Context-Aware Rule-Selection for Statistical Machine Translation

LIU Qun
CNGL @ Dublin City University &
ICT @ Chinese Academy of Sciences
Linguistic Knowledge in SMT

• Used:
  – Morphology: segmentation
  – Syntax: Constituent, Dependency

So limited!
Linguistic Knowledge in SMT

• Unused:
  – Morphology: Inflection, Compound word
  – Syntax: Movement
  – Semantic: Preference, Semantic Role
  – Ontology
  – Discourse: Co-reference, Coherence, Topic Structure, Anaphora
  – Pragmatic: Sentiment, Intention, Situation…

So Much!
New SMT Paradigm?

- Word-based Translation
- Phrase-based Translation
- Syntax-based Translation
- ......
- Semantic-based Translation ???
- Discourse-based Translation ???
- ......
Problem

- Some of the translation problem may never be resolved without using certain kind of linguistic knowledge.
Example

• 10天前玛丽丢了一辆自行车。
• Mary lost her bicycle 10 days ago.

• 刚才警察来通知车找到了。
• Just now the police come to tell her that her bicycle was found.

• Need ontology: “自行车” is-kind-of “车”
• Need coreference resolution for insert “her”
Problem

• Some linguistic theory only have effect on very specific language phenomenon

• Building a new SMT paradigm on a certain linguistic knowledge (x-based translation)
  – high cost
  – usually lead to decrease of BLEU scores
Our Solution: CARS

Context-Aware Rule Selection

- Compatible to current log-linear SMT framework
- Easy to integration various linguistic knowledge to current SMT system
- Working locally rather than globally
- Effect!
Example: mouse

| mouse | 老鼠 | 鼠标 |

- The **mouse** was found to have **escaped** two days later.
- 两天后发现这只**老鼠****逃跑**了。

- The **mouse** was found **damaged** two days later.
- 两天后发现这只**鼠标****坏**了。
The *mouse* was found to have escaped two days later.

The *mouse* was found damaged two days later.

*鼠标*两天后逃脱。

两天后发现损坏的*鼠标*。
Example: $X_1$ 的 $X_2$

<table>
<thead>
<tr>
<th></th>
<th>$X_1$ ’s $X_2$</th>
<th>$X_1$ $X_2$</th>
<th>$X_2$ of $X_1$</th>
<th>$X_2$ of $X_1$ ’s</th>
</tr>
</thead>
</table>

狐狸的尾巴
- the fox’s tail

木头的桌子
- wood table

地球的环境
- the environment of the earth

小王的一个朋友
- a friend of Xiao Wang’s
Google Translate

- 狐狸的尾巴
- Fox tail

- 地球的环境
- Earth’s environment

- 小王的一个朋友
- Wang a friend

- 木头的桌子
- Wood table
• **Language Expression:**

an expression used in statistical translation model as a description of a piece of language surface form or certain language structure.
Notions

• **Translation Rule:**
  a mapping from a *source language expression* to a *target language expression*

• **Translation Rule Selection:**
  to select the best *target language expression* for a given *source language expression*, by giving a score to each candidate translation rule
Language Expression

- Word
- Phrase
- CFG Rule
- CFG Tree
- Dependency Rule
- Dependency Treelet
- String of Terminals and Non-T.

......
## Translation Rules

<table>
<thead>
<tr>
<th>Translation Models</th>
<th>Translation Rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBM Model 1-5</td>
<td>Word $\rightarrow$ Word (word translation table)</td>
</tr>
<tr>
<td>Phrase-based Model</td>
<td>Phrase $\rightarrow$ Phrase (phrase table)</td>
</tr>
<tr>
<td>Hierarchical Phrase-based Model</td>
<td>CFG Rule $\rightarrow$ CFG Rule</td>
</tr>
<tr>
<td>String-to-Dependency (Shen 08)</td>
<td>CFG Rule $\rightarrow$ CFG rule with Dep.</td>
</tr>
<tr>
<td>Tree-to-String Model</td>
<td>CFG Tree $\rightarrow$ String</td>
</tr>
<tr>
<td>String-to-Tree Model</td>
<td>String $\rightarrow$ CFG Tree</td>
</tr>
<tr>
<td>Dependency Model (Quirk 05)</td>
<td>Dep. Treelet $\rightarrow$ Dep. Treelet</td>
</tr>
<tr>
<td>Dependency Model (Xiong 06)</td>
<td>Dep. Treelet $\rightarrow$ String</td>
</tr>
<tr>
<td>Dependency Model (Xie 11)</td>
<td>Dep. Rule $\rightarrow$ String</td>
</tr>
</tbody>
</table>
# Rule Selection

