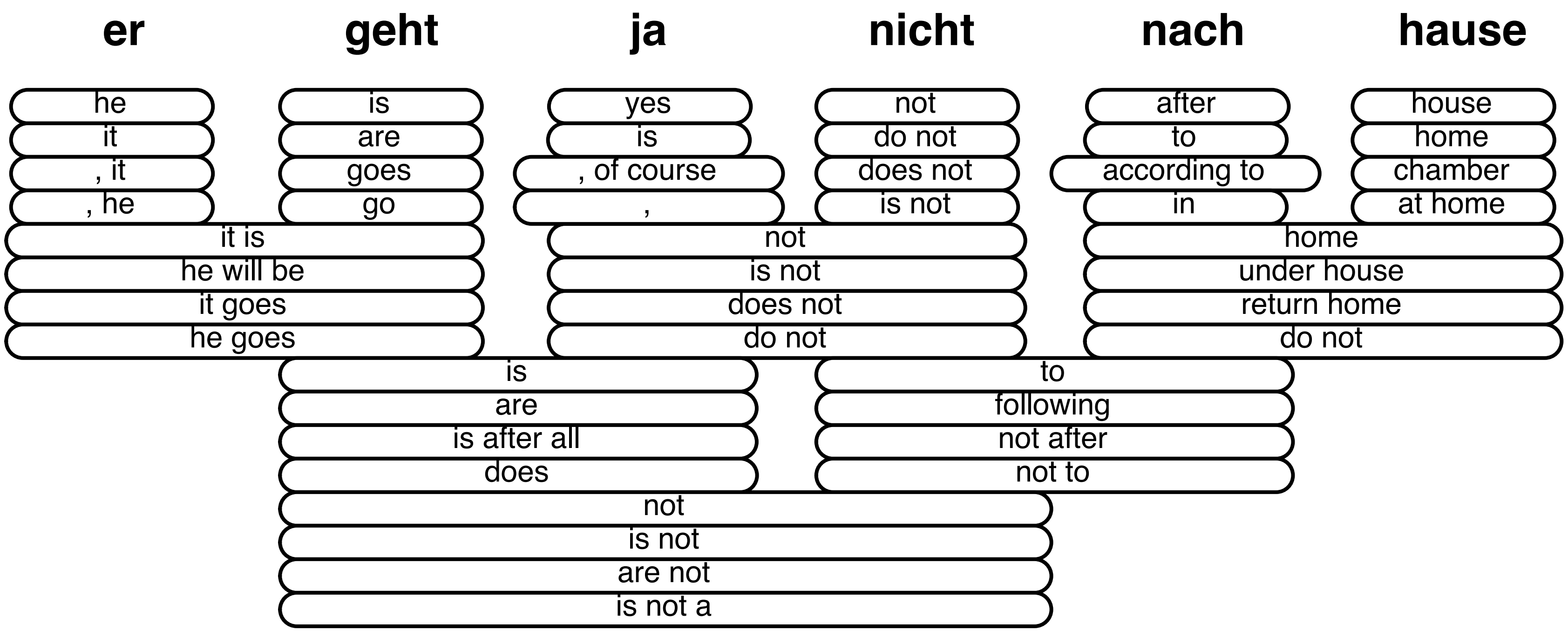
Translation Process

- Task: translate this sentence from German into English



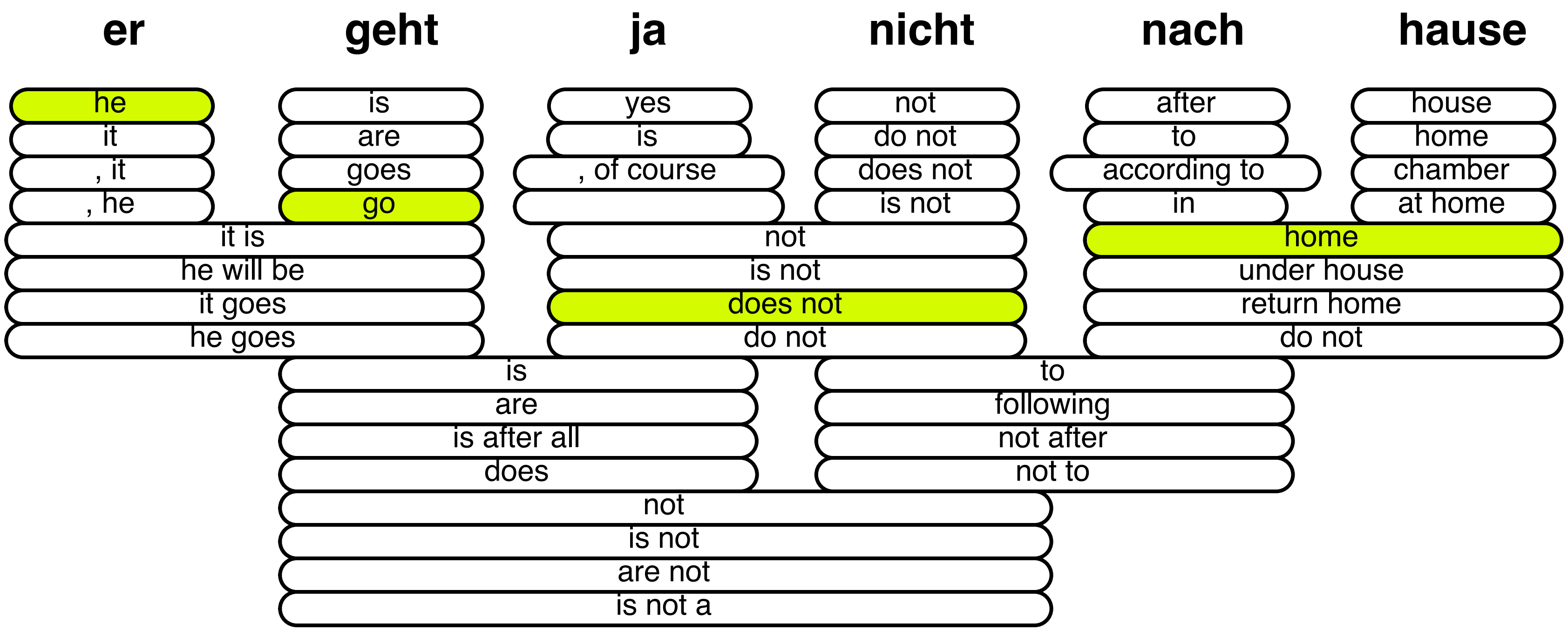
- Pick phrase in input, translate

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

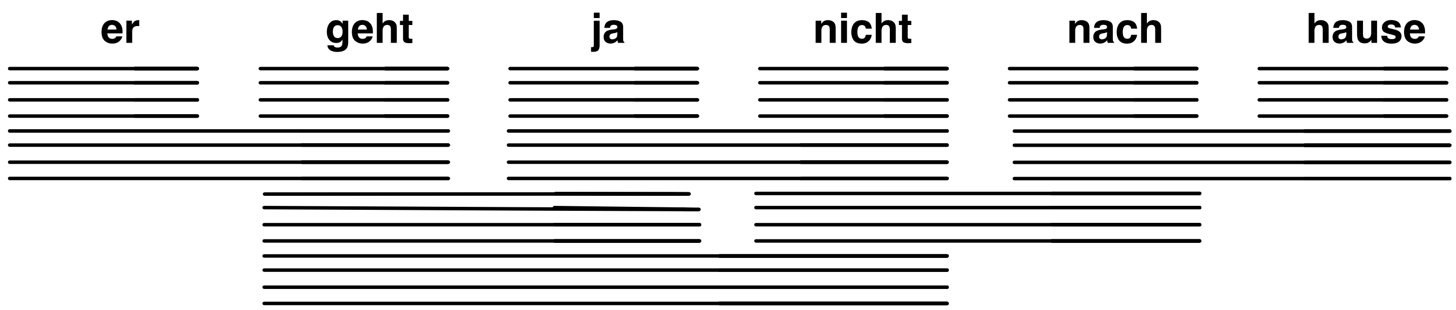
