Pharaoh: a Beam Search Decoder for Phrase-Based Statistical Machine Translation

Philipp Koehn

koehn@csail.mit.edu

Computer Science and Artificial Intelligence Lab

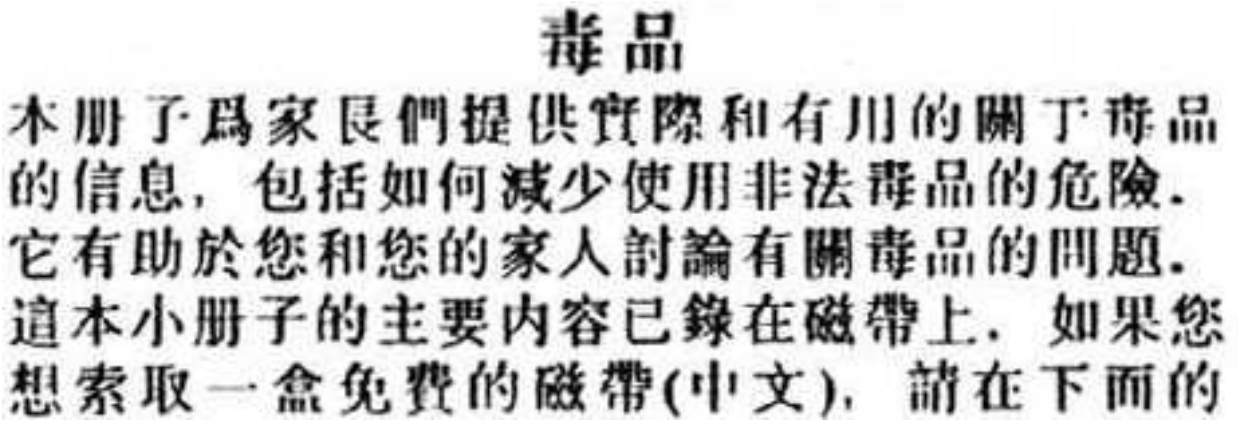
Massachusetts Institute of Technology

Outline

- **Phrase-Based Statistical MT**
- Beam Search Decoding
- Experiments
- Advanced Features

Machine Translation

- Task: Make sense of foreign text like

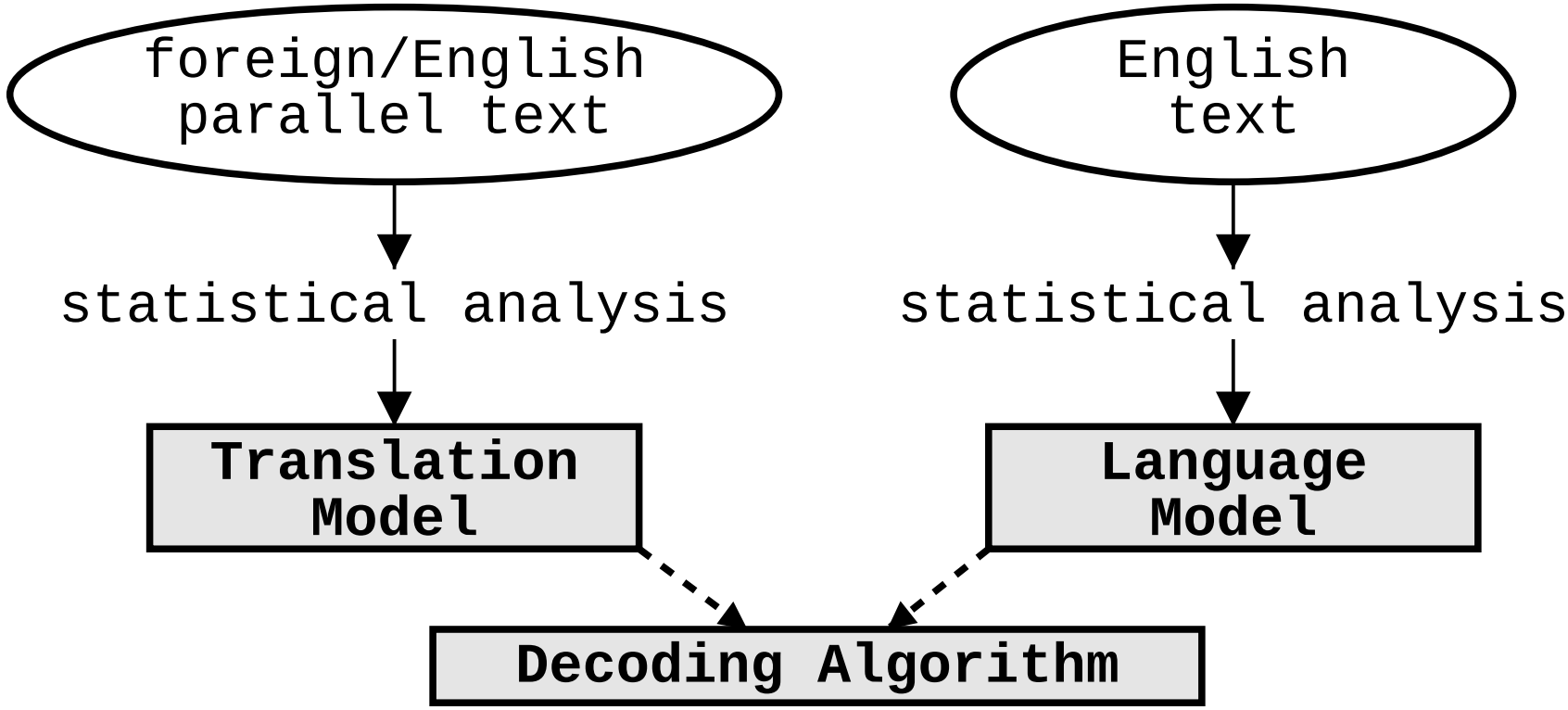

毒品

本册子爲家長們提供實際和有用的關于毒品
的信息，包括如何減少使用非法毒品的危險.
它有助於您和您的家人討論有關毒品的問題.
這本小册子的主要内容已錄在磁帶上. 如果您
想索取一盒免費的磁帶(中文)，請在下面的