<table>
<thead>
<tr>
<th>mouse</th>
<th>老鼠</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>鼠标</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>X₁ 的 X₂</th>
<th>X₁ X₂</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>X₁ ’s X₂</td>
</tr>
<tr>
<td></td>
<td>X₂ of X₁</td>
</tr>
<tr>
<td></td>
<td>X₂ of X₁ ’s</td>
</tr>
</tbody>
</table>
Rule Selection

Given $S$, select rule from:

$$
\begin{align*}
& r_1 : S \rightarrow T_1 \\
& r_2 : S \rightarrow T_2 \\
& \vdots \\
& r_n : S \rightarrow T_n
\end{align*}
$$
Rule Selection by Probability

\[ \hat{r} = \arg\max_{r_i} P(r_i|S) \]

where:
\[ \sum_{i} P(r_i|S) = \sum_{i} P(T_i|S) = 1 \]
Rule Selection by Probability

<table>
<thead>
<tr>
<th>mouse</th>
<th>老鼠</th>
<th>0.4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>鼠标</td>
<td>0.6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>X₁ 的 X₂</th>
<th>X₁ X₂</th>
<th>0.3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>X₁ ’s X₂</td>
<td>0.4</td>
</tr>
<tr>
<td></td>
<td>X₂ of X₁</td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td>X₂ of X₁ ’s</td>
<td>0.1</td>
</tr>
</tbody>
</table>
Problem

• All probabilities for rule selection are static values trained from the training corpus.

• No context information is able to be used for rule selection.

• Language model and reordering model only help a little for rule selection.
Motivation

Rule Selection by Dynamic Context Information
Context-Aware Rule Selection
—— CARS Model

\[ \text{Score}(r_i | C, S) \]

\[ r_i: S \rightarrow T_i: \text{Translation Rule} \]
\[ C: \text{Context} \]
\[ S: \text{Source Expression} \]

**Note**: CARS model is used as a feature of the log-linear model in SMT.
Probabilistic CARS Model

\[
\text{Score}(S, C) = P(r_i | C, S)
\]

where: \( \sum_i P(r_i | C, S) = 1 \)

Note: As a feature of log-linear model, CARS model is not necessary to be a probability.
Discriminative CARS Model

\[ P(r_i|S,C) = \frac{\exp(\sum_k \lambda_k h_k(r_i,C|S))}{\sum_r \exp(\sum_k \lambda_k h_k(r_j,C|S))} \]

\( h_k(r,C) \): Context Features

\( \lambda_k \): Weights of Context Features
Training CARS Model

• To training a CARS model, we need:
  – Count the number of the rules (as usual)
  – Reserve the context for each occurrence of the rule (new requirement)
Applicability of CARS Model

- CARS model may applicable only to part of the rules, for example:
  - only for lexicalized rules
  - only for un-lexicalized rules
  - only for verbs (SRL)
  - only for pronouns (Coreference)
  - only for to a single word (DE)
  - ......
CARS Utilization as a Feature

• An additional feature of CARS Utilization may be also necessary in log-linear model
  – To record the times of using CARS model in decoding
  – To balance between the rules using or not using CARS model
  – Not necessary if the CARS model is applicable to all rules
Outline

Introduction

Context-Aware Rule-Selection

CARS Application Examples

Conclusion and Future Work
CARS Application Examples

- CARS for Bracketing Transduction Grammar
- CARS for Hierarchical Phrase-based Model
- CARS for Tree-to-String Model
- CARS using Topic Model
- CARS for Agglutinative Language Translation

CARS for Bracketing Transduction Grammar
  - Deyi Xiong et al. COLING-ACL2006

CARS for Hierarchical Phrase-based Model
  - Zhongjun He et al. COLING2008
  - Qun Liu et al. EMNLP2008

