Breaking open Belief Tracking

Carel van Niekerk
Dialogue System

User Utterance

System Utterance

Spoken Language Understanding

Natural Language Generation

Policy Agent
Hey. I need a restaurant near the city centre.

hello(type=restaurant) 0.6
inform(type=restaurant, location=centre) 0.4

Where would you like the restaurant?

The City Centre!

inform(location=city) 0.6
inform(location=centre) 0.4

Do you want a restaurant?
Dialogue System

User Utterance

System Utterance

Spoken Language Understanding

Belief Tracker

Natural Language Generation

Policy Agent
Hey. I need a restaurant near the city centre.

The City Centre!

Where would you like the restaurant?

To confirm you want a restaurant near the city centre?
Dialogue Belief Tracking

- Belief State – Internal Distribution over states

- State - Information the agents needs to make decisions
  - Capture user intentions
  - Capture history of dialogue

- Aim: Predict Belief State
Datasets

- Two main comparative sets:
  - WOZ 2.0
  - MultiWOZ 2.1 (Most Challenging)

- Metrics
  - Slot accuracy – Proportion of domain-slot-value triplets correctly identified.
  - Joint-goal accuracy – Proportion of turn where all user goals correctly identified.
**WOZ 2.0**

- **Single Domain – Restaurants**
- **1200 Dialogues**

<table>
<thead>
<tr>
<th>Model</th>
<th>Slot Accuracy</th>
<th>Joint-goal Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>NBT</td>
<td>-</td>
<td>84.8%</td>
</tr>
<tr>
<td>MDBT</td>
<td>96.4%</td>
<td>85.5%</td>
</tr>
<tr>
<td>GLAD</td>
<td>97.1%</td>
<td>88.1%</td>
</tr>
<tr>
<td>StateNET</td>
<td>-</td>
<td>88.9%</td>
</tr>
<tr>
<td>GCE</td>
<td>97.4%</td>
<td>88.5%</td>
</tr>
<tr>
<td>GLAD + RC + FS</td>
<td>97.4%</td>
<td>89.2%</td>
</tr>
</tbody>
</table>
### MultiWOZ 2.0

- **Multiple** Domain – 7 domains
- **10000+** Dialogues
- **Richer & Noisier** Dialogues

<table>
<thead>
<tr>
<th>Model</th>
<th>Slot Accuracy</th>
<th>Joint-goal Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>MDBT</td>
<td>89.53%</td>
<td>15.57%</td>
</tr>
<tr>
<td>GCE</td>
<td>98.42%</td>
<td>36.57%</td>
</tr>
<tr>
<td>Neural Reading</td>
<td>98.42%</td>
<td>41.10%</td>
</tr>
<tr>
<td>HYST</td>
<td>96.44%</td>
<td>46.65%</td>
</tr>
<tr>
<td>SUMBT</td>
<td>96.42%</td>
<td>48.62%</td>
</tr>
<tr>
<td>TRADE</td>
<td>96.42%</td>
<td>48.62%</td>
</tr>
</tbody>
</table>
MultiWOZ 2.1

- Corrections:
  - Delayed annotation
  - Incorrect annotation
  - Missed annotations
  - Spelling errors in annotations

<table>
<thead>
<tr>
<th>Model</th>
<th>2.0</th>
<th>2.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neural Reading</td>
<td>41.10%</td>
<td>36.40%</td>
</tr>
<tr>
<td>HYST</td>
<td>44.24%</td>
<td>38.10%</td>
</tr>
<tr>
<td>TRADE</td>
<td>48.62%</td>
<td>45.60%</td>
</tr>
</tbody>
</table>
Generative vs Discriminative

Generative

The future observations are generated by the current state.

\[ P(sys, usr, slu|state) \]
**Generative vs Discriminative**

**Discriminative**

\[ \mathbb{P}(\text{state}|\text{features}) \]

**Discriminate** between the possible **states** using **features** of the dialogue.
Generative vs Discriminative

- **Generative models:**
  - Assumes turns are *independent* given the state.