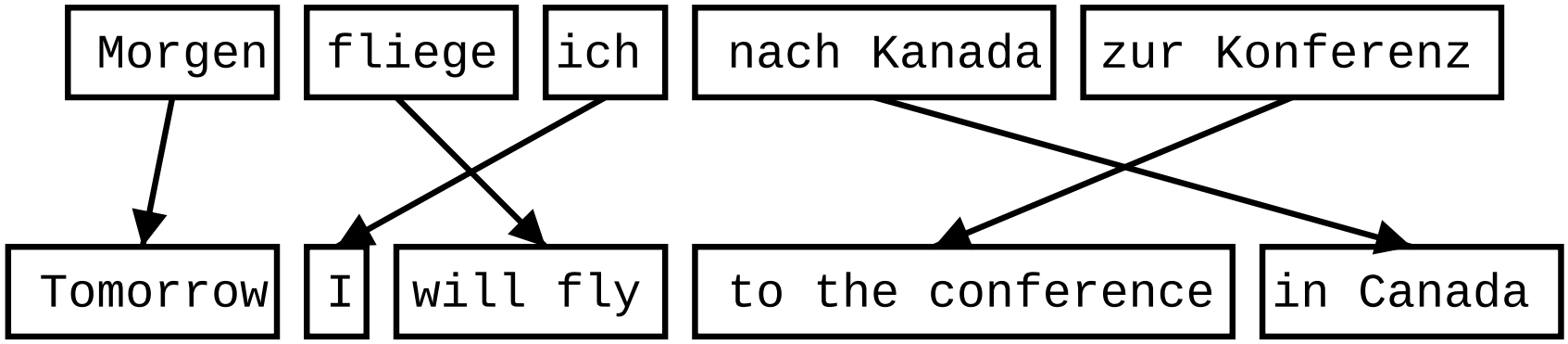
- Long-standing problem in artificial intelligence
- Ultimately requires syntax, semantics, pragmatics

Statistical Machine Translation

- Components: Translation model, language model, decoder



Phrase-Based Translation



- Foreign input is segmented in phrases
  - any sequence of words, not necessarily linguistically motivated
- Each phrase is translated into English
- Phrases are reordered

Phrase-Based Systems

- A number of research groups developed phrase-based systems (RWTH Aachen, USC/ISI, CMU, IBM, JHU, ITC-irst, MIT, ...)
- Systems differ in
  - training methods
  - model for phrase translation table
  - reordering models
  - additional feature functions
- Currently best method for SMT (MT?)
  - top systems in DARPA/NIST evaluation are phrase-based
  - best commercial system for Arabic-English is phrase-based

Pharaoh

- Translation engine
  - works with various phrase-based models
  - beam search algorithm
  - time complexity roughly linear with input length
  - good quality takes about 1 second per sentence
- Very good performance in DARPA/NIST Evaluation
- Freely available for researchers

  `http://www.isi.edu/licensed-sw/pharaoh/`

Outline

- Phrase-Based Statistical MT
- **Beam Search Decoding**
- Experiments
- Advanced Features

Decoding Process

- Build translation left to right
  - select foreign words to be translated

Decoding Process

- Build translation left to right
  - select foreign words to be translated
  - find English phrase translation
  - add English phrase to end of partial translation

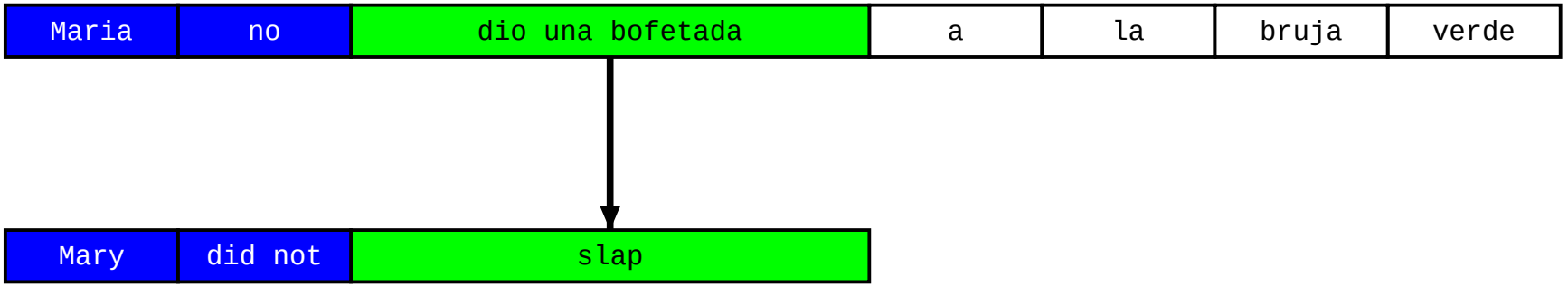
Decoding Process

- Build translation left to right
  - select foreign words to be translated
  - find English phrase translation
  - add English phrase to end of partial translation
  - mark foreign words as translated

