Position Information for Language Modeling in Speech Recognition

Hsuan-Sheng Chiu[1,2], *Guan-Yu Chen*[1], *Chun-Jen Lee*[3], *Berlin Chen*[1]

[1] Taiwan Normal University, Taipei
[2] Delta Electronics, Inc., Taipei
[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)

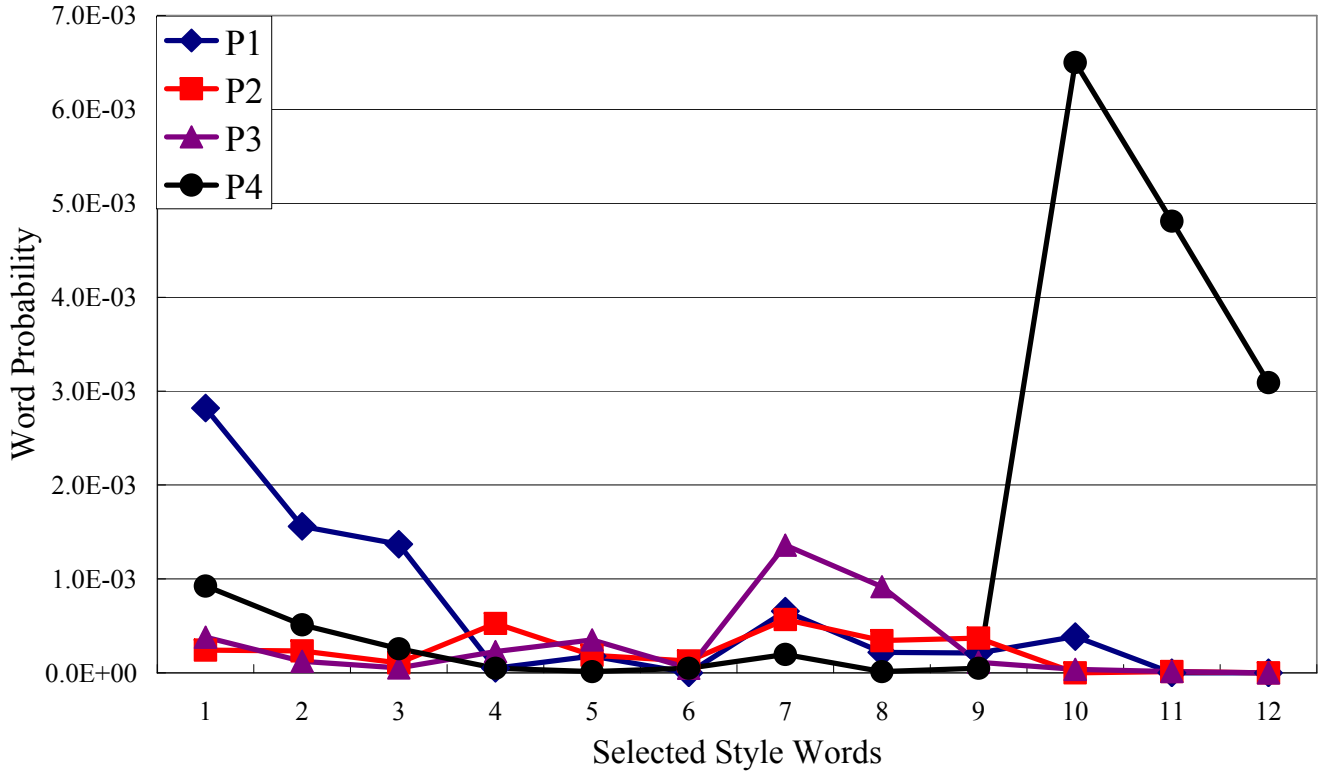
- The table below shows the style words with higher rank of *TF-IDF* scores on four partitions of the broadcast news corpus
  - The corpus was partitioned by a left-to-right HMM segmenter

P1	P2	P3	P4
1繼續 Continue	4醫師 Doctor	7學生 Student	10公視 TV station name
2現場 Locale	5網路 Internet	8老師 Teacher	11綜合報導 Roundup
3歡迎 Welcome	6珊瑚 Coral	9酒 Rice wine	12編譯 Edit and translate

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)



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:

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

  - Where S is the number of partitions, α_S is the weight for a specific position L_S

