DETECTING ACTIONABLE ITEMS IN MEETINGS BY CONVOLUTIONAL DEEP STRUCTURED SEMANTIC MODELS

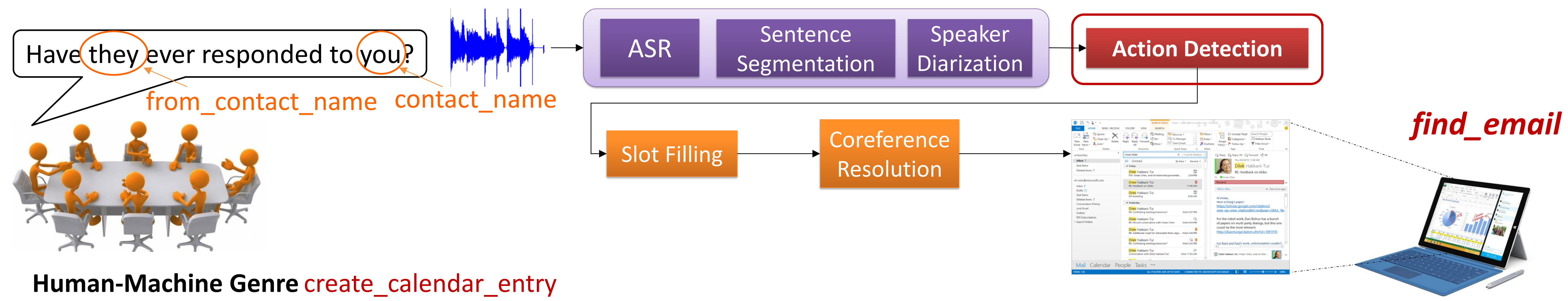
Yun-Nung (Vivian) Chen, Dilek Hakkani-Tür, and Xiaodong He

Microsoft Research

Dataset available: http://research.microsoft.com/projects/meetingunderstanding/

Introduction

- Motivation
  - Computing devices have been easily accessible and information search has been a common part of regular conversations, where these meetings include discussions for identifying participants' next actions.
- Main Idea
  - Human-machine interactions collected by existing intelligent systems (e.g. Cortana data) may help detect actionable items in human-human dialogues (e.g. meetings)
  - Learning action representations using a CDSSM architecture helps transfer high-level semantics across genres
- **Actionable Item Detection Task**
  - Goal: provide the easy access to information and perform actions a personal assistant can handle without interrupting the meetings
  - Assumption: some actions and associated arguments can be shared across genres



Human-Machine Genre create_calendar_entry
schedule a meeting with <contact_name>John</contact_name> <start_time>this afternoon</start_time>

Human-Human Genre create_calendar_entry
how about the <contact_name>three of us</contact_name> discuss this later <start_time>this afternoon</start_time>?

