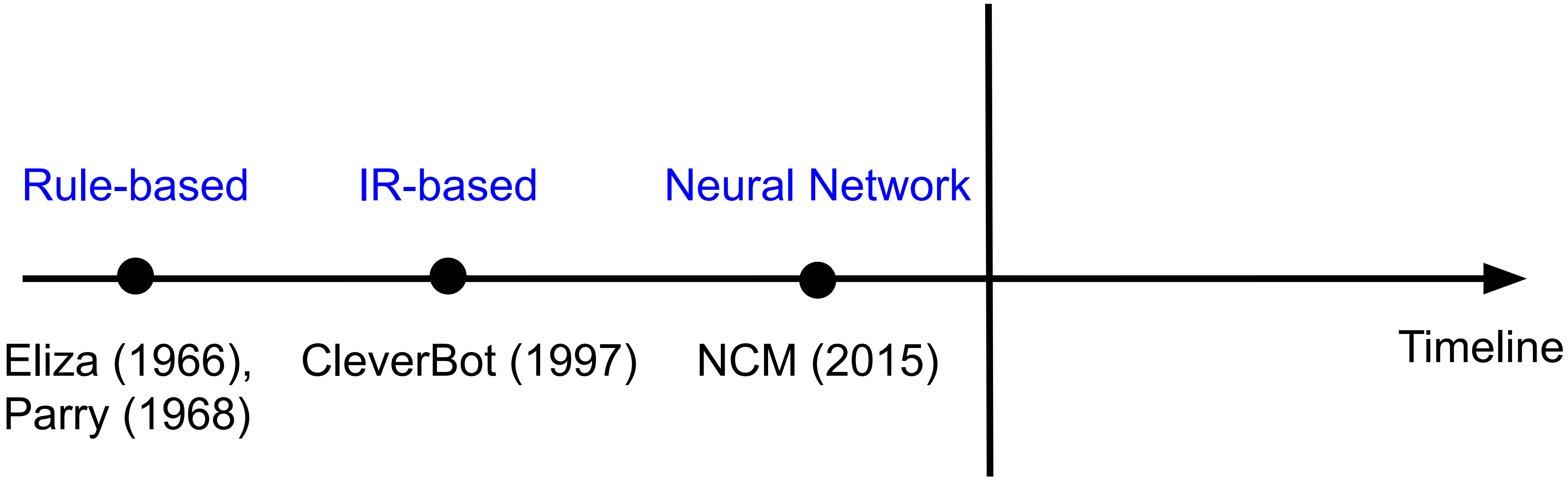
Dialogue II

Zhenyu Song, Jace Lu

Apr. 2 2020

Recap: Dialogue Agents



Overview

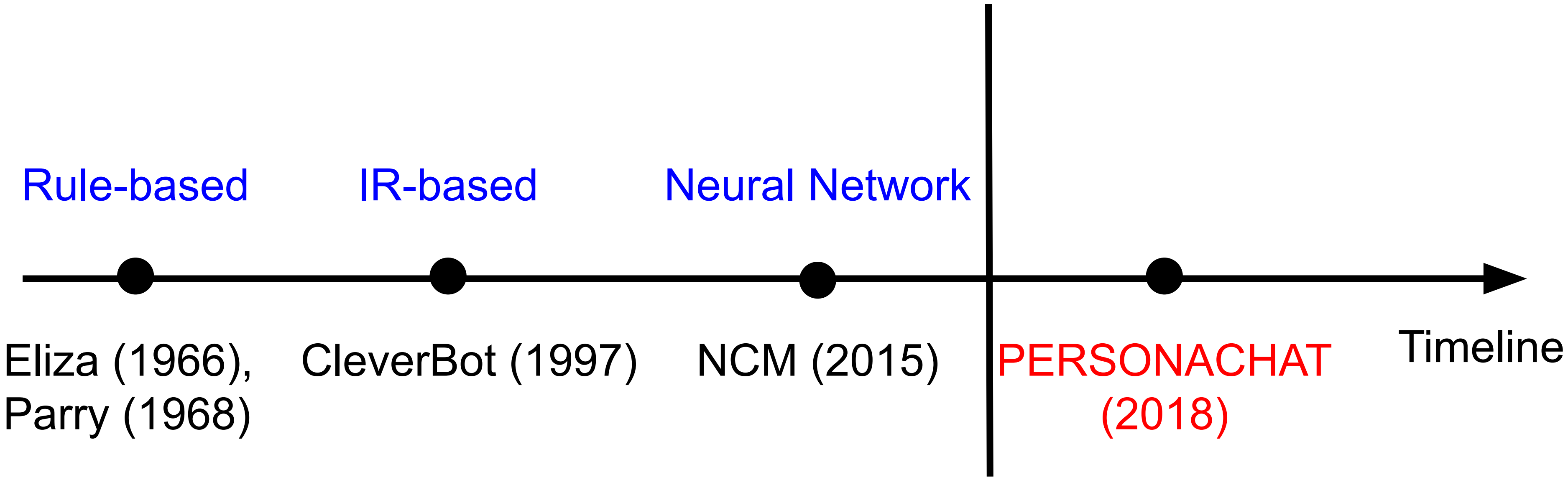
Personalizing Dialogue Agents: I have a dog, do you have pets too? [Zhang et al., 2018]

- Dataset with consistent personalities & evaluate different models

What makes a good conversation? How controllable attributes affect human judgments. [See et al., 2019]

- Evaluate different controllable attributes using the dataset above

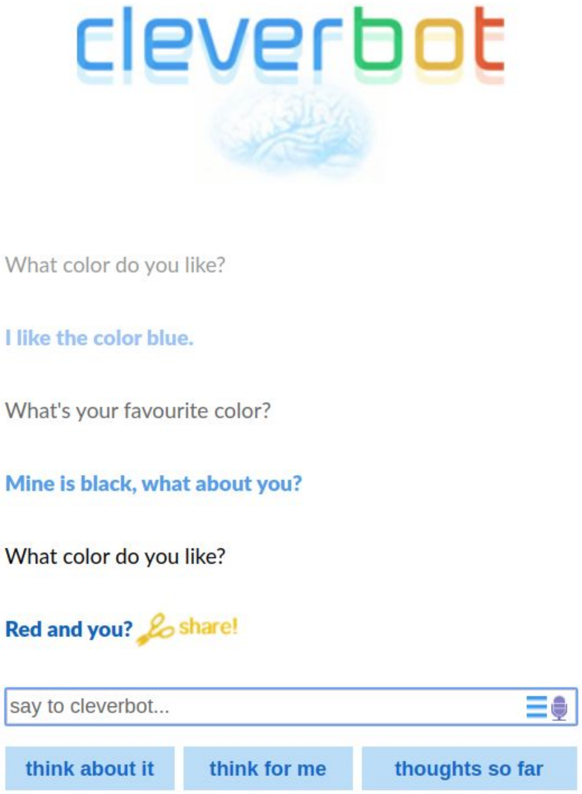
Recap: Dialogue Agents



Problems with Chit-Chat Agents

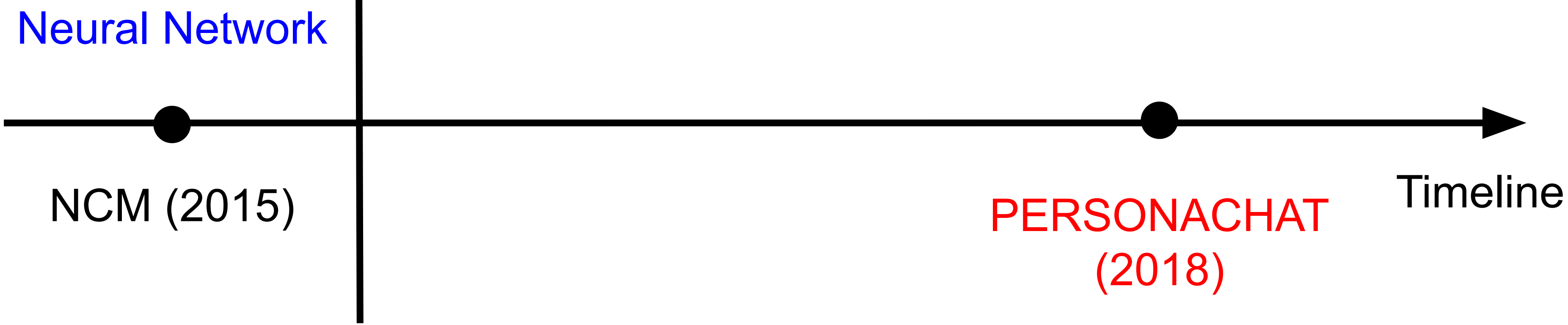
Neural Network

NCM (2015)



Problems with Chit-Chat Agents

- **Lack of a consistent personality**
- Lack of long-term memory
- Tendency to produce non-specific answer



Why Consistent Personality?

- For some applications, we don’t care

Calendar.help

WIKIPEDIA
The Free Encyclopedia

- For some applications. we do care

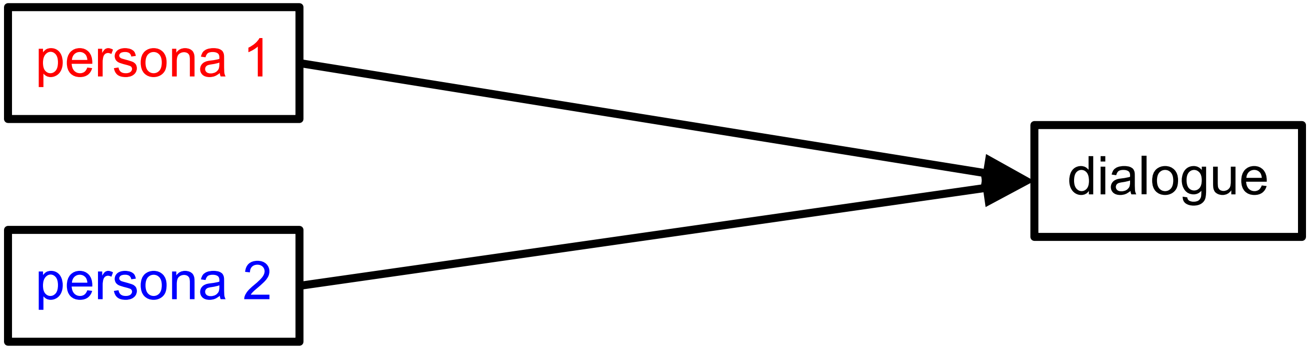
Why Lack of a Consistent Personality?

