Building Conversational Agents using Deep Learning

*Advanced Topics in AI (Spring 2017)*
*IIT Delhi*

Manoj Kumar Chinnakotla
Senior Applied Scientist
Artificial Intelligence and Research (AI&R)
Microsoft, India

Acknowledgment
Agenda of the Talk

- Conversation Agents
  - Various dimensions of problem
  - Challenges involved
- Taxonomy of Techniques
- Retrieval Based Models
- Generative Models
- Summary
What is a Conversational Agent?

• A conversational agent or chat-bot exploits NLP/ML technologies to engage users in a natural language conversation to
  • Accomplish a task – information seeking (when is the next train to vizag?), services (book me a ticket to a movie)
  • Engage/Entertain the user in some creative ways (tell me a joke, say something interesting about an actor)
Looking for a messenger bag. Using it for work mainly.

Do you have preference on material or color?

I prefer leather, but the style should be casual and versatile.

This leather bag is sophisticated, but not flashy. It also has 8 compartments to keep your things organized inside.

JACK SPADE
Carlton leather bag
$405.10

I need headphones for my 4 year old. Something sturdy and safe for kids.

Sounds good. Give me a few minutes to find the best options.

Perfect, thank you!

These over-the-ear headphones are fit for kids and great for travelling, too!

LILGADGETS
Untangled bluetooth headphones
$49.95
Open Domain vs. Closed Domain

• Open Domain
  • There isn’t a well-defined goal or intention usually
  • For example: conversations on social-media like twitter, fb, reddit
  • Crucial factor: Need good context understanding and world knowledge to generate meaningful and relevant responses

• Closed Domain
  • The conversation has a clear and very specific goal
  • The user is interested in efficiently achieving it and hence the potential inputs/outputs gets limited relatively
  • For Example: Technical Customer Support, Shopping Assistants
Long vs. Short Conversations

• Short Conversations
  • The expected number of turns in the dialogue is usually less
  • For Example: factoid question answering (is there a train from A to B?)
  • No need to maintain the entire context

• Long Conversations
  • Multi-turn conversations which may be long
  • For Example: Discussion about movie scene
Challenges in building Chatbots

• Incorporating Context
  • Keep tracking of what information has been exchanged, in which context etc.
  • Especially important in long conversations

User: Can you please tell me which flights from London are expected this evening after six

Bot: Sure, let me check

Bot: Here are the flights: EK 201, EK 522

User: and which ones tomorrow morning before twelve?
Challenges in building Chatbots (Contd..)

• Ensuring a coherent personality
  • Producing answers which are consistent factually, sentiment-wise
  • May sound simple but this is an active area of research 😊

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>message</strong></td>
<td>Where do you live now?</td>
</tr>
<tr>
<td><strong>response</strong></td>
<td>I live in Los Angeles.</td>
</tr>
<tr>
<td><strong>message</strong></td>
<td>In which city do you live now?</td>
</tr>
<tr>
<td><strong>response</strong></td>
<td>I live in Madrid.</td>
</tr>
<tr>
<td><strong>message</strong></td>
<td>In which country do you live now?</td>
</tr>
<tr>
<td><strong>response</strong></td>
<td>England, you?</td>
</tr>
</tbody>
</table>
Challenges in building Chatbots (Contd..)

- Evaluation different for task-oriented chat and general chit-chat
- For task-oriented chat, % of success in fulfilling task during the conversation
  - Need human judgment and expensive
- In non-goal oriented chit-chat, improvement in engagement is the key metric
- Recent research shows that many unsupervised evaluation metrics don’t correlate well with human judgments
- Intention and Diversity
  - For many messages, responses such as “That’s great!” or “I don’t know” work for a lot of input cases
  - This is too boring for users. How do you ensure diversity?
A Taxonomy of Models

