Learning ASR-Robust Contextualized Embeddings for Spoken Language Understanding

Chao-Wei Huang    Yun-Nung (Vivian) Chen

National Taiwan University

r07922069@ntu.edu.tw    y.v.chen@ieee.org

Code available at https://github.com/MiuLab/SpokenVec
Highlights

• Contextualized embeddings like ELMo do not transfer well to spoken domain w/ ASR errors.

• We propose a fine-tuning method to tackle this problem.
Motivation: Bridge between ASR and NLU

- Intuitive way for SLU: pipelined approach

- ASR errors affects downstream tasks
LU models

- The SOTA LU models are usually pre-trained LMs
- But they are pre-trained on written text

Do they transfer well to spoken domain w/ ASR errors?
Do they transfer well to spoken domain?

How can we transfer them to spoken domain?
Our method: Additional fine-tuning stage

- LM pre-training
  - Same as ELMo

- LM fine-tuning
  - make the embeddings acoustic-aware

- Training target task classifier
  - Pre-trained LM is used as a feature extractor
Make them acoustic-aware!

- Force embeddings of acoustically similar words to be closer

How to determine which words to bring closer?

Confusion loss

$$\mathcal{L}_{\text{conf}} = \frac{1}{|C|} \sum_{c \in C} \sum_{i=0}^{1} 1 - \frac{h_{t1,i}^x \cdot h_{t2,i}^x}{\|h_{t1,i}^x\| \|h_{t2,i}^x\|}$$
How to determine which words to bring closer?

- Case 1: we have paired ASR and manual transcripts (supervised)
- Case 2: we only have some ASR transcripts (unsupervised)
How to determine which words to bring closer?

- Case 1: we have paired ASR and manual transcripts (supervised)
How to determine which words to bring closer?

• Case 2: we only have some ASR transcripts (unsupervised)
Fine-tuning LMs (ELMo)

- **ULMFit:**
  Fine-tune w/ LM objective helps domain transfer (Howard and Ruder, 2018)

\[
\mathcal{L}_{LM} = \frac{1}{|x|} \sum_{t=1}^{|x|} -\log p(w_t | w_{<t}) - \log p(w_t | w_{>t}),
\]

\[
\mathcal{L}_{conf} = \frac{1}{|C|} \sum_{c \in C} \sum_{i=0}^{1} 1 - \frac{h_{t1,i} \cdot h_{t2,i}}{\| h_{t1,i} \| \cdot \| h_{t2,i} \|}
\]

\[
\mathcal{L}_{FT} = \mathcal{L}_{LM} + \beta \mathcal{L}_{conf},
\]
Results

![Table showing results for ATIS, SNIPS, and SMARTLIGHTS datasets with different methods]

- **ATIS (15.55%)**: Manual: 96.65, ASR: 93.27, ULMFit: 94.27, supervised: 95.65, unsupervised: 95.39
- **SNIPS (45.56%)**: Manual: 96.29, ASR: 77.86, ULMFit: 87.74, supervised: 88.52, unsupervised: 89.55
Conclusions

• Contextualized models (ELMo) do not transfer well to spoken domain with ASR errors.

• We introduce an additional fine-tuning stage to make embeddings more acoustic-aware.

• We achieve this by forcing embeddings of acoustically similar words to be closer. We propose two methods to extract these pairs

• The experiment results show that our method can make contextualized embeddings more robust to ASR errors.
Thanks for listening!

Code available at [https://github.com/MiuLab/SpokenVec](https://github.com/MiuLab/SpokenVec)

Chao-Wei Huang  
r07922069@ntu.edu.tw

Yun-Nung (Vivian) Chen  
y.v.chen@ieee.org