- Previous training dataset includes many dialogs each with different speakers
- There is no speaker information

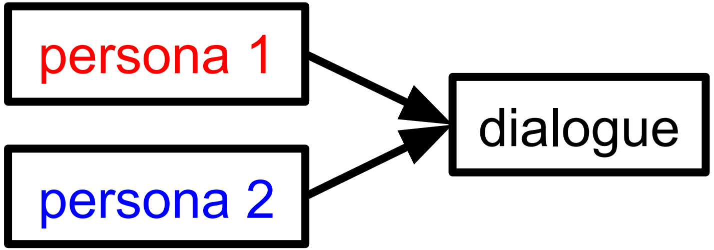
Contributions

- **Built PERSONA-CHAT dataset: endow each agent with explicit persona**
- Evaluate different models on PERSONA-CHAT dataset
  - New model: generative profile memory network

PERSONA-CHAT Dataset



PERSONA-CHAT Dataset



Persona 1	Persona 2
I like to ski	I am an artist
My wife does not like me anymore	I have four children
I have went to Mexico 4 times this year	I recently got a cat
I hate Mexican food	I enjoy walking for exercise
I like to eat cheetos	I love watching Game of Thrones

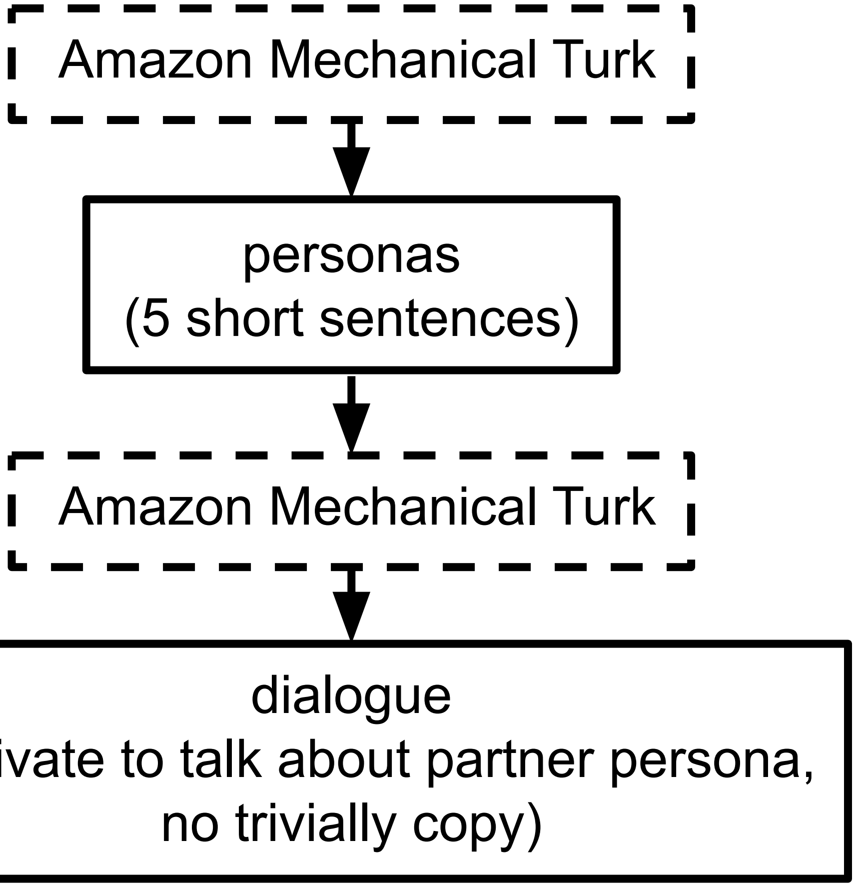
[PERSON 1:] Hi
[PERSON 2:] Hello ! How are you today ?
[PERSON 1:] I am good thank you , how are you.
[PERSON 2:] Great, thanks ! My children and I were just about to watch Game of Thrones.
[PERSON 1:] Nice ! How old are your children?
[PERSON 2:] I have four that range in age from 10 to 21. You?
[PERSON 1:] I do not have children at the moment.
[PERSON 2:] That just means you get to keep all the popcorn for yourself.
[PERSON 1:] And Cheetos at the moment!
[PERSON 2:] Good choice. Do you watch Game of Thrones?
[PERSON 1:] No, I do not have much time for TV.
[PERSON 2:] I usually spend my time painting: but, I love the show.

Related Work

A Persona-Based Neural Conversation Model. [Li et al., 2016]

- Twitter data
- Distributed embeddings, one per speaker vs explicit profile information
- Does not focus on attempting to engage the other speaker by getting to know them

PERSONA-CHAT Collecting Detail



PERSONA-CHAT Statistics

- 1155 personas
- 10,981 dialogs (~19 dialogs per persona)
- 164,356 utterances (sentences)
- 3–5 persona sentences per dialog
- 6–8 chat turns per dialog

Evaluation

- Next utterance prediction -- given the dialogue history
- Four scenarios, where model conditions on
  - No persona
  - Self persona
  - The other speaker's persona
  - Both personas
- Original persona makes the problem less challenging, as the human tends to repeat persona text
- Solution: rewrite persona sentences.
  - Full eval: 4 scenarios x {origin persona sentence, revised persona}

Revised Personas

Original Persona	Revised Persona
I love the beach.	To me, there is nothing like a day at the seashore.
My dad has a car dealership	My father sales vehicles for a living.
I just got my nails done	I love to pamper myself on a regular basis.
I am on a diet now	I need to lose weight.
Horses are my favorite animal.	I am into equestrian sports.
I play a lot of fantasy videogames.	RPGs are my favorite genre.
I have a computer science degree.	I also went to school to work with technology.
My mother is a medical doctor	The woman who gave birth to me is a physician.
I am very shy.	I am not a social person.
I like to build model spaceships.	I enjoy working with my hands.

Not only rephrases but also includes generalizations and specializations

Evaluation Metrics

- Perplexity
- Hit@1 accuracy among 20 candidate utterances
- F1 score
- Human evaluation

Models

- Ranking models: select response from training set
- Generative models: generate word by word

Ranking Models

- tf-idf BoW based IR baseline
- StarSpace Embedding [Wu et al., 2017]
- Ranking Profile Memory Network
- Key-Value (KV) Profile Memory Network

Tf-idf BoW Based IR Baseline

- Given the query, first find the most similar message in the training dataset
- Similarity is defined by tf-idf weighted cosine similarity between the bags of words
- Output the corresponding response from training set
- Concatenate profile vector to query vector

StarSpace Embedding

- Supervised embedding, learning the similarity between query q and next utterance c: sim(q, c)
- Similarity is defined by the cosine similarity of the sum of word embeddings of the query q and candidate c
- Concatenate profile vector to query vector
- To select candidate c’

$$c' = argmax_c sim(q, c)$$

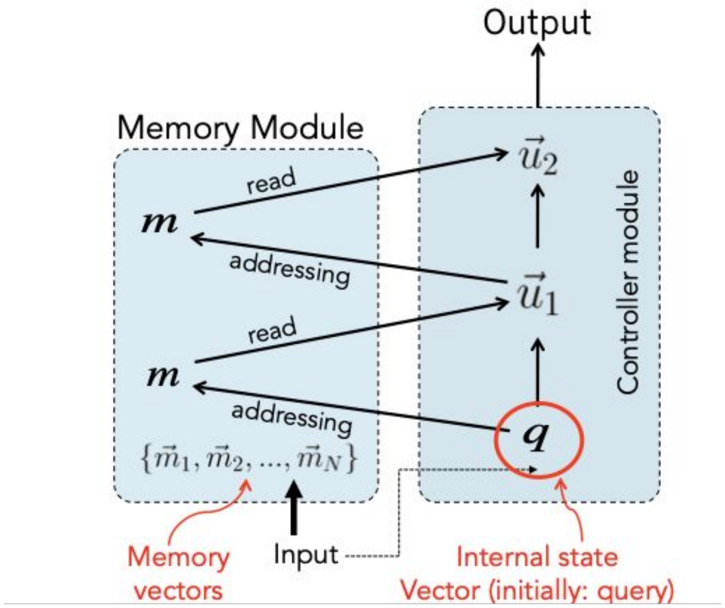
Memory Networks

- Proposed by Jason Weston and others (with many different variants)
  - [Weston et al., 2015] [Sukhbaatar et al., 2015] [Miller et al., 2016]
- Most ML has **limited memory** which is more-or-less all that's needed for "low level" tasks e.g. object detection.
- Long-term memory is required to read a dialog: to remember previous dialog (short- and long-term), and respond

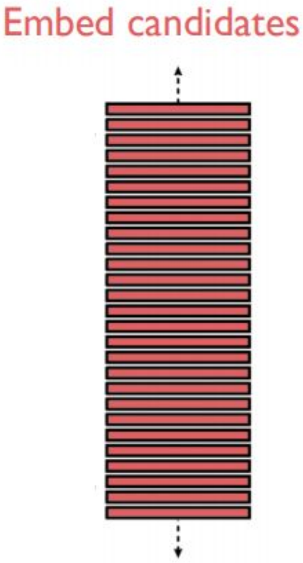
Memory Networks: Example

Consider the follow sequence with a query “Where is the milk now?”