Translation Options



- The machine translation decoder does not know the right answer
  - picking the right translation options
  - arranging them in the right order

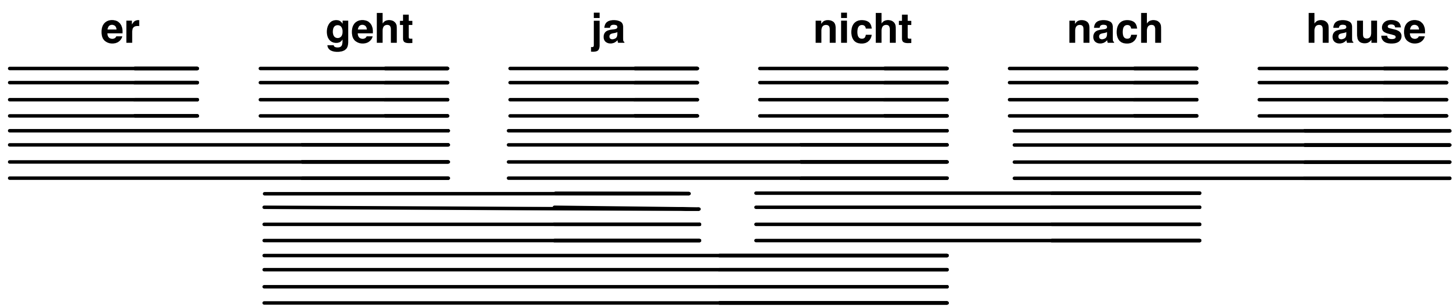
→ Search problem solved by heuristic beam search

