6 Tree-Based SMT

- Traditional statistical models operate on sequences of words
- Many translation problems can be best explained by pointing to syntax
  - reordering, e.g., verb movement in German–English translation
  - long distance agreement (e.g., subject-verb) in output

⇒ Translation models based on tree representation of language
  - significant ongoing research
  - state-of-the art for some language pairs

6.1 Synchronous Phrase Structure Grammar

- English rule

$$\text{NP} \rightarrow \text{DET JJ NN}$$

- French rule

$$\text{NP} \rightarrow \text{DET NN JJ}$$

- Synchronous rule (indices indicate alignment):

$$\text{NP} \rightarrow \text{DET}_1\ \text{NN}_2\ \text{JJ}_3 \mid \text{DET}_1\ \text{JJ}_3\ \text{NN}_2$$

Synchronous Grammar Rules

- Nonterminal rules
$$\text{NP} \rightarrow \text{DET}_1\ \text{NN}_2\ \text{JJ}_3 \mid \text{DET}_1\ \text{JJ}_3\ \text{NN}_2$$
- Terminal rules
$$\text{N} \rightarrow \text{maison} \mid \text{house}$$
$$\text{NP} \rightarrow \text{la maison bleue} \mid \text{the blue house}$$
- Mixed rules
$$\text{NP} \rightarrow \text{la maison JJ}_1 \mid \text{the JJ}_1\ \text{house}$$

Synchronous Grammar-Based Translation Model

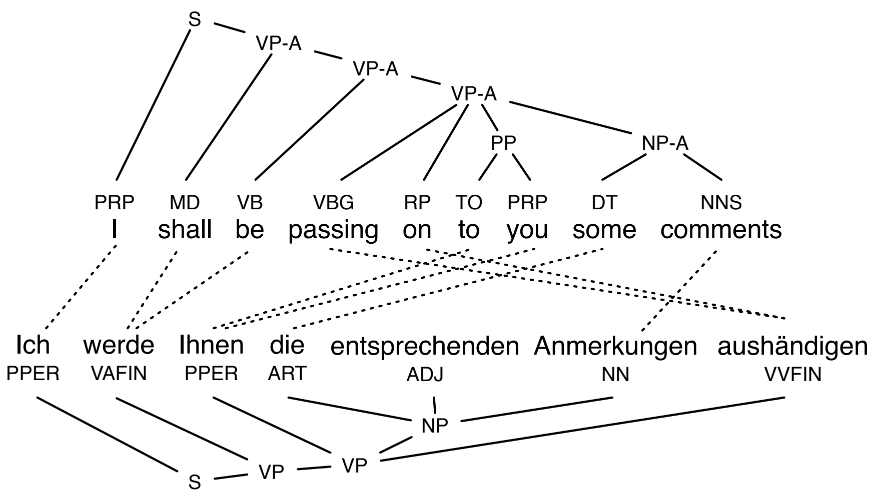
- Translation by parsing
  - synchronous grammar has to parse entire input sentence
  - output tree is generated at the same time
  - process is broken up into a number of rule applications
- Translation probability

$$\textsc{score}(\textsc{tree}, \textsc{e}, \textsc{f}) = \prod_i \textsc{rule}_i$$

- Many ways to assign probabilities to rules

6.2 Synchronous Tree-Substitution Grammars

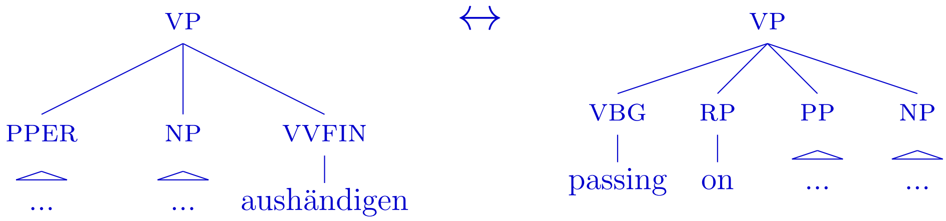
Aligned Tree Pair



Phrase structure grammar trees with word alignment
(German–English sentence pair.)

