Joint Tokenization and Translation

Xinyan Xiao¹, Yang Liu¹, Young-Sook Hwang², Qun Liu¹, and Shouxun Lin¹

¹Institute of Computing Technology
²SKTelecom
Separate Tokenization and Translation

Characters

0 tao 1 fei 2 ke 3 you 4 wang 5 duo 6 fen

Characters

陶 菲 克 有 望 夺 分
Separate Tokenization and Translation

Characters

<table>
<thead>
<tr>
<th>Characters</th>
<th>0 tao</th>
<th>1 fei</th>
<th>2 ke</th>
<th>3 you</th>
<th>4 wang</th>
<th>5 duo</th>
<th>6 fen</th>
</tr>
</thead>
</table>

Words

tao-fei-ke  you-wang  duo  fen
Separate Tokenization and Translation

Characters

0 tao 1 fei 2 ke 3 you 4 wang 5 duo 6 fen

Words

tao-fei-ke  you-wang  duo  fen

Translation

Taufik is expected to gain a point
Challenges

- Propagation of tokenization errors
- Hard to define optimal granularity (Chang et al., 2008; Zhang et al., 2008)
- Inconsistence tokenization between training and testing
Previous work

- Offering more tokenizations
  - Lattice (Xu et al., 2005; Dyer et al., 2008; Dyer, 2009)
Joint Tokenization and Translation

tao 1 fei 2 ke 3 you 4 wang 5 duo 6 fen 7
Joint Tokenization and Translation

$r_1$: tao-fei-ke
→ Taufik

0 tao 1 fei 2 ke 3 you 4 wang 5 duo 6 fen 7
Joint Tokenization and Translation

\[ r_1: \text{tao-fei-ke} \rightarrow \text{Taufik} \]
Joint Tokenization and Translation

$r_2$: duo fen
→ gain a point
Joint Tokenization and Translation

\[ r_3: x_1 \text{ you-wang } x_2 \rightarrow x_1 \text{ is expected to } x_2 \]

Taufik is expected to gain a point
Log-linear Model

\[
\text{Score}(c, w, e) = \sum \lambda_k h_k (c, w, e)
\]

Och and Ney 2002
Log-linear Model

\[ \text{Score}(c, w, e) = \sum \lambda_k h_k (c, w, e) \]

- **characters**
- **tokenization**
- **translation**

8 Tokenization Features

8 Translation Features

Och and Ney 2002
Chiang, 2007
Tokenization Features

- Maximum entropy model over label sequences (Xue and Shen, 2003; Ng and Low, 2004)

\[ P(\text{tao-fei-ke} | \text{tao fei ke}) = P(\text{bme} | \text{tao fei ke}) \]
Tokenization Features

- Maximum entropy model over label sequences (Xue and Shen, 2003; Ng and Low, 2004)

\[
P(\text{tao-fei-ke} \mid \text{tao fei ke}) = P(\text{bme} \mid \text{tao fei ke})
\]

- N-gram language model over words
- Word count
- OOV features
Experiment Setup

- Bilingual corpus
  - 1.5M sentence pairs from LDC
- 4-gram English language model
  - Xinhua portion of the GIGAWORD corpus
- Chinese segmentation corpus
  - People’s Daily (6M words)
  - Split into three parts: training, development, and testing
- Three Chinese word segmenters
  - ICTCLAS (ICT)
  - Stanford (SF)
  - Maximum Entropy model (ME)
  - All: combined corpus processed by the three segmenters to extract rules
Translation Result on MT2005

![Bar chart showing translation results for Joint and Separate tasks. The chart includes results for ICT, SF, ME, and All tasks.](chart.png)

**Joint**
- ICT: 33.06
- SF: 33.22
- ME: 30.91
- All: 33.95

**Separate**
- ICT: 30
- SF: 31
- ME: 32
- All: 33

Lattice (Dyer et al., 2008)
Translation Result on MT2005

<table>
<thead>
<tr>
<th>Joint</th>
<th>Separate</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICT</td>
<td>34.69</td>
</tr>
<tr>
<td>SF</td>
<td>34.56</td>
</tr>
<tr>
<td>ME</td>
<td>34.17</td>
</tr>
<tr>
<td>All</td>
<td>34.88</td>
</tr>
</tbody>
</table>

Lattice (Dyer et al., 2008)
Tokenization Result

<table>
<thead>
<tr>
<th></th>
<th>ICT</th>
<th>SF</th>
<th>ME</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joint</td>
<td>97.47</td>
<td>97.48</td>
<td>97.68</td>
<td>97.68</td>
</tr>
<tr>
<td>Separate</td>
<td>95.53</td>
<td>97.5</td>
<td>97.5</td>
<td>97.5</td>
</tr>
</tbody>
</table>

2010-08-24
**Better Tokenization = Better Translation?**

<table>
<thead>
<tr>
<th>Criterion for testing</th>
<th>Criterion for tuning weights</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>F-1</strong> on test-set of People’s Daily</td>
<td><strong>Max F-1</strong> on dev-set of People’s Daily</td>
</tr>
<tr>
<td><strong>BLEU</strong> on MT2005</td>
<td><strong>Max BLEU</strong> on MT2002</td>
</tr>
<tr>
<td>97.37</td>
<td>92.49</td>
</tr>
<tr>
<td>27.43</td>
<td>34.88</td>
</tr>
</tbody>
</table>
An Example from MT2002

**Character** 作为第二或第三单打的陶菲克就有望夺分

**Reference** Taufik, as the second or third singles player, will have hopes of scoring

**Stanford** 作为第二或第三单打的陶菲克就有望夺分
as the second or third hopefully fought alone

**Lattice** 作为第二或第三单打的陶菲克就有望夺分
as the second or third singles tao菲克 expected to win

**Joint** 作为第二或第三单打的陶菲克就有望夺分
as the second or third singles 陶菲克 expected to win points

Granularity: Too small / large

Tokenization error
Conclusion

- Better tokenization ≠ Better translation
- Joint tokenization and translation provide an elegant and effective way to
  - Optimize tokenization for translation
  - Improve tokenization by translation information
- Future work
  - Apply this method in other models
  - Improve the performance for morphologically rich languages
Better tokenization ≠ Better translation

Tokenization  Joint  Translation

Thank you!

Thanks to Wenbin Jiang, Zhiyang Wang, Zongcheng Ji and anonymous reviewers
Considering All Tokenizations

Constructing tokenization by the extracted rules will limit the search space of tokenization.

0-3 tao fei ke

0-2 tao fei

0-1 tao

1-2 fei

2-3 ke

tao fei ke

0 1 2 3
Considering All Tokenizations

When a span is not tokenized into a single word by the extracted rules, we considering the entire span as an OOV.
Considering All Tokenizations

Working with the glue rule (S \rightarrow SX, SX), we can construct all potential tokenization
Considering All Tokenizations

We use two types of features to control the generation of OOV

- OOV character count (OCC): the number of OOV characters
  - OCC = 3
  - OD_1 = 3

- OOV discount features: penalties for the OOV words with different number of characters (OD_i = 1,2,3,4+)
  - OCC = 3
  - OD_3 = 1

```
tao fei ke
```
```
tao-fei-ke
```