Iron Man (2008)
Language Empowering Intelligent Assistants

- Apple Siri (2011)
- Google Now (2012)
- Microsoft Cortana (2014)
- Amazon Alexa/Echo (2014)
- Facebook M & Bot (2015)
- Google Home (2016)
- Apple HomePod (2017)
Task-Oriented Dialogue Systems (Young, 2000)

**Language Understanding (LU)**
- Domain Identification
- User Intent Detection
- Slot Filling

**Dialogue Management (DM)**
- Dialogue State Tracking (DST)
- Dialogue Policy

**Natural Language Generation (NLG)**

**Speech Recognition**

**Hypothesis**
are there any action movies to see this weekend

**Text Input**
Are there any action movies to see this weekend?

**Speech Signal**

**Text response**
Where are you located?

**System Action/Policy**
request_location

**Semantic Frame**
request_movie
genre=action, date=this weekend

**Database**
Recent Advances in NLP

- Contextual Embeddings (ELMo & BERT)
  - Boost many understanding performance with pre-trained natural language
Listening...
Lift all lights to Morocco
List all flights tomorrow
Task-Oriented Dialogue Systems (Young, 2000)

Speech Recognition

Hypothesis
are there any action movies to see this weekend

Language Understanding (LU)
- Domain Identification
- User Intent Detection
- Slot Filling

Semantic Frame
request_movie
genre=action, date=this weekend

System Action/Policy
request_location

Natural Language Generation (NLG)

Dialogue Management (DM)
- Dialogue State Tracking (DST)
- Dialogue Policy

Database
Mismatch between Written and Spoken Languages

- **Goal: ASR-Robust Contextualized Embeddings**
  - ✓ learning contextualized word embeddings specifically for spoken language
  - ✓ achieves better performance on *spoken* language understanding tasks
    - shows better results on ASR transcripts
    - maintain similar results on manual transcripts
Solution 1: Adapting Transformer to ASR Lattices
BERT/GPT Pre-Training & Fine-Tuning

- **Pre-Training**
  - Transformer Encoder
    - $w_1, w_2, \ldots, w_{m-1}, w_m$
    - Linear

- **Fine-Tuning**
  - Transformer Encoder
    - $w_1, w_2, \ldots, w_{m-1}, w_m$
    - Linear
    - $<S>$, $w_1, w_2, \ldots, w_{m-1}, w_m <E>$

$w_2, w_3, \ldots, w_m, w_{m+1}$
ASR Lattices

- Idea: lattices may include correct words
- Goal: feed lattices into Transformer

1) Linearize
2) Binary mask
3) Probabilistic mask
Self-Attention (Vaswani+, 2017)

\[
A(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V(M)V
\]
Attention Masks \[ A(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}} + M)V \]

- **Binary masks**
  \[ M_{ij}^\text{bin} = \begin{cases} 
  0 & v_j \in \text{Pre}(v_i) \text{ or } v_i = v_j, \\
  -\infty & \text{otherwise}. 
\end{cases} \]

- **Probabilistic masks**
  \[ M_{ij}^\text{prob} = \begin{cases} 
  \log P(v_j \in \text{Pre}(v_i) \mid v_i) & v_j \in \text{Pre}(v_i), \\
  0 & v_i = v_j, \\
  -\infty & \text{otherwise}. 
\end{cases} \]
Spoken Language Understanding Results

- Airline Traveling Information System (ATIS)
- Word Error Rate: 15.5%
Spoken Language Understanding Results

- Airline Traveling Information System (ATIS)
  - Word Error Rate: 15.5%
Spoken Language Understanding Results

- Airline Traveling Information System (ATIS)
  - Word Error Rate: 15.5%
Spoken Language Understanding Results

- Airline Traveling Information System (ATIS)
  - Word Error Rate: 15.5%
Spoken Language Understanding Results

• Airline Traveling Information System (ATIS)
• Word Error Rate: 26.3%
What if we do not have ASR lattices?
Solution 2: Learning ASR-Robust Embeddings
ASR-Robust Contextualized Embeddings

- Confusion-Aware Fine-Tuning
  - Supervised

  Acoustic Confusion \( C = \{ w_3^{x_{trf}}, w_2^{x_{asr}} \} \)