Reordering Rule

- Subtree alignment



- Synchronous grammar rule

VP → PPER$_1$ NP$_2$ aushändigen | passing on PP$_1$ NP$_2$

Another Rule

- Subtree alignment

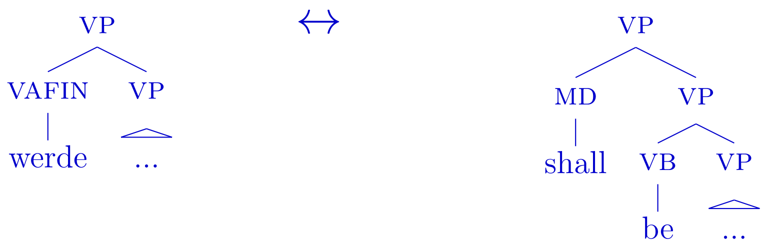
- Synchronous grammar rule (stripping out English internal structure)

PRO/PP → Ihnen | to you

- Rule with internal structure

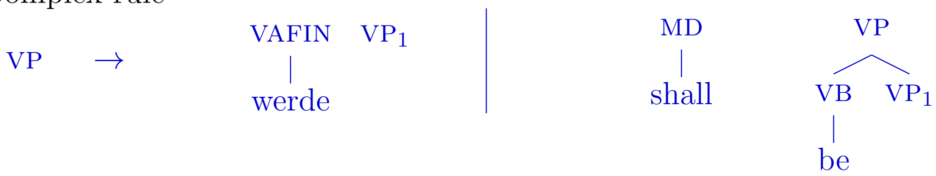
Another Rule

- Translation of German werde to English shall be



- Translation rule needs to include mapping of VP

⇒ Complex rule



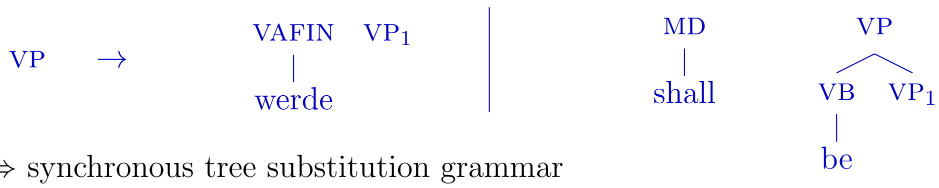
Internal Structure

- Stripping out internal structure

$$\text{VP} \rightarrow \text{werde VP}_1 \mid \text{shall be VP}_1$$

⇒ synchronous context free grammar

- Maintaining internal structure

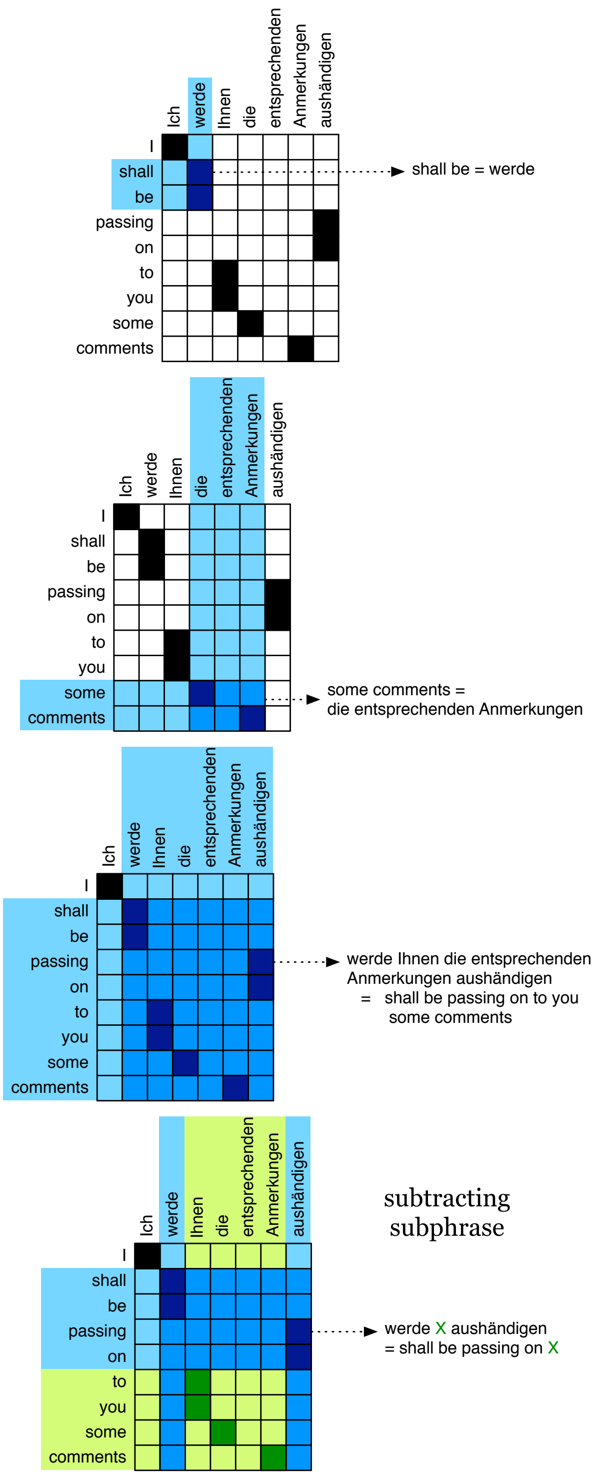


⇒ synchronous tree substitution grammar

6.3 Learning Synchronous Grammars

- Extracting rules from a word-aligned parallel corpus
- First: Hierarchical phrase-based model
  - only one non-terminal symbol X
  - no linguistic syntax, just a formally syntactic model
- Then: Synchronous phrase structure model
  - non-terminals for words and phrases: NP, VP, PP, ADJ, ...
  - corpus must also be parsed with syntactic parser

Extracting Phrase Translation Rules



Formal Definition

- Recall: consistent phrase pairs

$$
\begin{aligned}
(\bar{e}, \bar{f}) \text{ consistent with } A \Leftrightarrow \\
&\forall e_i \in \bar{e} : (e_i, f_j) \in A \rightarrow f_j \in \bar{f} \\
