Joint Tokenization, Parsing and Translation

Yang Liu

Institute of Computing Technology
Chinese Academy of Sciences

joint work with Xinyan Xiao, Qun Liu, Young-Sook Hwang, and Shouxun Lin
Tokenization

Mr. Smith, please pay 3,000 dollars for the computer.
Tokenization

Mr. Smith, please pay 3,000 dollars for the computer.

Mr. Smith, please pay 3,000 dollars for the computer.
Tokenization Ambiguity

- Tokenization is easy for English, but hard for many languages such as Chinese

下雨天地面积水
Tokenization Ambiguity

- Tokenization is easy for English, but hard for many languages such as Chinese

下雨天地面积

下雨天 地面 积水
Tokenization Ambiguity

- Tokenization is easy for English, but hard for many languages such as Chinese

下雨天地面面积水

在下雨天，地面积水

in rainy days the ground is wet
Tokenization Ambiguity

- Tokenization is easy for English, but hard for many languages such as Chinese.

下雨天地面积水

在雨天

地面

积水

下雨

天地

面积

水

in rainy days

the ground

is wet
Tokenization Ambiguity

• Tokenization is easy for English, but hard for many languages such as Chinese

下雨天 地面 积水
下雨 天地 面积 水

in rainy days the ground is wet
rain world area water
Tokenization Ambiguity

- Tokenization is easy for English, but hard for many languages such as Chinese.

- Different tokenizations may lead to different translations.
Phrase-based Translation on 1-best Tokenization

(Koehn et al., 2003)
Phrase-based Translation on 1-best Tokenization

(地面, the ground)

下雨天  地面  积水

(Koehn et al., 2003)
Phrase-based Translation on 1-best Tokenization

(地面, the ground)

(Koehn et al., 2003)
Phrase-based Translation on 1-best Tokenization

(Koehn et al., 2003)
Phrase-based Translation on 1-best Tokenization

(积水, is wet)

(Koehn et al., 2003)
Phrase-based Translation on 1-best Tokenization

(积水, is wet)

下雨天 地面 积水  
the ground is wet

(Koehn et al., 2003)
Phrase-based Translation on 1-best Tokenization

下雨天 地面 积水
the ground is wet

(Koehn et al., 2003)
Phrase-based Translation on 1-best Tokenization

(下雨天, in rainy days)

(Koehn et al., 2003)
Phrase-based Translation on 1-best Tokenization

(下雨天, in rainy days)

(Koehn et al., 2003)
Tokenization Mistake Propagation

source string → tokenize → translate → target string
Tokenization Mistake Propagation

(source) string tokenize token. translate (target) string

(source) string tokenize lattice translate (target) string
Phrase-based Translation on Lattice

(Dyer et al., 2008)
Phrase-based Translation on Lattice

(地面, the ground)

(Dyer et al., 2008)
Phrase-based Translation on Lattice

(地面, the ground)

(Dyer et al., 2008)
Phrase-based Translation on Lattice

(Dyer et al., 2008)
Phrase-based Translation on Lattice

(Dyer et al., 2008)
Phrase-based Translation on Lattice

(Dyer et al., 2008)
Phrase-based Translation on Lattice

(Dyer et al., 2008)
Phrase-based Translation on Lattice

（积水, is wet）

the ground is wet

(Dyer et al., 2008)
the ground is wet

(Dyer et al., 2008)
Phrase-based Translation on Lattice

(下雨天, in rainy days)

the ground is wet

(Dyer et al., 2008)
Phrase-based Translation on Lattice

(下雨天, in rainy days)

the ground is wet

(Dyer et al., 2008)
Phrase-based Translation on Lattice

(下雨天, in rainy days)

the ground is wet in rainy days

(Dyer et al., 2008)
Tokenization and Translation: **Separate Vs. Joint**

- **Source** string → **Tokenize** → **Translate** → **Target** string
- **Source** string → **Tokenize** → **Lattice** → **Translate** → **Target** string
Tokenization and Translation: **Separate** Vs. **Joint**

