Improved Statistical Models for SMT-Based Speaking Style Transformation

Graham Neubig, Yuya Akita, Shinsuke Mori, Tatsuya Kawahara

School of Informatics, Kyoto University, Japan
1. Overview of Speaking-Style Transformation
Speaking Style Transformation (SST)

- ASR is generally modeled to find the verbatim utterance $V$ given acoustic features $X$
- In many cases verbatim speech is difficult to read:
  ya know when I was asked earlier about uh the issue of coal uh you under my plan uh of a cap and trade system ...
- In order to create usable transcripts from ASR results, it is necessary to transform $V$ into clean text $W$
  When I was asked earlier about the issue of coal under my plan of a cap and trade system, ...
Previous Research

• Detection-Based Approaches
  • Focus on deletion of fillers, repeats, and repairs, as well as insertion of punctuation
  • Modeled using noisy-channel models [Honal & Schultz 03, Maskey et al. 06], HMMs, and CRFs [Liu et al. 06]

• SMT-Based Approaches
  • Treat spoken and written language as different languages, and “translate” between them
  • Proposed by [Shitaoka et al. 04] and implemented using WFSTs and log-linear models in [Neubig et al. 09]
  • Is able to handle colloquial expression correction, insertion of dropped words (important for formal settings)
Research Summary

- Propose **two enhancements of the statistical model** for finite-state SMT-based SST
  - **Incorporation of context** in a noisy channel model by transforming context-sensitive joint probabilities to conditional probabilities
  - Allowing **greater emphasis on frequent patterns** by log-linearly interpolating joint and conditional probability models
- Evaluation of the proposed methods on both verbatim transcripts and ASR output for the Japanese Diet (national congress)
2. Noisy-Channel and Joint-Probability Models for SMT
Noisy Channel Model

- Statistical models for SST attempt to maximize $P(W|V)$
- Training requires a parallel corpus of $W$ and $V$
  - It is generally easier to acquire a large volume of clean transcripts ($W$) than a parallel corpus ($W$ and $V$)
  - Bayes' law is used to decompose the probabilities

$$\hat{W} = \arg\max_{W} P(W|V)$$

$$= \arg\max_{W} P_t(V|W) P_l(W)$$

- Translation Model (TM)
- Language Model (LM)

- $P_l(W)$ is estimated using an $n$-gram (3-gram) model
Probability Estimation for the TM

- $P_t(V|W)$ is difficult to estimate for the whole sentence
  - Assume that the word TM probabilities are independent
  - Set the sentence TM probability equal to the product of the word TM probabilities
    \[
    P_t(V|W) \approx \prod_i P_t(v_i|w_i)
    \]
- However, the word TM probabilities are actually not context independent

I told him that I really like his new hairstyle.

$P_t(\text{like}|\varepsilon)$

$P_t(\text{like}|\varepsilon, H_1)$ (large)

$P_t(\text{like}|\varepsilon, H_2)$ (small)
Joint Probability Model
[Casacuberta & Vidal 2004]

- The joint probability model is an alternative to the noisy-channel model for speech translation

\[ \hat{W} = \arg\max_{W} P_t(W, V) \]

- Sentences are aligned into matching words or phrases

\[
\begin{array}{ccccccccc}
V = & \text{ironna} & e- & \text{koto} & de & \text{chumon} & \text{tsukeru} & to & \text{desu} & ne & \ldots \\
W = & \text{iroyo na} & \text{koto} & de & \text{chumon} & o & \text{tsukeru} & to & \ldots \\
\end{array}
\]

- A sequence \( \Gamma \) of word/phrase pairs is created

\[
\Gamma = \text{ironna/iroyo na e-} \varepsilon \text{koto/koto de/de chumon/chumon } \varepsilon/o \text{tsukeru/tsukeru to/to desu/} \varepsilon \text{ ne/} \varepsilon
\]
Joint Probability Model (2)

- The probability of $\Gamma$ is estimated using a smoothed $n$-gram model trained on $\Gamma$ strings:

$$P_t(W, V) = P_t(\Gamma) \approx \prod_{k=1}^{K} P_t(\gamma_k | \gamma_{k-n+1}^{k-1})$$

- Context information is contained in the joint probability

- However, this probability can only be trained on parallel text (an LM probability cannot be used)

$$\arg\max_{W} P_t(W|V) \neq \arg\max_{W} P_t(W, V)P_l(W)$$

- It is desirable to have a context-sensitive model that can be used with a language model
3. A Context-Sensitive Translation Model
Context-Sensitive Conditional Probability

- It is possible to model the conditional (TM) probability from right-to-left, similarly to the joint probability

\[
P_t(V|W) = \prod_{i=1}^{k} P_t(v_i|v_1, \ldots, v_{i-1}, w_1, \ldots, w_k)
\]

\[
= \prod_{i=1}^{k} P_t(v_i|y_1, \ldots, y_{i-1}, w_i, \ldots, w_k)
\]
Independence Assumptions