Decoding Process

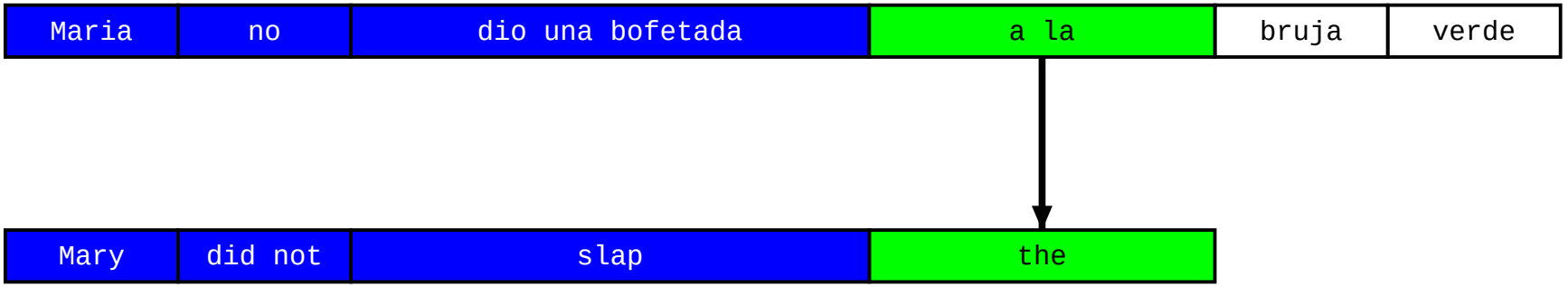
- One to many translation

Decoding Process



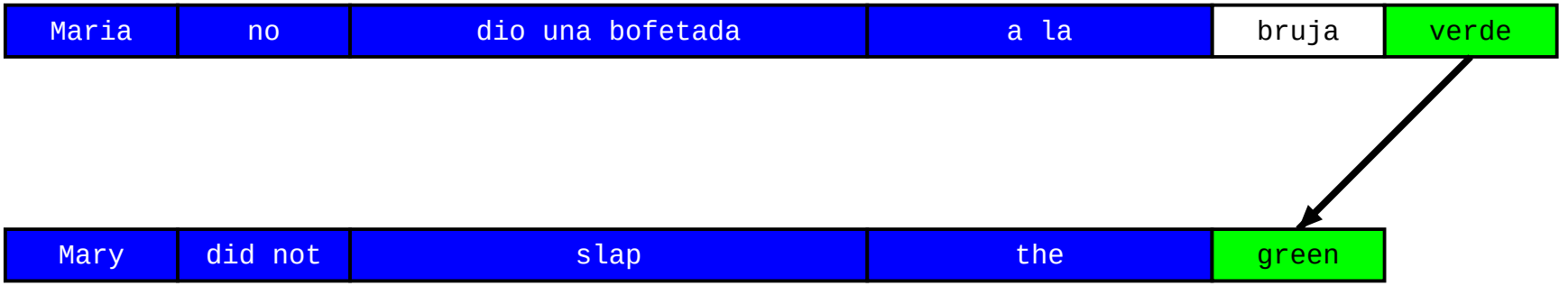
- Many to one translation

Decoding Process



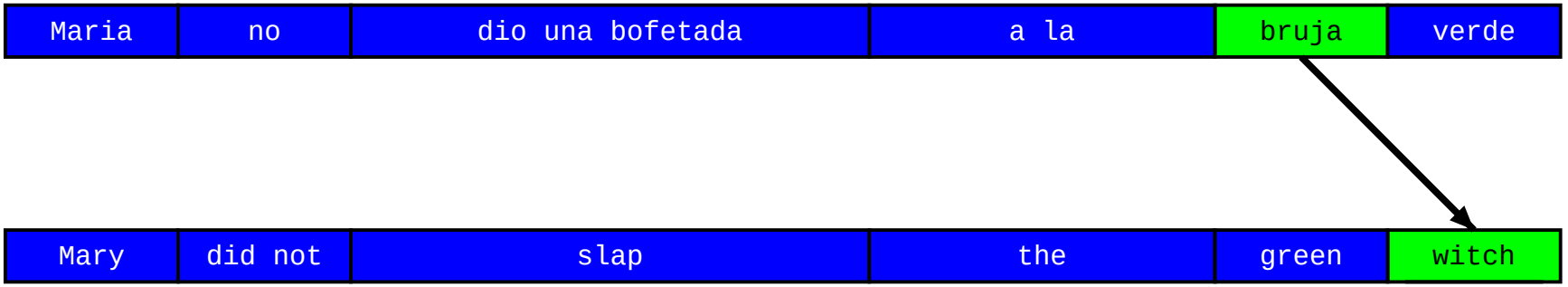
- Many to one translation

Decoding Process



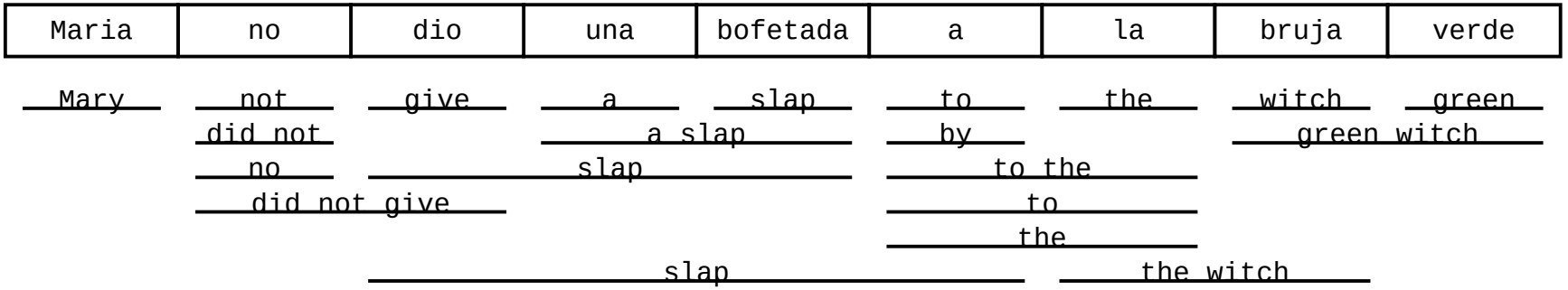
- Reordering

Decoding Process



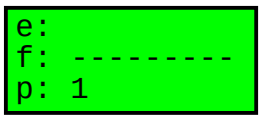
- Translation finished

Translation Options



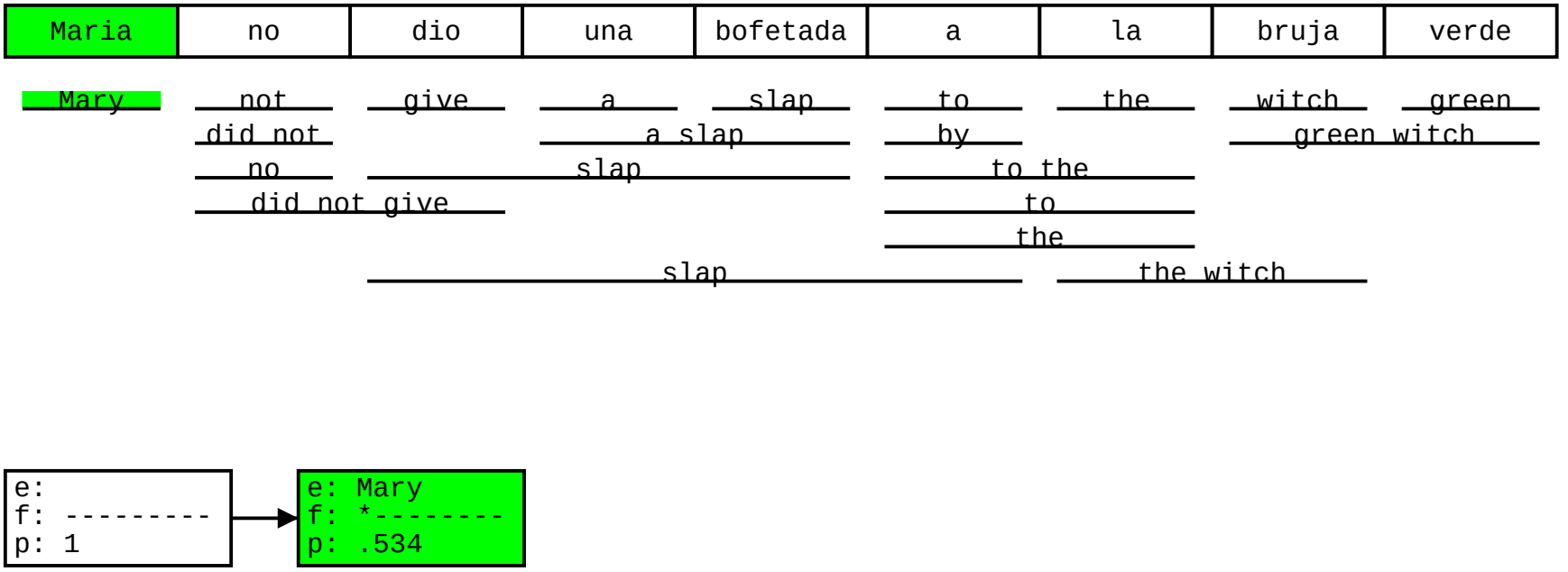
- Look up possible phrase translations
  - many different ways to segment words into phrases
  - many different ways to translate each phrase

