The University of Washington Machine Translation System for IWSLT 2006

Katrin Kirchhoff, Kevin Duh, Chris Lim
{katrin,duh,chrislim}@ee.washington.edu
University of Washington, Seattle
System Overview

- Multi-pass phrase-based statistical MT system

- Adding heterogeneous data
- Using ASR N-best / ConfusionNet as input
- Exploring new features
Outline

1. Basic System & Data
   • Data
   • 1st-pass system & features

2. 2nd-pass Rescoring (novel features)

3. Adding heterogeneous data


5. Official results and conclusions
Data

• Task: Italian-English open-data track
  • Input conditions: ASR-Output & Corrected transcriptions

• TRAIN SET:
  • BTEC training data + devset1,2,3 (190K words)
  • Europarl (European parliamentary proceedings)
    • (17M words) – for translation model
  • Fisher (Conversational telephone speech)
    • (2.3M words) – for 2\textsuperscript{nd} pass language models

• DEV SET:
  • devset4 – 350 sentences (to optimize 2\textsuperscript{nd}-pass rescorer)

• HELD-OUT SET:
  • devset4 – 139 sentences

Additional heterogeneous data
First-Pass Translation System

• Log-linear model:

\[ e^* = \arg \max_e p(e \mid f) = \arg \max_e \left\{ \sum_{k=1}^{K} \lambda_k \phi_k (e, f) \right\} \]

• Weights optimized on BLEU (minimum error rate training)
• Pharaoh decoder w/ monotone decoding
• 9 Features:
  • 2 phrase-based translation scores
  • 2 lexical translation scores
  • BTEC/Europarl data source indicator feature
  • word transition probability
  • phrase penalty
  • distortion penalty
  • language model score (3gram w/ KN smoothing, trained on BTEC)
Translation models

• 2 separate BTEC & Europarl phrase tables
  • Run GIZA++ and obtain heuristic alignments separately for each corpus
  • Decoder uses both phrase tables, without re-normalization of probabilities

Example:

\[
\begin{align*}
P(e1|f1) &= 0.4 \\
P(e2|f1) &= 0.6 \\
P(e1|f1) &= 0.1 \\
P(e3|f1) &= 0.9
\end{align*}
\]

From BTEC

From Europarl

• An additional binary feature indicates the data source
Outline

1. Basic System & Data
   - Data
   - 1st-pass system & features
   - Postprocessing

2. 2nd-pass Rescoring (novel features)

3. Adding heterogeneous data (Europarl, Fisher)


5. Official results and conclusions
2nd-pass Rescoring model

• Rescore N-best lists (N=2000max)

• Log-linear model, weights trained by downhill simplex to optimize BLEU

• 14 Features
  • 9 1st-pass model scores
  • 4-gram language model score
  • POS 5-gram score [mxpost tagger]
    • Rank in N-best list
    • Factored language model score ratio
    • Focused language model score
Rank in N-best list (2\textsuperscript{nd}-pass feature)

- Idea1: Leverage 1\textsuperscript{st}-pass decoder rankings in N-best

- Idea2: Hypotheses with same surface string should be tied together

**Rank feature**
- indicates rank of hypothesis in N-best
- ties together identical surface strings

**Example N-best list**
1. The store is open today \hspace{5em} \text{rank}=1
2. The store is open today \hspace{5em} \text{rank}=1
3. The shop is open now \hspace{5em} \text{rank}=2
4. The store is open today \hspace{5em} \text{rank}=1
5. The store it is open \hspace{5em} \text{rank}=3
Factored Language Model Ratio (2\textsuperscript{nd}-pass feature)

- Factored LM: flexible framework for incorporating diverse information (e.g. morphology, POS) [Bilmes&Kirchhoff03]
  - We model \( P(\text{word}_t|\text{word}_{t-1},\text{pos}_{t-1},\text{cluster}_{t-1}) \)
  & various backoffs e.g. \( P(\text{word}_t|\text{pos}_{t-1},\text{cluster}_{t-1}) \), \( P(\text{word}_t|\text{word}_{t-1}) \)

- Data-driven FLM backoff selection [Duh&Kirchhoff04]
  - Use a Genetic Algorithm search
  - FLM1: optimize on N-best oracle 1-best sentences
  - FLM2: optimize on N-best oracle worst sentences

- Feature score: \[
\frac{\log \text{prob}\{FLM_1(e)\}}{\log \text{prob}\{FLM_2(e)\}}
\]
  - Log-likelihood ratio: discriminate between good vs. bad sentences
Focused LM (2\textsuperscript{nd}-pass feature)

- **Motivation:** LM trained on BTEC (BTEC+Fisher) \textit{wastes probability mass} on words that never occur in the N-best list.
- **Solution:** train restricted-vocabulary n-grams

