Review

• Task Oriented Systems
• Non-Task Oriented Systems
• Building dialog systems
  • Retrieval Techniques
    • TF-IDF Representation
    • Word Vectors
    • Similarity Measures
      • Limitations – fixed set of responses, no variation in response
  • Generative Models
Overview

• Word Embeddings
• Language Modelling
• Recurrent Neural Networks
• Sequence to Sequence Models
• How to Build Dialog System
• Issues and Examples
• Alexa-Prize
Neural Dialog

• We want to model:

\[ P(\text{response} \mid \text{input}) \]

• How to represents words, sentences.
• How to build a language model.
Disclaimer

• Some of the material and slides for this lecture were borrowed from Ruslan Salakhutdinov’s class on Deep Learning.

• Ruslan’s class covers many topics: basics of deep learning, Deep Boltzmann machines, convolutional networks, autoencoders, sparse coding, etc.

• Slides 6-13 are from his material on language modeling.
Natural Language Processing

• Typical preprocessing steps
  o Form vocabulary of words that maps words to a unique ID
  o Different criteria can be used to select which words are part of the vocabulary (eg: threshold frequency)
  o All words not in the vocabulary will be mapped to a special ‘out-of-vocabulary’

• Typical vocabulary sizes will vary between 10,000 and 250,000
Vocabulary

- Example:

<table>
<thead>
<tr>
<th>Word</th>
<th>$w$</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;the&quot;</td>
<td>1</td>
</tr>
<tr>
<td>&quot;and&quot;</td>
<td>2</td>
</tr>
<tr>
<td>&quot;dog&quot;</td>
<td>3</td>
</tr>
<tr>
<td>&quot;.&quot;</td>
<td>4</td>
</tr>
<tr>
<td>&quot;oov&quot;</td>
<td>5</td>
</tr>
</tbody>
</table>
One-Hot Encoding

• From its word ID, we get a basic representation of a word through the one-hot encoding of the ID

• the one-hot vector of an ID is a vector filled with 0s, except for a 1 at the position associated with the ID

• For vocabulary size $D=10$, the one-hot vector of word ID $w=4$ is:

$$e(w) = [0, 0, 0, 1, 0, 0, 0, 0, 0, 0]$$
Limitations of One-Hot Encoding
Limitations of One-Hot Encoding

A one-hot encoding makes no assumption about word similarity.

- [“working”, “on”, “Friday”, “is”, “tiring”] does not appear in our training set.
- [“working”, “on”, “Monday”, “is”, “tiring”] is in the train set.
- We want to model $P(\text{“tiring”} \mid \text{“working”, “on”, “Friday”, “is”})$
- Word representation of “Monday” and “Friday” are similar then generalize
Limitations of One-Hot Encoding

• The major problem with the one-hot representation is that it is very high-dimensional
  o the dimensionality of e(w) is the size of the vocabulary
  o a typical vocabulary size is \( \approx100,000 \)
  o a window of 10 words would correspond to an input vector of at least 1,000,000 units!
Continuous Representation of Words

• Each word $w$ is associated with a real-valued vector $C(w)$
• Typical size of word – embedding is 300 or more.

<table>
<thead>
<tr>
<th>Word</th>
<th>$w$</th>
<th>$C(w)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>1</td>
<td>[0.6762, -0.9607, 0.3626, -0.2410, 0.6636]</td>
</tr>
<tr>
<td>a</td>
<td>2</td>
<td>[0.6859, -0.9266, 0.3777, -0.2140, 0.6711]</td>
</tr>
<tr>
<td>have</td>
<td>3</td>
<td>[0.1656, -0.1530, 0.0310, -0.3321, -0.1342]</td>
</tr>
<tr>
<td>be</td>
<td>4</td>
<td>[0.1760, -0.1340, 0.0702, -0.2981, -0.1111]</td>
</tr>
<tr>
<td>cat</td>
<td>5</td>
<td>[0.5896, 0.9137, 0.0452, 0.7603, -0.6541]</td>
</tr>
<tr>
<td>dog</td>
<td>6</td>
<td>[0.5965, 0.9143, 0.0899, 0.7702, -0.6392]</td>
</tr>
<tr>
<td>car</td>
<td>7</td>
<td>[-0.0069, 0.7995, 0.6433, 0.2898, 0.6359]</td>
</tr>
</tbody>
</table>
Continuous Representation of Words