CARS for Tree-to-String Model
  - Xinyan Xiao et al. ACL2012
  - unpublished
CARS Application Examples

- CARS for Bracketing Transduction Grammar
- CARS for Hierarchical Phrase-based Model
- CARS for Tree-to-String Model
- CARS using Topic Model
- CARS for Agglutinative Language Translation
**ITG: Inversion Transduction Grammar**

(Wu, Dekai 1995)

- Synchronized Grammar
- Binary Rules (CNF style)

<table>
<thead>
<tr>
<th>ITG rules</th>
<th>Source</th>
<th>Target</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A \rightarrow [B \ C ]$</td>
<td>$A \rightarrow BC$</td>
<td>$A \rightarrow BC$</td>
</tr>
<tr>
<td>$A \rightarrow &lt;B \ C &gt;$</td>
<td>$A \rightarrow BC$</td>
<td>$A \rightarrow CB$</td>
</tr>
<tr>
<td>$A \rightarrow x/y$</td>
<td>$A \rightarrow x$</td>
<td>$A \rightarrow y$</td>
</tr>
</tbody>
</table>
ITG Non-Terminal Rules

```

target

source

monotone  reverse
```

Diagram:

- Monotone:
  - Target node connected to source node.
  - No reverse direction.

- Reverse:
  - Target node connected to source node.
  - Reverse direction indicated by red line.

```
ITG Based Translation (1)
ITG Based Translation (2)
ITG Based Translation (3)

• Pros:
  – Recursive
  – Linguistic-style grammar
  – Limited search space

• Cons:
  – Need human annotated bi-lingual corpus for training
BTG: Bracketing Transduction Grammar

• BTG:
  A simplified ITG with only one non-terminal

• Only two non-terminal rules:
  \[ X \rightarrow \left[ X_1 \ X_2 \right] \] (monotone rule)
  \[ X \rightarrow \left< X_1 \ X_2 \right> \] (reverse rule)
Stochastic BTG

(Wu, Dekai 96)

• Static Rule Selection

• Only one parameter for non-terminal rules

\[ X \rightarrow [ X_1 \ X_2 ] : \ p(\text{monotone rule})=0.7 \]
\[ X \rightarrow \langle X_1 \ X_2 \rangle : \ p(\text{reverse rule})=0.3 \]

• Too week discriminability

• Our Approach: CARS
Our Method

- Given bilingual phrase $X_1$ and $X_2$
  
  $X_1 = \text{“with them ◇ 与他们”}$
  
  $X_2 = \text{“keep contact ◇ 保持联系”}$

- Calculate the probabilities using $X_1$ and $X_2$:

  $$P(\text{Monotone}, X_1, X_2) = 0.05$$
  
  $$P(\text{Reverse}, X_1, X_2) = 0.95$$
Maximum-Entropy BTG

• Modeling: Maximum Entropy

\[ \Omega = p_\theta(o \mid X_1, X_2) = \frac{\exp(\sum_i \theta_i h_i(o, X_1, X_2))}{\sum_o \exp(\sum_i \theta_i h_i(o', X_1, X_2))} \]

\[ h_i(o, X_1, X_2) = \begin{cases} 
  1 & \text{if } f(X_1, X_2) = \text{True}, \ o = 0 \\
  0 & \text{otherwise}
\end{cases} \]

0 \in \{\text{monotone, reverse}\}
We ONLY use monolingual or bilingual left boundary words as features.
Feature Templates

\[ X_1 = "with \text{ them}\Diamond 与他们" \]
\[ X_2 = "keep \text{ contact}\Diamond 保持联系" \]

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>C1=与</td>
<td></td>
</tr>
<tr>
<td>C2</td>
<td>C2=保持</td>
<td></td>
</tr>
<tr>
<td>E1</td>
<td>E1=with</td>
<td></td>
</tr>
<tr>
<td>E2</td>
<td>E2=keep</td>
<td></td>
</tr>
<tr>
<td>C1C2</td>
<td>C1=与 &amp; C2=保持</td>
<td></td>
</tr>
<tr>
<td>C1E1</td>
<td>C1=与 &amp; E1=with</td>
<td></td>
</tr>
<tr>
<td>C2E2</td>
<td>C2=保持 &amp; E2=keep</td>
<td></td>
