Automatic Evaluation of Generation and Parsing for MT with Automatically Induced Transfer Rules

Yvette Graham, Deirdre Hogan and Josef van Genabith
National Centre for Language Technology
School of Computing
Dublin City University
Overview

1. Transfer-based MT

2. MT Evaluation

3. Conventional Parsing/Generation Evaluation Methods

4. New Parsing/Generation Evaluation Method

5. Experiments
   • Hand-crafted LFG technologies
   • Automatically-induced LFG technologies

6. Conclusions
1. Transfer-based Machine Translation
Transfer-based MT

An oft-cited future application of parsing and generation.

Relies heavily on the employed
• parsing technologies
• generation technologies
• how well these technologies work together
MT Evaluation in General

- Conventional methods evaluate MT system giving an overall result for the system on a given test set.
Transfer-based MT Evaluation

To understand results more fully

- Need to evaluate parsing & generation technologies in isolation from MT system
- But still in a relevant setting for transfer-based MT system

Previous work

- Furuse & Iida 1992, Meyers et al. 1998, Menezes & Richardson 2001
- Riezler & Maxwell 2006
  - Training data 80% full parses, 20% fragment parses
  - Test data “only 44% were in coverage of both parsing and generation technologies”
3. Conventional Parsing and Generation

Evaluation Methods and Transfer-based MT
Conventional Evaluation Methods

Why not use Conventional Methods of Evaluation?

**Conventional Generation Evaluation:**

1: 

```
| ref_t1 | compare | t1 |
```

2: 

```
| ref_t1 | compare | t1 |
```

**Conventional Parsing Evaluation:**

1: 

```
| ref_t1 | parse | f(t1) |
```

2: 

```
| f(t1) | compare | gold standard f(t1) |
```
Conventional Parsing Evaluation

Why not use Conventional Methods of Parsing Evaluation methods?

- **Gold standard** abstract structures of test data required
- Difficult to apply to **new test set**
- Results from different test set, probably **different language domain**
- Does not evaluate how well the **parsing and generation** components work **together**
4. New Evaluation Method for Parsing and Generation
New Evaluation Method

New Method of Parsing / Generation Evaluation

- Evaluate parsing/generation technologies that are cited as having transfer-based MT as a future application
- Include the dependence of the generation technologies on parsing technologies
- Easy evaluation on real MT test data
- Estimate the upper bound imposed on the MT system by these technologies
New Evaluation Method

If we assume every part of the transfer MT system is perfect except for the employed TL parsing and generation technologies, What would $f(t_1)$, i.e. the input to the generator be?
New Evaluation Method

If we assume every part of the transfer MT system is perfect except for the employed TL parsing and generation technologies,

What would $f(t_1)$, i.e. the input to the generator be?

**Answer:** Parsed ref$_t_1$
New Evaluation Method

To estimate the upper bound imposed by parsing/generation components of MT system:

1. \[ f(t_1) \]
   - parse \[ \text{ref}_t \]
   - generate \[ t_1 \]

2. \[ \text{ref}_t \]
   - compare \[ t_1 \]
5. Experiment Results
Results Styles

Entire Test Set Results:

\[
\begin{array}{c}
t_1 \\
t_2 \\
\ldots \\
t_n \\
\text{ref}_t_1 \\
\text{ref}_t_2 \\
\ldots \\
\text{ref}_t_n \\
\end{array}
\]

\[\text{score}\]

Example: NIST=9.1

In-coverage only Results:

\[
\begin{array}{c}
t_1 \\
t_2 \\
\ldots \\
t_m \\
\text{ref}_t_1 \\
\text{ref}_t_2 \\
\ldots \\
\text{ref}_t_m \\
\end{array}
\]

\[\text{score, coverage}= \frac{m}{n}\]

Example: NIST=10.4, coverage=90%
5.1 Experiment 1
Experiment 1: Aim

Compare results of evaluating using new method with results of evaluating using conventional generation method.