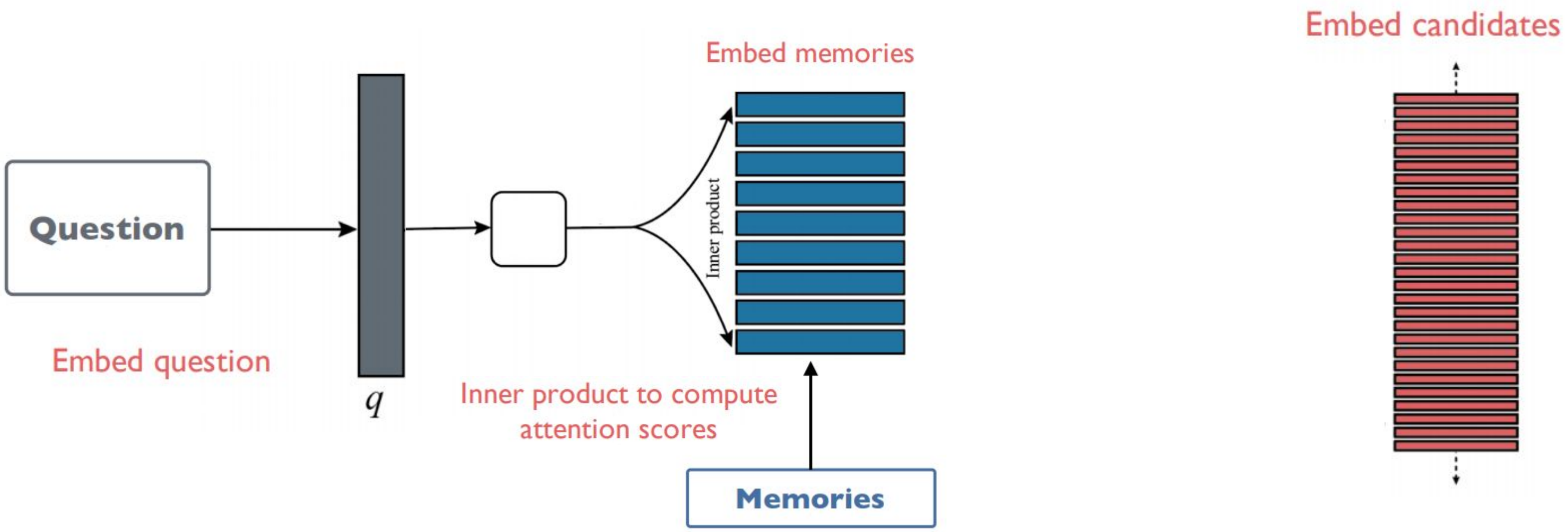
- Joe went to the kitchen.
- Fred went to the kitchen.
- Joe picked up the milk.
- Joe traveled to the office.
- Joe left the milk.
- Joe went to the bathroom.