Cf. Mixture-based Language Model

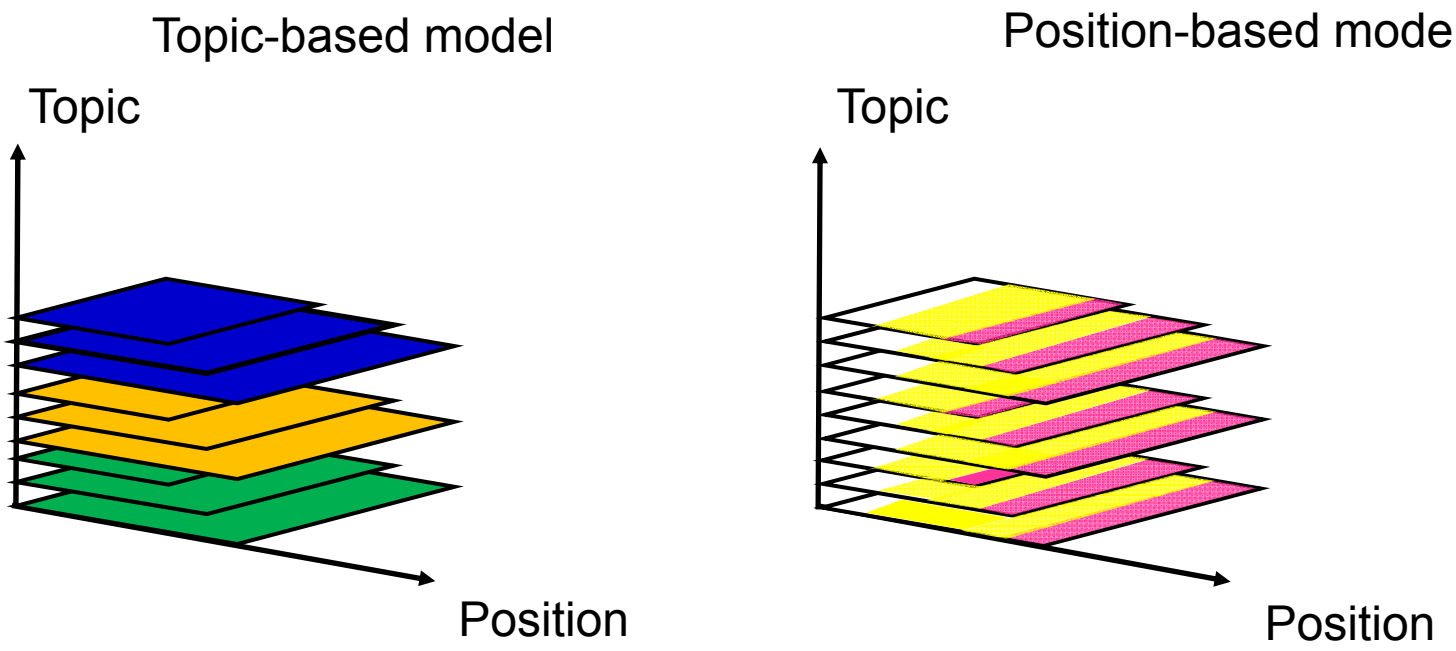
$$P_{MIX}\left(w_i \mid w_{i-2}, w_{i-1}\right) = \sum_{k=1}^{K} \beta_k P\left(w_i \mid w_{i-2}, w_{i-1}, T_k\right)$$

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.

Comparisons- Positional N-gram & Mixture-based LM



- 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

$$P_{PosPLSA}\left(w_i \middle| M_{H_{w_i}}\right) = \sum_{s=1}^{S}\sum_{k=1}^{K} P\left(w_i \middle| T_k, L_s\right) P\left(L_s \middle| M_{Hw_i}\right) P\left(T_k \middle| M_{Hw_i}\right)$$

position latent topic

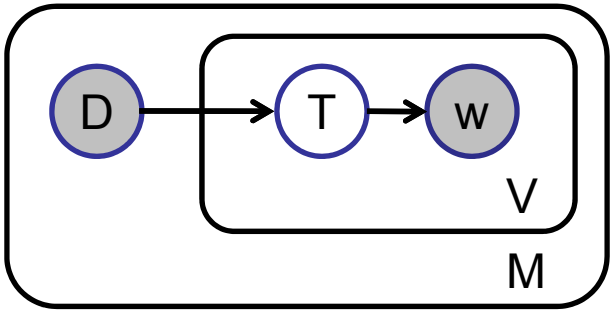
Cf. Probabilistic Latent Semantic Analysis (PLSA)

$$P_{PLSA}\left(w_i \middle| M_{H_{w_i}}\right) = \sum_{k=1}^{K} P\left(w_i \middle| T_k\right) P\left(T_k \middle| M_{Hw_i}\right)$$

