Deep Chit-Chat: Deep Learning for Chatbots

Wei Wu and Rui Yan
Microsoft Xiaolce and Peking University
2019.7.15
Chatbots are "HOT"

**Academia**

ConvAI2: Overview of the competition

There are currently few datasets appropriate for training and evaluating models for non-goal-oriented dialogue systems (chatbots), and equally problematic, there is currently no standard procedure for evaluating such models beyond the classic Turing test.

The aim of our competition is therefore to establish a concrete scenario for testing chatbots that aim to engage humans, and become a standard evaluation tool in order to make such systems directly comparable.

**Industry**

Virtual Assistants
- Microsoft Cortana
- Apple Siri
- Baidu Duer

Smart Speakers
- Amazon Echo
- Google Home
- Tmall Genie

Social Bots & Customer Service
- Microsoft Xiaolce
- Microsoft Rinna
- Alime Assistant
Conversations in a Chatbot

• Analysis is done from 100 sessions randomly sampled from the log of Microsoft Rinna, a commercial chatbot with 6M+ users in Japan.
Building A Dialogue System for Human-Machine ``Chat''

- **Chit-chat**: casual and non-goal-oriented.

- **Open domain**: the topic of the conversation could be any. Users may jump from topic to topic in the conversation.

- **Relevance & Diversity**: the research focuses on automatic replying with relevant (to the context) and diverse (informative) responses to make conversations engaging.

---

Chit-chat: casual and non-goal-oriented.

Open domain: the topic of the conversation could be any. Users may jump from topic to topic in the conversation.

Relevance & Diversity: the research focuses on automatic replying with relevant (to the context) and diverse (informative) responses to make conversations engaging.

Different From Task-Oriented Dialogues
Why People Build Chatbots (In History)

• To prove that an AI program can chat like a human.

• One of the first chatbots created at MIT AI Lab in 1966.

• Conversation is realized by matching user prompts to scripted responses with patterns.

• Simulation of a Rogerian psychotherapist.

ELIZA: is something troubling you?
Human: men are all alike.
ELIZA: what is the connection, do you suppose?
Human: They’re always bugging us about something or other.
ELIZA: Can you think of a specific example?
Human: Well, my boyfriend made me here.
ELIZA: Is it important to you that your boyfriend made you come here?
Human: He says I’m depressed much of the time.
ELIZA: I am sorry to hear that you are depressed.
Human: It’s true. I am unhappy.
...
Why People Build Chatbots (Recently)

• First party characters powered by Microsoft Xiaolce.

- China
  - 2014 小冰
- Japan
  - 2015 すんな
- U.S.
  - 2016 Zo
- India
  - 2017 Ruuh
- Indonesia
  - 2017 Rinna

• Third party IP characters powered by Microsoft Xiaolce.

And More...
Why People Build Chatbots (Recently)

800M+ Users

114M MAU

CPS=23

Knowledge Base

End-To-End Dialogue Modeling

Text-To-Speech (TTS)
Two Approaches for Non-Task-Oriented Dialogue Systems

- **Retrieval-base Methods**
- **Neural Networks**
- **Generation-base Methods**

**Input**

**Index**

**Output**

**Neural Networks**
Response Selection for Retrieval-based Chatbots
Roadmap

• Message-Response Matching for Single-Turn Response Selection
  • Framework I: matching with sentence embeddings.
  • Framework II: matching with message-response interaction.
  • Insights from empirical studies.

• Context-Response Matching for Multi-Turn Response Selection
  • Framework I: embedding->matching.
  • Framework II: representation->matching->aggregation.
  • Insights from empirical studies.

• Knowledge Enhanced Response Selection

• Emerging Trends in Research of Response Selection
  • Matching with better representations.
  • Learning a matching model for response selection.
Single-Turn Response Selection
Do you like cats?

- Yes, of course.  
- I am a girl.  
- Yes, more than dogs.  
- I lost my job!  
- Are you kidding me?  
- I love most animals!

Candidate responses

Do you like cats?

Yes, more than dogs.
System Architecture for Single-Turn Response Selection

- **Retrieval**
  - Message
  - Retrieved message-response pairs
  - Index of message-response pairs
  - Online

- **Feature Generation**
  - Message-response pairs with features

- **Ranking/Classification**
  - Ranked/Classified responses
  - Gradient boosted tree
  - Ranking Model/Classifier

- **Matching Models**
  - Deep learning based message-response matching

Example messages:

- **Query:** well I hope you felt better smiley
  - **Response:** thanks

- **Query:** I was so excited about dying my hair blond
  - **Response:** do it
Message-Response Matching: Framework I

**Sentence embedding layer**: representing a message and a response as vectors

**Matching layer**: calculating similarity of the message vector and the response vector

\[ f(\cdot;\cdot) \]

**Matching function**

\[ g(\cdot) \]

**Representation function**

\[ S \]

**Representation function**
Representation Functions

**Input**

\[ x \]

**Word Embedding**

\[ e_x \]

**Sentence Embedding**

\[ g(\cdot) \]

**Output**

\[ g(x) \]

**Sentence Vector**

**Mean/Max Pooling**

**Conv**

**CNN**

**Max Pooling**

**RNN-LSTM / RNN-GRU**

**LSTM (GRU)**

**LSTM (GRU)**

**LSTM (GRU)**

**Word/Char Sequence**

\[ v_1 \]

\[ v_2 \]

\[ v_3 \]

\[ v_4 \]
Matching Functions

\[ f(q, r) \]

Output

Matching Score

Cosine

Bilinear

Multi-layer Perceptron (MLP)

Neural Tensor Network

Matching Functions

\[ \frac{q \cdot r}{\|q\| \cdot \|r\|} \]

\[ \sigma(q^T \cdot W \cdot r) \]

\[ \sigma_2(b_2 + W_2 \cdot \sigma_1(b_1 + W_1 \cdot [q^T, r^T]^T)) \]

\[ \sigma(q^T \cdot W^{[1:n]}, r + V \cdot [q^T, r^T]^T + b) \]

Input

\[ q, r \]

Sentence Vector

Sentence Vector
Special Case 1: Arc-I

\[ g(\cdot) = \text{CNN} \]

\[ f(\cdot;\cdot) = \text{MLP} \]

Hu et al., Convolutional Neural Network Architectures for Matching Natural Language Sentences, NIPS'14
Special Case 2: Attentive LSTM

\[ f(\cdot, \cdot) = \text{Cosine} \]

\[ g(\cdot) = \text{LSTM + Attention + Pooling} \]

Ming Tan et al., Improved Representation Learning for Question Answer Matching, ACL'16
<table>
<thead>
<tr>
<th>Model</th>
<th>Representation</th>
<th>Matching</th>
<th>Paper</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLP</td>
<td>Mean pooling</td>
<td>MLP</td>
<td>N/A</td>
</tr>
<tr>
<td>Arc-I</td>
<td>CNN</td>
<td>MLP</td>
<td>Baotian Hu et al., Convolutional Neural Network Architectures for Matching Natural Language Sentences, <em>NIPS’14</em></td>
</tr>
<tr>
<td>CNTN</td>
<td>CNN</td>
<td>Neural Tensor Network</td>
<td>Xipeng Qiu et al., Convolutional Neural Tensor Network Architecture for Community-based Question Answering, <em>IJCAI’15</em></td>
</tr>
<tr>
<td>CNN with GESD &amp; AESD</td>
<td>CNN</td>
<td>Dot product &amp; Euclidean distance</td>
<td>Minwei Feng et al., Applying Deep Learning to Answer Selection: A Study and An Open Task, <em>IEEE ASRU 2015</em></td>
</tr>
<tr>
<td>Attentive LSTM</td>
<td>BiLSTM (+attention) + mean pooling</td>
<td>Cosine</td>
<td>Ming Tan et al., Improved Representation Learning for Question Answer Matching, <em>ACL’16</em></td>
</tr>
<tr>
<td>IARNN</td>
<td>GRU (+attention)</td>
<td>Cosine</td>
<td>Binging Wang et al., Inner Attention based Recurrent Networks for Answer Selection, <em>ACL’16</em></td>
</tr>
</tbody>
</table>
Message-Response Matching: Framework II

Input Representation: words are represented as vectors

Interaction Representation: encoding interaction of word pairs in (q,r)

Interaction Function $f(\cdot, \cdot)$

Interaction transformation and compression $g(\cdot)$

Matching function $h(\cdot)$

Matching Score $S$
Two Types of Interaction

- Similarity matrix-based
  - Convolution: $\sigma(W \cdot z + b)$
  - Pooling: $\text{max/ave}_{z \in \text{window}}(z)$
  - Word-word similarity: $f(w_i, v_j)$

- Attention-based
  - Attention: $f(\sum_{k=1}^{2} \alpha(v_i, w_k) w_k, v_i)$

$g(\cdot) = \text{RNN}$
### Implementation of Framework II

<table>
<thead>
<tr>
<th><strong>Input Representation</strong></th>
<th>$f(\cdot;\cdot)$</th>
<th><strong>Interaction Representation</strong></th>
<th>$g(\cdot)$</th>
<th>$h(\cdot)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word embedding with random initialization</td>
<td>Cosine/Dot product</td>
<td>Similarity matrices</td>
<td>2D CNN</td>
<td>MLP</td>
</tr>
<tr>
<td>Word embedding initialized with Glove</td>
<td>Linear and non-linear transformation</td>
<td>Weighted average (Attention)</td>
<td>RNN</td>
<td>SoftMax</td>
</tr>
<tr>
<td>Hidden states of an RNN</td>
<td>Tensor</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Indicator function</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Euclidean distance</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Special Case 1: Match Pyramid

**Input Representation:**

\[ f(\cdot, \cdot) : \text{indicator function / dot product / cosine} \]

**Interaction Representation:**

\[ g(\cdot) : \text{2D CNN} \]

\[ h(\cdot) : \text{MLP} \]

**Interaction Representation:** similarity matrix

**Input Representation:** word embedding with random initialization

Pang et al., Text Matching as Image Recognition, AAAI'16
Special Case 2: Match LSTM

**Input Representation:** hidden states of an RNN that processes the word sequences

\[ f(;\cdot) : w \cdot \tanh(A_1 \cdot h^j + A_2 \cdot h^k + A_3 \cdot h^{m}_{k-1}) \]

**Interaction Representation:** weighted average

\[ g(\cdot) : \text{LSTM} \]

\[ h(\cdot) : \text{softmax} \]

Wang & Jiang, Learning Natural Language Inference with LSTM, NAACL’16
## More Methods in Framework II

<table>
<thead>
<tr>
<th>Model</th>
<th>Interaction Representation</th>
<th>g(·)</th>
<th>Paper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arc-II</td>
<td>Similarity matrix (linear &amp; non-linear transformation)</td>
<td>2D CNN</td>
<td>Baotian Hu et al., Convolutional Neural Network Architectures for Matching Natural Language Sentences, <em>NIPS’14</em></td>
</tr>
<tr>
<td>DeepMatch_topic</td>
<td>Similarity matrix (linear &amp; non-linear transformation)</td>
<td>MLP</td>
<td>Zhengdong Lu et al., A Deep Architecture for Matching Short Texts, <em>NIPS’13</em></td>
</tr>
<tr>
<td>Match Pyramid</td>
<td>Similarity matrix</td>
<td>2D CNN</td>
<td>Liang Pang et al., Text Matching as Image Recognition, <em>AAAI’16</em></td>
</tr>
<tr>
<td>Match LSTM</td>
<td>Weighted average</td>
<td>LSTM</td>
<td>Shuohang Wang &amp; Jing Jiang, Learning Natural Language Inference with LSTM, <em>NAACL’16</em></td>
</tr>
</tbody>
</table>
Comparison between Framework I and Framework II

• Efficacy
  • In general, models in Framework II are better than models in Framework I on published datasets (see later), because *matching information in a message-response pair is sufficiently preserved by the interaction in Framework II.*

• Efficiency
  • *Because the sufficient (and heavy) interaction*, models in Framework II in general is more costly than models in Framework I.
  
