Building A User-Centric and Content-Driven Socialbot

Hao Fang

Committee: Mari Ostendorf (Chair)  Hannaneh Hajishirzi
Leah M. Ceccarelli (GSR)          Eve Riskin
Yejin Choi             Geoffrey Zweig
Agenda

- Background
- Sounding Board System – 2017 Alexa Prize Winner
- A Graph-Based Document Representation for Dialog Control
- Multi-Level Evaluation for Socialbot Conversations
- Summary and Future Directions
Agenda

- Background
- Sounding Board System – 2017 Alexa Prize Winner
- A Graph-Based Document Representation for Dialog Control
- Multi-Level Evaluation for Socialbot Conversations
- Summary and Future Directions
Sci-Fi Movies
Daily Life
Types of Conversational AI

**Socialbots**

“converse coherently and engagingly with humans on popular topics and current events”

**Task Definition**

- **Task-oriented**
- **Non-task-oriented**

**Domain Coverage**

- **Single-domain**
- **Multi-domain**
- **Open-domain**

**Dialog Initiative**

- **System-initiative**
- **User-initiative**
- **Mixed-initiative**
Socialbot Applications

- Entertainment, education, healthcare, companionship, ...
- A conversational gateway to online content
Agenda

○ Background

○ Sounding Board System – 2017 Alexa Prize Winner

○ A Graph-Based Document Representation for Dialog Control

○ Multi-Level Evaluation for Socialbot Conversations

○ Summary and Future Directions
Design Objectives

User-Centric

• Users can control the dialog flow and switch topics at any time
• Bot responses are adapted to acknowledge user reactions

Content-Driven

• Content cover the wide range of user interests
• Dialog strategies to lead or contribute to the dialog flow
2017 Alexa Prize Finals
Dialog Control for Many Miniskills?

Conversation Activities (Miniskills)

- Greet
- List Topics
- Tell Fun Facts
- Tell Jokes
- Tell Headlines
- Discuss Movies
- Personality Test
- ...
Hierarchical Dialog Management

- Dialog Context Tracker
  - dialog state, topic/content/miniskill history, user personality

- Master Dialog Manager
  - miniskill polling
  - topic and miniskill backoff

- Miniskill Dialog Managers
  - miniskill dialog control as a finite-state machine
  - retrieve content & build response plan
Social Chat Knowledge

An important type of social chat knowledge is online content.

How to organize content to facilitate the dialog control?

A framework that allows dialog control to be defined in a consistent way.
Knowledge Graph

- **Nodes**
  - content post (fact, movie, news article, ...)
  - topic (entity or generic topic)

- **Relational edges between content post and topic**
  - topic mention (NER, noun phrase extraction)
  - category tag (Reddit meta-information)
  - movie name, genre, director, actor (IMDB)

- **Dialog Control**: move along edges

---

**UT Austin and Google AI** use machine learning on data from NASA's Kepler Space Telescope to discover an eighth *planet* circling a *distant star*.
Agenda

- Background
- Sounding Board System – 2017 Alexa Prize Winner
- A Graph-Based Document Representation for Dialog Control
- Multi-Level Evaluation for Socialbot Conversations
- Summary and Future Directions
Motivation

- Dialog control defined based on moves on the graph
  - lead the conversation
  - handle user initiatives

- Challenges for unstructured document (e.g., news articles)
  - not all sentences are equally interesting to a listener
  - need to figure out a coherent presenting order
  - answer questions about the document
  - need a smooth transition between sentences
  - handle entity-based information seeking requests
  - handle opinion-seeking requests
Graph-Based Document Representation

Entity 1
Entity 2
Entity 3

Storytelling Chain
Sent 1
Sent 2
Sent 3
Sent 4

Subject
Comment
Answer

Opinion 1
Opinion 2
Question 1
Question 2
Question 3
Document Representation Construction

Text Pre-processing
Sentence Node Creation
Entity Node Creation
Subject Edge Creation
Storytelling Chain Creation
Question Generation
Comment Collection

NLP Tools

Tokenization
Sentence Split
Sentence Filtering
Part-of-Speech Tagging
Constituency Parsing
Named Entity Recognition
Entity Linking
Coreference Resolution
Dependency Parsing
Storytelling Chain Creation

- **Problem formulation**
  - context sentence sequence \((s_1, s_2, ..., s_L)\)
  - candidate sentence set \(\{y_1, y_2, ..., y_N\}\)
  - candidate sentence chain \((y_i | s_1, s_2, ..., s_L)\)

- **Data collection:** 550 news articles
  - Train/Validation/Test: 3/1/1 based on article ID

Data:
- \(L=1, N=4\):
  - Positive: 662
  - Negative: 1538
- \(L=2, N=3\):
  - Positive: 865
  - Negative: 1064

|       | Positive | Negative |
|-------|----------|----------|
| Sent 1|           |          |
| Sent 2|           |          |
| Sent 3|           |          |

chart: bars with labels for positive and negative counts by sentence length and number of sentences.
Model and Features