- **During N-best optimization:**
  1. Collect vocabulary from N-best lists (DEV set)
  2. Train n-gram on BTEC with restricted vocabulary
  3. Generate scores and optimize feature weight

- **During evaluation:**
  1. Collect vocabulary from N-best lists (EVAL set)
  2. Train \textit{new} n-gram on BTEC with restricted vocabulary
  3. Generate scores for rescoring

- **BIG Assumption:** optimal feature weight in training is suitable in testing

\begin{tabular}{|c|c|c|}
\hline
& LM vs. Focused LM (ASR-output) & LM vs. Focused LM (correct trans.) \\
\hline
DEV & +0.7 bleu & +1.2 bleu \\
HELD-OUT & +3.0 bleu & -1.7 bleu \\
\hline
\end{tabular}
### Rescoring Results on DEV set

#### Correct transcription task

<table>
<thead>
<tr>
<th></th>
<th>#f</th>
<th>BLEU</th>
<th>PER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rescoring w/ 1st-pass features</td>
<td>9</td>
<td>44.8</td>
<td>30.8</td>
</tr>
<tr>
<td>+4gram</td>
<td>10</td>
<td>44.9</td>
<td>31.0</td>
</tr>
<tr>
<td>+FLM</td>
<td>10</td>
<td>45.0</td>
<td>31.4</td>
</tr>
<tr>
<td>+focus</td>
<td>10</td>
<td>45.1</td>
<td>31.6</td>
</tr>
<tr>
<td>+pos</td>
<td>10</td>
<td>45.9</td>
<td>30.8</td>
</tr>
<tr>
<td>+rank</td>
<td>10</td>
<td>46.8</td>
<td>28.5</td>
</tr>
</tbody>
</table>

**Observations:**
- Rank is the strongest feature
- Combination of 14 features outperforms 1st-pass

#### ASR-output task

<table>
<thead>
<tr>
<th></th>
<th>#f</th>
<th>BLEU</th>
<th>PER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rescoring w/ 1st-pass features</td>
<td>9</td>
<td>34.6</td>
<td>39.6</td>
</tr>
<tr>
<td>Rescoring w/ ALL FEATURES</td>
<td>14</td>
<td>37.0</td>
<td>37.8</td>
</tr>
</tbody>
</table>
Outline

1. Basic System & Data
   • Data
   • 1\textsuperscript{st}-pass system & features
   • Postprocessing

2. 2\textsuperscript{nd}-pass Rescoring (novel features)

3. Adding heterogeneous data (Europarl, Fisher)


5. Official results and conclusions
Adding Europarl to 1\textsuperscript{st}-pass Translation Model (1/2)

• Does adding Europarl improve translation models, despite domain/style difference?
• Answer:
  • Yes, for correct transcription task
  • No, for ASR-output task
Adding Europarl to 1\textsuperscript{st}-pass Translation Model (1/2)

- **Does adding Europarl improve translation models, despite domain/style difference?**
- **Answer:**
  - Yes, for correct transcription task
  - No, for ASR-output task

### Phrase coverage (%) on DEV
[correct transcription task]

<table>
<thead>
<tr>
<th></th>
<th>BTEC</th>
<th>Europarl</th>
<th>Both</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>84.0</td>
<td>88.3</td>
<td>94.0</td>
</tr>
<tr>
<td>2</td>
<td>40.8</td>
<td>48.1</td>
<td>60.1</td>
</tr>
<tr>
<td>3</td>
<td>13.6</td>
<td>11.9</td>
<td>20.1</td>
</tr>
<tr>
<td>4</td>
<td>3.4</td>
<td>1.5</td>
<td>4.5</td>
</tr>
<tr>
<td>5</td>
<td>1.1</td>
<td>0.2</td>
<td>1.3</td>
</tr>
</tbody>
</table>

### 1\textsuperscript{st}-pass translation result on DEV
[correct transcription task]

<table>
<thead>
<tr>
<th></th>
<th>BLEU(%)</th>
<th>PER</th>
</tr>
</thead>
<tbody>
<tr>
<td>BTEC</td>
<td>44.5</td>
<td>29.9</td>
</tr>
<tr>
<td>Both</td>
<td><strong>46.8</strong></td>
<td><strong>28.0</strong></td>
</tr>
</tbody>
</table>
Adding Europarl to 1st-pass Translation Model (2/2)

• Does adding Europarl improve translation models, despite domain/style difference?
• Answer:
  • Yes, for correct transcription task
  • No, for ASR-output task

<table>
<thead>
<tr>
<th>Phrase coverage (%) on DEV [ASR-output task]</th>
<th>1st-pass translation result on DEV [ASR-output task]</th>
</tr>
</thead>
<tbody>
<tr>
<td>BTEC</td>
<td>Europarl</td>
</tr>
<tr>
<td>1 84.0</td>
<td>87.7</td>
</tr>
<tr>
<td>2 38.9</td>
<td>43.0</td>
</tr>
<tr>
<td>3 13.6</td>
<td>9.9</td>
</tr>
<tr>
<td>4 4.2</td>
<td>1.0</td>
</tr>
<tr>
<td>5 1.4</td>
<td>0.2</td>
</tr>
</tbody>
</table>
Adding Fisher to 2nd-pass Language Models