  • *Because one can pre-compute the representations of messages and responses and store them in index with text*, models in Framework I is more preferable when there is strict requirement to online responding time.
A Dataset for Empirical Studies

• **Ubuntu Dialogue Corpus**
  • Dyadic English human-human conversations from Ubuntu chat logs.
  • Each conversation has multiple turns (Avg. # turns=7.71). For single-turn studies, we keep the last turn and the response to form a message-response pair.
  • Task = distinguishing the positive response from negative ones for a given message.
  • Train : validation: test = 1M : 0.5M : 0.5M.
  • Positive responses = human responses, negative responses = randomly sampled ones.
  • Positive : negative = 1:1 (train), 1:9 (validation), 1:9 (test).
  • Evaluation metrics = $R_n@k$. For each message, if the only positive response is ranked within top $k$ positions of $n$ candidates, then $R_n@k = 1$. The final result is the average on messages.

Lowe et al., The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-turn Dialogue Systems. *SIGDIAL’15*
## Empirical Comparison

<table>
<thead>
<tr>
<th>Model</th>
<th>Ubuntu</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$R_2@1$</td>
</tr>
<tr>
<td>MLP (I)</td>
<td>0.651</td>
</tr>
<tr>
<td>Arc-I (I)</td>
<td>0.665</td>
</tr>
<tr>
<td>LSTM (I)</td>
<td>0.725</td>
</tr>
<tr>
<td>CNTN (I)</td>
<td>0.743</td>
</tr>
<tr>
<td>Attentive-LSTM (I)</td>
<td>0.758</td>
</tr>
<tr>
<td>DeepMatch_topic (II)</td>
<td>0.593</td>
</tr>
<tr>
<td>Match LSTM (II)</td>
<td>0.685</td>
</tr>
<tr>
<td>Arc-II (II)</td>
<td>0.736</td>
</tr>
<tr>
<td>Match Pyramid (II)</td>
<td>0.743</td>
</tr>
<tr>
<td>MV-LSTM (II)</td>
<td>0.767</td>
</tr>
</tbody>
</table>
Insights from the Comparison

• **Neural tensor is a powerful matching function** (CNTN is much better than Arc-I).

• **Attentive LSTM achieves good performance** because the attention mechanism is inherently a kind of interaction.

• **Similarity matrix based interaction is better** than attention based interaction for Framework II (Match LSTM performs badly).

• In similarity matrix based methods, **dot product or cosine is better** as an interaction function (Match Pyramid is better than Arc-II).

• **Input representation that encodes more contextual information can improve** the performance of matching (MV-LSTM is better than Match Pyramid).
References


• Baotian Hu, Zhengdong Lu, Hang Li, and Qingcai Chen. Convolutional Neural Network Architectures for Matching Natural Language Sentences. *NIPS’14*

• Shengxian Wang, Yanyan Lan, Jiafeng Guo, Jun Xu, Liang Pang, and Xueqi Cheng. A Deep Architecture for Semantic Matching with Multiple Positional Sentence Representations. *AAAI’16*


• Ming Tan, Cicero dos Santos, Bing Xiang, and Bowen Zhou. Improved Representation Learning for Question Answer Matching. *ACL’16.*

• Bingning Wang, Kang Liu, and Jun Zhao. Inner Attention based Recurrent Neural Networks for Answer Selection. *ACL’16.*

• Shuohang Wang and Jing Jiang. Learning Natural Language Inference with LSTM. *NAACL’16.*

• Aliaksei Severyn and Alessandro Moschitti. Learning to Rank Short Text Pairs with Convolutional Deep Neural Networks. *SIGIR’15.*
References


• Shengxian Wan, Yanyan Lan, Jun Xu, Jiafeng Guo, Liang Pang, and Xueqi Cheng. Match-SRNN: Modeling the Recursive Matching Structure with Spatial RNN. *IJCAI’16*.


• Yu Wu, Wei Wu, Can Xu, and Zhoujun Li. Knowledge Enhanced Hybrid Neural Network for Text Matching. *AAAI’18*.
Multi-Turn Response Selection
Multi-Turn Response Selection

Candidate responses

- Yes, of course.
- What lesson?
- No, no free lessons.
- Yes, please bring your drum
- We do not have coaches.
- Our English lessons are free

Candidates in red are good without the context!
System Architecture for Multi-Turn Response Selection

Context (message + history)

Retrieval

Retrieved context-response pairs

Feature Generation

Context-response pairs with features

Ranking

Ranked responses

Index of message-response pairs

Matching Models

Deep learning based context-response matching

Gradient boosted tree

online

do it

$q>$ well I hope you fell better smiley</q>

$q>$ I was so excited about dying my hair blond</q>

$q>$ thanks</q>

$q>$ do it</q>

$q>$ do it</q>
New Challenges with Contexts

• A hierarchical data structure
  • Words -> utterances -> session

• Information redundancy
  • Not all words and utterances are useful for response selection

• Logics
  • Order of utterances matters in response selection
  • Long-term dependencies among words and utterances
  • Constraints to proper responses
Context-Response Matching: Framework I

Context embedding layer: modeling the semantics/relationship among utterances

Utterance embedding layer: modeling the semantics in and among language units (word/character)

Matching layer: calculating a matching score with the two vectors

- Average/Concatenation of unit embedding
- Multi-layer perceptron (MLP)
- Convolutional neural network (CNN)
- Recurrent neural network (RNN)

- Multi-layer perceptron (MLP)
- Bilinear model
- Neural tensor (multiple bilinear models)
Special Case 1: Dual-LSTM

- **Context embedding:** LSTM on the concatenation of word vectors
- **Utterance embedding:** word embedding vectors
- **Matching:** Bilinear model
- **Response vector:** last hidden vector of an LSTM

Lowe et al., The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-turn Dialogue Systems. SIGDIAL 2015
Special Case 2: Multi-view Response Selection Model

Context embedding (word view): GRU on the concatenation of word vectors

Utterance embedding (word view): word embedding vectors

Response vector (word view): last hidden vector of GRU

Matching: Bilinear

Utterance embedding (utterance view): CNN

Context embedding (utterance view): GRU

Response vector (utterance view): CNN

<table>
<thead>
<tr>
<th>Model</th>
<th>Utterance Embedding</th>
<th>Context Embedding</th>
<th>Matching</th>
<th>Paper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dual-LSTM</td>
<td>Word embedding</td>
<td>LSTM</td>
<td>Bilinear</td>
<td>Ryan Lowe et al., The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-turn Dialogue Systems, <em>SIGDIAL’15</em></td>
</tr>
<tr>
<td>DL2R</td>
<td>BiLSTM+CNN</td>
<td>Identity</td>
<td>MLP</td>
<td>Rui Yang et al., Learning to Respond with Deep Neural Networks for Retrieval-based Human-Computer Conversation System, <em>SIGIR’16</em></td>
</tr>
<tr>
<td>Multi-View</td>
<td>Word embedding/CNN</td>
<td>GRU</td>
<td>Bilinear</td>
<td>Xiangyang Zhou et al., Multi-view Response Selection for Human-Computer Conversation, <em>EMNLP’16</em></td>
</tr>
<tr>
<td>Any single-turn model</td>
<td>Word embedding</td>
<td>Message embedding of the model</td>
<td>Matching function of the model</td>
<td>N/A</td>
</tr>
</tbody>
</table>
Analysis of Framework I

- Modeling the hierarchy of conversation sessions? ✔
- Modeling the relationship/dependency among words? ✔
- Modeling the relationship/dependency among utterances? ✔
- Modeling word/utterance importance ❌

- Everything in a context is compressed to a fixed length vector before matching with the response.
- Utterance representation is independent with the response/other utterances.
Context-Response Matching: Framework II

Wu et al., A Sequential Matching Framework for Multi-turn Response Selection in Retrieval-based Chatbots. *Computational Linguistics*, 1-50, 2018
Special Case 1: Sequential Matching Network (SMN)

\[ f(u_i, r) = W \cdot [\text{CNN}(M_1) \oplus \text{CNN}(M_2)] + b \]

Representations:
- word embedding
- hidden vectors of a GRU

Wu et al., Sequential Matching Network: A New Architecture for Multi-turn Response Selection in Retrieval-based Chatbots. ACL’17
Special Case 2: Sequential Attention Network (SAN)

Representations:
• word embedding
• hidden vectors of a GRU

$$f(u_i, r) = GRU([\text{Att}(r \rightarrow u_i) \odot r]_w \oplus [\text{Att}(r \rightarrow u_i) \odot r]_h)$$

Special Case 3: Deep Utterance Aggregation (DUA)

Representations:

- hidden vectors of a GRU
- Fusion from the last utterance to other utterances/the response
- Self-attention within each utterance/the response

$$f(u_i, r) = [\text{CNN}(M_1) \oplus \text{CNN}(M_2)]$$

How Framework II Models Word/Utterance Importance

$u_1$: How can unzip many rar files (\textit{number} for example) at once?

$u_2$: Sure, you can do it in bash.

$u_3$: OK, how?

$u_4$: Are the files all in the same directory?

$u_5$: Yes, they all are.

Response: Then the command \texttt{glebihan} should extract them all from/to that directory.
Comparison between Framework I and Framework II

• Efficacy
  • In general, models in Framework II are better than models in Framework I on published datasets (see later), because *important information in a context is sufficiently distilled and preserved in matching in Framework II.*

• Efficiency
  • *Because the sufficient (and heavy) interaction,* models in Framework II in general is more costly than models in Framework I.