</tr>
<tr>
<td>E1E2</td>
<td>E1=with &amp; E2=keep</td>
<td></td>
</tr>
</tbody>
</table>
Training Samples Extraction

Word Alignment

\[ f \]

\[ e \]

monotone

reverse
## Experiment Result

<table>
<thead>
<tr>
<th>Systems</th>
<th>NIST MT-05</th>
<th>IWSLT-04</th>
</tr>
</thead>
<tbody>
<tr>
<td>monotone</td>
<td>20.1 ± 0.8</td>
<td>37.8 ± 3.2</td>
</tr>
<tr>
<td>NONE</td>
<td>19.6 ± 0.8</td>
<td>36.3 ± 2.9</td>
</tr>
<tr>
<td>Distortion</td>
<td>20.9 ± 0.8</td>
<td>38.8 ± 3.0</td>
</tr>
<tr>
<td>Flat</td>
<td>20.5 ± 0.8</td>
<td>38.7 ± 2.8</td>
</tr>
<tr>
<td>Pharaoh</td>
<td>20.8 ± 0.8</td>
<td>38.9 ± 3.3</td>
</tr>
<tr>
<td>MaxEnt (lex)</td>
<td>22.0 ± 0.8</td>
<td>42.4 ± 3.3</td>
</tr>
<tr>
<td>MaxEnt (lex + col)</td>
<td><strong>22.2 ± 0.8</strong></td>
<td><strong>42.8 ± 3.3</strong></td>
</tr>
</tbody>
</table>
Summary

• We proposed MEBTG to compute the probability of two BTG non-terminal rules.
• Only boundary word features are used in MEBTG model.
• MEBTG model is very effective as a reordering model for phrase-based translation.
• A lot of citations and follow-up works.
CARS Application Examples

- CARS for Bracketing Transduction Grammar
- CARS for Hierarchical Phrase-based Model
- CARS for Tree-to-String Model
- CARS using Topic Model
- CARS for Agglutinative Language Translation
Hierarchical Phrase-Based Model

David Chiang. ACL2005

\[ X \rightarrow< \text{在} X_1 \text{ 的 } X_2, \text{ X}_2 \text{ in } X_1 > \]
Rule Selection in HPB Model
Static Rule Selection

在 $X_1$ 的 $X_2$

在 $X_1$ 的 $X_2$

在今天的会议上

在今天的会议上

$X_2$ in $X_1$

0.4

$X_2$ in $X_1$

0.3

at $X_1$ ’s $X_2$

0.2

with $X_2$ of $X_1$

meeting in today
0.3

在 \( X_1 \) 的 \( X_2 \) 与 

0.4

在 \( X_2 \) in \( X_1 \) 会上

0.3

在 \( X_2 \) on \( X_1 \)’s 会议

0.2

在 \( X_2 \) of \( X_1 \) 会

0.2

在 \( X_2 \) of \( X_1 \) 会
Static Rule Selection

- The corresponding string of $X_1$ and $X_2$ have strong preference for rule selection.
- CARS should be helpful.
Maximum Entropy RS Model

\[ P(r_i | S, C) = \left( \frac{\exp(\sum_k h_k(R, X_1^N))}{\sum_{r_j} \exp(\sum_k (\lambda_k h_k(R, X_1^N)))} \right) \]

*\(R\): Neighbour Context

*\(X_1^N\): Variables Context
Context for Rule Selection

\[
X \quad \text{context}
\]

\[
X_i \quad \text{sub-tree}
\]

\[
X_j \quad \text{sub-tree}
\]

\[
\text{context}
\]
Translation Rule

在 \( X_1 \) 的 \( X_2 \) → at \( X_1 \) ’s \( X_2 \)
加强在经济领域的合作。
加强在经济领域的合作。
加强在经济领域的合作。
Source Variable Lengths

Variable | Feature | Value
---|---|---
$X_1$ | Length | 2
$X_2$ | Length | 1
Source Neighbour Words and POS

加强在经济领域的合作。

<table>
<thead>
<tr>
<th>Neighbour</th>
<th>Feature</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>左边词</td>
<td>Left Word</td>
<td>加强</td>
</tr>
<tr>
<td>左边POS</td>
<td>Left POS</td>
<td>VERB</td>
</tr>
<tr>
<td>右边词</td>
<td>Right Word</td>
<td>。</td>
</tr>
<tr>
<td>右边POS</td>
<td>Right POS</td>
<td>PUNCT</td>
</tr>
</tbody>
</table>
Target Expression Instantiation

X₂ in the X₁

strengthen the cooperation in the economic field.
strengthen the cooperation in the economic field.
