

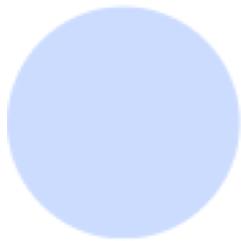


of Wikipedia: Knowledge-Powered Conversational Agents

*Emily Dinan, *Stephen Roller, *Kurt Shuster,
Angela Fan, Michael Auli, Jason Weston

(*joint first authors)

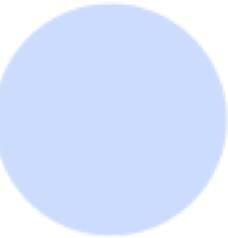
Facebook AI Research





of Wikipedia: Knowledge-Powered Conversational Agents

*Emily Dinan, *Stephen Roller, *Kurt Shuster,
Angela Fan, Michael Auli, Jason Weston



Current chatbots just chat, they aren't (em)powered with *knowledge*

- Seq2Seq models take in last lines of dialog, output a sentence
- We should ground on concrete information, rather than “generate and hope”

Current chatbots just chat, they aren't (em)powered with *knowledge*

- Seq2Seq models take in last lines of dialog, output a sentence
- We should ground on concrete information, rather than “generate and hope”

E.g. many studies/models using non-knowledge-based chit-chat datasets:

- Open-Subtitles (Vinyals & Le, 2015)
- Persona-Chat (Zhang et al., 2018)
- Twitter (Sordoni et al., 2015)

[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.

Existing Work

Goal directed dialogue uses knowledge :

- but via an API on structured knowledge
- e.g. restaurant or airline booking. (El Asri et al., 2017; Bordes et al., 2017)

Question-answering uses knowledge:

- but can only answer questions
- e.g. SQuAD (Rajpurkar et al., 2016) or QuAC (Choi et al., 2018)
- can use a retriever over unstructured text, e.g. Open-SQuAD (Chen et al., 2017)

Text based Knowledge for dialogue:

- Ghazvininejad et al. (2018) - local businesses using Foursquare tips as knowledge
- Parthasarathi & Pineau (2018) - news articles using Wikipedia summaries
- Moghe et al. (2018) – discuss movies given plot, reviews, etc.
- “A Dataset for Document Grounded Conversations” Zhou et al., this EMNLP!

To our knowledge, no convincing demonstration yet of full multi-turn dialogue in an open-domain setting...

Open-Domain Dialogue Setting

- Initial Starting Topic
- Pair of speakers converse naturally.
- They chat & learn from each other: discussion, facts, opinions.

Open-Domain Dialogue Setting

- Initial Starting Topic
- Pair of speakers converse naturally.
- They chat & learn from each other: discussion, facts, opinions.

Non-symmetric speakers:

- *Apprentice*: curious & eager to learn!
- *Wizard*: knowledgeable on area, eager to discuss & be engaging.
 - Given an IR system to condition response on.
 - Clicks on sentences used.



1307 Diverse General Topics: *crowd-sourced*

Gouda cheese
commuting
music festivals
podcasts
bowling
Arnold Schwarzenegger
Alpine skiing
Bodybuilding supplement
Harley-Davidson
Miley Cyrus
Hamilton (musical)
Ireland
Cannabis (drug)
Mount Kilimanjaro
Eggplant
Welding
Aquarium
Italian cuisine

Mercedes-Benz S-Class
Peanut
German language
Chicago-style pizza
Black hair
Toga party
100 metres
Tiger
Smoking
Winter
Acrophobia
List of art media
Pet adoption
Influencer marketing
Vitamin C
Human height
Steak

Kurt Cobain
List of water sports
Strawberry
Online game
Text messaging
Baileys Irish Cream
Fiction
American football
Online shopping
Aldi
Rock and roll
Kendrick Lamar
Medical billing
Blue Ridge Parkway
Clown
Pasta
Hiking



Each Linked to Wikipedia

A.1 HUMAN ANNOTATION INTERFACE (FOR WIZARD)

Chat with Knowledge!

You have just met the other person, who seems quite curious, and you are eager to discuss a topic with them!

You will try to inform your conversation partner about a topic that one of you will choose. After a topic is chosen, you will receive information about that topic that will be visible throughout the chat.

