Empirical Methods in Natural Language Processing
Lecture 18
Machine translation (V): Syntax-Based Models

Philipp Koehn & Adam Lopez
6 March 2008
Phrase-based SMT

- Already works pretty well.

- Are there any problems that we need to solve here?
Phrase-based SMT

- **Computational**: computing all possible reorderings is NP-complete.

- **Linguistic**: language is not finite-state.
## Syntax-based SMT

- What's going on here? A whole lot of things...

- Chiang (2005) makes a distinction between *formally* syntax-based and *linguistically* syntax-based.

<table>
<thead>
<tr>
<th></th>
<th><em>formally</em> syntax-based</th>
</tr>
</thead>
<tbody>
<tr>
<td>phrase-based</td>
<td>hierarchical phrase-based [Chiang 2005]</td>
</tr>
</tbody>
</table>
Linguistic Advantages of Syntax-Based Translation

- *Generalized* reordering for syntactic reasons
  - e.g., move German object to end of sentence

- Better explanation for *function words*
  - e.g., prepositions, determiners

- Conditioning to *syntactically related words*
  - translation of verb may depend on subject or object

- Use of *syntactic language models*
  - ensuring grammatical output
**Clause Level Restructuring [Collins et al.]**

<table>
<thead>
<tr>
<th></th>
<th>formally syntax-based</th>
</tr>
</thead>
<tbody>
<tr>
<td>phrase-based</td>
<td>hierarchical phrase-based [Chiang 2005]</td>
</tr>
<tr>
<td>linguistically</td>
<td>reordering + phrase-based [Collins et al. 2005]</td>
</tr>
<tr>
<td>syntax-based</td>
<td>syntax-based SCFG</td>
</tr>
<tr>
<td></td>
<td>[Yamada &amp; Knight 2002]</td>
</tr>
</tbody>
</table>
Clause Level Restructuring [Collins et al.]

- Why clause structure?
  - languages differ vastly in their clause structure (English: SVO, Arabic: VSO, German: fairly free order; a lot details differ: position of adverbs, sub clauses, etc.)
  - large-scale restructuring is a problem for phrase models

- Restructuring
  - reordering of constituents (main focus)
  - add/drop/change of function words

- Details see [Collins, Kucerova and Koehn, ACL 2005]
• Syntax tree from German parser
  – statistical parser by Amit Dubay, trained on TIGER treebank
Reordering When Translating

- **Reordering** when translating into English
  - tree is *flattened*
  - clause level constituents line up
Clause level reordering is a *well defined task*

- label German constituents with their *English order*
- done this for 300 sentences, two annotators, high agreement
Systematic Reordering German → English

- Many types of reorderings are systematic
  - move verb group together
  - subject - verb - object
  - move negation in front of verb

⇒ Write rules by hand
  - apply rules to test and training data
  - train standard phrase-based SMT system

<table>
<thead>
<tr>
<th>System</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline system</td>
<td>25.2%</td>
</tr>
<tr>
<td>with manual rules</td>
<td>26.8%</td>
</tr>
</tbody>
</table>
Improved Translations

• we must also this criticism should be taken seriously.
   → we must also take this criticism seriously.

• i am with him that it is necessary, the institutional balance by means of a political revaluation of both the commission and the council to maintain.
   → i agree with him in this, that it is necessary to maintain the institutional balance by means of a political revaluation of both the commission and the council.

• thirdly, we believe that the principle of differentiation of negotiations note.
   → thirdly, we maintain the principle of differentiation of negotiations.

• perhaps it would be a constructive dialog between the government and opposition parties, social representative a positive impetus in the right direction.
   → perhaps a constructive dialog between government and opposition parties and social representative could give a positive impetus in the right direction.
Other Linguistically Syntax-Based Approaches

- **Reranking** phrase-based SMT output with syntactic features
  - create n-best list with phrase-based system
  - POS tag and parse candidate translations
  - rerank with syntactic features
  - see [Koehn, 2003] and JHU Workshop [Och et al., 2003]

- Incorporate syntax into decoder [Tillman and Ney, 2003]
  - Add finite-state control structure to allow long-distance movement of verbs in German-English translation.
Formal Advantages of Syntax-Based Translation

• Foundation in *well-understood* models from formal language theory (theoretical computer science).
  – Maybe they have some use after all

• *Computational complexity* is (in principle) just as much as we need to model linguistic phenomena, and no more.
  – Polynomial even with full reordering.
  – Caveat: no easy trick to speed it up as with phrase-based models.