strengthen the cooperation in the economic field.
strengthen the cooperation in the economic field.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Feature</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(X_1)</td>
<td>Length</td>
<td>2</td>
</tr>
<tr>
<td>(X_2)</td>
<td>Length</td>
<td>2</td>
</tr>
</tbody>
</table>
Inapplicable because we use a bottom-up decoding manner
Experiment Settings

- Chinese-to-English translation
- **Baseline**: Reimplementation of Hiero (Chiang 2005)
- Corpus:

<table>
<thead>
<tr>
<th>Task Name</th>
<th>Training corpus</th>
<th>Dev. set</th>
<th>Test set</th>
</tr>
</thead>
<tbody>
<tr>
<td>IWSLT05</td>
<td>BTEC (40k sent. 354k + 378k)</td>
<td>IWSLT04 (500 sent.)</td>
<td>IWSLT05 (506 sent.)</td>
</tr>
<tr>
<td>NIST03</td>
<td>FBIS (239k sent. 6.9M + 8.9M)</td>
<td>NIST02 (878 sent.)</td>
<td>NIST03 (919 sent.)</td>
</tr>
</tbody>
</table>
## Experiment Results

<table>
<thead>
<tr>
<th>System</th>
<th>NIST03 (BLEU-4%)</th>
<th>IWSLT05 (BLEU-4%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>28.05</td>
<td>56.20</td>
</tr>
<tr>
<td>lexical features (source-side)</td>
<td>28.26</td>
<td>56.51</td>
</tr>
<tr>
<td>POS features</td>
<td>28.78</td>
<td>56.95</td>
</tr>
<tr>
<td>lexical features (source-side) + POS features</td>
<td>28.89</td>
<td>56.99</td>
</tr>
<tr>
<td>lexical features (source-side) + POS features + length features (source-side)</td>
<td>28.96</td>
<td>57.10</td>
</tr>
<tr>
<td>All features (source + target)</td>
<td><strong>29.02</strong></td>
<td><strong>57.20</strong></td>
</tr>
</tbody>
</table>

* case insensitive

↑ 0.97  
↑ 1.0
Better Phrase Translation: for *terminal* rules

<table>
<thead>
<tr>
<th>Source</th>
<th>恐怕 这趟 航班 已经 订满 了。</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>$X \rightarrow &lt; X_1 \text{订 满}, X_1 \text{booked}&gt;$</td>
</tr>
<tr>
<td></td>
<td>I’m afraid already booked for this flight.</td>
</tr>
<tr>
<td>Baseline +MERS</td>
<td>$X \leftrightarrow &lt; X_1 \text{订 满}, X_1 \text{full}&gt;$</td>
</tr>
<tr>
<td></td>
<td>I’m afraid this flight is full.</td>
</tr>
</tbody>
</table>
Better Phrase Reordering: for *nonterminal* rules

<table>
<thead>
<tr>
<th>source</th>
<th>… 联合国 安全 理事会 的 五个 常任 理事国 …</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>( X \rightarrow \langle X_1 \text{ 的 } X_2, \text{ the } X_1 X_2 \rangle )</td>
</tr>
<tr>
<td>Baseline</td>
<td>… the <em>United Nations Security Council</em> five permanent members …</td>
</tr>
<tr>
<td>+MERS</td>
<td>( X \rightarrow \langle X_1 \text{ 的 } X_2, X_2 \text{ of } X_1 \rangle )</td>
</tr>
<tr>
<td>+MERS</td>
<td>… the five permanent members of the <em>UN Security Council</em> …</td>
</tr>
</tbody>
</table>
Summary

• A MERS model was proposed for hierarchical phrase-based model
• Features used in MERS model:
  – Boundary words and POS tags of internal variables
  – Boundary words and POS tags of neighbours
• MERS help to improve the system performance significantly
CARS Application Examples

- CARS for Bracketing Transduction Grammar
- CARS for Hierarchical Phrase-based Model