→ more casual, include conversational terms

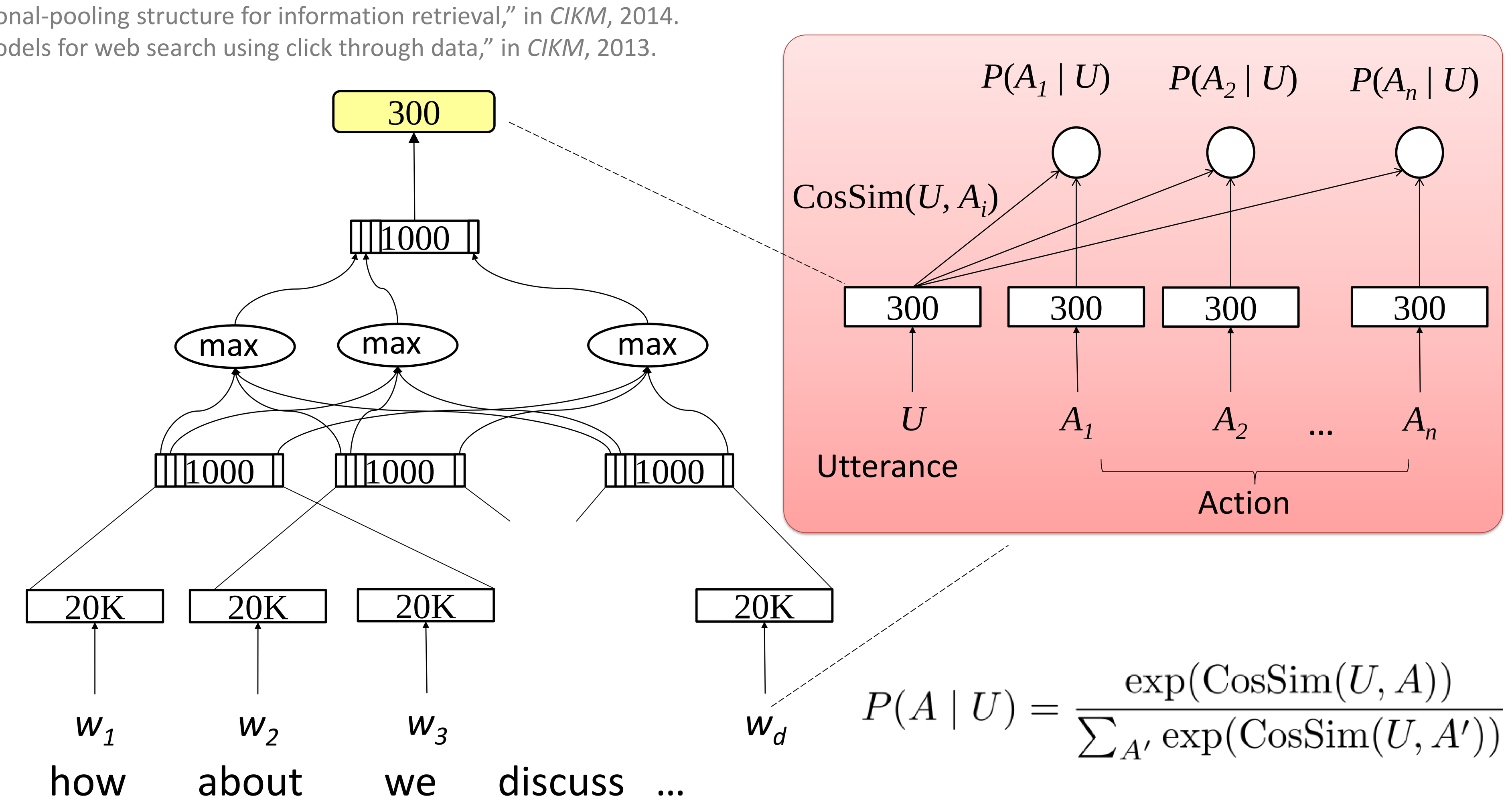
- Task: multi-class utterance classification
  - train on the available human-machine genre
  - test on the human-human genre
- Adaptation
  - model adaptation
  - embedding vector adaptation

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



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

Training Procedure

- Predictive model: $\Lambda(\theta_1) = \log \prod_{(U, A^+)} P(A^+ \mid U)$
- Generative model: $\Lambda(\theta_2) = \log \prod_{(U^+, A)} P(U^+ \mid A)$

Bidirectional Score Estimation

- incorporate the effectiveness of predictive and generative models

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

- During training, utterances and action embeddings are learned.
- During estimation, utterance embeddings are generated.

Adaptation

- Issue: source-target genre mismatch
- Solution: 1) **Adapting CDSSM**; 2) **Adapting Action Embeddings**
- Learning adapted action embeddings via the objective considering action and utterance embeddings together:

$$\Phi_{\mathrm{act}}(\hat{Q}, \hat{R}) = \sum_{i=1}^{n} \left[\alpha_i \|\hat{q}_i - q_i\|^2 + \sum_{l(r_j)=i} \beta_{ij} \|\hat{q}_i - \hat{r}_j\|^2 \right]$$

The distance between original and new action embeddings

The distance between new action embeddings and corresponding utterance embeddings