Hypothesis Expansion



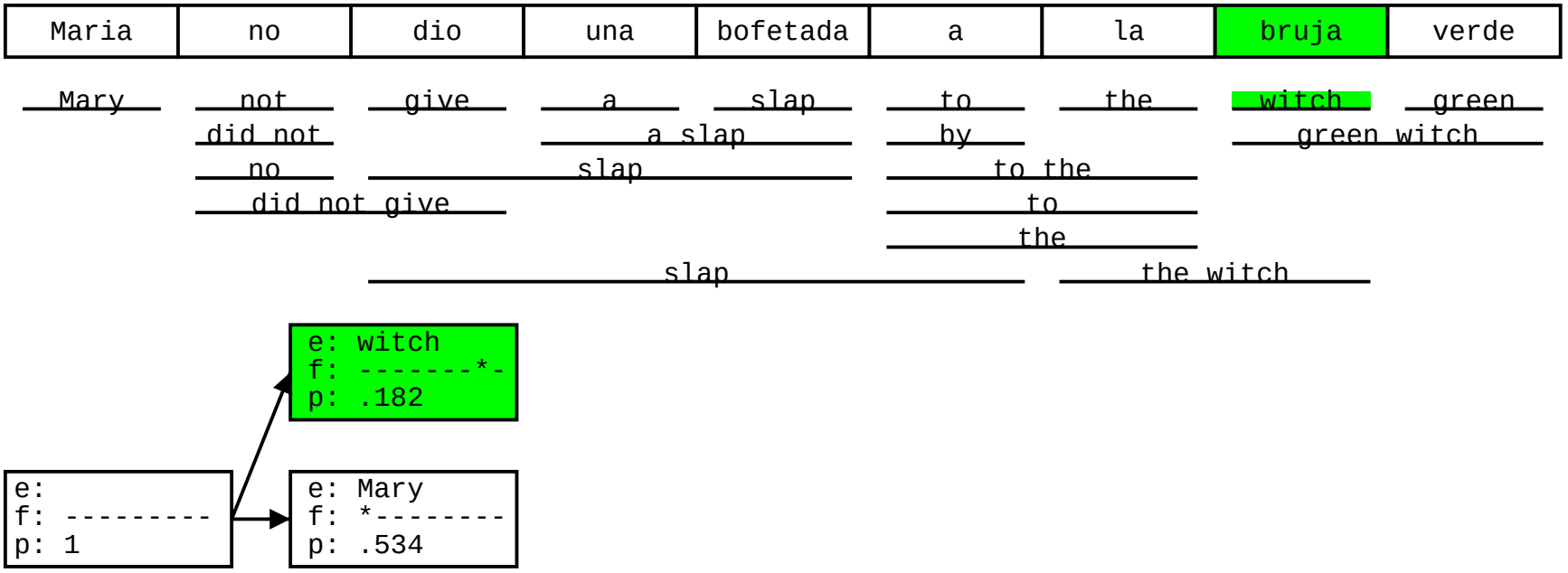
- Start with null hypothesis
  - e: no English words
  - f: no foreign words covered
  - p: probability 1

Hypothesis Expansion



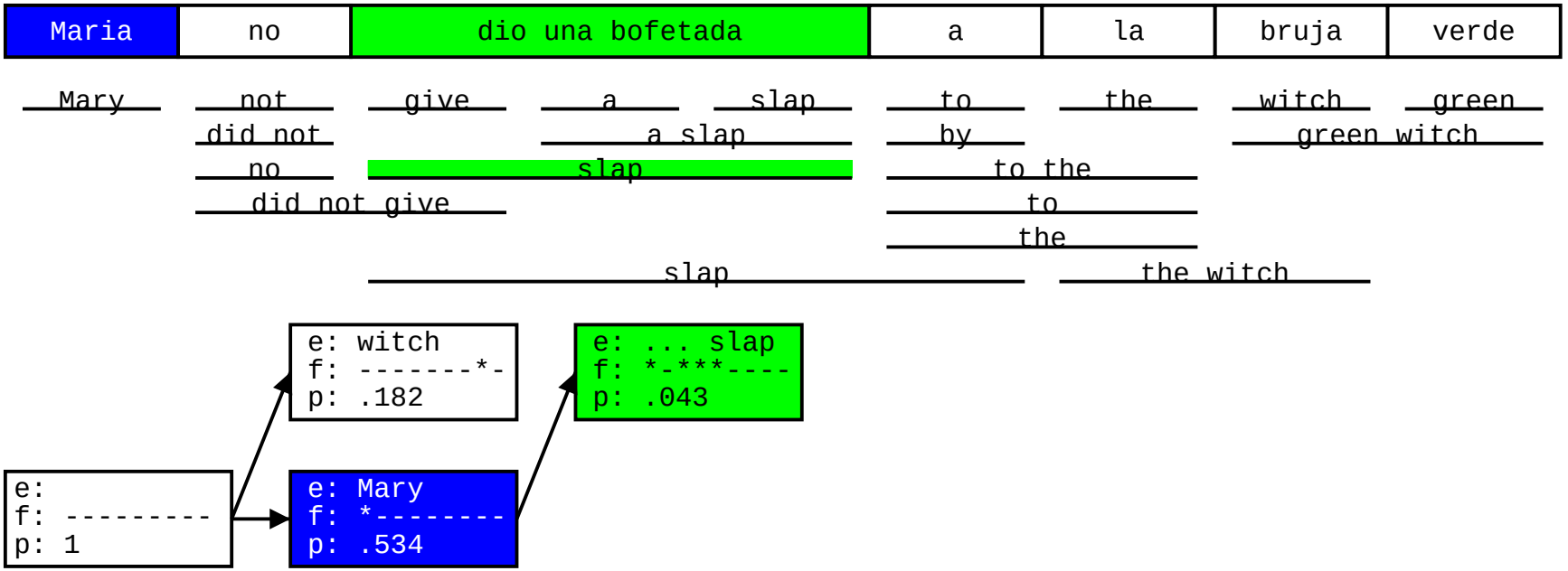
- Pick translation option
- Create hypothesis
  - e: add English phrase `Mary`
  - f: first foreign word covered
  - p: probability 0.534

