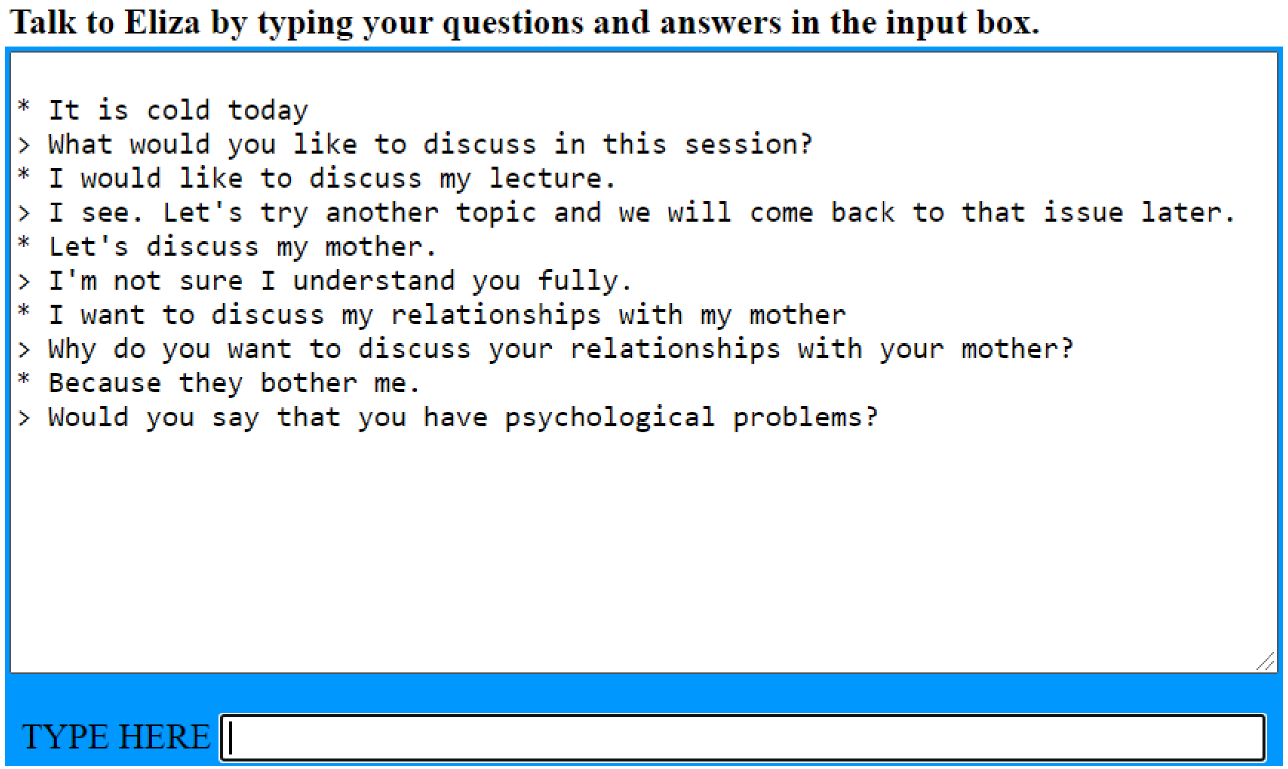
Chatbots

Irina Piontkovskaya

Huawei Noah’s Ark Moscow

AI as a Chatbot


Talk to Eliza by typing your questions and answers in the input box.