- CARS for Tree-to-String Model
- CARS using Topic Model
- CARS for Agglutinative Language Translation
Tree-to-String Model

Yang Liu et al. ACL2006
Liang Huang et al. AMTA2006

A tree-to-string translation rule

Source syntax tree

Target string

Lexical translation

Phrase reordering
Rule Selection Problem

工业产品的制造水平

DNP

NP

NPB

NP

DEG

NN

NN

X₁

的

X₂

水平

X₁  X₂ levels

industrial products  manufacturing  levels

比赛的总体水平

DNP

NP

NPB

NP

DEG

NN

NN

X₁

的

X₂

水平

X₂ standard of X₁

overall standard of the match
Maximum Entropy RS Model

\[ P(r_i | S, C) = \left( \frac{\exp\left( \sum_k h_k(R,Y,X_1^N) \right)}{\sum_{r_j} \exp\left( \sum_k \lambda_k h_k(R,Y,X_1^N) \right)} \right) \]

- **R**: Neighbours
- **Y**: Syntax Tree Context
- **X_1^N**: Internal Variables in Rules
Context for Rule Selection

\[ X \]

\[ X_1 \]

\[ X_2 \]

\[ X_3 \]

\[ X_i \]

\[ X_j \]

context

sub-tree

context
Feature Definition

... improving industrial products ’s manufacturing ...

... industrial products ’s manufacturing ...

VP

... VV

提高

VP

... VV

提高

VP

... VV

提高
Feature Definition: Lexical Features (LF)

... improving industrial products ’s manufacturing ...

VP

... VV

DNP

X1:NP

NP

DEG

NN

NN

工业

产品

的

NPB

NN

制造
Feature Definition: POS Features (POSF)

... VP VV ... NP DNP X:NP NN NN DEG NP NPB NN ... 

提高 工业 产品的 制造

improving industrial products ’s manufacturing ...
Feature Definition: Span Features (SPF)

...  
VP  
...  
VV  

VP  
...  
VV  

VP  
...  
VV  

DNP  
X₁:NP  

NPB  
NN  
...  

NP  

VP  
...  
VV  

X₁:NP  

NN  

NN  

DEG  

NN  

NN  

的  

的  

的  

制造

...  
improving  
industrial  
products  
’s  
manufacturing  
...
Feature Definition: Parent Feature (PF)

... improving industrial products ’s manufacturing ...

... improving industrial products ’s manufacturing ...

VP

... VV

提高

DNP

X₁:NP

NN

工业

NN

产品

DEG

的

NPB

NN

制造

...
Feature Definition: Sibling Features (SBF)

... improving industrial products ’s manufacturing ...

提高工业产品的制造
Experiments

• Chinese-to-English translation
• Baseline: Lynx (Liu Yang, et al., 2006), the state-of-the-art syntax-based SMT system
• Corpus:

<table>
<thead>
<tr>
<th>Training corpus</th>
<th>Dev. set</th>
<th>Test set</th>
</tr>
</thead>
<tbody>
<tr>
<td>FBIS (239k sent.</td>
<td>NIST02</td>
<td>NIST03 (919 sent.)</td>
</tr>
<tr>
<td>6.9M + 8.9M)</td>
<td>(878 sent.)</td>
<td>(1082 sent.)</td>
</tr>
</tbody>
</table>
## Results

<table>
<thead>
<tr>
<th>System</th>
<th>NIST03 (BLEU-4%)</th>
<th>NIST05 (BLEU-4%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lynx</td>
<td>26.15</td>
<td>26.09</td>
</tr>
<tr>
<td>LF</td>
<td>26.12</td>
<td>26.32</td>
</tr>
<tr>
<td>POSF</td>
<td>26.36</td>
<td>26.21</td>
</tr>
<tr>
<td>PF</td>
<td>26.17</td>
<td>25.90</td>
</tr>
<tr>
<td>SBF</td>
<td>26.47</td>
<td>26.08</td>
</tr>
<tr>
<td>LF+POSF</td>
<td>26.61</td>
<td>26.59</td>
</tr>
<tr>
<td>LF+POSF+SPF</td>
<td>26.70</td>
<td>26.44</td>
</tr>
<tr>
<td>LF+POSF+PF</td>
<td>26.81</td>
<td>26.56</td>
</tr>
<tr>
<td>LF+POSF+SBF</td>
<td>26.68</td>
<td>26.89</td>
