The RWTH Statistical Machine Translation System for the IWSLT 2007 Evaluation

Arne Mauser, David Vilar, Gregor Leusch, Yuqi Zhang, Hermann Ney

{lastname}@i6.informatik.rwth-aachen.de

International Workshop on Spoken Language Translation (IWSLT) 2007
October 16, 2007

Human Language Technology and Pattern Recognition
Lehrstuhl für Informatik 6
Computer Science Department
RWTH Aachen University, Germany
Outline

RWTH System 2007

- Main facts
- Models
- System combination
- Systems and Results
Main Facts

Combination of different SMT system.
Each system: log-linear combination of a main translation model with several additional models.

Translation models
- Phrase-based model
- Bilingual $n$-gram model
- Hierarchical phrase model

Additional models
- Target language model
- Word lexicon
- Phrase count features
- Word penalty
- Phrase penalty
- Reordering model

No additional training data.
From 2006 to 2007: Improvements

- System combination on site (last year: across TC-Star project sites)
- More robust system combination
- New hierarchical phrase model
- Add syntactical reordering model

In terms of automatic measures for Chinese-English on IWSLT 2005 (dev3):

<table>
<thead>
<tr>
<th>System</th>
<th>BLEU [%]</th>
<th>NIST</th>
<th>WER [%]</th>
<th>PER [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>2004</td>
<td>40.4</td>
<td>8.59</td>
<td>52.4</td>
<td>42.2</td>
</tr>
<tr>
<td>2005</td>
<td>46.3</td>
<td>8.73</td>
<td>47.4</td>
<td>39.7</td>
</tr>
<tr>
<td>2006</td>
<td>48.8</td>
<td>8.56</td>
<td>47.3</td>
<td>39.2</td>
</tr>
<tr>
<td>2006 (40k)</td>
<td>51.4</td>
<td>9.00</td>
<td>40.0</td>
<td>33.2</td>
</tr>
<tr>
<td>2007</td>
<td>62.4</td>
<td>9.64</td>
<td>30.7</td>
<td>26.0</td>
</tr>
<tr>
<td>2007 comb.</td>
<td>63.4</td>
<td>10.14</td>
<td>30.8</td>
<td>25.3</td>
</tr>
</tbody>
</table>
Models
Hierarchical model

Generalization of phrase-based-models

- Allow for “gaps” in the phrases.
- Integration of reordering in the translation model.

Similar to Chiang’s approach

- Rules of the form \( X \rightarrow \langle \gamma, \alpha, \sim \rangle \), where
  - \( X \) is a non-terminal.
  - \( \gamma \) and \( \alpha \) are strings of terminals and non-terminals.
  - \( \sim \) is a one-to-one correspondence between the non-terminals of \( \alpha \) and \( \gamma \).

Here: Rules are transformed into GNF-like form to allow for left-to-right generation
Extraction Process

► Basic idea:
  ▶ Extract standard phrases.
  ▶ If the extracted phrases contain further sub-phrases, create “holes”.

► Main restrictions:
  ▶ Maximum of two non-terminals.
  ▶ Non-terminals must be non-adjacent in the source side.
  ▶ Rules must have at least a terminal symbol.
System Combination
System Combination

- Compute consensus translation by combining their outputs
- **Idea**: select words present in the majority of translations
  - Similar idea as ROVER for ASR
- Consider possible reordering of words/phrases
- High-quality alignment of different hypotheses required for voting
- Build confusion network from alignment
- Repeat Process $N$ times for $N$ systems
Improvements over the 2006 system combination

- Direct language model rescoring on the union of confusion networks
  - More efficient and better Bleu/TER score than N-best list rescoring
- use special language model in rescoring:
  - Train trigram LM on the outputs of all systems
  - Boost probability of $n$-grams present in the original phrases
- Optimize parameters automatically for Bleu
- Use 10-best lists from each system
Systems and Results
Italian-English Setup

Preprocessing

- lowercase and remove punctuation on the Italian training data
- bilingual $n$-gram model: keep punctuation in train and insert in test
- reduce case to the most frequent per word for English
- split punctuation and contractions: dell’albergo $\rightarrow$ dell’ albergo

Models and Training

- add dev1-3 to the training data
- same alignment for all models, refined heuristic
- use dev4 or dev5b for minimum error training of the systems
- used dev5a for tuning the system combination
- local or no reordering
- ASR: first-best
## Italian-English Results