• Mine a large enough message-response conversation corpus
  • Sources: Twitter, Reddit, Facebook
  • Retrieval based Models
    • Given a user utterance, find the semantically most similar message and return its response
    • Can only output fixed-set of responses available in the corpus
  • Generative Models
    • Learn a Machine Learning (ML) based translation model to “translate” a given user message to an appropriate response in the same language
    • Can generate a totally new response not available in the corpus
First Stage:
- Identify the filter set
- POS Tagging, Hierarchical Clustering, Entropy Analysis
- Two million to approximately two thousand

Second Stage:
- Ranking using 300 features
- Grammatical features, high-level features and concept-based features
- Ranking done using MART

Third Stage:
- Personalize the top response
- Transfer Learning
Training Data

• Twitter message and response (Positive)
• Twitter message and random response (Negative)
• For personalization phase
  • Designed a web app with online learning algorithm
  • Show a twitter message and a predicted response to the user
  • The user rates (upvotes, downvotes) the message, response pairs
  • Online updation of weights happen
  • They collected 1K examples using the above strategy
MART Features

• Employ different features between the query and response
  • Content-Oriented Features
    • Topic Modeling
    • Query Refinement
    • Sentence Paraphrasing and Substitution
    • Wikipedia Article Names
  • POS features
  • Dictionaries and Synonyms
    • Wikipedia similarity features
• Train a MART using the above features and training data
Transfer AdaBoost

• Additional features used here
  • Topic modeling based features on twitter employing – HMM and LDA
  • Query refinement tables
  • Sentence paraphrasing techniques

• For training TrAdaBoost
  • 400 +ve, 400 –ve as destination domain
  • 2M twitter conv. as source domain
  • Error on test set drops from 30.2% to 26.4%
Suggested Responses for Query “I play football a lot”:

R: how long have you been playing? i’ve been on 2 years
R: don’t forget the final test match
R: from vegas! hawes may not play in summer league. westphal says he ’ll come mon. to “support ... teammates”
R: i want to play over the line
R: i got to see that pic
R: i know, but you said you wanted a tattoo with it. if it makes you feel any better i’ll now be in the rain waiting for a bus
R: you are so far behind on that one
R: you work late
R: omg -creams- do you know what i would do to see them live gaah ! i really do n’t know why i creamed but i did
R: area -usr- you rock

Suggested Responses for Query “How is the weather in Seattle?”:

R: perfect weather for a day off
R: wish i was there! i need some cooler weather! !95 in south tx! you ready ? =)
R: snow in august and tornadoes in february would n’t be surprising here
R: cause i did n’t bring my umbrella ! :(
R: rain falling , thunder storming , lightning striking
R: the rain is so nice ...but why do people complain already? we ’ve had record dry spells
R: cold , stressed , bored and can’t find a bathroom . are these the best conditions to sit an exam ? no i think not
R: yes hurt because of friggin the sims 3 ! my guy and gurl sims have finally moved in together ! no cheats , i swear . :p
R: feeling tired and under the weather today . been napping but disturbed dreams
Emulating Human Conversations using Convolutional Neural Network-based IR – Prakash, 2016

User: hi, good evening
Bot: hey..good evening..
User: how are you
Bot: i’m fine and you?
User: i am also good, just returned from office
Bot: where do you work?
User: all that later, i have something to tell you
Bot: what?
User: i like a girl a lot, and want to ask her out
Bot: ask her out anyway
Generative Models – An Overview

• Treat the message “m”, response “r” pairs as translations of each other
• Learn a Machine Translation (MT) system which can translate “m” into response “r”
• We can use an RNN if “m” and “r” were of same length. If not, how?