- **Separate** tokenization and translation:
  - Source string → Tokenize → Translate → Target string
  - Target string → Tokenize → Translate → Source string

- **Joint** tokenization and translation:
  - Source string → Tokenize + Translate → Target string
  - Target string → Tokenize + Translate → Source string
Tokenization and Translation: **Separate** Vs. **Joint**

- **Separate**: Source → tokenize → translate → Target
  - Source string → tokenize → translation string → Target string

- **Joint**: Source → tokenize + translate → Target
  - Source string → tokenize + translate → Target string
Tokenization and Translation: **Separate Vs. Joint**

- **Separate**
  - Source string \(\rightarrow\) Tokenize \(\rightarrow\) Translate \(\rightarrow\) Target string

- **Joint**
  - Source string \(\rightarrow\) Tokenize \(\rightarrow\) Lattice \(\rightarrow\) Translate \(\rightarrow\) Target string
  - Source string \(\rightarrow\) Tokenize + Translate \(\rightarrow\) Target string
Tokenization and Translation: **Separate Vs. Joint**

1. **Separate**
   - **Source** string → **tokenize** → **translate** → **Target** string

2. **Separate**
   - **Source** string → **tokenize** → **lattice** → **translate** → **Target** string

3. **Joint**
   - **Source** string → **tokenize + translate** → **Target** string
Joint Tokenization and Translation

(Xiao et al., 2010)
Joint Tokenization and Translation

(地面, the ground)

(Xiao et al., 2010)
Joint Tokenization and Translation

(地面, the ground)

(Xiao et al., 2010)
Joint Tokenization and Translation

(地面, the ground)

(Xiao et al., 2010)
Joint Tokenization and Translation

(Xiao et al., 2010)
Joint Tokenization and Translation

(积水, is wet)

(Xiao et al., 2010)
Joint Tokenization and Translation

(Xiao et al., 2010)
Joint Tokenization and Translation

下 雨 天 地 面 积 水

the ground is wet

(積水, is wet)

(Xiao et al., 2010)
Joint Tokenization and Translation

the ground is wet

(Xiao et al., 2010)
Joint Tokenization and Translation

(下雨天, in rainy days)

the ground is wet

(Xiao et al., 2010)
Joint Tokenization and Translation

the ground is wet

(Xiao et al., 2010)
Joint Tokenization and Translation

(下雨天, in rainy days)

下雨天，地面是湿的

the ground is wet in rainy days

(Xiao et al., 2010)
Joint Tokenization and Translation

(下雨天, in rainy days)

the ground is wet in rainy days

(Xiao et al., 2010)
Model and Training

\[
score(f, t, e) = \sum_k \lambda_k h_k(f, t, e)
\]

- We use a linear model to combine tokenization and translation models as features
- The minimum-error-rate training algorithm is used for training feature weights

(Och, 2003)
Translation Evaluation

**BLEU**

<table>
<thead>
<tr>
<th>Method</th>
<th>30</th>
<th>31</th>
<th>32</th>
<th>33</th>
<th>34</th>
<th>35</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICT</td>
<td>33.06</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stanford</td>
<td>33.22</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MaxEnt</td>
<td>30.91</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>33.95</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(Xiao et al., 2010)
Translation Evaluation

BLEU

separate

ICT

Stanford

MaxEnt

All

30
31
32
33
34
35

joint

33.06
33.22
30.91
33.95

(Xiao et al., 2010)
Translation Evaluation

- ICT: 33.06
- Stanford: 33.22
- MaxEnt: 30.91
- All: 33.95

(Xiao et al., 2010)
Translation Evaluation

- **ICT**: 33.06
- **Stanford**: 33.22
- **MaxEnt**: 30.91
- **All**: 33.95

**Separate** vs **Joint**: +1.63

(Xiao et al., 2010)
Translation Evaluation

<table>
<thead>
<tr>
<th>Method</th>
<th>BLEU Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICT</td>
<td>33.06</td>
</tr>
<tr>
<td>Stanford</td>
<td>33.22</td>
</tr>
<tr>
<td>MaxEnt</td>
<td>30.91</td>
</tr>
<tr>
<td>All</td>
<td>33.95</td>
</tr>
</tbody>
</table>