<table>
<thead>
<tr>
<th>Model</th>
<th>opt-corpus</th>
<th>reorder</th>
<th>BLEU[%]</th>
<th>TER[%]</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Phrase-based</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dev4</td>
<td>no</td>
<td>41.6</td>
<td>44.5</td>
<td></td>
</tr>
<tr>
<td>local</td>
<td>41.7</td>
<td>44.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>dev5b</td>
<td>no</td>
<td>42.9</td>
<td>43.0</td>
<td></td>
</tr>
<tr>
<td>local</td>
<td>42.8</td>
<td>43.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Hierarchical</strong></td>
<td>dev5b</td>
<td></td>
<td>42.5</td>
<td>43.7</td>
</tr>
<tr>
<td><strong>n-gram</strong></td>
<td>dev4</td>
<td>no</td>
<td>33.5</td>
<td>50.5</td>
</tr>
<tr>
<td><strong>System Combination</strong></td>
<td></td>
<td></td>
<td>45.3</td>
<td>41.4</td>
</tr>
</tbody>
</table>

*n*-gram model performed comparably on dev sets → error in phrase extraction

### 2.4 Bleu-points improvement by system combination
Chinese-English Setup

Preprocessing and Training

► Chinese word segmentation using ICTClas or given segmentation
► Split punctuation marks and contractions: he’ll → he will

Models and Training

► Different alignment training
  ▶ Number of wordclasses in GIZA++ training
  ▶ Models used (HMM, HMM+IBM4)
► Phrase Extraction
  ▶ different heuristic
  ▶ Use minimum-weight edge cover algorithm on HMM and IBM4 probabilities
► All systems optimized on dev2 (IWSLT 2004)
► Optimization criterion: Bleu with minimum nearest/average reference length
Chinese-English - Results

Two combinations submitted (the wrong one as primary)

<table>
<thead>
<tr>
<th>Model</th>
<th>Alignment</th>
<th>BLEU[%]</th>
<th>TER[%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phrase-based</td>
<td>IBM4, MWEC-Alignment</td>
<td>37.2</td>
<td>48.0</td>
</tr>
<tr>
<td></td>
<td>HMM, MWEC-Alignment</td>
<td>36.7</td>
<td>48.4</td>
</tr>
<tr>
<td></td>
<td>HMM, refined</td>
<td>34.7</td>
<td>52.8</td>
</tr>
<tr>
<td>+syntax-reordering</td>
<td>IBM4, MWEC-Alignment</td>
<td>33.6</td>
<td>54.2</td>
</tr>
<tr>
<td>Hierarchical</td>
<td>IBM4, refined</td>
<td>33.3</td>
<td>51.4</td>
</tr>
<tr>
<td>System Combination</td>
<td></td>
<td>38.5</td>
<td>47.2</td>
</tr>
</tbody>
</table>

Hierarchical system not fully optimized

► Recent results: Bleu: 35.0 TER: 50.5
## Results

### Evaluation submissions

<table>
<thead>
<tr>
<th>Translation Direction</th>
<th>Input</th>
<th>Accuracy Measures</th>
<th>Error Rates</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>BLEU [%]</td>
<td>NIST</td>
</tr>
<tr>
<td><strong>Italian-to-English</strong></td>
<td>Clean</td>
<td>45.3</td>
<td>8.21</td>
</tr>
<tr>
<td></td>
<td>ASR</td>
<td>41.3</td>
<td>7.74</td>
</tr>
<tr>
<td><strong>Chinese-to-English</strong></td>
<td>Correct</td>
<td>37.1</td>
<td>6.75</td>
</tr>
<tr>
<td></td>
<td>Correct</td>
<td>38.5</td>
<td>6.80</td>
</tr>
</tbody>
</table>
New approaches

Combination of different translation systems / models

Robust, on-site system combination

good results on both language pairs (Italian-English, Chinese-English)

Hierarchy phrase model

Use syntax

Summary
Thank you for your attention

Arne Mauser

mauser@i6.informatik.rwth-aachen.de

http://www-i6.informatik.rwth-aachen.de/
Translation Models

Phrase-Based Model

▶ standard phrase-based model
▶ training: bilingual phrase pairs extracted from word-aligned training data
▶ main features: log probabilities $p(\tilde{f}|\tilde{e})$ and $p(\tilde{e}|\tilde{f})$ estimated by relative frequencies
Bilingual \( n \)-gram model

- model joint probability of source and target sentence \( Pr(f^J_1, e^I_1) \)
- segment the source and target sentences with the same restrictions given for the phrase-based model
- find the smallest units such that the resulting phrase segmentation is monotonic

extracted sentence:

```plaintext
ci_sono|are_there partite_di_baseball|any_baseball_games oggi|today → train smoothed \( n \)-gram language model [?]
```
Chiang’s approach