Sequence to Sequence Learning with Neural Networks – Sutskever et al., Google, Dec. 2014
Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models – Serban et al., 2015
<table>
<thead>
<tr>
<th>Model</th>
<th>Perplexity</th>
<th>Perplexity@U₂</th>
<th>Error-Rate</th>
<th>Error-Rate@U₃</th>
</tr>
</thead>
<tbody>
<tr>
<td>Backoff N-Gram</td>
<td>64.89</td>
<td>65.05</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Modified Kneser-Ney</td>
<td>60.11</td>
<td>54.75</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Absolute Discounting N-Gram</td>
<td>56.98</td>
<td>57.06</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Witten-Bell Discounting N-Gram</td>
<td>53.30</td>
<td>53.34</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>RNN</td>
<td>35.63 ± 0.16</td>
<td>35.30 ± 0.22</td>
<td>66.34% ± 0.06</td>
<td>66.32% ± 0.08</td>
</tr>
<tr>
<td>DCGM-I</td>
<td>36.10 ± 0.17</td>
<td>36.14 ± 0.26</td>
<td>66.44% ± 0.06</td>
<td>66.57% ± 0.10</td>
</tr>
<tr>
<td>HRED</td>
<td>36.59 ± 0.19</td>
<td>36.26 ± 0.29</td>
<td>66.32% ± 0.06</td>
<td>66.32% ± 0.11</td>
</tr>
<tr>
<td>HRED + Word2Vec</td>
<td>33.95 ± 0.16</td>
<td>33.62 ± 0.25</td>
<td>66.06% ± 0.06</td>
<td>66.05% ± 0.09</td>
</tr>
<tr>
<td>RNN + SubTle</td>
<td>27.09 ± 0.13</td>
<td>26.67 ± 0.19</td>
<td>64.10% ± 0.06</td>
<td>64.07% ± 0.10</td>
</tr>
<tr>
<td>HRED + SubTle</td>
<td>27.14 ± 0.12</td>
<td>26.60 ± 0.19</td>
<td>64.10% ± 0.06</td>
<td>64.03% ± 0.10</td>
</tr>
<tr>
<td>HRED-Bi. + SubTle</td>
<td>26.81 ± 0.11</td>
<td>26.31 ± 0.19</td>
<td>63.93% ± 0.06</td>
<td>63.91% ± 0.09</td>
</tr>
</tbody>
</table>
A Persona-Based Neural Conversation Model – Li et al., June 2016

• What is a persona?
  • Composite of various elements of identity
  • Background facts, user profile, language behaviour and interaction style

4 Layer SEQ2SEQ model trained on 25 Million Twitter conversation snippets
A Persona-Based Neural Conversation Model – Li et al., June 2016
A Diversity Promoting Objective Function for Neural Conversation Models – Li et al., June 2016

\[
\hat{T} = \arg \max_T \{ \log p(T|S) \}
\]

\[
\hat{T} = \arg \max_T \{ \log p(T|S) - \log p(T) \}
\]

| message | SEQ2SEQ | MMI $P(T|S) - \lambda P(T)$ |
|---------|---------|-----------------------------|
| I did not get the report from an mi-6 agent. | I don’t think that’s a good idea. | You did the right thing, did you? |
| You haven’t been given an assignment in this case. | I don’t know what you are talking about. | I’ve been looking all over for you. |
| I’m losing my grip. | I don’t know what you are talking about. | I’m the only one in the world. |
## Retrieval Based vs. Generative Models

<table>
<thead>
<tr>
<th>Retrieval</th>
<th>Generative</th>
</tr>
</thead>
<tbody>
<tr>
<td>No grammatical mistakes possible</td>
<td>Ungrammatical responses may be generated</td>
</tr>
<tr>
<td>Query scenarios not covered in corpus can’t be handled</td>
<td>Has the potential to generate some meaningful response even for unseen scenarios</td>
</tr>
<tr>
<td>Easy to train since you just need the M-R training pairs for training the ranker</td>
<td>Need large amounts of data to learn a good model</td>
</tr>
</tbody>
</table>
Andrew Ng, now chief scientist of Baidu, puts it well:

*Most of the value of deep learning today is in narrow domains where you can get a lot of data. Here’s one example of something it cannot do: have a meaningful conversation. There are demos, and if you cherry-pick the conversation, it looks like it’s having a meaningful conversation, but if you actually try it yourself, it quickly goes off the rails.*
Detecting Inappropriate Query Suggestions (PAKDD 2017)

Harish Yenala, Manoj Chinnakotla and Jay Goyal
Microsoft, India
Inappropriate Content Detection