• Interpretability
  • It is easy to understand what models in Framework II have learned via some visualization techniques.
Benchmark Datasets for Empirical Studies
(Lowe et al., SIGDIAL’15, Wu et al., ACL’17, Zhang et al., COLING’18)

<table>
<thead>
<tr>
<th>Source</th>
<th>Ubuntu Dialogue Corpus</th>
<th>Douban Conversation Corpus</th>
<th>E-Commerce Dialogue Corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Ubuntu Forum</td>
<td>Douban group, popular social networking service in China</td>
<td>Taobao, the largest e-commerce platform in China</td>
</tr>
<tr>
<td>Conversation Types</td>
<td>Multi-turn technical discussions</td>
<td>Multi-turn open domain chit-chat</td>
<td>Multi-turn conversations covering commodity consultation, recommendation, logistics express, negotiation, and chit-chat</td>
</tr>
<tr>
<td>Train : Val : Test (pairs)</td>
<td>1M : 0.5M : 0.5M</td>
<td>1M : 50K : 6670</td>
<td>1M : 10K : 10K</td>
</tr>
<tr>
<td>Dataset Construction</td>
<td>Random sampling</td>
<td>Train &amp; Val: random sampling Test: response retrieval &amp; human judgment</td>
<td>Response retrieval</td>
</tr>
<tr>
<td>Evaluation Metrics</td>
<td>$R_{10}@1$, $R_{10}@2$, $R_{10}@5$, $R_2@1$</td>
<td>MAP, MRR, P@1</td>
<td>$R_{10}@1$, $R_{10}@2$, $R_{10}@5$</td>
</tr>
</tbody>
</table>
### Empirical Comparison
*(Zhang et al., COLING 2018)*

<table>
<thead>
<tr>
<th>Model</th>
<th>Ubuntu</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$R_2@1$</td>
<td>$R_{10}@1$</td>
<td>$R_{10}@2$</td>
<td>$R_{10}@5$</td>
<td>MAP</td>
<td>MRR</td>
<td>P@1</td>
</tr>
<tr>
<td>TF-IDF</td>
<td>0.659</td>
<td>0.410</td>
<td>0.545</td>
<td>0.708</td>
<td>0.331</td>
<td>0.359</td>
<td>0.180</td>
</tr>
<tr>
<td>CNN</td>
<td>0.848</td>
<td>0.549</td>
<td>0.684</td>
<td>0.896</td>
<td>0.417</td>
<td>0.440</td>
<td>0.226</td>
</tr>
<tr>
<td>BiLSTM</td>
<td>0.895</td>
<td>0.630</td>
<td>0.780</td>
<td>0.944</td>
<td>0.479</td>
<td>0.514</td>
<td>0.313</td>
</tr>
<tr>
<td>MV-LSTM</td>
<td>0.906</td>
<td>0.653</td>
<td>0.804</td>
<td>0.946</td>
<td>0.498</td>
<td>0.538</td>
<td>0.348</td>
</tr>
<tr>
<td>Match LSTM</td>
<td>0.904</td>
<td>0.653</td>
<td>0.799</td>
<td>0.944</td>
<td>0.500</td>
<td>0.537</td>
<td>0.345</td>
</tr>
<tr>
<td>Attentive-LSTM</td>
<td>0.903</td>
<td>0.633</td>
<td>0.789</td>
<td>0.943</td>
<td>0.495</td>
<td>0.523</td>
<td>0.331</td>
</tr>
<tr>
<td>Dual LSTM</td>
<td>0.901</td>
<td>0.638</td>
<td>0.784</td>
<td>0.949</td>
<td>0.485</td>
<td>0.527</td>
<td>0.320</td>
</tr>
<tr>
<td>DL2R</td>
<td>0.899</td>
<td>0.626</td>
<td>0.783</td>
<td>0.944</td>
<td>0.488</td>
<td>0.527</td>
<td>0.330</td>
</tr>
<tr>
<td>Multi-View</td>
<td>0.908</td>
<td>0.662</td>
<td>0.801</td>
<td>0.951</td>
<td>0.505</td>
<td>0.543</td>
<td>0.342</td>
</tr>
<tr>
<td>SMN</td>
<td>0.926</td>
<td>0.726</td>
<td>0.847</td>
<td>0.961</td>
<td>0.529</td>
<td>0.569</td>
<td>0.397</td>
</tr>
<tr>
<td>SAN</td>
<td>0.932</td>
<td>0.734</td>
<td>0.852</td>
<td>0.962</td>
<td>0.532</td>
<td>0.575</td>
<td>0.387</td>
</tr>
<tr>
<td>DUA</td>
<td>N/A</td>
<td>0.752</td>
<td>0.868</td>
<td>0.962</td>
<td>0.551</td>
<td>0.599</td>
<td>0.421</td>
</tr>
</tbody>
</table>
Empirical Comparison (Cont’)
(Zhang et al., COLING 2018)

<table>
<thead>
<tr>
<th>Model</th>
<th>E-Commerce</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$R_{10}@1$</td>
</tr>
<tr>
<td>TF-IDF</td>
<td>0.159</td>
</tr>
<tr>
<td>CNN</td>
<td>0.328</td>
</tr>
<tr>
<td>BiLSTM</td>
<td>0.355</td>
</tr>
<tr>
<td>MV-LSTM</td>
<td>0.412</td>
</tr>
<tr>
<td>Match LSTM</td>
<td>0.410</td>
</tr>
<tr>
<td>Attentive-LSTM</td>
<td>0.401</td>
</tr>
<tr>
<td>Dual LSTM</td>
<td>0.365</td>
</tr>
<tr>
<td>DL2R</td>
<td>0.399</td>
</tr>
<tr>
<td>Multi-View</td>
<td>0.421</td>
</tr>
<tr>
<td>SMN</td>
<td>0.453</td>
</tr>
<tr>
<td>SAN</td>
<td>0.463</td>
</tr>
<tr>
<td>DUA</td>
<td>0.501</td>
</tr>
</tbody>
</table>
Insights from the Comparison

• **Utterance-response interaction at the beginning is crucial**, as models in Framework II (SMN & SAN) are much better than models in Framework I.

• **Ensemble of matching from multiple views are effective**, as multi-view outperforms all other models in Framework I.

• **Selection of interaction functions does not make a clear difference on efficacy**, as long as the function can model word/phrase importance. This is indicated by the comparable performance of SMN and SAN.

• **SMN is more efficient and easier to parallelize than SAN**, due to the characteristics of CNN and RNN.

• **Self-attention is powerful in matching**, as DUA achieves the best performance on all the three benchmarks.
References


• Rudolf Kadlec, Martin Schmid, and Jan Kleindienst. Improved Deep Learning Baselines for Ubuntu Corpus Dialogs. NIS’15 workshop.

• Rui Yan, Yiping Song, and Hua Wu. Learning to Respond with Deep Neural Networks for Retrieval-Based Human-Computer Conversation System. SIGIR’16.


• Bowen Wu, Baoxun Wang, and Hui Xue. Ranking Responses Oriented to Conversational Relevance in Chatbots. COLING’16

• Yu Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhoujun Li. Sequential Matching Network: A New Architecture for Multi-turn Response Selection in Retrieval-based Chatbots. ACL’17.


• Zhen Xu, Bingquan Liu, Baoxun Wang, Chengjie Sun, and Xiaolong Wang. Incorporating Loose-Structured Knowledge into LSTM with Recall Gate for Conversation Modeling. IJCNN’17

• Liu Yang, Minghui Qiu, Chen Qu, Jiafeng Guo, Yongfeng Zhang, W.Bruce Croft, Jun Huang, and Haiqing Chen. Response Ranking with Deep Matching Networks and External Knowledge in Information-seeking Conversation Systems. SIGIR’18.

• Zhuosheng Zhang, Jiangtong Li, Pengfei Zhu, Hai Zhao, Guosheng Liu. Modeling Multi-turn Conversation with Deep Utterance Aggregation. COLING’18
Knowledge Enhanced Response Selection
Knowledge Enhanced Message-Response Matching: Framework I

**Topic Aware Attentive Recurrent Neural Network (TAARNN)**

- **Message representation**
  - Message hidden vectors
  - Message word embedding

- **Response representation**
  - Response hidden embedding
  - Response word embedding

**Knowledge:**
- Topic words (from LDA)
- Syntax patterns
- Triples in knowledge bases
- ......

Wu et al., Response Selection with Topic Clues for Retrieval-based Chatbots. *Neurocomputing Volume 316, p251-261*
Knowledge Enhanced Message-Response Matching: Framework II

Knowledge Enhanced Hybrid Neural Network (KEHNN)

Knowledge:
- Topic words
- Syntax patterns
- Triples in knowledge bases
- …

Wu et al., Knowledge Enhanced Hybrid Neural Network for Text Matching. AAAI’18
Empirical Results on the Ubuntu Data

<table>
<thead>
<tr>
<th>Model</th>
<th>Ubuntu</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$R_2@1$</td>
</tr>
<tr>
<td>MLP (I)</td>
<td>0.651</td>
</tr>
<tr>
<td>Arc-I (I)</td>
<td>0.665</td>
</tr>
<tr>
<td>LSTM (I)</td>
<td>0.725</td>
</tr>
<tr>
<td>CNTN (I)</td>
<td>0.743</td>
</tr>
<tr>
<td>Attentive-LSTM (I)</td>
<td>0.758</td>
</tr>
<tr>
<td>DeepMatch_topic (II)</td>
<td>0.593</td>
</tr>
<tr>
<td>Match LSTM (II)</td>
<td>0.685</td>
</tr>
<tr>
<td>Arc-II (II)</td>
<td>0.736</td>
</tr>
<tr>
<td>Match Pyramid (II)</td>
<td>0.743</td>
</tr>
<tr>
<td>MV-LSTM (II)</td>
<td>0.767</td>
</tr>
<tr>
<td>TAARNN (II)</td>
<td>0.770</td>
</tr>
<tr>
<td>KEHNN (II)</td>
<td>0.786</td>
</tr>
</tbody>
</table>
Knowledge Enhanced Context-Response Matching

Knowledge is incorporated into matching through **Pseudo Relevance Feedback**

Knowledge is incorporated into matching through an **Extra Matching Channel**
### Empirical Results

