Role Specific Lattice Rescoring for Speaker Role Recognition from Speech Recognition Outputs

Nikolaos Flemotomos[1], Panayiotis Georgiou[1], David C. Atkins[2], Shrikanth Narayanan[1]

[1]Signal Analysis and Interpretation Lab (SAIL), University of Southern California
[2]Department of Psychiatry and Behavioral Sciences, University of Washington

Problem & Motivation

- Definition of SRR: map every speaker turn to some speaker role
- Examples:
  - Business meetings (CEO, Graphics Designer, HR Specialist, etc)
  - Broadcast news programs (Anchor, Presenter, Guest, etc)
  - Psychotherapy (Therapist, Patient)
- Typically exploit linguistic features after ASR
  - **Problem**: Generic ASR system leads to information loss
  - **Solution**: Build role-specific ASR systems

Method: Turn-level SRR

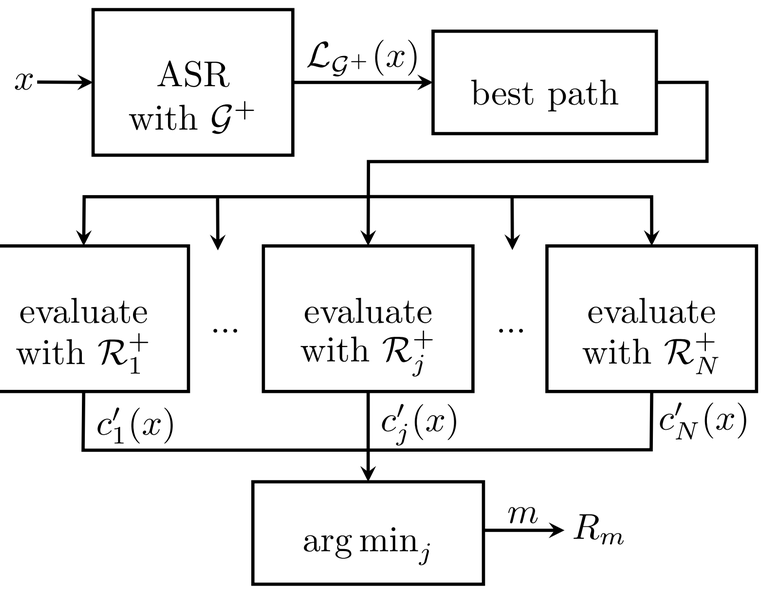
- Build a backround LM $\mathcal{G}$ and N role-specific LMs $\mathcal{R}_1, \mathcal{R}_2, \cdots, \mathcal{R}_N$ corresponding to the N roles
- Interpolate the LMs to recognize the same vocabulary

$$\mathcal{G}^+ = w_g\mathcal{G} \oplus (1 - w_g)\tilde{\mathcal{R}}$$

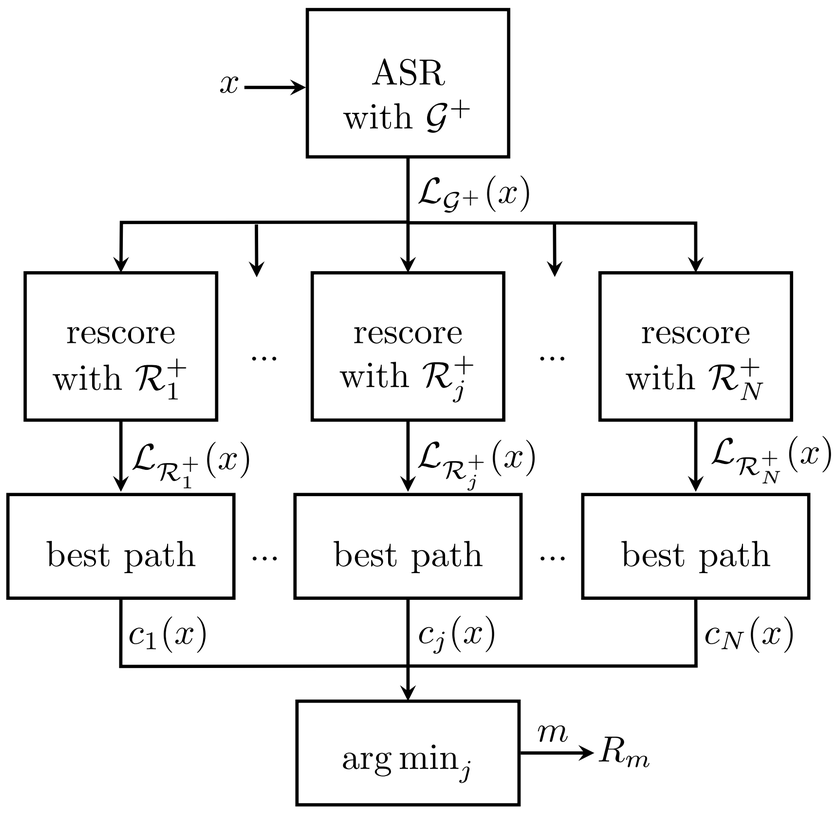
$$\mathcal{R}_i^+ = w_{g_i}\mathcal{G} \oplus w_{r_i}\mathcal{R}_i \oplus (1 - w_{g_i} - w_{r_i})\tilde{\mathcal{R}}_i$$