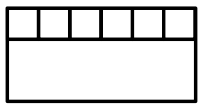
Decoding: Precompute Translation Options



consult phrase translation table for all input phrases

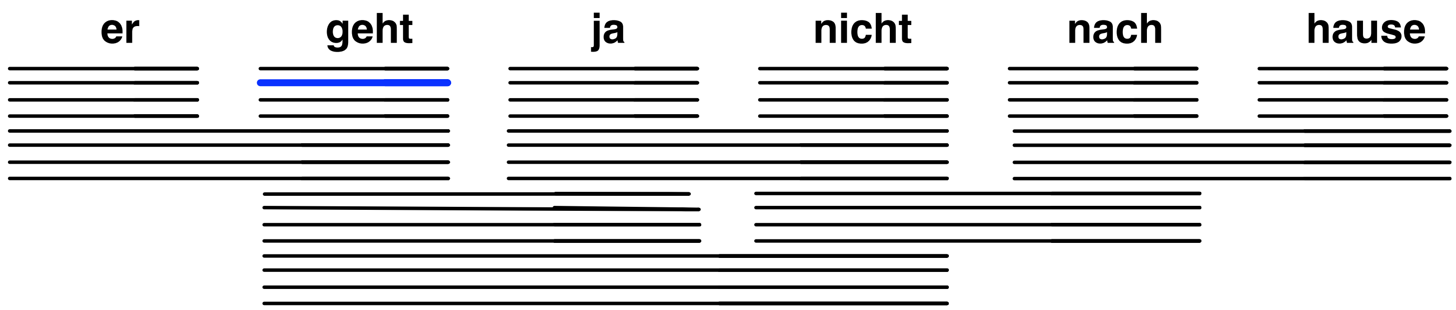
Decoding: Start with Initial Hypothesis

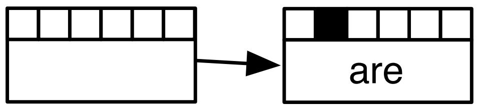




