Lecture 12: Syntactic Machine Translation

Slav Petrov – Google

Thanks to David Chiang, Jacob Devlin, Thang Luong, Kyunghyun Cho, Christopher Manning for many of today’s slides!

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>Yamada &amp; Knight ‘01</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>

Hot Topic: Neural MT!

Syntax Bet

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

Australia is one of the few countries that have diplomatic relations with North Korea.

Human translation:

澳大利亚是与北韩有邦交的少数国家之一。
**Phrase-Based Translation**

- Australia is with North Korea in diplomatic relations.
- Australia is one of the few countries with diplomatic relations with North Korea.

**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**

- 'Australia' is a few countries with diplomatic relations with North Korea.

**Phrases of Phrases**

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

Rewrite linked nonterminals to generate pairs of trees

Grammar Extraction

Example Rules

澳大利亚与北韩有邦交的少数国家之一

Australia is one of the few countries that have diplomatic relations with North Korea

(与北韩有邦交, have diplomatic relations with North Korea)

邦交, diplomatic relations

(邦交, diplomatic relations)

北韩, North Korea

(X → 与 X1 有 X2, X → have X2 with X1)

X → 的, X → ’s

X → X1 的 X2, X → the X2 of X1

X → X1 的 X2, X → the X2 that X1

X → 在, X → in

X → 在 X1 下, X → under X1

X → 在 X1 前, X → before X1

X → 今年 X1, X → X1 this year

X → X1 之一, X → one of X1

X → X1 总统, X → president X1
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

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:
  - ‘Phrase-Based’: 31.5
  - Hiero: 34.3

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

- \( VP \rightarrow PP^{(1)} VP^{(2)} \)
- \( VP \rightarrow juxing le huitan, \) held a meeting
- \( PP \rightarrow yu Shalong, \) with Sharon

- \( yu Shalong \) juxing le huitan held a meeting with Sharon
Translation by Parsing

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

Syntactic Constraints

<table>
<thead>
<tr>
<th>VP</th>
<th>PP(1) VP(2)</th>
<th>held a talk with Sharon</th>
</tr>
</thead>
<tbody>
<tr>
<td>VP</td>
<td>juxing le huitan, held a meeting with Sharon</td>
<td></td>
</tr>
</tbody>
</table>

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)

Australia is one of the few countries that have diplomatic relations with North Korea.

(NP → the few countries that VP)

(?? → the few countries that)
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>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>
</tr>
<tr>
<td>5-nodes</td>
<td>135</td>
<td>1227</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Chinese</th>
<th>BLEU</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>

Results

CKY-style Bottom-Up Parsing

For each span 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 [i,j]:
- Apply all grammar rules to [i,j]

$S \rightarrow No\ se\ VP\ NP\ PP$

Problems: Applying adjacent non-terminals is slow

Binarizing Translation Rules

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

Original:
- $S \rightarrow VB\ NP\ NP\ PP$

Binarization options:
- $S \rightarrow VB\ NP\ NP\ PP$
- $S \rightarrow VB\ NP\ NP\ PP$

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

LM Integration (for ITG)

Monotonic Combination
- Generation
- Inverted Combination

[Wu '97]

Inversion Transduction Grammars
LM Integration (for ITG)

Compact Forests

Neural Machine Translation

Neural Network Language Models (NNLMs)
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

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

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

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

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

![Diagram of Neural Net Lexical Translation Model](image)

**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

![Pre-Computed Hidden Layer Diagram](image)

![Compute Equation](image)
seq2seq Machine Translation

[Sutskever et al. 2014, Bahdanau et al. 2014]

Google Neural MT System

Progress in MT

Benefits

- End-to-end training: all parameters are simultaneously optimized to minimize a loss function on the network’s output
- Distributed representations share strength: Better exploitation of word and phrase similarities
- Better exploitation of context: NMT can use a much bigger context – both source and partial target text – to translate more accurately
Challenges

- Vocabulary Size
- ‘Memory’
- Language Complexity
- Data

Encoder-Decoder-Attention

8 Layer Deep

Bi-Directional First-Layer Encoder
Residual Connections

- 'Plain' Stacking
- +Residual Connections

Vocabulary Size

- Limit vocabulary size, use word piece models, copy mechanisms, etc.
- Softmax computation is expensive.

Attention

- seq2seq: last hidden state needs to 'remember entire sentence'
- attention: allows model to 'look back at the source'

Multi-Task Learning

- German (translation) → English (translation)
- German (lang model)
- German (parsing)
Multi-Lingual Models

Zero-Resource Translation?

Make Models Compact

Conclusions

- Neural Networks have taken the field by storm
- Everybody is now focusing on Neural MT systems and the results are astonishing

- 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 room for improvement
- Many possible views, see Syntax Bet
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: 12+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