- **Model**: binary logistic regression
  - input: candidate sentence chain \((y_i \mid s_1, s_2, \ldots, s_L)\)
  - output: probability score \(s(y_i \mid s_1, s_2, \ldots, s_L) \in \mathbb{R}^{[0,1]}\)

- **Features**
  - **SentImportance**: \(r(y_i \mid D)\)
  - **SentDistance**: \(d(y_i \mid s_1, s_2, \ldots, s_L) = SentIdx(y_i) - SentIdx(s_L)\)
  - **SentEmbedding**: \(e(y_i)\)
  - **ChainEmbedding**: \(c(y_i \mid s_1, s_2, \ldots, s_L)\)

TextRank unsupervised summarization on the document \(D\)

Pre-trained BERT

used for ranking sentences given \(s_1, s_2, \ldots, s_L\)
Test Set Results

- Next sentence is not always good

% the highest-ranked sentence has a positive label
Test Set Results

sentence embedding alone may capture some features about importance / style (e.g., length, informativeness)

% the highest-ranked sentence has a positive label
% the highest-ranked sentence has a positive label

sentence importance (document context) is very useful
Test Set Results

dialog context is important as the chain gets longer

% the highest-ranked sentence has a positive label

| SentDistance | SentEmbedding | SentImportance | ChainEmbedding | All |
|--------------|--------------|----------------|----------------|-----|
| L=1, N=4     | 62.1         | 63.2           | 64.8           | 66.3 |
| L=2, N=3     | 69.3         | 71.9           | 73.7           | 70.2 |

+2.7  +4.4
Test Set Results

using all features (2050-dimensional) overfits for L=2 (1239 training samples)

% the highest-ranked sentence has a positive label

|          | L=1, N=4 | L=2, N=3 |
|----------|----------|----------|
| SentDistance | 54.7     | 62.3     |
| SentEmbedding | 62.1     | 69.3     |
| SentImportance | 63.2     | 71.9     |
| ChainEmbedding | 64.8     | 73.7     |
| All       | 66.3     | 70.2     |

Note: The highest-rank sentence has a positive label.
Question Generation

- Dependency Parsing
- Dependent Selection for Answer
- Question Type Classification
- Clause/Question Planning
- Clause/Question Realization

Question Interestingness/Importance
- Hand-Crafted Decision Tree
- Template-Based Planning
- Dependency-Based Realization

Universal Dependencies

Question 1

Question 2
Among leading U.S. carriers, Sprint was the only one to throttle Skype, the study found.
Evaluation of Generated Questions

- As a transition clause for introducing Sent2 given Sent1
  - *do you want to know _____?*

- 4 question generation methods
  - generic: *more about this article*
  - constituency-based (Heilman, 2011)
  - dependency-based
  - human-written

- Human judgments on question pairs (A, B, cannot tell)
  - 134 sentences, 5 judgments per pair
Overall Quality

**vs. Generic**
- Win: Constituency - 59, Dependency - 35
- Tie: Constituency - 44, Dependency - 52
- Loss: Constituency - 6, Dependency - 4

**vs. Human**
- Win: Constituency - 73, Dependency - 49
- Tie: Constituency - 18, Dependency - 7
- Loss: Constituency - 9, Dependency - 44

*dependency-based outperforms constituency-based, but does not achieve “human performance”*
dependency-based method generates much more informative questions (better than human)
Transition Smoothness

vs. Generic
- Win: 73%
- Tie: 58%
- Loss: 5%

vs. Human
- Win: 79%
- Tie: 57%
- Loss: 5%

dialog context is important!
Agenda

- Background
- Sounding Board System – 2017 Alexa Prize Winner
- A Graph-Based Document Representation for Dialog Control
- Multi-Level Evaluation for Socialbot Conversations
- Summary and Future Directions
Motivation: Evaluation & Diagnosis

- Users only give an optional conversation rating
- Aspects that influence user ratings?
  - prior model-free metrics do not outperform conversation length
- Structure of socialbot conversations?
  - prior models of dialog structure are not suitable
- Diagnosis calls for more than conversation scores
  - a conversation can involve good and bad segments/topics/policies/...
Conversation Acts for User Turns

- AskQuestion
- RequestHelpOrRepeat
- ProposeTopic
- AcceptTopic
- RejectTopic
- FollowAndNonNegative

**Rule-Base Tagging**

- InterestedInContent
- NotInterestedInContent
- PositiveToContent
- NegativeToContent
- PositiveToBot
- NegativeToBot

**Model-Base Tagging**
Correlation Analysis

For each act $A$
- number of turns $N_A$
- percentage of turns $P_A$

$N_A$ cannot tell any negative correlation

Conversation Length $r = 0.15$

Pearson $r$ with conversation user ratings

$r_{num}$  $r_{pct}$  -0.2

- AskQuestion
- RequestHelpOrRepeat
- ProposeTopic
- AcceptTopic
- RejectTopic
- FollowAndNonNegative
- InterestedInContent
- NotInterestedInContent
- PositiveToContent
- NegativeToContent
- PositiveToBot
- NegativeToBot

$r_{num}$ and $r_{pct}$ for each act $A$ show different correlations with user ratings.
It is a good sign that user follows the conversation flow when the bot is the primary speaker.