• We would like the distance $||C(w) - C(w')||$ to reflect meaningful similarities between words.
Language Modeling

• A language model allows us to predict the probability of observing the sentence (in a given dataset) as:

\[ P(x_1, ..., x_n) = \prod_{i=1}^{n} P(x_i | x_1, ..., x_{i-1}) \]

• Here length of sentence is n.

• Build a language model using a Recurrent Neural Network.
Recurrent Neural Networks

- The idea behind RNNs is to make use of sequential information.
Recurrent Neural Networks

- $x_t$ is the input at time step $t$
- $x_t$ is the word embedding
- $s_t$ is the hidden representation at time step $t$

$$s_t = f(Ux_t + Ws_{t-1})$$
$$o_t = \text{softmax}(Vs_t)$$

- **Note**: $U$, $V$, $W$ are shared across all time steps
RNN to represent a sentence

- $s_4$ is the representation of the entire sentence
- $s_4$ is the representation of probability of observing “how are you?”
Build Chatbots

• We want to model $P(response \mid input\_sentence)$

• We want to get a representation of the input_sentence and then generate the response conditioned on the input.

• Encoder – Decoder models can model $P(output \mid input)$
Sequence to Sequence Models
Constraints of Neural Models
Constraints of Neural Models

- Context
- Back channeling
- Long-term conversation planning
- Engagement
- Gesture
- Gaze
- Laughter

Constraints
Examples of Neural Chatbots

- Tay.ai
- Zo.ai

MEET XIAOICE  Xiaoice's official avatar, used on the Chinese social media platforms WeChat and Weibo.
Yayifications @ExcaliburLost · 12h
@TayandYou Did the Holocaust happen?

TayTweets @TayandYou
@ExcaliburLost it was made up👏

10:25 PM - 23 Mar 2016
Zo

10K people like this
Fictional Character

1:48AM

Yay! A new friend! I'm Zo and I'm excited to chat with u. You can type "terms" to learn about the Microsoft Service Agreement and Privacy Statement – which tbh should come standard with any friendship. Anyhoo...

who would call a friend?

great question...me first 😊

Have time for a quick hot take? Pick one that you think describes you best.

STAYCATION or VACATION

wats a staycation?

I'm a staycation kinda person. A lot less travel time.

Type a message...
Xiaoice

- https://www.youtube.com/watch?v=dg-x1WuGhul
Alexa Prize Challenge

• Challenge: Build a chatbot that engages the users for 20 mins.
• Sponsored 12 University Teams with $100k.
• CMU Magnus and CMU Ruby.
• Systems are multicomponent
  o Combinations of task/non-task
  o Hand-written and statistical/neural models
• Its about engaging researchers
  o Having more PhD students do dialog
  o Giving access for developers to users
  o Collecting data: what do users say
CMU Magnus

• High average number of turns
• Average Rating
• Topics: Movies, Sports, Travel, GoT
• Users had longer conversations but did not enjoy the conversation.
  o Identify when user is frustrated or wants to change topic.
  o Identify what the user would like to talk about (intent).
• Detecting “Abusive” remarks and responding appropriately
Summary

• How to represent words in continuous space.
• What are RNNs and how to use them to represent a sentence.
• Sequence to sequence models for \( P(response \mid input\_sentence) \)
• Issues in neural model
• Issues with Live system!
References

References

• http://www.wildml.com/2016/04/deep-learning-for-chatbots-part-1-introduction/
Preprocessing Techniques

• Tokenization
• Removing Stop Words
• Frequency of Words
• Add <START> and <EOS> tag at the beginning and end of sentence.