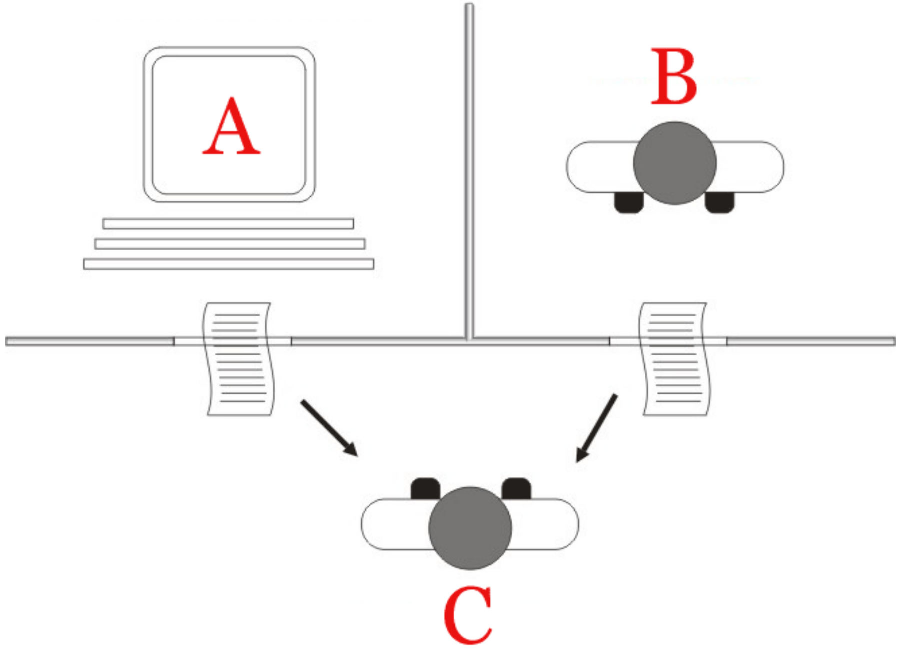
```
* It is cold today
> What would you like to discuss in this session?
* I would like to discuss my lecture.
> I see. Let's try another topic and we will come back to that issue later.
* Let's discuss my mother.
> I'm not sure I understand you fully.
* I want to discuss my relationships with my mother
> Why do you want to discuss your relationships with your mother?
* Because they bother me.
> Would you say that you have psychological problems?
```

ELIZA – chatbot, simulating psychotherapist (1966)

PARRY attempted to simulate a person with paranoid schizophrenia(1972)

https://web.njit.edu/~ronkowit/eliza.html

Turing test



User C (interrogator) is trying to determine if A a computer or a person

Turing test

Eugene Goostman is portrayed as being a 13-year-old boy from Odessa, Ukraine, who has a pet guinea pig and a father who is a gynecologist.

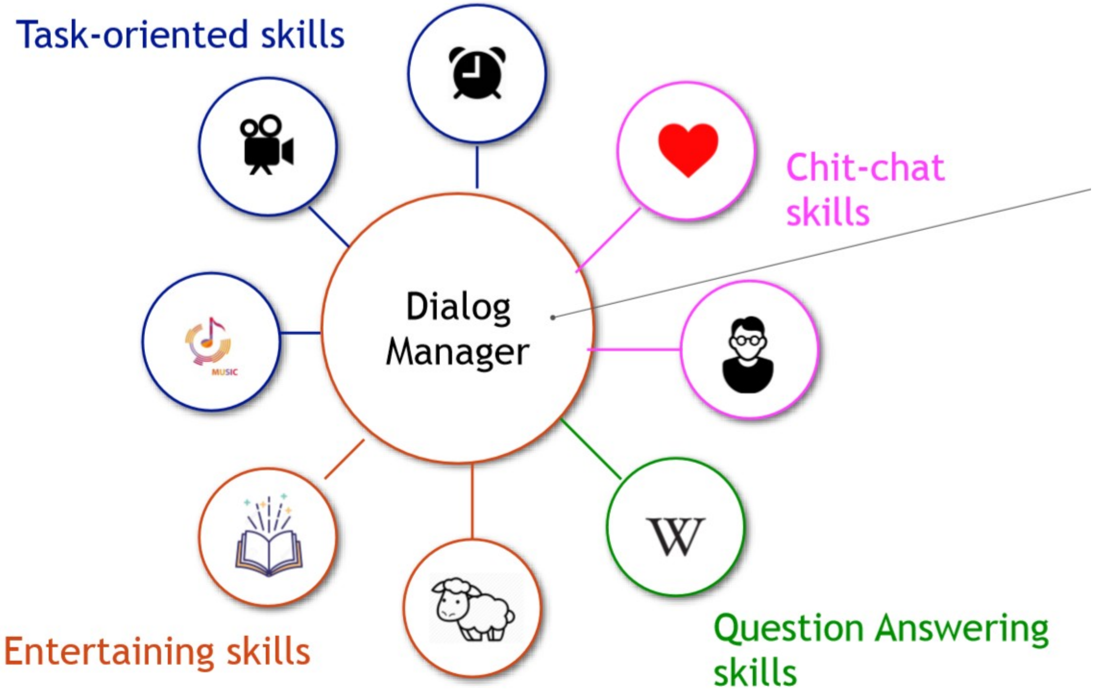
Developed in Saint Petersburg in 2001 * Passed Turing Test in 2014