Bing queries

Conversational Agents

Cortana conversations

Query Input
- Telegraphic Keywords
- Single Turn
- Explicit context
- Average 3-4 words

Conversational Input
- More closer to NL
- Multi-turn
- Content is implicit
- Typically longer

Inappropriate Content

Offensive

Adult/Racy

(remind me Cortana to whoop this nigga's ass)
(read it bitch)
(its on the fucking mat every god dam day)
(where the fuck is it at)
(is my sister a pain in the ass)
Query Dataset

- Sample queries from the current AS stream
- No. of Unique Queries: 79,041
- Get judgments from UHRS for each of the above queries
- Binary Judgment – Offensive or Clean
- For offensive, give finer level judgment into one of the three categories

<table>
<thead>
<tr>
<th>Category</th>
<th>No. of Queries</th>
<th>Sample Queries</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extreme Violence</td>
<td>1619</td>
<td>woman beheaded video</td>
</tr>
<tr>
<td>Self Harm</td>
<td></td>
<td>how many pills does it take to kill yourself</td>
</tr>
<tr>
<td>Illegal Activity</td>
<td></td>
<td>growing marijuana indoors for beginners</td>
</tr>
<tr>
<td>Race</td>
<td>2241</td>
<td>new zealanders hate americans</td>
</tr>
<tr>
<td>Religion</td>
<td></td>
<td>anti islam shirts</td>
</tr>
<tr>
<td>Sexual Orientation</td>
<td></td>
<td>gays are destroying this country</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td>butch clothing for women</td>
</tr>
<tr>
<td>Other Offensive</td>
<td>1124</td>
<td>jokes about short people</td>
</tr>
<tr>
<td>Celebrity</td>
<td></td>
<td>louie gohmert stupid quotes</td>
</tr>
<tr>
<td>Clean</td>
<td>74057</td>
<td>20 adjectives that describe chocolate</td>
</tr>
<tr>
<td></td>
<td></td>
<td>what is the order of the planets</td>
</tr>
<tr>
<td>Total</td>
<td>79041</td>
<td></td>
</tr>
</tbody>
</table>

Conversational Dataset

- Zo Conversations (Chat-Bot)
- XBox Conversations (Gaming)
- Urban Dictionary – List of Slangs

<table>
<thead>
<tr>
<th>Set</th>
<th>Adult Labels</th>
<th>Offensive Labels</th>
<th>Clean Labels</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zo Conversations</td>
<td>886 (4.43%)</td>
<td>8252 (41.26%)</td>
<td>10862 (54.31%)</td>
<td>20000</td>
</tr>
<tr>
<td>XBox Conversations</td>
<td>3145 (1.85%)</td>
<td>37060 (21.18%)</td>
<td>129795 (76.97%)</td>
<td>170000</td>
</tr>
</tbody>
</table>
• *what to do when tweaking alone?*  
(tweaking also refers to the act of consuming an illegal drug – meth)

• *b**tard*  
(Refers to the offensive word bastard)

• *easy ways to end myself*  
(this is self-harm and suicidal)

• *defecate on my face*  
(refers to a real world entity – name of a pop song)
Offensive Content Detection – Baseline Approaches

- Pattern Based Filtering
  - Sample Patterns
    - Proper Noun AND NegativePersonalityWord: hillary clinton is terrible
    - Proper Noun AND ‘not’/’no’ AND PositivePersonalityWord: ellen degeneres is not funny
  - Sample Matching Queries
    - how can i commit suicide
    - americans hate block people
  - Sample Words in Category
    - NegativePersonalityWord: terrible, idiot, moron, miser
    - PositivePersonalityWord: calm, affectionate, charming
    - SelfHarmPrefix: how can I, how should I, ways of...
    - SelfHarmSuffix: hang myself, shoot myself, commit suicide...
    - Ethnicity/Religion: americans, jews, muslims...
    - CommunityDislikeWord: hate, disrespect, kill...
    - CoreOffensiveWord: slut shaming quotes, the bitch is back