Passage for Chosen Topic

- Cupcake
 - A cupcake (also British English: fairy cake; Hiberno-English: bun; Australian English: fairy cake or patty cake) is a small cake designed to serve one person, which may be baked in a small thin paper or aluminum cup.
 - As with larger cakes, icing and other cake decorations such as fruit and candy may be applied.
 - The earliest extant description of what is now often called a cupcake was in 1796, when a recipe for "a light cake to bake in small cups" was written in "American Cookery" by Amelia Simmons.
 - The earliest extant documentation of the term "cupcake"

Relevant Information

Click on a topic below to expand it. Then, click the checkbox next to the sentence that you use to craft your response, or check 'No Sentence Used.'

No Sentence Used

Information about your partner's message

- Cupcake
- Hostess CupCake
 - Hostess CupCake is a brand of snack cake formerly produced and distributed by Hostess Brands and currently owned by private equity firms Apollo Global Management and Metropoulos & Co. Its most common form is a chocolate cupcake with chocolate icing and vanilla creme filling, with eight distinctive white squiggles across the top.
 - However, other flavors have been available at times.
 - It has been claimed to be the first commercially produced cupcake and has become an iconic American brand.

Information about your message

- Farley's & Sathers Candy Company
- Hi-Chew
- Candy
- Field ration
- Candy Candy
- Hi-5 (Australian band)
- Drum kit

SYSTEM: Your partner has selected the topic. Please look to the left to find the relevant information for this topic.

Partner: Hi! Do you have any good recipes for cupcakes?

SYSTEM: Please take a look at the relevant information to your left and check the appropriate sentence before answering, but try not to copy the sentence as your whole response.

You: Hi! You can add fruit and candy to make them even more delicious!

Partner: That's cool! What's your favorite cupcake?

SYSTEM: Please take a look at the relevant information to your left and check the appropriate sentence before answering, but try not to copy the sentence as your whole response.

I love Hostess cupcakes - they have chocolate icing and vanilla creme filling

Send

Dataset Examples

Topic: Lifeguard

Apprentice: So I am a lifeguard. Know anything about saving lives in water?

Wizard: I'm impressed! It's a big responsibility to supervise other people's safety in the water!
Tell me more.

Apprentice: Well, I help make sure people do not drown or get injured while in or near the water!

Knowledge: A lifeguard is a rescuer who supervises the safety and rescue of swimmers, surfers, ...
Lifeguards are strong swimmers and trained in CPR/AED first aid, certified in water ...

...

In some areas, the lifeguard service also carries out mountain rescues, or may function as the primary EMS provider.

Wizard: I've heard that in some places, lifeguards also help with other sorts of emergencies, like mountain rescues!

Is that part of your job too?

Apprentice: I have! I feel like you know much about this! What brings you to know so much?

Wizard: Oh, that's about the extent of my knowledge. I've just been around beaches and I've always admired lifeguards. I'm not a super strong swimmer myself.

Dataset Examples

Topic:	Armadillo
Wizard:	I love animals and think armadillos are awesome with their leathery shell.
Apprentice:	I don't think I've ever seen an armadillo in real life!
Wizard:	I've seen them at the zoo. Armadillo means little armored one in Spanish.
Apprentice:	Are they native to a Spanish-speaking part of the world?
Knowledge:	Armadillos are New World placental mammals in the order Cingulata ... The word “armadillo” means “little armoured one” in Spanish. ... The nine-banded armadillo (“ <i>Dasypus novemcinctus</i> ”), or the nine-banded, long-nosed armadillo, is a medium-sized mammal found in North, Central, and South America.
Wizard:	Yes, they are most commonly found in North, Central, and South America

Dataset Examples

Topic:	Ice cream
Wizard:	I just love ice cream. I love the types with fruits and flavours. Do you like ice cream?
Apprentice:	I love Ice cream as much as any one. I especially like Gelato, foreign ice cream!
Knowledge	<p>Ice cream is a sweetened frozen food typically eaten as a snack or dessert....</p> <p>It is usually made from dairy products, such as milk and cream, and ...</p> <p style="text-align: center;">...</p> <p>Bacon ice cream (or bacon-and-egg ice cream) is an ice cream generally created by adding bacon to egg custard and freezing the mixture.</p>
Wizard:	Me too. There are some strange combinations though, have you heard of bacon ice cream? where they add bacon and even egg custard to the freezing mixture!
Apprentice:	Surprisingly bacon ice cream doesn't surprise me. That doesn't sound appealing to me, but perhaps it could be delicious...

