Position Information for Language Modeling in Speech Recognition

Hsuan-Sheng Chiu\textsuperscript{1,2}, Guan-Yu Chen\textsuperscript{1}, Chun-Jen Lee\textsuperscript{3}, Berlin Chen\textsuperscript{1}

\textsuperscript{1} Taiwan Normal University, Taipei
\textsuperscript{2} Delta Electronics, Inc., Taipei
\textsuperscript{3} Telecommunication Labs., Chunghwa Telecom Co., Ltd., Taoyuan
Outline

• Introduction
• Position Information
• Positional Language Modeling
• Comparisons with Other Models
• Experimental Results
• Conclusions
Introduction (1/2)

• Language model (LM) plays a decisive role in many research fields of natural language processing, such as machine translation, information retrieval, speech recognition.

• The $n$-gram model, which aims at capturing only the local contextual information, or the lexical regularity of a language,
  – Inevitably faced with the problem of missing the information (either semantic or syntactic information) conveyed in the history before the immediately preceding $n$-1 words of the newly decoded word.
Introduction (2/2)

• According to different levels of linguistic information being utilized, language models can be roughly classified into the following several categories:
  – Word-based models (n-gram)
  – Word class- or topic based models (class based n-gram, WTM)
  – Sentence structure -based models (structured LM)
  – Document topic-based models (PLSA, LDA)

• Are there any other alternatives beyond the above LMs?
  – Position-dependent language models
    • In order to verify our belief of the usefulness of word position information, we try to analyze the word usage of a broadcast news corpus partitioned by the structure of documents
Position Information (1/3)

- The table below shows the style words with higher rank of TF-IDF scores on four partitions of the broadcast news corpus
  - The corpus was partitioned by a left-to-right HMM segmenter

<table>
<thead>
<tr>
<th>P1</th>
<th>P2</th>
<th>P3</th>
<th>P4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Continue</td>
<td>Doctor</td>
<td>Student</td>
<td>TV station name</td>
</tr>
<tr>
<td>Locale</td>
<td>Internet</td>
<td>Teacher</td>
<td>Roundup</td>
</tr>
<tr>
<td>Welcome</td>
<td>Coral</td>
<td>Rice wine</td>
<td>Edit and translate</td>
</tr>
</tbody>
</table>

Style words: introductions, topical words, footnotes
**Position Information (2/3)**

- We can observe that the word usage with respect to different partitions (or positions) of the broadcast news stories is apparently quite different (for the 12 style words)

![Graph showing word probabilities for different partitions](image-url)
Position Information (3/3)

• We could conclude that words in the marginal positions of documents are more specific while words in the middle positions are more comprehensive for the broadcast news documents.

• Hence, we first propose a **positional n-gram model** to explore the positional information inherent in the broadcast news documents for better speech recognition performance.
**Positional Language Modeling - Positional N-gram Model**

- The \( n \)-gram language model is trained respectively for each partition, and finally a positional \( n \)-gram model is constructed as a composite \( n \)-gram language model:

\[
P_{POS}(w_i \mid w_{i-2}, w_{i-1}) = \sum_{s=1}^{S} \alpha_s P(w_i \mid w_{i-2}, w_{i-1}, L_s)
\]

- Where \( S \) is the number of partitions, \( \alpha_s \) is the weight for a specific position \( L_s \)

Cf. Mixture-based Language Model

\[
P_{MIX}(w_i \mid w_{i-2}, w_{i-1}) = \sum_{k=1}^{K} \beta_k P(w_i \mid w_{i-2}, w_{i-1}, T_k)
\]
Comparisons - Positional N-gram & Mixture-based LM

• For modeling training, the mixture-based language model requires additional clustering being performed
  – While the positional $n$-gram model assumes that the documents in the collection share the similar structure
    • Determined by an HMM segmenter

• The model complexity of both models are equal to $V^n \times U$, where $V$ is vocabulary size, $n$ denotes the length of the window of words considered by the $n$-gram model, and $U$ is either the topic number or the position number.
The major difference between topic- and position-based models is that they are conceptually orthogonal.

- That is, the training corpus is either divided by topic or by position.
**Positional Language Modeling - Positional PLSA Model**

- Word position information also has been integrated into the PLSA model as a complement of the topic (or concept) information that has already been modeled by PLSA
  - The resulting model is referred to as the positional PLSA model

\[
P_{\text{PosPLSA}} \left( w_i \left| M_{H_{wi}} \right. \right) = \sum_{s=1}^{S} \sum_{k=1}^{K} P \left( w_i \left| T_k, L_s \right. \right) P \left( L_s \left| M_{H_{wi}} \right. \right) P \left( T_k \left| M_{H_{wi}} \right. \right)
\]