$$\tilde{\mathcal{R}} = \frac{1}{N}\bigoplus_{i=1}^{N}\mathcal{R}_i, \quad \tilde{\mathcal{R}}_i = \frac{1}{N-1}\bigoplus_{\substack{j=1 \\ j\neq i}}^{N}\mathcal{R}_j$$

- Typical approach:



- Proposed approach:



$c_j(x)$: LM-cost of the best path in the lattice $\mathcal{L}_{\mathcal{R}_j^+}(x)$

Method: Speaker-level SRR

- First, apply speaker clustering
- Define the costs $c(S_i|R_j) \triangleq \sum_{x\in T_i} c_j(x)$
  T_i the set of turns corresponding to the speaker S_i

Algorithm Speaker-level SRR given costs for each (speaker,role)

Inputs: speakers $S_1, S_2, \cdots, S_N$
roles $R_1, R_2, \cdots, R_N$
costs $c(S_i|R_j) \forall i, j$

$\tilde{S} \leftarrow \{S_i\}_{i=1}^N$; $\tilde{R} \leftarrow \{R_i\}_{i=1}^N$
while $\tilde{S} \neq \phi$ **do**
  for $S_i \in \tilde{S}$ **do**
    $l_i \leftarrow \arg\min_m c(S_i|R_m),\ R_m \in \tilde{R}$
    $C_i \leftarrow \min_n |c(S_i|R_{l_i}) - c(S_i|R_n)|,\ R_n \in \tilde{R} \setminus \{R_{l_i}\}$
  end for
  $k \leftarrow \arg\max_i C_i$
  assign R_{l_k} to S_k
  $\tilde{S} \leftarrow \tilde{S} \setminus \{S_k\}$; $\tilde{R} \leftarrow \tilde{R} \setminus \{R_{l_k}\}$
end while

Datasets

- PSYCH: dyadic interactions in psychotherapy
  Therapist (49.0h) vs. Client (43.0h)
- AMI: business meetings
  Project Manager (22.9h), Marketing Expert (15.3h),
  User Interface Designer (13.8h), Industrial Designer (15.2h)

Results

- Turn-level SRR: Misclassification Rate

	rescoring	no rescoring	majority class
PSYCH	23.58	**10.75**	50.67
AMI	64.70	63.40	**62.22**

- Speaker-level SRR: Misclassification Rate

	rescoring	no rescoring	clustering (BIC-HAC)
PSYCH†	**0.00**	7.46	–
PSYCH	**4.41**	5.83	4.08
AMI†	**29.46**	55.52	–
AMI	**46.16**	60.94	15.63

† denotes oracle speaker clustering

- Effect on speech recognition: WER

	rescoring, turn-level	rescoring, speaker-level	generic
PSYCH	37.84	**37.54**	37.99
AMI	29.35	**29.27**	29.29

Conclusion

- short speech segments: insufficient observations to infer speaker role
  $\Rightarrow$ speaker-level SRR
- even small role-specific improvements in the text (small decrease in WER) can be of high value for SRR
- future work:
  time-dependent speaker roles or multiple speakers with the same role

flemotom@usc.edu
May 15, 2019
Work supported by the NIH, NSF, and DoD.