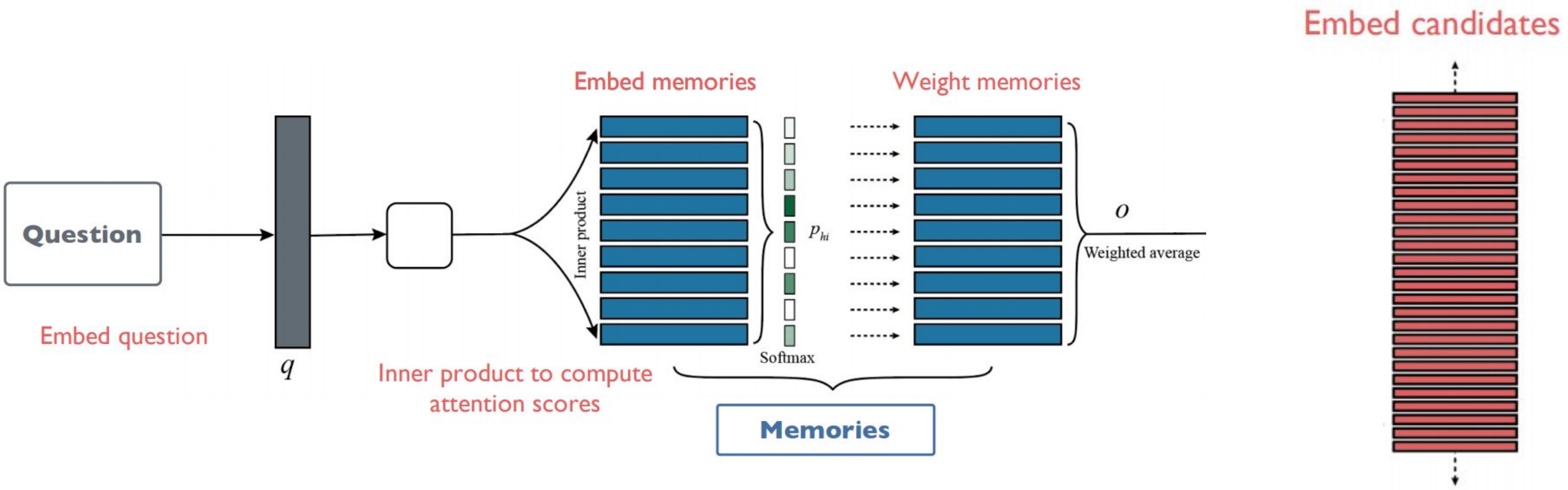
Ranking Profile Memory Network



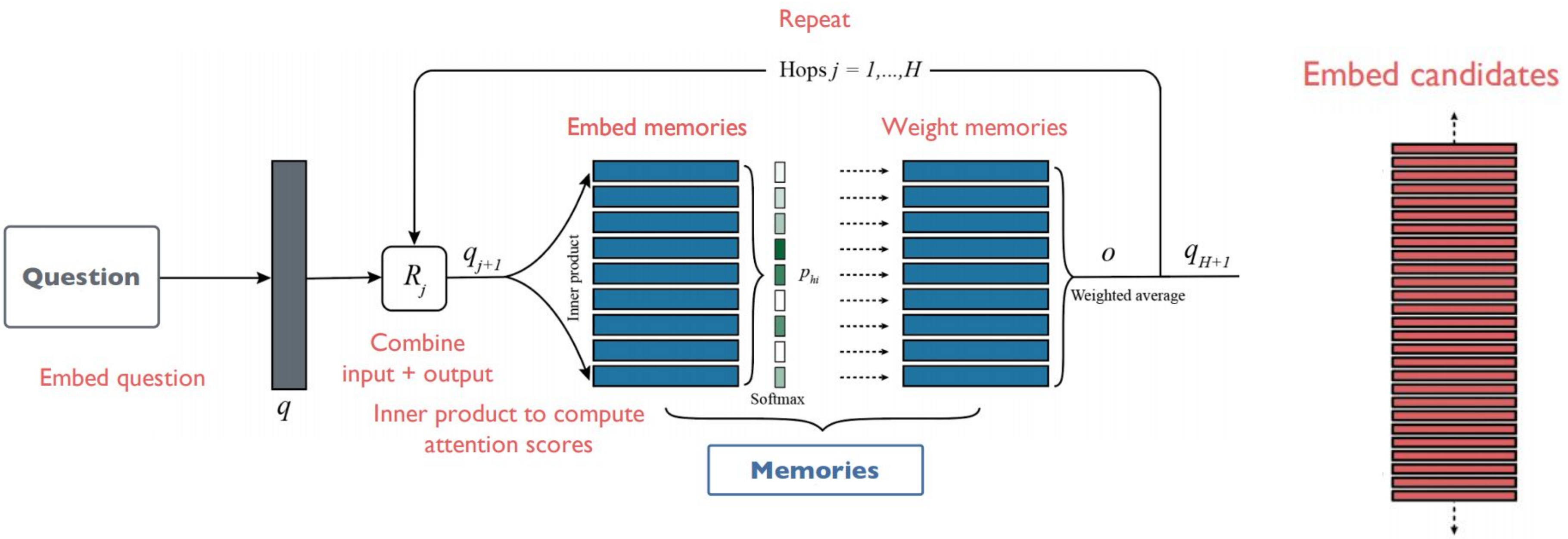
Ranking Profile Memory Network



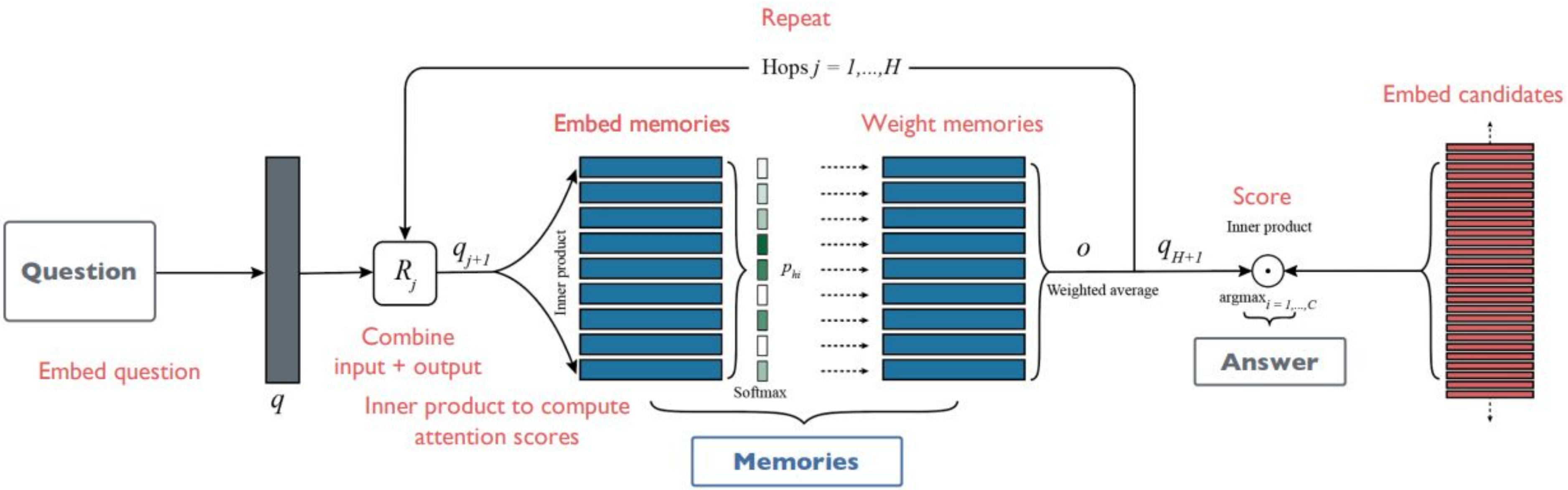
Ranking Profile Memory Network



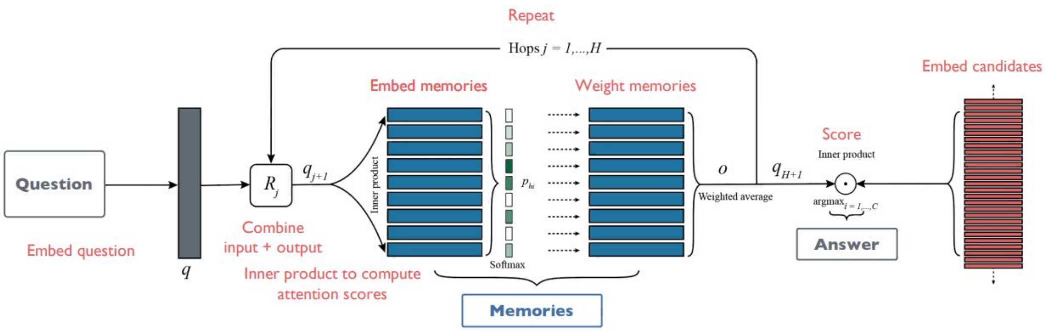
Ranking Profile Memory Network



Ranking Profile Memory Network



Ranking Profile Memory Network



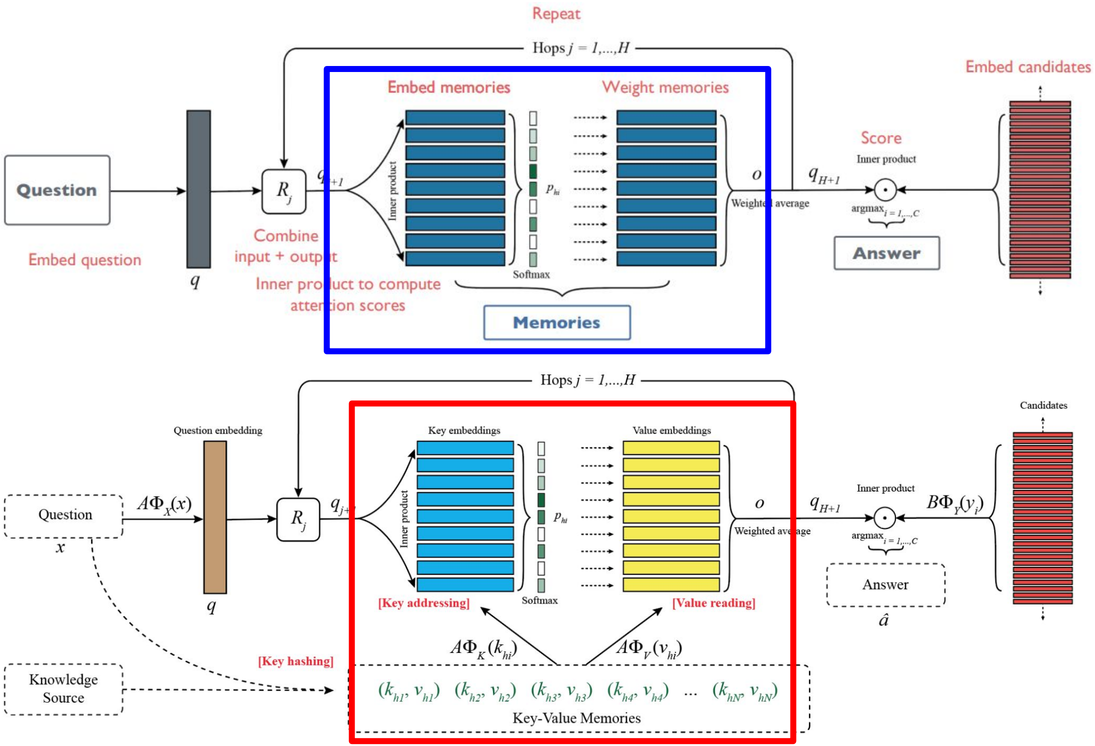
- This paper uses 0 hop. Use Starspace embedding.
- Given profile p, query q, candidate c’

$$q^+ = q + \sum_i s_i p_i \qquad s_i = Softmax(sim(q, p_i))$$

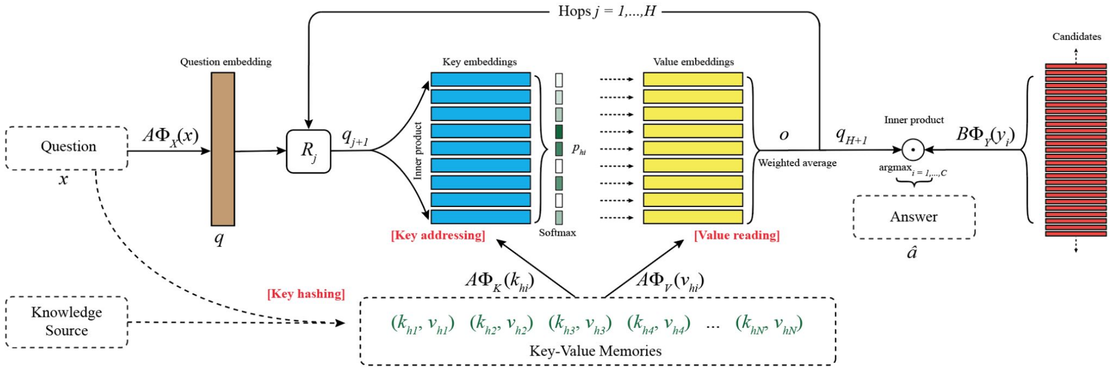
$$c' = argmax_c sim(q^+, c)$$

Key-Value (KV) Profile Memory Network

Use different embeddings to match query and candidates



Key-Value (KV) Profile Memory Network



Question embedding

$$p_{h_i} = \text{Softmax}(A\Phi_X(x) \cdot A\Phi_K(k_{h_i}))$$

$$o = \sum_i p_{h_i} A\Phi_V(v_{h_i})$$

Hops

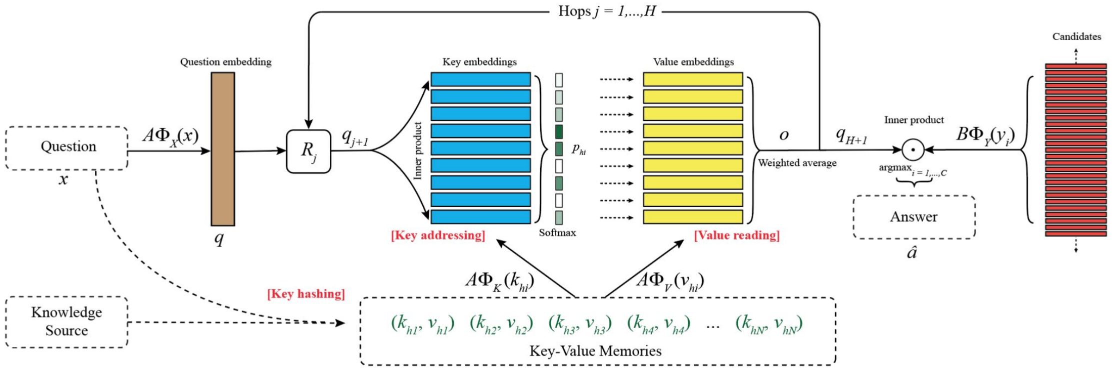
$$p_{h_i} = \text{Softmax}(q_{j+1}^\top A\Phi_K(k_{h_i}))$$

$$\hat{a} = \text{argmax}_{i=1,\ldots,C}\text{Softmax}(q_{H+1}^\top B\Phi_Y(y_i))$$

Φ: feature map

A, B: weight

Key-Value (KV) Profile Memory Network



- This paper uses 0 hop.
- The output from memory network (q^+) as input. The parameters are same as memory network
- (Key, value): (dialog histories, next dialogue utterances) in the training set

Generative Models

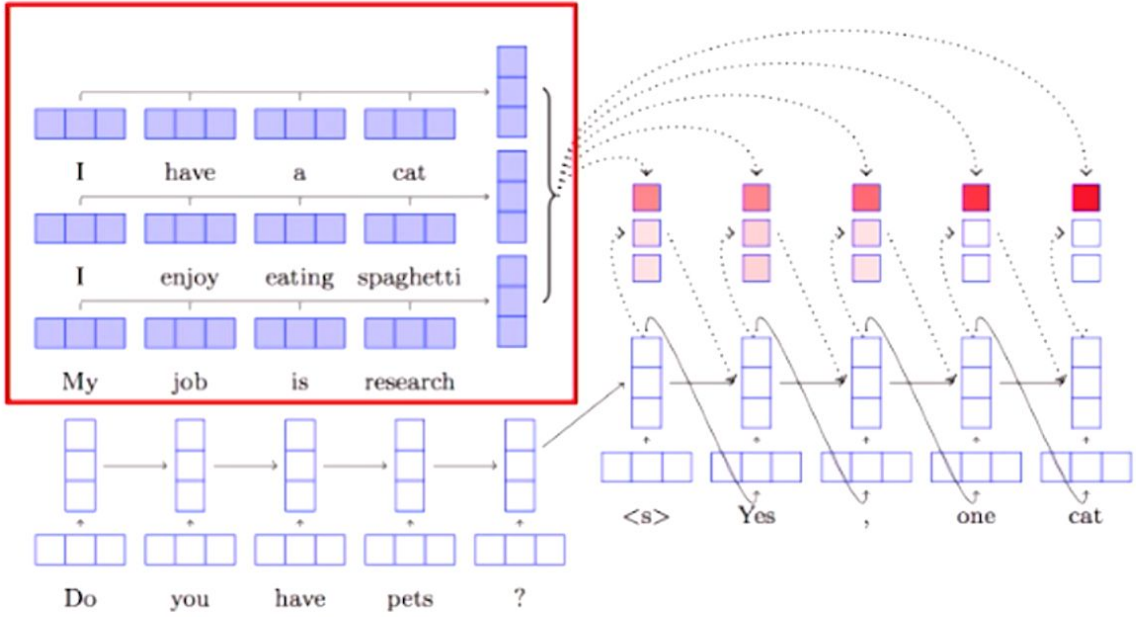
- Seq2Seq
- Generative Profile Memory Network

Seq2Seq

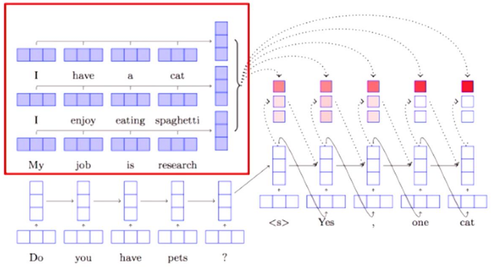
- Classic Seq2Seq model
- Prepend persona to the input sequence

Generative Profile Memory Network

- Modified Seq2Seq, make profile “closely” to output



Generative Profile Memory Network



Memory

$$F_i = \sum_{j}^{|p_i|} \alpha_i p_{i,j}$$

$$\alpha_i = \frac{1}{1+log(1+tf)}$$

Seq2Seq

$$a_t = softmax(FW_a h_t^d),$$
$$c_t = a_t^\intercal F; \;\; \hat{x}_t = tanh(W_c[c_{t-1}, x_t]).$$

Evaluation: Ranking Model

Method	No Persona		Self Persona		Their Persona		Both Personas	
	Orig	Rewrite	Orig	Rewrite	Orig	Rewrite	Orig	Rewrite
IR baseline	0.214	0.214	0.410	0.207	0.181	0.181	0.382	0.188
Training on original personas								
Starspace	0.318	0.318	0.481	0.295	0.245	0.235	0.429	0.258
Profile Memory	0.318	0.318	0.473	0.302	0.283	0.267	0.438	0.266
Training on revised personas								
Starspace	0.318	0.318	0.491	0.322	0.271	0.261	0.432	0.288
Profile Memory	0.318	0.318	0.509	0.354	0.299	0.294	0.467	0.331
KV Profile Memory	0.349	0.349	0.511	0.351	0.291	0.289	0.467	0.330

Table 6: **Evaluation of dialog utterance prediction with ranking models** using hits@1 in four settings: conditioned on the speakers persona (”self persona”), the dialogue partner’s persona (”their persona”), both or none. The personas are either the original source given to Turkers to condition the dialogue, or the rewritten personas that do not have word overlap, explaining the poor performance of IR in that case.