- **Discriminative models:**
  - *No assumptions* about the independence.
  - **Outperforms** Generative models.
Models without a independent SLU

- Independent SLU Problems:
  - Accumulation of **errors**
  - Requires **additional** annotated training **data**.

Combining the **SLU** and **Feature extractors** into one.
Dialogue System

User Utterance

System Utterance

Belief Tracker

Natural Language Generation

Policy Agent
Delexicalisation for single SLU and Belief Tracker model.
- Requires large dictionaries of semantic lexicons.

Word Embeddings and Feature Extractors
- Convolutional Neural Networks
- Recurrent cell NN
- More scalable
- Equivalent or better performance
Fully Statistical Trackers

- Statistical Dialogue Trackers:
  - Recurrent Cell
  - More adaptable
  - Superior performance
Statistical Discriminative Belief Tracking

Turn Utterances

Utterance encoder / Feature extractor

\[ f \]

Statistical Prediction Model

\[ \mathbb{P}_{1:t}(s|f) \]

Statistical Update Model
Multi-Domain Belief Tracker

Overview:

- Feature Extractor
- State Prediction Model
- Statistical Update Model
- Ontology (d-s-v)
Multi-Domain Belief Tracker

Utterance Encoder

- **Semantic similarity** to identifies the presence of a state in a utterance.

- **Slot-Value Features:**
  - **User confirm** (System slot-value + User affirm)
    
    *System: So you want a *restaurant* near the *centre* of town?*

    *User: Yes*
Multi-Domain Belief Tracker

Utterance Encoder

- **User request** (System slot + User value)
  
  System: Where would you like the hotel to be?

  User: Near the Rhine river.

- **User inform** (User slot-value)

  User: I need a taxi to the airport at 10.
Multi-Domain Belief Tracker

Utterance Encoder

System Utterance

User Utterance

Ontology

Domain Encoder

Slot-Value Encoder

Input Encoders
Multi-Domain Belief Tracker

Statistical Prediction Model

\[ \mathbb{P}_t(d) \]

\[ \sum \]

\[ \mathbb{P}_t(s, v) \]
Multi-Domain Belief Tracker

Statistical Prediction Model

- Two **independent** models. **Share knowledge** across domains.

- **Shares parameters** across all ontology terms -> **Scalable**

- **Multi-class** classification - individual **binary** classification
  - Allows **adaption** to new domains
Multi-Domain Belief Tracker

Statistical Update Model

\[ p_t(s, v) \]

Split by slots

\[ p_t(v) \]

\[ p_{1:t}(v) \]

RNN with Memory cell

\[ p_{1:t}(v) \]

\[ \theta \]

Model parameters

\[ \bigoplus \]

\[ p_{1:t}(s, v) \]
Multi-Domain Belief Tracker

Statistical Update Model

\[ \mathbb{P}_t(d) \]

\[ \mathbb{P}_{1:t}(d) \]

\[ \mathbb{P}_{1:t}(s, v) \]

\[ \mathbb{P}_{1:t}(d, s, v) \]

RNN with Memory cell

Independence
Multi-Domain Belief Tracker

Overview

- **Shortcomings** **Adapting**
  - Assumes Known Ontology – Scalability Issues

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Slot Accuracy</th>
<th>Joint-goal Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>WOZ 2.0</td>
<td>96.4%</td>
<td>85.5%</td>
</tr>
<tr>
<td>MultiWOZ 2.0</td>
<td>89.53%</td>
<td>15.57%</td>
</tr>
</tbody>
</table>
Improved Feature Extraction

- Slot *conditioned*

- *Global* parameter sharing

- Self-Attention *contextual* embeddings
Overview:

Globally-Conditioned Encoder (GCE)

Slot Conditioned Feature Extraction

Global Encoder

State Prediction

Ontology (d-s-v)
Globally-Conditioned Encoder (GCE)

Utterance Encoder

- Bidirectional LSTM model $\rightarrow$ Contextual token embeddings

- Convolutional self-attention $\rightarrow$ Contextual utterance embedding.