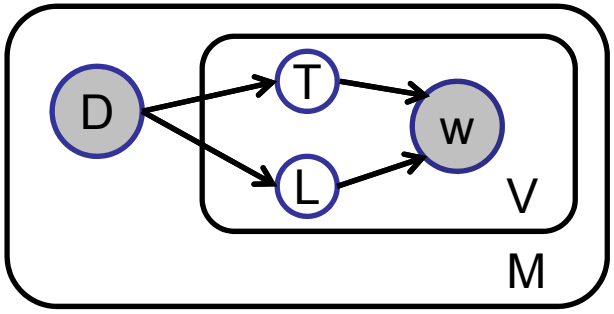
Comparisons- PLSA & Positional PLSA model

- Graphical model representations

PLSA



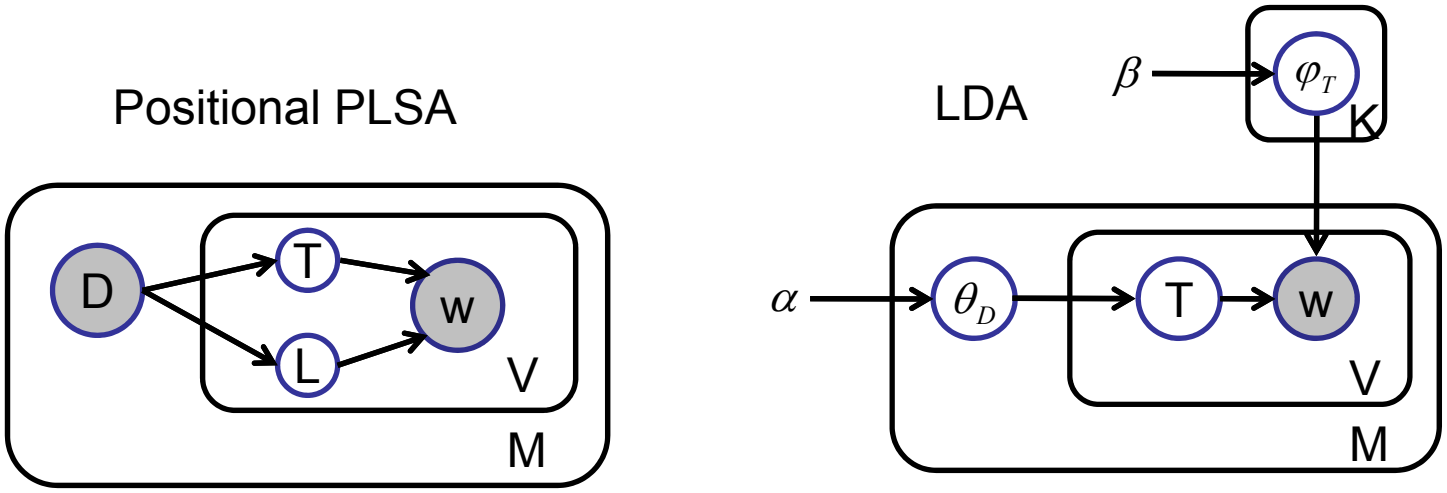
Positional PLSA



- 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 $P(L_S \mid M_{H_{w_i}})$ 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

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

CER: character error rate, PP: perplexity

Experimental Results – Positional N-gram & Mixture-based LM

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

- 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

partition \ topic	1 topic	2 topics	4 topics	8 topics	16 topics
1 partition	19.23	19.12	19.17	18.95	18.80
2 partitions	19.12	19.17	19.05	19.10	18.89
4 partitions	18.94	19.09	18.94	18.96	18.90
8 partitions	19.05	19.15	19.15	19.03	-
16 partitions	18.97	19.23	19.21	-	-

- The CER performance is problematic probably due to data sparseness

	T1	T2
P1	P1T1	P1T2
P2	P2T1	P2T2

Experiments and Results – LDA & PLSA

PLSA	CER(%)	PP
8 topics	19.76	563.70
16 topics	19.77	554.07
32 topics	19.60	545.14
64 topics	19.71	539.61
128 topics	19.55	533.29

LDA	CER(%)	PP
8 topics	19.80	561.13
16 topics	19.83	549.46
32 topics	19.73	538.86
64 topics	19.46	537.35
128 topics	19.57	535.78

- 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

CER(%)	Topics		
	8	16	32
2 partitions	19.76	19.57	19.63
3 partitions	19.73	19.68	19.68
4 partitions	19.69	19.64	19.66
PP	Topics		
	8	16	32
2 partitions	555.97	546.27	538.73
3 partitions	547.90	544.28	537.77
4 partitions	552.22	554.66	557.70

- 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