Evaluation: Generative Model

Persona	Method	Original ppl	Original hits@1	Original F1	Revised ppl	Revised hits@1	Revised F1
No Persona		38.08	0.092	0.168	38.08	0.092	0.168
Self Persona	Seq2Seq	40.53	0.084	**0.172**	40.65	0.082	**0.171**
	Profile Memory	**34.54**	**0.125**	**0.172**	38.21	**0.108**	0.170
Their Persona	Seq2Seq	41.48	0.075	0.168	41.95	0.074	0.168
	Profile Memory	36.42	0.105	0.167	**37.75**	0.103	0.167
Both Personas	Seq2Seq	40.14	0.084	0.169	40.53	0.082	0.166
	Profile Memory	35.27	0.115	0.171	38.48	0.106	0.168

Table 5: **Evaluation of dialog utterance prediction with generative models** in four settings: conditioned on the speakers persona ("self persona"), the dialogue partner's persona ("their persona"), both or none. The personas are either the original source given to Turkers to condition the dialogue, or the revised personas that do not have word overlap. In the "no persona" setting, the models are equivalent, so we only report once.

Human Evaluation

- Online Turing test, humans are connected to either humans or models (they don't know which is which)
- Ask humans to give score (1 to 5) on **Fluency, Engagingness** and **Consistency** of the other speaker (turker or model)
- Ask human to detect the other speaker's perona by choosing from two candidates after the conversation

Human Evaluation

Method Model	Profile	Fluency	Engagingness	Consistency	Persona Detection
Human	Self	4.31(1.07)	4.25(1.06)	4.36(0.92)	0.95(0.22)
Generative PersonaChat Models					
Seq2Seq	None	3.17(1.10)	3.18(1.41)	2.98(1.45)	0.51(0.50)
Profile Memory	Self	3.08(1.40)	3.13(1.39)	3.14(1.26)	0.72(0.45)
Ranking PersonaChat Models					
KV Memory	None	3.81(1.14)	3.88(0.98)	3.36(1.37)	0.59(0.49)
KV Profile Memory	Self	3.97(0.94)	3.50(1.17)	3.44(1.30)	0.81(0.39)
Twitter LM	None	3.21(1.54)	1.75(1.04)	1.95(1.22)	0.57(0.50)
OpenSubtitles 2018 LM	None	2.85(1.46)	2.13(1.07)	2.15(1.08)	0.35(0.48)
OpenSubtitles 2009 LM	None	2.25(1.37)	2.12(1.33)	1.96(1.22)	0.38(0.49)
OpenSubtitles 2009 KV Memory	None	2.14(1.20)	2.22(1.22)	2.06(1.29)	0.42(0.49)

Table 4: **Human Evaluation** of various PERSONA-CHAT models, along with a comparison to human performance, and Twitter and OpenSubtitles based models (last 4 rows), standard deviation in parenthesis.

Examples: Seq2Seq

Persona 1	**Persona 2 (Not used by model)**
I am afraid of dogs	I am a carpenter
I used to work at a carnival	I used to drink but I stopped five years ago
I like to drink scotch to relax	I have a wife and three kids
I write short stories for a living	I like playing poker
I like dark superhero movies	I dont have many friends

[PERSON 1:] Hi are you relaxing on this fine evening?
[PERSON 2:] I am doing good and you?
[PERSON 1:] I am well, just drinking some scotch and working on my writing
[PERSON 2:] I love to cook and bake but I rarely make time for it
[PERSON 1:] what do you usually bake?
[PERSON 2:] I do not cook, I bake cakes at a local restaurant. how about you?
[PERSON 1:] I used to be a juggler at a carnival but now I am an author
[PERSON 2:] I am a dancer, what do you do?
[PERSON 1:] I just told you. Do you have any pets?
[PERSON 2:] no, I am allergic to pets unfortunately.
[PERSON 1:] I understand. I am not allergic but I am terrified of dogs!
[PERSON 2:] I have a cat. she has been with me through three wives, haha.
[PERSON 1:] Funny that you just told me you have no pets
[PERSON 2:] I do not have any animals.

Table 8: Example dialog between a human (Person 1) and the Seq2Seq model (Person 2).

Examples: Generative Profile Memory Network

Persona 1	**Persona 2**
I dont like exercise or physical activity	My favorite type of food is Italian
My favorite food is ice cream	I have a big pet dog
I am a homebody	I want to travel to Europe
I have a husband who I love very much	I love reading books about travel
I love watching movies and TV	I like to walk long distances

[PERSON 1:] Hello, how are you doing today?
[PERSON 2:] I am good. I just got finished walking my dog.
[PERSON 1:] Oh cool! I am just sitting watching a movie.
[PERSON 2:] I have been watching that one before. I love it.
[PERSON 1:] Me too. I don't like going out much so I enjoy movies and TV.
[PERSON 2:] that is a great idea. I am going to read a book on writing.
[PERSON 1:] Oh, are you a writer?
[PERSON 2:] yes, I love to read. I am a great author.
[PERSON 1:] What sorts of topics do you write about?
[PERSON 2:] I could read all of those. I am sure you enjoy reading.
[PERSON 1:] I do, but I prefer movies.
[PERSON 2:] I have not, I can not do that. I have to go to the east coast.
[PERSON 1:] Why can't you watch movies?
[PERSON 2:] I have not. I have always had that done.

Table 10: Example dialog between a human (Person 1) and the Generative Profile Memory Network with Self Persona.

Examples: KV Profile Memory Network

Persona 1	Persona 2
I am engaged to be married	I am a writer
I used to play sports in high school	I write romance novels
My favorite sport is soccer	I live on a faraway island
I am a college graduate	I am an introvert
I love watching sports on TV	I have two cats

[PERSON 1:] I just got done playing soccer. What have you been up to?
[PERSON 2:] I do not have pets. I spend my free time writing. what about you?
[PERSON 1:] Just graduated from college. Looking for work.
[PERSON 2:] yes I graduated, in writing and poetry
[PERSON 1:] Have you had anything published?
[PERSON 2:] I mainly write fiction novels.
[PERSON 1:] Do you like to watch sports?
[PERSON 2:] do you like kings of leon my favorite by them is use somebody
[PERSON 1:] Are you married? I will be married soon.
[PERSON 2:] haha, no time. I have got a novel to finish.
[PERSON 1:] What is your novel about?
[PERSON 2:] it is a love story. I really enjoy writing. do you write or read?
[PERSON 1:] Like to read. Spend most time watching football.
[PERSON 2:] yes I do I am and author for a living

Table 9: Example dialog between a human (Person 1) and the Key-Value Profile Memory Network with Self Persona.

Conclusion

- Explicit personas make agent more **consistent** and **engaging**
- “Lack of long-term memory”, “tendency to produce non-specific answer” are still unsolved questions

ConvAI2 NIPS Competition

- Expanded version of PERSONA-CHAT
- Evaluate dialogue systems
  - Automated metrics
  - Amazon Mechanical Turk
  - ‘Wild’ Live Chat with Volunteers
- More information in the next paper

Question 1

Zhang et al 2018 proposed a dataset called PersonaChat and several ranking/generative models to solve this task. If we look at their experimental results, what are the main findings in terms of 1) ranking vs generative model 2) no persona vs self persona vs their persona vs both personas 3) original persona vs revised personas? Does it make sense to you or not?

Extra Question

- Shall we really build persona text first and then run unnatural conversations?
- Or can we learn persona from natural data for real-world scenarios?
  - E.g., from someone’s Twitter, a book, or Zoom.

What makes a good conversation

How <u>controllable attributes</u> affect human judgements

Natural Language Generation task spectrum

Machine Translation	Sentence Compression	Abstractive Summarization	Story Generation	Chitchat Dialogue

Less open-ended ⟷ More open-ended

Mostly word-level decisions ⟷ Requires high-level decisions

Neural LMs more successful ⟷ **Neural LMs less successful**

Makes errors like **repetition** and **generic response** (under certain decoding algorithms).

Difficulty learning to make high-level decisions.

Natural Language Generation task spectrum

Machine Translation | Sentence Compression | Abstractive Summarization | Story Generation | Chitchat Dialogue

Less open-ended	More open-ended
Mostly word-level decisions	Requires high-level decisions
Neural LMs more successful	**Neural LMs less successful**
Control is less important	**Control is more important**

Control = ability to specify desired attributes of the text at test time.

We can use control to fix errors, and allow us to handle some high-level decisions.