Embeddings:

- Current User Utterance
- Previous j System acts
- Value candidates
Globally-Conditioned Encoder (GCE)

Utterance Encoder – Slot Conditioned token embeddings

Utterance
\[ e_1, e_2, \ldots, e_n \]

Slot
\[ c_1, c_2, \ldots, c_n \]

\[ S_K \]
Globally-Conditioned Encoder (GCE)

Utterance Encoder – Bi-directional LSTM with Self-Attention

\[ h_1, h_2, \ldots, h_n \]

\[ c_1, c_2, \ldots, c_n \]

\[ H_K \]

\[ s_K \]

\[ a_1, a_2, \ldots, a_n \]

\[ c_k \]

\[ \sum c_k \]

\[ \varnothing \]

\[ \oplus \]

\[ \odot \]
The Globally-Conditioned Encoder (GCE)
The Globally-Conditioned Encoder (GCE)

User Utterance

System Actions

Slot Values

Slot

GCE

User Utterance

System Actions

Slot Values

Features

$H_{utt}, C_{utt}$

$H_{act}, C_{act}$

$H_{val}, C_{val}$
Globally-Conditioned Encoder (GCE)

Statistical Prediction Model

- **Utterance scoring** model
  - **User token** embeddings + **value** embeddings
  - Degree to which the slot-value pair was **mentioned** by the **user**.

*User: I want a **Italian** restaurant.*
Globally-Conditioned Encoder (GCE)

Statistical Prediction Model

- **Action scoring** model
  - **System** utterance embeddings + **User** utterance + **value**
  - Degree to which the slot-value was **mentioned** by the **system**.

System: *Would you like the restaurant to be in the *east* of town?*

**Location not east**

User: *No.*
Globally-Conditioned Encoder (GCE)

**Statistical Prediction Model**

- **Utterance Scoring Model**
  - $h_{1}^{utt}$, $h_{2}^{utt}$, $h_{m}^{utt}$
  - $\sum_{\text{Linear}} y_{utt}$
  - $c_{val}$

- **Action Scoring Model**
  - $c_{1}^{act}$, $c_{2}^{act}$, $c_{f}^{act}$
  - $\sum_{y_{act}}$
  - $c_{val}$

$\mathbb{P}_{1:t}(s,v)$
Overview

- **Limitations:**
  - *Past J* system utterances used.
  - Assumes Known Ontology – Scalability Issues

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Slot Accuracy</th>
<th>Joint-goal Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>WOZ 2.0</td>
<td>97.38%</td>
<td>88.51%</td>
</tr>
<tr>
<td>MultiWOZ 2.0</td>
<td>98.42%</td>
<td>36.57%</td>
</tr>
</tbody>
</table>
Dynamic Ontology

- Value generation models
Transferable Dialogue State Generator (TRADE)

Overview:

- History based Feature Extractor
- Domain – slot Conditioned encoder
- State Prediction

Ontology (d-s) No Values
Transferable Dialogue State Generator (TRADE)

Utterance Encoder

- Bidirectional **GRU** model $\rightarrow$ Contextual **token** embeddings

- Domain-slot **conditioned** GRU $\rightarrow$ Contextual **history** embedding.

- Encodes past $l$ turns **jointly**.
Transferable Dialogue State Generator (TRADE)

Utterance Encoder

User Utterance \( t - l \)  
System Utterance \( t - l \)

User Utterance \( t - 1 \)  
System Utterance \( t - 1 \)

User Utterance \( t \)  
System Utterance \( t \)

\[ \bigoplus \]

\( t_1 \ t_2 \ ... \ t_m \)
Transferable Dialogue State Generator (TRADE)

Utterance Encoder

\[ t_1, t_2, \ldots, t_m \]

\[ h_1, h_2, h_m \]

\[ H_t \]