(Xiao et al., 2010)
Translation Evaluation

(Xiao et al., 2010)
Translation Evaluation

<table>
<thead>
<tr>
<th>Method</th>
<th>BLEU</th>
<th>Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICT</td>
<td>33.06</td>
<td>+1.63</td>
</tr>
<tr>
<td>Stanford</td>
<td>33.22</td>
<td>+1.34</td>
</tr>
<tr>
<td>MaxEnt</td>
<td>30.91</td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>33.95</td>
<td></td>
</tr>
</tbody>
</table>

(Xiao et al., 2010)
Translation Evaluation

<table>
<thead>
<tr>
<th>Model</th>
<th>BLEU (separate)</th>
<th>BLEU (joint)</th>
<th>Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICT</td>
<td>33.06</td>
<td>33.06</td>
<td>+1.63</td>
</tr>
<tr>
<td>Stanford</td>
<td>33.22</td>
<td>33.22</td>
<td>+1.34</td>
</tr>
<tr>
<td>MaxEnt</td>
<td>30.91</td>
<td>34.17</td>
<td>+3.26</td>
</tr>
<tr>
<td>All</td>
<td>33.95</td>
<td>35.64</td>
<td>-</td>
</tr>
</tbody>
</table>

(Xiao et al., 2010)
Translation Evaluation

<table>
<thead>
<tr>
<th>Method</th>
<th>BLEU Score</th>
<th>Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICT</td>
<td>33.06</td>
<td>+1.63</td>
</tr>
<tr>
<td>Stanford</td>
<td>33.22</td>
<td>+1.34</td>
</tr>
<tr>
<td>MaxEnt</td>
<td>30.91</td>
<td>+3.26</td>
</tr>
<tr>
<td>All</td>
<td>33.95</td>
<td></td>
</tr>
</tbody>
</table>

(Xiao et al., 2010)
## Translation Evaluation

<table>
<thead>
<tr>
<th>Model</th>
<th>BLEU (separate)</th>
<th>BLEU (joint)</th>
<th>Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICT</td>
<td>33.06</td>
<td>33.22</td>
<td>+1.63</td>
</tr>
<tr>
<td>Stanford</td>
<td>33.22</td>
<td>33.34</td>
<td>+1.34</td>
</tr>
<tr>
<td>MaxEnt</td>
<td>30.91</td>
<td>30.91</td>
<td>0.00</td>
</tr>
<tr>
<td>All</td>
<td>33.95</td>
<td>34.08</td>
<td>+0.93</td>
</tr>
</tbody>
</table>

(Xiao et al., 2010)
<table>
<thead>
<tr>
<th>Model</th>
<th>F1 Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICT</td>
<td>97.47</td>
</tr>
<tr>
<td>Stanford</td>
<td>97.48</td>
</tr>
<tr>
<td>MaxEnt</td>
<td>95.53</td>
</tr>
</tbody>
</table>

(Xiao et al., 2010)
Tokenization Evaluation

(Xiao et al., 2010)
Tokenization Evaluation

<table>
<thead>
<tr>
<th>Method</th>
<th>separate</th>
<th>joint</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICT</td>
<td>97.47</td>
<td></td>
</tr>
<tr>
<td>Stanford</td>
<td>97.48</td>
<td></td>
</tr>
<tr>
<td>MaxEnt</td>
<td>95.53</td>
<td></td>
</tr>
</tbody>
</table>

