Lecture 13:
Syntactic Machine Translation

Slav Petrov – Google

Thanks to David Chiang and Jacob Devlin for many of today’s slides!

Syntax Bet

Based on slide from mt-class.org

Every time I hire a linguist my BLEU score goes up

Every time I fire a linguist my performance goes up

Simpler models will always win

Syntax will improve translation

Overview of Approaches

<table>
<thead>
<tr>
<th>Phrase Based</th>
<th>Hierarchical</th>
<th>Tree-to-String</th>
<th>String-to-Tree</th>
<th>Tree-to-Tree</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brown et al. ’93</td>
<td>Wu ’97 (ITG)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Koehn et al. ’03</td>
<td>Chiang ’05 (Hiero)</td>
<td>Huang et al ’06</td>
<td>Galley et al. ’04/’06 (ISI / GHKM )</td>
<td>Lavie et al ’08</td>
</tr>
</tbody>
</table>

Syntax Bet

Based on slide from mt-class.org

Every time I hire a linguist my BLEU score goes up

Every time I fire a linguist my performance goes up

Simpler models will always win

Syntax will improve translation

Word-Based Translation

Human translation:

Australia is one of the few countries that have diplomatic relations with North Korea.
Phrase-Based Translation

Australia is with North Korea is one of the few countries.

Machine Translation Preordering

- English has Subject-Verb-Object word order, while Japanese has Subject-Object-Verb order.
- Use hand-written or automatically learned rules to change word order prior to translation [Collins et al. '05].
- Dependency-based reordering for English-Japanese [Xu et al. '09].

Syntax-Based Translation

'S Australia' is 'is' "few" 'country' 'one of' 'country'.

Phrases of Phrases

'Hiero' [Chiang '05]
Australia is one of the few countries that have diplomatic relations with North Korea.

(X → with, X → have X_2 with X_1)

(X → North Korea)

(X → diplomatic relations)
Synchronous CFG

Rewrite linked nonterminals to generate pairs of trees

Grammar Extraction

Example Rules

<table>
<thead>
<tr>
<th>X → 的</th>
<th>X → 's</th>
</tr>
</thead>
<tbody>
<tr>
<td>X → X_1 的 X_2</td>
<td>X → the X_2 of X_1</td>
</tr>
<tr>
<td>X → X_1 的 X_2</td>
<td>X → the X_2 that X_1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>X → 在</th>
<th>X → in</th>
</tr>
</thead>
<tbody>
<tr>
<td>X → 在 X_1 下</td>
<td>X → under X_1</td>
</tr>
<tr>
<td>X → 在 X_1 前</td>
<td>X → before X_1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>X → 今年 X_1</th>
<th>X → X_1 this year</th>
</tr>
</thead>
<tbody>
<tr>
<td>X → X_1 之一</td>
<td>X → one of X_1</td>
</tr>
<tr>
<td>X → X_1 总统</td>
<td>X → president X_1</td>
</tr>
</tbody>
</table>
Glue Rules

- Plus “glue” rules:
- \( S \rightarrow S_1X_2, \ S \rightarrow S_1X_2 \)
- \( S \rightarrow X_1, \ S \rightarrow X_1 \)
- Acts as fallback like phrase-based systems

\[
\begin{array}{c}
S \\
S_1 \\
X_2 \\
X_1
\end{array}
\]

Experimental setup

- Baseline system: ISI ATS (Och et al)
- Training: 30M words for Hiero, 150M words for baseline
- Language model: 2.8G words, 3-grams
- Max-BLEU training: MT Eval 2002 dry-run
- Test: MT Eval 2003

Results

- MT Eval 2003 data, case-insensitive BLEU-4:

<table>
<thead>
<tr>
<th></th>
<th>Chinese</th>
</tr>
</thead>
<tbody>
<tr>
<td>‘Phrase-Based’</td>
<td>31.5</td>
</tr>
<tr>
<td>Hiero</td>
<td>34.3</td>
</tr>
</tbody>
</table>

Syntax-Based MT

- synchronous context-free grammars (SCFGs)
- context-free grammar in two dimensions
- generating pairs of strings/trees simultaneously
- co-indexed nonterminal further rewritten as a unit

\[
\begin{align*}
\text{VP} & \rightarrow \text{PP}^{(1)} \text{VP}^{(2)} \\
\text{VP} & \rightarrow \text{juxing le huitan, held a meeting} \\
\text{PP} & \rightarrow \text{yu Shalong, with Sharon}
\end{align*}
\]

\[
\begin{align*}
\text{VP} & \rightarrow \text{juxing le huitan} \\
\text{PP} & \rightarrow \text{yu Shalong, held a meeting} \\
\text{PP} & \rightarrow \text{yu Shalong, held a meeting with Sharon}
\end{align*}
\]
Translation by Parsing

- translation with SCFGs => monolingual parsing
- parse the source input with the source projection
- build the corresponding target sub-strings in parallel

Translating with Tree Transducers

Learning MT Grammars

- syntax-directed, English to Chinese (Huang, Knight, Joshi, 2006)
- first parse input, and then recursively transfer

Syntactic Constraints

- synchronous tree-substitution grammars (STSG) (Galley et al., 2004; Eisner, 2003)
Extracting syntactic rules

Extract rules (Galley et al. '04, '06)

Rules can...