<table>
<thead>
<tr>
<th>Data</th>
<th>UDC</th>
<th></th>
<th></th>
<th>MSDialog</th>
<th></th>
<th></th>
<th>AliMe</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Methods</strong></td>
<td><strong>MAP</strong></td>
<td><strong>Recall@5</strong></td>
<td><strong>Recall@1</strong></td>
<td><strong>Recall@2</strong></td>
<td><strong>MAP</strong></td>
<td><strong>Recall@5</strong></td>
<td><strong>Recall@1</strong></td>
<td><strong>Recall@2</strong></td>
</tr>
<tr>
<td>BM25</td>
<td>0.6504</td>
<td>0.8206</td>
<td>0.5138</td>
<td>0.6439</td>
<td>0.4387</td>
<td>0.6329</td>
<td>0.2626</td>
<td>0.3933</td>
</tr>
<tr>
<td>BM25-PRF</td>
<td>0.6620</td>
<td>0.8292</td>
<td>0.5289</td>
<td>0.6554</td>
<td>0.4419</td>
<td>0.6423</td>
<td>0.2652</td>
<td>0.3970</td>
</tr>
<tr>
<td>ARC-II</td>
<td>0.6855</td>
<td>0.8978</td>
<td>0.5350</td>
<td>0.6959</td>
<td>0.5398</td>
<td>0.8662</td>
<td>0.3189</td>
<td>0.5413</td>
</tr>
<tr>
<td>MV-LSTM</td>
<td>0.6611</td>
<td>0.8936</td>
<td>0.4973</td>
<td>0.6733</td>
<td>0.5059</td>
<td>0.8516</td>
<td>0.2768</td>
<td>0.5000</td>
</tr>
<tr>
<td>DRMM</td>
<td>0.6749</td>
<td>0.8776</td>
<td>0.5287</td>
<td>0.6773</td>
<td>0.5704</td>
<td>0.9003</td>
<td>0.3507</td>
<td>0.5854</td>
</tr>
<tr>
<td>Duet</td>
<td>0.5692</td>
<td>0.8272</td>
<td>0.4756</td>
<td>0.5592</td>
<td>0.5158</td>
<td>0.8481</td>
<td>0.2934</td>
<td>0.5046</td>
</tr>
<tr>
<td>SMN</td>
<td>0.7327</td>
<td>0.9273</td>
<td>0.5948</td>
<td>0.7523</td>
<td>0.6188</td>
<td>0.8374</td>
<td>0.4529</td>
<td>0.6195</td>
</tr>
<tr>
<td>DMN</td>
<td>0.7363</td>
<td>0.9196</td>
<td>0.6056</td>
<td>0.7509</td>
<td>0.6415</td>
<td>0.9155</td>
<td>0.4521</td>
<td>0.6673</td>
</tr>
<tr>
<td>DMN-KD</td>
<td><strong>0.7655</strong></td>
<td><strong>0.9351</strong></td>
<td><strong>0.6443</strong></td>
<td><strong>0.7841</strong></td>
<td><strong>0.6728</strong></td>
<td><strong>0.9304</strong></td>
<td><strong>0.4908</strong></td>
<td><strong>0.7089</strong></td>
</tr>
<tr>
<td>DMN-PRF</td>
<td><strong>0.7719</strong></td>
<td><strong>0.9343</strong></td>
<td><strong>0.6552</strong></td>
<td><strong>0.7893</strong></td>
<td><strong>0.6792</strong></td>
<td><strong>0.9356</strong></td>
<td><strong>0.5021</strong></td>
<td><strong>0.7122</strong></td>
</tr>
</tbody>
</table>

Yang et al., Response Ranking with Deep Matching Networks and External Knowledge in Information-seeking Conversation Systems. *SIGIR’18*
**Document-Grounded Response Selection**

- **Scenarios: open domain conversation with a document as background**
  - One introduces the documents to the other.
  - The two discuss the content of the documents.

- **Tasks:**
  - Response selection according to both the conversation context and the background document

- **Challenges:**
  - **Context grounding:** not all utterances are related to the document.
  - **Document comprehension:** information in the document is redundant.
  - **Matching:** multiple sources

---

Zhou et al., A Dataset for Document Grounded Conversations. *EMNLP’18*
Document-Grounded Matching Network

- **Context grounding**
  - Fusing the document into the context via **CROSS ATTENTION**
    \[ \bar{U}_{ij} = \text{Attention}(U_i, D_j, D_j) \]
  - Representing an utterance with **RESIDUE CONNECTION**
    \[ U_i \rightarrow [U_i = \bar{U}_{i0}, \bar{U}_{i1}, \cdots, \bar{U}_{im}] \]
  - Determining if grounding is necessary via **HIERARCHICAL INTERACTION**
    First-level: \( r_j \rightarrow_{\text{att}} \bar{h}_{i,k} \rightarrow h_{i,j,k} \)
    Second-level: \( r_j \rightarrow_{\text{att}} \bar{h}_{i,j} = \{h_{i,j,k}\} \rightarrow h_{i,j} \)

- **Document comprehension**
  - Selecting important sentences via context-document **CROSS ATTENTION**
    \[ \bar{D}_{ji} = \text{Attention}(D_j, U_i, U_i) \]

- **Matching**
  - Merging **MULTIPLE CHANNELS**

Zhao et al., A Document-Grounded Matching Network for Response Selection in Retrieval-based Chatbots. *IJCAI’19*
## Empirical Evaluation on Benchmarks

<table>
<thead>
<tr>
<th></th>
<th>PERSONA-CHAT</th>
<th>CMUDoG</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Original Persona</td>
<td>Revised Persona</td>
</tr>
<tr>
<td><strong>r@1</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Memory Network</td>
<td>51.1</td>
<td>35.1</td>
</tr>
<tr>
<td>Transformer</td>
<td>54.2</td>
<td>42.1</td>
</tr>
<tr>
<td><strong>r@2</strong></td>
<td>61.8</td>
<td>45.7</td>
</tr>
<tr>
<td><strong>r@5</strong></td>
<td>77.4</td>
<td>66.3</td>
</tr>
<tr>
<td><strong>r@1</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DGMN</td>
<td>67.6</td>
<td>58.8</td>
</tr>
<tr>
<td><strong>r@2</strong></td>
<td>80.2</td>
<td>62.5</td>
</tr>
<tr>
<td><strong>r@5</strong></td>
<td>92.9</td>
<td>87.7</td>
</tr>
</tbody>
</table>

**PERSONA-CHAT**: Zhang et al., Personalizing Dialogue Agents: I have a dog, do you have pets too? *ACL’18. ACL’18*

**CMUDoG**: Zhou et al., A Dataset for Document Grounded Conversations. *EMNLP’18*
Matching With Better Representations
Representations Go Deep

- Representing utterances and responses by stacking multiple attention modules

**Deep Attention Matching Network (DAM)**

<table>
<thead>
<tr>
<th>Model</th>
<th>Ubuntu</th>
<th>Douban</th>
</tr>
</thead>
</table>
|       | $R_2@1$ | $R_{10}@1$ | $R_{10}@2$ | $R_{10}@5$ | MAP | MRR | $P@1$
| SMN   | 0.926   | 0.726   | 0.847   | 0.961   | 0.529   | 0.569 | 0.397 |
| DAM   | 0.938   | 0.767   | 0.874   | 0.969   | 0.550   | 0.601 | 0.427 |

$+1.2\%$ $+4.1\%$ $+2.7\%$ $+0.8\%$ $+2.1\%$ $+3.1\%$ $+3.0\%$

- **Cross-attention**: let an utterance and a response attend each other

Zhou et al., Multi-turn Response Selection for Chatbots with Deep Attention Matching Network. ACL’18
Representations Go Wide

• Fusing multiple types of representations are helpful, but how to fuse matters.

Multi-Representation Fusion Network (MRFN)

<table>
<thead>
<tr>
<th>Model</th>
<th>Ubuntu</th>
<th>Douban</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$R_2@1$</td>
<td>$R_{10}@1$</td>
<td>$R_{10}@2$</td>
<td>$R_{10}@5$</td>
<td>MAP</td>
<td>MRR</td>
</tr>
<tr>
<td>SMN</td>
<td>0.926</td>
<td>0.726</td>
<td>0.847</td>
<td>0.961</td>
<td>0.529</td>
<td>0.569</td>
</tr>
<tr>
<td>DAM</td>
<td>0.938</td>
<td>0.767</td>
<td>0.874</td>
<td>0.969</td>
<td>0.550</td>
<td>0.601</td>
</tr>
<tr>
<td>MRFN(FES)</td>
<td>0.930</td>
<td>0.742</td>
<td>0.857</td>
<td>0.963</td>
<td>0.538</td>
<td>0.583</td>
</tr>
<tr>
<td>MRFN(FIS)</td>
<td>0.936</td>
<td>0.762</td>
<td>0.870</td>
<td>0.967</td>
<td>0.558</td>
<td>0.605</td>
</tr>
<tr>
<td>MRFN(FLS)</td>
<td><strong>0.945</strong></td>
<td><strong>0.786</strong></td>
<td><strong>0.886</strong></td>
<td><strong>0.976</strong></td>
<td><strong>0.571</strong></td>
<td><strong>0.617</strong></td>
</tr>
<tr>
<td></td>
<td>+0.7%</td>
<td>+1.9%</td>
<td>+1.2%</td>
<td>+0.7%</td>
<td>+2.1%</td>
<td>+1.6%</td>
</tr>
</tbody>
</table>

Tao et al., Multi-Representation Fusion Network for Multi-turn Response Selection in Retrieval-based Chatbots. WSDM’19
### Representations from Pre-Training

- Pre-training neural networks on large scale data sets as representations significantly improves the existing models.