(Xiao et al., 2010)
Tokenization Evaluation

<table>
<thead>
<tr>
<th>Method</th>
<th>separate</th>
<th>joint</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICT</td>
<td>97.47</td>
<td>+0.21</td>
</tr>
<tr>
<td>Stanford</td>
<td>97.48</td>
<td></td>
</tr>
<tr>
<td>MaxEnt</td>
<td>95.53</td>
<td></td>
</tr>
</tbody>
</table>

(Xiao et al., 2010)
Tokenization Evaluation

<table>
<thead>
<tr>
<th>Model</th>
<th>Separate F1</th>
<th>Joint F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICT</td>
<td>97.47</td>
<td>97.68</td>
</tr>
<tr>
<td>Stanford</td>
<td>97.48</td>
<td>97.79</td>
</tr>
<tr>
<td>MaxEnt</td>
<td>95.53</td>
<td>95.74</td>
</tr>
</tbody>
</table>

(Xiao et al., 2010)
Tokenization Evaluation

<table>
<thead>
<tr>
<th>Method</th>
<th>Separate F1</th>
<th>Joint F1</th>
<th>Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICT</td>
<td>97.47</td>
<td>97.68</td>
<td>+0.21</td>
</tr>
<tr>
<td>Stanford</td>
<td>97.48</td>
<td>97.68</td>
<td>+0.20</td>
</tr>
<tr>
<td>MaxEnt</td>
<td>95.53</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(Xiao et al., 2010)
Tokenization Evaluation

(Xiao et al., 2010)
### Tokenization Evaluation

<table>
<thead>
<tr>
<th>Model</th>
<th>Separate F1</th>
<th>Joint F1</th>
<th>Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICT</td>
<td>97.47</td>
<td></td>
<td>+0.21</td>
</tr>
<tr>
<td>Stanford</td>
<td>97.48</td>
<td></td>
<td>+0.20</td>
</tr>
<tr>
<td>MaxEnt</td>
<td>95.53</td>
<td></td>
<td>+2.17</td>
</tr>
</tbody>
</table>

(Xiao et al., 2010)
Search Space Comparison

(Xiao et al., 2010)
Search Space Comparison

tokenization-based

(Xiao et al., 2010)
Search Space Comparison

tokenization-based

(Xiao et al., 2010)
Search Space Comparison

tokenization-based

(Xiao et al., 2010)
Search Space Comparison

tokenization-based

(Xiao et al., 2010)
Search Space Comparison

tokenization-based  lattice-based

(Xiao et al., 2010)
Search Space Comparison

(tokenization-based)  (lattice-based)

(Xiao et al., 2010)
Search Space Comparison

(tokenization-based)  (lattice-based)  (string-based)

(Xiao et al., 2010)
Parsing

布什与沙龙举行了会谈
Parsing

布什 与 沙龙 举行 了 会谈

NBP
Parsing

NPB  P
布什  与  沙龙  举行  了  会谈
Parsing

<table>
<thead>
<tr>
<th>NPB</th>
<th>P</th>
<th>NPB</th>
</tr>
</thead>
<tbody>
<tr>
<td>布什</td>
<td>与</td>
<td>沙龙 举行 了 会谈</td>
</tr>
<tr>
<td>NPB</td>
<td>P</td>
<td>NPB</td>
</tr>
<tr>
<td>------</td>
<td>---</td>
<td>------</td>
</tr>
<tr>
<td>布什</td>
<td>与</td>
<td>沙龙</td>
</tr>
</tbody>
</table>
Parsing

NPB | P | NPB VS AS
布什 与 沙龙 举行 了 会谈
Parsing

NPB | P | NPB | VS | AS | NPB
布什 | 与 | 沙龙 | 举行 | 了 | 会谈
Parsing
Parsing
Parsing

[Diagram of a parse tree with nodes labeled NPB, PP, VP, VS, AS, and NPB, with the words Bush, with,沙龙,举行,了,和会谈.
Parsing

IP
  VP
    PP
      NPB
      P
      NPB
    VS
    AS
      NPB

布什 与 沙龙 举行 了 会谈
Structural Ambiguity
Structural Ambiguity

布什 与 沙龙 举行 了 会谈

布什 与 沙龙 举行 了 会谈
Structural Ambiguity
Bush held a talk with Sharon
Structural Ambiguity