initial hypothesis: no input words covered, no output produced

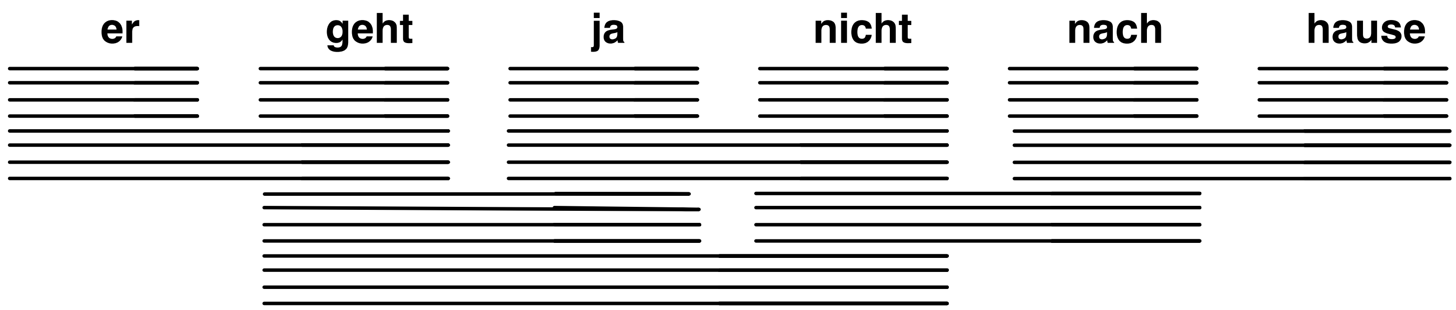
Decoding: Hypothesis Expansion

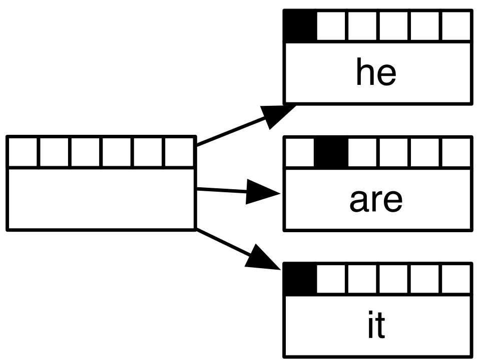




pick any translation option, create new hypothesis

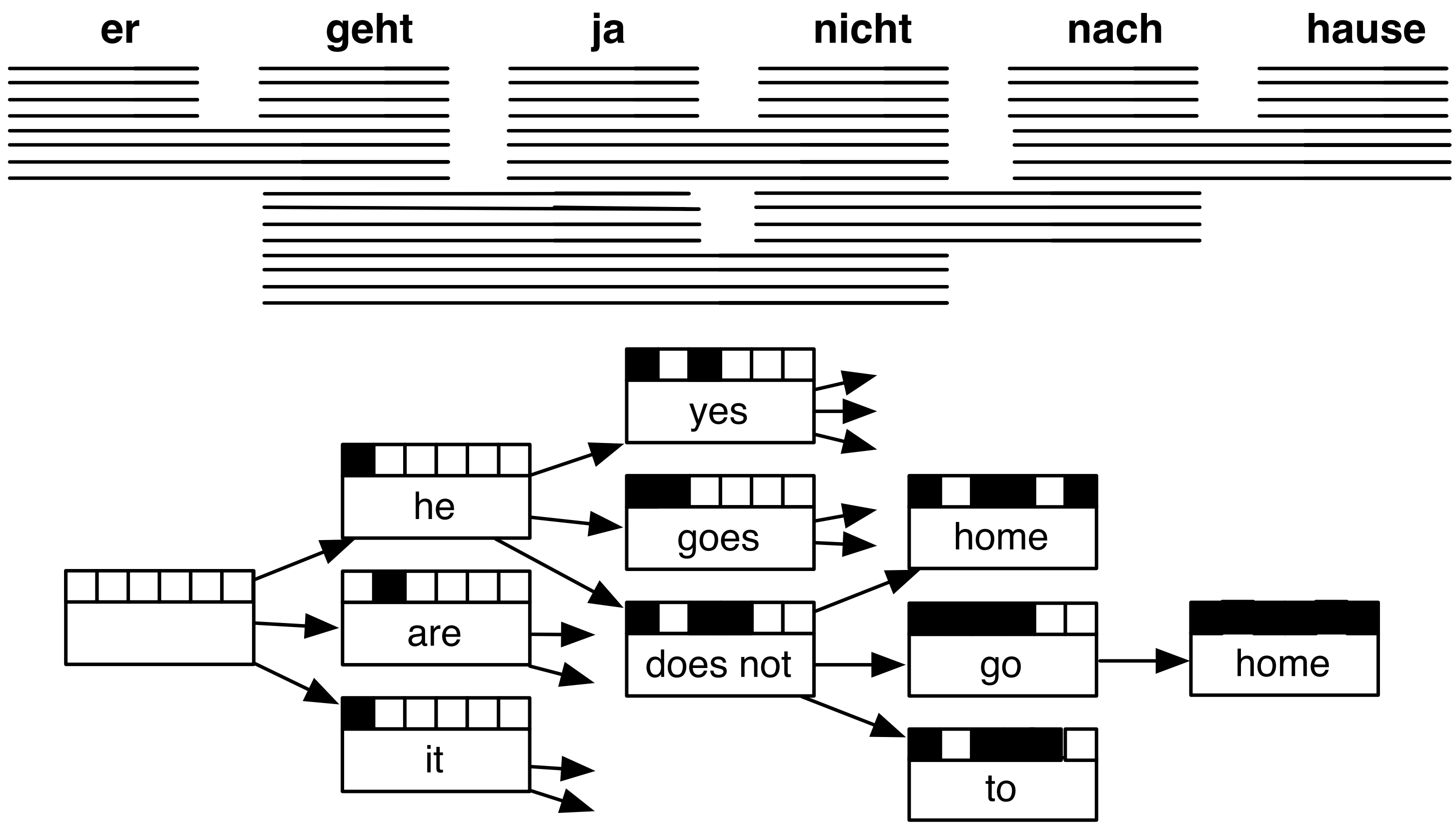
Decoding: Hypothesis Expansion





