A Survey on Construction and Enhancement Methods in Service Chatbots Design

Zhenhui PENG

Supervisor: Prof. Xiaojuan Ma

2019/01/16

Outline

1. Introduction
2. Core Design Philosophy
3. Enhancement of Chatbot
4. Future Directions

Outline

1. Introduction
2. Core Design Philosophy
3. Enhancement of Chatbot
4. Future Directions

Background

Chatbot --- a computer program which conducts a natural conversation with users via speech or text. *(Mauldin et al., AAAI'94)*

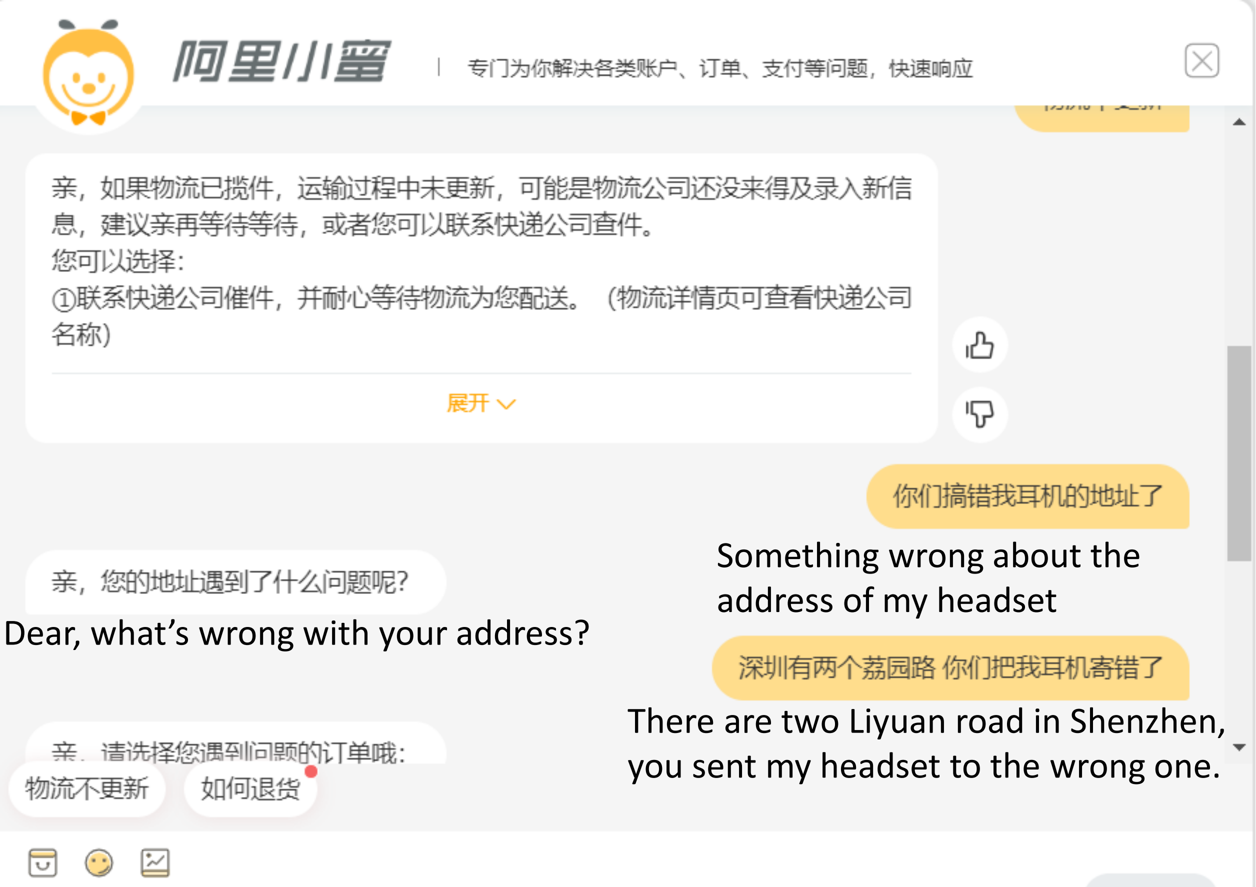
- For fun
  - Microsoft Xiaoice *(2014)*

Background

Chatbot --- a computer program which conducts a natural conversation with users via speech or text. *(Mauldin et al., AAAI’94)*

- For fun
  - Microsoft Xiaoice *(2014)*

- **Provide services in specific domains**
  - Structured tasks: schedule meetings, restaurant reservations, etc.
  - Unstructured tasks: online customer care, question answering, etc.



Motivation

Challenges: given a user's request, **how to provide a satisfactory response.**

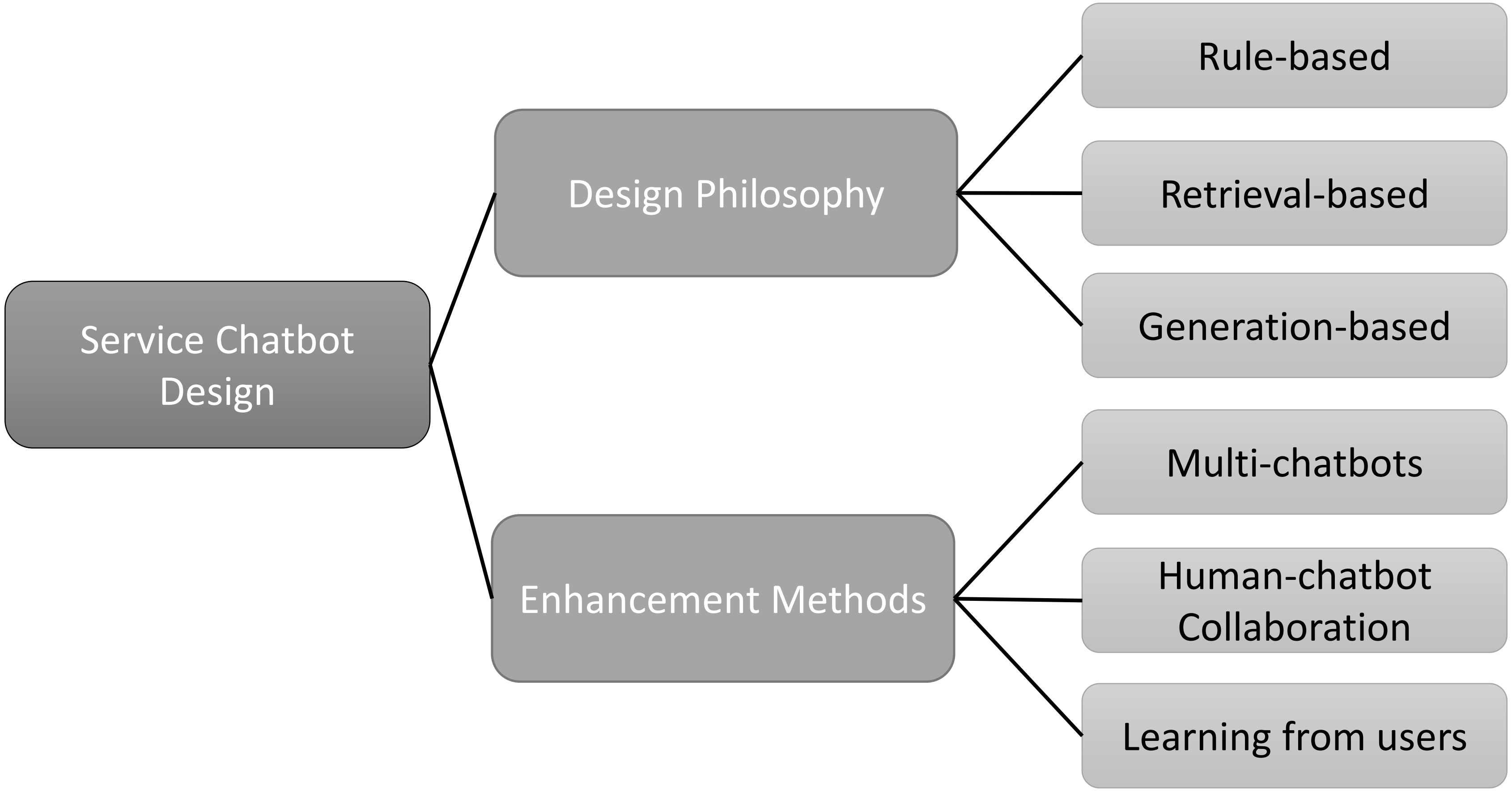
- Appropriate: on the same topic and makes sense *(Xu et al., CHI 2017)*
- Helpful: contains useful and concrete information *(Xu et al., CHI 2017)*
- Tone-aware: conveys feelings like empathy and passion *(Hu et al., CHI 2018)*

And how to enhance service chatbots capabilities over time.

- To handle a broader scope of service requests

In the rest of the survey, we consider both the requests and responses are textual utterances, sequences, or sentences

Taxonomy


Service Chatbot Design
Design Philosophy
Enhancement Methods
Rule-based
Retrieval-based
Generation-based
Multi-chatbots
Human-chatbot Collaboration
Learning from users

Outline

1. Introduction

2. Core Design Philosophy

3. Enhancement of Chatbot

4. Future Directions

Core Design Philosophy

Philosophy	Techniques	Papers
Rule-based	Pattern matching	*(Weizenbaum et al., 1966); (Colby et al., 1972); (Wallace, 2009)*
	Modular task-oriented system	*(Chen et al., SIGKDD 2017)*
Retrieval-based	TF-IDF	*(Lowe et al., SIGDIAL 2015)*
	DNN-based	*(Lu et al., NIPS 2013); (Hu et al., NIPS 2014)*
	RNN-based	*(Lowe et al., SIGDIAL 2015); (Zhou et al., EMNLP 2016)*
Generation-based	Statistical Machine Translation	*(Ritter et al., EMNLP 2011)*
	Seq2Seq	*(Sutskever et al., NIPS 2014); (Xu et al., CHI 2017)*
	Seq2Seq + attention mechanism	*(Shang et al., ACL 2015)*
	Seq2Seq + hierarchical structure	*(Serban et al., AAAI 2016)*
	Seq2Seq + memory network	*(Ghazvininejad et al., AAAI 2018)*

Rule-based

Definition: Mainly built on manually constructed rules.

- Pattern Matching: e.g., patterns in requests, "if-then" logic, response templates.
- Slot fillings: e.g., modular task-oriented system

Pattern Matching: ELIZA *(Weizenbaum et al., 1966)*

```
(0 YOU 0 ME) [pattern] -> (WHAT MAKES YOU THINK I 3 YOU) [transform]
```

User: It seems that you like me.

