Speech Translation: from Single-best to N-Best to Lattice Translation

Ruiqiang ZHANG                 Genichiro KIKUI

Spoken Language Communication Laboratories
Speech Translation Structure

- Single-best only
  - ASR \rightarrow \text{Single-best} \rightarrow \text{MT}

- N-best hypothesis translation
  - ASR \rightarrow \text{N-best} \rightarrow \text{MT}

- Word lattice
  - ASR \rightarrow J \rightarrow \text{Word lattice} \rightarrow \text{WLT} \rightarrow E
References

- Saleem (ICSLP 2004). Using word lattice information for a tight coupling in speech translation systems.
- Matusov (Eurospeech 2005). On the Integration of Speech Recognition and SMT.
Outline

- N-best translation
- Word lattice translation
- IWSLT 2005 evaluation
- Conclusions
N-best Hypothesis Translation

- **J₁, J₂ and J₇**: N-best speech recognition hypotheses
- **E₂,1, E₂,2, ..., E₂,K**: K-best translation hypotheses produced from J₂
- **Rescore**: to rescore all NxK translations
Rescore: Integration of ASR and SMT

- Statistical theory

\[
\hat{E} = \arg \max_E P(E|X) = \arg \max_E P(E)P(X|E) = \arg \max_E \left\{ P(E) \sum_J P(X, J|E) \right\} = \arg \max_E \left\{ P(E) \sum_J P(X|J)P(J|E) \right\}
\]

- Make approximations

\[
\langle \hat{E}, J \rangle = \arg \max_{E,J} \{ P(E)P(X \mid J)P(J \mid E) \}
\]
Rescore: Log-linear models

\[ \hat{E} = \arg \max_E \sum_{m=1}^{M} \lambda_m \log P_m(X, E) \]

\[ P(E \mid X) = \frac{\exp\left(\sum_{m=1}^{M} \lambda_m f_m(X, E)\right)}{\sum_{E'} \exp\left(\sum_{m=1}^{M} \lambda_m f_m(X, E)\right)} \]

\( E \) : all possible translation hypotheses

\( P_m(X, E) \) : m-th feature in log value

\( \lambda \) : weight of each feature
Parameter optimization

Objective function:

\[ \lambda^M = \text{optimize } D(R_s, \hat{E}_s) \]

\( \hat{E}_s \) translation output after log-linear model rescoring

\( R_s \) References of English sentences. 16 reference sentences for each English sentence

\( D(R_s, \hat{E}_s) \) Automatic translation quality metrics. BLEU, NIST, mWER and mPER
Translation Assessment $D(R_s, \hat{E}_s)$

- **N-gram methods**
  - **BLEU**: A weighted geometric mean of the n-gram matches between test and reference sentences plus a short sentence penalty
  - **NIST**: An arithmetic mean of the n-gram matches between test and reference sentences

- **Word error rate**
  - **mWER**: multiple reference word error rate.
  - **mPER**: multiple reference position independent word error rate
Optimization: Direction Set Methods

- Change initial lambda
- Change direction
- Local optimization
- Local lambda
- Best lambda

$D(R_s, \hat{E}_s)$
Features from ASR

- Acoustic Model (AM) scores
  - Gaussian mixture output probability density function (pdf)

- Language Model (LM) scores
  - N-gram language model
Features from Phrase-based SMT

- Target language model (trigram)
- Target class language model: SRILM cluster (5-gram)
- Target phrase language model:
- Phrase translation model:
- Distortion model:
- Length model:
- NULL word translation model:
- Jump model:
- Long distance target LM: (9-gram) for rescore
- Long distance class LM: (11-gram)
An Experimental Results of N-best Translation

![Bar chart showing BLEU scores for different numbers of hypotheses.]