**Cf. Probabilistic Latent Semantic Analysis (PLSA)**

\[
P_{\text{PLSA}} \left( w_i \left| M_{H_{wi}} \right. \right) = \sum_{k=1}^{K} P \left( w_i \left| T_k \right. \right) P \left( T_k \left| M_{H_{wi}} \right. \right)
\]
Comparisons- PLSA & Positional PLSA model

- Graphical model representations

![Graphical Model Representations](image)

- If the position of a decoded word is observable, positional PLSA can be easily reduced to original PLSA with respect to a certain position, which means \( P(L_S | M_{H_w}) \) will be 1 for a certain position \( L_S \)

- The model complexities for positional PLSA and PLSA are \( V \times K \times S + (K + S) \times H \) and \( V \times K + K \times H \).
Comparisons - Positional PLSA & LDA models

- Latent Dirichlet Allocation (LDA) is an extension to PLSA model

- LDA use a prior knowledge to constrain
  - The distribution of the documents over the latent topics
  - The unigram distribution of each topic
Experimental Results - Setting

• The speech corpus consists of about 200 hours of MATBN Mandarin broadcast news (Mandarin Across Taiwan Broadcast News)
  – A subset of 25-hour speech collected was used to bootstrap the acoustic model training

• Another subset of 3-hour speech data was reserved for development (1.5 hours) and evaluation (1.5 hours)

• A background text news corpus consists of 170 million Chinese characters and an adaptation corpus consists of broadcast news transcription of 1 million characters
  – The vocabulary size is about 72k words

• Experiments were conducted in word graph rescoring stage
Experimental Results – Aspects

- The application of the positional $n$-gram model for language model adaptation can be discussed from three aspects:
  1) Whether the language model training corpus is segmented uniformly or segmented by the HMM segmenter
  2) Whether the word position of a decoded word in the search process is deterministic or nondeterministic
  3) The number of partitions being used

- We evaluate the performance of our proposed positional $n$-gram model, for which the order of $n$ is set to three
## Experimental Results – Positional n-gram

<table>
<thead>
<tr>
<th></th>
<th>CER(%)</th>
<th>PP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Background Trigram</td>
<td>20.32</td>
<td>682.10</td>
</tr>
<tr>
<td>Adapted Trigram</td>
<td>19.23</td>
<td>434.46</td>
</tr>
<tr>
<td>+ Deterministic Positional n-gram</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Uniform/HMM Segmentation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 partitions</td>
<td>19.09/19.44</td>
<td>402.31/387.08</td>
</tr>
<tr>
<td>4 partitions</td>
<td>19.29/19.54</td>
<td>408.02/382.78</td>
</tr>
<tr>
<td>8 partitions</td>
<td>19.62/19.37</td>
<td>416.41/378.20</td>
</tr>
<tr>
<td>16 partitions</td>
<td>19.85/19.36</td>
<td>453.59/387.48</td>
</tr>
<tr>
<td>+ Nondeterministic Positional n-gram</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Uniform/HMM Segmentation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 partitions</td>
<td>19.08/19.12</td>
<td>392.48/389.33</td>
</tr>
<tr>
<td>4 partitions</td>
<td>19.08/18.94</td>
<td>399.67/392.93</td>
</tr>
<tr>
<td>8 partitions</td>
<td>19.19/19.05</td>
<td>408.54/401.47</td>
</tr>
<tr>
<td>16 partitions</td>
<td>19.35/18.97</td>
<td>423.13/405.99</td>
</tr>
</tbody>
</table>

CER: character error rate, PP: perplexity
**Experimental Results – Positional N-gram & Mixture-based LM**

<table>
<thead>
<tr>
<th>+ Nondeterministic Positional $n$-gram</th>
<th>CER(%)</th>
<th>PP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Uniform/HMM Segmentation</td>
<td>Uniform/HMM Segmentation</td>
</tr>
<tr>
<td>2 partitions</td>
<td>19.08/19.12</td>
<td>392.48/389.33</td>
</tr>
<tr>
<td>4 partitions</td>
<td>19.08/18.94</td>
<td>399.67/392.93</td>
</tr>
<tr>
<td>8 partitions</td>
<td>19.19/19.05</td>
<td>408.54/401.47</td>
</tr>
<tr>
<td>16 partitions</td>
<td>19.35/18.97</td>
<td>423.13/405.99</td>
</tr>
<tr>
<td>+Mixture-Based LM</td>
<td>CER(%)</td>
<td>PP</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 topics</td>
<td>19.12</td>
<td>388.00</td>
</tr>
<tr>
<td>4 topics</td>
<td>19.17</td>
<td>384.26</td>
</tr>
<tr>
<td>8 topics</td>
<td>18.95</td>
<td>377.64</td>
</tr>
<tr>
<td>16 topics</td>
<td>18.80</td>
<td>372.26</td>
</tr>
</tbody>
</table>

- The CER performance of positional $n$-gram model is comparable with mixture-based language model when the number of topics or partitions is small (e.g., 2 or 4)
Experimental Results – Combination of Positional & Topical N-gram LM

Combined Model: Retrain LMs with each topical and positional block of the corpus.