&\text{AND } \forall f_j \in \bar{f} : (e_i, f_j) \in A \rightarrow e_i \in \bar{e} \\
&\text{AND } \exists e_i \in \bar{e}, f_j \in \bar{f} : (e_i, f_j) \in A
\end{aligned}
$$

- Let P be the set of all extracted phrase pairs $(\bar{e}, \bar{f})$

- Extend recursively:

$$
\begin{aligned}
&\text{if } (\bar{e}, \bar{f}) \in P \text{ AND } (\bar{e}_{\text{SUB}}, \bar{f}_{\text{SUB}}) \in P \\
&\quad \text{AND } \bar{e} = \bar{e}_{\text{PRE}} + \bar{e}_{\text{SUB}} + \bar{e}_{\text{POST}} \\
&\quad \text{AND } \bar{f} = \bar{f}_{\text{PRE}} + \bar{f}_{\text{SUB}} + \bar{f}_{\text{POST}} \\
&\quad \text{AND } \bar{e} \neq \bar{e}_{\text{SUB}} \text{ AND } \bar{f} \neq \bar{f}_{\text{SUB}} \\
&\text{add } (e_{\text{PRE}} + \text{X} + e_{\text{POST}}, f_{\text{PRE}} + \text{X} + f_{\text{POST}}) \text{ to } P
\end{aligned}
$$

(note: any of e_{PRE}, e_{POST}, f_{PRE}, or f_{POST} may be empty)

- Set of hierarchical phrase pairs is the closure under this extension mechanism

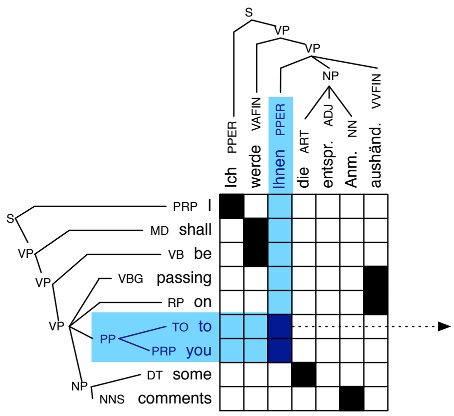
Comments

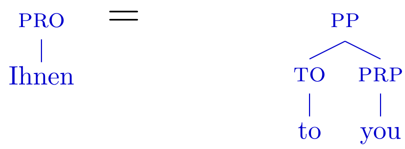
- Removal of multiple sub-phrases leads to rules with multiple non-terminals, such as:

$$\text{Y} \rightarrow \text{X}_1 \ \text{X}_2 \ | \ \text{X}_2 \ \textit{of} \ \text{X}_1$$

- Typical restrictions to limit complexity [Chiang, 2005]
  - at most 2 nonterminal symbols
  - at least 1 but at most 5 words per language
  - span at most 15 words (counting gaps)