<table>
<thead>
<tr>
<th>Model</th>
<th>Ubuntu</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$R_2@1$</td>
</tr>
<tr>
<td>SMN</td>
<td>0.926</td>
</tr>
<tr>
<td>SMN+CoVe</td>
<td>0.930</td>
</tr>
<tr>
<td></td>
<td>+0.4%</td>
</tr>
<tr>
<td>SMN+ELMo</td>
<td>0.926</td>
</tr>
<tr>
<td></td>
<td>+0.0%</td>
</tr>
<tr>
<td>SMN+ELMo (fine-tune)</td>
<td>0.930</td>
</tr>
<tr>
<td></td>
<td>+0.4%</td>
</tr>
<tr>
<td>SMN+ECMo</td>
<td>0.934</td>
</tr>
<tr>
<td></td>
<td>+0.8%</td>
</tr>
</tbody>
</table>
## Results with BERT

<table>
<thead>
<tr>
<th>Model</th>
<th>Ubuntu</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$R_2@1$</td>
</tr>
<tr>
<td>SMN</td>
<td>0.926</td>
</tr>
<tr>
<td>DAM</td>
<td>0.938</td>
</tr>
<tr>
<td>MRFN(FLS)</td>
<td>0.945</td>
</tr>
<tr>
<td>BERT (base)</td>
<td>0.951</td>
</tr>
</tbody>
</table>
Deep Matching
From Shallow Representation to Deep Representation

Shallow Representation (e.g., SMN)

Deep Representation (e.g., DAM)
From Shallow Interaction to Deep Interaction

One Time of Interaction (e.g., attention-based)

Interaction-Over-Interaction (e.g., attention-based)

Tao et al., One Time of Interaction May Not Be Enough: Go Deep with an Interaction-over-Interaction Network for Response Selection in Dialogues. ACL ’19
## Empirical Studies

<table>
<thead>
<tr>
<th>Model</th>
<th>SMN</th>
<th>DAM</th>
<th>MRFN(FLS)</th>
<th>lol</th>
<th>SMN</th>
<th>DAM</th>
<th>MRFN(FLS)</th>
<th>lol</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Ubuntu</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$R_2@1$</td>
<td>0.926</td>
<td>0.938</td>
<td>0.945</td>
<td>0.947</td>
<td>0.453</td>
<td>0.526</td>
<td>0.526</td>
<td>0.563</td>
</tr>
<tr>
<td>$R_{10}@1$</td>
<td>0.726</td>
<td>0.767</td>
<td>0.786</td>
<td>0.796</td>
<td>0.654</td>
<td>0.727</td>
<td>0.768</td>
<td>0.768</td>
</tr>
<tr>
<td>$R_{10}@2$</td>
<td>0.847</td>
<td>0.874</td>
<td>0.886</td>
<td>0.894</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$R_{10}@5$</td>
<td>0.961</td>
<td>0.969</td>
<td>0.976</td>
<td>0.974</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAP</td>
<td>0.529</td>
<td>0.550</td>
<td>0.571</td>
<td>0.573</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MRR</td>
<td>0.569</td>
<td>0.601</td>
<td>0.617</td>
<td>0.621</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$P@1$</td>
<td>0.397</td>
<td>0.427</td>
<td>0.448</td>
<td>0.444</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>E-Commerce</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>E-Commerce</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$R_{10}@1$</td>
<td>0.453</td>
<td>0.526</td>
<td>0.526</td>
<td>0.563</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$R_{10}@1$</td>
<td>0.654</td>
<td>0.727</td>
<td>0.768</td>
<td>0.768</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$R_{10}@5$</td>
<td>0.886</td>
<td>0.933</td>
<td>0.950</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Does Depth of Interaction Matter?

![Graphs showing R@1 and P@1 metrics for Ubuntu, E-Commerce, and Douban across different # Interaction Blocks.]

- **R@1** for Ubuntu increases from 0.778 to 0.794.
- **R@1** for E-Commerce increases from 0.467 to 0.561.
- **R@1** for Douban remains relatively stable around 0.441.

The graphs illustrate how interaction depth impacts recommendation performance across different platforms.
Learning a Matching Model for Response Selection
Why Pay Special Attention to Learning?

- Matching models are becoming more and more complicated

<table>
<thead>
<tr>
<th>Test Data</th>
<th>Model</th>
<th>( R_{10@1} )</th>
<th>( R_{10@2} )</th>
<th>( R_{10@5} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \text{Built by } \text{heuristics} )</td>
<td>SMN</td>
<td>0.453</td>
<td>0.654</td>
<td>0.886</td>
</tr>
<tr>
<td></td>
<td>DAM</td>
<td>0.501</td>
<td>0.700</td>
<td>0.921</td>
</tr>
<tr>
<td>( \text{Built by } \text{human labeling} )</td>
<td>SMN</td>
<td>0.302 (↓ 15%)</td>
<td>0.464 (↓ 19%)</td>
<td>0.757 (↓ 13%)</td>
</tr>
<tr>
<td></td>
<td>DAM</td>
<td>0.325 (↓ 18%)</td>
<td>0.491 (↓ 21%)</td>
<td>0.772 (↓ 15%)</td>
</tr>
</tbody>
</table>
Learning with Unlabeled Data – Weak Annotator

- Existing learning approach:
  \[ i = 1 \sum_{n} n \gamma = 1 \log f(x)_{i}, \gamma = 1 + 1 - \gamma \log [1 - f(x)_{i}, \gamma] \]
  - Positive example: human response, \( y = 1 \)
  - Negative examples: randomly sampled responses, \( y = 0 \)

- New learning approach:
  \[ n \sum_{i=1}^{n} \max \left[ 0, f(x)_{i}, \gamma - f(x)_{i}, \gamma + s_{i} \right] \]
  - Large margin between a candidate and human response.
  - Candidates are retrieved from an index.
  - Margin is defined by a generative model as weak supervision.

Wu et al., Learning Matching Models with Weak Supervision for Response Selection in Retrieval-based Chatbots. ACL’18

<table>
<thead>
<tr>
<th>Model</th>
<th>Douban</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MAP</td>
<td>MRR</td>
</tr>
<tr>
<td>Dual-LSTM</td>
<td>0.485</td>
<td>0.527</td>
</tr>
<tr>
<td>Dual-LSTM+Weak</td>
<td>0.519</td>
<td>0.559</td>
</tr>
<tr>
<td>Multi-View</td>
<td>0.505</td>
<td>0.543</td>
</tr>
<tr>
<td>Multi-View+Weak</td>
<td>0.534</td>
<td>0.575</td>
</tr>
<tr>
<td>SMN</td>
<td>0.526</td>
<td>0.571</td>
</tr>
<tr>
<td>SMN+Weak</td>
<td>0.565</td>
<td>0.609</td>
</tr>
</tbody>
</table>

- Candidates are retrieved from an index.
Learning with Co-Teaching – Denoising with Your Peer

- **Key Ideas**
  - **Teaching:** two models judge quality of training examples mutually. The knowledge is transferred between the two models through learning protocols.
  - **Learning:** two models learn from their peers via the transferred learning protocols.
  - **Co-evolving:** through teaching and learning, the two models get improved together.
  - **Resemble:** two peer students who learn from different but related materials inspire each other during learning through knowledge exchange.

Feng et al., Learning a Matching Model with Co-Teaching for Multi-turn Response Selection in Retrieval-based Dialogue Systems. ACL'19
What to Teach: Study on Teaching Strategies

• STRATEGY 1: Teaching with Dynamic Margins; Learning Protocol=Loss Function

➢ Measuring how likely an example is a false negative through margins.

\[ \Delta_{B,i} = \max(0, \lambda[s_A(x_{B,i}^+) - s_A(x_{B,i}^-)]) \]
\[ \Delta_{A,i} = \max(0, \lambda[s_B(x_{A,i}^+) - s_B(x_{A,i}^-)]) \]

➢ Optimization with the margins

\[ L_A = \sum_{i=1}^{N_A} \max\{0, \Delta_{A,i} - s_A(x_{A,i}^+) + s_A(x_{A,i}^-)\} \]
\[ L_B = \sum_{i=1}^{N_B} \max\{0, \Delta_{B,i} - s_B(x_{B,i}^+) + s_B(x_{B,i}^-)\} \]
What to Teach: Study on Teaching Strategies (Cont.)

• STRATEGY II: Teaching with Dynamic Instance Weighting; Learning Protocol=Loss Function
  ➢ Assigning small weights to potential false negatives.
    
    $\omega_{B,i} = \begin{cases} 
    1, & \text{if } y_{B,i} = 1 \\
    1 - s_A(x_{B,i}), & \text{if } y_{B,i} = 0 
    \end{cases}$
    
    $\omega_{A,i} = \begin{cases} 
    1, & \text{if } y_{A,i} = 1 \\
    1 - s_B(x_{A,i}), & \text{if } y_{A,i} = 0 
    \end{cases}$

  ➢ Optimization with weighted losses
    
    $L_A = \sum_{i=1}^{N_A} \omega_{A,i} L(y_{A,i}, x_{A,i})$

    $L_B = \sum_{i=1}^{N_B} \omega_{B,i} L(y_{B,i}, x_{B,i})$
What to Teach: Study on Teaching Strategies (Cont.)

• STRATEGY III: Teaching with Dynamic Data Curriculum; Learning Protocol=Data
  ➢ Selecting high-confidence examples (with small loss) for learning

\[
\tilde{D}_B = \arg\min_{|\tilde{D}_B|=\delta|D_B|, \tilde{D}_B \subset D_B} L(\tilde{D}_B)
\]

\[
\tilde{D}_A = \arg\min_{|\tilde{D}_A|=\delta|D_A|, \tilde{D}_A \subset D_A} L(\tilde{D}_A)
\]
## Empirical Results

<table>
<thead>
<tr>
<th>Test Data</th>
<th>Model</th>
<th>E-Commerce</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$R_{10}@1$</td>
</tr>
<tr>
<td>Built by <em>heuristics</em></td>
<td>SMN</td>
<td>0.453</td>
</tr>
<tr>
<td></td>
<td>DAM</td>
<td>0.501</td>
</tr>
<tr>
<td></td>
<td>SMN (CoT-S1)</td>
<td>0.302 ($\downarrow$ 15%)</td>
</tr>
<tr>
<td></td>
<td>SMN (CoT-S2)</td>
<td>0.311 ($\uparrow$ 1.6%)</td>
</tr>
<tr>
<td></td>
<td>SMN (CoT-S3)</td>
<td>0.323 ($\uparrow$ 2.1%)</td>
</tr>
<tr>
<td></td>
<td>DAM</td>
<td>0.325 ($\downarrow$ 18%)</td>
</tr>
<tr>
<td></td>
<td>DAM (CoT-S1)</td>
<td>0.337 ($\uparrow$ 1.2%)</td>
</tr>
<tr>
<td></td>
<td>DAM (CoT-S2)</td>
<td>0.343 ($\uparrow$ 1.8%)</td>
</tr>
<tr>
<td></td>
<td>DAM (CoT-S3)</td>
<td>0.345 ($\uparrow$ 2.0%)</td>
</tr>
</tbody>
</table>

Built by *human labeling*
Co-Evolving Under Co-Teaching

- Model = DAM
- Data = E-Commerce Test Set
Short Summary

• Research of response selection focuses on how to build a matching model to recognize semantic relationship between an input and a response.

• Two frameworks for building a matching model.
  • Framework I: vectorization->matching.
  • Framework II: representation->matching->aggregation.

• Extra knowledge can enhance the performance of existing matching models.

• New advances.
  • Better representations.
  • Deep Matching
  • Better learning methods.
Response Generation
Roadmap

• Basic models.

• Challenges in response generation & representative models.

• Response generation with extra knowledge.

• Controllable response generation.

• Comparison between retrieval and generation.
Basic Model: Encoder-Decoder

- Sequence-to-Sequence from machine translation

Sutskever et al., Sequence to Sequence Learning with Neural Networks, NIPS’14
Vinyals & Le, A Neural Conversational Model, ICML’15 Deep Learning Workshop
Basic Model: Encoder-Attention-Decoder

• Just as machine translation, the RNNs in encoder/decoder can be replaced with CNN/Transformer.