Design, learn, & maintain engaging conversation flows (≠ system-initiative)
Correlation Analysis

AskQuestion and ProposeTopic slightly impact user ratings in the negative direction.

Improve the bot’s capability of handling user questions and topic requests.
Limitations

- Conversation ratings and conversation-act-based metrics do not tell
  - which topics are handled badly by the bot
  - which dialog policies need improvement
  - which content sources have less suitable quality

- Segment-level scores can tell us more, but
  - how to segment a socialbot conversation?
  - how to compute a segment-level score?
Hierarchical Dialog Model

A conversation is a sequence of topical subdialogs, each of which is a sequence of microsegments, each of which contains posts.

Subdialog: SmallTalk, Cats, Batman, Robots

Microsegment: Batman vs. Superman, Henry Cavill, Ben Affleck

Post: fun fact, amusing thought, news headline
Automatic Segment Scoring

- Labels: conversation-level user ratings

- Features
  - conversation-act-based metrics
  - other features such as bag-of-words, verbosity, ...

- Two different model hypotheses
  - H1: segment scores are predicted just like conversation scores
  - H2: a conversation score is some aggregation of segment scores
Automatic Segment Scoring

- **H1: Linear Scoring Model**
  - segment score = \( f(\text{segment features}) \)
  - conversation score = \( f(\text{conversation features}) \)
  - \( f(x_1, \ldots, x_d) = \sum_{i=1}^{d} u_i x_i + u_0 \)

- **H2: BiLSTM Scoring Model**
  - segment score \( s_t = h_t(\text{segment features}) \)
    - \( h_1, h_2, \ldots, h_T \): BiLSTM over individual segments
    - \( s_{\text{mean}} = \text{mean}(s_1, s_2, \ldots, s_T), \ldots \)
  - conversation score = \( g(s_{\text{mean}}, s_{\text{max}}, s_{\text{min}}) \)
    - \( g(s_{\text{mean}}, s_{\text{max}}, s_{\text{min}}) = \sum v_i s_m + v_0 \)
Evaluation of Subdialog Scores

- Human judgments on subdialog pairs (A, B)
  - 250 within-conversation pairs (same user)
  - 250 cross-conversation pairs (same topic)
  - 5 judgments per pair

- Spearman rank correlation $\rho$ between $x$ and $y$
  - $x = \text{votes on A} - \text{votes on B}$
  - $y = \text{score of A} - \text{score of B}$

BiLSTM may learn features about the user by using surrounding context

![Graph showing Spearman $\rho$ for NumTurns, Linear, and Subdialog BiLSTM]
Agenda

- Background
- Sounding Board System – 2017 Alexa Prize Winner
- A Graph-Based Document Representation for Dialog Control
- Multi-Level Evaluation for Socialbot Conversations
- Summary and Future Directions
Summary: Sounding Board System

- A mixed-initiative and open-domain socialbot
  - user-centric and content-driven dialog strategies
  - it is a new and fast-growing area and we are one of the pioneers
  - several strategies have influenced 2018 socialbots

- System architecture
  - a hierarchical DM framework for efficient dialog control
  - social chat knowledge graph
  - several 2018 socialbots follow a similar DM architecture and acknowledge the importance of content
Summary: Graph-Based Representation

- Extended conversations grounded on a document
  - a graph-based document representation
  - bridge machine reading and dialog control

- Automatic document representation construction
  - a model for storytelling chain creation
  - an unsupervised dependency-based question generation
  - new NLP tasks that emphasize both dialog context and sentence/question interestingness
Summary: Multi-Level Evaluation

- In-depth analysis on aspects that influence user ratings
  - conversation acts for socialbot conversations
  - valuable insights for socialbot evaluation
  - better metrics than the conversation length baseline

- Automatic segment scoring for system diagnosis
  - a new hierarchical dialog model for socialbot conversations
  - two scoring models with different hypotheses for segments scores
Future Directions

- Open-domain and mixed-initiative conversational AI
  - large-scale knowledge base & computational dialog control
  - switch between two roles (primary speaker & active listener)

- Document/content analysis for conversational AI
  - unstructured text to structured representation
  - understand interestingness and socially appropriateness

- Human-in-the-loop for conversational AI
  - data collection & evaluation
  - crowd-powered system
Acknowledgements

- PhD Advisor: Mari Ostendorf
- Committee Members
  - Leah M. Ceccarelli, Yejin Choi, Hannaneh Hajishirzi, Eve Riskin, Geoffrey Zweig
- Sounding Board Team & TIAL Lab Members & Alumni
  - Hao Cheng, Elizabeth Clark, Ari Holtzman, Maarten Sap, Noah Smith
  - Amittai Axelrod, Sangyun Hahn, Ji He, Jingyong Hou, Brian Hutchinson, Aaron Jaech, Yuzong Liu, Roy Lu, Yi Luan, Kevin Lybarger, Alex Marin, Julie Medero, Farah Nadeem, Nicole Nichols, Sining Sun, Trang Tran, Ellen Wu, Victoria Zayats
- Mentors and collaborators during Internships
- Amazon Alexa Prize organizers
Thank You