- To simplify the model, we make two assumptions
  - Assume that word probabilities rely only on preceding words

\[
P_t(V|W) \approx \prod_{i=1}^{k} P_t(v_i|\gamma_1, \ldots, \gamma_{i-1}, w_i)
\]

- Limit the history length

\[
P_t(V|W) \approx \prod_{i=1}^{k} P_t(v_i|\gamma_{i-n+1}, \ldots, \gamma_{i-1}, w_i)
\]
Calculating Conditional Probabilities from Joint Probabilities

- It is possible to decompose this equation into its numerator and denominator

\[
P_t(v_i | y_{i-n+1}, \ldots, y_{i-1}, w_i) = \frac{P_t(y_i | y_{i-n+1}, \ldots, y_{i-1})}{P_t(w_i | y_{i-n+1}, \ldots, y_{i-1})}
\]

- The numerator is equal to the joint \( n \)-gram probability, while the denominator can be marginalized

\[
P_t(v_i | y_{i-n+1}, \ldots, y_{i-1}, w_i) = \frac{P_t(y_i | y_{i-n+1}, \ldots, y_{i-1})}{\sum_{\tilde{y} \in \{\tilde{y} : (\tilde{v}, w_i)\}} P_t(\tilde{y} | y_{i-n+1}, \ldots, y_{i-1})}
\]

- This conditional probability uses context information and can be combined with a language model
Training the Proposed Model

Clean Corpus

- Clean Transcripts (W)

Parallel Corpus

- Clean Transcripts (W)
- Verbatim Transcripts or ASR Results (V)

Train

- $P_t(W, V)$ Joint Prob.

Calculate

- $P_t(V|W)$ Context-Sensitive TM

Noisy-Channel Model

- $P_t(W)$ LM

Clean Corpus

- Train $P_t(W)$
Log-Linear Interpolation with the Joint Probability

- The joint probability contains information about pattern frequency not present in the conditional probability

\[
c(\gamma_1) = 100 \quad c(\gamma_2) = 1 \quad P_t(v_1|w_1) = P_t(v_2|w_2)
\]
\[
c(w_1) = 1000 \quad c(w_2) = 10 \quad P_t(\gamma_1) \neq P_t(\gamma_2)
\]

- High-frequency patterns are more reliable
- The strong points of both models can be utilized through log-linear interpolation

\[
\log(P(W|V)) \propto \lambda_1 \log(P_t(V|W)) + \lambda_2 \log(P_t(W)) + \lambda_3 \log(P_t(V,W))
\]

Noisy-Channel Model

Joint Probability
Training the Proposed Model

Clean Corpus

- Clean Transcripts (W)

Parallel Corpus

- Clean Transcripts (W)
- Verbatim Transcripts or ASR Results (V)

Train

- $P_t(W, V)$ Joint Prob.

- $P_t(V|W)$ Context-Sensitive TM

Calculate

- $P_t(W|V)$

Log-Linear Model

- $P_t(W)$
- $P_t(V|W)$

$\lambda_1$, $\lambda_2$, $\lambda_3$
4. Evaluation
Experimental Setup

- Verbatim transcripts and ASR output of meetings from the Japanese Diet were used as a target

<table>
<thead>
<tr>
<th>Data Type</th>
<th>Size</th>
<th>Time Period</th>
</tr>
</thead>
<tbody>
<tr>
<td>TM Training</td>
<td>2.31M</td>
<td>1/2003 - 10/2006</td>
</tr>
<tr>
<td>Weight Training</td>
<td>66.3k</td>
<td>10/2006-12/2006</td>
</tr>
<tr>
<td>Testing</td>
<td>300k</td>
<td>10/2007</td>
</tr>
</tbody>
</table>

- TM training:
  - Verbatim system: Verbatim transcripts and clean text
  - ASR system: **ASR output and clean text**
  - Baseline: noisy channel, 3-gram LM, 1-gram TM
Improved Statistical Models for SMT-Based Speaking Style Transformation

Effect of Translation Models (Verbatim Transcripts)

• 4 models were compared
  
  A) The context-sensitive noisy-channel model
  B) A with log-linear interpolation of the LM and TM
  C) The joint-probability model
  D) B and C log-linearly interpolated

• Evaluated using edit distance from the clean transcript (WER), with no editing, the WER was 18.62%

<table>
<thead>
<tr>
<th>Model</th>
<th>LL</th>
<th>TM n-gram order</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1-gram</td>
</tr>
<tr>
<td>A. Noisy-Channel (Noisy)</td>
<td>6.51%</td>
<td>5.33%</td>
</tr>
<tr>
<td>B. Noisy-Channel (Noisy LL)</td>
<td>★</td>
<td>5.99%</td>
</tr>
<tr>
<td>C. Joint Probability (Joint)</td>
<td>9.89%</td>
<td>4.70%</td>
</tr>
<tr>
<td>D. B+C (Noisy+Joint LL)</td>
<td>★</td>
<td>5.81%</td>
</tr>
</tbody>
</table>
Effect of Translation Models (ASR Output)