- 1 hypothesis: 0.55
- 3 hypotheses: 0.56
- 5 hypotheses: 0.57
- 10 hypotheses: 0.58
- 50 hypotheses: 0.59
- 100 hypotheses: 0.6
Word Lattice Translation
Recognition Word Lattice

ASR First-best: 経験者を呼ぶでもらえますか
First-best translation: Could I get a job
ASR correct recognition: 救急車を呼ぶでもらえますか
Word Lattice translation: Could you call an ambulance
could you call an ambulance
Machine Translation for Lattice

NULL model

Fertility model
Machine Translation for Lattice

Lexical model

Distortion model

経験者
救急車

を 呼ぶ で もらえ ます
か
How We Translate Word Lattice

- Two-step decoding: beam-search + A* search
- Beam search: construct translation word graph (TWG)
  - An edge in the word lattice is mapped to an edge in the TWG
  - A path in the TWG corresponds to a path in the word lattice
  - Lower-scored edges are pruned.
  - Simple translation models are used.
- A* search:
  - Search the TWG with a higher-grade translation models (IBM model4)
Beam-search: threshold pruning
Translation Word Graph (example)
A* Search

- A* search
  - Forward score: Accumulated from the start node to current node, using IBM Model4 model
  - Heuristic score: Accumulated from the current node to the end node
- Approximations are made on the models dependent on the length of source sentence:
  - distribution model
  - NULL word
Features in Speech Translation Models

\[
\hat{E} = \arg \max_E \left\{ \lambda_0 \log P_{pp} + \lambda_1 \log P_{lm}(E) \\
+ \lambda_2 \log P_{lm}(POS(E)) + \lambda_3 \log P(\phi_0) \\
+ \lambda_4 \log N(\Phi \mid E) + \lambda_5 \log T(J \mid E) \\
+ \lambda_6 \log D(E, J) \right\}
\]
Effect of Word Lattice Translation

![Graph showing the effect of word lattice translation on BLEU score with varying Nbest in Lattice.]
Beam-size effect in WLT

1. A translation of 1\textsuperscript{st} best ASR hypotheses
2. A translation of 2\textsuperscript{nd} best ASR hypotheses
3. A translation of 3\textsuperscript{rd} best ASR hypotheses

Single-best translation

Word lattice translation
Beam-size Effects in WLT (N-best=20)

- Promising hypotheses pruned in WLT but saved in single-best translation under the same beam size

![Graph showing BLEU scores for 1-best and 20-best translations as beam size increases.](image)
Why Word Lattice Minimization

- Raw lattice is too huge
- A lot of duplicated word IDs in the lattice
- Significant are the top N-best hypotheses
- Minimization under the light of machine translation
- Minimization can make decoding fast
- Minimization can reduce translation error; reduce pruning error in decoding
Word Lattice Minimization

Raw SWL

Hypotheses

Downsized SWL

Transfer rules

- \( j_0 \rightarrow l_0(W_a) \)
- \( j_1, j_2 \rightarrow l_1(W_b) \)
- \( j_3, j_5 \rightarrow l_2(W_c) \)
- \( j_4 \rightarrow l_3(W_e) \)
- \( j_6, j_7 \rightarrow l_4(W_d) \)
Word Lattice ?? N-best ??

- After lattice minimization, the output is not a lattice again. Only N-best with new assigned edge ids.
- After lattice minimization, the ASR score lost in single edge. Instead, we use ASR path score to represent single edge’s score.
Effect of Lattice Minimization

#Nbest in lattice
Posterior Probability

- Integrating acoustic model and language model probabilities
- Indicating relative accuracy of N-best hypotheses

\[
p(J_j \mid X) = \frac{e^{\lambda \log \text{score}_j}}{\sum_{i=1}^{N} e^{\lambda \log \text{score}_i}}
\]

\( \log \text{score}_i \) : log-scale ASR score (AM+LM)
Confidence Measure Filtering

- ASR hypotheses with very low posterior probability degrade translations
- A predefined confidence threshold, T, is applied to remove the most unlikely ASR hypotheses
- By comparing a hypothesis’s posterior probability to the single-best hypothesis’s posterior probability multiplied by T, $P_{\text{first-best}} \cdot T$, remove the smaller.
## Confidence Measure Filtering