<table>
<thead>
<tr>
<th>partition</th>
<th>1 topic</th>
<th>2 topics</th>
<th>4 topics</th>
<th>8 topics</th>
<th>16 topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 partition</td>
<td>19.23</td>
<td>19.12</td>
<td>19.17</td>
<td>18.95</td>
<td>18.80</td>
</tr>
<tr>
<td>2 partitions</td>
<td>19.12</td>
<td>19.17</td>
<td>19.05</td>
<td>19.10</td>
<td>18.89</td>
</tr>
<tr>
<td>4 partitions</td>
<td>18.94</td>
<td>19.09</td>
<td>18.94</td>
<td>18.96</td>
<td>18.90</td>
</tr>
<tr>
<td>8 partitions</td>
<td>19.05</td>
<td>19.15</td>
<td>19.15</td>
<td>19.03</td>
<td>-</td>
</tr>
<tr>
<td>16 partitions</td>
<td>18.97</td>
<td>19.23</td>
<td>19.21</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

- The CER performance is problematic probably due to data sparseness.
### Experiments and Results – LDA & PLSA

<table>
<thead>
<tr>
<th>PLSA</th>
<th>CER(%)</th>
<th>PP</th>
</tr>
</thead>
<tbody>
<tr>
<td>8 topics</td>
<td>19.76</td>
<td>563.70</td>
</tr>
<tr>
<td>16 topics</td>
<td>19.77</td>
<td>554.07</td>
</tr>
<tr>
<td>32 topics</td>
<td>19.60</td>
<td>545.14</td>
</tr>
<tr>
<td>64 topics</td>
<td>19.71</td>
<td>539.61</td>
</tr>
<tr>
<td>128 topics</td>
<td>19.55</td>
<td>533.29</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>LDA</th>
<th>CER(%)</th>
<th>PP</th>
</tr>
</thead>
<tbody>
<tr>
<td>8 topics</td>
<td>19.80</td>
<td>561.13</td>
</tr>
<tr>
<td>16 topics</td>
<td>19.83</td>
<td>549.46</td>
</tr>
<tr>
<td>32 topics</td>
<td>19.73</td>
<td>538.86</td>
</tr>
<tr>
<td>64 topics</td>
<td>19.46</td>
<td>537.35</td>
</tr>
<tr>
<td>128 topics</td>
<td>19.57</td>
<td>535.78</td>
</tr>
</tbody>
</table>

- The PP will be slightly improved when the number of topic increases
  - However, the CER does not have such tendency
- The performance of LDA and PLSA model are almost indistinguishable in our task
### Experiments and Results – PLSA & Positional PLSA

<table>
<thead>
<tr>
<th>CER(%)</th>
<th>Topics</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>8</td>
<td>16</td>
<td>32</td>
<td></td>
</tr>
<tr>
<td>2 partitions</td>
<td>19.76</td>
<td>19.57</td>
<td>19.63</td>
<td></td>
</tr>
<tr>
<td>3 partitions</td>
<td>19.73</td>
<td>19.68</td>
<td>19.68</td>
<td></td>
</tr>
<tr>
<td>4 partitions</td>
<td>19.69</td>
<td>19.64</td>
<td>19.66</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PP</th>
<th>Topics</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>8</td>
<td>16</td>
<td>32</td>
<td></td>
</tr>
<tr>
<td>2 partitions</td>
<td>555.97</td>
<td>546.27</td>
<td>538.73</td>
<td></td>
</tr>
<tr>
<td>3 partitions</td>
<td>547.90</td>
<td>544.28</td>
<td>537.77</td>
<td></td>
</tr>
<tr>
<td>4 partitions</td>
<td>552.22</td>
<td>554.66</td>
<td>557.70</td>
<td></td>
</tr>
</tbody>
</table>

- We compare the original PLSA language model with the positional PLSA language model under different numbers of topics and partitions.
- The performance of positional PLSA seems not to be significantly different from that of PLSA.
Experiments and Results – Discussions

- Why CER does not get significantly improved?
  - Possibly the document structure of the evaluation set is not complicated enough to be split into such many partitions
  - The use of a specific language model for the last partition might provide an additional benefit; however its short duration (compared to the other positions) will make its contribution to the overall CER improvement insignificant
    - Durations of the four partitions (P1 to P4) of the corpus are 31%, 35%, 28% and 6% on average
  - The information over topic (cluster) and position (partition) might be overlapped
Conclusions and Future work

• An alternative document topic (or style) modeling approach was proposed

• Although the performance gains are not very significant for our proposed positional n-gram model and positional PLSA model, we believe that the use of position information still has its potential

• In the meantime, we are also investigating the discriminative N-best reranking technique by utilizing the word positional information