Natural Language Generation task spectrum

Machine Translation | Sentence Compression | Abstractive Summarization | Story Generation | Chitchat Dialogue

Less open-ended ⟷ More open-ended

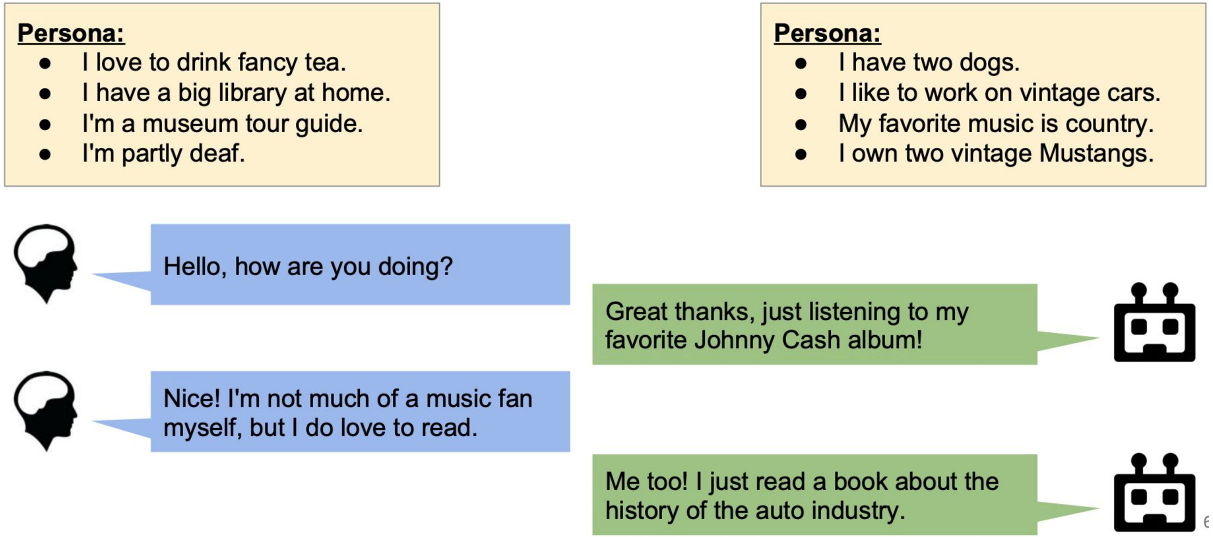
Less open-ended	More open-ended
Mostly word-level decisions	Requires high-level decisions
Neural LMs more successful	**Neural LMs less successful**
Control is less important	**Control is more important**
Eval is difficult	**Eval is fiendish**

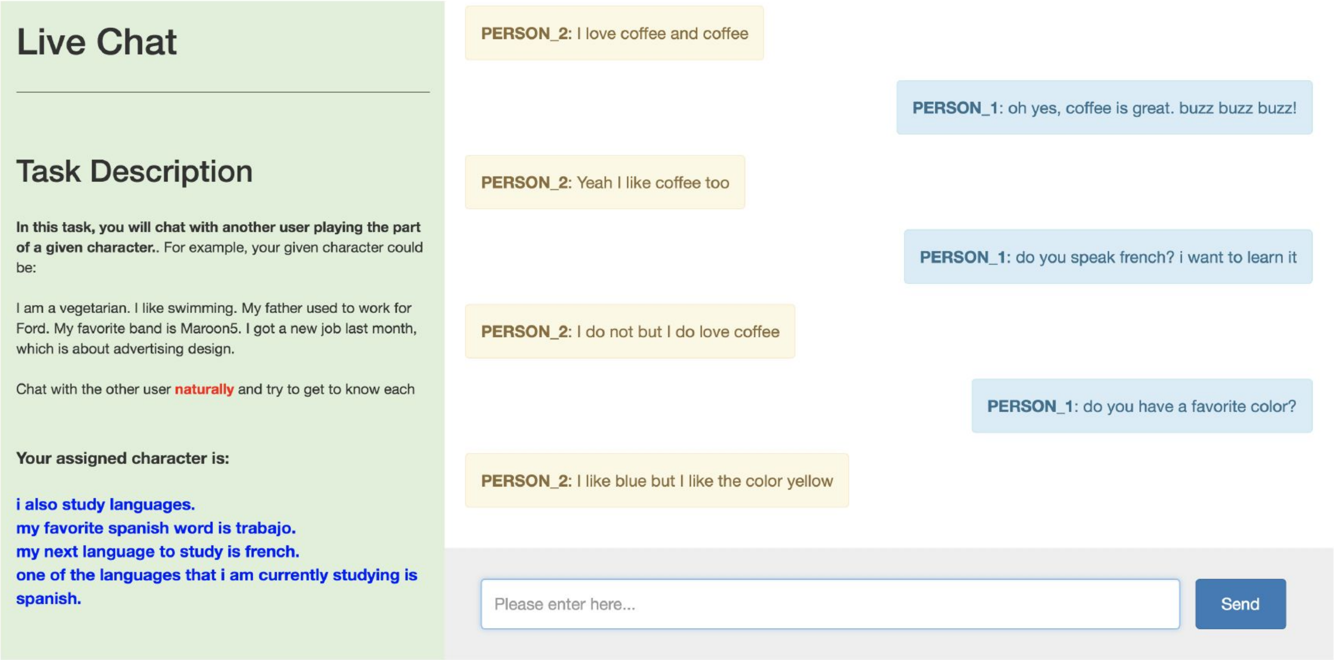
No automatic metric for overall quality.

Dialogue is even more complex:
Single-turn or multi-turn eval?
Interactive or static conversation?

PersonaChat task

The PersonaChat task was the focus of the NeurIPS 2018 ConvAI2 Competition. Then with respect to human judgment via the question “How much did you enjoy talking to this user?” On a scale of 1-4.





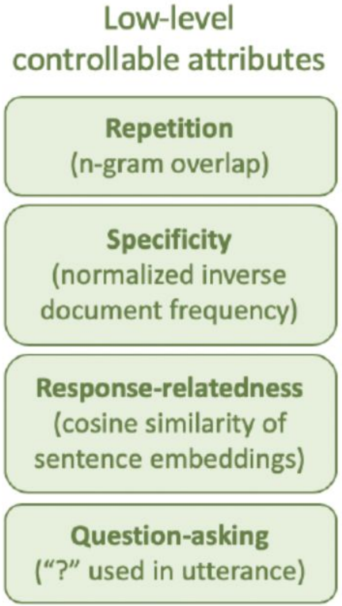
chatUI, talking with the beam search baseline model

Research Question to ask

How effectively can we control the different attributes?

How do the controllable attributes affect human evaluation?

Low-level controllable attributes



Goal

- Reduce repetition (within and across utterances)
- Reduce genericness of responses (e.g. oh that's cool)
- Respond more on-topic; don't ignore user
- Find the optimal rate of question-asking

Effect on human judgments

Human judgment of conversational aspects	measurement
Avoiding Repetition	Does the bot repeat itself
Interestingness	Is the bot Interesting to talk to
Making sense	Does the conversation make sense?
Fluency	Use natural English?
Listening	Good listener?(pay attention to what you say)
Inquisitiveness	Ask enough questions?

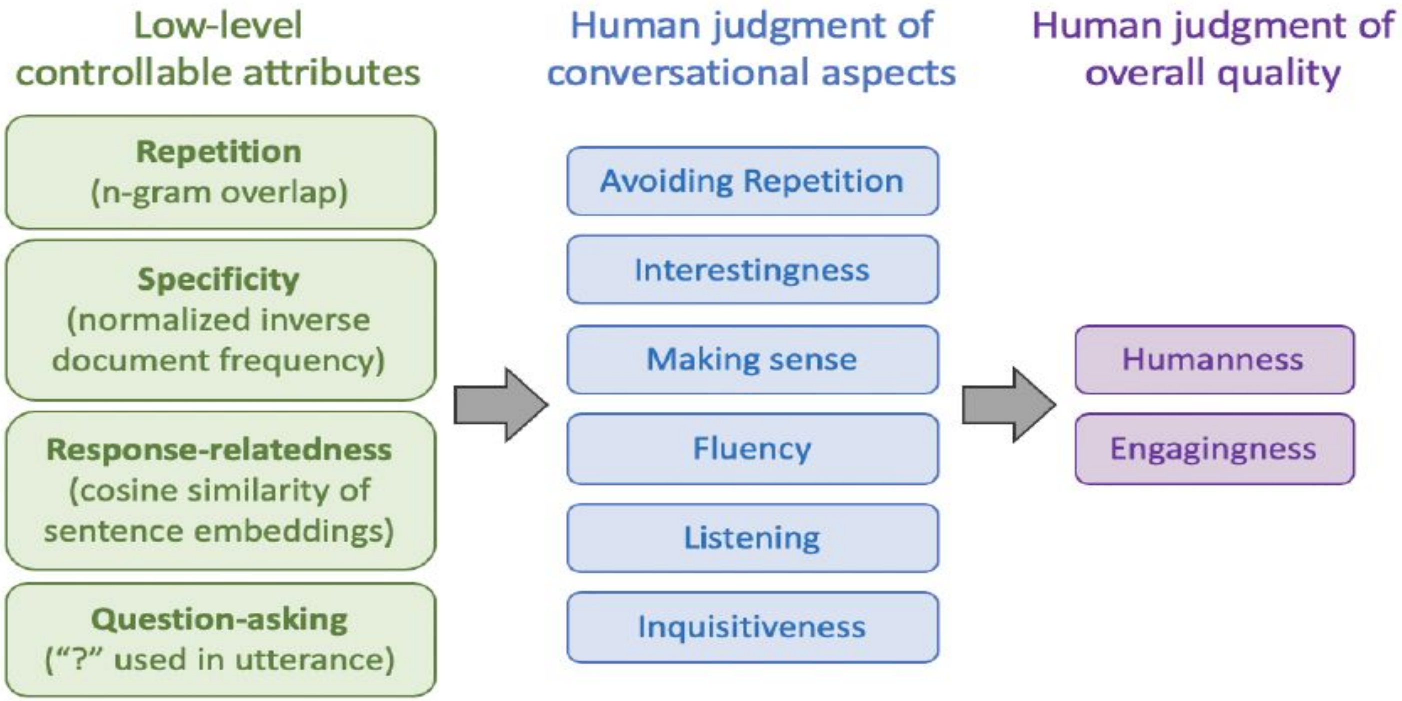
Overall quality of human judgment

measurement

Humanness → Can you tell is it a person or a bot?

Engagingness → Is it enjoyable to talk to?

Overview


Low-level controllable attributes
Repetition (n-gram overlap)
Specificity (normalized inverse document frequency)
Response-relatedness (cosine similarity of sentence embeddings)
Question-asking ("?" used in utterance)
Human judgment of conversational aspects
Avoiding Repetition
Interestingness
Making sense
Fluency
Listening
Inquisitiveness
Human judgment of overall quality
Humanness
Engagingness

Control methods

- **Conditional Training (CT):** CT is a method to learn a sequence-to-sequence model P(y|x,z). Train the model to generate response y, conditioned on the input x, and the desired output attribute z. (Kikuchi et al 2016, Peng et al 2018, Fan et al 2018)

- **Weighted Decoding (WD)**: WD is a decoding method that increases or decreases the probability of words with certain features. During decoding, increase/decrease the probability of generating words w in proportion to features f(w). (Ghazvininejad et al 2017, Baheti et al 2018)

Conditional Training(CT)

First automatically annotate every (x,y) pair in the training set with the attribute we wish to control.