<table>
<thead>
<tr>
<th>ASR Output</th>
<th>ASR score</th>
<th>PP=Posterior probability</th>
<th>cmf=PP/PP_{1-st}</th>
<th>Decision cmf&gt;0.5?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st cand.</td>
<td>0.55</td>
<td>0.215</td>
<td>1</td>
<td>PASS</td>
</tr>
<tr>
<td>2nd cand.</td>
<td>0.50</td>
<td>0.196</td>
<td>0.912</td>
<td>PASS</td>
</tr>
<tr>
<td>3rd cand.</td>
<td>0.45</td>
<td>0.176</td>
<td>0.818</td>
<td>PASS</td>
</tr>
<tr>
<td>4th cand.</td>
<td>0.40</td>
<td>0.157</td>
<td>0.730</td>
<td>PASS</td>
</tr>
<tr>
<td>5th cand.</td>
<td>0.30</td>
<td>0.118</td>
<td>0.549</td>
<td>PASS</td>
</tr>
<tr>
<td>6th cand.</td>
<td>0.20</td>
<td>0.078</td>
<td>0.363</td>
<td>FAIL</td>
</tr>
<tr>
<td>7th cand.</td>
<td>0.10</td>
<td>0.039</td>
<td>0.181</td>
<td>FAIL</td>
</tr>
<tr>
<td>8th cand.</td>
<td>0.05</td>
<td>0.020</td>
<td>0.09</td>
<td>FAIL</td>
</tr>
</tbody>
</table>

SUM=2.25
Effect of CM filtering

![Graph showing the effect of CM filtering with various N-best and Lattice Min. metrics.](image)
IWSLT 2005 Evaluation
## IWSLT 2005 Evaluation (training data)

<table>
<thead>
<tr>
<th>Language pair</th>
<th>Data track</th>
<th>Data size</th>
<th>perplexity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Testset</td>
</tr>
<tr>
<td>C/E</td>
<td>Supplied +tagger</td>
<td>20K</td>
<td>65.4</td>
</tr>
<tr>
<td></td>
<td>C-star</td>
<td>172K</td>
<td>69.3</td>
</tr>
<tr>
<td>J/E</td>
<td>Supplied +tagger</td>
<td>20K</td>
<td>54.9</td>
</tr>
<tr>
<td></td>
<td>C-star</td>
<td>463K</td>
<td>22.5</td>
</tr>
</tbody>
</table>
Test Data Analysis

ASR Recognition Accuracy

Japanese

Chinese

N=1  N=20

N=1  N=20
Test Data Results (J/E BLEU C-star track)
Test Data Results (J/E NIST C-star track)
Test Data Results (J/E WER CSTAR track)
## Evaluation Results (CE)

<table>
<thead>
<tr>
<th>Data track</th>
<th>Input</th>
<th>BLEU</th>
<th>NIST</th>
<th>WER</th>
<th>PER</th>
<th>METEOR</th>
<th>GTM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supplied+tools</td>
<td>Text</td>
<td>0.305</td>
<td>7.20</td>
<td>0.607</td>
<td>0.494</td>
<td>0.574</td>
<td>0.471</td>
</tr>
<tr>
<td></td>
<td>Nbest</td>
<td>0.267</td>
<td>6.19</td>
<td>0.645</td>
<td>0.546</td>
<td>0.506</td>
<td>0.421</td>
</tr>
<tr>
<td></td>
<td>Sbest</td>
<td>0.251</td>
<td>5.93</td>
<td>0.683</td>
<td>0.581</td>
<td>0.479</td>
<td>0.395</td>
</tr>
<tr>
<td>Cstar</td>
<td>Text</td>
<td>0.421</td>
<td>8.17</td>
<td>0.518</td>
<td>0.422</td>
<td>0.642</td>
<td>0.547</td>
</tr>
<tr>
<td></td>
<td>Nbest</td>
<td>0.375</td>
<td>6.80</td>
<td>0.561</td>
<td>0.486</td>
<td>0.560</td>
<td>0.493</td>
</tr>
<tr>
<td></td>
<td>Sbest</td>
<td>0.340</td>
<td>6.76</td>
<td>0.619</td>
<td>0.525</td>
<td>0.531</td>
<td>0.461</td>
</tr>
</tbody>
</table>
# Evaluation Results (JE)