Bush held a talk with Sharon

Bush and Sharon held a talk
Tree-to-String Translation

(Liu et al., 2006; Huang et al., 2006)
Tree-to-String Translation

(Liu et al., 2006; Huang et al., 2006)
Tree-to-String Translation

(Liu et al., 2006; Huang et al., 2006)
Tree-to-String Translation

(Liu et al., 2006; Huang et al., 2006)
Tree-to-String Translation

NPB  |
布什 |
 PP   |
P    |
与    |
NPB  |
沙龙  |
VP   |
VS   |
举行  |
AS   |
了  |
NPB  |
会谈 |

(Liu et al., 2006; Huang et al., 2006)
Tree-to-String Translation

(Liu et al., 2006; Huang et al., 2006)
Tree-to-String Translation

(Liu et al., 2006; Huang et al., 2006)
Tree-to-String Translation

(Liu et al., 2006; Huang et al., 2006)
Tree-to-String Translation

(Liu et al., 2006; Huang et al., 2006)
Tree-to-String Translation

(Liu et al., 2006; Huang et al., 2006)
Tree-to-String Translation

(Liu et al., 2006; Huang et al., 2006)
Tree-to-String Translation

(Liu et al., 2006; Huang et al., 2006)
Tree-to-String Translation

Bush held a (Liu et al., 2006; Huang et al., 2006)
Tree-to-String Translation

Bush held a

会谈 与 沙龙

(Liu et al., 2006; Huang et al., 2006)
Tree-to-String Translation

Bush held a talk

(Liu et al., 2006; Huang et al., 2006)
Tree-to-String Translation

Bush held a talk

(Liu et al., 2006; Huang et al., 2006)
Bush held a talk

Tree-to-String Translation

(Liu et al., 2006; Huang et al., 2006)
Bush held a talk

(Liu et al., 2006; Huang et al., 2006)
Bush held a talk with NPB 沙龙
(Liu et al., 2006; Huang et al., 2006)
Bush held a talk with Sharon

(Liu et al., 2006; Huang et al., 2006)
Bush held a talk with Sharon

(Liu et al., 2006; Huang et al., 2006)
Bush held a talk with Sharon

(Liu et al., 2006; Huang et al., 2006)
Parsing Mistake Propagation

source \(\rightarrow\) string \(\rightarrow\) parse \(\rightarrow\) tree \(\rightarrow\) translate \(\rightarrow\) string

target
 Parsing Mistake Propagation

source -> string -> parse -> tree -> translate -> string

source -> string -> parse -> forest -> translate -> string
Packed Forest

(Billot and Lang, 1989)
Matching on Forest

(Mi et al., 2008)
Matching on Forest

"Matching on Forest" (Mi et al., 2008)
Matching on Forest

(Mi et al., 2008)
Matching on Forest

(Mi et al., 2008)
Matching on Forest

(Mi et al., 2008)
Matching on Forest

(Mi et al., 2008)
Matching on Forest

(Mi et al., 2008)
Translation Forest

(Mi et al., 2008)
Translation Forest

(Mi et al., 2008)
Translation Forest

(Mi et al., 2008)
Translation Forest

(Mi et al., 2008)
Translation Forest

(Mi et al., 2008)
Translation Forest

(NPB,1)布什 (NPB,2)沙龙 (NPB,3)布什 (NPB,4)举行 (NPB,5)了 (NPB,6)会谈 (Mi et al., 2008)
Translation Forest

(Mi et al., 2008)
Translation Forest

(Mi et al., 2008)
Translation Forest

(Mi et al., 2008)
Translation Forest

(Mi et al., 2008)
Translation Forest

(Mi et al., 2008)
Translation Forest

NPB  |  Bush
---|---
NPB  |  Sharon
NPB  |  talk

VPB

VS

AS

X1:NPB

NPB

Bush

Sharon

talk

held  a  X1

(Mi et al., 2008)
Translation Forest

Bush held a talk with Sharon and Sharon talked with Bush.