- Does additional conversational-style Fisher data improve (1) 4gram LM, (2) POS LM, (3) Focus LM?
- Answer:
  - No, in general
  - Yes, for Focus LM in correct transcription task (BLEU only)
  - Yes, for POS LM in ASR-output task

2nd-pass translation result on DEV

**[correct transcription task]**

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
<th>PER</th>
</tr>
</thead>
<tbody>
<tr>
<td>4gram LM</td>
<td>44.9</td>
<td>31.0</td>
</tr>
<tr>
<td>+ Fisher</td>
<td>44.8</td>
<td>31.0</td>
</tr>
<tr>
<td>POS LM</td>
<td>45.8</td>
<td>30.8</td>
</tr>
<tr>
<td>+ Fisher</td>
<td>45.9</td>
<td>30.8</td>
</tr>
<tr>
<td>Focus LM</td>
<td>44.4</td>
<td>31.3</td>
</tr>
<tr>
<td>+ Fisher</td>
<td>45.1</td>
<td>31.6</td>
</tr>
</tbody>
</table>

2nd-pass translation result on DEV

**[ASR-output task]**

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
<th>PER</th>
</tr>
</thead>
<tbody>
<tr>
<td>4gram LM</td>
<td>34.3</td>
<td>39.2</td>
</tr>
<tr>
<td>+ Fisher</td>
<td>34.1</td>
<td>39.6</td>
</tr>
<tr>
<td>POS LM</td>
<td>35.4</td>
<td>40.2</td>
</tr>
<tr>
<td>+ Fisher</td>
<td>35.7</td>
<td>40.0</td>
</tr>
<tr>
<td>Focus LM</td>
<td>35.2</td>
<td>39.8</td>
</tr>
<tr>
<td>+ Fisher</td>
<td>34.3</td>
<td>40.9</td>
</tr>
</tbody>
</table>
Outline

1. Basic System & Data
   • Data
   • 1st-pass system & features
   • Postprocessing
2. 2nd-pass Rescoring (novel features)
3. Adding heterogeneous data (Europarl, Fisher)
5. Official results and conclusions
ASR-outputs for machine translation

1. ASR 1-best → M-best translation hypotheses
2. ASR N-best → NxM-best translation hypotheses

3. Confusion Networks 1-best

- Idea: 1-best drawn from ConfusionNet may be more accurate than ASR 1-best
- [Post-evaluation] Significant DEV set improvement over ASR 1-best (37.0 vs. 38.0 BLEU)
Outline

1. Basic System & Data
   • Data
   • 1st-pass system & features
   • Postprocessing
2. 2nd-pass Rescoring (novel features)
3. Adding heterogeneous data (Europarl, Fisher)
5. Official results and conclusions
Official Results, (Rank)

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
<th>NIST</th>
<th>METEOR</th>
<th>WER</th>
<th>PER</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Correct Transcription Task</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Official</td>
<td>35.43 (2nd)</td>
<td>8.19 (1st)</td>
<td>70.17 (1st)</td>
<td>48.34</td>
<td>38.92</td>
</tr>
<tr>
<td>No case/punc</td>
<td>42.06 (1st)</td>
<td>9.24 (1st)</td>
<td>70.19 (1st)</td>
<td>42.86</td>
<td>31.75</td>
</tr>
<tr>
<td><strong>ASR-Output Task</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Official</td>
<td>27.87 (2nd)</td>
<td>6.93 (1st)</td>
<td>58.53 (1st)</td>
<td>55.87</td>
<td>46.76</td>
</tr>
<tr>
<td>No case/punc</td>
<td>31.68 (2nd)</td>
<td>7.69 (1st)</td>
<td>58.53 (1st)</td>
<td>53.17</td>
<td>42.11</td>
</tr>
</tbody>
</table>

Summary of submitted system:
1st pass Pharoah decoder
- Monotone decoding
- Translation table uses additional Europarl data
2nd pass Rescorer
- 14 features (incl. N-best rank, Factored LM, Focus LM)
Input for ASR-Output Task: 1-best ASR hypothesis
Conclusions

Adding heterogeneous data (Europarl, Fisher)
- Europarl helps TM for correct transcription task
- Fisher did not help LM in general

Exploring new features:
- Rank, Factored LM ratio, Focus LM
- 14 features beneficial in combination
- Rank alone gives large improvements

Using ASR N-best / ConfusionNet as input
- Direct translation of N-best not useful
- Confusion network 1-best is promising
THANKS!

Questions, suggestions, comments?
woof! ワン！bau!

UW Husky