Domain \[ d_j \]  \[ s_k \]

\[ f_{jK} \]

\[ h_{jk}^{dec} \]
Transferable Dialogue State Generator (TRADE)

Statistical Prediction Model – The TRADEMARK!

\[ h_{jk}^{dec} \]

\[ f_{jk} \]

\[ E \]

\[ H_t \]

\[ \mathbb{P}^{vocab}_{jk} \]

\[ \mathbb{P}^{hist}_{jk} \]

\[ \mathbb{P}^{gen}_{jk} \]

\[ \mathbb{P}^{final}_{jk} \]

\[ \mathbb{P}^{final}_{jk} \]

\[ W_g \]

\[ G_j \]

Probability of Significance
Overview

<table>
<thead>
<tr>
<th>Model</th>
<th>Slot Accuracy</th>
<th>Joint-goal Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>GCE</td>
<td>98.42%</td>
<td>36.57%</td>
</tr>
<tr>
<td>TRADE</td>
<td>96.42%</td>
<td>48.62%</td>
</tr>
</tbody>
</table>

- **Positives:**
  - **Generates values** with great success.
  - Shows promise with **few-shot** learning.

- **Limitation:**
  - **Zero-shot** performance not great.
  - **Past L** turns used. (Inefficient)
  - Requires domain-slots to be defined
Improved latent space mappings

- **Transformer** based **contextual** mappings

- Truly **statistical latent** space belief tracker
Slot-Utterance Matching for Universal BT(SUMBT)

Overview:

Contextual embeddings for utterances and ontology terms

Multi-Head attention and RNN based latent belief tracker

Statistical Prediction Model

Ontology (d-s-v)
Utterance Encoder

- Two fine-tuned BERT models:
  - **Utterance** embedding
  - **Domain-slot-value** embedding
- Use of **contextual** embeddings
Slot-Utterance Matching for Universal BT(SUMBT)

Utterance Encoder

[CLS] [User Utterance] [SEP]  
[System Utterance] [SEP]

\[ h_0 \  h_1 \  h_2 \ ... \  h_m \]

\[ H_t \]

[CLS] [Domain-Slot] [SEP]

\[ h_{[CLS]} \]

\[ q_{slot} \]

[CLS] [Value] [SEP]

\[ h_{[CLS]} \]

\[ y_t^s \]
Multi-head Attention

- Input:
  - **Query** – What is the encoder *asking*?
  - **Key** – The state of the encoder. Key *unlocks* the *answer*.
  - **Value** – How much *attention* should we give?

- Passed through *multiple attention heads*.
- Returns *context* embedding.
Statistical Update Model

- **Query** = *System* utterance
- **Key** = *User* utterance
- **Value** = *Domain-slot*

The **attention heads** provides the **context** of the dialogue.

**RNN** tracks context over dialogue.

Provides a **estimated contextual value** embedding.
Slot-Utterance Matching for Universal BT(SUMBT)

Statistical Update Model

\[ H_t \]

\[ K \quad Q \quad V \]

Multi-head Attention

\[ h_t^s \]

RNN

\[ d_t^s \]

LayerNorm

\[ \hat{y}_t^s \]

Estimated value embedding
Statistical Prediction Model

\[ \hat{y}_t^s \rightarrow \text{distance} \rightarrow y_t^s \]

\[ \mathbb{P}_{1:t}(d, s, v) \]

\[ d_t^1, d_t^2, d_t^m \]
Slot-Utterance Matching for Universal BT(SUMBT)

Overview

<table>
<thead>
<tr>
<th>Model</th>
<th>Slot Accuracy</th>
<th>Joint-goal Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRADE</td>
<td>96.42%</td>
<td>48.62%</td>
</tr>
<tr>
<td>SUBMT</td>
<td>96.44%</td>
<td>46.65%</td>
</tr>
</tbody>
</table>

- **Positives:**
  - True latent space fully statistical belief tracker.