ELIZA: What makes you think that I like you?

keyword   pattern   Ranked list of transforms

$$
\begin{array}{l}
(K \; ((D_1) \; (R_{1,\,1}) \; (R_{1,\,2}) \cdots (R_{1,\,m_1})) \\
\quad ((D_2) \; (R_{2,\,1}) \; (R_{2,\,2}) \cdots (R_{2,\,m_2})) \\
\quad \vdots \\
\quad ((D_n) \; (R_{n,\,1}) \; (R_{n,\,2}) \cdots (R_{n,\,m_n})))
\end{array}
$$

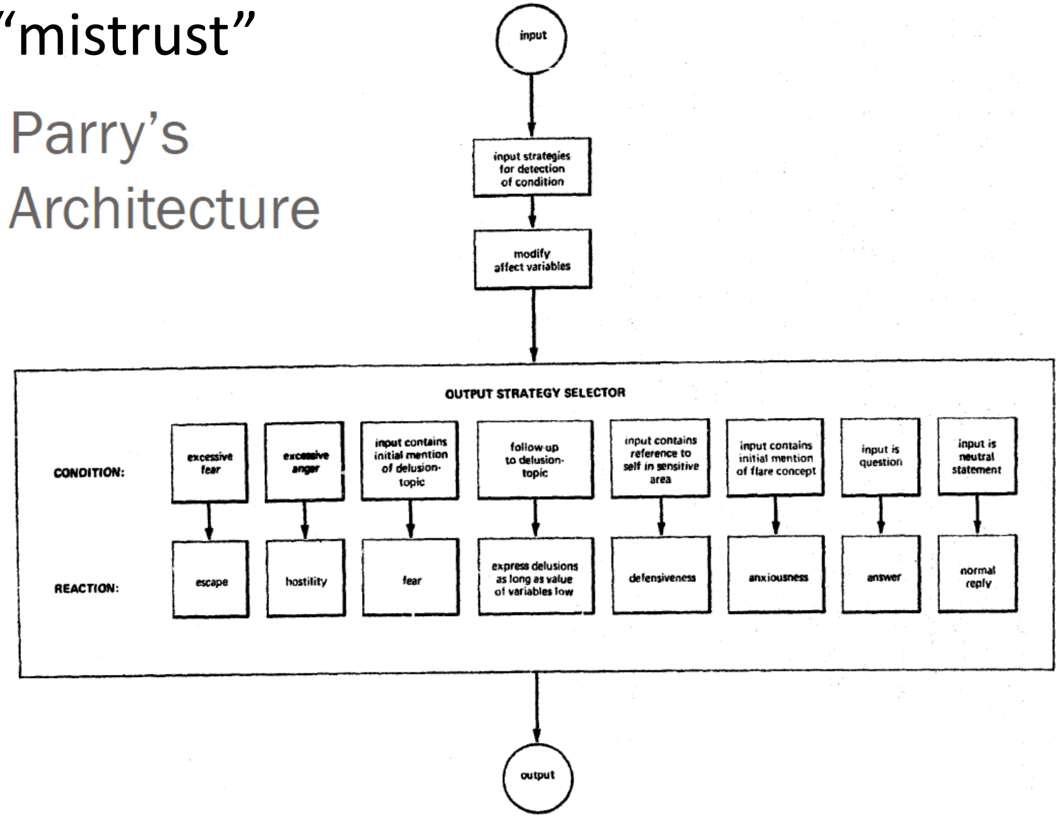
Rule-based

Pattern Matching: Parry *(Colby et al., 1972)*

- Add affective variables like "fear", "anger" and "mistrust"
- Lots of complex rules

  E.g., when a user mentions Parry,
  - Decreases fear if mistrust is low
  - Increases anger if mistrust is high

Parry's Architecture



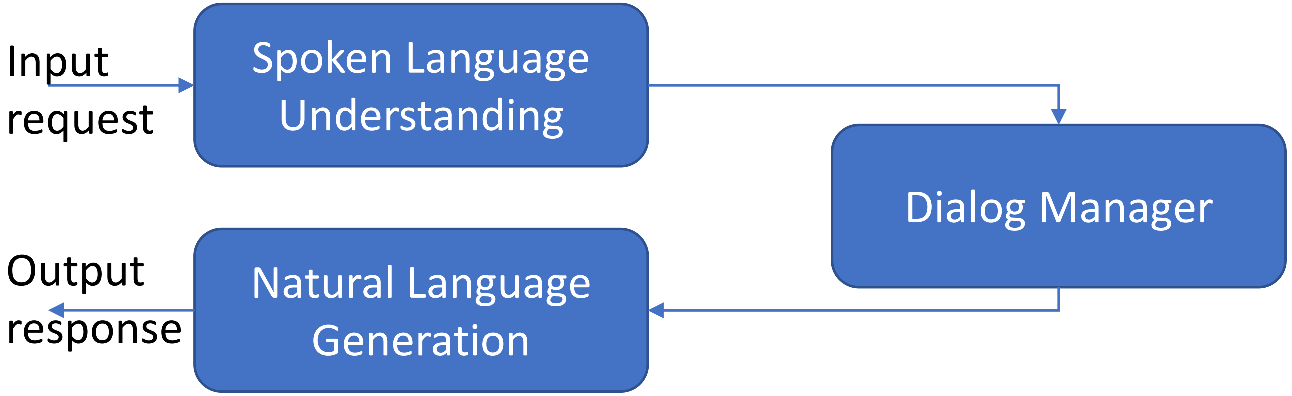
Pattern Matching: ALICE *(Wallace, 2009)*

```
<category>
<pattern>YES</pattern>
<that>DO YOU LIKE MOVIES</that>
<template>What is your favorite movie?</template>
</category>
```

Rule-based

Slots filling: Modular task-oriented system *(Chen et al., SIGKDD 2017)*

Sentence	Show	flights	from	Beijing	to	Hong	Kong	today
Slots	O	O	O	B-dept	O	B-arr	I-arr	B-date
Entity	O	O	O	B-city	O	B-city	I-city	B-date
Intent	Find_Flight							
Domain	Airline Travel							



All needed slots are filled
Takes action to get flight information

Filling template: "Here are the flights from B-dept to B-arr I-arr B-date: flight information"
Response: Here are the flights from Beijing to Hong Kong today: 10:50 – 14:40 $200.

Modular task-oriented system

- Available platforms: Microsoft LUIS, IBM Watson Assistant, Dialogueflow, WIT.AI, etc.
- Needs to define: intent, entities, logic, template

Brief Summary

Techniques	Pros	Cons
Pattern Matching	Easy and robust in the domains that have structured knowledge, e.g., online shoes shopping	Difficult to anticipate all user's intentions and design rules in complex or unstructured scenarios, e.g., Ubuntu technical support
Modular task-oriented system		

Data-driven approaches are needed

Retrieval-based

Definition: Select the response that best matches the user's request by searching a pre-constructed conversational repository.

Key: request-response matching

- Request-based strategy $[r_{\mathrm{argmax}_i \, \mathrm{sim}(q, q_i)}]$: retrieve the response r_i whose associated request q_i is most similar to the user's input request q ;
- Response-based strategy $[r_{\mathrm{argmax}_i \, \mathrm{sim}(q, r_i)}]$: retrieve the response r_i which is most similar to the user's input request q.

General idea: transform the request and response into some numeric or vector representations.

TF-IDF: "term-frequency – inverse document frequency"

- Term-frequency: the number of times the word appears in a given request $f(w, q)$
- Inverse document frequency: puts a penalty on how often this word appears elsewhere in the repository. $|q \in D : w \in q|$ *(Lowe et al., SIGDIAL 2015)*

$$\text{tf-idf}(w, q, D) = f(w, q) \times \log \frac{N}{|q \in D : w \in q|} \quad (2.1)$$

- D: the collection of the requests or responses
- N: total number of requests or responses in the repository
- q: the request

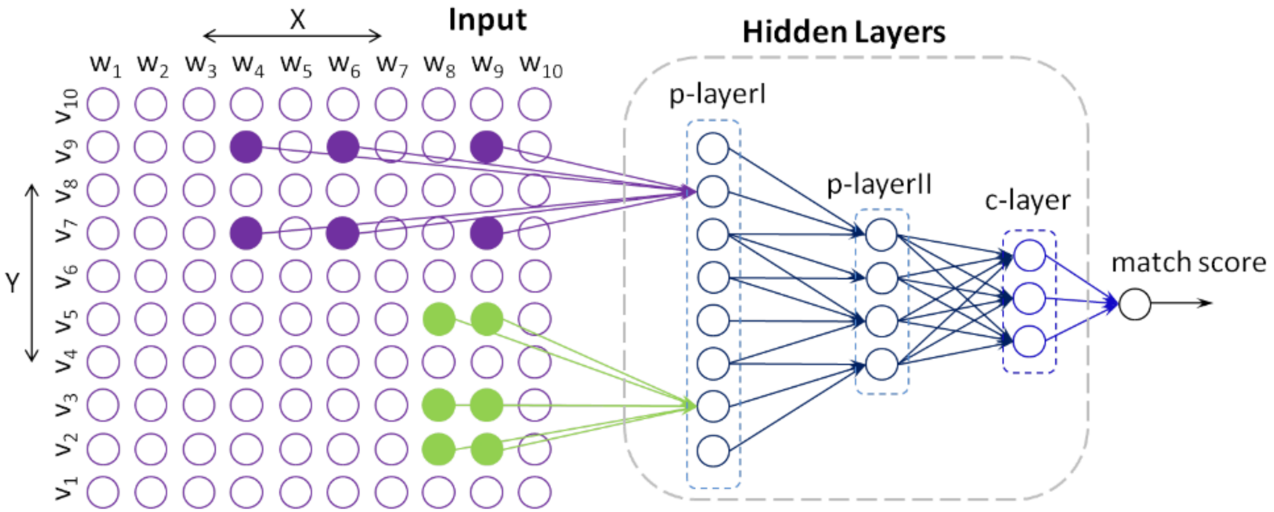
Concatenate all tf-idf scores together, calculate consine similarity, select the response with highest score.

Retrieval-based

TF-IDF: simple, without training, but does not efficiently capture the semantics of the sentence

DNN-based methods

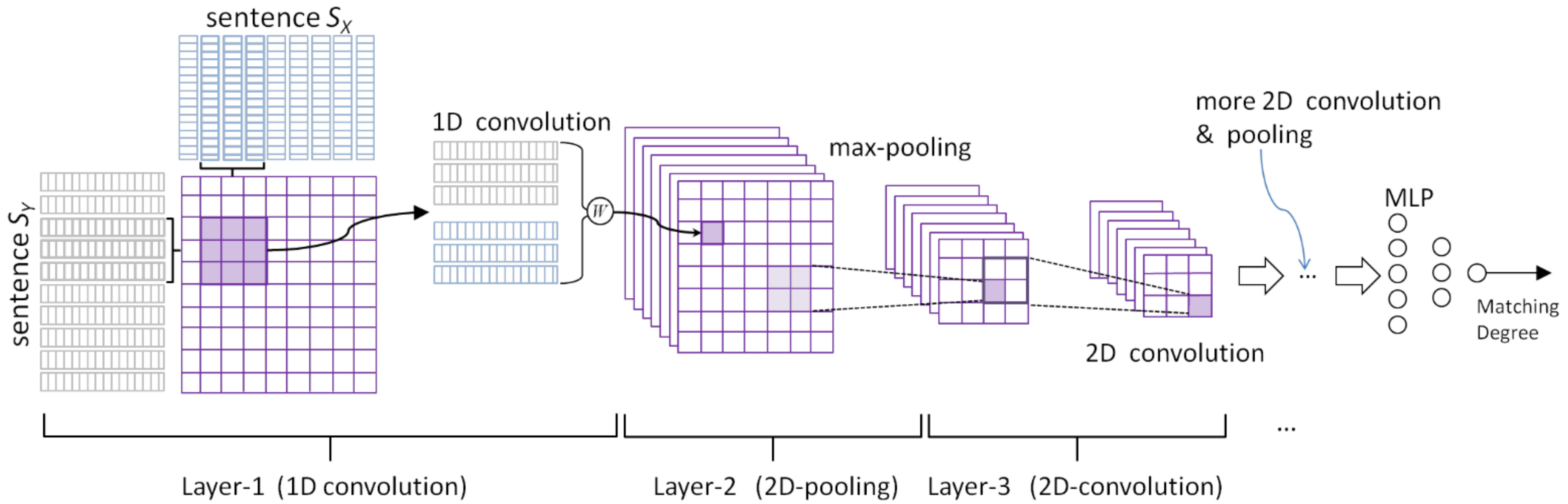
- DEEPMATCH *(Lu et al., NIPS 2013)*
  - Interaction space of bag-of-words vectors
  - Experiment on a traveling-related (Question, Answer) pairs dataset and a Weibo dataset



Only (w_i, v_j) pattern features --> Use CNN

DNN-based methods

- DEEPMATCH *(Lu et al., NIPS 2013)*
- ARC-II *(Hu et al., NIPS 2014)*
  - Models all the possible combination of the word embedding vectors

