Chart-Based Decoding

Kenneth Heafield

University of Edinburgh

6 September, 2012

Most slides courtesy of Philipp Koehn
Overview of Syntactic Decoding

Input Sentence

Decoding

SCFG Parsing

Search

Output Sentence
Overview of Syntactic Decoding

- Parallel Corpus
- Translation Model
- Input Sentence
- SCFG Parsing
- Search
- Decoding
- Output Sentence
- Language Model
- Monolingual Corpus
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.

```
Sie
PPER
will
VAFIN
eine
ART
Tasse
NN
Kaffee
NN
trinken
VVINF
NP
VP
S
```
Syntax Decoding

German input sentence with tree

Sie
PPER
will
VAFIN
eine
ART
Tasse
NN
Kaffee
NN
trinken
VVINF

S
VP
NP
Syntax Decoding

Purely lexical rule: filling a span with a translation (a constituent)
Syntax Decoding

Purely lexical rule: filling a span with a translation (a constituent)
Syntax Decoding

Purely lexical rule: filling a span with a translation (a constituent)
Complex rule: matching underlying constituent spans, and covering words
Sie will eine Tasse Kaffee trinken.
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
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.
Applying rule creates new hypothesis

apply rule:
NP → NP Kaffee ; NP → NP+P coffee
Another hypothesis

Both hypotheses are indistinguishable in future search
→ can be recombined
Recombinable States

Recombinable?

- NP: a cup of coffee
- NP: a cup of coffee
- NP: a mug of coffee

Yes, if max. 2-gram language model is used.
Recombinable States

Recombinable?

NP: a cup of coffee
NP: a cup of coffee
NP: a mug of coffee

Yes, if max. 2-gram language model is used
Recombinability

Hypotheses have to match in

- span of input words covered
- output constituent label
- first \( n-1 \) output words

not properly scored, since they lack context

- last \( n-1 \) output words

still affect scoring of subsequently added words,
just like in phrase-based decoding

\((n \text{ is the order of the n-gram language model})\)
Language Model Contexts

When merging hypotheses, internal language model contexts are absorbed.

S

(minister of Germany met with Condoleezza Rice)
the foreign ... ... in Frankfurt

NP

(minister)
the foreign ... ... of Germany

VP

(Condoleezza Rice)
met with ... ... in Frankfurt

relevant history

un-scored words

$\text{p}_{LM}(\text{met I of Germany})$

$\text{p}_{LM}(\text{with I Germany met})$
Stack Pruning

- Number of hypotheses in each chart cell explodes
  ⇒ need to discard bad hypotheses
  e.g., keep 100 best only
- Different stacks for different output constituent labels?
- Cost estimates
  - translation model cost known
  - language model cost for internal words known
    → estimates for initial words
  - outside cost estimate?
    (how useful will be a NP covering input words 3–5 later on?)
Naive Algorithm: Blow-ups

- Many subspan sequences
  for all sequences \( s \) of hypotheses and words in span \([\text{start}, \text{end}]\)
- Many rules
  for all rules \( r \)
- Checking if a rule applies not trivial
  rule \( r \) applies to chart sequence \( s \)

\(\Rightarrow\) Unworkable
Solution

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

- First concern: do they apply to span?
  → have to match available hypotheses and input words
- Example rule
  \[
  \text{NP} \rightarrow x_1 \text{ des } x_2 \mid \text{NP}_1 \text{ of the NN}_2
  \]
- Check for applicability
  - is there an initial sub-span that with a hypothesis with constituent label \text{NP}?
  - is it followed by a sub-span over the word \text{des}?
  - is it followed by a final sub-span with a hypothesis with label \text{NN}?
- Sequence of relevant information
  \[
  \text{NP} \bullet \text{ des} \bullet \text{NN} \bullet \text{NP}_1 \text{ of the NN}_2
  \]
Trying to cover a span of six words with given rule

NP • des • NN → NP: NP of the NN

das Haus des Architekten Frank Gehry
Rule Applicability Check

First: check for hypotheses with output constituent label NP

NP • des • NN → NP: NP of the NN

das  Haus  des  Architekten  Frank  Gehry
Rule Applicability Check

Found NP hypothesis in cell, matched first symbol of rule

NP • des • NN → NP: NP of the NN

das Haus des Architekten Frank Gehry
Matched word *des*, matched second symbol of rule

**NP • des • NN → NP: NP of the NN**

das Haus des Architekten Frank Gehry
Rule Applicability Check

Found a **NN** hypothesis in cell, matched last symbol of rule

NP • des • NN → NP: NP of the NN

das Haus des Architekten Frank Gehry
Rule Applicability Check

Matched entire rule → apply to create a NP hypothesis

NP • des • NN → NP: NP of the NN

das Haus des Architekten Frank Gehry
Rule Applicability Check

Look up output words to create new hypothesis
(note: there may be many matching underlying NP and NN hypotheses)