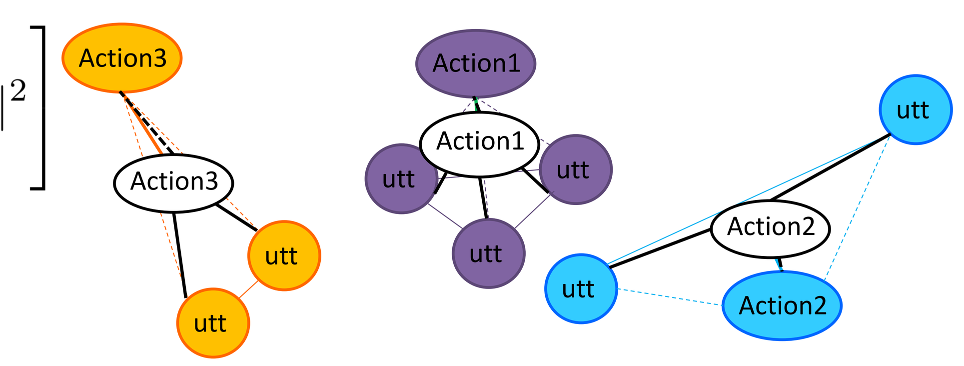
+

$$\Phi_{\mathrm{utt}}(\hat{R}) = \sum_{i:l(r_i)=1}^{n} \left[\alpha_i \|\hat{r}_i - r_i\|^2 + \sum_{l(r_j)=l(r_i)} \beta_{ij} \|\hat{r}_i - \hat{r}_j\|^2 \right]$$

$$\Phi(\hat{Q}, \hat{R}) = \Phi_{\mathrm{act}}(\hat{Q}, \hat{R}) + \Phi_{\mathrm{utt}}(\hat{R})$$

adapted action embeddings

$$\Delta\hat{q}_i = \frac{\alpha q_i + \sum \beta_{ij} \hat{r}_j}{\alpha + \sum \beta} \qquad \Delta\hat{r}_i = \frac{\alpha r_i + \sum \beta_{ij} \hat{r}_j}{\alpha + \sum \beta}$$



- The adapted action vectors are close to the corresponding utterance vectors for the target genre.

Experiments

- Dataset: 22 meetings from the ICSI meeting corpus
- Evaluation metrics: the average AUC for 10 actions+*others*

Approach (%)		#dim	Mismatch-CDSSM			Adapt-CDSSM			Match-CDSSM		
			P(A|U)	P(U|A)	Bidir	P(A|U)	P(U|A)	Bidir	P(A|U)	P(U|A)	Bidir
w/o SVM	Sim		47.5	48.2	49.1	48.7	50.1	50.4	56.3	43.4	50.6
	AdaptSim		54.0	53.9	**55.8**	59.5	57.0	**60.1**	64.2	60.4	**62.3**
w/ SVM	Embeddings	300	53.1	48.1	55.7	60.1	59.0	64.0	64.3	65.6	**69.3**
	+ Sim	311	52.8	55.0	59.1	60.8	60.3	65.1	64.5	64.8	68.9
	+ AdaptSim	311	52.8	55.2	**59.2**	61.6	61.1	**65.7**	64.7	65.4	69.1

	Model	AUC (%)
Baseline	N-gram (N=1,2,3)	52.84
	Paragraph Vector (doc2vec)	59.79
Proposed	CDSSM: P(A|U)	64.33
	CDSSM: P(U|A)	65.58
	CDSSM: Bidirectional	**69.27**

Conclusion

- The latent semantic features generated by CDSSM show the effectiveness of detecting actions in meetings compared to lexical features, and also outperform the state-of-the-art semantic features.
- The adaptation techniques are proposed to adjust the learned embeddings to fit the target genre when the source genre does not match well with target genre, showing significant improvements in detecting actionable items.
- The proposed bidirectional estimation outperforms unidirectional one, because predictive and generative models compensate each other.
- The paper highlights a future research direction by releasing an annotated dataset and the trained embeddings for actionable item detection.

http://vivianchen.idv.tw yvchen@cs.cmu.edu vivian.ynchen