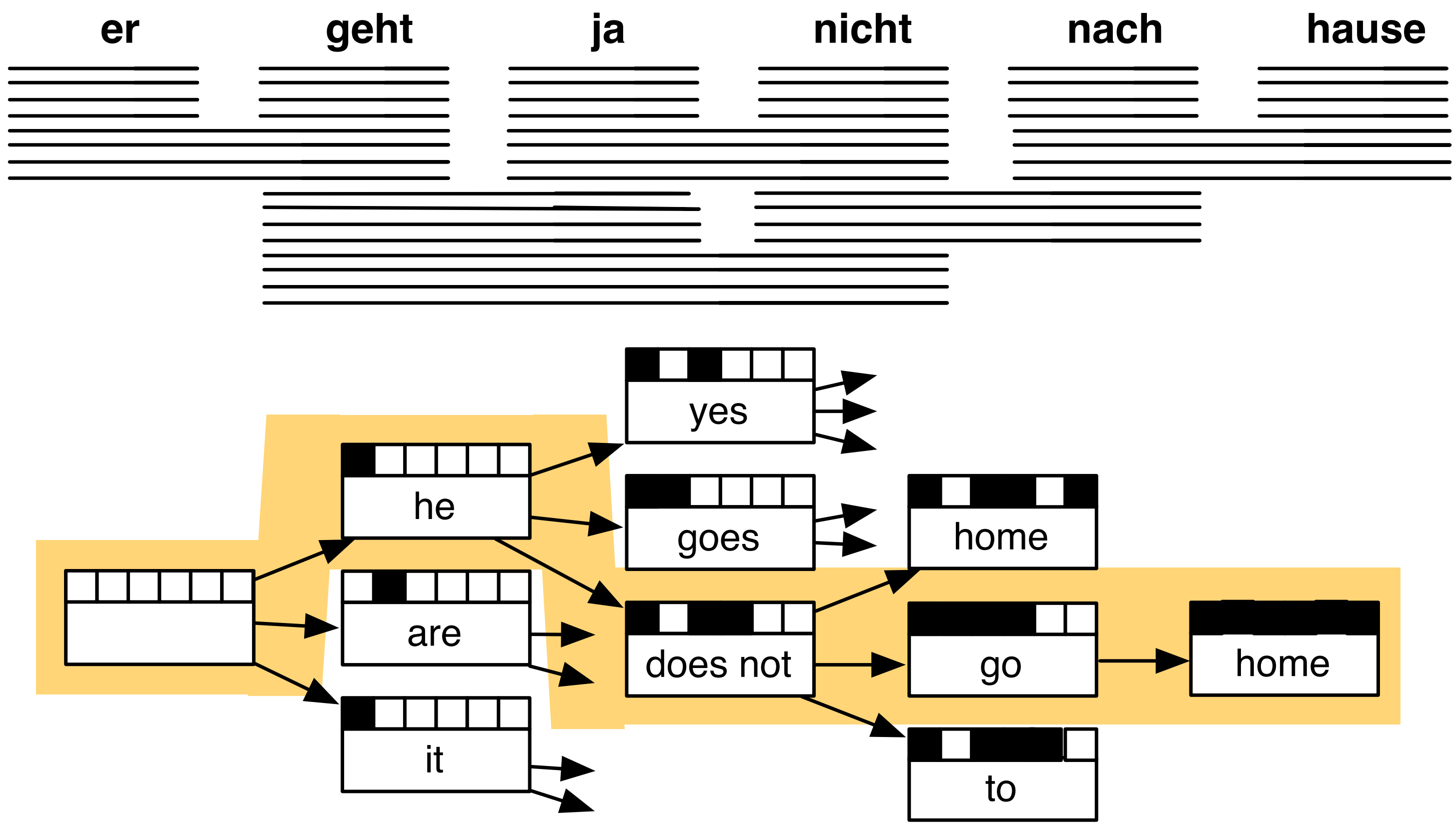
create hypotheses for all other translation options

Decoding: Hypothesis Expansion



also create hypotheses from created partial hypothesis

Decoding: Find Best Path



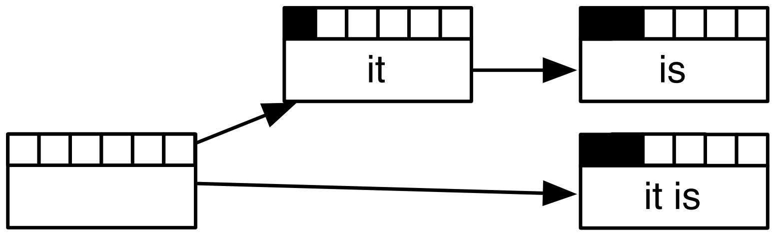
backtrack from highest scoring complete hypothesis