NP • des • NN → NP: NP of the NN

NP: the house of the architect Frank Gehry

NP: the house

NN: architect Frank Gehry

das Haus des Architekten Frank Gehry
Checking Rules vs. Finding Rules

- What we showed:
  - given a rule
  - check if and how it can be applied
- But there are too many rules (millions) to check them all
- Instead:
  - given the underlying chart cells and input words
  - find which rules apply
Highlighted Rules

\[
\begin{align*}
\text{NP} & \rightarrow \text{NP}_1 \ \text{DET}_2 \ \text{NN}_3 \ | \ \text{NP}_1 \ \text{IN}_2 \ \text{NN}_3 \\
\text{NP} & \rightarrow \text{NP}_1 \ | \ \text{NP}_1 \\
\text{NP} & \rightarrow \text{NP}_1 \ \text{des} \ \text{NN}_2 \ | \ \text{NP}_1 \ \text{of the} \ \text{NN}_2 \\
\text{NP} & \rightarrow \text{NP}_1 \ \text{des} \ \text{NN}_2 \ | \ \text{NP}_2 \ \text{NP}_1 \\
\text{NP} & \rightarrow \ \text{DET}_1 \ \text{NN}_2 \ | \ \text{DET}_1 \ \text{NN}_2 \\
\text{NP} & \rightarrow \ \text{das} \ \text{Haus} \ | \ \text{the house}
\end{align*}
\]
Dotted Rules: Key Insight

- If we can apply a rule like
  \[ p \rightarrow A \; B \; C \; | \; x \]
  to a span

- Then we could have applied a rule like
  \[ q \rightarrow A \; B \; | \; y \]
  to a sub-span with the same starting word

⇒ We can re-use rule lookup by storing \( A \; B \bullet \) (dotted rule)
Finding Applicable Rules in Prefix Tree

das | Haus | des | Architekten | Frank | Gehry
Covering the First Cell

das Haus des Architekten Frank Gehry
Looking up Rules in the Prefix Tree

das Haus des Architekten Frank Gehry
Taking Note of the Dotted Rule

das Haus des Architekten Frank Gehry
Checking if Dotted Rule has Translations

das ❶ DET: the

Haus des Architekten Frank Gehry
Applying the Translation Rules

Das Haus des Architekten Frank Gehry

Das ➟ Haus ➟ des ➟ Architekten ➟ Frank ➟ Gehry
Looking up Constituent Label in Prefix Tree

det 1

Das ①

Det ②

Das

Das Haus des Architekten Frank Gehry

Det: that
Det: the

Das ①

Das Haus des Architekten Frank Gehry

Det: that
Det: the
Add to Span’s List of Dotted Rules

das ❶
DET ❷
das ❶
das Haus des Architekten Frank Gehry
DET : the
DET : that
Moving on to the Next Cell

das

Haus des Architekten Frank Gehry

DET: that
DET: the
DET: das
Looking up Rules in the Prefix Tree

das ❶
DET ❷
Haus ❸

Das Haus des Architekten Frank Gehry

Das Haus des Architekten Frank Gehry
Taking Note of the Dotted Rule

das ①
DET ②
Haus ③

das ①
das ①
das Haus des Architekten Frank Gehry

DET ②
DET: the
DET: that

DET ②
das ①
das Haus des Architekten Frank Gehry

DET ②
DET: the
DET: that

das ①
Haus ③
des ②
Architekten ③
Frank ④
Gehry ⑤
Checking if Dotted Rule has Translations

das ❶
DET ❷
Haus ❸
NN: house
NP: house

das ❶
das Haus des Architekten Frank Gehry
DET : the
DET : that
house ❸
Applying the Translation Rules

Das Haus des Architekten Frank Gehry

Det: that
Det: the
Det
Das
Haus
Des
Architekten
Frank
Gehry

NN: house
NP: house
Looking up Constituent Label in Prefix Tree

- DET ❷
- das ❶
- Haus ❸
- NN ❹
- NP ❺
- DET: that
- DET: the
- DET ❷
- das ❶
- house ❸
- NP: house
- NN: house
- des
- Architekten
- Frank
- Gehry
Add to Span’s List of Dotted Rules

das ❶
DET ❷
Haus ❸
NN ❹
NP ❺

DET ❷
das ❶
das Haus des Architekten Frank Gehry
DET : the
DET : that
NN ❹  NP ❺ 
house ❸
NN : house
NP : house
More of the Same
Moving on to the Next Cell

das das Haus des Architekten Frank Gehry
Covering a Longer Span

Cannot consume multiple words at once
All rules are extensions of existing dotted rules
Here: only extensions of span over \textit{das} possible
Extensions of Span over das