► Formalization as a synchronous CFG.
► Rules of the form $X \rightarrow \langle \gamma, \alpha, \sim \rangle$, where
  ▶ $X$ is a non-terminal.
  ▶ $\gamma$ and $\alpha$ are strings of terminals and non-terminals.
  ▶ $\sim$ is a one-to-one correspondence between the non-terminals of $\alpha$ and $\gamma$.
► Example:

\[
X \rightarrow \langle \text{yu } X_1 \text{ you } X_2, \text{ have } X_2 \text{ with } X_1 \rangle \\
X \rightarrow \langle X_1 \text{ de } X_2, \text{ the } X_2 \text{ that } X_1 \rangle \\
\]

► Additionally: Glue rules

\[
X \rightarrow \langle S_1 X_2, S_1 X_2 \rangle \\
X \rightarrow \langle X_1, X_1 \rangle \\
\]
Phrase with two gaps
Experiment Results

IWSLT 2007, Chinese-to-English task

<table>
<thead>
<tr>
<th>System</th>
<th>BLEU</th>
<th>TER</th>
<th>WER</th>
<th>PER</th>
</tr>
</thead>
<tbody>
<tr>
<td>PBT monoton</td>
<td>29.6</td>
<td>56.0</td>
<td>58.3</td>
<td>48.9</td>
</tr>
<tr>
<td>best PBT</td>
<td>37.2</td>
<td>48.0</td>
<td>48.7</td>
<td>44.3</td>
</tr>
<tr>
<td>Hierarchical</td>
<td>35.0</td>
<td>50.5</td>
<td>51.3</td>
<td>46.4</td>
</tr>
</tbody>
</table>

Example translations:

<table>
<thead>
<tr>
<th></th>
<th>The sightseeing. Where is it?</th>
</tr>
</thead>
<tbody>
<tr>
<td>PBT</td>
<td>The sightseeing. Where is it?</td>
</tr>
<tr>
<td>Hierarchical</td>
<td>Where is the tourist information office?</td>
</tr>
<tr>
<td>Reference</td>
<td>Where is the tourist information office?</td>
</tr>
<tr>
<td>PBT</td>
<td>Please tell me about this form, I do not know what to fill out?</td>
</tr>
<tr>
<td>Hierarchical</td>
<td>Please tell me how to fill out this form?</td>
</tr>
<tr>
<td>Reference</td>
<td>Can you tell me how to fill out this form?</td>
</tr>
</tbody>
</table>
Phrase count features

Motivation:

- rare phrases are overestimated
- estimated probabilities not reliable

Idea:

- adjust probabilities of rare phrases
- “mark” phrases with a occurrence count below a given threshold
- include these marker as a binary feature in the log-linear translation model

\[ h_{C,\tau}(f_1^J, e_1^I, s_1^K) = \sum_{k=1}^{K} [N(\tilde{f}_k, \tilde{e}_k) \leq \tau] \]

\( \tau \): threshold, \( N(\tilde{f}_k, \tilde{e}_k) \): bilingual phrase count, \( i_1^K \): segmentation of the source sentence
Algorithm: Idea

- Align different MT system outputs for each source sentence:
  - Allow word reordering
  - Take context of whole test document into account
  - Get more reliable alignment by using an iterative alignment procedure

- Construct confusion network from the reordered hypotheses

- Use system prior probabilities and other statistical models to select consensus translation from network
Alignment and Reordering

Alignment

- Pairwise alignment of the output of $M$ systems for $N$ test sentences
- IBM Model 1 and HMM alignment models

Note: Alignment can be improved by adding more translated data from involved MT systems

Reordering

- Select a primary hypothesis $E_m$
- Reorder each other MT output $E_n$ based on alignment with $E_m$
- Construction of confusion network from alignments
Example: Confusion Network
### Example

| system hypotheses | 0.25 would your like coffee or tea  
|                   | 0.35 have you tea or coffee  
|                   | 0.10 would like your coffee or  
|                   | 0.30 I have some coffee tea would you like  
| alignment and reordering | have would you your $ like coffee coffee or or tea tea  
|                       | would would your your like like coffee coffee or or $ tea  
|                       | I $ would would you your like like have $ some $ coffee coffee $ or tea tea  
| confusion network | $ would your like $ $ coffee or tea  
|                       | $ have you $ $ coffee or tea  
|                       | I would your like have some coffee $ tea  
| voting | $/0.7 would/0.65 you/0.65 $/0.35 $/0.7 $/0.7 coffee/1.0 or/0.7 tea/0.9  
|         | l/0.3 have/0.35 your/0.35 like/0.65 have/0.3 some/0.3 $/0.3 $/0.1  
| consensus translation | would you like coffee or tea  