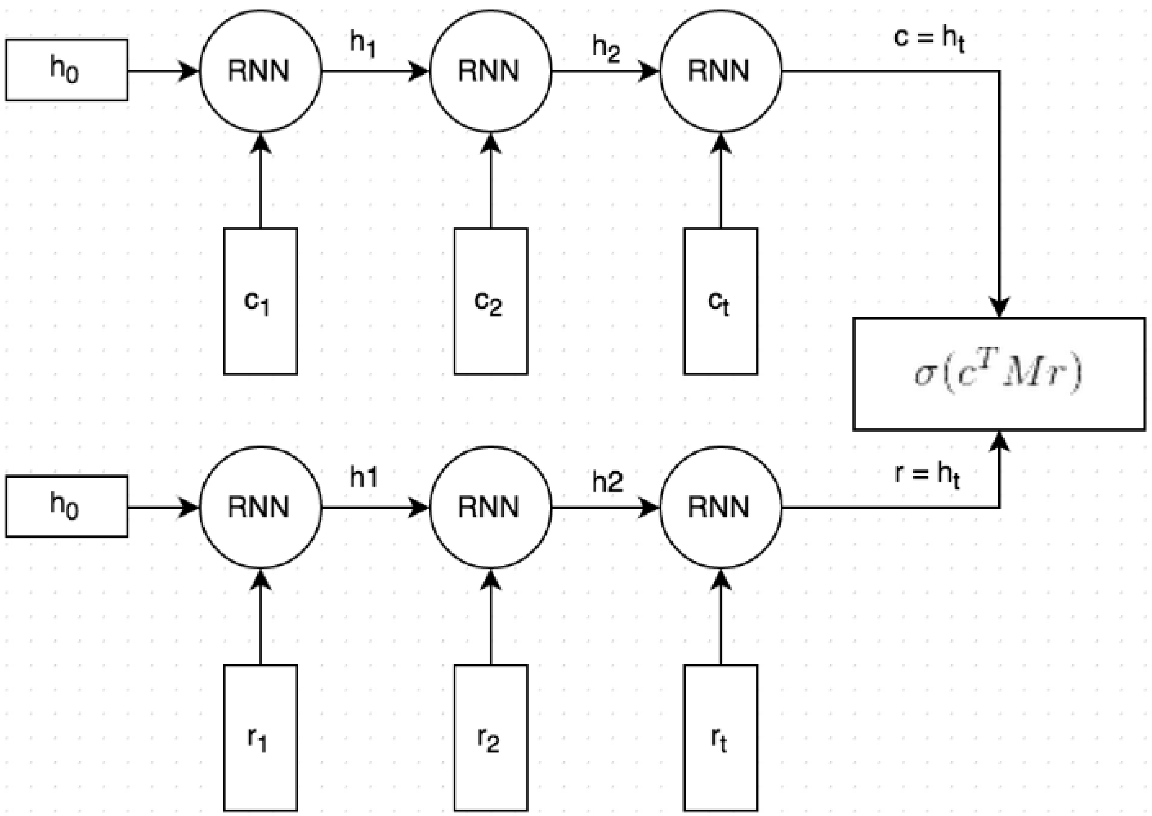


Retrieval-based

DNN-based methods: May not efficiently capture the sequential information

RNN-based methods

- RNN model in *(Lowe et al., SIGDIAL 2015)*
  - LSTM, GRU are commonly used as hidden units
  - Experiment in Ubuntu technical support dataset



$$h_t = f(h_{t-1}, x_t) = f(W_h h_{t-1} + W_x x_t) \qquad (2.2)$$

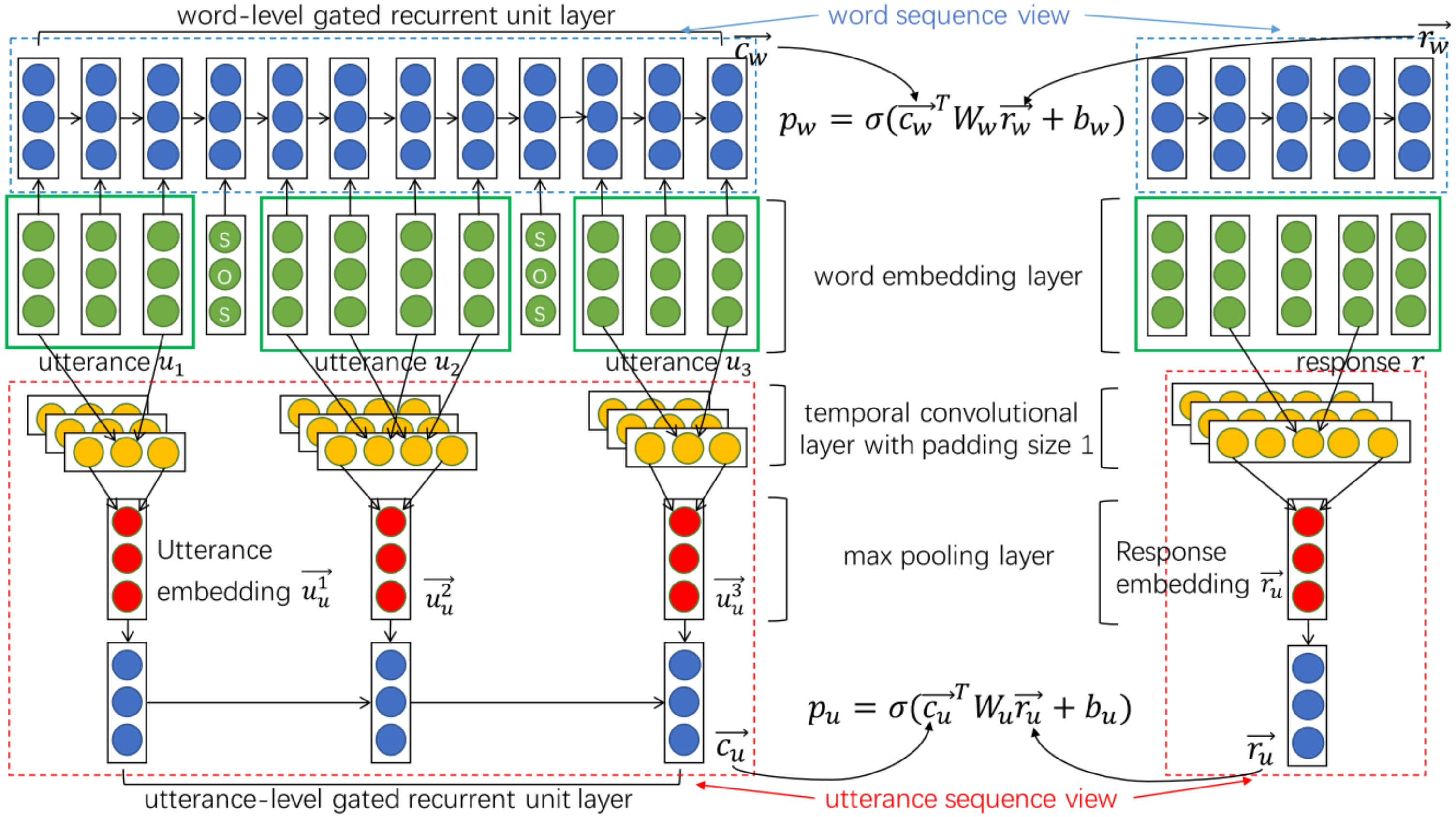
- h_t : the hidden state at time step t
- x_t : observed variable (e.g., word) at time step t
- W_h, W_x : weights

Only single-turn information --> Use previous turns

Retrieval-based

RNN-based methods

- RNN model in *(Lowe et al., SIGDIAL 2015)*
- Multi-view model *(Zhou et al., EMNLP 2016)*
  - **Word-level semantics** and dependencies in the connected utterances
  - **Utterance-level semantic** and discourse information



Retrieval-based

Brief Summary

Techniques	Pros	Cons
TF-IDF	Easy to use without training	Can not efficiently represent the importance of words
DNN-based	Learn the conncections between two utterances more efficiently	May not efficiently capture the sequential information
RNN-based	Can efficiently represent the sequential information	Needs a lot of training data

In general

Pros	Cons
Can handle more requests; Literal human utterance; Various expressions	Inconsistent personality; Easily out of context; Limited by size of repository

Needs Generation-based

Generation-based

Definition: synthesize a new sentence word by word as the response to the users' requests

Phrase-based statistical machine translation (SMT) *(Ritter et al., EMNLP 2011)*

- Strong relation between many request-response pairs
- Experiment on Twitter dataset

Stimulus: I'm slowly making this soup ...... and it smells gorgeous!

Response: I'll bet it looks delicious too! Haha

Could work badly since the responses are often not semantically matched to the requests as in translations

Generation-based

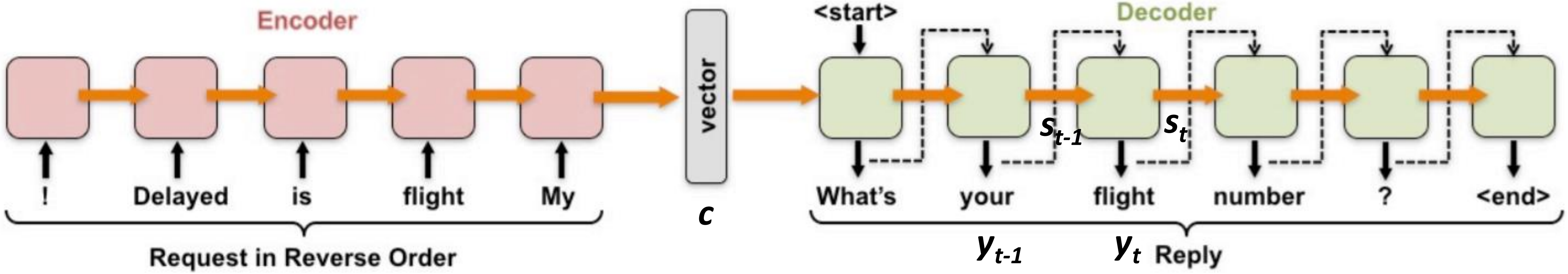
Seq2Seq (Sequence-to-Sequence) *(Sutskever et al., NIPS 2014)*

$$s_t = f(y_{t-1}, s_{t-1}, c) \tag{2.3}$$

$$p_t = \mathrm{softmax}(s_t, y_{t-1}) \tag{2.4}$$

$$p(y_1, \ldots, y_{T'} | x_1, \ldots, x_T) = p(y_1|c) \prod_{t=2} p(y_t|c, y_1, \ldots, y_{t-1}) \tag{2.5}$$

Application example in online customer care *(Xu et al., CHI 2017)*



Only use the hidden state of the last word as the context vector --> Use attention mechanism