Computational Complexity

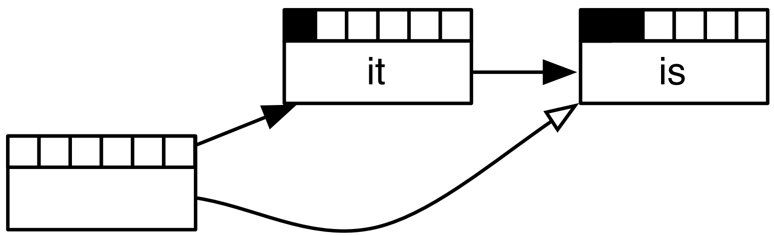
- The suggested process creates exponential number of hypothesis
- Machine translation decoding is NP-complete
- Reduction of search space:
  - recombination (risk-free)
  - pruning (risky)

Recombination

- Two hypothesis paths lead to two matching hypotheses
  - same number of foreign words translated
  - same English words in the output
  - different scores



- Worse hypothesis is dropped



Evaluation

- How good is a given machine translation system?
- Hard problem, since many different translations acceptable
  $\rightarrow$ semantic equivalence / similarity
- Evaluation metrics
  - subjective judgments by human evaluators
  - automatic evaluation metrics
  - task-based evaluation, e.g.:
    - how much post-editing effort?
    - does information come across?

Ten Translations of a Chinese Sentence

这个 机场 的 安全 工作 由 以色列 方面 负责 .

Israeli officials are responsible for airport security.
Israel is in charge of the security at this airport.
The security work for this airport is the responsibility of the Israel government.
Israeli side was in charge of the security of this airport.
Israel is responsible for the airport's security.
Israel is responsible for safety work at this airport.
Israel presides over the security of the airport.
Israel took charge of the airport security.
The safety of this airport is taken charge of by Israel.
This airport's security is the responsibility of the Israeli security officials.

(a typical example from the 2001 NIST evaluation set)

Adequacy and Fluency

- Human judgement
  - given: machine translation output
  - given: source and/or reference translation
  - task: assess the quality of the machine translation output

- Metrics

  Adequacy: Does the output convey the same meaning as the input sentence? Is part of the message lost, added, or distorted?

  Fluency: Is the output good fluent English? This involves both grammatical correctness and idiomatic word choices.

Fluency and Adequacy: Scales

Adequacy	
5	all meaning
4	most meaning
3	much meaning
2	little meaning
1	none

Fluency	
5	flawless English
4	good English
3	non-native English
2	disfluent English
1	incomprehensible

Annotation Tool

Judge Sentence

You have already judged 14 of 3064 sentences, taking 86.4 seconds per sentence.

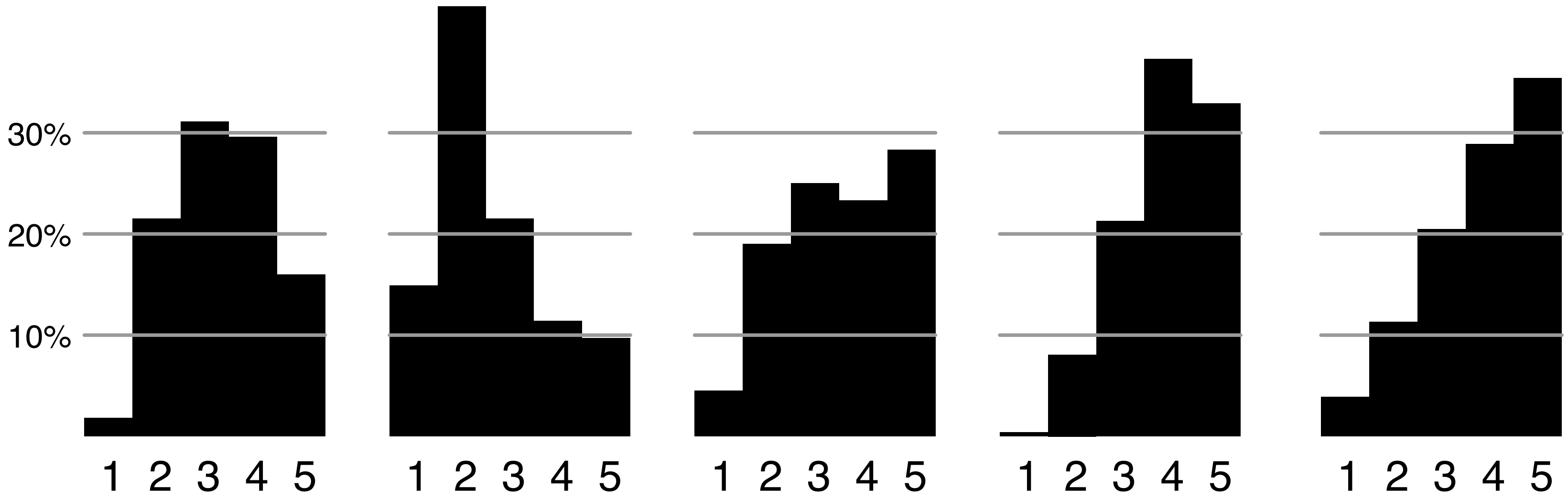
Source: les deux pays constituent plutôt un laboratoire nécessaire au fonctionnement interne de l ' ue .