Conventional Generation Evaluation:
1: 
   [Diagram showing process]
   \[f(t1) \xrightarrow{\text{generate}} t1\]

2: 
   \[\text{ref}_t1 \xleftarrow{\text{compare}} t1\]

New Parsing & Generation Evaluation:
1: 
   [Diagram showing process]
   \[\text{parse} \xrightarrow{f(t1)} \text{generate} \]

2: 
   \[\text{ref}_t1 \xleftarrow{\text{compare}} t1\]
Experiment 1: Details

Evaluated parsing and generation of technologies

- parsing: Cahill et al. (2004)
- generation: Cahill et al. (2006), Hogan et al. (2007)
- automatically induced
- LFG
- English
- trained on WSJ Sections 2-21

Test Data:

- WSJ Section 23

Generator Input:

- Generation Evaluation: gold standard parse tree
- New Evaluation method: raw text sentence
Experiment 1: Entire Test set Results

- Cahill et al. (2004), Cahill et al. (2006), Hogan et al. (2007)
- Entire test set results

<table>
<thead>
<tr>
<th>Section 23 (2416 sentences)</th>
<th>NIST</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>From Gold-standard Trees</td>
<td>13.29</td>
<td>0.6680</td>
</tr>
<tr>
<td>From Parser Trees</td>
<td>13.01</td>
<td>0.6511</td>
</tr>
</tbody>
</table>

- Significant difference for NIST and BLEU scores

Note: domain of language is same as training data
Experiment 1: In-coverage only Results

- Cahill et al. (2004), Cahill et al. (2006), Hogan et al. (2007)
- In-coverage only results

<table>
<thead>
<tr>
<th>Section 23 (2416 sentences)</th>
<th>NIST</th>
<th>BLEU</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>From Gold-standard Trees</td>
<td>13.31</td>
<td>0.6693</td>
<td>99.88%</td>
</tr>
<tr>
<td>From Parser Trees</td>
<td>13.02</td>
<td>0.6515</td>
<td>99.96%</td>
</tr>
</tbody>
</table>

- High coverage
- Scores similar to entire testset scores
5.2 Experiment 2
Experiment 2: Aim

Results of a transfer-based MT system compared to TL parser/generator upper bound.

MT System Results:

1: parse \( s_1 \) to transfer \( f(s_1) \) to generate \( f(t_1) \)

2: \( \text{ref}_t_1 \leftarrow \text{compare} \rightarrow t_1 \)

New Parsing & Generation Evaluation:

1: parse \( \text{ref}_t_1 \) to \( f(t_1) \) to generate \( t_1 \)

2: \( \text{ref}_t_1 \leftarrow \text{compare} \rightarrow t_1 \)
Experiment 2: Details

**MT System Evaluation**

- Riezler & Maxwell (2006)
- MT system uses parsing / generation technologies of Riezler et al. (2002)
- Published results: NIST

**Employed TL parsing/ generation technologies of MT System**

- hand-crafted
- LFG f-structures
- English

**Test Data:**

- English Europarl test set (Koehn et al 2003) sentences length 5-15
## Experiment 2: Entire Test set Results

<table>
<thead>
<tr>
<th>Europarl (5-15)</th>
<th>NIST</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MT System</strong> (Riezler &amp; Maxwell 2006)</td>
<td>5.62</td>
<td></td>
</tr>
<tr>
<td><strong>TL Parsing/Generation Upper bound</strong>  (Riezler et al. 2002)</td>
<td>12.08</td>
<td>0.7431</td>
</tr>
</tbody>
</table>

- High Upper bound → good news for transfer-based MT 😊
5.3 Experiment 3
Experiment 3: Aims A

parsing/generation results of **hand-crafted** technologies using new method on

parsing/generation results of **treebank-induced** technologies using new method on
Experiment 3: Aims B

parsing/generation results of **hand-crafted** technologies using new method on

<table>
<thead>
<tr>
<th></th>
<th>MT test set X</th>
</tr>
</thead>
<tbody>
<tr>
<td>parsing/generation results of <strong>treebank-induced</strong> technologies using new method on</td>
<td></td>
</tr>
</tbody>
</table>

MT test set X
Experiment 3: Details A

Parsing and generation technologies 1
- Riezler et al. (2002)
- hand-crafted
- LFG f-structures
- English

Parsing and generation technologies 2
- Cahill et al. (2004), Cahill et al. (2006), Hogan et al. (2007)
- Treebank-induced
- LFG f-structures
- English
**Test Data:**

- Europarl test set (Koehn et al 2003) length 5-15 (1755 sentences)
- Europarl test set (Koehn et al 2003) all lengths (500 sentences)
- Homecenter Corpus all lengths (766 sentences)

**String Comparison**

- BLEU
- NIST
### Experiment 3: Entire Test Set Results

<table>
<thead>
<tr>
<th></th>
<th>NIST</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Riezler et al. 2002 Hand-crafted</td>
<td>Cahill et al. 2004 Treebank-Induced</td>
</tr>
<tr>
<td>Europarl (5-15)</td>
<td>12.08</td>
<td>11.72</td>
</tr>
<tr>
<td>Europarl (all lengths)</td>
<td>6.33</td>
<td>10.24</td>
</tr>
<tr>
<td>Homecentre</td>
<td>10.75</td>
<td>10.06</td>
</tr>
</tbody>
</table>

- Results vary greatly from one test set to the next for both hand-crafted and induced technologies
  - but more dramatically for hand-crafted technologies
### Experiment 3: Entire Test set Results cont.