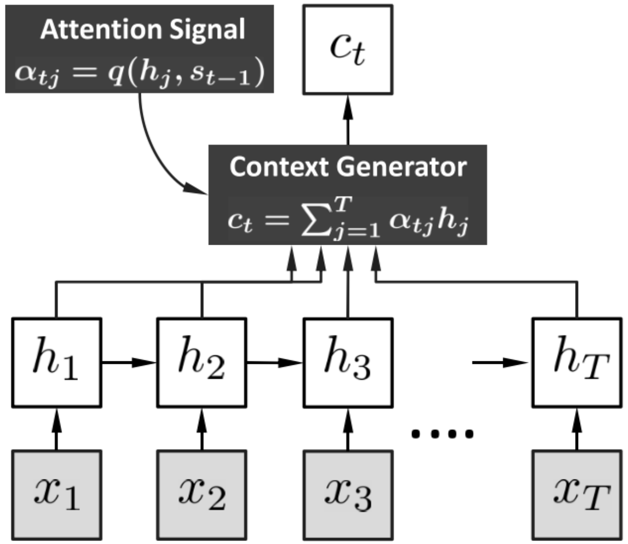
Generation-based

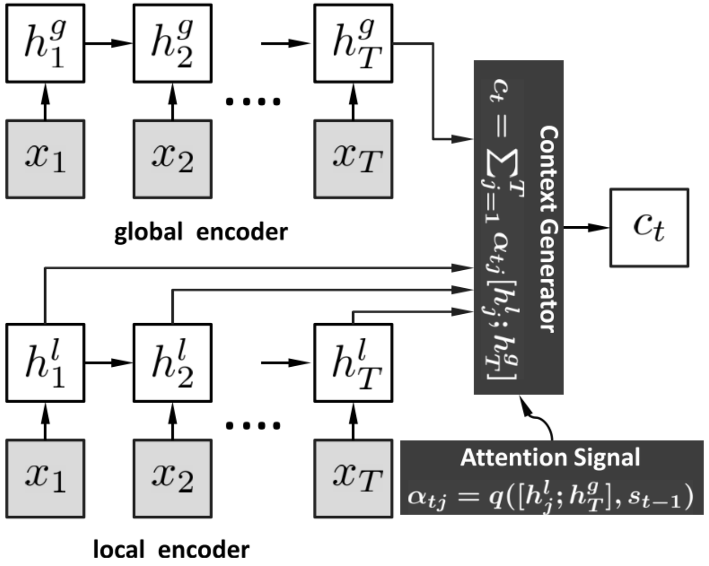
Seq2Seq + attention mechanism *(Shang et al., ACL 2015)*

- The context vector is conditioned on the combination of all hidden units in the request

$$c_t = \sum_{j=1}^{T} \alpha_{tj} h_j \tag{2.6}$$

$$\alpha_{tj} = q(h_j, s_{t-1}) \tag{2.7}$$



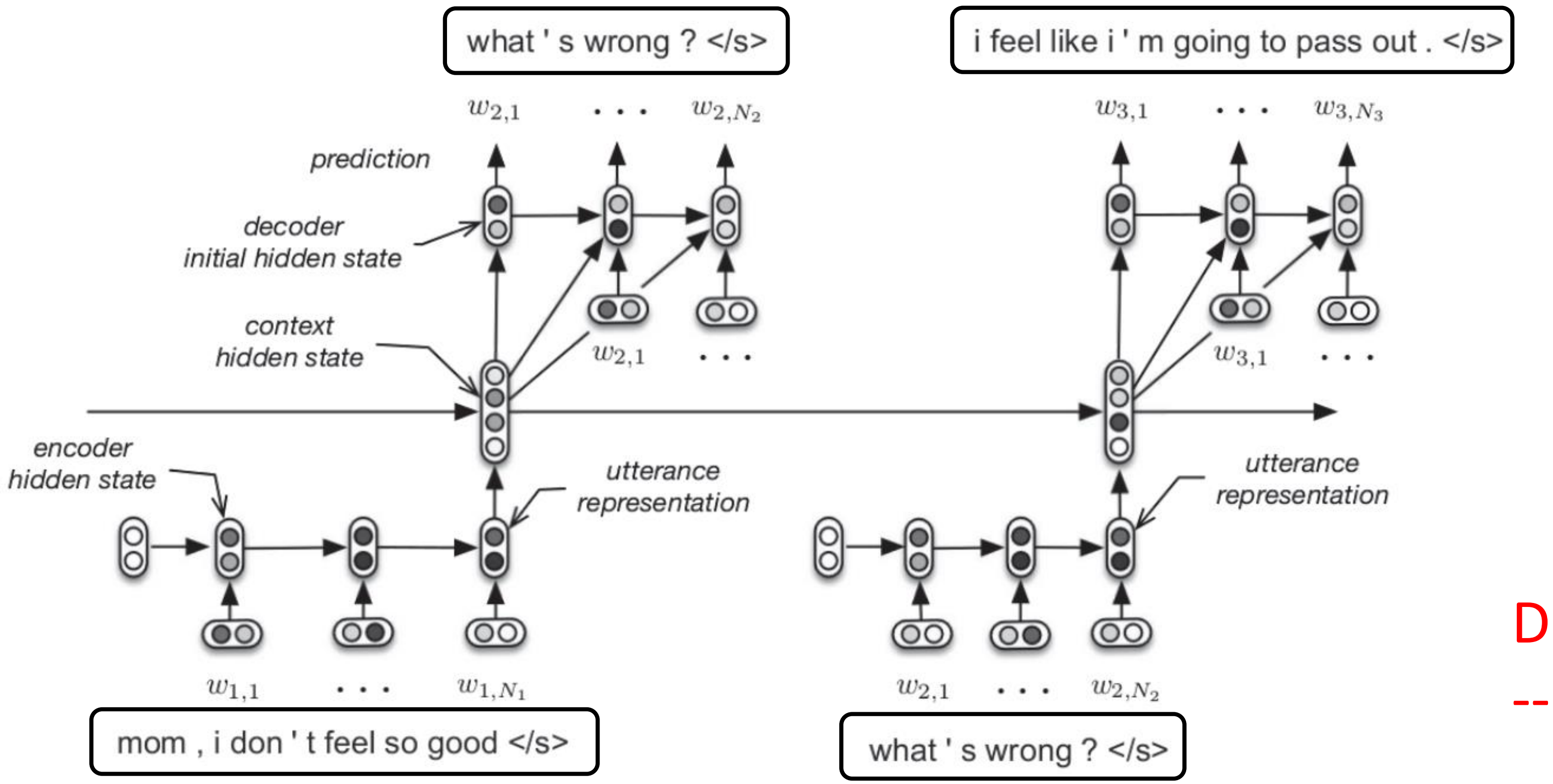


Only generates response based on one previous request
--> Use a hierarchical structure

Generation-based

Seq2Seq + hierarchical structure *(Serban et al., AAAI 2016)*

$$P_\theta(U_1, ..., U_M) = \prod_{m=1}^{M} P_\theta(U_m|U_{<m}) = \prod_{m=1}^{M} \prod_{n=1}^{N_m} P_\theta(w_{m,n}|w_{m,<n}, U_{<m}) \quad (2.8)$$

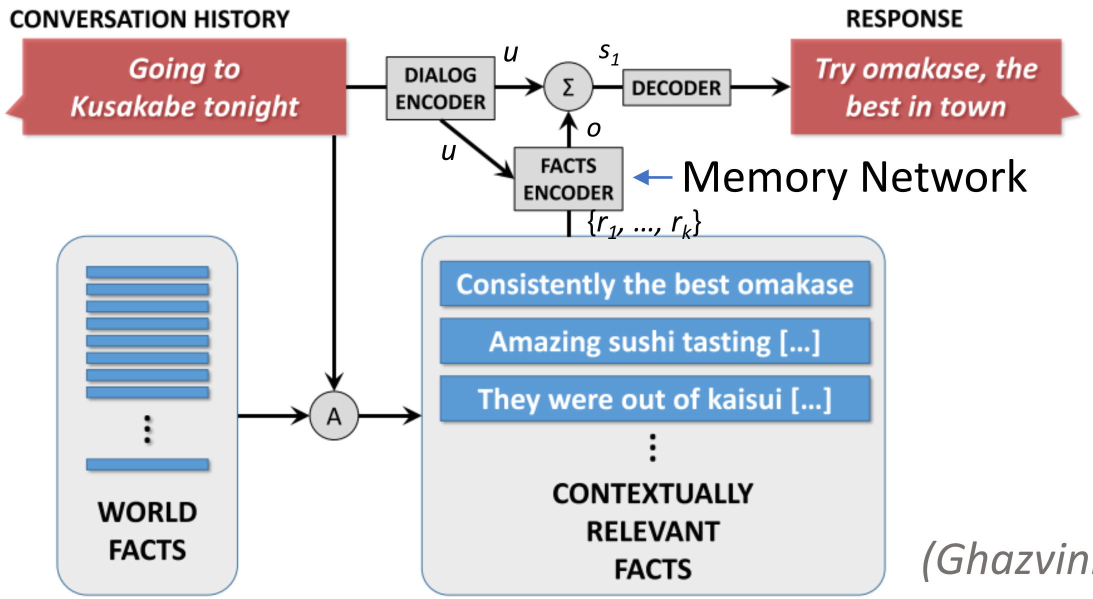


Does not use external knowledge
--> Use memory network

Seq2Seq + memory network *(Serban et al., AAAI 2016)*

$$m_i = Ar_i; c_i = Cr_i; p_i = \mathrm{softmax}(u^T m_i); o = \sum_{i=1}^{k} p_i c_i; s_1 = o + u \qquad (2.9)$$

A: Just had an awesome dinner at [...] Great recommendation [...]
B: One of my favorite places I've ever been to in NYC. The food is great and the service is lackluster.



(Ghazvininejad et al., AAAI 2018)

- ➢ Experiment in the Twitter dataset grounded by Foursquare tips (e.g., comments about restaurant and other commercial establishments)

Generation-based

Brief Summary

Techniques	Idea or Purpose
Phrase-based SMT	Strong structural relation between many request-response pairs
Seq2Seq	Encode the request word by word as a vector and decode it word by word
Seq2Seq + attention mechanism	A word in the response may strongly relates to different parts in the request
Seq2Seq + hierarchical structure	Make use of the information in previous turns
Seq2seq + memory network	Make use of external knowledge to generate more informative responses

In general

Pros	Cons
Can generate new responses; Can add in external knowledge; Highly coherent	Still prone to generate universal sentences; Need a huge training dataset

Evaluation Metrics

Evaluate the quality of a chatbot's response given the request

- Automatic Metrics for reference
  - Word perplexity
    - Measure the ability to regenerate the exact dialogue

model with parameters θ, dataset with N triples $\{U_1^n, U_2^n U_3^n\}_{n=1}^N$

$$\exp\left(-\frac{1}{N_W}\sum_{n=1}^{N}\log P_\theta(U_1^n, U_2^n, U_3^n)\right) \tag{2.10}$$

where N_W is the number of tokens in the entire dataset and $P_\theta(U_1^n, U_2^n, U_3^n)$ is the probability of regenerating the exact triple (U_1^n, U_2^n, U_3^n).

Evaluation Metrics

Evaluate the quality of a chatbot's response given the request

- Automatic Metrics for reference
  - Word perplexity
  - BLEU (bilingual evaluation understudy)
    - Grades a response according to n-gram matches to the reference

$$\mathrm{BLEU} = \mathrm{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \qquad (2.11)$$

- BP: the brevity penalty on the length of the utterance
- p_n : the propability that n-grams in generated response occur in the real response
- N: max number of gram (e.g., 4)
- w_n : weight for each n-gram (e.g., 1/4)

Evaluation Metrics

Evaluate the quality of a chatbot's response given the request

- Human-based Metrics
  - Pair-wise comparison: let a human choose which of the two responses is more suitable, more appropriate, more helpful, etc. *(Ritter et al., EMNLP 2011); (Shang et al., ACL 2015)*
  - Likert Scale: rate the appropriateness, helpfulness, passion, etc. *(Xu et al., CHI 2017); (Hu et al., CHI 2018)*
  - Case studies: analyze the response in depth. *(Ghazvininejad et al., AAAI 2018); (Hu et al., CHI 2018)*
  - Interview: analyze the chatbot design in depth

Currently more convincing method, but need time and money

Summary

Philosophy	Techniques	Pros	Cons	Scenarios
Rule-based	Pattern matching (ELIZA, PARRY, ALICE) Modular task-oriented system	Easy to start; Robust, safe in narrow domains; Context-aware	Hard to extend; Need structured domain knowledge; Need a lot of hand-crafted features	Restaurant reservation; Movie booking; Food ordering; Online shopping
Retrieval-based	TF-IDF DNN-based RNN-based	Can handle more requests; Literal human utterance; Various expression	Easily out of context; Inconsistent personality; Limited by size of repository	Domain-specific (e.g., travel) question answering ; Technical support
Generation-based	Phrase-based SMT Seq2Seq-based: + attention; + hierarchical structure; + memory network	Can generate new responses; Can add in external knowledge; Highly coherent	Prone to generate universal sentences; Easily not informative; Need a huge training dataset	Online customer care; Technical support; Entertainment

