Learning Bidirectional Intent Embeddings by Convolutional Deep Structured Semantic Models for Spoken Language Understanding

Yun-Nung (Vivian) Chen, Dilek Hakkani-Tür, and Xiaodong He

Microsoft **Research**

Summary

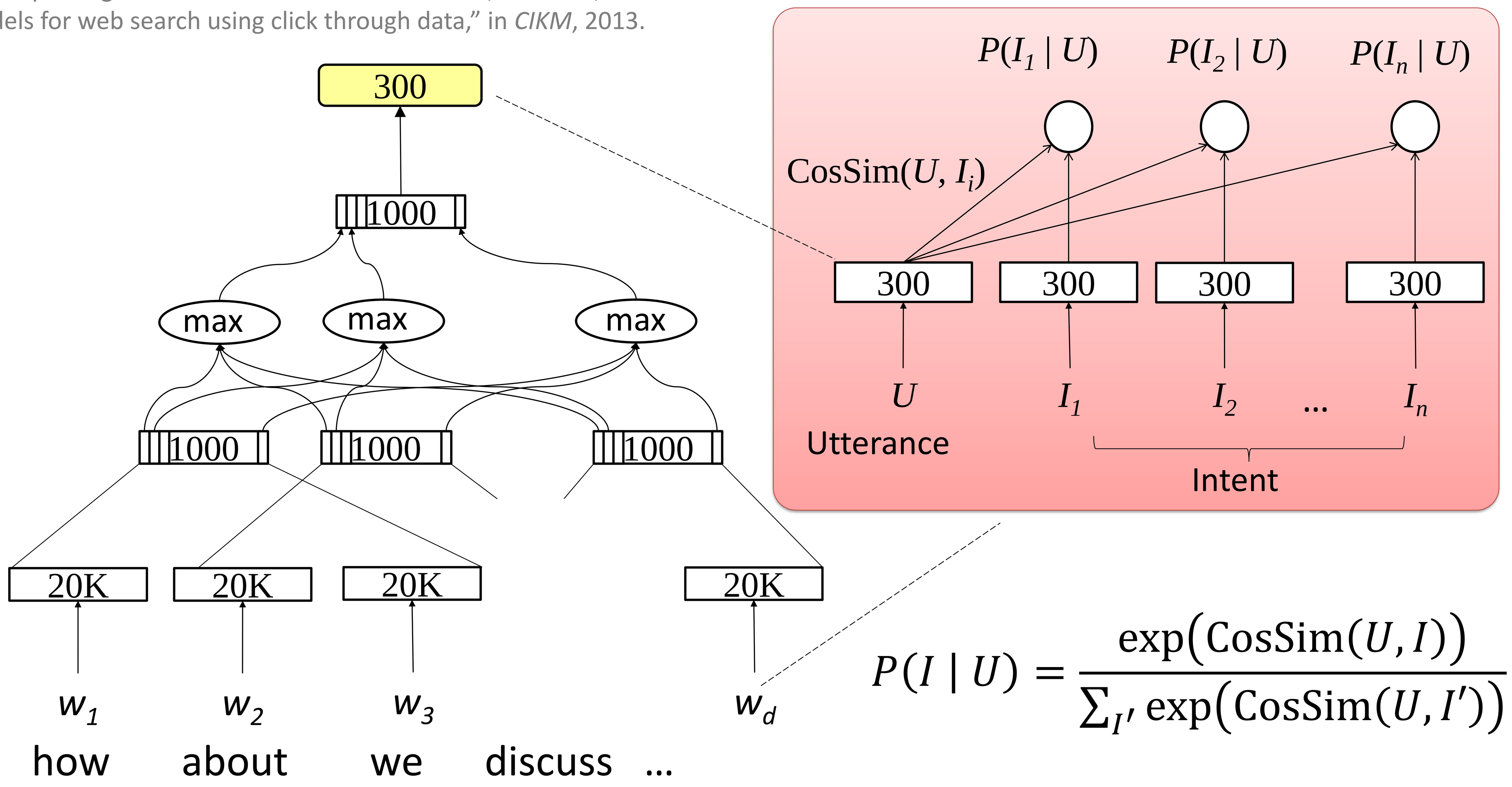
- Motivation: ***Inflexible Intent Schema***
  - Intents are usually predefined and inflexible to expand and transfer across domains and genres, where re-designing a semantic schema with intents for different domains or genres requires human effort for annotation and model re-training.
- Approach: ***Learning Intent Representation***
  - Applying CDSSM to learn high-level semantic representations to bridge the semantic relation across domains and across genres for intent expansion and actionable item detection tasks respectively, (e.g. "find movie" and "find weather" belong to different domains, but they share the semantics about "find".)
- Result
  - CDSSM is capable of generating more flexible intent embeddings to remove the domain constraint in dialogue systems for intent expansion. The intent embeddings can also be transferred to different genres, showing the robustness to genre mismatch.