Shang et al., Neural Responding Machine for Short-Text Conversation, ACL’15
Challenges in Response Generation (Different from MT)

• **Safe response**
  - Basic models tend to generate generic responses like ```I see``` and ```OK```.  
  - The generic responses lead to boring, short, and meaningless conversations.

• **Context modeling**
  - Conversation contexts are hierarchical, chronological, and interdependent.

• **Evaluation**
  - Diverging from references does NOT mean BAD responses.
Challenges in Response Generation (Different from MT)

• **Safe response**
  • Basic models tend to generate generic responses like ``I see'' and ``OK''.
  • The generic responses lead to boring, short, and meaningless conversations.

• **Context modeling**
  • Conversation contexts are hierarchical, chronological, and interdependent.

• **Evaluation**
  • Diverging from references does NOT mean BAD responses.
Why There Exists “Safe Response” Problem

• One-to-many relationship between messages and responses

As a result, the basic models are prone to memorize high-frequency patterns in data and generate safe responses with the patterns.
Attacking The “Safe Response” Problem

• Modeling the “one-to-many” relationship with latent variables.

• Content introducing.

• Generation with specially designed decoders.

• Learning beyond maximum likelihood estimation (MLE).
Response Generation with Latent Variables

Conditional Variational Autoencoder (CVAE): Continuous Variables [Zhao et al., ACL’17]

\[ P(y|x) \rightarrow P(y|z)P(z|x) \]

\[ \downarrow \text{Maximize ELBO} \]

\[ E_Q(z|x,y)[P(y|z,x)] - KL(P(z|x)||Q(z|x,y)) \]

Mechanism-Aware Response Machine (MARM): Discrete Variables [Zhou et al., AAAI’17]

\[ P(y|x) \rightarrow \sum_{i=1}^{M} P(y|m_i)P(m_i|x) \]

\[ \downarrow \text{Maximize MLE} \]

\[ \sum_{(x,y) \in D} \log[\sum_{i=1}^{M} P(y|m_i)P(m_i|x)] \]
Pros of the hard way:
• Explicitly control the content of responses.
• The keyword is guaranteed to appear in the response

Cons of the hard way:
• Only one keyword can be introduced into a response.

\[ P(y_1, \cdots y_{k-1} | w, x) \]

\[ P(y_{k+1}, \cdots y_n | w, y_1, \cdots y_{k-1}, x) \]
**Topic Introducing:** Soft Way I

**Topic attention:** selecting topic words for generation

\[
\alpha_{o|j}^i = \frac{\exp(\eta_o(s_{i-1}, k_j, h_T))}{\sum_{j'=1}^{n} \exp(\eta_o(s_{i-1}, k_{j'}, h_T))}.
\]

**Biased generation probability:** encouraging use of topic words in responses

\[
p_V(y_i = w) = \begin{cases} 
\frac{1}{Z} e^{\psi_V(s_i, y_{i-1}, w)}, & w \in V \cup K \\
0, & w \notin V \cup K
\end{cases}, \quad p_K(y_i = w) = \begin{cases} 
\frac{1}{Z} e^{\psi_K(s_i, y_{i-1}, c_i, w)}, & w \in K \\
0, & w \notin K
\end{cases}
\]

Xing et al., Topic-Aware Neural Response Generation. AAAI’17
Content Introducing: Soft Way II

Implicit Content Introducing with Hierarchical Gated Fusion Unit

Content: cue words estimated with PMI

\[
\text{PMI}(w_{q1}, \ldots, w_{qn}, w_r) \approx \log \frac{\prod_i p(w_{qi}|w_r)}{\prod_i p(w_{qi})} = \sum_i \log \frac{p(w_{qi}|w_r)}{p(w_{qi})} = \sum_i \text{PMI}(w_{qi}, w_r)
\]

Cue word GRU: fusing cue word into hidden states of the decoder

\[
\begin{align*}
    r_w &= \sigma(W_r C_w + U_r h_{t-1} + U_{cr} C_t + b_r) \\
    z_w &= \sigma(W_z C_w + U_z h_{t-1} + U_{cz} C_t + b_z) \\
    \tilde{h}_w &= \tanh(W_h C_w + U_h (r_w \circ h_{t-1}) + U_{ch} C_t + b_h) \\
    h_w &= (1 - z_w) \circ h_{t-1} + z_w \circ \tilde{h}_w
\end{align*}
\]
Content Introducing: Soft Way III

Retrieval Augmented Response Generation

Content: response candidates retrieved from an index

Pros of the soft way:
- No constraints on the form of content to be introduced.

Cons of the soft way:
- It is not guaranteed that the content will appear in the responses

Song et al., An Ensemble of Retrieval-Based and Generation-Based Human-Computer Conversation Systems. *IJCAI’18*
Decoding with Multi-Head Attention

Each head corresponds to a semantic space, and thus represents a kind of relationship between input and output.

Tao et al., Get The Point of My Utterance! Learning Towards Effective Responses with Multi-Head Attention Mechanism. IJCAI’18
Decoding with Specificity Control Variables

Traditional Decoder
\[ p_M(y_t = w) = w^T \sigma(W_M^h y_{t-1} + W_M^e e_{t-1} + b_M) \]

Specificity Control Decoder
\[ p_S(y_t = w) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(w^T(U \cdot W_U + b_U) - s)^2}{\sigma^2}\right) \]

Semantic-based & Specificity-based Generation

Zhang et al., Learning to Control the Specificity in Neural Response Generation. ACL’18
Learning to Diversify Responses

• From MLE to Maximum Mutual Information (MMI) [Li et al., NAACL-HLT’16]

Training: \[ \arg\max_y \frac{P(x, y)}{p(x)p(y)} \Rightarrow \arg\max_y [\lambda p(y|x) + (1 - \lambda)p(x|y)] \]

Prediction: beam search + re-ranking

• From MLE to Adversarial Learning [Li et al., EMNLP’17]

Update the Discriminator \( D \): \((x, y^+) = \) human dialogues, \((x, y^-) \sim G(y|x)\)

Update the Generator \( G \): \([D(x, y) - b(x, y)] \nabla \log(G(y|x))\)
Learning to Diversify Responses (Cont’)

• From MLE to Reinforcement Learning [Li et al., EMNLP’16]

Policy Gradient: \( \sum_{t=1}^{T} \nabla \log(p(y_t|x_t)) \left[ \sum_{i=t}^{T} R(y_i, x_i) \right] \)

Reward:
- Penalizing generic responses
- Penalizing repetition
- Encouraging mutual information

• From MLE to Adversarial Information Maximization [Zhang et al., NIPS’18]

Objective: \( \mathcal{L}_{GAN}(\theta, \psi) + \lambda \mathcal{L}_{MI}(\theta, \psi) \)

\( \mathcal{L}_{GAN}(\theta, \psi): \)

Update Discriminator \( \mathbb{E}_{T, \tilde{T}, S}[D(T, S) - D_{\psi}(\tilde{T}, S)] \)

\( \nabla_{\theta} \mathcal{L}_{MI}(\theta, \psi) = \mathbb{E}_{p_{\theta}(T|S)}[q_{\psi}(S|T) - b] \nabla_{\theta} \log(p_{\theta}(T|S)) \)

Update Generator \( \mathbb{E}_{p(Z)} \nabla_{\tilde{T}} D(\tilde{T}, S) \nabla_{\theta} \tilde{T}(S, Z) \)

\( \nabla_{\phi} \mathcal{L}_{MI}(\theta, \psi) = \mathbb{E}_{p_{\theta}(T|S)} \nabla_{\phi} q_{\phi}(S|T) \)
Challenges in Response Generation (Different from MT)

• **Safe response**
  - Basic models tend to generate generic responses like “I see” and “OK”.
  - The generic responses lead to boring, short, and meaningless conversations.

• **Context modeling**
  - Conversation contexts are hierarchical, chronological, and interdependent.

• **Evaluation**
  - Diverging from references does NOT mean BAD responses.
Incorporating Conversation History into Generation

- **Dynamic-Context Generative Models**

\[
DGM-I: k_1 = [c; m]^T f_1
\]

\[
DGM - II: k_1 = [c^T f_1^1; m^T f_1^1]
\]

\[
k_l = \sigma(k_{l-1}^T W_{f}^{l-1})
\]

\[
h_t = \sigma(h_{t-1}^T W_{hh} + k_l + s_t^T W_{in})
\]

\[
o_t = h_t^T W_{out}
\]

\[
p(s_{t+1}|s_1, \ldots s_t) = \text{softmax}(o_t)
\]

Sordoni et al., A Neural Network Approach to Context-Sensitive Generation of Conversational Responses. *NAACL-HLT’15*
Context Modeling Checklist

- Modeling the hierarchy of conversation sessions? ✗
- Modeling the relationship/dependency among words? ✗
- Modeling the relationship/dependency among utterances? ✗
- Modeling word/utterance importance ✗
Hierarchical Architectures for Context Modeling

**Hierarchical Recurrent Encoder-Decoder**

[Serban et al., AAAI’16]

h_{n,m}^{enc} = \text{RNN}(h_{n,m-1}^{enc}, x_{n,m})

h_{n}^{con} = \text{RNN}(h_{n-1}^{con}, h_{n}^{enc})

h_{n,m}^{dec} = \text{RNN}(h_{n,m-1}^{dec}, y_{n,m}, h_{n}^{con})

**Latent Variable Hierarchical Recurrent Encoder-Decoder**

[Serban et al., AAAI’17]

h_{n,m}^{enc} = \text{RNN}(h_{n,m-1}^{enc}, x_{n,m})

h_{n}^{con} = \text{RNN}(h_{n-1}^{con}, h_{n}^{enc})

h_{n,m}^{dec} = \text{RNN}(h_{n,m-1}^{dec}, y_{n,m}, h_{n-1}^{con}, z_{n})

**Hierarchical Recurrent Attention Network**

[Xing et al., AAAI’18]

\begin{align*}
    r_n &= \sum \alpha_{n,m} h_{n,m}^{enc} \\
    h_n^{con} &= \text{RNN}(h_{n+1}^{con}, r_n) \\
    h_n^{con} &= \sum \beta_n h_n^{con} \\
    h_{n,m}^{dec} &= \text{RNN}(h_{n,m-1}^{dec}, y_{n,m}, c)
\end{align*}
Challenges in Response Generation (Different from MT)

• **Safe response**
  - Basic models tend to generate generic responses like “I see” and “OK”.
  - The generic responses lead to boring, short, and meaningless conversations.

• **Context modeling**
  - Conversation contexts are hierarchical, chronological, and interdependent.

• **Evaluation**
  - Diverging from references does NOT mean BAD responses.
Evaluating Open Domain Dialogue Systems is Difficult

• Due to the one-to-many nature, responses diverging from references are NOT always bad responses.