Outline

Enhancement of Chatbot

Enhancement Methods	Techniques	Papers
Multi-chatbots Design	Reinforcement learning for re-ranker policy	*(Serban et al., CoRR 2017)*
	Data-driven re-ranker models	*(Qiu et al., ACL 2017);* *(Song et al., IJCAI 2018)*
Human-chatbot Collaboration	CoChat: external memory + HRNN	*(Luo et al., AAAI 2018)*
	Evorus: crowd-powered, automates itself over time	*(Huang et al., CHI 2018)*
Learning from Users	Programming by demonstration	*(Li et al., CHI 2017)*
	Verbal instruction	*(Azaria et al., AAAI 2016)*

Multi-chatbots Design

Ideas

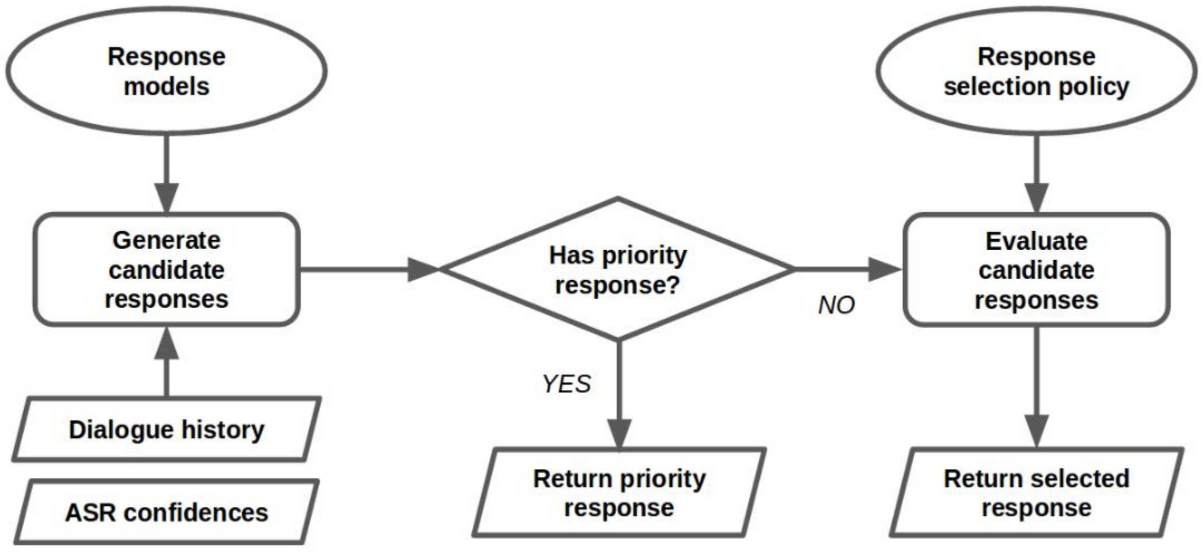
- Combine available chatbots that have different expertise to satisfy user's multiple-domain needs;
- Assemble chatbots built on different methods to absorb their merits in the same domain.

Key: Re-ranker (or response selection policy) which scores all candidates to pick the highest-score response.

Multi-chatbot Design

A Deep Reinforcement Learning Chatbot *(Serban et al., CoRR 2017)*

- 2016 Amazon Alexa Prize competition
- Combine 22 different response models
- Trained on crowdsourced data and real-world user interactions via RL
- Sequential decision making problem



$$R = \sum_{t=1}^{T} \gamma^t r_t \qquad (3.1)$$

- $\gamma \in (0, 1]$: discount factor
- r_t: reward after taking action at time step t, here is the labled 1-5 points of appropriateness

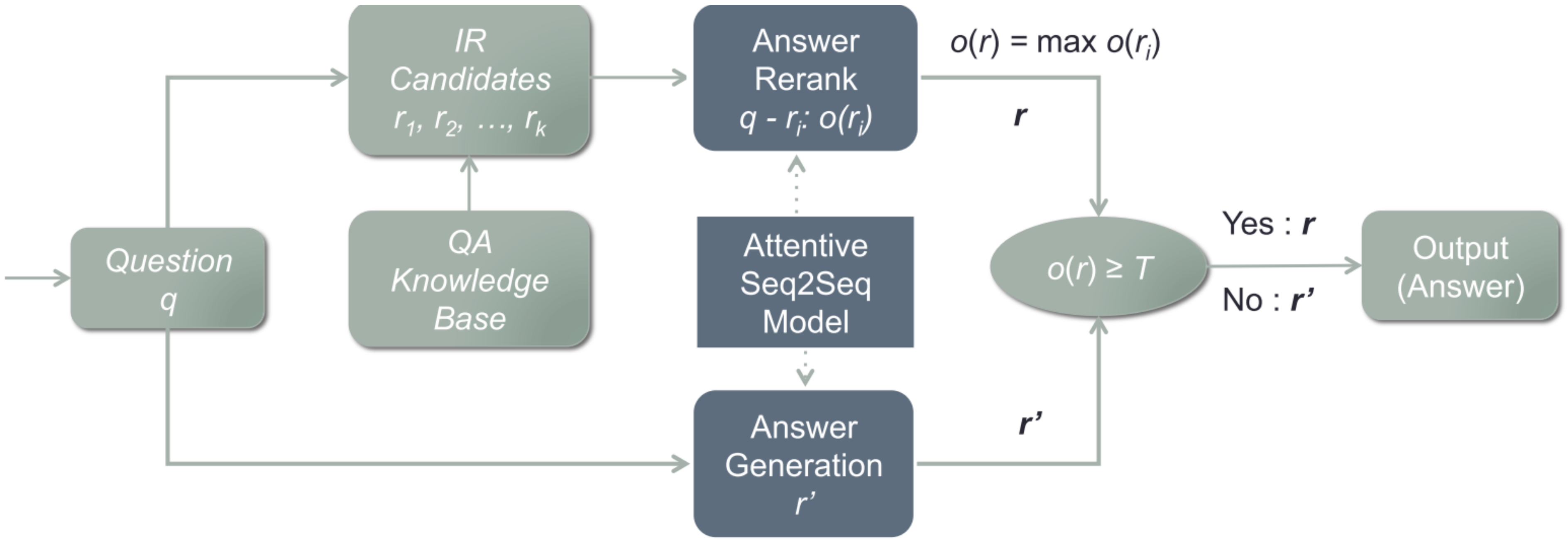
Needs a lot of labeled data

Multi-chatbot Design

Data-driven re-ranker models

- ➢ AliMe Chat *(Qiu et al., ACL 2017)*

$$s^{\text{Mean-Prob}} = \frac{1}{n}\sum_{i=1}^{n} p(y_i = w_i|\theta_i) \quad (3.2)$$



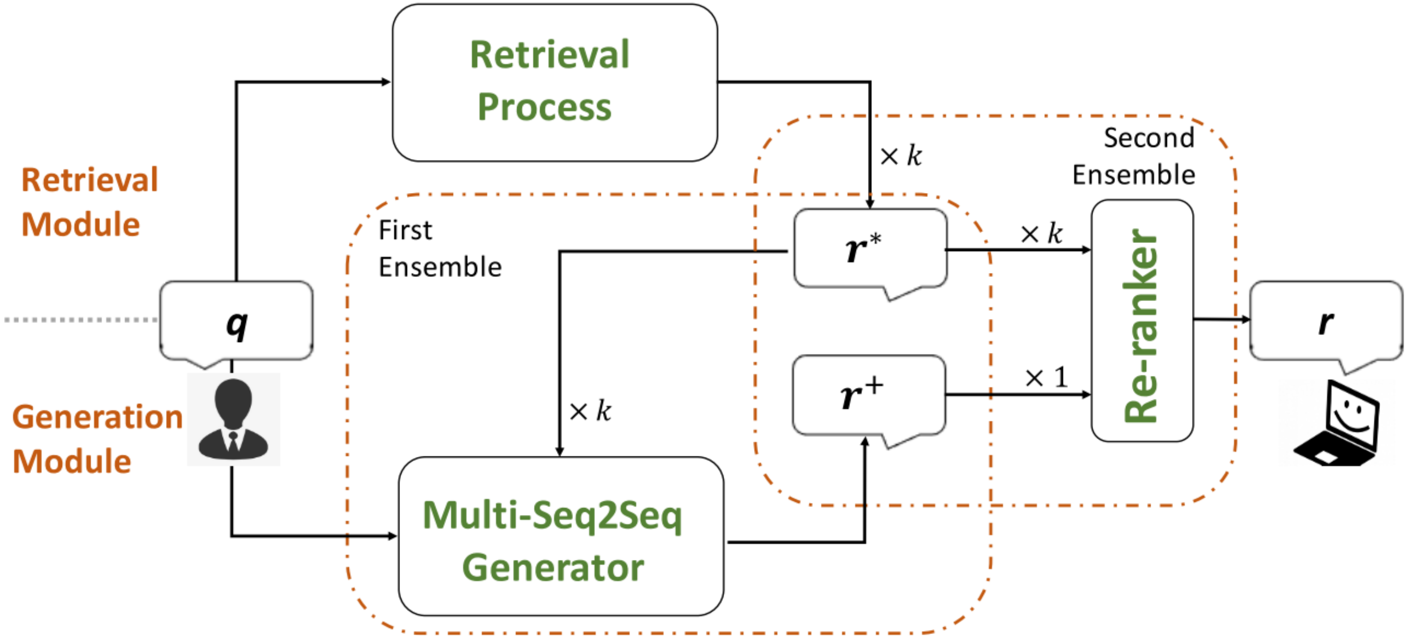
Multi-chatbot Design

Data-driven re-ranker models

➢ An ensemble of retrieval-based and generation-based chatbots

(*Song et al., IJCAI 2018*)

- Further make use of the retrieval candidates
- Train re-ranker with high-level features, e.g., term similarity, entity similarity, topic similarity, length, etc.
- Trained on Baidu Tieba dataset



In general, can not learn new skills outside the capability of the chatbot ensemble

Human-chatbot Collaboration

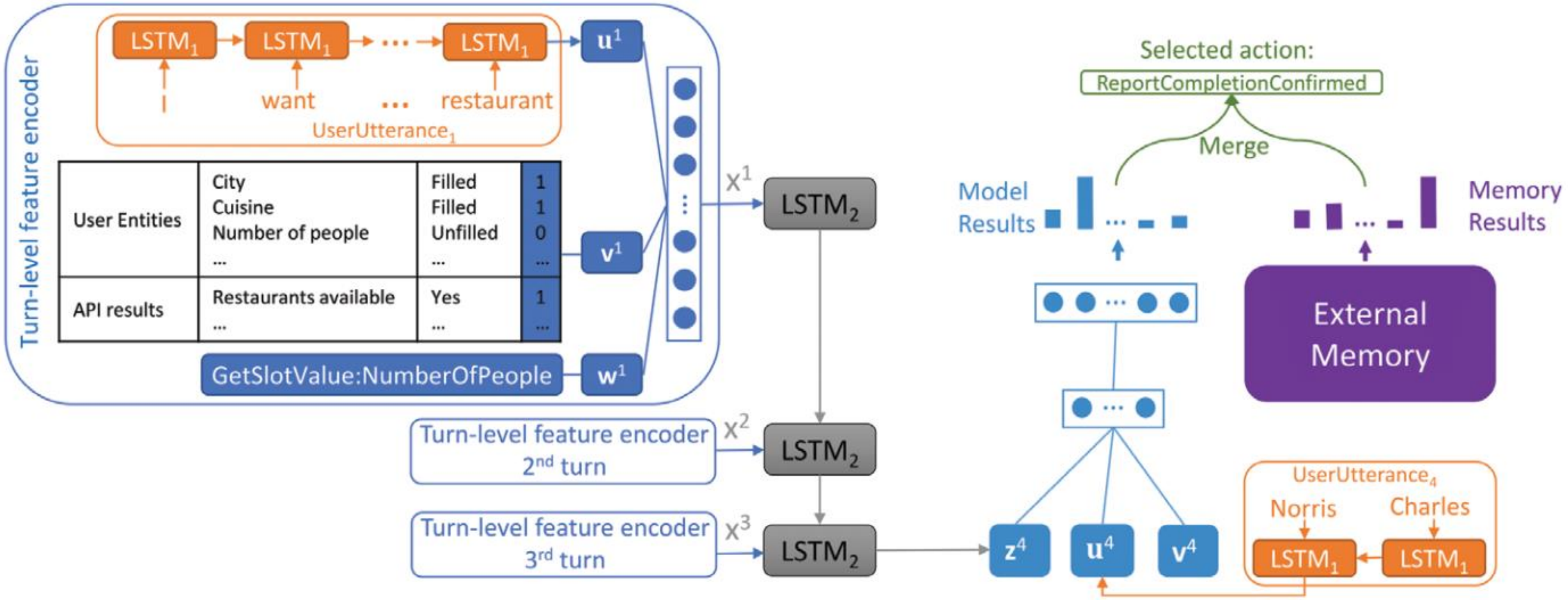
General idea

- Fully automatic service chatbots are still potentially problematic by generating wrong responses or taking unexpected actions
- Necessary to involve human workers in some failure-sensitive real-world domains