Reference: rather , the two countries form a laboratory needed for the internal working of the eu .

Translation	**Adequacy**	**Fluency**
both countries are rather a necessary laboratory the internal operation of the eu .	1 2 3 4 5	1 2 3 4 5
both countries are a necessary laboratory at internal functioning of the eu .	1 2 3 4 5	1 2 3 4 5
the two countries are rather a laboratory necessary for the internal workings of the eu .	1 2 3 4 5	1 2 3 4 5
the two countries are rather a laboratory for the internal workings of the eu .	1 2 3 4 5	1 2 3 4 5
the two countries are rather a necessary laboratory internal workings of the eu .	1 2 3 4 5	1 2 3 4 5
Annotator: Philipp Koehn **Task:** WMT06 French-English	Annotate	
Instructions	5= All Meaning 4= Most Meaning 3= Much Meaning 2= Little Meaning 1= None	5= Flawless English 4= Good English 3= Non-native English 2= Disfluent English 1= Incomprehensible

Evaluators Disagree

- Histogram of adequacy judgments by different human evaluators



(from WMT 2006 evaluation)

Goals for Evaluation Metrics

Low cost: reduce time and money spent on carrying out evaluation

Tunable: automatically optimize system performance towards metric

Meaningful: score should give intuitive interpretation of translation quality

Consistent: repeated use of metric should give same results

Correct: metric must rank better systems higher

Automatic Evaluation Metrics

- Goal: computer program that computes the quality of translations
- Advantages: low cost, tunable, consistent
- Basic strategy
  - given: machine translation output
  - given: human reference translation
  - task: compute similarity between them

Word Error Rate

- Minimum number of editing steps to transform output to reference

  match: words match, no cost
  substitution: replace one word with another
  insertion: add word
  deletion: drop word

- Levenshtein distance

$$\text{WER} = \frac{\textit{substitutions} + \textit{insertions} + \textit{deletions}}{\textit{reference-length}}$$

Example

		Israeli	officials	responsibility	of	airport	safety
	0	1	2	3	4	5	6
Israeli	1	0	1	2	3	4	5
officials	2	1	0	1	2	3	4
are	3	2	1	1	2	3	4
responsible	4	3	2	2	2	3	4
for	5	4	3	3	3	3	4
airport	6	5	4	4	4	3	4
security	7	6	5	5	5	4	4

		airport	security	Israeli	officials	are	responsible
	0	1	2	3	4	5	6
Israeli	1	1	2	2	3	4	5
officials	2	2	2	3	2	3	4
are	3	3	3	3	3	2	3
responsible	4	4	4	4	4	3	2
for	5	5	5	5	5	4	3
airport	6	5	6	6	6	5	4
security	7	6	5	6	7	6	5

Metric	System A	System B
word error rate (WER)	57%	71%

BLEU

- N-gram overlap between machine translation output and reference translation
- Compute precision for n-grams of size 1 to 4
- Add brevity penalty (for too short translations)

$$\text{BLEU} = \min\left(1, \frac{\textit{output-length}}{\textit{reference-length}}\right) \, \Big(\prod_{i=1}^{4} \textit{precision}_i\Big)^{\frac{1}{4}}$$

- Typically computed over the entire corpus, not single sentences

Example

SYSTEM A: [Israeli officials] responsibility of [airport] safety
2-GRAM MATCH — 1-GRAM MATCH

REFERENCE: Israeli officials are responsible for airport security

SYSTEM B: [airport security] [Israeli officials are responsible]
2-GRAM MATCH — 4-GRAM MATCH

