**On Language Model Integration for RNN Transducer based Speech Recognition**

Weizhou Zhou, Zuoyun Zheng, Ralf Schütter, Hermann Ney

**Human Language Technology and Pattern Recognition, RWTH Aachen University, Aachen, Germany**

**Paper ID:** 4436

---

**References**


---

**Decoding Interpretation & Verification**

2 Major Reasons: Improvement with ILM Correction

- R1. prior removal rebalances label distribution of \( P_{\text{RNN}} \) to rely more on external LM for context modeling (desired)
- R2. division by \( P_{\text{ILM}} \) boosts the label probability against (usually high) blank probability \( \rightarrow \) increase importance of external LM (\( \lambda_j \)) without suffering huge deletion errors
- limitation of shallow fusion (\( \lambda_j = 0 \))
- no need of decoding heuristics, e.g. length reward

Experimental Verification

- verify effect of R2 without R1: shallow fusion + length reward
- verify effect of R1 without R2: \( P_{\text{ILM}} \) + re enr- \( P_{\text{RNN}} \)

for each \( y_i \neq \epsilon \):

\[
\text{search score: } -1 + \frac{P_{\text{RNN}}(y_i | R^{\text{ILM}}(y_i)) \cdot P_{\text{ILM}}(y_i | R^{\text{ILM}}(y_i))}{\sum_{y_i} P_{\text{RNN}}(y_i | R^{\text{ILM}}(y_i)) \cdot P_{\text{ILM}}(y_i | R^{\text{ILM}}(y_i))}
\]

Evaluation

<table>
<thead>
<tr>
<th>( \lambda_j )</th>
<th>( \lambda_i )</th>
<th>Librispeech WER %</th>
<th>Sub</th>
<th>Del</th>
<th>Ins</th>
</tr>
</thead>
<tbody>
<tr>
<td>shallow fusion</td>
<td>0.63</td>
<td>5.1</td>
<td>7.7</td>
<td>3.9</td>
<td>0.7</td>
</tr>
<tr>
<td>+ length reward</td>
<td>0.65</td>
<td>4.8</td>
<td>3.8</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>( \lambda_j ) + re enr</td>
<td>0.61</td>
<td>4.9</td>
<td>3.6</td>
<td>0.9</td>
<td>0.4</td>
</tr>
<tr>
<td>+ length reward</td>
<td>0.65</td>
<td>4.3</td>
<td>3.7</td>
<td>0.5</td>
<td>0.4</td>
</tr>
</tbody>
</table>
---

**R1.**

**R2.** division by \( P_{\text{ILM}} \) boosts the label probability against (usually high) blank probability \( \rightarrow \) increase importance of external LM (\( \lambda_j \)) without suffering huge deletion errors
- limitation of shallow fusion (\( \lambda_j = 0 \))
- no need of decoding heuristics, e.g. length reward

---

**Highlights**

- various ILM-correction based LM integration methods
- formalization in a common RNN-T framework
- systematic comparison: both in-domain and cross-domain
- propose an exact-ILM training framework
- extension upon hybrid autoregressive transducer (HAT) (Variani et al. 2020)
- enable theoretical justification for other approaches
- further improve the best ILM method
- decoding interpretation: why improvement with ILM correction

---

**Human Language Technology and Pattern Recognition, RWTH Aachen University, Aachen, Germany**

**Email:** zhou@hlt.rwth-aachen.de, zuo.yun.zheng@hlt.rwth-aachen.de

**Website:** https://www-i6.informatik.rwth-aachen.de