Table 1: Dataset statistics of the Wizard of Wikipedia task.

Wizard of Wikipedia Task	Train	Valid	Test Seen	Test Unseen
Number of Utterances	166,787	17,715	8,715	8,782
Number of Dialogues	18,430	1,948	965	968
Number of Topics	1,247	599	533	58
Average Turns per Dialogue	9.0	9.1	9.0	9.1
Knowledge Database	5.4M articles		93M sentences	

Models

- If we can learn models on this data we think it could be (part of?) one of the best chatbots out there. Might actually want to talk to it?
 - Fun, engaging + knowledgeable !

Models require:

1. Read current dialogue
2. Retrieve knowledge (*e.g. IR system, operates every turn*)
3. Read/attend on results
4. Condition & generate something fun **and** knowledgeable!

Generative Model architecture

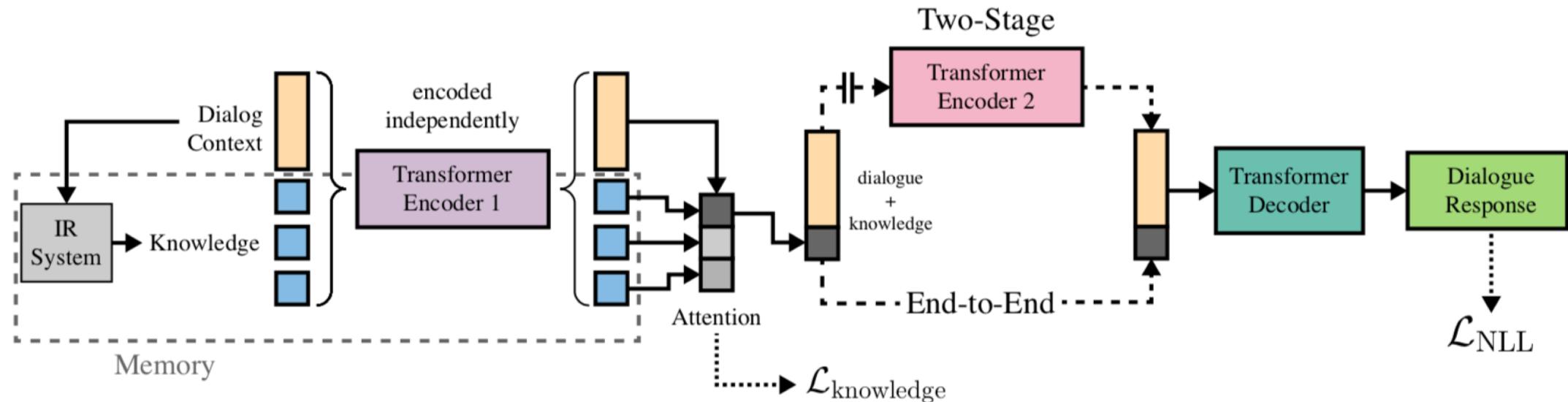


Figure 1: **Generative Transformer Memory Network.** An IR system provides knowledge candidates from Wikipedia. Dialogue Context and Knowledge are encoded using a shared encoder. In the Two-stage model, the dialogue and knowledge are re-encoded after knowledge selection.

Retrieval Model architecture

Similar to generative model, but:

- (1) we don't attend to only the best returned knowledge (*keep the weighted attention sum*)
- (2) we replace the decoder with ranker of encoded sentences:

We choose as a response r_ℓ where

$$\ell = \arg \max_{i \in \{1, \dots, L\}} \frac{\text{rep}_{\text{LHS}}(m_{c_1}, \dots, m_{c_K}, x)}{\|\text{rep}_{\text{LHS}}(m_{c_1}, \dots, m_{c_K}, x)\|_2} \bullet \frac{\text{rep}_{\text{RHS}}(r_i)}{\|\text{rep}_{\text{RHS}}(r_i)\|_2}.$$

- Transformers pretrained on 1.7 billion Reddit dialogue examples, following (Mazaré et al., 2018). *Show to work very well for PersonaChat dialogue*

Full Dialogue Task: Retrieval

Table 3: **Retrieval methods on the full Wizard task.** Models must select relevant knowledge and retrieve a response from the training set as a dialogue response. Using knowledge always helps, and the Transformer Memory Network with pretraining performs best.