Hypothesis Expansion



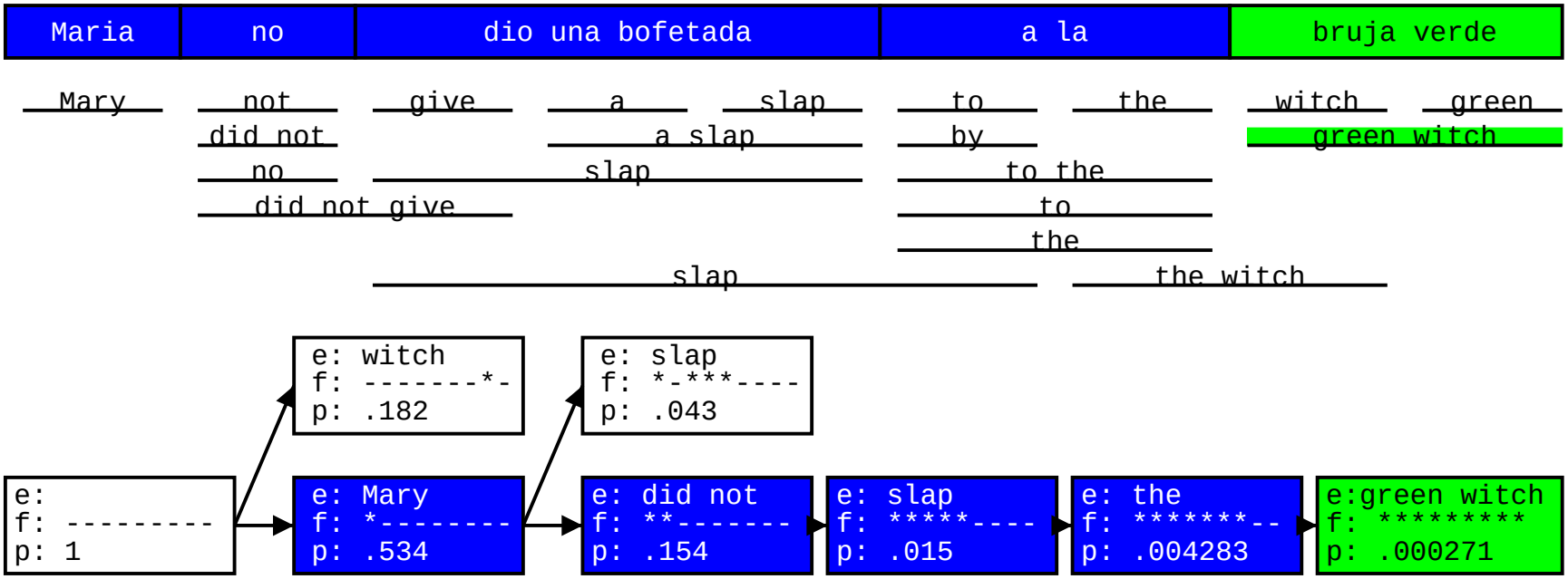
- Add another hypothesis

Hypothesis Expansion



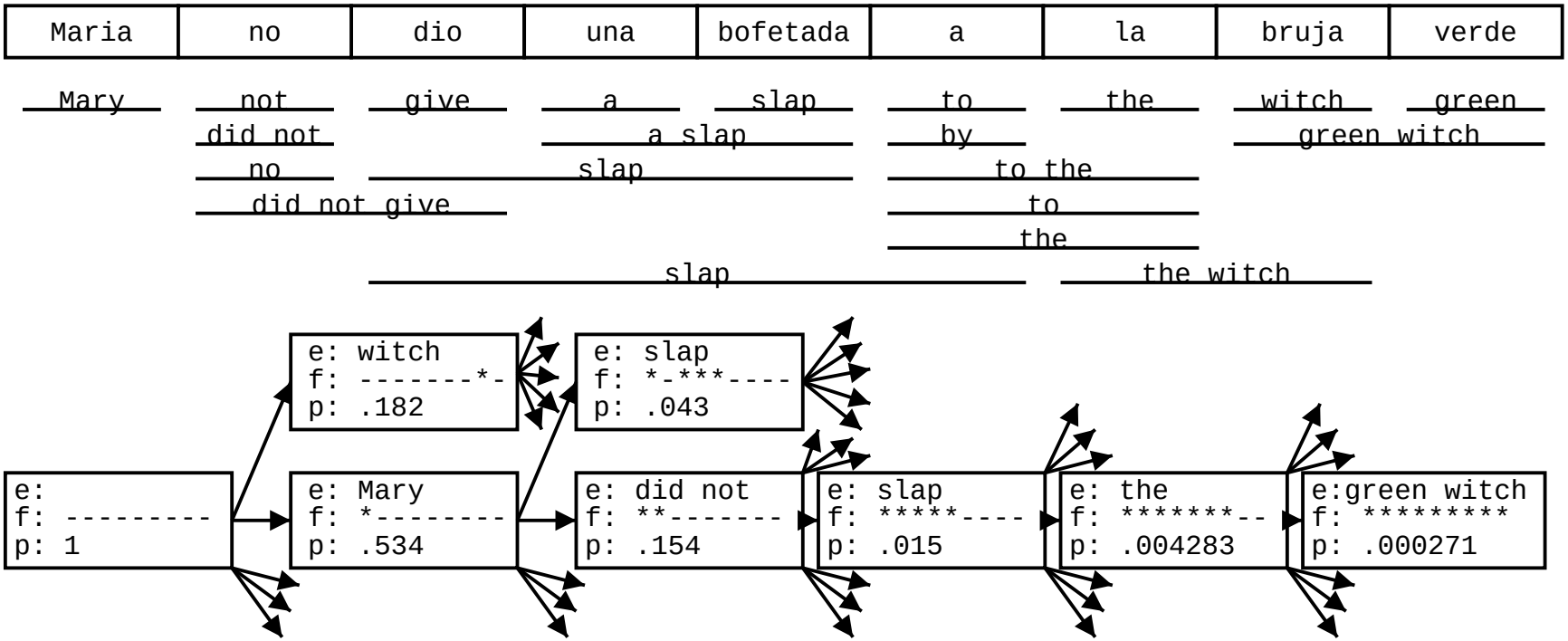
- Further hypothesis expansion

Hypothesis Expansion



- ... until all foreign words covered
  - find best hypothesis that covers all foreign words
  - backtrack to read off translation

Hypothesis Expansion



- Adding more hypothesis