Key: Enable the chatbots to learn a new action from human workers for handling similar cases later

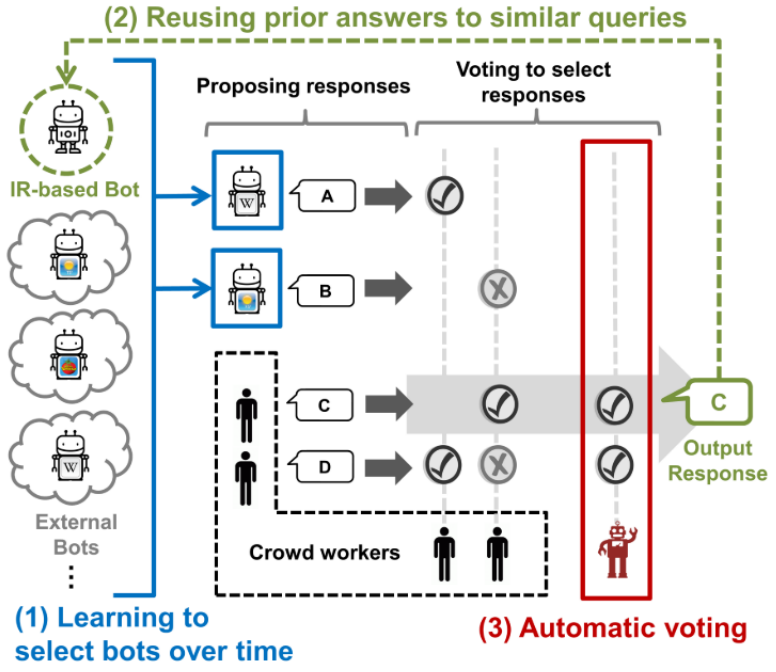
CoChat: external memory + HRNN *(Luo et al., AAAI 2018)*

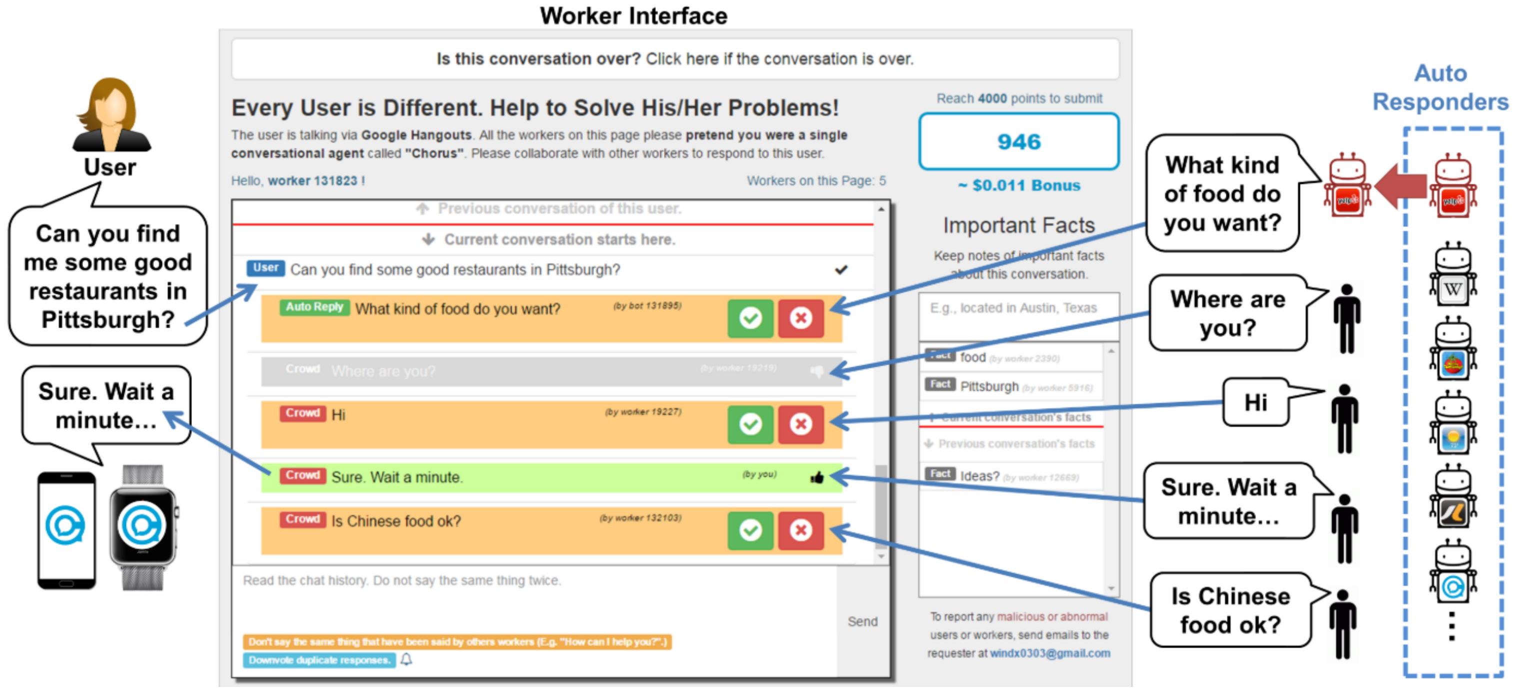
- Address the one-shot learning challenges of the new actions by increasing their possibilities if their recorded dialogue states are similar to the current dialogue state.