• *Apply advances* made algorithms for statistical parsing.
  – Earley, CKY, etc.
# Synchronous Context-Free Grammars

<table>
<thead>
<tr>
<th>formally syntax-based</th>
<th>phrase-based</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>linguistically</strong></td>
<td>reordering + phrase-based</td>
</tr>
<tr>
<td>syntax-based</td>
<td>[Collins et al. 2005]</td>
</tr>
<tr>
<td></td>
<td>syntax-based SCFG</td>
</tr>
<tr>
<td></td>
<td>[Yamada &amp; Knight 2002]</td>
</tr>
</tbody>
</table>
Synchronous Context-Free Grammars

- Finite-state transducers model *regular* language
- Regular tree transducers model *context-free* language
- Various guises of SCFG
  - Syntax-directed Transduction (Lewis and Stearns 1968)
  - Inversion Transduction Grammar (Wu 1995-1998)
  - Head Transducers (Alshawi et al. 2000)
  - Multitext Grammar (Melamed 2003)
Inversion Transduction Grammars

• Generation of both English and foreign trees [Wu, 1997]

• Rules (binary and unary)
  - $A \rightarrow A_1A_2 \parallel A_1A_2$
  - $A \rightarrow A_1A_2 \parallel A_2A_1$
  - $A \rightarrow e \parallel f$
  - $A \rightarrow e \parallel *$
  - $A \rightarrow * \parallel f$

⇒ Common binary tree required

  - limits the complexity of reorderings – polynomial in length, exponential in arity
Syntax Trees

Mary did not slap the green witch

- English binary tree
Syntax Trees

Maria no daba una bofetada a la bruja verde

- Spanish binary tree
• Combined tree with reordering of Spanish
Chiang: Hierarchical Phrase-based Model

- **Chiang** [ACL, 2005] (best paper award!)
  - context free bi-grammar
  - *one non-terminal* symbol
  - right hand side of rule may include non-terminals and terminals

- *Competitive* with phrase-based models in 2005 DARPA/NIST evaluation
Types of Rules

- **Word** translation
  - $X \rightarrow \text{maison} \parallel \text{house}$

- **Phrasal** translation
  - $X \rightarrow \text{daba una bofetada} \mid \text{slap}$

- Mixed non-terminal / terminal – **hierarchial phrases**
  - $X \rightarrow X_1 \text{ bleue} \parallel \text{blue } X_1$
  - $X \rightarrow \text{ne } X_1 \text{ pas} \parallel \text{not } X_1$
  - $X \rightarrow X_1 \ X_2 \parallel X_2 \text{ of } X_1$

- **Technical rules**
  - $S \rightarrow S_1 \ X_2 \parallel S_1 \ X_2$
  - $S \rightarrow X_1 \parallel X_1$
Learning Hierarchical Rules

Maria no daba una botefada a la bruja verde Mary witch green the slap not did

X → X verde || green X
Learning Hierarchical Rules

<table>
<thead>
<tr>
<th>María no daba una botefada a la bruja verde</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Mary</th>
<th>did</th>
<th>not</th>
<th>slap</th>
<th>the</th>
<th>green</th>
<th>witch</th>
</tr>
</thead>
</table>

\[ X \rightarrow a \ la \ X \parallel \ the \ X \]
Details of Chiang’s Model

- Too many rules
  - filtering of rules necessary

- **Efficient** parse decoding possible
  - hypothesis stack for each span of foreign words
  - only *one non-terminal* → hypotheses comparable
  - *length limit* for spans that do not start at beginning
  - *m*-gram language model integration increases complexity by $O(n^{2m})$
Language is not Context-Free!