(Mi et al., 2008)
Translation Forest

(Mi et al., 2008)
Translation Forest

NPB_0,1  P_1.2  CC_1.2  NPB_5.6  NPB_2.3  VPB_3.6  NPB_0,1  NPB_2.3  NPB_5.6

Bush  Sharon  talk  held  a  X_1

(Mi et al., 2008)
Translation Forest

(Mi et al., 2008)
Translation Forest

(Mi et al., 2008)
Translation Forest

NPB布什 → Bush
NPB沙龙 → Sharon
NPB会谈 → talk

Translation:
Bush held a x1 with x2
(Bush)’

(Mi et al., 2008)
Translation Forest

(Mi et al., 2008)
Translation Forest

(Mi et al., 2008)
Translation Forest

NPB
布什
Bush

NPB
沙龙
Sharon

NPB
会谈
talk

VPB
held a X1

IP
X1:NPB

NP
X3:VPB

CC
与

NPB
0,1

NPB
2,3

NPB
5,6

PP
1,3

PP
1,6

IP
0,6

VPB
3,6

VS
举行

AS
了

(Mi et al., 2008)
"Bush held a talk with Sharon"
Parsing and Translation: **Separate Vs. Joint**

- **Source** string → **Parse** tree → **Translate** string
- **Target** string → **Parse** forest → **Translate** string
Parsing and Translation: **Separate Vs. Joint**

**Separate**
- Source string → Parse → Tree → Translate → Target string

**Joint**
- Source string → Parse → Forest → Translate → Target string

**parse + translate**
- Source string → Parse + Translate → Tree → Target string
Parsing and Translation: **Separate** Vs. **Joint**

**Separate**

- **Source** string → **Parse** → **Tree** → **Translate** → **Target** string

**Joint**

- **Source** string → **Parse** → **Forest** → **Translate** → **Target** string
- **Source** string → **Parse + Translate** → **Tree** → **Target** string
Parsing and Translation: **Separate** Vs. **Joint**

- **Separate**:
  - Source → String → Parse → Tree → Translate → String
  - Target → String → Parse → Tree → Translate → String

- **Joint**:
  - Source → String → Parse + Translate → Tree → String
  - Target → String → Parse + Translate → Tree → String
Parsing and Translation: Separate Vs. Joint

source → string → parse → tree → translate → string → separate

source → string → parse → forest → translate → string → separate

source → string → parse + translate → tree → string → joint
Joint Parsing and Translation

布什 与 沙龙 举行 了 会谈

(Liu and Liu, 2010)
Joint Parsing and Translation

布什与沙龙举行了会谈

(Liu and Liu, 2010)
Joint Parsing and Translation

NPB
| 布什
→ Bush

NPB
| 布什 与 沙龙 举行 了 会谈

(Liu and Liu, 2010)
Joint Parsing and Translation

NPB
布什
→ Bush

布什 与 沙龙 举行 了 会谈

(Liu and Liu, 2010)
Joint Parsing and Translation

(Liu and Liu, 2010)
Joint Parsing and Translation

(Liu and Liu, 2010)
Joint Parsing and Translation

(NPB)