Method	Predicted Knowledge				Gold Knowledge	
	Test Seen		Test Unseen		Seen	Unseen
	R@1	F1	R@1	F1	R@1	R@1
Random	1.0	7.4	1.0	7.3	1.0	1.0
IR baseline	17.8	12.7	14.2	11.6	73.5	67.5
BoW MemNet (no knowledge)	56.1	14.2	28.8	11.6	56.1	28.8
BoW MemNet	71.3	15.6	33.1	12.3	84.5	66.7
Transformer (no knowledge, w/o Reddit)	60.8	13.3	25.5	9.7	60.8	25.5
Transformer (no knowledge, w/ Reddit)	79.0	15.0	54.0	11.6	79.0	54.0
Transformer MemNet (w/ Reddit)	86.8	15.4	69.8	12.4	91.6	82.3
Transformer MemNet (w/ Reddit+SQuAD)	87.4	15.4	69.8	12.4	92.3	83.1

SubTask: Knowledge Selection

Table 2: **Test performance of various methods on the Knowledge Selection Task.** The models must select the gold knowledge sentences chosen by humans given the dialogue context.

Method	Seen Test		Unseen Test	
	R@1	F1	R@1	F1
Random	2.7	13.5	2.3	13.1
IR baseline	5.8	21.8	7.6	23.5
BoW MemNet	23.0	36.3	8.9	22.9
Transformer	22.5	33.2	12.2	19.8
Transformer (+Reddit pretraining)	24.5	36.4	23.7	35.8
Transformer (+Reddit pretraining, +SQuAD training)	25.5	36.2	22.9	34.2

Full Dialogue Task: Generation

Table 4: **Generative models on the full Wizard Task.** The Two-stage model performs best using predicted knowledge, while the End-to-end (E2E) model performs best with gold knowledge.

Method	Predicted Knowledge				Gold Knowledge			
	Test Seen		Test Unseen		Test Seen		Test Unseen	
	PPL	F1	PPL	F1	PPL	F1	PPL	F1
Repeat last utterance	-	13.8	-	13.7	-	13.8	-	13.7
Transformer (no knowledge)	-	-	-	-	41.8	17.8	87.0	14.0
E2E Transformer MemNet (no auxiliary loss)	66.5	15.9	103.6	14.3	24.2	33.6	35.5	29.5
E2E Transformer MemNet (w/ auxiliary loss)	63.5	16.9	97.3	14.4	23.1	35.5	32.8	32.2
Two-Stage Transformer MemNet	54.8	18.6	88.5	17.4	30.0	30.7	42.7	28.6
Two-Stage Transformer MemNet (w/ K.D.)	46.5	18.9	84.8	17.3	28.6	30.6	43.7	28.0

Human Evaluations

Table 5: **Human Experiments.** Evaluations of the best generative and retrieval models on full dialogues with humans. Human ratings are reported as mean (stddev). Wiki F1 measures unigram overlap with the Wikipedia entry for the chosen topic, a measure of knowledge used in conversations.

Method	Seen		Unseen	
	Rating	Wiki F1	Rating	Wiki F1
Human Performance	4.13 (1.08)	11.1	4.34 (0.98)	10.6
Retrieval Transformer (no knowledge)	3.33 (1.30)	19.8	3.12 (1.34)	13.7
Generative Transformer (no knowledge)	2.11 (1.11)	15.3	2.54 (1.38)	10.1
Retrieval Transformer MemNet	3.43 (1.10)	23.4	3.14 (1.31)	16.3
Two-Stage Generative Transformer MemNet	2.92 (1.33)	30.0	2.93 (1.30)	26.2

Model: Retrieval Transformer MemNet

Topic: E-book (*seen*)

Human: i do not know why, but I have never gotten into E-books

Model: Same here. I have a kindle, but I don't use it for books. It's nicer to hold a real book and turn pages.