<table>
<thead>
<tr>
<th>Data track</th>
<th>Input</th>
<th>BLEU</th>
<th>NIST</th>
<th>WER</th>
<th>PER</th>
<th>METEOR</th>
<th>GTM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supplied+tagger</td>
<td>Text</td>
<td>0.388</td>
<td>4.39</td>
<td>0.563</td>
<td>0.519</td>
<td>0.520</td>
<td>0.431</td>
</tr>
<tr>
<td></td>
<td>Nbest</td>
<td>0.383</td>
<td>4.27</td>
<td>0.574</td>
<td>0.530</td>
<td>0.513</td>
<td>0.422</td>
</tr>
<tr>
<td></td>
<td>Lattice</td>
<td>0.378</td>
<td>4.18</td>
<td>0.578</td>
<td>0.534</td>
<td>0.511</td>
<td>0.420</td>
</tr>
<tr>
<td></td>
<td>Sbest</td>
<td>0.366</td>
<td>4.50</td>
<td>0.576</td>
<td>0.527</td>
<td>0.508</td>
<td>0.412</td>
</tr>
<tr>
<td>Cstar</td>
<td>Text</td>
<td>0.727</td>
<td>10.94</td>
<td>0.289</td>
<td>0.243</td>
<td>0.80</td>
<td>0.716</td>
</tr>
<tr>
<td></td>
<td>Nbest</td>
<td>0.679</td>
<td>10.04</td>
<td>0.324</td>
<td>0.281</td>
<td>0.760</td>
<td>0.670</td>
</tr>
<tr>
<td></td>
<td>Lattice</td>
<td>0.670</td>
<td>9.86</td>
<td>0.329</td>
<td>0.289</td>
<td>0.763</td>
<td>0.665</td>
</tr>
<tr>
<td></td>
<td>Sbest</td>
<td>0.646</td>
<td>9.68</td>
<td>0.352</td>
<td>0.304</td>
<td>0.741</td>
<td>0.645</td>
</tr>
</tbody>
</table>
Remarks

- Text translation (0.727) > N-best translation (0.679)
- N-best translation (0.679) > lattice translation (0.67)
- Lattice translation (0.670) > single-best translation (0.646)

- Training data size influences speech translation
Analysis: Lattice Translation Worse than N-best Translation

- We used the same number of ASR hypotheses in N-best translation and lattice translation.
- In beam search, N-best translation and lattice translation used the same beam size and threshold in pruning.
- Model approximations and inaccuracy: distortion, null, acoustic model, language model.
Comparisons of the structures

- **Single-best translation**
  - ✔ Simple, direct
  - ✔ ASR and SMT isolated optimization
  - ✔ MT flexible, easy to upgrade, multiple translation engines
  - ✗ Non-robust to ASR WER

- **N-best hypothesis translation**
  - ✔ Robust, resistant to ASR WER
  - ✔ MT flexible, multiple translation engines
  - ✗ Slow, duplicate calculation

- **Word lattice translation**
  - ✔ Reduce computing cost, efficient
  - ✔ Speech translation system, ASR and SMT, overall optimized
  - ✗ MT inflexible
Conclusions

- We applied two approaches to improve ASR single-best translation.
- By applying a log-linear model, N-best translation approach can improve single-best translation effectively.
- We observed improved speech translation performance in word lattice translation:
  - Confidence measure filtering
  - Word lattice reduction