$\Rightarrow$ Explosion of search space

Explosion of Search Space

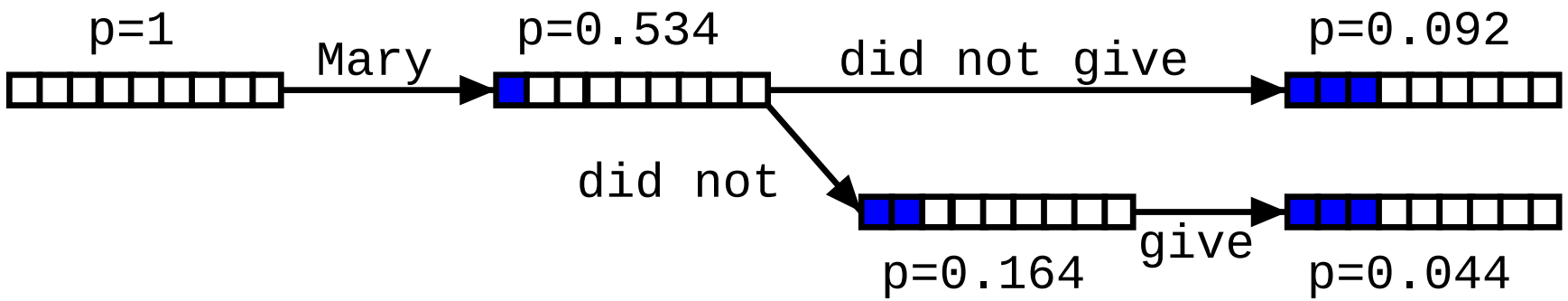
- Number of hypotheses is exponential with respect to sentence length

$\Rightarrow$ Decoding is NP-complete [Knight, 1999]

$\Rightarrow$ Need to reduce search space

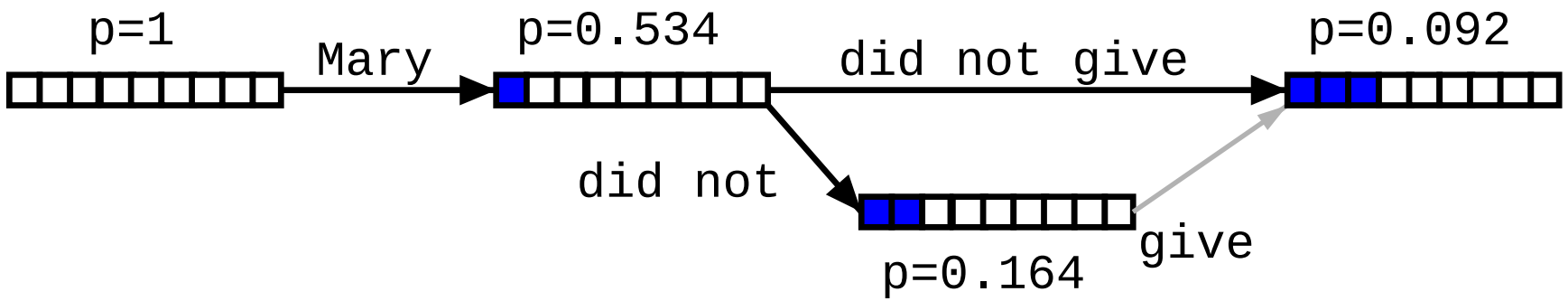
  - risk free: hypothesis recombination
  - risky: histogram/threshold pruning