- The WER between ASR output and verbatim transcripts (ASR WER) was 17.10%
- ASR output and clean transcripts was 36.10%

<table>
<thead>
<tr>
<th>Model</th>
<th>LL</th>
<th>TM n-gram Order</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1-gram</td>
</tr>
<tr>
<td>A. Noisy-Channel (Noisy)</td>
<td>21.83%</td>
<td>21.00%</td>
</tr>
<tr>
<td>B. Noisy-Channel (Noisy LL)</td>
<td>★ 21.63%</td>
<td>20.97%</td>
</tr>
<tr>
<td>C. Joint Probability (Joint)</td>
<td>28.61%</td>
<td>22.62%</td>
</tr>
<tr>
<td>D. B+C (Noisy+Joint LL)</td>
<td>★ 21.32%</td>
<td>20.04%</td>
</tr>
</tbody>
</table>

- The noisy-channel model was more effective than the joint-probability model for ASR output
Comparison with Phrase-Based SMT (New Results)

- The proposed techniques were also compared with Moses, a popular system for phrase-based SMT.

<table>
<thead>
<tr>
<th>Model</th>
<th>Verbatim WER</th>
<th>ASR WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>6.51%</td>
<td>21.83%</td>
</tr>
<tr>
<td>Noisy LL (2-gram or 3-gram)</td>
<td>5.13%</td>
<td>20.97%</td>
</tr>
<tr>
<td>Noisy+Joint (2-gram or 3-gram)</td>
<td>4.05%</td>
<td>20.03%</td>
</tr>
<tr>
<td>Moses</td>
<td>5.45%</td>
<td>20.97%</td>
</tr>
</tbody>
</table>

- **Noisy LL** is able to achieve performance as good or better than **Moses**, while **Noisy+Joint** greatly outperforms it.
Effect of Corpus Size (Verbatim Transcripts)

- The noisy-channel model is more effective with small data sizes, but the joint model improves rapidly.

- Combining both allows for greater accuracy at all sizes.
Conclusion

• We proposed two improved statistical models for SMT-based SST

• The proposed methods showed a significant improvement over the baseline for verbatim transcripts and ASR results

• Models transforming ASR output can be trained without using verbatim transcripts

• A promising future direction is tight coupling with a WFST-based ASR decoder
Thank you for listening.
Target Phenomena

- **Deletion of Extraneous Words:** These include fillers (“um”), context-dependent deletions (“like”), repeats

- **Colloquial Expressions:** Expressions used in speech but less in writing (“ya'know” → “you know”, “ironna” → “iroiro-na”)

- **Insertion of Words and Punctuation:** Words are omitted in speech, but not in writing (“[did you] talk to the boss?”, “chumon [o] tsukeru”)

<table>
<thead>
<tr>
<th>various</th>
<th>ahh</th>
<th>things by order</th>
<th>-obj</th>
<th>make</th>
<th>if</th>
<th>it is</th>
</tr>
</thead>
<tbody>
<tr>
<td>いろんな</td>
<td>あー</td>
<td>こと で 注文</td>
<td>つける と</td>
<td>です ね</td>
<td>desu ne</td>
<td></td>
</tr>
<tr>
<td>ironna</td>
<td>a-</td>
<td>koto de chu-mon</td>
<td>tsukeru to</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>いろいろな</td>
<td>な</td>
<td>こと で 注文</td>
<td>つける と</td>
<td>す で</td>
<td>non-fill</td>
<td></td>
</tr>
<tr>
<td>iroiro</td>
<td>na</td>
<td>koto de chu-mon</td>
<td>to</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **Other Phenomena:** order reversal, repairs, fragments
Effect of Corpus Size (ASR Results)

- Baseline
- Noisy LL 3-gram
- Joint 3-gram
- Noisy/Joint 3-gram

Words in TM Training Data

Wer (%) vs Words in TM Training Data

- 27
- 3.2k
- 4.9k
- 8.7k
- 17k
- 35k
- 66k
- 143k
- 287k
- 573k
- 1.14M
- 2.32M

19.5% 20.5% 21.5% 22.5% 23.5% 24.5% 25.5%
Accuracy by Transformation Type (Verbatim Transcript)

- **Fillers**
- **Deletions**
- **Insertions**
- **Substitutions**
- **Commas**
- **Periods**

Legend:
- Noisy LL 1
- Joint 3
- Noisy LL 3
- Noisy+Joint LL 3
Accuracy by Transformation Type (ASR Output)

- **Fillers**
- **Deletions**
- **Insertions**
- **Substitutions**
- **Commas**
- **Periods**

Accuracy Chart:
- Noisy LL 1
- Joint 3
- Noisy LL 3
- Noisy+Joint LL 3

F-Measure:
- Noisy LL 1: 90%
- Joint 3: 80%
- Noisy LL 3: 70%
- Noisy+Joint LL 3: 60%