  \( x_{trf} : \text{Show me the fares from Dallas to Boston} \)

  \( x_{asr} : \text{Show me affairs from Dallas to Boston} \)

- Unsupervised

  Acoustic Confusion

  Top hypothesis \( x_1 \)

  Alternative hypothesis \( x_2 \)

  List
  Least
  Lift
  all
  * slips
  flights
  tomorrow
  to
  * Monaco
  Morocco
Spoken Language Understanding Results

- Airline Traveling Information System (ATIS)
  - Word Error Rate: 16.4%
Spoken Language Understanding Results

- Airline Traveling Information System (ATIS)
- Word Error Rate: 16.4%
Spoken Language Understanding Results

- Airline Traveling Information System (ATIS)
  - Word Error Rate: 16.4%
Spoken Language Understanding Results

- Airline Traveling Information System (ATIS)
- Word Error Rate: 16.4%
Task-Oriented Dialogue Systems (Young, 2000)

Speech Signal

Text Input

Hypothesis

Language Understanding (LU)
- Domain Identification
- User Intent Detection
- Slot Filling

Semantic Frame
request_movie
genre=action, date=this weekend

System Action/Policy
request_location

Natural Language Generation (NLG)

Dialogue Management (DM)
- Dialogue State Tracking (DST)
- Dialogue Policy

Database

Text response
Where are you located?

Speech Recognition

are there any action movies to see this weekend

Are there any action movies to see this weekend?

Text response
Where are you located?
Structured Data vs Unstructured Data

- Can be displayed in rows, columns and relational databases
- Numbers, dates and strings
- Estimated 20% of enterprise data (Gartner)
- Requires less storage
- Easier to manage and protect with legacy solutions

- Cannot be displayed in rows, columns and relational databases
- Images, audio, video, word processing files, emails, spreadsheets
- Estimated 80% of enterprise data (Gartner)
- Requires more storage
- More difficult to manage and protect with legacy solutions
Structured Data vs Unstructured Data

Can be displayed in rows, columns and relational databases

Can not be displayed in rows, columns and relational databases

Numbers, dates, and strings

Images, audio, video, word processing files, emails, spreadsheets

Estimated 20% of enterprise data (Gartner)

Estimated 80% of enterprise data (Gartner)

Requires less storage

Requires more storage

Easier to manage and protect with legacy solutions

More difficult to manage and protect with legacy solutions
Conversational AI for Unstructured Knowledge

• A machine reads big text data
  • serves as a teacher
• A user can ask questions
  • serves as a student
  • in a conversational manner
→ Conversational QA
FlowDelta: Information Gain in Dialogue Flow

• Idea: model the *difference* of hidden states in multi-turn dialogues
FlowDelta: Information Gain in Dialogue Flow

• Idea: model the difference of hidden states in multi-turn dialogues
Conversational QA Results

- Data: QuAC, CoQA
Conversational QA Results

- Data: QuAC, CoQA
Conversational QA Results

- Data: QuAC, CoQA
## QuAC Leaderboard

<table>
<thead>
<tr>
<th>Rank</th>
<th>Model</th>
<th>F1</th>
<th>HEQQ</th>
<th>HEQD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Human Performance (Choi et al. EMNLP '18)</td>
<td>81.1</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td>TransBERT (single model) Anonymous</td>
<td>69.4</td>
<td>65.4</td>
<td>9.3</td>
</tr>
<tr>
<td>3</td>
<td>Bert-FlowDelta (single model) Anonymous</td>
<td>67.8</td>
<td>63.6</td>
<td>12.1</td>
</tr>
<tr>
<td>3</td>
<td>Context-Aware-BERT (single model) Anonymous</td>
<td>69.6</td>
<td>65.7</td>
<td>8.1</td>
</tr>
</tbody>
</table>
Summary

- **Spoken language embeddings** are needed for better conversational AI
  - Written texts enough for pre-training embeddings
  - Mismatch when applying to spoken language

1) Adapting Transformer to ASR lattices
2) Adapting contextualized embeddings robust to misrecognition

- **Conversational QA** enables unstructured information access
  - **FlowDelta**: information gain in dialogue flow guides better understanding
Yun-Nung (Vivian) Chen
Assistant Professor, National Taiwan University
y.v.chen@ieee.org / http://vivianchen.idv.tw