- Maybe it’s mildly context-sensitive?
  - Synchronous Tree-Adjoining Grammar [Shieber 1992, others]
  - Generalized Multitext Grammar [Melamed 2004]

- Various transducer formalisms – [Knight & Graehl 2005] for overview.
## Syntactic Language Model

<table>
<thead>
<tr>
<th>Syntactic Approach</th>
<th>Model Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>formally</em> syntax-based</td>
<td>phrase-based</td>
</tr>
<tr>
<td><em>linguistically</em> syntax-based</td>
<td>reordering + phrase-based (Collins et al. 2005)</td>
</tr>
<tr>
<td>hierarchical phrase-based (Chiang 2005)</td>
<td><em>syntax-based SCFG</em> (Yamada &amp; Knight 2002)</td>
</tr>
</tbody>
</table>
Syntactic Language Model

- **Good syntax tree** → good English
- Allows for **long distance constraints**

- Left translation preferred by syntactic LM
String to Tree Translation

- Use of English *syntax trees* [Yamada and Knight, 2001]
  - exploit *rich resources* on the English side
  - obtained with statistical parser [Collins, 1997]
  - *flattened tree* to allow more reorderings
  - works well with syntactic language model
Yamada and Knight [2001]

Kare ha ongaku wo kiku no ga daisuki desu

[from Yamada and Knight, 2001]
# Reordering Table

| Original Order | Reordering       | $p(\text{reorder}|\text{original})$ |
|----------------|------------------|-------------------------------------|
| PRP VB1 VB2    | PRP VB1 VB2      | 0.074                               |
| PRP VB1 VB2    | PRP VB2 VB1      | 0.723                               |
| PRP VB1 VB2    | VB1 PRP VB2      | 0.061                               |
| PRP VB1 VB2    | VB1 VB2 PRP      | 0.037                               |
| PRP VB1 VB2    | VB2 PRP VB1      | 0.083                               |
| PRP VB1 VB2    | VB2 VB1 PRP      | 0.021                               |
| VB TO VB       | VB TO            | 0.107                               |
| VB TO TO       | TO VB            | 0.893                               |
| TO NN TO       | TO NN            | 0.251                               |
| TO NN NN       | NN TO            | 0.749                               |
Decoding as Parsing

- Chart Parsing

```
PRP
he
```

- kare ha ongaku wo kiku no ga daisuki desu

- Pick Japanese *words*

- Translate into *tree stumps*
Decoding as Parsing

- Chart Parsing

- Pick Japanese words

- Translate into tree stumps
Decoding as Parsing

- Adding some *more entries*...
Decoding as Parsing

- Combine entries

kare ha ongaku wo kiku no ga daisuki desu
Decoding as Parsing

-he music

Listening
Decoding as Parsing

PRP
he

NN
music

TO
to

VB
listening

VB2

PP

PRP
he

NN
music

TO
to

VB
listening

VB1
adores

kare ha ongaku wo kiku no ga daisuki desu
Decoding as Parsing

- *Finished* when all foreign words covered
Yamada and Knight: Training

- **Parsing** of the English side
  - using Collins statistical parser

- **EM training**
  - translation model is used to map training sentence pairs
  - EM training finds low-perplexity model
  → *unity of training and decoding* as in IBM models
Is the Model Realistic?

- Do English trees match foreign strings?

- Crossings between French-English [Fox, 2002]
  - 0.29-6.27 per sentence, depending on how it is measured

- Can be reduced by
  - flattening tree, as done by [Yamada and Knight, 2001]
  - detecting phrasal translation
  - special treatment for small number of constructions

- Most coherence between dependency structures
Other Syntax-Based Approaches

• ISI: extending work of Yamada/Knight
  – more complex rules
  – performance approaching phrase-based

• Prague: Translation via dependency structures
  – parallel Czech–English dependency treebank
  – tecto-grammatical translation model [EACL 2003]

• U.Alberta/Microsoft: treelet translation
  – translating from English into foreign languages
  – using dependency parser in English
  – project dependency tree into foreign language for training
  – map parts of the dependency tree (“treelets”) into foreign languages
Syntax: Does it help?

• Getting there
  – for some languages competitive with best phrase-based systems

• Some evidence
  – work on reordering German
  – ISI: better for short sentences Chinese–English
  – automatically trained tree transfer systems promising

• Why not yet?
  – if real syntax, we need good parsers — are they good enough?
  – syntactic annotations add a level of complexity
    → difficult to handle, slow to train and decode
  – few researchers good at statistical modeling and syntactic theories