Learning Syntactic Translation Rules

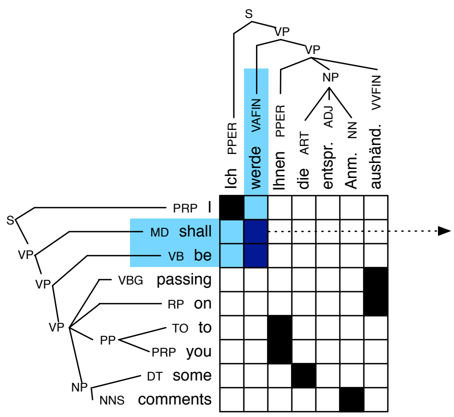




Constraints on Syntactic Rules

- Same word alignment constraints as hierarchical models
- Hierarchical: rule can cover any span
  ⇔ syntactic rules must cover constituents in the tree
- Hierarchical: gaps may cover any span
  ⇔ gaps must cover constituents in the tree
- Much less rules are extracted (all things being equal)

Impossible Rules



English span not a constituent
no rule extracted

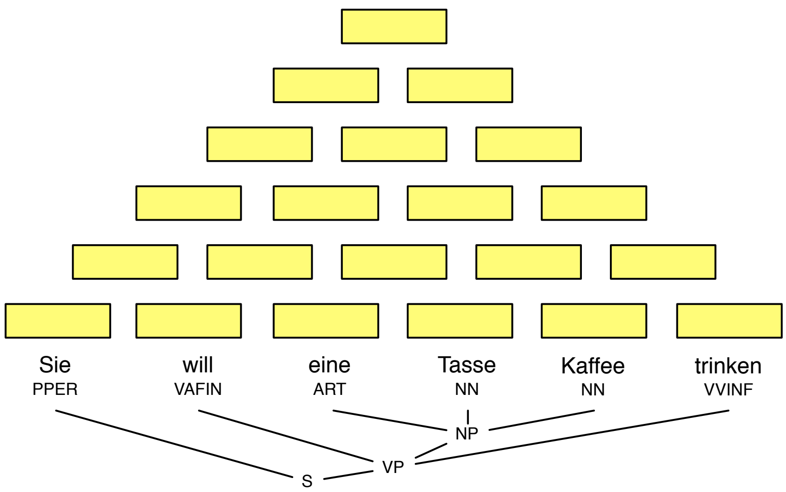
6.4 Scoring Translation Rules

- Extract all rules from corpus
- Score based on counts
  - joint rule probability: $p(\text{LHS}, \text{RHS}_f, \text{RHS}_e)$
  - rule application probability: $p(\text{RHS}_f, \text{RHS}_e|\text{LHS})$
  - direct translation probability: $p(\text{RHS}_e|\text{RHS}_f, \text{LHS})$
  - noisy channel translation probability: $p(\text{RHS}_f|\text{RHS}_e, \text{LHS})$
  - lexical translation probability: $\prod_{e_i \in \text{RHS}_e} p(e_i|\text{RHS}_f, a)$