Convolutional Deep Structured Semantic Models (CDSSM)

Model Architecture

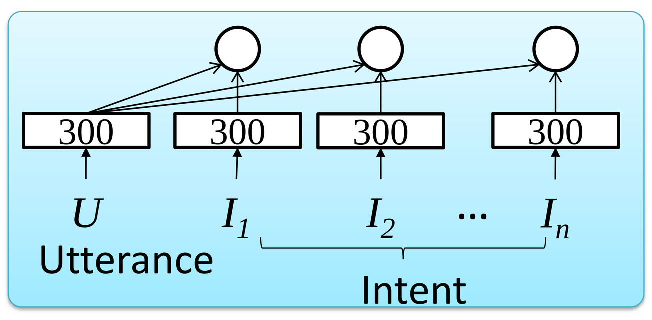
Shen et al., "A latent semantic model with convolutional-pooling structure for information retrieval," in *CIKM*, 2014.
Huang et al., "Learning deep structured semantic models for web search using click through data," in *CIKM*, 2013.

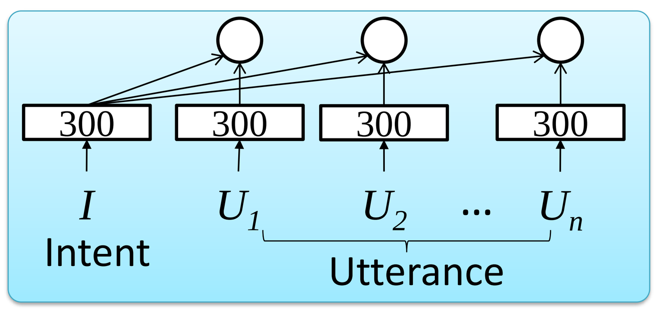
- **Semantic Layer:** y — feed-forward neural network layers outputs the final non-linear semantic features
  Projection Matrix: W_s
- **Max Pooling Layer:** l_m — only retain the most prominent local features by applying the max operation over each dimension of l_c to keep the max activation of hidden topics across the whole word sequence
  Max Pooling Operation
- **Convolutional Layer:** l_c — contextual features c_i for each target word
  $l_{ci} = \tanh(W_c^T c_i)$
  Convolution Matrix: W_c
- **Word Hashing Layer:** l_h — one-hot word vector → tri-letter vector (e.g. "email" → "#em", "ema", "mai", "ail", "il#")
  Word Hashing Matrix: W_h
- **Word Sequence:** x — user utterance / intent



$$P(I \mid U) = \frac{\exp(\mathrm{CosSim}(U, I))}{\sum_{I'} \exp(\mathrm{CosSim}(U, I'))}$$

Training Procedure





- The objective maximizes the likelihood of associated intents given utterances using mini-batch SGD updates:

→ **Predictive Model** $\Lambda(\theta_1) = \log \prod_{(U,I^+)} P(I^+ \mid U)$

- Similarly, reversing intents and utterances induces

→ **Generative Model** $\Lambda(\theta_2) = \log \prod_{(I,U^+)} P(U^+ \mid I)$

Bidirectional Score Estimation

- incorporate the effectiveness of predictive and generative models

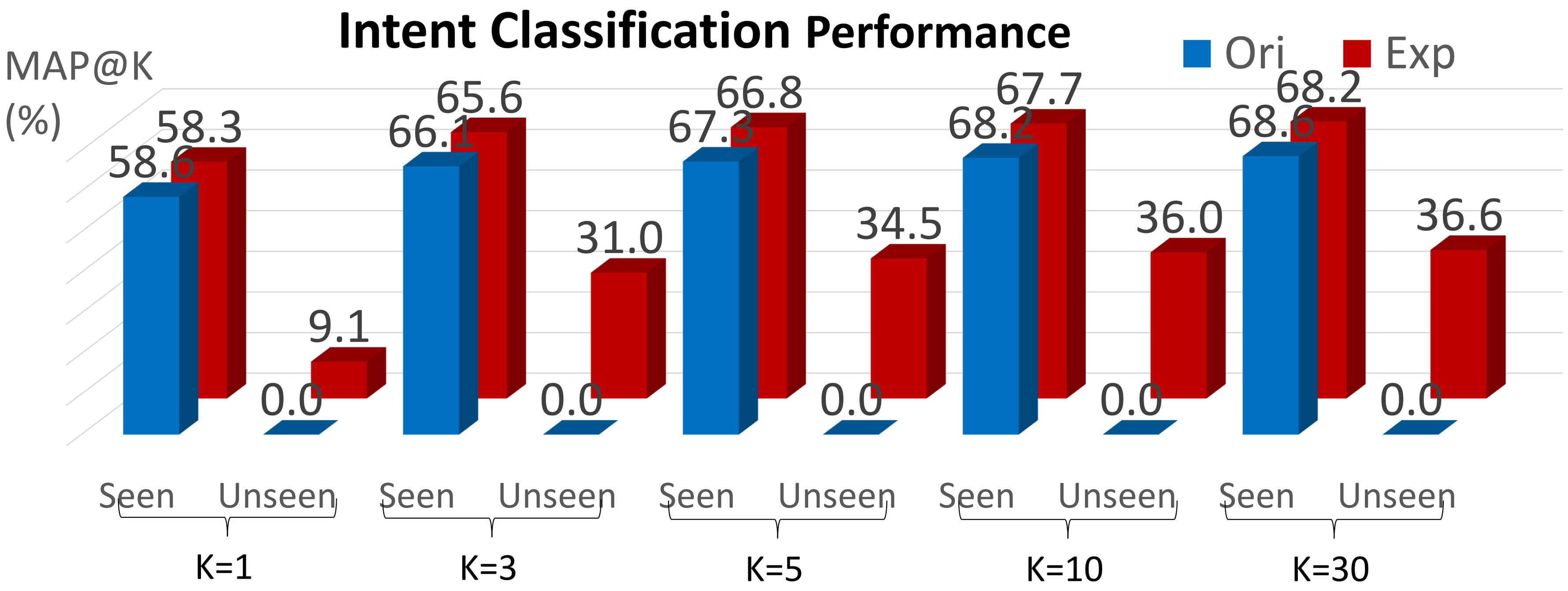
$$S_{\mathrm{Bi}}(U, I) = \gamma S_{\mathrm{P}}(U, I) + (1 - \gamma) S_{\mathrm{G}}(U, I)$$

- The trained model is able to generate representations given word sequences without model re-training, so the intent embeddings can generalize to different domains and genres.

Intent Expansion

- For each utterance vector y_U, the semantic similarity can be estimated using vectors for both seen and unseen intents.
- The unseen intent vectors can be generated from CDSSM by feeding the tri-letter vectors of the new intent as input without model re-training.

Intent Classification Performance



- The expanded models consider new intents without training samples, and produces similar but slightly worse than original models for seen intents due to higher uncertainty from more intent candidates.
- For unseen intents, expanded models are able to capture the correct intents and achieve higher than 30% of MAP when K ≥ 3, which indicates the encouraging performance when considering more than 100 intents.

Expand	Seen				Unseen			
	K=1	K=3	K=5	K=10	K=1	K=3	K=5	K=10
Predictive	**58.9**	**65.9**	**67.1**	**67.9**	5.2	18.7	23.4	26.1
Generative	44.7	52.0	53.5	54.6	6.7	23.2	26.5	28.7
Bidirectional	58.3	65.6	66.8	67.7	**9.1**	**31.0**	**34.5**	**36.0**

- Although the predictive model performs better for seen intents, the bidirectional estimation is more robust to unseen intents, which is crucial to this intent expansion task.

Actionable Item Detection

- This task investigates actionable item detection in meetings (human-human genre), where the intelligent assistant dynamically provides the participants access to information (e.g. scheduling a meeting, taking notes) without interrupting the meetings.
- A CDSSM is applied to learn the latent semantics for human actions and utterances from human-machine and human-human interactions.

Approach	Mismatch	Match
Predictive	52.8	64.5
Generative	55.0	64.8
Bidirectional	**59.1**	**68.9**

- A CDSSM is applied to learn the latent semantics for human actions and utterances from human-machine and human-human interactions.
- The improvement of bidirectional estimation suggests that the predictive and generative model can compensate each other, and then provide more robust estimated scores for the goal of actionable item detection.

Conclusion

- The experiments show that the learned embeddings
  - capture the semantics borrowed from other domains and can be used to flexibly expand the intents through high-level representations.
  - carry the crucial high-level semantics and can be applied to different genres for easy adaptation and extension.

Carnegie Mellon University