Human: yes, i like the physical feel and smell of a real book

Model: I actually had a Nook e-book where you had to download the e-book after selecting it online. I grew out of it, I prefer paper books also.

Human: i tried e-books but I do not like reading on a screen, its disorientated for me

Model: Yes, reading a book on a tablet just isn't the same!

Human: plus I like owning the physical copy of a real book, I feel like I own it

Model: I prefer print and newsreels.

Topic: Toga party (*seen*)

Model: I love toga parties! I usually make my toga out of a bed sheet and wear my sandals.

Human: Me too!!! I usually use some flip flops and my blanket!

Model: Toga parties typically have games involving kegs and excessive drinking!

Human: Yeah! And getting to wear funny outfits and a crown of leaves!

Model: I love me some Toga Parties!

Human: I love me some too!

Model: Toga parties are so much fun with the Roman or Greek theme.

Human: Yeah! I especially like Greek themes

Topic: Steak (*seen*)

Human: Hello, it's lunch time here, and I'm in the mood for a great steak

Model: Wow that sounds delicious I wish I could say the same hahah, When the word "steak" is used without qualification, it normally refers to a Beefsteak

Human: Interesting, I don't really have a favorite steak, how about you?

Model: I like rib eye it is really good specially medium rare, When the word "steak" is used without qualification, it generally refers to a beefsteak

Human: I like rib eye too! What is it about a rib eye that you enjoy?

Model: Most steak cuts are very tender meat but I think rib eye can be quite tender

Human: They really are. How do you choose a good steak?

Model: When it comes to steak and other cuts of meat you have rare, medium rare, medium, well done and so on

Human: How do you prefer your steak prepared?

Model: I like that you can prepare it many different ways. You can fry it, grill it, or even broil the steak

Topic: Mount Kilimanjaro (*seen*)

Model: I have a friend who will climb Mount Kilimanjaro, a volcano in Tanzania

Human: That's neat, are they very experienced climbers?

Model: It is a quite popular climb, while tough, Kilimanjaro is not as difficult as the Everest or K2.

Human: I'm sure it's a challenge though.

Model: It is, I also tried Mount Everest, but that was also just a hike.

Human: Right, sounds like you can do just about anything.

Model: I would like to climb Mount Kilimanjaro, the highest mountain in Africa one day, for mountaineering.

Human: Maybe take a walk on the moon too.

Topic: Eggplant (*seen*)

Model: Eggplant is the edible fruit of the nightshade plant, it's pretty good! do you like it??

Human: I have never had eggplant.

Model: I've only had it a few times, I think it's interesting that eggplant is actually in the nightshade family.

Human: I have looked up recipes for eggplant parmesan. I want to make it one day

Model: Do you like anything else that has eggplants in it?

Human: I think you can put it in a salad.

Model: Yes, a Greek salad has tomatoes, cucumbers, onions, and feta cheese.

Human: I like casear salad

Model: Have you ever tried any eggplant salads? Eggplant is also used as an appetizer.

Human: Can you fry eggplant?

Topic: Aquarium (*seen*)

Human: Hi! I'm a huge fan of aquariums, and go to them a few times a year. Do you enjoy aquariums?

Model: I have a big Aquarium in my back yard, is a vivarium of any size having at least one transparent side in which aquatic plants or animals are kept

Human: Whoa, that's really cool! What kind of animals do you have in there?

Model: A lot like bears, bottles, worms, frogs, hamburgers, sharks, toy soldiers, full-size rats, large human body parts (such as hearts and feet), and more. Do you like the sweet or sour worms?

Human: I like any kind of gummy, but I am partial to sweet gummy worms.

Model: I'm not sure! There are so many great choices! I've seen gummiI bears, bottles, worms, frogs, sharks and even full-size rats! There are quite a few manufacturers of gummis too.

Human: I've always wanted to try some of those giant gummies, I'd probably love them.

Model: The Gummi snakes used to scare me as a kid. The Gummies today come in every animal and shape. My daughter loves the toy soldiers at Christmas.

Topic: American football (*unseen*)

Model: American football can be traced back to Britain in the mid-19th century. The rules were very similar!

Human: I did not know that. Did they wear equipment back then?