Metric	**System A**	**System B**
precision (1gram)	3/6	6/6
precision (2gram)	1/5	4/5
precision (3gram)	0/4	2/4
precision (4gram)	0/3	1/3
brevity penalty	6/7	6/7
BLEU	0%	52%

Critique of Automatic Metrics

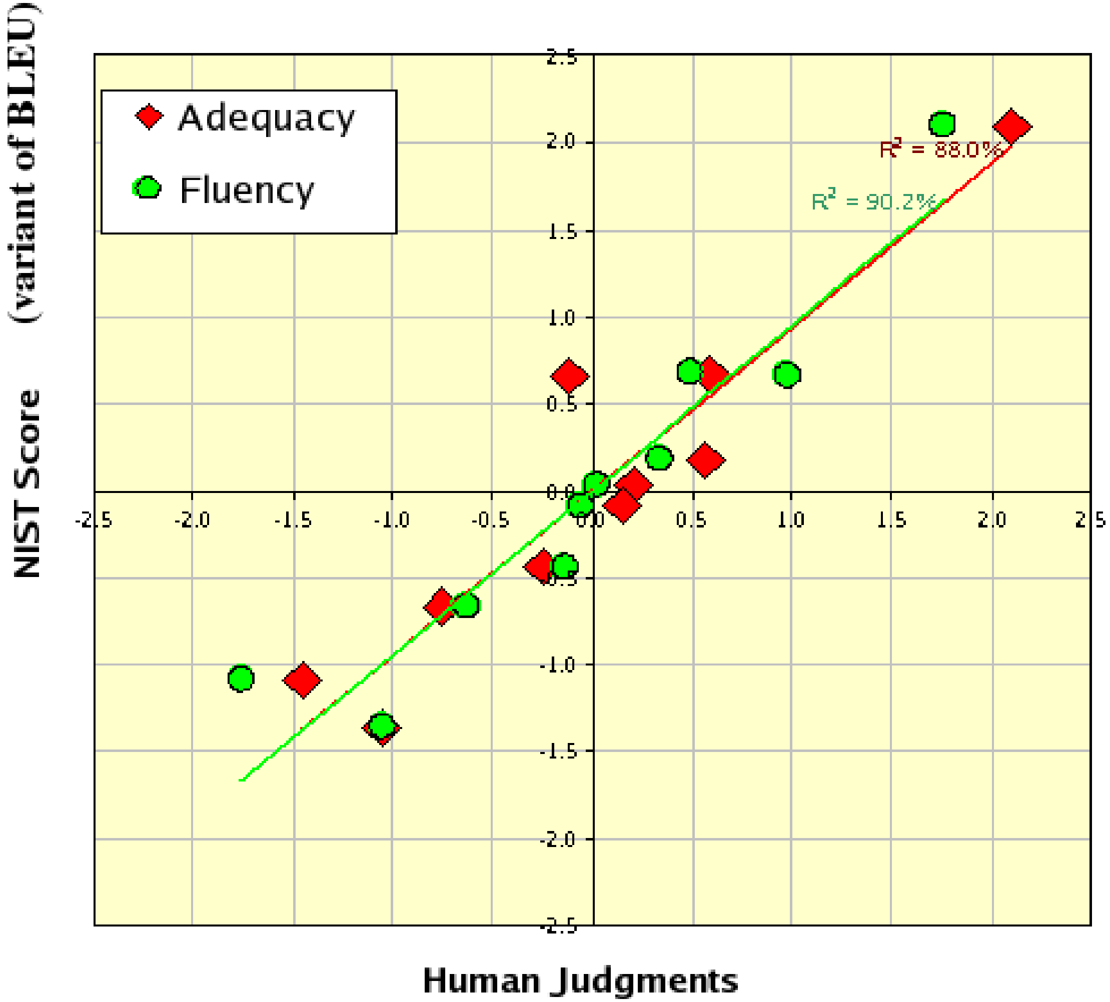
- Ignore relevance of words
  (names and core concepts more important than determiners and punctuation)
- Operate on local level
  (do not consider overall grammaticality of the sentence or sentence meaning)
- Scores are meaningless
  (scores very test-set specific, absolute value not informative)
- Human translators score low on BLEU
  (possibly because of higher variability, different word choices)

Evaluation of Evaluation Metrics

- Automatic metrics are low cost, tunable, consistent
- But are they correct?

$\rightarrow$ Yes, if they correlate with human judgement

Correlation with Human Judgement



Automatic Metrics: Conclusions

- Automatic metrics essential tool for system development
- Not fully suited to rank systems of different types
- Evaluation metrics still open challenge