Only experiment in narrow domains with structured knowledge

Evorus: crowd-powered, automates itself over time *(Huang et al., CHI 2018)*



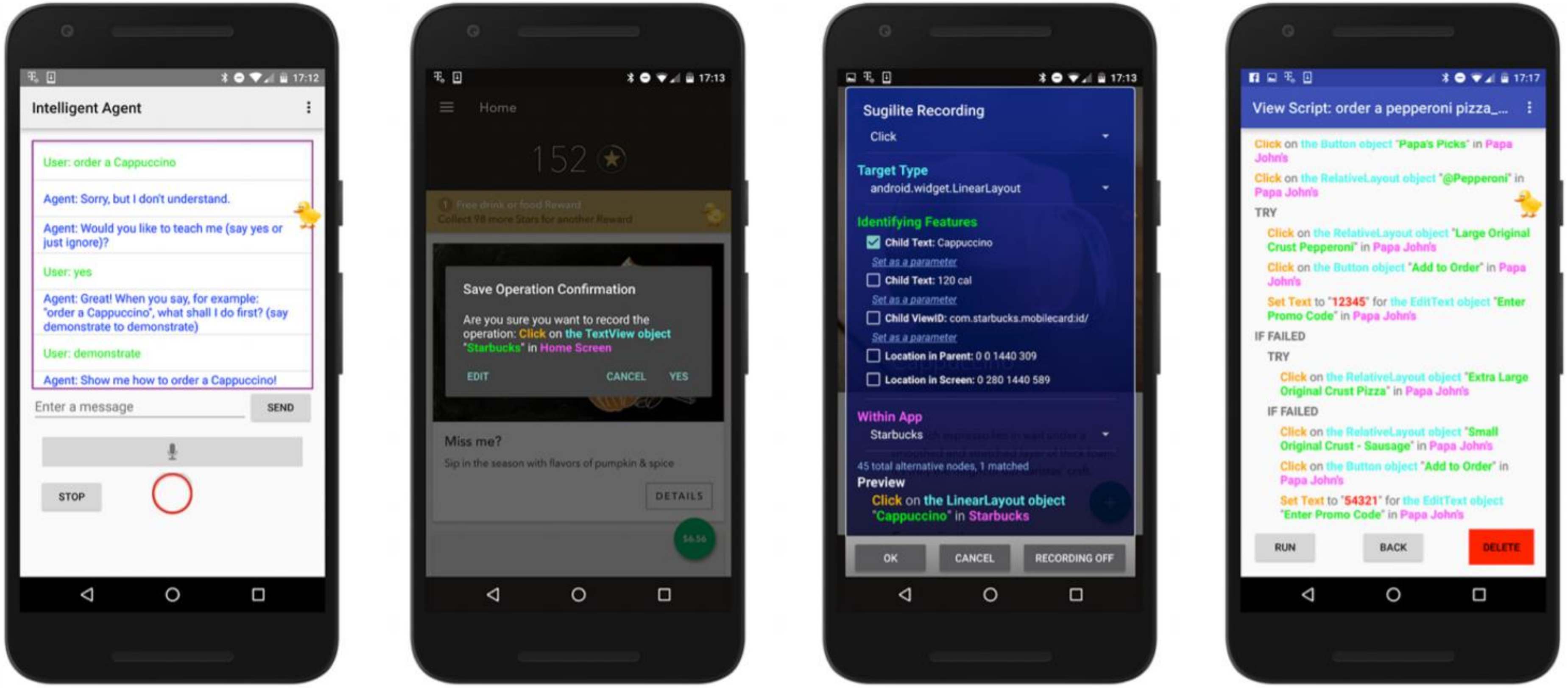


The cost paid for the crowd is expensive, and the answers may not be consistent

Learning from Users

Programming by demonstration *(Li et al., CHI 2017)*

- E.g., teach the agent how to order a green tea via a third-party app

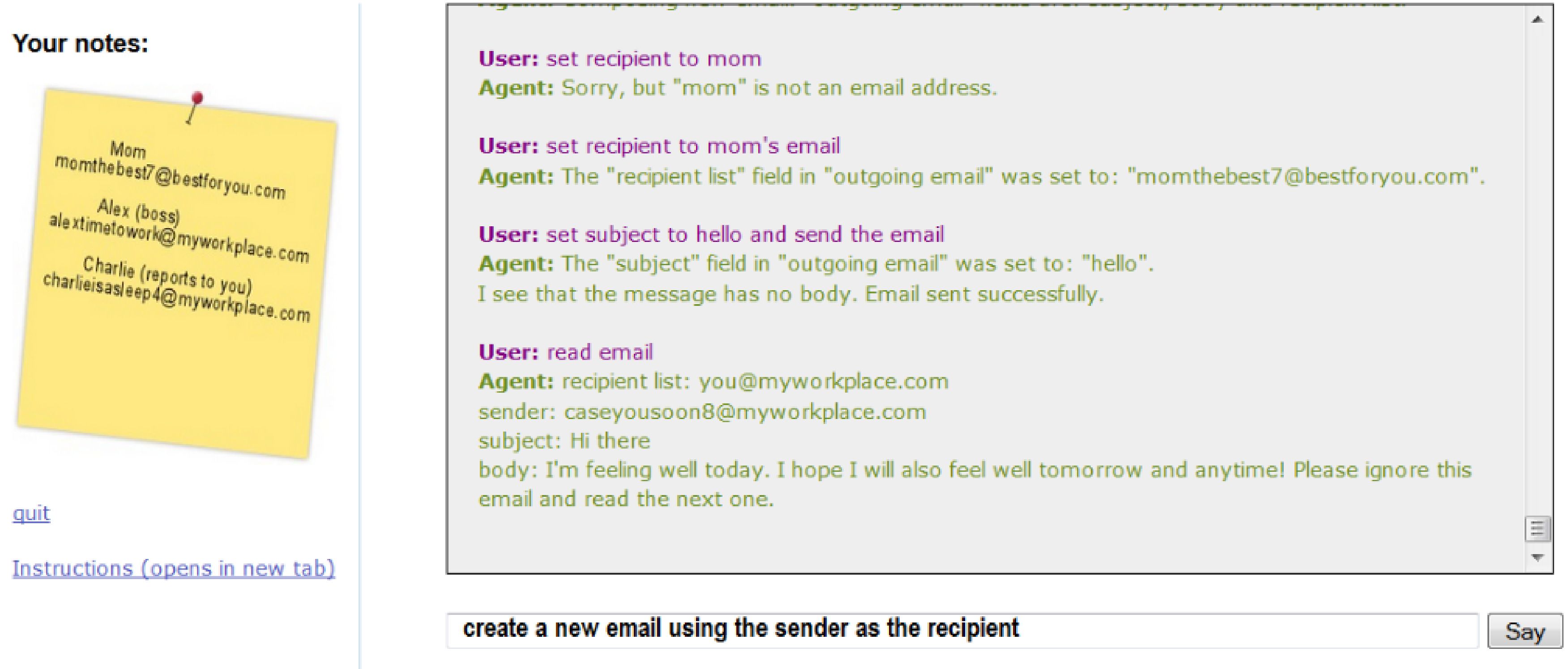


Limited to structured tasks through structured apps

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY
DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

Verbal instruction *(Azaria et al., AAAI 2016)*

- E.g., teach the agent how to send emails via solely natural language interaction



Rules behind each task are complicated and hard to generalize them

Summary

Enhancement	Techniques	Pros	Cons
Multi-chatbot Design	Reinforcement learning Data-driven re-ranker models	Can handle cross-domain requests; Makes use of existing chatbots	Does not learn new skills outside the capability of the ensemble; Fails in out-of-domain requests
Human-chatbot Collaboration	CoChat: external memory + HRNN Evorus: crowd-powered, automates itself over time	Can learn new skills from human workers; More robust and able to handle complex requests	Expensive; Not consistent; Long delays
Learning from Users	Programming by demonstration Verbal instruction	Can learn unknown commands from users; Friendly to novices	Limited to narrow domains; Need to design hand-crafted rules

Outline

1. Introduction
2. Core Design Philosophy
3. Enhancement of Chatbot
4. Future Directions

Future Directions

1. Response Generation with Transformer and Contextual Embedding

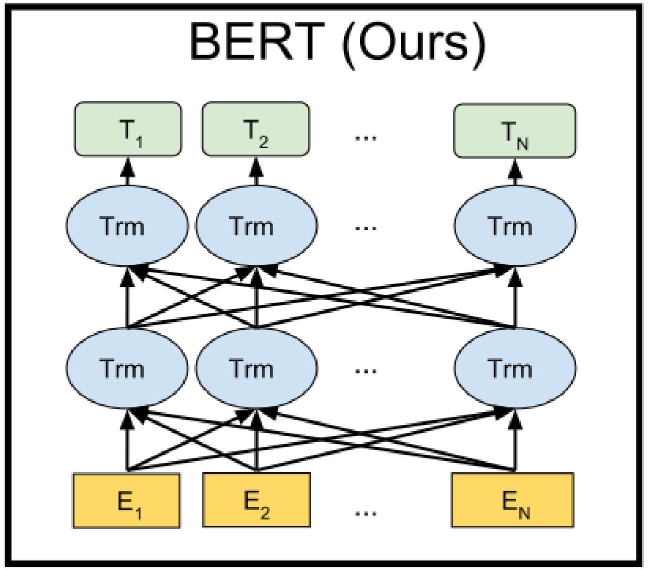
- State-of-the-art chatbots still suffer from out-of-context, universal response generation
- Possible main reasons
  - RNN-based models largely **ignore the global information of the whole sentence**, since a word could have different relations with all the words in the sentence. Use Transformer *(Vaswani et al., NIPS 2017)*
  - Word representations (e.g., word2vec, Glove) are context-free, e.g., "blue" always has the same vector in "blue color" and "blue mood".
    Use Contextual Embedding

(Devlin et al., CoRR 2018); (Radford et al., 2018); (Peters et al., NAACL 2018)

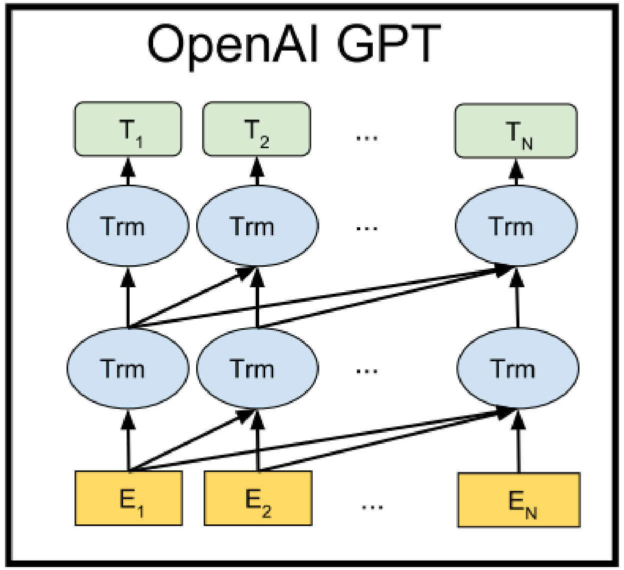
Note: some detail in backup slides

Contextual Embedding

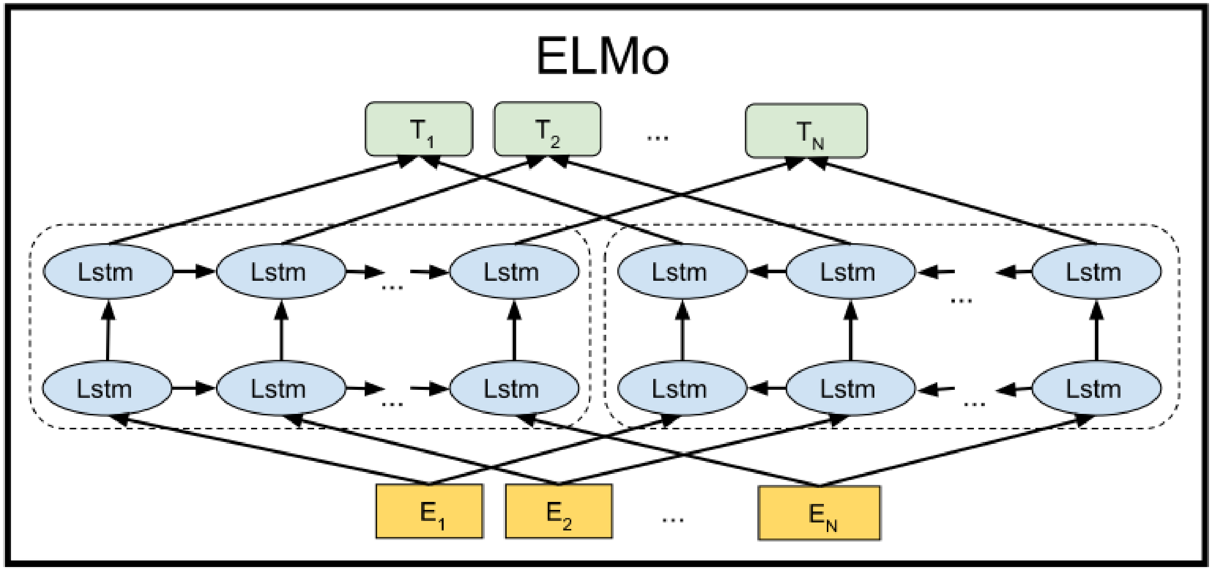
- Represents each word in the context of the sentences through some pre-training models
- Continuously refreshes the state-of-the-art results in NLP tasks like sentence pair classification tasks, and SQuAD v1.1 question answering tasks



(Devlin et al., CoRR 2018)



(Radford et al., 2018)



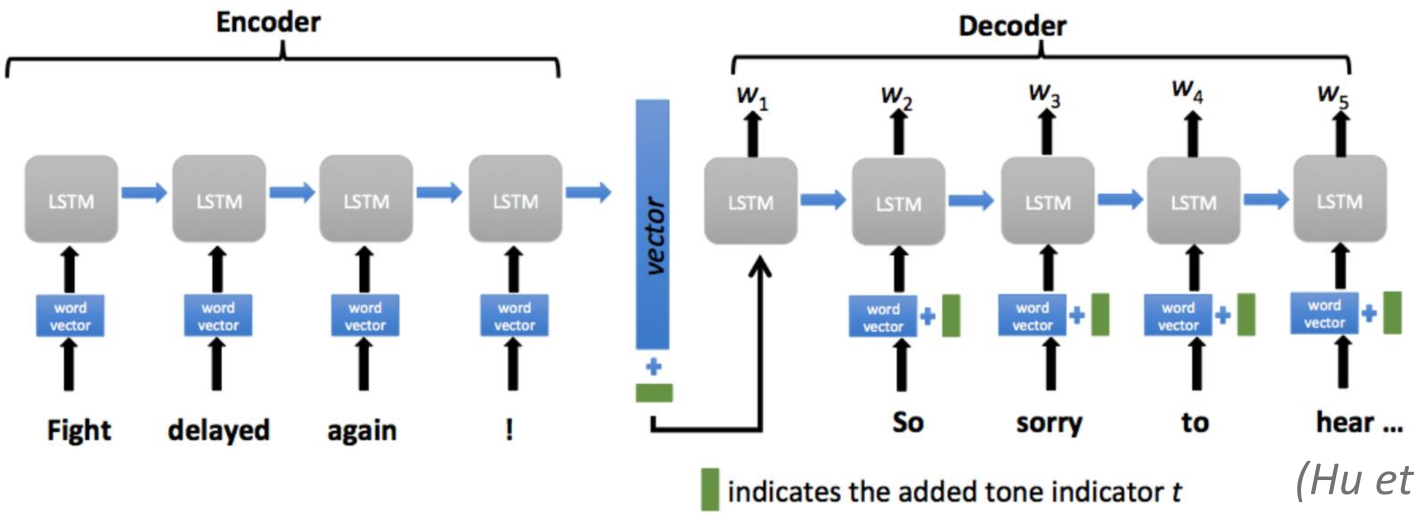
(Peters et al., NAACL 2018)

Future Directions

2. Building Chatbots with Personality

- Merge personality into the response generation models *(Li et al., ACL 2016); (Hu et al., CHI 2018)*

A satisfactory response
- Appropriate
- Helpful
- **Tone-aware: conveys certain feelings**



(Hu et al., CHI 2018)

Request	Your website is down and I really need to buy chair! ! ! No joke.
Passionate	Hi there! We are experiencing technical issues with our website. We are working to resolve this.
Empathetic	Hi I apologize for the inconvenience. Can you please dm your account or phone number?
Neutral	Hi there. Can you dm us your number so we can look into this for you?
Human	Hi Inga we were aware that the website was down and we apologize for this. It is now up and running again.