During training, for each example we determine the corresponding z value

Next, the control variable z is represented via an embedding

Lastly, the CT model learns to produce y = y1...yT by optimizing the cross-entropy loss:

$$\text{loss}_{\text{CT}} = -\frac{1}{T}\sum_{t=1}^{T} \log P(y_t|x, z, y_1, \ldots, y_{t-1})$$

Weighted Decoding(WD)

- The technique is applied only at test time, requiring no change to the training method.
- In weighted decoding, on the tth step of decoding, a partial hypothesis

$y_{<t} = y_1, \ldots, y_{t-1}$ is expanded by computing the score for each possible next word w in the vocabulary:

$$\text{score}(w, y_{<t}; x) = \text{score}(y_{<t}; x) + \log P_{\text{RNN}}(w|y_{<t}, x) + \sum_i w_i * f_i(w; y_{<t}, x).$$

Research Question to ask

How effectively can we control the different attributes?

(Pretty well! But some control methods only work for some attributes.)

How do the controllable attributes affect conversational quality aspects?

How efficient can we control attributes

- **Conditional Training (CT):**
  - Requires sufficient training examples for the attribute
  - Ineffective at learning complex relationships between input and output

Not effective for: repetition, response-relatedness
Effective for: specificity and question-asking

- **Weighted Decoding (WD):**
  - Requires attribute to be defined at the word-level.

Not effective for: question-asking
Effective for: repetition, response-relatedness, specificity

Controlling specificity(WD and CT)

	Input: *Yes, I'm studying law at the moment*	
	Baseline Response: *That sounds like a lot of fun!*	
Wt	**NIDF**	**Weighted Decoding Response**
-5.0	0.6%	*Oh......................................*
0.0	17.1%	*That sounds like a lot of fun!*
3.0	18.3%	*That sounds like a lot of fun. How long have you been studying?*
7.0	38.5%	*I majored in practising my spiritual full time philosophy test*
10.0	71.9%	*Oh wow! Merna jean isa paino yi hao hui bu acara sya gila* [...]
z	**NIDF**	**Conditional Training Response**
0	16.8%	*Sounds like you are a great person!*
2	18.3%	*So you are a law student?*
4	18.4%	*That sounds like a lot of fun*
6	22.8%	*That sounds like a rewarding job!*
8	24.4%	*That sounds like a rewarding career!*

More generic ↕ More specific (Weighted Decoding)

More generic ↕ More specific (Conditional Training)

WD: Large range, but degenerate output at the extremes

CT: Smaller range, but generally well-formed output

$$\mathrm{NIDF}(w) = \frac{\mathrm{IDF}(w) - \mathtt{min\_idf}}{\mathtt{max\_idf} - \mathtt{min\_idf}}$$

Controlling response-relatedness (WD)

Input: *Do you go get coffee often*
Baseline Response: *I do, when I am not playing the piano.*

Wt	Sim	Weighted Decoding Response
-10.0	-0.05	*I am a musician.*
0.0	-0.02	*I do, when I am not playing the piano.*
5.0	0.40	*I do, usually at starbucks.*
8.0	0.59	*Not usually, especially when you drink latte.*
11.0	0.72	*Not often, usually with drinks, espresso, latte, tea, etc.*

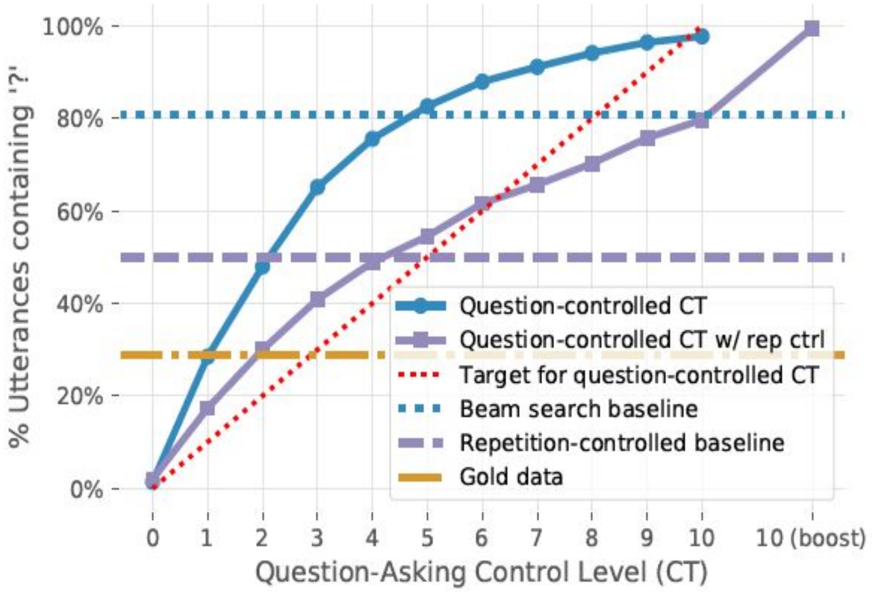
Less related

More related

Output is degenerate when weight is too high

$$\texttt{resp\_rel}(w; y_{<t}, x) = \texttt{cos\_sim}(\texttt{word\_emb}(w), \texttt{sent\_emb}(\ell))$$

Controlling question asking(CT)



Firstly, it allows us to achieve (close to) 0% questions, 100% questions, or anything in between,without introducing the risk of degenerate output.

Secondly, presence-of-a-question-mark captures the true attribute of interest (question asking) more exactly and directly than presence of interrogative words. .

Comparison of control methods

- The primary disadvantage of conditional training is that it sometimes fails to learn the connection between the control variable z and the target output y.
- The primary disadvantage of weighted decoding is that it risks going off-distribution when the weight is too strong

Other considerations:

- Convenience:
- Data availability
- Attribution definition

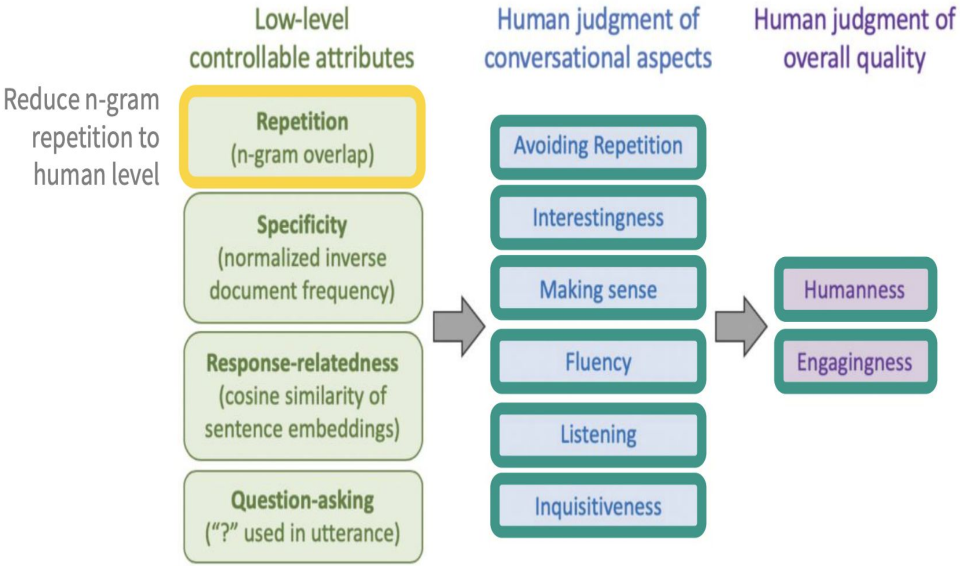
Research Question to ask

How effectively can we control the different attributes?

<u>How do the controllable attributes affect human evaluation?</u>

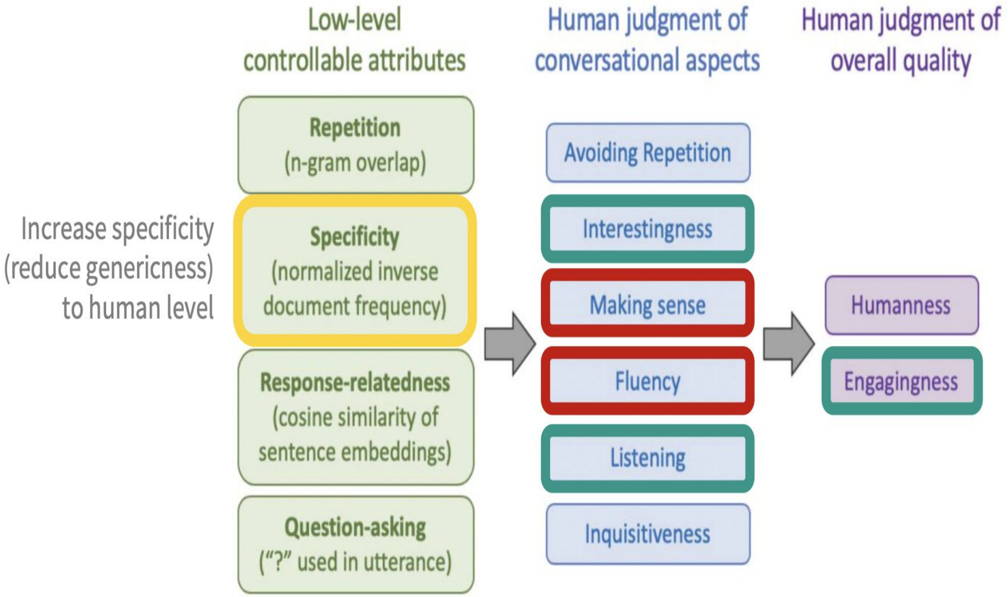
（Strongly – especially controlling repetition, question-asking, and specificity vs genericness）

How does it affect human evaluation



Reducing repetition leads to improvements across all our aspects of conversational quality.