</tr>
<tr>
<td>ALL</td>
<td><strong>27.05</strong></td>
<td><strong>27.28</strong></td>
</tr>
</tbody>
</table>

↑ 0.9

↑ 1.19
Three kinds of TATs

Lexicalized: Lexical translation

Partially lexicalized: Lexical translation & Phrase reordering

Unlexicalized: Phrase reordering
More than 78% source trees are ambiguous!
<table>
<thead>
<tr>
<th>Source</th>
<th>马耳他 位于 欧洲 南部</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lynx</td>
<td>Malta in southern Europe</td>
</tr>
<tr>
<td>Lynx+MERS</td>
<td>Malta is located in southern Europe</td>
</tr>
</tbody>
</table>

VV

位于

in

VV

位于

is located in
## Better Phrase Reordering

<table>
<thead>
<tr>
<th>Source</th>
<th>按照 在 中国 市场 的 发展 战略 , ...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lynx</td>
<td>Accordance with <strong>the Chinese market development strategy</strong>, ...</td>
</tr>
<tr>
<td>Lynx+MERS</td>
<td>According to the <strong>development strategy in the Chinese market</strong> , ...</td>
</tr>
</tbody>
</table>
(in) the Chinese market development strategy
Summary

• A MERS model was proposed for tree-to-string model

• Features used in MERS model:
  – Boundary words and POS tags of internal variables
  – Boundary words and POS tags of neighbours
  – Syntax labels of parent node and sibling node

• MERS help to improve the system performance significantly
CARS Application Examples

- CARS for Bracketing Transduction Grammar
- CARS for Hierarchical Phrase-based Model
- CARS for Tree-to-String Model
- CARS using Topic Model
- CARS for Agglutinative Language Translation
Rule Selection by Topic

Bank

Mouse
Rule Selection by Topic

Bank

河岸  Geography

Mouse

鼠标  Computer

银行  Finance

老鼠  Biology
Topic Distribution of Rules

给予 $X_1 \Rightarrow$ give $X_1$

作战 能力 $\Rightarrow$ operational capacity

授予 $X_1 \Rightarrow$ grants $X_1$

$X_1$ 举行 会谈 $X_2 \Rightarrow$ held talks $X_1 X_2$
Topic Similarity and Sensitivity

• Topic **Similarity** Model
  – Describe the relatedness of rules to topics of given documents

• Topic **Sensitivity** Model
  – Distinguish topic-insensitive rules and topic-sensitive rules
Topic Similarity Model

Source Document
Topic Similarity Model

Source Document

给予 $X_1 \Rightarrow give X_1$

給予 $X_1 \Rightarrow grants X_1$
Topic Similarity Model

\[ \sum_{k=1}^{K} \left( \sqrt{\hat{p}(z = k|d)} - \sqrt{\hat{p}(z = k|r)} \right)^2 \]
Topic Sensitivity Model

Topic-sensitive Rule
Applied in few topics

Topic-insensitive Rule
Applied in many topics

• Describe by Entropy as a metric
Topic Sensitivity Model

- Topic-insensitive rules are always penalized
- But common, sometime more preferable
- Sensitivity as a complement
Estimation
Estimation

\((r_1, 0.1, \ldots)\) \hspace{2cm} \((r_2, 0.5, \ldots)\)

\[\text{source}\]
Estimation

\[(r_1, 0.1, \quad , \quad , \quad ) \quad (r_2, 0.5, \quad , \quad , \quad ) \quad \ldots \ldots \]
One-to-many Topic Projection

Target Distribution
One-to-many Topic Projection

Target Distribution

Topic Assignment

<table>
<thead>
<tr>
<th>Topic Assignment Alignment</th>