• A common case is that multiple references are NOT available in multi-turn dialogues.

• Relevance (including validity) does NOT mean everything.

• The worse thing is that relevance is probably the easiest dimension to measure the quality of responses.
Metrics Used by Existing Work

• **Human judgment**
  • Labeling with absolute scores (e.g., 0-2) [Shang et al., ACL’15, Xing et al., AAAI’17].
  • Labeling with relative scores (e.g., pairwise comparison) [Sordoni et al., NAACL-HLT’15, Serban et al., AAAI’17].

• **Automatic evaluation**
  • Perplexity (language model) [Serban et al., AAAI’16, Xing et al., AAAI’18].
  • BLEU (MT) [Sordoni et al., NAACL-HLT’15, Li et al., NAACL-HLT’16].
  • Embedding-based similarity [Serban et al., AAAI’17].
  • Diversity: distinct-1, distinct-2 [Li et al., NAACL-HLT’16].
  • Dialogue length [Li et al., EMNLP’16].

• **Case study**
However......

• Human judgment: subjective & expensive.

• Case study: cherry-picking.

• Automatic metrics: weak correlation with human judgment.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Speaker</th>
<th>p-value</th>
<th>Pearson</th>
<th>p-value</th>
<th>Speaker</th>
<th>p-value</th>
<th>Pearson</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Greedy</td>
<td>0.2119</td>
<td>0.034</td>
<td>0.1994</td>
<td>0.047</td>
<td>0.05276</td>
<td>0.6</td>
<td>0.02049</td>
<td>0.84</td>
</tr>
<tr>
<td>Average</td>
<td>0.2259</td>
<td>0.024</td>
<td>0.1971</td>
<td>0.049</td>
<td>-0.1387</td>
<td>0.17</td>
<td>-0.1631</td>
<td>0.10</td>
</tr>
<tr>
<td>Extrema</td>
<td>0.2103</td>
<td>0.036</td>
<td>0.1842</td>
<td>0.067</td>
<td>0.09243</td>
<td>0.36</td>
<td>-0.002903</td>
<td>0.98</td>
</tr>
<tr>
<td>METEOR</td>
<td>0.1887</td>
<td>0.06</td>
<td>0.1927</td>
<td>0.055</td>
<td>0.06314</td>
<td>0.53</td>
<td>0.1419</td>
<td>0.16</td>
</tr>
<tr>
<td>BLEU-1</td>
<td>0.1665</td>
<td>0.098</td>
<td>0.1288</td>
<td>0.2</td>
<td>-0.02552</td>
<td>0.8</td>
<td>0.01929</td>
<td>0.85</td>
</tr>
<tr>
<td>BLEU-2</td>
<td>0.2076</td>
<td>&lt; 0.01</td>
<td>0.3874</td>
<td>&lt; 0.01</td>
<td>0.03819</td>
<td>0.71</td>
<td>0.0586</td>
<td>0.56</td>
</tr>
<tr>
<td>BLEU-3</td>
<td>0.3423</td>
<td>&lt; 0.01</td>
<td>0.1443</td>
<td>0.15</td>
<td>0.0878</td>
<td>0.38</td>
<td>0.1116</td>
<td>0.27</td>
</tr>
<tr>
<td>BLEU-4</td>
<td>0.3417</td>
<td>&lt; 0.01</td>
<td>0.1392</td>
<td>0.17</td>
<td>0.1218</td>
<td>0.23</td>
<td>0.1132</td>
<td>0.26</td>
</tr>
<tr>
<td>ROUGE</td>
<td>0.1235</td>
<td>0.22</td>
<td>0.09714</td>
<td>0.34</td>
<td>0.05405</td>
<td>0.5933</td>
<td>0.06401</td>
<td>0.53</td>
</tr>
<tr>
<td>Human</td>
<td>0.9476</td>
<td>&lt; 0.01</td>
<td>1.0</td>
<td>0.0</td>
<td>0.9550</td>
<td>&lt; 0.01</td>
<td>1.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>

Table 3: Correlation between each metric and human judgements for each response. Correlations shown in the human row result from randomly dividing human judges into two groups.

Model-based Evaluation Metrics

**Automatic Dialogue Evaluation Model**

<table>
<thead>
<tr>
<th>Metric</th>
<th>Spearman</th>
<th>Pearson</th>
<th>Spearman</th>
<th>Pearson</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLEU-2</td>
<td>0.039 (0.013)</td>
<td>0.081 (&lt;0.001)</td>
<td>0.051 (0.254)</td>
<td>0.120 (&lt;0.001)</td>
</tr>
<tr>
<td>BLEU-4</td>
<td>0.051 (0.001)</td>
<td>0.025 (0.113)</td>
<td>0.063 (0.156)</td>
<td>0.073 (0.103)</td>
</tr>
<tr>
<td>ROUGE</td>
<td>0.062 (&lt;0.001)</td>
<td>0.114 (&lt;0.001)</td>
<td>0.096 (0.031)</td>
<td>0.147 (&lt;0.001)</td>
</tr>
<tr>
<td>METEOR</td>
<td>0.021 (0.189)</td>
<td>0.022 (0.165)</td>
<td>0.013 (0.745)</td>
<td>0.021 (0.601)</td>
</tr>
<tr>
<td>T2V</td>
<td>0.140 (&lt;0.001)</td>
<td>0.141 (&lt;0.001)</td>
<td>0.140 (&lt;0.001)</td>
<td>0.141 (&lt;0.001)</td>
</tr>
<tr>
<td>VHRED</td>
<td>-0.035 (0.062)</td>
<td>-0.030 (0.106)</td>
<td>-0.091 (0.023)</td>
<td>-0.010 (0.805)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Validation set</th>
<th>Test set</th>
</tr>
</thead>
<tbody>
<tr>
<td>C-ADEM</td>
<td>0.338 (&lt;0.001)</td>
</tr>
<tr>
<td>R-ADEM</td>
<td>0.404 (&lt;0.001)</td>
</tr>
<tr>
<td>ADEM (T2V)</td>
<td>0.252 (&lt;0.001)</td>
</tr>
<tr>
<td>ADEM</td>
<td><strong>0.410</strong> (&lt;0.001)</td>
</tr>
</tbody>
</table>

Lowe et al., Towards An Automatic Turing Test: Learning to Evaluate Dialogue Responses. ACL’17
Model-based Evaluation Metrics (Cont’)

![Model-based Evaluation Metrics Table](image)


\[ J = \max \left[ 0, \Delta - s_{U}q, r + s_{U}q, r \right] \]

Model learning does NOT need human labeling
Unfortunately......

• Model-based evaluation metrics are NOT as good as expected (ADEM as an example):
  • Scores of various sets of responses (including adversarial examples) spread around a relatively stable mean with small variance.
  • Easy to be fooled by simple adversarial attacks (e.g., specific generic responses).
  • Easy to be manipulated by a white box attack.

There still a long way to go for evaluating open domain dialogue generation!

Sai et al., Re-evaluating ADEM: A Deeper Look at Scoring Dialogue Responses. AAAI’19
References

- Oriol Vinyals and Quoc V. Le. A Neural Conversational Model. *ICML’15 Deep Learning Workshop*
- Lifeng Shang, Zhengdong Lu, and Hang Li. Neural Responding Machine for Short-Text Conversation. *ACL’15*
- Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. Learning Discourse-level Diversity for Neural Dialog Models using Conditional Variational Autoencoders. *ACL’17*
- Tiancheng Zhao, Kyusong Lee, and Maxine Eskenazi. Unsupervised Discrete Sentence Representation Learning for Interpretable Neural Dialog Generation. *ACL’18*
- Ganbin Zhou, Ping Luo, Rongyu Cao, Fen Lin, Bo Chen, and Qing He. Mechanism-Aware Neural Machine for Dialogue Response Generation. *AAAI’17*
- Ganbin Zhou, Ping Luo, Yijun Xiao, Fen Lin, Bo Chen, and Qing He. Elastic Responding Machine for Dialog Generation with Dynamically Mechanism Selecting. *AAAI’18*
- Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. Topic-Aware Neural Response Generation. *AAAI’17*
- Lili Yao, Yaoyuan Zhang, Yansong Feng, Dongyan Zhao, and Rui Yan. Towards Implicit Content-Introducing for Generative Short-Text Conversation Systems. *EMNLP’17*
- Yiping Song, Cheng-Te Li, Jian-Yun Nie, Ming Zhang, Dongyan Zhao, and Rui Yan. An Ensemble of Retrieval-Based and Generation-Based Human-Computer Conversation Systems. *IJCAI’18.*
- Lili Mou, Yiping Song, Rui Yan, Ge Li, Lu Zhang, and Zhi Jin. Sequence to Backward and Forward Sequences: A Content-Introducing Approach to Generative Short-Text Conversation. *COLING’16*
References

• Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao and Bill Dolan. A Diversity-Promoting Objective Function for Neural Conversation Models. *NAACL-HLT’16*

• Jiwei Li, Will Monroe, Tianlin Shi, Sebastien Jean, Alan Ritter, and Dan Jurafsky. Adversarial Learning for Neural Dialogue Generation. *EMNLP’17*


• Yizhe Zhang, Michel Galley, Jianfeng Gao, Zhe Gan, Xiujun Li, Chris Brockett, and Bill Dolan. Generating Informative and Diverse Conversational Responses via Adversarial Information Maximization. *NIPS’18*

• Chongyang Tao, Shen Gao, Mingyue Shang, Wei Wu, Dongyan Zhao, and Rui Yan. Get The Point of My Utterance! Learning Towards Effective Responses with Multi-Head Attention Mechanism. *IJCAI’18*

• Xiang Gao, Sungjin Lee, Yizhe Zhang, Chris Brockett, Michel Galley, Jianfeng Gao, and Bill Dolan. Jointly Optimizing Diversity and Relevance in Neural Response Generation. *NAACL-HLT’19*

• Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao and Bill Dolan. A Neural Network Approach to Context-Sensitive Generation of Conversational Responses. *NAACL-HLT’15*


• Zhiliang Tian, Rui Yan, Lili Mou, Yiping Song, Yansong Feng and Dongyan Zhao. How to Make Contexts More Useful? An Empirical Study to Context-Aware Neural Conversation Models. *ACL’17*

• Ananya B. Sai, Mithun Das Gupta, Mitesh M. Khapra, and Mukundhan Srinivasan. Re-evaluating ADEM: A Deeper Look at Scoring Dialogue Responses. *AAAI’19*
References

• Iulian V. Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yushua Bengio. A Hierarchical Latent Variable Encoder-Decoder Model for Generating Dialogues. AAAI’17

• Iulian V. Serban, Tim Klinger, Gerald Tesauro, Kartik Talamadupula, Bowen Zhou, Yushua Bengio, and Aaron Courville. Multiresolution Recurrent Neural Networks: An Application to Dialogue Response Generation. AAAI’17