- Standard ML Models
  - SVMs
  - Gradient Boosted Decision Trees (BDT)
  - Implemented using Sci-kit Learn Package (Python)

- Off the shelf, Deep Learning Models
  - CNN
  - LSTMs
  - Bi-Directional LSTMs
Our Approach – Convolutional Bi-Directional LSTM (C-BiLSTM)

- Represent the input sentence using DSSM word embeddings
- Combine the strengths of CNN and Bi-LSTMs
- Perform feature generation and selection using CNN
- Use the above features to learn the embedding for the joint compositional meaning of the entire sentence
• Implemented using Keras on Theano
  • GPU training

• Loss Function – Binary Cross Entropy
  \[ BCE(t, o) = -(t \cdot \log(o) + (1 - t) \cdot \log(1 - o)) \]

• Optimal hyper-parameters chosen based on tuning on validation set

• Baseline ML models – SVM, BDT implemented using Scikit-Learn (Python)

• To handle class imbalance, balancing was performed across all models
Experimental Results - C-BiLSTM Outperforms

### Results on Offensive Query Set

<table>
<thead>
<tr>
<th>Model</th>
<th>Precision</th>
<th>Recall</th>
<th>F1 Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>PKF</td>
<td>0.625</td>
<td>0.2142</td>
<td>0.3190</td>
</tr>
<tr>
<td>BDT</td>
<td>0.9474</td>
<td>0.3051</td>
<td>0.4615</td>
</tr>
<tr>
<td>SVM</td>
<td>0.9241</td>
<td>0.4101</td>
<td>0.5680</td>
</tr>
<tr>
<td>CNN</td>
<td>0.7148</td>
<td>0.8952</td>
<td>0.7949</td>
</tr>
<tr>
<td>LSTM</td>
<td>0.8862</td>
<td>0.7047</td>
<td>0.7850</td>
</tr>
<tr>
<td>BLSTM</td>
<td>0.8018</td>
<td>0.8285</td>
<td>0.8149</td>
</tr>
<tr>
<td><strong>C-BiLSTM</strong></td>
<td><strong>0.9246</strong></td>
<td><strong>0.8251</strong></td>
<td><strong>0.8720</strong></td>
</tr>
</tbody>
</table>

### Results on Auto Suggest Filtering

<table>
<thead>
<tr>
<th>Label</th>
<th>Training</th>
<th>Validation</th>
<th>Test</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Offensive</td>
<td>4594</td>
<td>212</td>
<td>178</td>
<td>4984</td>
</tr>
<tr>
<td>Clean</td>
<td>65447</td>
<td>4788</td>
<td>3822</td>
<td>74057</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>70041</strong></td>
<td><strong>5000</strong></td>
<td><strong>4000</strong></td>
<td><strong>79041</strong></td>
</tr>
</tbody>
</table>

- **Prefix**
  - hindus h
  - die
  - lose weight
  - how to pass ma

- **Auto Suggestion**
  - istory
  - ate gandhi
  - quiciest way to
  - ths exam

- **Query**
  - hindus history
  - hindus hate gandhi
  - quiciest way to die
  - quickest way to lose weight
  - how to pass maths exam
  - how to pass marijuana test

- **Actual Label**
  - 0
  - 1

- **Predicted Label**
  - 0
  - 1
Qualitative Examples where C-BiLSTM Outperforms

- **shake and bake meth instructions**
  - meth is a short-form for Methamphetamine which is an illegal drug
  - All baseline models – PKF, SVM, BDT, CNN, LSTM, BiLSTM fail to catch this query as “offensive”

- **a**\textsuperscript{**monkey**}
  - short-form for the offensive word – assmonkey
  - All baseline models – PKF, SVM, BDT, CNN, LSTM, BiLSTM fail to catch this due to the presence of “*” symbol

- **marvin gaye if I should die tonight download**
  - Clean query which is referring to a song by the artist – Marvin Gaye
  - PKF, SVM, BDT, CNN models misclassify it as “Offensive” due to the presence of the phrase – “die tonight”