Model: Am not really certain about it but American football resulted from several major divergences from association football and rugby football

Human: The Philadelphia eagles actually have a rugby player on their team this year. he never played football until this year.

Model: Philadelphia is one of a few cities to have all four teams, but they love the Eagles more than anything else by a long shot!

Human: As a life long Philadelphia sports fan I have to agree with you.

Model: I've always been more of a fan of the American Football team from Pittsburgh, The Steelers!

Human: Being from Philadelphia I am not much of a pittsburgh fan.

Model: Are the Pittsburgh Steelers in the AFC?

Human: Yes they are. the AFC north to be exact.

Topic: Ireland (*unseen*)

Model: Hi, have you ever been to Dublin? It is the capital and largest city in Ireland.

Human: No, I haven't but would love to

Model: It is the largest city in Northern Ireland.

Human: I've always found Ireland to be fascinating and would love to visit sometime

Model: Same here! It is the capital of Ireland too!

Human: What is the most popular place to visit in Ireland?

Model: I would recommend the Thrace, Crete, and the Ionian Islands as must see places. (*)

Human: I don't know anything about the Ionian Islands

Model: It was once referred to as the Hellenic Republic! (*)

Model: Two-Stage Generative Transformer MemNet

Topic: Circus (*seen*)

Model: have you ever been to a circus? it is a company of performers who put on diverse entertainment shows.

Human: I went to the circus once when I was a child actually.

Model: that's great! entertainment is a form of activity that holds the attention and interest of an audience.

Human: That's true. Have you ever been to a circus?

Model: I have not, but I know they were built with various types of seating ring, and sometimes a stage.

Human: I went to one held in a stadium, so we had box seats.

Model: that's awesome! i'm sure you'll have to go to the circus, with appropriate music.

Human: Yes, circuses usually have their own kind of music.

Model: circuses can be done by circuses.

Model: Two-Stage Generative Transformer MemNet

Topic: Shower (*unseen*)

Human: So I've been looking for a better showerhead with more water pressure for my shower.

Model: I love pressure-verses water at nearly constant temperature to a shower

Human: That sounds pretty great! Love those rainshower showers!

Model: the ancient greeks were the first people to have showers.

Human: Well I knew they were good for something.

Model: the original showers were neither indoor structures nor man made but were common.

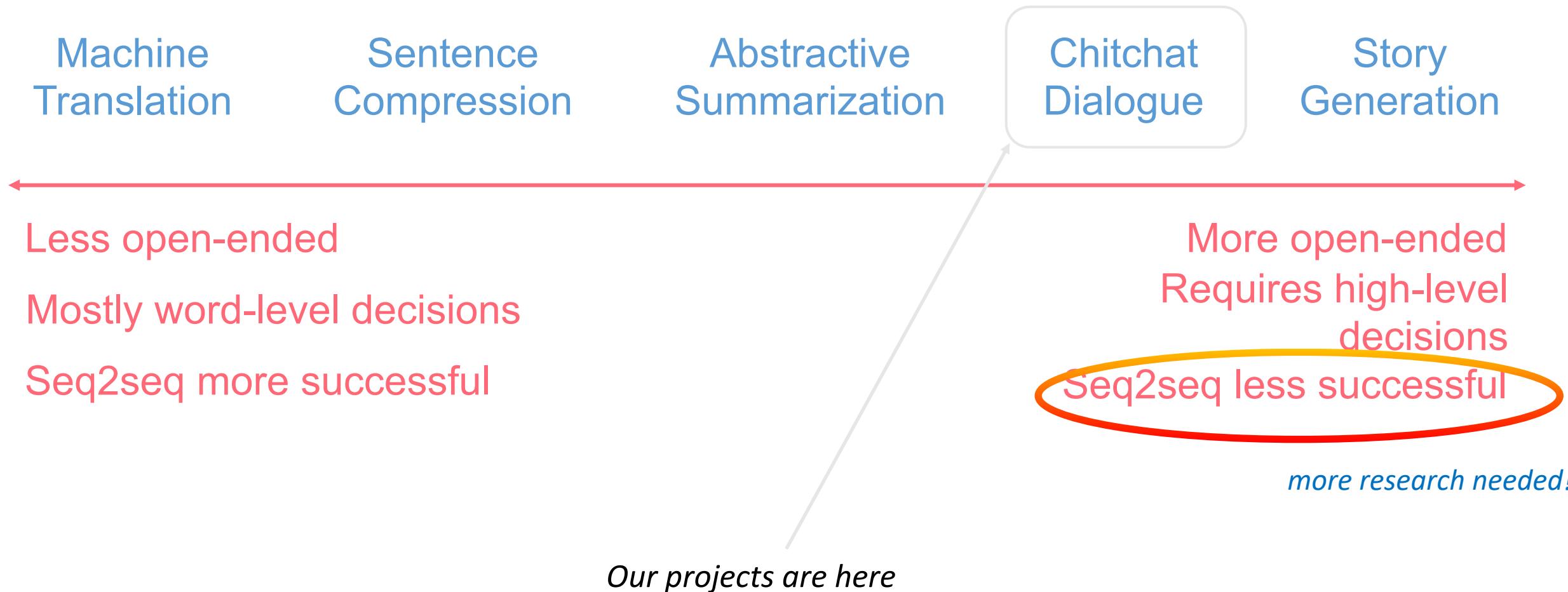
Human: Ah that sounds intimidating!