Hypothesis Recombination



- Different paths to the same partial translation

Hypothesis Recombination

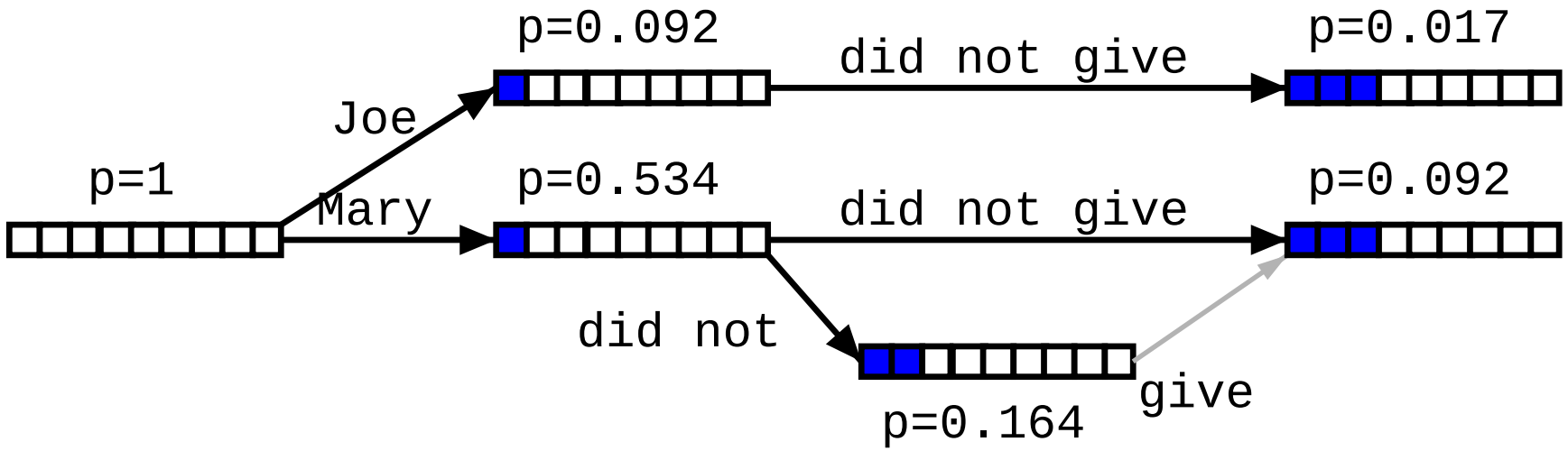


- Different paths to the same partial translation

⇒ Combine paths

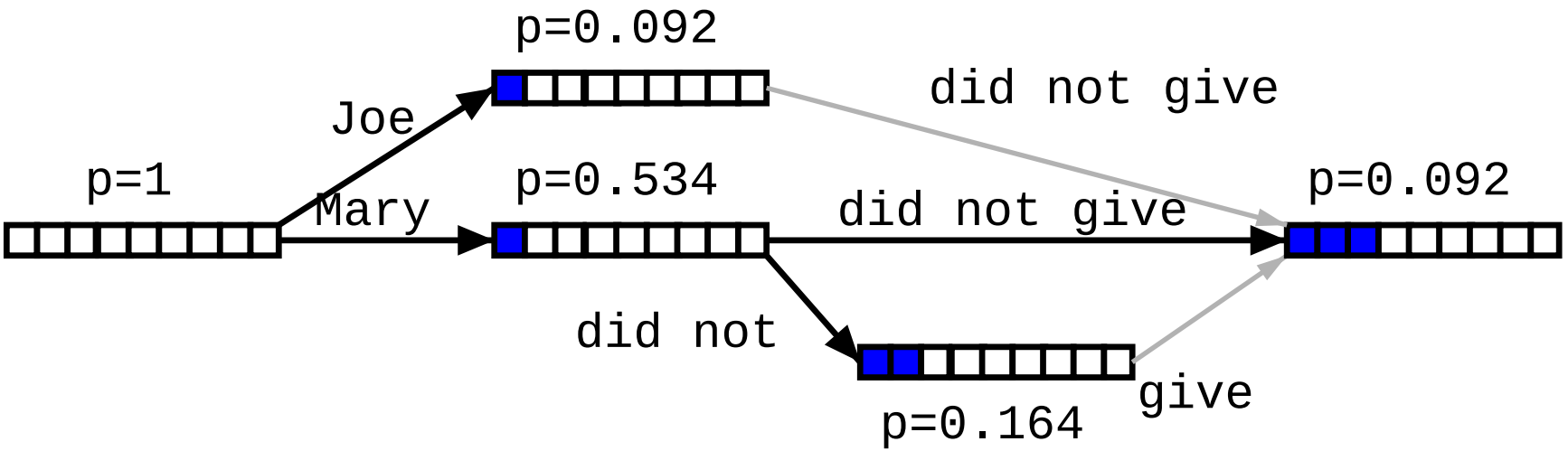
  - drop weaker hypothesis
  - keep pointer from worse path

Hypothesis Recombination



- Recombined hypotheses do not have to match completely
- No matter what is added, weaker path can be dropped, if:
  - last two English words match (matters for language model)
  - foreign word coverage vectors match (effects future path)

Hypothesis Recombination



- Recombined hypotheses do not have to match completely
- No matter what is added, weaker path can be dropped, if:
  - last two English words match (matters for language model)
  - foreign word coverage vectors match (effects future path)

$\Rightarrow$ Combine paths

Pruning

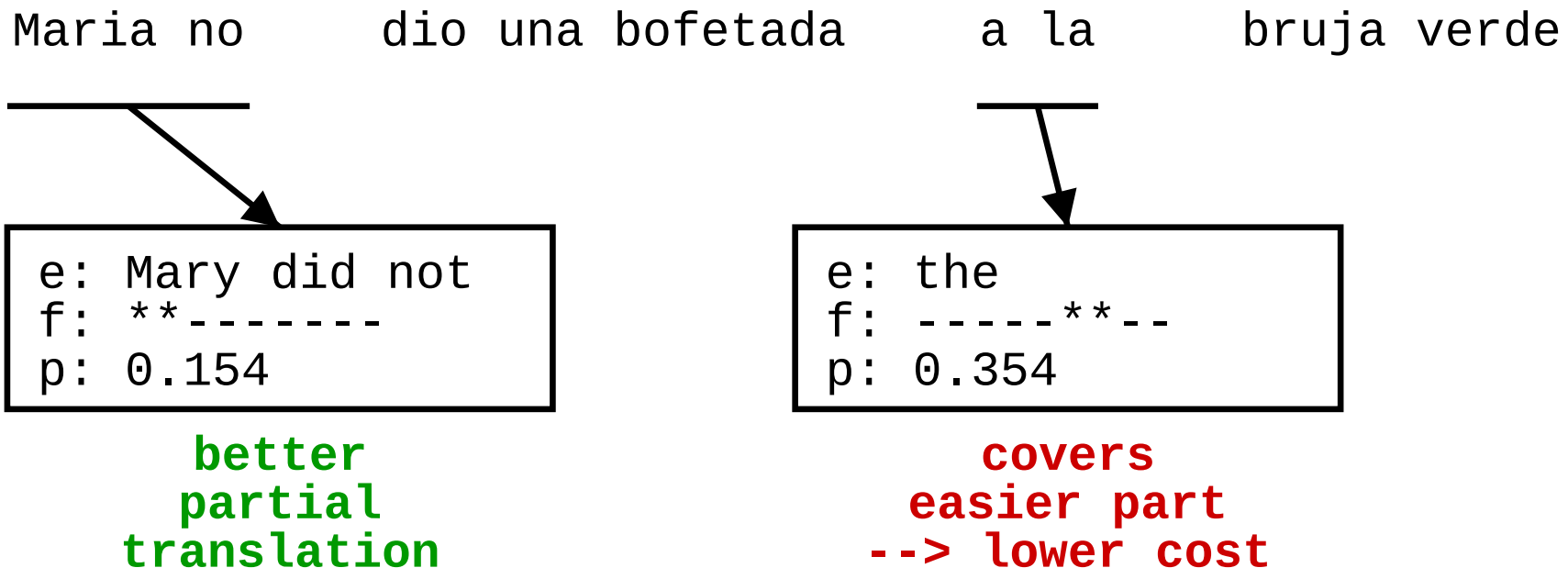
- Hypothesis recombination is not sufficient