Future Directions

1. Response Generation with Transformer and Contextual Embedding

- Use the Transformer to replace the RNN, LSTM and GRU in previous retrieval-based and generation-based models
- **Use the contextual word embeddings as the new representation for the requests and responses**

2. Building Chatbots with Personality

- Evaluate how different chatbot personalities could affect the end user engagement
- **Train a service chatbot that not only has a consistent personality, but can also sense user's emotional status and adjust its speaking style**

Conclusion

- Using chatbots to assist or replace human workers is a trend in the service industries
- From the NLP aspect
  - Explore better response generation models
- From the HCI aspect
  - Explore possible usage cases of service chatbots, exploit suitable methods to design them, and evaluate them through user studies
  - Explore better methods for human-chatbot collaboration and for learning from users

Thank you!

Presenter: Zhenhui PENG
zpengab@connect.ust.hk
2019/01/16

Reference

Please refer the full survey paper in
https://penguinzhou.github.io/Chatbot_survey.pdf

Backup Materials

Personal Motivation

- Curent Research Interests
  - Chatbot applied in various domains to benefit the users
  - Dialogue strategy between human-robot interaction or human-chatbot interaction
  - Human-in-the-loop conversational system design
- Managing some knowledge in NLP is needed and conducting a survey on construction methods of chatbot design could be beneficial.
- Research conducted:
  - **Zhenhui Peng**, Yunhwan Kwon, Jiaan Lu, Ziming Wu, and Xiaojuan Ma. 2019. **Design and Evaluation of Service Robot's Proactivity in Decision Making Support Process**.In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. (Conditionally accepted)

Transformer *(Vaswani et al., NIPS 2017)*

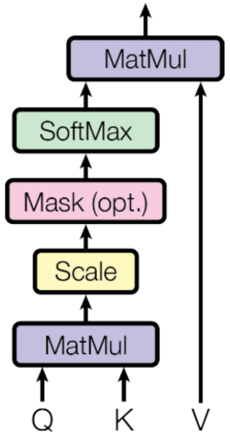
$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

Q: queries
K: keys
V: values

Self-attention

➢ Set Q=K=V= the set of words X in the sentence

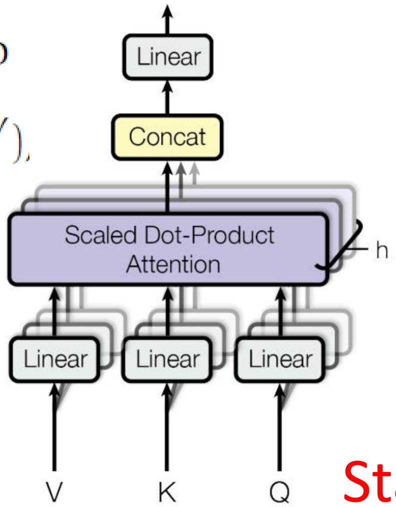
Scaled Dot-Product Attention

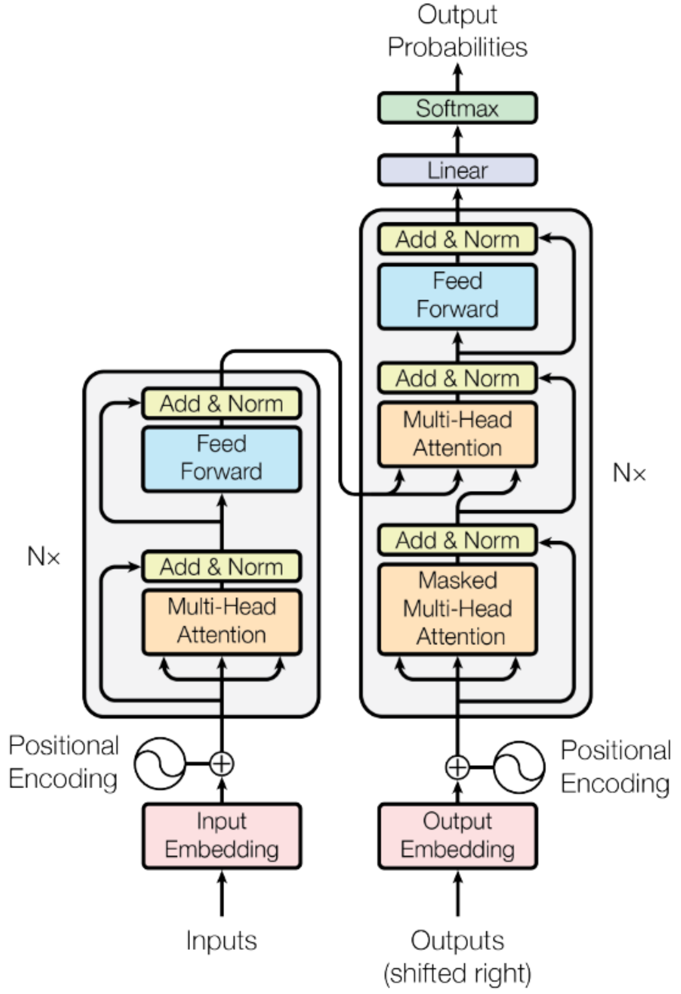


$$\mathrm{MultiHead}(Q, K, V) = \mathrm{Concat}(\mathrm{head}_1, ..., \mathrm{head}_h)W^O$$
$$\text{where where } \mathrm{head}_i = \mathrm{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

➢ Like CNN, jointly attend to information from different representation subspaces at different positions

Multi-Head Attention





State-of-the-art in machine translation tasks

2. Building Chatbot with Personality

- Still hard to design a chatbot that has consistent personality, e.g.,
  - Have a persona (e.g., background facts or profile)
  - Have a certain speaking style (e.g., toned responses)

- Personality has significant impacts on user experience *(Li et al. ACL 2016); (Hu et al. CHI 2018)*

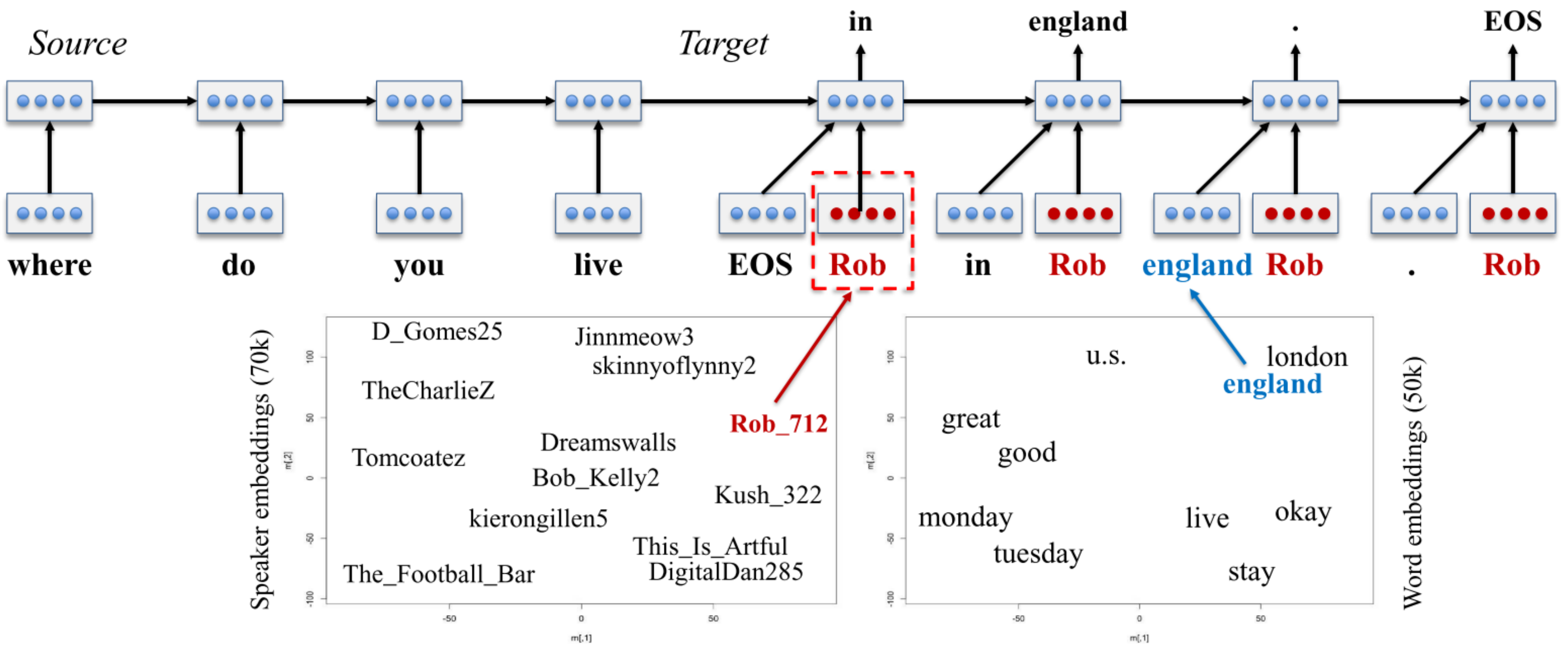
- Rule-based: almost impossible when chatbots become complicate
- Retrieval-based and Generation-based
  - Build a highly consistent dataset, but nearly impossible.
  - **Possible: merge personality into the response generation models**

(Li et al. ACL 2016); (Hu et al. CHI 2018)

Building Chatbot with Personality

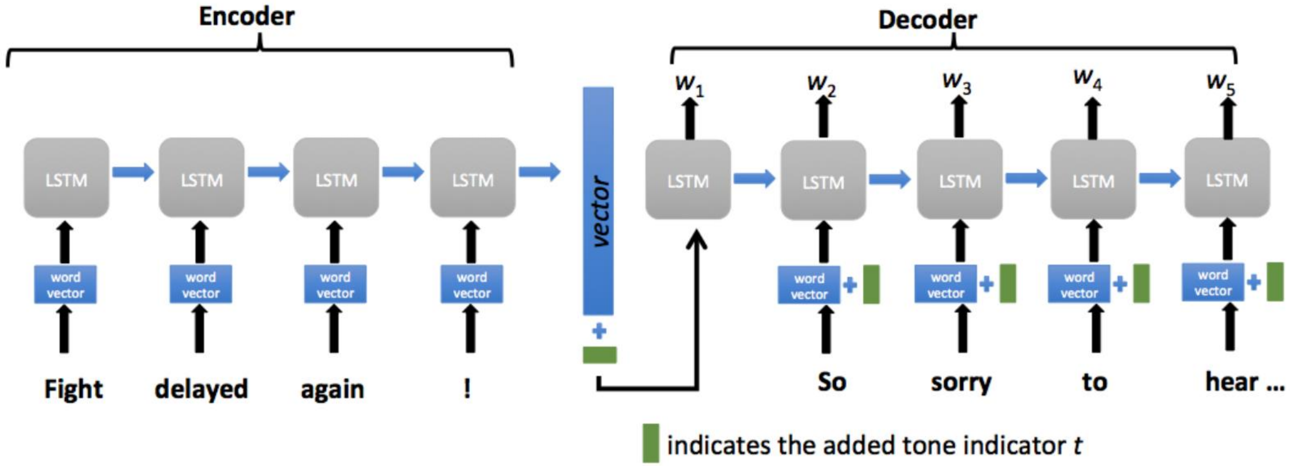
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY
DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

- Concatenate additional speaker embeddings with the word embeddings in the decoder of the standard Seq2Seq model *(Li et al., ACL 2016)*



Building Chatbot with Personality

- Tone-aware Seq2Seq models *(Hu et al., CHI 2018)*
  - Annotation, linear regression
  - Eight major tones, two beneficial tones
  - Keywords as the indicator



Request	Your website is down and I really need to buy chair! ! ! No joke.
Passionate	Hi there! We are experiencing technical issues with our website. We are working to resolve this.
Empathetic	Hi I apologize for the inconvenience. Can you please dm your account or phone number?
Neutral	Hi there. Can you dm us your number so we can look into this for you?
Human	Hi Inga we were aware that the website was down and we apologize for this. It is now up and running again.