das ❶ — NN, NP, Haus?
DET ❷ — NN, NP, Haus?
Haus ❸ — NN
NP ❺ —

DET ❷
das ❶
das Haus des Architekten Frank Gehry
DET : the
DET : that
NN ❹  NP ❺ 
house ❸
NP : house
IN : of
NN: architect
NP : architect
NNP : Frank
NP : architect
NNP : Gehry
Looking up Rules in the Prefix Tree

das ①  Haus ⑥
DET ②
Haus ⑧
NN ⑨

Das Haus des Architekten Frank Gehry

Das ① das Haus des Architekten Frank Gehry

Das ① das Haus des Architekten des Frank Gehry

Das ① das Haus des Architekten des Frank Gehry
Taking Note of the Dotted Rule

Das Haus des Architekten Frank Gehry.
Checking if Dotted Rules have Translations

- das
- Haus
- des
- Architekten
- Frank
- Gehry

- DET: that
- DET: the
- DET: das

- NP: house
- NN: house
- des
- Architekten

- NNP: Frank
- NNP: Gehry

- NP: the house
- NP: the NN
- NP: DET house
- NP: DET NN
Applying the Translation Rules
Looking up Constituent Label in Prefix Tree

```
NP: that house
NP: the house
DET NN 9
DET Haus 6
das NN 7
das Haus 8
```

```
DET: that
DET: the
DET 2
das 1
```

```
IN: of
```

```
NP: architect
```

```
NN: architect
```

```
NNP: Frank
```

```
NNP: Gehry
```

```
房子
```

```
Frank
```

```
Gehry
```

```
NP: DET house
NP: DET NN
```

```
NP: the house
NP: the NN
```

```
das Haus des Architekten Frank Gehry
```

```
das Haus
```

```
das
das Haus
```

```
the house
the
```

```
the
```

```
the
```

```
that house
```

```
that
```

```
house
```

```
house
```

```
architect
```

```
architect
```

```
Frank
```

```
Gehry
```
Add to Span’s List of Dotted Rules

NP: that house
NP: the house

DET NN 9  NP 5
DET Haus 6
das NN 7
das Haus 6

DET: that
DET: the

NP: house
NN: house

IN: of
DET: the

NP: architect
NN: architect

NN 4  NP 5
Architekten

NNP: Frank

NNP: Gehry
Even Larger Spans

Extend lists of dotted rules with cell constituent labels

span’s dotted rule list (with same start)
plus neighboring
span’s constituent labels of hypotheses (with same end)
Reflections

- Complexity $O(rn^3)$ with sentence length $n$ and size of dotted rule list $r$
  - may introduce maximum size for spans that do not start at beginning
  - may limit size of dotted rule list (very arbitrary)
- Does the list of dotted rules explode?
- Yes, if there are many rules with neighboring target-side non-terminals
  - such rules apply in many places
  - rules with words are much more restricted
Difficult Rules

- Some rules may apply in too many ways
- Neighboring input non-terminals

  \[ VP \rightarrow \text{gibt } X_1 \ X_2 \mid \text{gives } NP_2 \ \text{to } NP_1 \]

  - non-terminals may match many different pairs of spans
  - especially a problem for hierarchical models (no constituent label restrictions)
  - may be okay for syntax-models

- Three neighboring input non-terminals

  \[ VP \rightarrow \text{trifft } X_1 \ X_2 \ X_3 \ \text{heute} \mid \text{meets } NP_1 \ \text{today } PP_2 \ PP_3 \]

  - will get out of hand even for syntax models
Where are we now?

- We know which rules apply
- We know where they apply (each non-terminal tied to a span)
- But there are still many choices
  - many possible translations
  - each non-terminal may match multiple hypotheses
  - number choices exponential with number of non-terminals
Rules with One Non-Terminal

Found applicable rules $PP \rightarrow \text{des } X \mid \ldots \text{ NP } \ldots$

- Non-terminal will be filled any of $h$ underlying matching hypotheses
- Choice of $t$ lexical translations

$\Rightarrow$ Complexity $O(ht)$

(note: we may not group rules by target constituent label, so a rule $NP \rightarrow \text{des } X \mid \text{the } NP$ would also be considered here as well)
Rules with Two Non-Terminals

Found applicable rule \( NP \rightarrow X_1 \; \text{des} \; X_2 \mid NP_1 \; \ldots \; NP_2 \)

- Two non-terminal will be filled any of \( h \) underlying matching hypotheses each
- Choice of \( t \) lexical translations

\[ \Rightarrow \] Complexity \( O(h^2 t) \) — a three-dimensional "cube" of choices

(note: rules may also reorder differently)
Filling a Constituent

- Hyp: seen, Score: -3.8
- Hyp: saw, Score: -4.0
- Hyp: view, Score: -4.0

- Hyp: man, Score: -3.6
- Hyp: the man, Score: -4.3
- Hyp: some men, Score: -6.3
## Beam Search