<th>Topic-to-Topic Projection Matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td>e3</td>
<td>f10  f15  f10  f26</td>
</tr>
<tr>
<td>e3</td>
<td>0.1   0.4   0.1    ⋯</td>
</tr>
<tr>
<td>e8</td>
<td>0.3   0.2   0.1    ⋯</td>
</tr>
<tr>
<td>e8</td>
<td>0.4   0.1   0.3    ⋯</td>
</tr>
<tr>
<td></td>
<td>⋯     ⋯     ⋯      ⋯</td>
</tr>
</tbody>
</table>
## One-to-many Topic Projection

<table>
<thead>
<tr>
<th>e-topic</th>
<th>f-topic 1</th>
<th>f-topic 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>enterprises</td>
<td>农业 (agricultural)</td>
<td>企业 (enterprise)</td>
</tr>
<tr>
<td>rural</td>
<td>农村 (rural)</td>
<td>市场 (market)</td>
</tr>
<tr>
<td>state</td>
<td>农民 (peasant)</td>
<td>国有 (state)</td>
</tr>
<tr>
<td>agricultural</td>
<td>改革 (reform)</td>
<td>公司 (company)</td>
</tr>
<tr>
<td>market</td>
<td>财政 (finance)</td>
<td>金融 (finance)</td>
</tr>
<tr>
<td>reform</td>
<td>社会 (social)</td>
<td>银行 (bank)</td>
</tr>
</tbody>
</table>

\[
P(z_f|z_e) = \begin{array}{c}
0.38 \\
0.28
\end{array}
\]
One-to-many Topic Projection

Topic Assignment

Alignment

Target Distribution

Topic-to-Topic Projection Matrix

Projected Target Distribution
Topic-based Rule Selection Model

- Similarity (source)
- Similarity (target)
- Sensitivity (source)
- Sensitivity (target)

Used as four features in log-linear model for SMT

Xinyan Xiao et al. ACL 2012
Experiment Setup

- In-house implementation of HPB model
- Topic Tool: GibbsLDA++
- Bilingual corpus: FBIS 239K sentence pairs
  - With document boundary
  - For both LDA training and rule extraction
- Report Average BLEU on test sets NIST06, NIST08
Effect of Topic Similarity Model

- Baseline
- Topic lex
- SimSrc
- SimTgt
- SimSrc+SimTgt
- Sim+Sen

Scores:
- Baseline: 25.8
- Topic lex: 26
- SimSrc: 26.2
- SimTgt: 26.4
- SimSrc+SimTgt: 26.6
- Sim+Sen: 26.8
Effect of Sensitivity Model

- Baseline
- Topic lex
- SimSrc
- SimTgt
- SimSrc+SimTgt
- Sim+Sen

Values:
- Baseline: 25.8
- Topic lex: 26
- SimSrc: 26.2
- SimTgt: 26.4
- SimSrc+SimTgt: 26.6
- Sim+Sen: 26.8

X-axis: Values from 25.8 to 27
One-to-many Topic Projection

- baseline
- one-to-one
- one-to-many

Values:
- 25.9
- 26.1
- 26.3
- 26.5
Summary

• Compared with word-level WSD, our Topic-based Rule Selection Model is more effective.

• A topic similarity model and a topic sensitive model are used in both source side and target side.

• Document boundary is necessary in training corpus.
CARS Application Examples

- CARS for Bracketing Transduction Grammar
- CARS for Hierarchical Phrase-based Model
- CARS for Tree-to-String Model
- CARS using Topic Model
- CARS for Agglutinative Language Translation
Outline

- Introduction
- Context-Aware Rule-Selection
- CARS Application Examples
- Conclusion and Future Work
Conclusion

• An idea of Context-Aware Rule-Selection is proposed
• CARS is very effective on various translation models
• CARS is compatible with log-linear model for SMT
• CARS is very convenient for incorporating various context features and linguistic knowledge.
Future Work

- CARS by Semantic Role Labeling
- CARS by Coreference
- ……
Thanks!

Q & A