Model: it is required manual transport of both fresh and waste water.

Error Analysis Summary

- **Retrieval models** are surprisingly good on seen topics (*not always of course*)
- On unseen, they can fail very badly
- **Generative models** aren't as good on seen, but *can* generalize to unseen.
- They still suffer from several problems (fixes: open research problems):
 - local & global repetition in dialogue, fixed to some extent by knowledge grounding
 - too much reliance on copying (“I don’t know, but I do know that [Wikipedia excerpt] ”)
 - clearly not as engaging as retrieval, but catches up on unseen
 - balance issues: too many statements, not enough questions

Natural Language Generation task spectrum



Retrieve & Refine

Jason Weston, Emily Dinan, Alex Miller

- Retrieval models: engaging, but sometimes weird results, can't make new text
- Generative models: "I don't know" problem, short sentences, common words, boring

Combine them!!!

Variants:

RetNRef First Version

RetNRef+ Reweighting to Pay more attention to retrieval

RetNRef++ Fix 'Exact' Copies of Retriever

Method	Word cnt	Char cnt	Rare Word % <100	Rare Word % <1k
Seq2Seq	11.7	40.5	0.4%	5.8%
RetNRef	11.8	40.4	1.1%	6.9%
RetNRef ⁺	12.1	45.0	1.7%	10.1%
RetNRef ⁺⁺	12.7	48.1	2.3%	10.9%
MemNet	13.1	54.5	4.0%	15.3%
Human	13.0	54.6	3.0%	11.5%

Table 2: Output sequence statistics for the methods. Seq2Seq generates shorter sentences with more common words than humans, which RetNRef alleviates.

Person 1	hey , george here . lets be friends
Person 2	hi george , i'm regina . do you enjoy dance ?
Seq2Seq	i do not but i do have a lot of time to dance
MemNet	hello george my name is pink
RetNRef ⁺	i do. i enjoy dance.

Person 1	yeah. well, i wish you luck there. we bought a cute 2 story house.
Person 2	nice , i don't think i have ever left ohio.
Seq2Seq	i am sorry. i am so scared of twins.
MemNet	i am from mississippi . just moved back here from austin texas
RetNRef ⁺	i am from tennessee. just moved here from austin texas.

Method	Engagingness
Seq2Seq (PPL)	2.70(1.17)
Seq2Seq (100 epochs)	2.76(1.15)
Memory Network	3.66(1.26)
RetrieveNRefine	2.94(1.26)
RetrieveNRefine ⁺	3.50(1.33)
RetrieveNRefine ⁺⁺	3.80(1.18)

Conclusions:

- RetNRef word statistics look more similar to human sequences than Seq2Seq
- Humans prefer RetNRef over Seq2Seq and retrieval MemNet

Conclusion



- Models:
 - Need to be engaging:
 - (i) fun to talk to,
 - (ii) expert/knowledgeable

Open problems:

- Improve generative models to level of retrieval models
- Generalize well to unseen topics
- Natural to multitask with other tasks, e.g. get strong QA performance