<table>
<thead>
<tr>
<th></th>
<th>man</th>
<th>the man</th>
<th>some men</th>
</tr>
</thead>
<tbody>
<tr>
<td>seen</td>
<td>-3.6</td>
<td>-4.3</td>
<td>-6.3</td>
</tr>
<tr>
<td>saw</td>
<td>-3.8</td>
<td>-4.0</td>
<td>-3.6</td>
</tr>
<tr>
<td>view</td>
<td>-4.0</td>
<td>-4.0</td>
<td>-3.6</td>
</tr>
</tbody>
</table>

**Hypothesis Sum**

- seen man: -8.8
- seen the man: -7.6
- seen some men: -9.5
- saw man: -8.3
- saw the man: -6.9
- saw some men: -8.5
- view man: -8.5
- view the man: -8.9
- view some men: -10.8
Cube Pruning [Chiang, 2007]

- man -3.6
- the man -4.3
- some men -6.3
- seen -3.8
- Queue
- saw -4.0
- view -4.0

Queue

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>seen man</td>
<td>-3.8-3.6=-7.4</td>
</tr>
<tr>
<td>Word</td>
<td>Value</td>
</tr>
<tr>
<td>----------</td>
<td>-------</td>
</tr>
<tr>
<td>man</td>
<td>-3.6</td>
</tr>
<tr>
<td>the man</td>
<td>-4.3</td>
</tr>
<tr>
<td>some men</td>
<td>-6.3</td>
</tr>
<tr>
<td>seen</td>
<td>-3.8</td>
</tr>
<tr>
<td>seen man</td>
<td>-8.8</td>
</tr>
<tr>
<td>saw</td>
<td>-4.0</td>
</tr>
<tr>
<td>view</td>
<td>-4.0</td>
</tr>
<tr>
<td>saw man</td>
<td>-4.0</td>
</tr>
<tr>
<td>seen the man</td>
<td>-3.8</td>
</tr>
</tbody>
</table>
Cube Pruning [Chiang, 2007]

<table>
<thead>
<tr>
<th>verb</th>
<th>score</th>
<th>phrase</th>
<th>Queue</th>
</tr>
</thead>
<tbody>
<tr>
<td>seen</td>
<td>-3.8</td>
<td>seen man</td>
<td>Queue</td>
</tr>
<tr>
<td>saw</td>
<td>-4.0</td>
<td>saw man</td>
<td>Queue</td>
</tr>
<tr>
<td>view</td>
<td>-4.0</td>
<td>Queue</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>view man</td>
<td>-4.0-3.6=-7.6</td>
</tr>
<tr>
<td>seen the man</td>
<td>-3.8-4.3=-8.1</td>
</tr>
<tr>
<td>saw the man</td>
<td>-4.0-4.3=-8.3</td>
</tr>
</tbody>
</table>
Cube Pruning versus Beam Search

Same  Bottom-up with fixed-size beams
Different  Beam filling algorithm
Queue of Cubes

- Several groups of rules will apply to a given span
- Each of them will have a cube
- We can create a queue of cubes
  - Always pop off the most promising hypothesis, regardless of cube

- May have separate queues for different target constituent labels
Bottom-Up Chart Decoding Algorithm

1: for all spans (bottom up) do
2:   extend dotted rules
3:   for all dotted rules do
4:     find group of applicable rules
5:     create a cube for it
6:     create first hypothesis in cube
7:     place cube in queue
8:   end for
9: for specified number of pops do
10: pop off best hypothesis of any cube in queue
11: add it to the chart cell
12: create its neighbors
13: end for
14: extend dotted rules over constituent labels
15: end for
Two-Stage Decoding

- **First stage:** decoding without a language model (-LM decoding)
  - may be done exhaustively
  - eliminate dead ends
  - optionally prune out low scoring hypotheses

- **Second stage:** add language model
  - limited to packed chart obtained in first stage

- **Note:** essentially, we do two-stage decoding for each span at a time
Coarse-to-Fine

- Decode with increasingly complex model

Examples
  - reduced language model [Zhang and Gildea, 2008]
  - reduced set of non-terminals [DeNero et al., 2009]
  - language model on clustered word classes [Petrov et al., 2008]
Outside Cost Estimation

- Which spans should be more emphasized in search?
- Initial decoding stage can provide outside cost estimates

- Use min/max language model costs to obtain admissible heuristic
  (or at least something that will guide search better)
Open Questions

- Where does the best translation fall out the beam?
- Are particular types of rules too quickly discarded?
- Are there systemic problems with cube pruning?
Summary

- Synchronous context free grammars
- Extracting rules from a syntactically parsed parallel corpus
- Bottom-up decoding
- Chart organization: dynamic programming, stacks, pruning
- Prefix tree for rules
- Dotted rules
- Cube pruning