- capture phrasal translation
- reorder parts of the tree
- traverse the tree without reordering
- insert (and delete) words

Bad alignments make bad rules

Sometimes they’re really bad

This isn’t very good, but let’s look at a worse example...

One bad link makes a totally unusable rule!
What's wrong with extraction constraints?

20M words Chinese-English training data

<table>
<thead>
<tr>
<th></th>
<th>rules (millions)</th>
<th>size (gzip MB)</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>minimal</td>
<td>1.9</td>
<td>62</td>
<td>35</td>
</tr>
<tr>
<td>5-nodes</td>
<td>135</td>
<td>1227</td>
<td>40</td>
</tr>
</tbody>
</table>

Results

- MT Eval 2003 data, case-insensitive BLEU-4:

<table>
<thead>
<tr>
<th></th>
<th>Chinese</th>
</tr>
</thead>
<tbody>
<tr>
<td>‘Phrase-Based’</td>
<td>31.5</td>
</tr>
<tr>
<td>Hiero</td>
<td>34.3</td>
</tr>
<tr>
<td>+Syntax</td>
<td>36.2</td>
</tr>
</tbody>
</table>

CKY-style Bottom-Up Parsing

For each length:

For each span [i,j]:

- Apply all grammar rules to [i,j]

Binary rule: \( X \rightarrow Y Z \)

Split points: \( i < k < j \)

Operations: \( O(j - i) \)

Time scales with: Grammar constant

Many untransformed lexical rules can be applied in linear time
CKY-style Bottom-Up Parsing

For each span length:
For each span \([i,j]\):
Apply all grammar rules to \([i,j]\)

\[ S \rightarrow \text{No se VP NP PP} \]

Problem: Applying adjacent non-terminals is slow

Inversion Transduction Grammars

Generation
Monotonic Combination
Inverted Combination

[Wu '97]

Binarizing Translation Rules

We must select a binary derivation for each non-terminal sequence

Original:
\[ S \rightarrow \text{VB NP NP PP} \]

Binarization options:
\[ S \rightarrow \text{VB~NP~NP~PP} \]
\[ S \rightarrow \text{VB NP~NP~PP} \]
\[ S \rightarrow \text{VB~NP NP~PP} \]

Objective function:
The minimum number of grammar symbols, such that all non-terminal sequences have binary derivations

[Resumption of the session]

LM Integration (for ITG)
**LM Integration (for ITG)**

![Diagram of LM Integration](image)

**Compact Forests**

![Diagram of Compact Forests](image)

**Neural Machine Translation**

![Diagram of Neural Machine Translation](image)

**Neural Network Language Models (NNLMs)**

![Diagram of NNLMs](image)
Neural Net Joint Model (NNJM)

Source: werde ich das mit der bank morgen endlich klaeren koennen
Target: i will finally be able to clarify that with the bank tomorrow

\[ P(\text{tomorrow | with the bank; mit der bank morgen endlich klaeren koennen}) \]

Neural Net Lexical Translation Model (NNLTM)

Source: werde ich das mit der bank morgen endlich klaeren koennen
Target: i will finally be able to clarify that with the bank tomorrow

\[ P(\text{be\_able\_to | morgen endlich klaeren koennen }<s> </s> </s>) \]

Pre-Computation

- The "pre-computation trick": The matrix-vector product between the word embedding and a section of the first hidden layer can be computed offline

The matrix-vector product between the word embedding and a section of the first hidden layer can be computed offline.
Neural Translation Models

- Alternative approach: “Pure” neural network models

  Sutskever et al. 2014
  “Sequence to Sequence Learning with Neural Networks”
  Encode source into fixed length vector, use it as initial recurrent state for target decoder model

  Bahdanau et al. 2014
  “Neural Machine Translation by Jointly Learning to Align and Translate”
  Recurrent model responsible for producing target words and picking the next source word to give “attention” to

Conclusions

- Syntax-based systems can outperform many phrase-based systems
- Explicit syntactic information is optional but helpful
- Induction of synchronous grammars can be done successfully, but there is much room for improvement
- This is just one possible view, see Syntax Bet

Class Summary: Machine Translation

- You now know how to build this!

Class Summary: Information Extraction

- You now know how to build this!
Hypothesis

- Understanding arises from machine learning of relationships implicit in web content and use
  - Some expert annotation may be needed to start
  - Most evidence is not explicitly annotated: text “in the wild”
  - Aggregate information from multiple unstructured sources into a structured “knowledge base”
  - Exploit user interactions and implicit user feedback

Final Project

- Talk to me if you are unsure
- Presentations:
  - Everybody should talk:
    - individual projects: 4+1 min
    - 2 people projects: 6+2 min
    - 3 people projects: 9+2 min
    - 4 people projects: 11+2 min
  - Explain:
    - (1) the problem
    - (2) a baseline
    - (3) your approach
    (OK not to have final results)
  - I will bring pizza

Paper (6/8/10/12 pages):
- Introduction
- Related Work
- Your Approach
- Experiments
- Conclusions

Stay in touch!

slav@petrovi.de