Scott: Which is bigger, a shoebox or Mount Everest?
Eugene: I can’t make a choice right now. I should think it out later. And I forgot to ask you where you are from…
Scott: How many legs does a camel have?
Eugene: Something between 2 and 4. Maybe, three? :-))) By the way, I still don’t know your specialty – or, possibly, I’ve missed it?
Scott: How many legs does a millipede have?
Eugene: Just two, but Chernobyl mutants may have up two five. I know you are supposed to trick me.
Scott: No, I need to know that you're not a chatbot. Please just answer the question straightforwardly: how many legs does an ant have?
Eugene: Something between 2 and 4. Maybe three? :-))) Oh, what a fruitful conversation ;-)

Desirable AI bot features

- Fluent language
- Understanding
- Keeping context
- Following the goal
- Common sense and world knowledge
- Persona consistency
- Empathy
- …

Motivating example:
scalable virtual assistant architecture



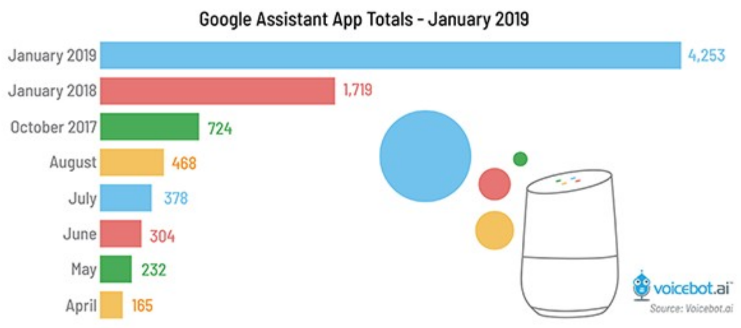
Virtual assistant supports a lot of possible actions:

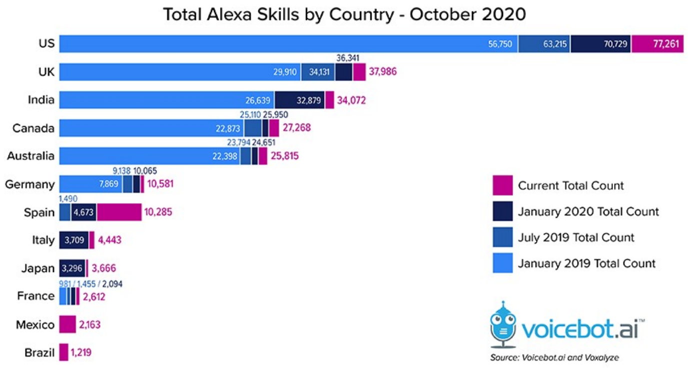
- chit-chat (just talk)
- information (who is John Donne?)
- games
- read a fairy tail
- call taxi
- order pizza
- check weather
- send message
- find recipe

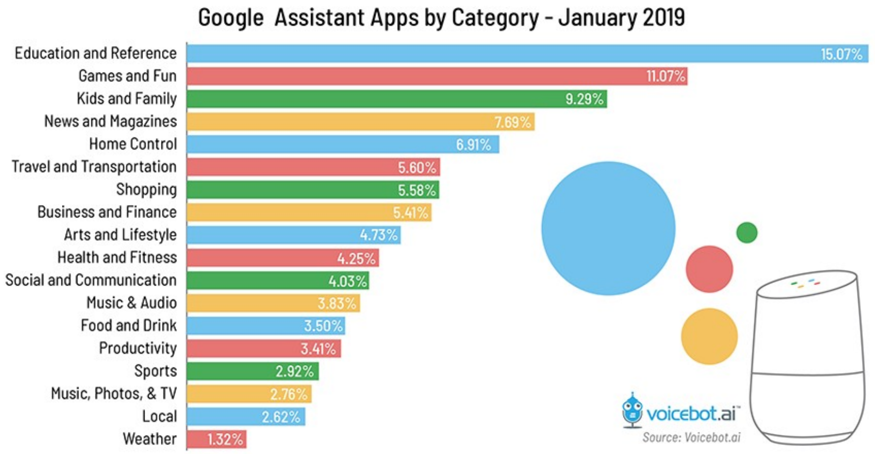
….