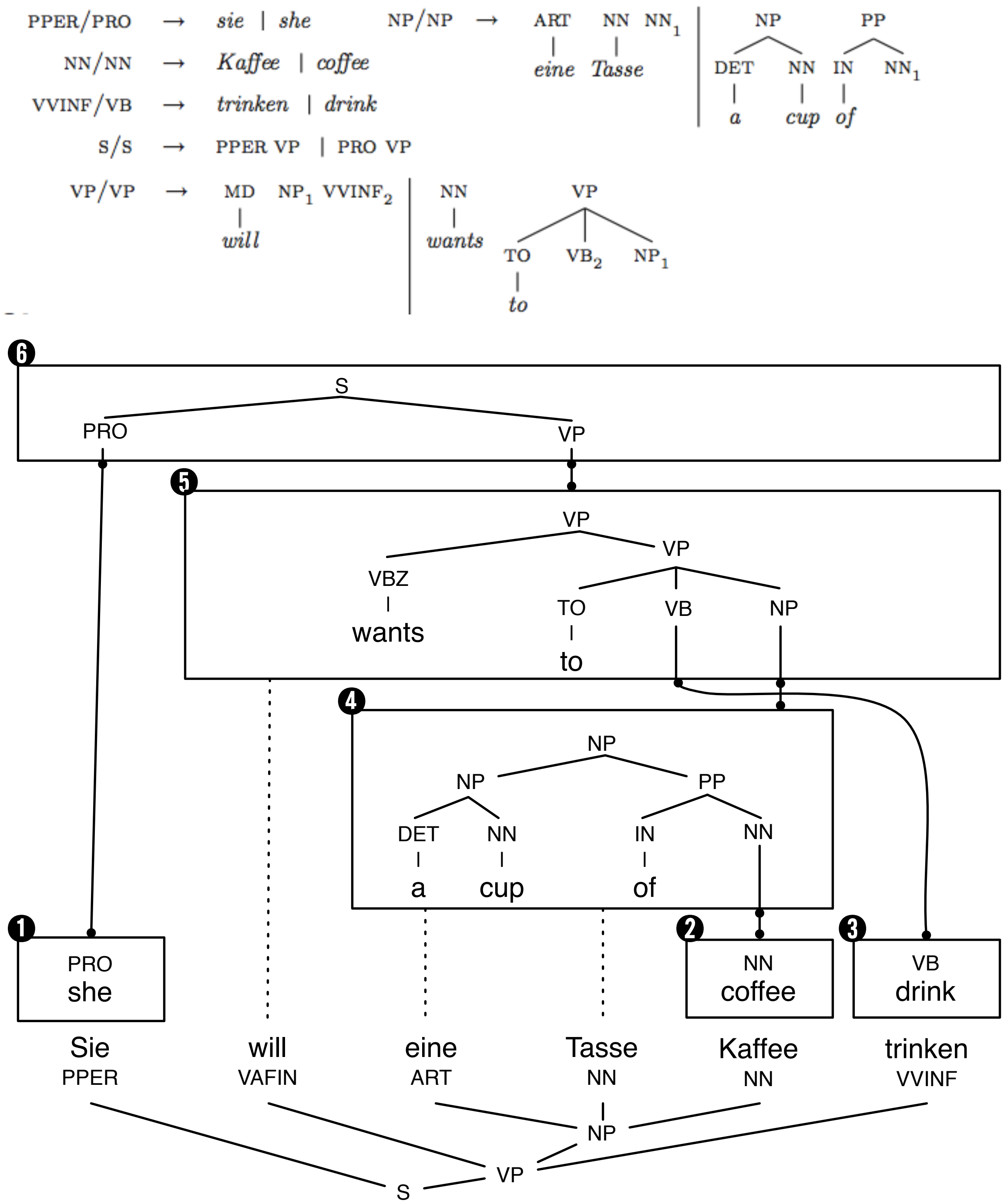
6.5 Syntactic Decoding

Inspired by monolingual syntactic chart parsing:

During decoding of the source sentence,
a chart with translations for the $O(n^2)$ spans has to be filled

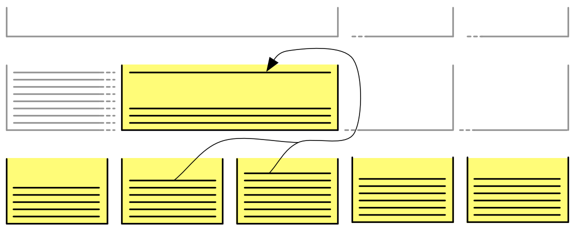


Syntax Decoding



Bottom-Up Decoding

- For each span, a stack of (partial) translations is maintained
- Bottom-up: a higher stack is filled, once underlying stacks are complete



Naive Algorithm

Input: Foreign sentence $\mathbf{f} = f_1, ...f_{l_f}$, with syntax tree
Output: English translation $\mathbf{e}$

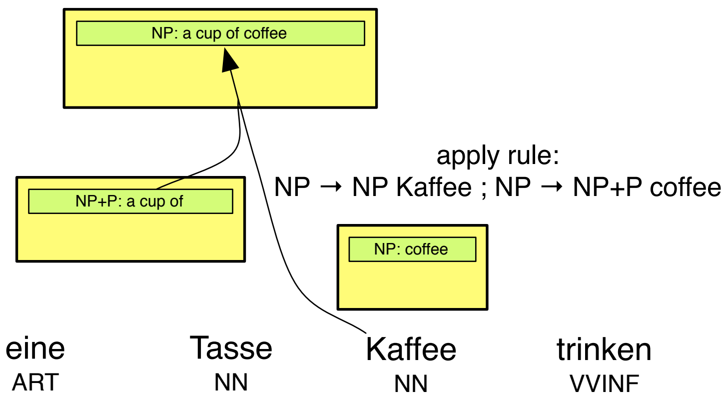
for all spans [start,end] (bottom up) **do**
  for all sequences s of hypotheses and words in span [start,end] **do**
    for all rules r **do**
      if rule r applies to chart sequence s **then**
        create new hypothesis c
        add hypothesis c to chart
      end if
    end for
  end for
end for
return English translation $\mathbf{e}$ from best hypothesis in span $[0,l_f]$