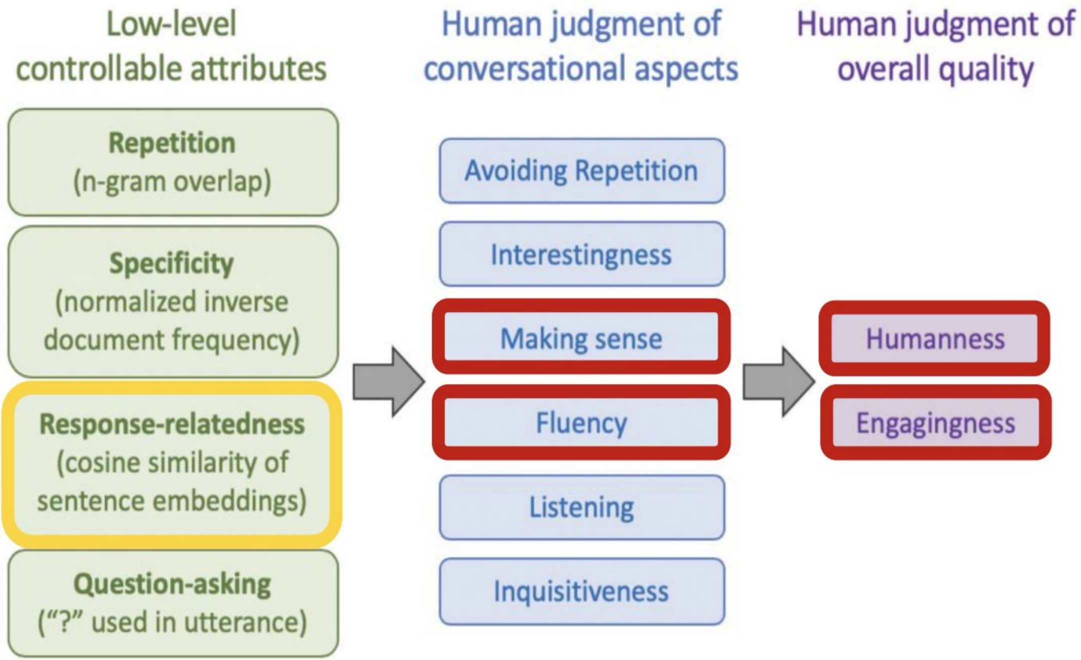
How does it affect human evaluation



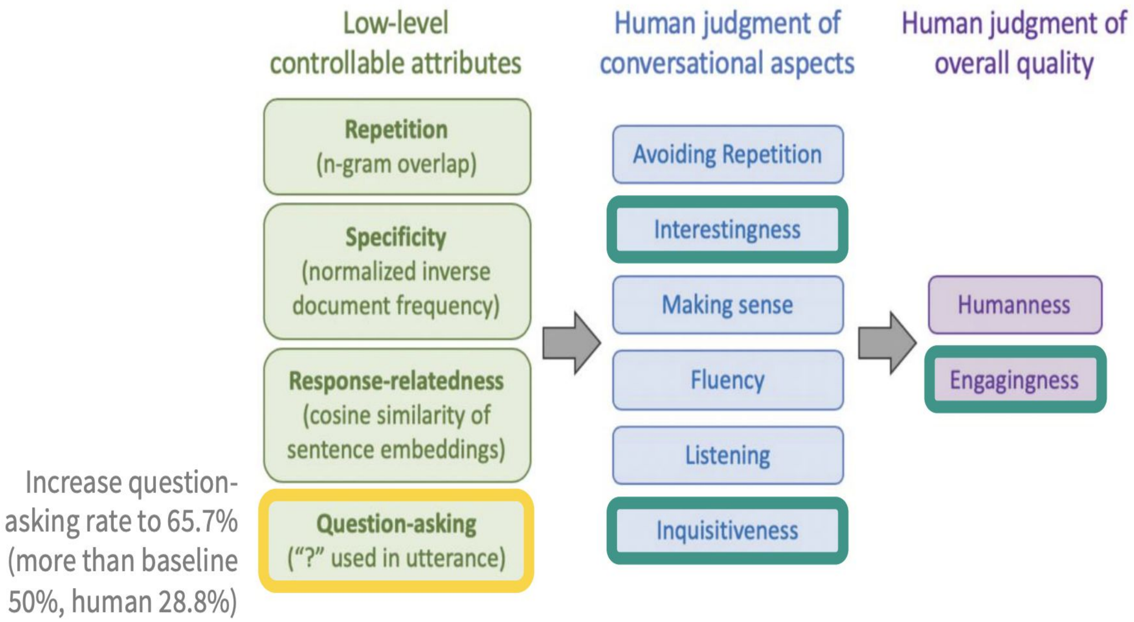
Increasing specificity shows improvements in interestingness and listening ability over the repetition-controlled baseline

Q2: How does control affect human eval?

Increase response-relatedness (similarity to last utterance)

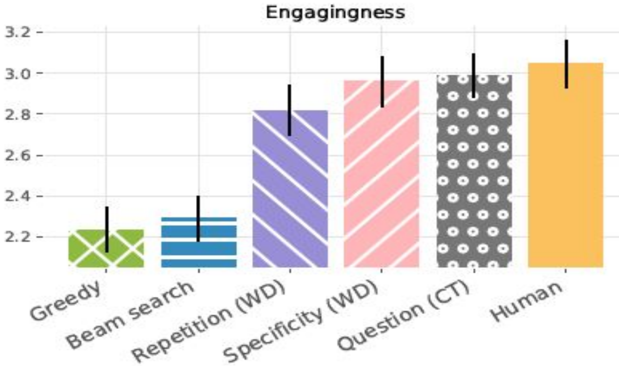


How does it affect human evaluation



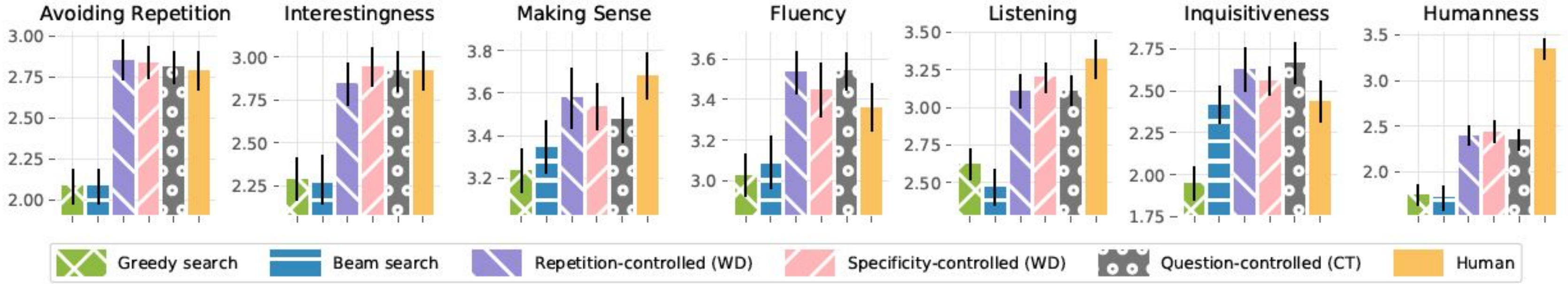
increasing question-asking shows improvements in inquisitiveness and interestingness over the repetition-controlled baseline.

Calibrated human judgments of engagingness for the baselines and best controlled models

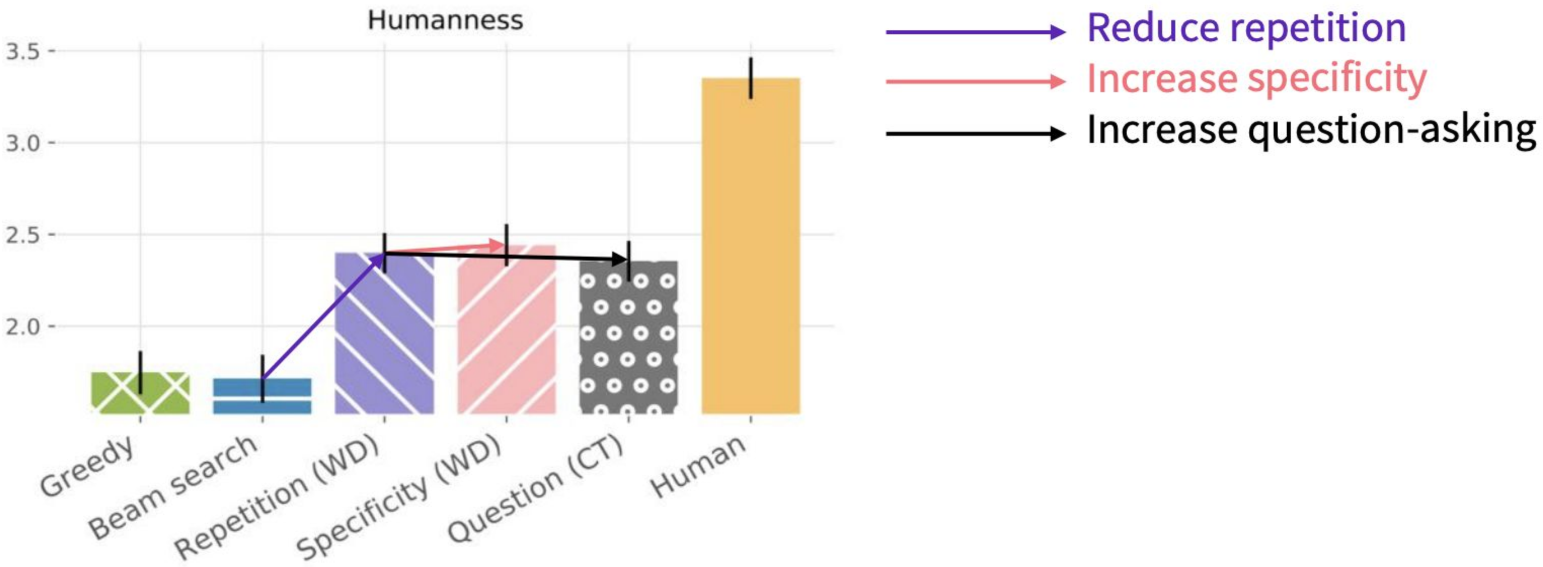


Our raw engagingness score matches the **ConvAI2 competition winner**'s GPT-based model, even though ours is:

- **much smaller** (2 layers vs 12)
- trained on **12x less data**

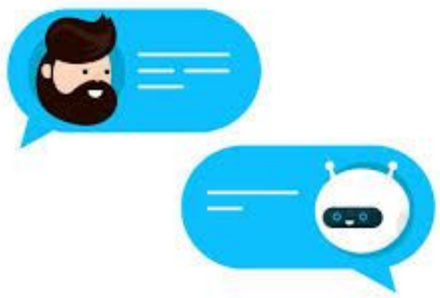


However: On the **humanness** (i.e. Turing test) metric, **our models are nowhere near human-level**!



What makes a good chatbot?

Chatbot = Human ?



- Engagingness is not equal to Humanness
- Bots are almost as engaging as human but non-human yet!
- Engagingness depends on the situation

Conclusions

- **Control is a good idea** for your neural sequence generation dialogue system.
- Using simple control, **we matched performance of GPT-based contest winner**.
- **Don't repeat yourself. Don't be boring. Ask more questions.**
- **Multi-turn phenomena** (repetition, question-asking frequency) are important – so need **multi-turn eval** to detect them.
- **Engagingness ≠ Humanness**, so think carefully about which to use.
- **Paid Turkers** are **not engaging conversationalists**, or good judges of engaging conversation. Humans chatting for fun may be better.
- **Problem**: Manually finding the best combination of control settings is **painful**.

Thank you.