$\Rightarrow$ Heuristically discard weak hypotheses

- Organize Hypothesis in stacks, e.g. by
  - same foreign words covered
  - same number of foreign words covered (Pharaoh does this)
  - same number of English words produced
- Compare hypotheses in stacks, discard bad ones
  - histogram pruning: keep top n hypotheses in each stack (e.g., n=100)
  - threshold pruning: keep hypotheses that are at most α times the cost of best hypothesis in stack (e.g., α = 0.001)

Comparing Hypotheses

- Comparing hypotheses with same number of foreign words covered



- Hypothesis that covers *easy* part of sentence is preferred

$\Rightarrow$ Need to consider future cost

Future Cost Estimation

- Estimate cost to translate remaining part of input
- Step 1: find cheapest translation options
  - find cheapest translation option for each input span
  - compute translation model cost
  - estimate language model cost (no prior context)
  - ignore reordering model cost
- Step 2: compute cheapest cost
  - for each contiguous span:
  - find cheapest sequence of translation options
- Precompute and lookup
  - precompute future cost for each contiguous span
  - future cost for any coverage vector:
    sum of cost of each contiguous span of uncovered words

  $\rightarrow$ no expensive computation during run time

Outline

- Phrase-Based Statistical MT
- Beam Search Decoding
- **Experiments**
- Advanced Features

Experiments

- Decoder has to be evaluated in terms of search errors
  - translation errors not due to search errors are a challenge to the translation model
  - do not rely on search errors for good translation quality!
- Experimental setup
  - German to English
  - Europarl training corpus (30 million words)
  - 1500 sentence test corpus (avg. length 28.9 words)
  - 3 Ghz Linux machine, needs 512 MB RAM
  - Focus: illustrate trade-off speed / search errors
- Not measuring true search error
  - it is not tractable to find truly best translation

  $\rightarrow$ relative to best translation found with high beam and different settings

Threshold Pruning

Threshold	0.0001	0.001	0.01	0.05	0.08
Time per Sentence	149 sec	119 sec	70 sec	27 sec	18 sec
Search Errors	-	+0%	+0%	+0%	+0%

Threshold	0.1	0.15	0.2	0.3	
Time per Sentence	15 sec	13 sec	10 sec	7 sec	
Search Errors	+1%	+3%	+6%	+12%	

- Low ratio of search errors for threshold $\alpha \leq 0.1$
- Results depend on weights for models

Histogram Pruning

Beam Size	1000	200	100	50	20	10	5
Time	15s	15s	14s	10s	9s	9s	7s
Search Errors	+1%	+1%	+2%	+4%	+8%	+20%	+35 %

- Low ratio of search errors for beam size $n \geq 200$

Translation Table Entries per Input Phrase

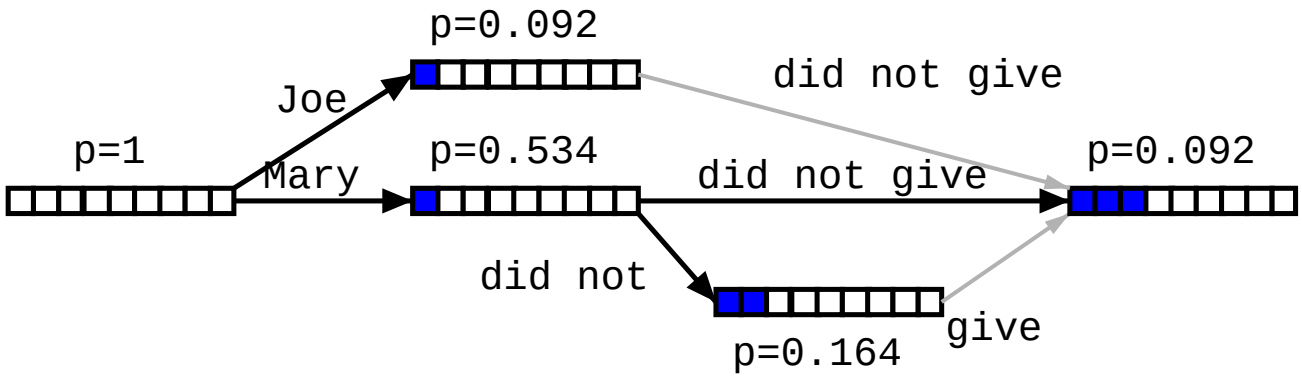
T-Table Limit	1000	500	200	100	50	20	10	5
Time	15.0s	7.6s	3.8s	1.9s	0.9s	0.4s	0.2s	0.1s
Search Errors	+1%	+1%	+1%	+1%	+1%	+2%	+7%	+18%

- Low ratio of search errors for limit of ≥ 50 entries in the translation table for each source language phrase
- About 1 second per sentence (30 words per second)
- Your mileage may vary

Outline

- Phrase-Based Statistical MT
- Beam Search Decoding
- Experiments
- **Advanced Features**

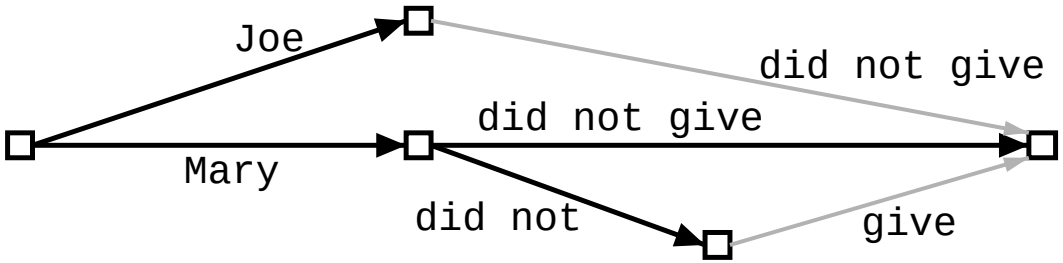
Word Lattice Generation



- Search graph can be easily converted into a word lattice
  - can be further mined for n-best lists

→ enables reranking approaches

→ enables discriminative training



XML Interface

```
Er erzielte <NUMBER english='17.55'>17,55</NUMBER>
Punkte .
```

- Add additional translation options
  - number translation
  - noun phrase translation [Koehn, 2003]
  - name translation
- Additional options
  - provide multiple translations
  - provide probability distribution along with translations
  - allow bypassing of provided translations

Thank You!

- Questions?