• Chen Xing, Wei Wu, Yu Wu, Ming Zhou, Yalou Huang, and Wei-Ying Ma. Hierarchical Recurrent Attention Network for Response Generation. AAAI’18


• Ryan Lowe, Michael Noseworthy, Iulian V. Serban, Nicolas A.-Gontier, Yushua Bengio, and Joelle Pineau. Towards An Automatic Turing Test: Learning to Evaluate Dialogue Responses. ACL’17


• Ruqing Zhang, Jiafeng Guo, Yixing Fan, Yanyan Lan, Jun Xu, and Xueqi Cheng. Learning to Control the Specificity in Neural Response Generation. ACL’18

• Elia Bruni and Raquel Fernandez. Adversarial Evaluation for Open-Domain Dialogue Generation. SIGDIAL’17
Response Generation According to Extra Knowledge
Extra Knowledge=Personality

• Speaker model: injecting speaker ID into decoding.

\[ h_{t}^{dec} = \text{LSTM}(h_{t-1}^{dec}, e_t, v_i) \]  embedding of user i

• Speaker-Addressee model: injecting both speaker ID and receiver ID into decoding.

\[ h_{t}^{dec} = \text{LSTM}(h_{t-1}^{dec}, e_t, V_{i,j}) \]  speaker – addressee embedding

\[ V_{i,j} = \tanh(W_1 \cdot v_i + W_2 \cdot v_j) \]  embedding of user i embedding of user j

Li et al., A Persona-Based Neural Conversation Model. ACL’16
Beyond Speaker IDs: Grounding Chat by User Profiles

The **PERSONA-CHAT** Dataset

- **Profiles:**
  - 1155 from Amazon Mechanical Turk.
  - $\geq 5$ sentences per profile.
  - Revised by additional Turkers.

- **Dialogues:**
  - 10,981 dyadic human-human conversations according to randomly assigned profiles.
  - 164,354 utterances.

- **Tasks:**
  - Persona-based response selection.
  - Persona-based response generation.
  - Conversational Intelligence (ConvAI) Challenge

---

Zhang et al., Personalizing Dialogue Agents: I have a dog, do you have pets too? *ACL’18*
Extra Knowledge=Documents

Knowledge Grounded Generation Model

Memory Network

CONVERSATION HISTORY

Going to Kusakabe tonight

RESPONSE

Try omakase, the best in town

fact (document): \{r_i\}

\[ m_i = Ar_i \]

\[ c_i = Cr_i \]

\[ p_i = \text{softmax}(u_i^T m_i) \]

\[ o = \sum_{i=1}^{k} p_i c_i \]

Ghazvininejad et al., A Knowledge Grounded Neural Conversation Model. AAAI’18
Datasets for Document Grounded Chat

**CMU Document Grounded Conversations (CMU_DoG) [Zhou et al., EMNLP’18]**

- **Documents:**
  - 30 movie-related wiki articles covering various movie types.
  - 4 sections selected from each document.

- **Scenarios:**
  - One introduces the documents to the other.
  - The two discuss the content of the documents.

- **Dialogues:**
  - 4112 conversations (2128 for Scenario I and 1984 for Scenario II) collected from Amazon Mechanical Turk.
  - 21.43 turns per conversation.

- **Tasks:**
  - Response selection/generation.
Datasets for Document Grounded Chat (Cont’)

Wizard of Wikipedia [Dinan et al., ICLR’19]

• **Documents:**
  • 1365 open domain topics with each linked to a wiki article.
  • Document (paragraph) retrieval for the last two turns of dialogues.

• **Scenarios:**
  • One introduces the documents to the other.

• **Dialogues:**
  • 22,311 conversations collected from crowdsourcing. 20,999 turns in total.

• **Tasks:**
  • Response selection/generation.

---

**Difference from CMU_DoG**

• Broader topics, closer to open-domain conversation.

• Dynamic knowledge rather than static knowledge.

• Larger size.
Extra Knowledge = Emotions

Zhou et al., Emotional Chatting Machine: Emotional Conversation Generation with Internal and External Memory. AAAI’18
Extra Knowledge=Knowledge Base

Commonsense Knowledge Aware Conversational Model

Encoder

Decoder

Moonlight lacks the ultraviolet rays of sunlight. I don’t think that’s a lack of uv.

Zhou et al., Commonsense Knowledge Aware Conversation Generation with Graph Attention. IJCAI’18
References

• Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing Dialogue Agents: I have a dog, do you have pets too? ACL’18

• Jiwei Li, Michel Galley, Chris Brockett, Georgios P.Spithourakis, Jianfeng Gao, and Bill Dolan. A Persona-Based Neural Conversation Model. ACL’16

• Pierre-Emmanuel Mazar´e, Samuel Humeau, Martin Raison, and Antoine Bordes. Training Millions of Personalized Dialogue Agents. EMNLP’18

• Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. A Knowledge Grounded Neural Conversation Model. AAAI’18

• Kangyan Zhou, Shrimai Prabhumoye, and Alan W Black. A Dataset for Document Grounded Conversations. EMNLP’18

• Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. Wizard of Wikipedia: Knowledge-Powered Conversational Agents. ICLR’19

• Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. Emotional Chatting Machine: Emotional Conversation Generation with Internal and External Memory. AAAI’18

• Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. Commonsense Knowledge Aware Conversation Generation with Graph Attention. IJCAI’18


• Pierre Colombo, Wojciech Witon, Ashutosh Modi, James Kennedy, and Mubbasir Kapadia. Affect-Driven Dialog Generation. NAAACL-HLT’19
Controllable Response Generation via Meta-Words

Definition (Meta-Word):
A structured record that characterizes the response to generate with a group of variables (attributes).

Example: meta-word=dialogue act | response length | if copy message | if multiple utterances | specificity level

<table>
<thead>
<tr>
<th>Message:</th>
<th>last week I have a nice trip to New York!</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meta-word:</td>
<td>Act: yes-no question</td>
</tr>
<tr>
<td>Response 1:</td>
<td>Is New York more expensive than California?</td>
</tr>
<tr>
<td>Response 2:</td>
<td>Cool, sounds great! What is the tallest building in this city, Chrysler building?</td>
</tr>
<tr>
<td>Response 3:</td>
<td>I don’t know what you are talking about. But it seems good.</td>
</tr>
</tbody>
</table>

Xu et al., Neural Response Generation with Meta-Words. ACL’19
Advantages of Response Generation with Meta-Words

• Explainable
  • Meta-words tell developers and end users what responses they may have before the responses are generated.

• Controllable
  • Meta-word is an interface that allows developers to customize responses by tailoring the attributes.

• General
  • Different kinds of extra knowledge, such as user ids, topics, and emotions, can be treated as attributes of a meta-word.

• Scalable
  • Feature engineering of generation models.
Goal Tracking Memory Network

Tracking status of meta-word expression

Key $\mathcal{M}_i.k = \text{Rep}(m_i.k)$,
Goal $\mathcal{M}_i.g = \text{Rep}(m_i.v)$,
Status $\mathcal{M}_i.v_0 = 0$.

Updating state memory with ADD (for sub-expression) or SUB (for over-expression) operations

$\hat{V}_t(i) = \mathcal{M}_i.v_{t-1} - g_t(i) \odot \Delta^{SU}_t(i)$,
$\mathcal{M}_i.v_t = \hat{V}_t(i) + (1 - g_t(i)) \odot \Delta^{ADD}_t(i)$.

Monitoring difference between goal and status

Difference Vector $d_t^i = (\mathcal{M}_i.g - \mathcal{M}_i.v_t) \oplus (\mathcal{M}_i.g \odot \mathcal{M}_i.v_t)$.

Attention

$a_t = \sum_{i=1}^{t} a_i \cdot (Ud_t^i)$,
$a_t^i = \text{softmax}((s_t)^T(Ud_t^i)),$
## Feature Engineering on Meta-Words

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Multiple utterances</th>
<th>Dialog Act</th>
<th>Length</th>
<th>Copy Ratio</th>
<th>Specificity</th>
<th>PPL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Twitter</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>70.19</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td></td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>67.23</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>62.13</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>50.36</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>42.05</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>38.57</td>
</tr>
<tr>
<td>Reddit</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>72.43</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td></td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>65.17</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>61.92</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>49.67</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>41.78</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>37.96</td>
</tr>
</tbody>
</table>
Retrieval v.s. Generation

Retrieval

• Pros
  • Informative & fluent responses.
  • Easy to update (e.g., by updating index).
  • Easy to evaluate (e.g., evaluated as a learning to rank/classification task).

• Cons
  • No new responses.
  • System quality is bonded to index quality.

Generation

• Pros
  • End-to-end learning.
  • New responses (e.g., with specific personas or emotions).

• Cons
  • Hard to evaluate.
  • Safe responses.
  • Difficult to update.
Takeaway Messages

• Chatbot, or broadly speaking, conversational AI is emerging as the next opportunity AI brings to industry.

• Two approaches to implementing an open domain conversation system.
  • Retrieval-based methods -> response re-using.
    • Sentence embedding based matching.
    • Interaction based matching.
  • Generation-based methods -> response synthesis.
    • Three challenges.
      • Safe response.
      • Context modelling.
      • Evaluation.
    • Generation with extra knowledge & controllable generation.
Looking Forward

• Model Design -> Model Learning

• Knowledge Agnostic -> Knowledge Aware

• Single Modality -> Multiple Modalities
Single Modality to Multiple Modalities

**Conversation + Scene + Sentiment**
Caption and Question
Image Scene + Sentiment
Facial Expression

Conversation Turn 1: “Hanging out on Saturday.”
Conversation Turn 2: “Did you have a good time?”

Sweet Boy
Easy Life
Awesome Times
Smile
Lips Part
Cheek Raise

Sun-screen
Hat

**Query:** you are my creator.
A1: you are my savior.
A2: i am the something.
A3: and i will show you my power.

Q: i’d suffer the consequence.
A1: i’m sorry.
A2: i can’t.
A3: i would rather die than be with you.

Q: he was supposed to kick back for or five mil, was it?
A1: yeah.
A2: i’m saying he’s a good guy
A3: do you know how much i love you?

**Huber et al., Emotional Dialogue Generation using Image-Grounded Language Models. CHI’18**

**Chu et al., A Face-to-Face Neural Conversation Model. CVPR’18**

**Pasunuru et al., Game-based Video-Context Dialogue. EMNLP’18**
Looking Forward

• Model Design -> Model Learning

• Knowledge Agnostic -> Knowledge Aware

• Single Modality -> Multiple Modalities

• Big Data -> Small Data
  • Big data: mixed intentions, styles, characters, ......
  • Small data: specific intention, style, character, ......
Thanks!