<table>
<thead>
<tr>
<th></th>
<th>NIST</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Riezler et al. 2002 Hand-crafted</td>
<td>Cahill et al. 2004 Treebank-Induced</td>
</tr>
<tr>
<td>Europarl (5-15)</td>
<td><strong>12.08</strong></td>
<td><strong>11.72</strong></td>
</tr>
<tr>
<td>Europarl (all lengths)</td>
<td>6.33</td>
<td>10.24</td>
</tr>
<tr>
<td>Homecentre</td>
<td>10.75</td>
<td>10.06</td>
</tr>
</tbody>
</table>

- Short sentences: hand-crafted better
### Experiment 3: Entire Test set Results cont.

<table>
<thead>
<tr>
<th></th>
<th>NIST</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Riezler et al. 2002 Hand-crafted</td>
<td>Cahill et al. 2004 Treebank-Induced</td>
</tr>
<tr>
<td>Europarl (5-15)</td>
<td>12.08</td>
<td>11.72</td>
</tr>
<tr>
<td>Europarl (all lengths)</td>
<td><strong>6.33</strong></td>
<td><strong>10.24</strong></td>
</tr>
<tr>
<td>Homecentre</td>
<td>10.75</td>
<td>10.06</td>
</tr>
</tbody>
</table>

- Unrestricted length: automatically-induced technologies better
### Experiment 3: Entire Test Set Results cont.

<table>
<thead>
<tr>
<th></th>
<th>NIST</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Riezler et al. 2002 Hand-crafted</td>
<td>Cahill et al. 2004 Treebank-Induced</td>
</tr>
<tr>
<td>Europarl (5-15)</td>
<td>12.08</td>
<td>11.72</td>
</tr>
<tr>
<td>Europarl (all lengths)</td>
<td>6.33</td>
<td>10.24</td>
</tr>
<tr>
<td>Homecentre</td>
<td><strong>10.75</strong></td>
<td><strong>10.06</strong></td>
</tr>
</tbody>
</table>

- **Homecenter Corpus**
  - development data for hand-crafted technologies
- **Domain of language: printer manual**
  - many imperative sentences
  - not common in WSJ text
Experiment 3: In-coverage only Results

- Effect of **change of domain** on automatically induced resources

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NIST</td>
<td>BLEU</td>
</tr>
<tr>
<td>WSJ Section 23</td>
<td>13.02</td>
<td>0.6511</td>
</tr>
<tr>
<td>Europarl (5-15)</td>
<td>11.72</td>
<td>0.6968</td>
</tr>
<tr>
<td>Europarl (all lengths)</td>
<td>10.24</td>
<td>0.5716</td>
</tr>
<tr>
<td>Homecentre</td>
<td>10.06</td>
<td>0.6640</td>
</tr>
</tbody>
</table>

**WSJ** newspaper text → Europarl parliamentary proceedings
### Experiment 3: In-coverage only Results

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NIST</td>
<td>BLEU</td>
<td>Coverage</td>
<td>NIST</td>
</tr>
<tr>
<td>WSJ Section 23</td>
<td>13.02</td>
<td>0.6511</td>
<td>99.96%</td>
<td></td>
</tr>
<tr>
<td>Europarl (5-15)</td>
<td>11.72</td>
<td>0.6968</td>
<td>100%</td>
<td>12.26</td>
</tr>
<tr>
<td>Europarl (all lengths)</td>
<td>10.24</td>
<td>0.5716</td>
<td>100%</td>
<td>12.1</td>
</tr>
<tr>
<td>Homecentre</td>
<td>10.06</td>
<td>0.6640</td>
<td>100%</td>
<td>10.81</td>
</tr>
</tbody>
</table>

- Treebank-induced technologies → better coverage
- Cannot compare NIST and BLEU scores here → different sentences when coverage not 100%
Conclusions

• Presented a **new method of evaluation** for parsing and generation technologies

• Useful for work that cites transfer-based MT as a **future application**

• Easily applied to any test set – **no gold standard** needed for this evaluation method

• Evaluation Method estimates the **upper bound** imposed by the quality of the TL parsing and generation technologies on a transfer-based MT system

• Provides a **realistic evaluation** in the context of transfer-based MT

• Means simple and quick means of investigating the **viability of a transfer-based MT system** given a particular pair of TL parsing / generation technologies
Thanks!