布什与沙龙举行了会谈

Bush

(Liu and Liu, 2010)
Joint Parsing and Translation

(Liu and Liu, 2010)
Joint Parsing and Translation

Bush NPB 与 NPB 莎朗 举行了 会谈

(Liu and Liu, 2010)
Joint Parsing and Translation

(NPB)布什与沙龙举行了会谈

(Liu and Liu, 2010)
Joint Parsing and Translation

Bush NPB Sharon

会谈

→ talk

(Liu and Liu, 2010)
Joint Parsing and Translation

NPB
| 会谈
→ talk

NPB | NPB | NPB
布什 | 与 | 沙龙 | 举行 | 了 | 会谈
Bush | Sharon | talk

(Liu and Liu, 2010)
Joint Parsing and Translation

(NPB) Bush 与 (NPB) Sharon 举行了 (NPB) 会谈

(Liu and Liu, 2010)
Joint Parsing and Translation

(Liu and Liu, 2010)
Joint Parsing and Translation

(Liu and Liu, 2010)
Joint Parsing and Translation

Bush NPB with Sharon talk

(Liu and Liu, 2010)
Joint Parsing and Translation

Bush NPB with Sharon talk

(Liu and Liu, 2010)
Joint Parsing and Translation

Bush NPB with Sharon talk

NPB PP

布什 与 举行 了 会谈 会话

P  NPB

(Liu and Liu, 2010)
Joint Parsing and Translation

Bush NPB with Sharon talk

(Liu and Liu, 2010)
Joint Parsing and Translation

Bush NPB with Sharon talk

(Liu and Liu, 2010)
Joint Parsing and Translation

Bush NPB with Sharon NPB talk

(Liu and Liu, 2010)
Joint Parsing and Translation

Bush held a talk with Sharon (Liu and Liu, 2010)
Joint Parsing and Translation

Bush with Sharon held a talk

(Liu and Liu, 2010)
Joint Parsing and Translation

Bush held a talk with Sharon (Liu and Liu, 2010)
Bush held a talk with Sharon. (Liu and Liu, 2010)
Joint Parsing and Translation

(Liu and Liu, 2010)
Bush held a talk with Sharon.
Joint Parsing and Translation

Bush held a talk with Sharon

(Liu and Liu, 2010)
Joint Parsing and Translation

Bush held a talk with Sharon

(Liu and Liu, 2010)
Bush held a talk with Sharon

(Liu and Liu, 2010)
Joint Parsing and Translation

\[(Liu \text{ and Liu}, 2010)\]
Bush held a talk with Sharon

(Liu and Liu, 2010)
Model and Training

\[ \text{score}(f, \pi, e) = \sum_{k} \lambda_k h_k(f, \pi, e) \]

- We use a linear model to combine parsing and translation models as features.
- The minimum-error-rate training algorithm is used for training feature weights.

(Liu and Liu, 2010)
Translation Evaluation

- tree-based: 29.8
- forest-based: 31.6
- string-based: 32.7

(BLEU) (Liu and Liu, 2010)
Parsing Evaluation

(Liu and Liu, 2010)
Search Space Comparison

(Liu and Liu, 2010)
Search Space Comparison

tree-based

(Liu and Liu, 2010)
Search Space Comparison

tree-based

(Liu and Liu, 2010)
Search Space Comparison

tree-based

(Liu and Liu, 2010)
Search Space Comparison

(Liu and Liu, 2010)
Search Space Comparison

tree-based

forest-based

(Liu and Liu, 2010)
Search Space Comparison

- tree-based
- forest-based

(Liu and Liu, 2010)
Search Space Comparison

- tree-based
- forest-based
- string-based

(Liu and Liu, 2010)
Conclusion

- Joint decoding integrates separate steps in a pipeline to eliminate the mistake propagation problem.

- Tokenization, parsing, and translation can interact with each other in a discriminative framework.
Future Work

tokenize + parse + translate

source string target

token. and tree string
Thanks
Backup Slides for Q&A
Better Tokenization = Better Translation?

<table>
<thead>
<tr>
<th></th>
<th>decoding</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max F</td>
<td></td>
</tr>
<tr>
<td>Max BLEU</td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>BLEU</td>
</tr>
<tr>
<td>97.37</td>
<td>27.43</td>
</tr>
<tr>
<td>92.49</td>
<td>34.88</td>
</tr>
</tbody>
</table>

(Xiao et al., 2010)
Speed Comparison

- tokenization: 2.6
- lattice: 6.8
- string: 17.7
Speed Comparison (cont.)

- tree: 0.6
- forest: 9.5
- string: 91.7