- **Limitation:**
  - Very large model. Expensive to train.
  - Requires ontology to be defined.
Slot-Utterance Matching for Universal BT(SUMBT)

Overview

Dialog Example

Turn 1, U: Hello, I'm looking for a restaurant, either Mediterranean or Indian, it must be reasonably priced though.

Turn 2, S: Sorry, we don't have any matching restaurants.

U: How about Indian?

Turn 3, S: We have plenty of Indian restaurants. Is there a particular place you'd like to stay in?

U: I have no preference for the location, I just need an address and phone number.
Success stories

- Word embeddings
  - Improved performance
  - Better Scalability
  - Successes of contextual embeddings

- Recurrent models
  - Fully statistical
  - Learns cross-turn dependencies
  - No rules needed
Success stories

- Semantic similarity
  - Leveraged from word **embeddings**
  - Does the user/system **mention** a concept?
- Projecting **dialogue history** onto latent **representation**.
- Knowledge sharing
  - **Parameter** sharing
  - Domains **share** slots, Slots **share** values
  - Improved **performance**
  - **Adaptability**
  - **Scalability**
Success stories

- Value generation methods
  - More **scalable**
  - Improved **performance**
  - **Adaptability**
  - Ontology only needs **domains and slots**
- **Joint** Belief Tracking and Policy Learning
  - Promises to improve **performance**
Shortcomings

- **Predefined ontology**
  - Not scalable
  - Not possible - new values can constantly be added (Restaurant names)

- **Zero-shot adaption**
  - Very little success

- **Rare slot-value combinations**
  - Difficulty accurately predicting these
  - Negatively impacts joint goal accuracy
  - Limits adaptability
Shortcomings

- **Utilising non-dialogue data**
  - Utilising non dialogue data through word embeddings.

- **Representation** of states
  - How to represent states
  - Is **domain-slot-value** sufficient
  - Could **graph** structures states be **embedded**
  - **Efficient** use of data for **rare states**

- Joint goal on **rich and noisy** datasets
Resources

- **HyST: A Hybrid Approach for Flexible and Accurate Dialogue State Tracking**
  R Goel, S Paul and D Hakkani-Tür, 2019

- **Neural Belief Tracker: Data-Driven Dialogue State Tracking**
  N Mrkšić, D Séaghdha, T Wen, B Thomson and S Young 2016

- **The Dialog State Tracking Challenge: A Review**
  JD Williams, A Raux, D Ramachandran, and A Black 2013

- **SUMBT: Slot-Utterance Matching for Universal and Scalable Belief Tracking**
  H Lee, J Lee and T Kim 2019

- **Dialog State Tracking: A Neural Reading Comprehension Approach**
  S Gao, A Sethi, S Aggarwal, T Chung and D Hakkani-Tür 2019

- **Improving Dialogue State Tracking by Discerning the Relevant Context**
  S Sharma, PK Choubey and R Huang 2019

- **Large-Scale Multi-Domain Belief Tracking with Knowledge Sharing**
  O Ramadan, P Budzianowski and M Gašić 2018
Resources

- **Toward Scalable Neural Dialogue State Tracking Model**
  E Nouri and E Hosseini-Asl 2018

- **Transferable Multi-Domain State Generator for Task-Oriented Dialogue Systems**
  A Madotto, E Hosseini-Asl and C Xiong 2018

- **Towards Universal Dialogue State Tracking**
  L Ren, K Xie, L Chen and K Yu 2019

- **BERT-DST : Scalable End-to-End Dialogue State Tracking with Bidirectional Encoder Representations from Transformer**
  G Chao and I Lane 2019

- **Global-Locally Self-Attentive for Dialogue State Tracking**
  V Zhong, C Xiong and R Socher 2019

- **Fully Statistical Neural Belief Tracking**
  N Mrkšić Nikola and I Vulić 2018

- **Word-Based Dialog State Tracking with Recurrent Neural Networks**
  M Henderson, B Thomson and S Young 2015