Chart Organization

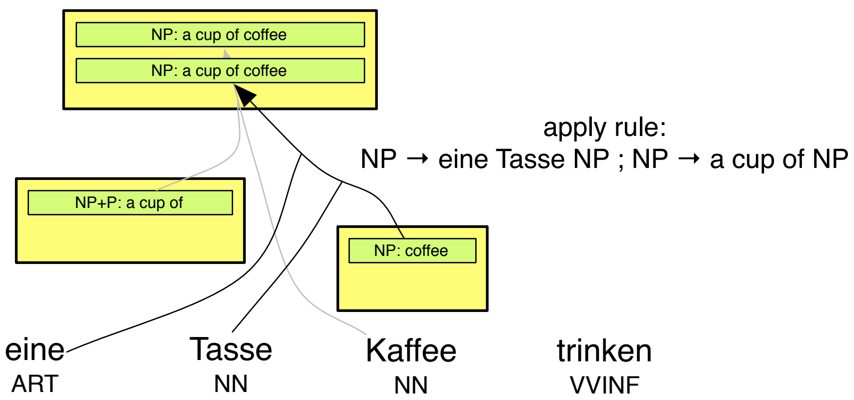
- Chart consists of cells that cover contiguous spans over the input sentence
- Each cell contains a set of hypotheses
- Hypothesis = translation of span with target-side constituent

Dynamic Programming

Applying rule creates new hypothesis



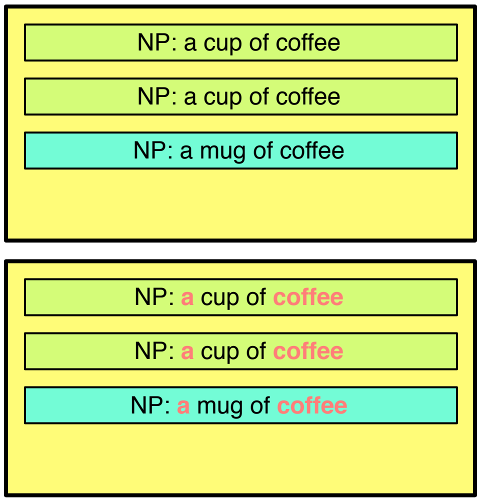
Another hypothesis



Both hypotheses are indistiguishable in future search
→ can be recombined

Recombinable States

Recombinable?

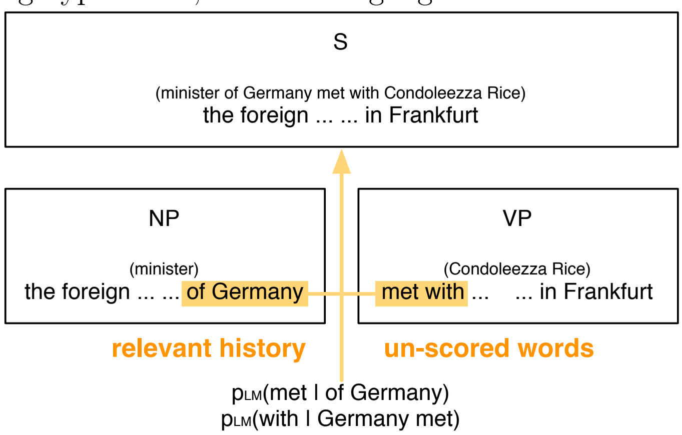


Yes, iff max. 2-gram language model is used

Hypotheses have to match in

- span of input words covered
- output constituent label
- first n–1 output words
- last n–1 output words

When merging hypotheses, internal language model contexts are absorbed



Stack Pruning

- Number of hypotheses in each chart cell explodes

⇒ need to discard bad hypotheses
e.g., keep 100 best only

Naive Algorithm: Blow-ups

- Many subspan sequences

for all sequences s of hypotheses and words in span [start,end]

- Many rules

for all rules r

- Checking if a rule applies not trivial

rule r applies to chart sequence s

⇒ Unworkable

Solution

- Prefix tree data structure for rules
- Dotted rules
- Cube pruning