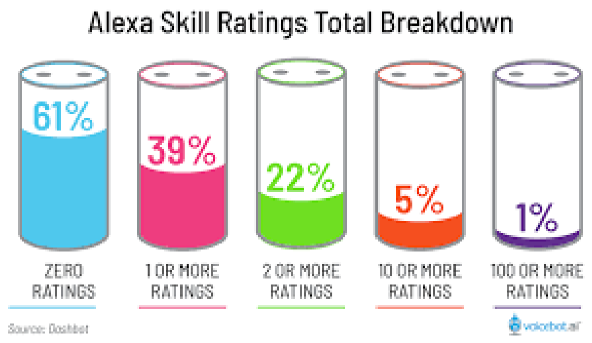
Virtual assistants statistics

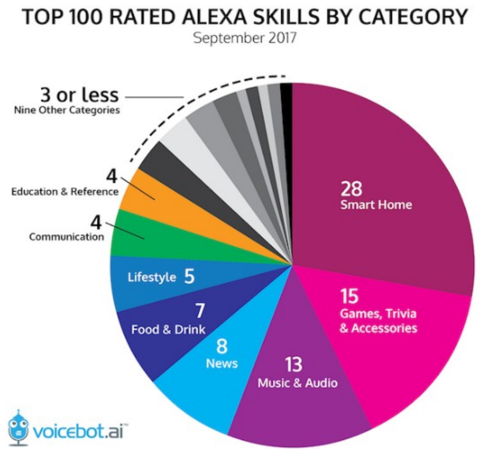
HUAWEI





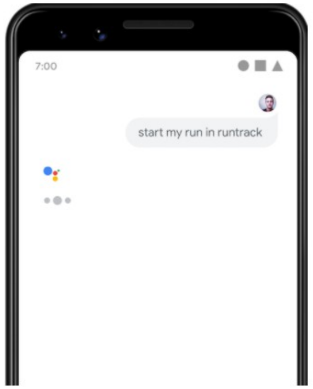






How to add new skill?

Google Actions



start my run in runtrack

Figure 1. Invoking START_EXERCISE.

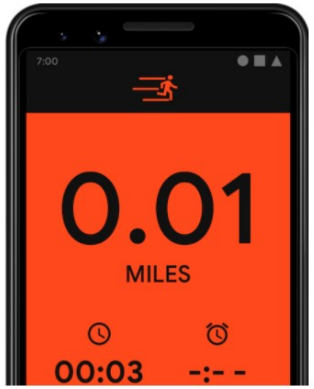
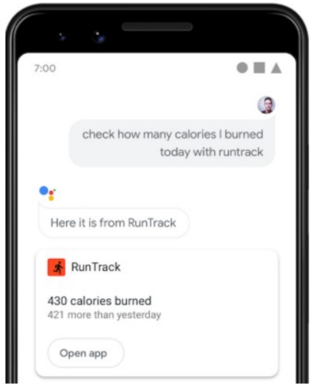


Figure 2. Deep linking into the app.



check how many calories I burned today with runtrack

Here it is from RunTrack

430 calories burned

421 more than yesterday

Figure 3. Launching a Slice in Assistant.

Framework to support your app in Google Assistant

https://developers.google.com/assistant/app

Custom intents and Built-in intents are available

Built-in intents: ~60 intents in 11 categories

Category: Finance

Pay Invoice Initiate payment of a user's bill or invoice

Example queries:

Make the minimum payment on my account using my debit card.

I want to pay my monthly bill with my checking account.

Category: Health and Fitness

Start Exercise Initiate the user's fitness-related physical activity in real-time (for example, live tracking of a run).

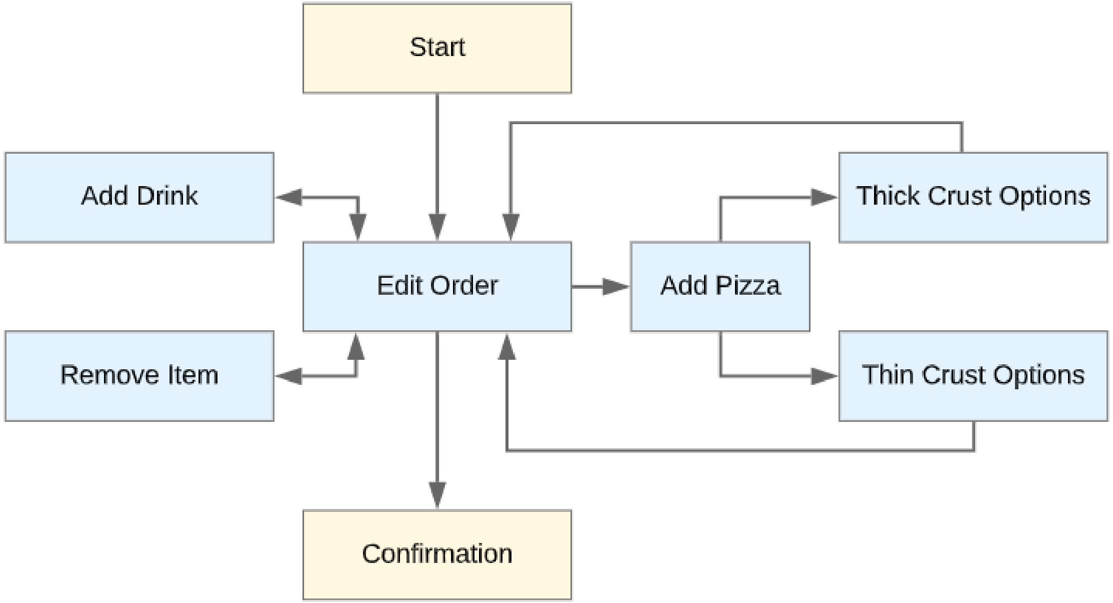
Example queries:

Start my run

Track my hike

Dialogflow

Dialog is a state machine



Routes (state transitions) defined by:

- Intent requirements
- Conditions

Task-oriented dialogs

- Well-defined task: clear metric of success

- Restricted scope: fixed number of intents + OOD (out-of-domain) class

- ML tasks: intent classification, slot filling, dialog state tracking

Datasets:
Wizard-of-Oz setup
Dialog self-play setup

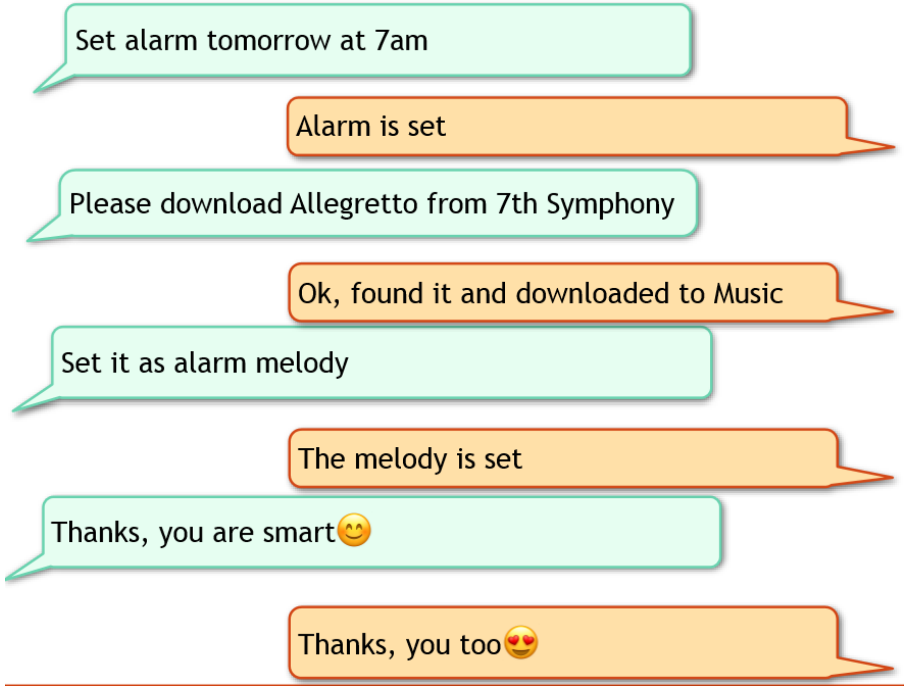
Metrics:
Intent classification accuracy
Slot filling (F1)
Joint Goal accuracy

Issues

- Skewed, low quality, insufficient data

Please *get me the weather in Moscow for tomorrow*

I want *o watch the most popular show on Netflix*



Set alarm tomorrow at 7am

Alarm is set

Please download Allegretto from 7th Symphony

Ok, found it and downloaded to Music

Set it as alarm melody

The melody is set

Thanks, you are smart😊

Thanks, you too😍

- Hard to support new intents, the whole model’s re-training is required

Question: how to add new intent without data collection and re-training?

Schema-guided dialog

Service: Payment
Description: “Digital wallet to make and request payments”

Intents:
MakePayments
Description: “Send money to your contact”

RequestPayment
Description: “Request money from a contact”

Please put **$10** in **John Doe**'s checking account

Slots:
amount
Description: “Amount of money to transfer or request”

contact_name
Description: “Name of contact for transaction”

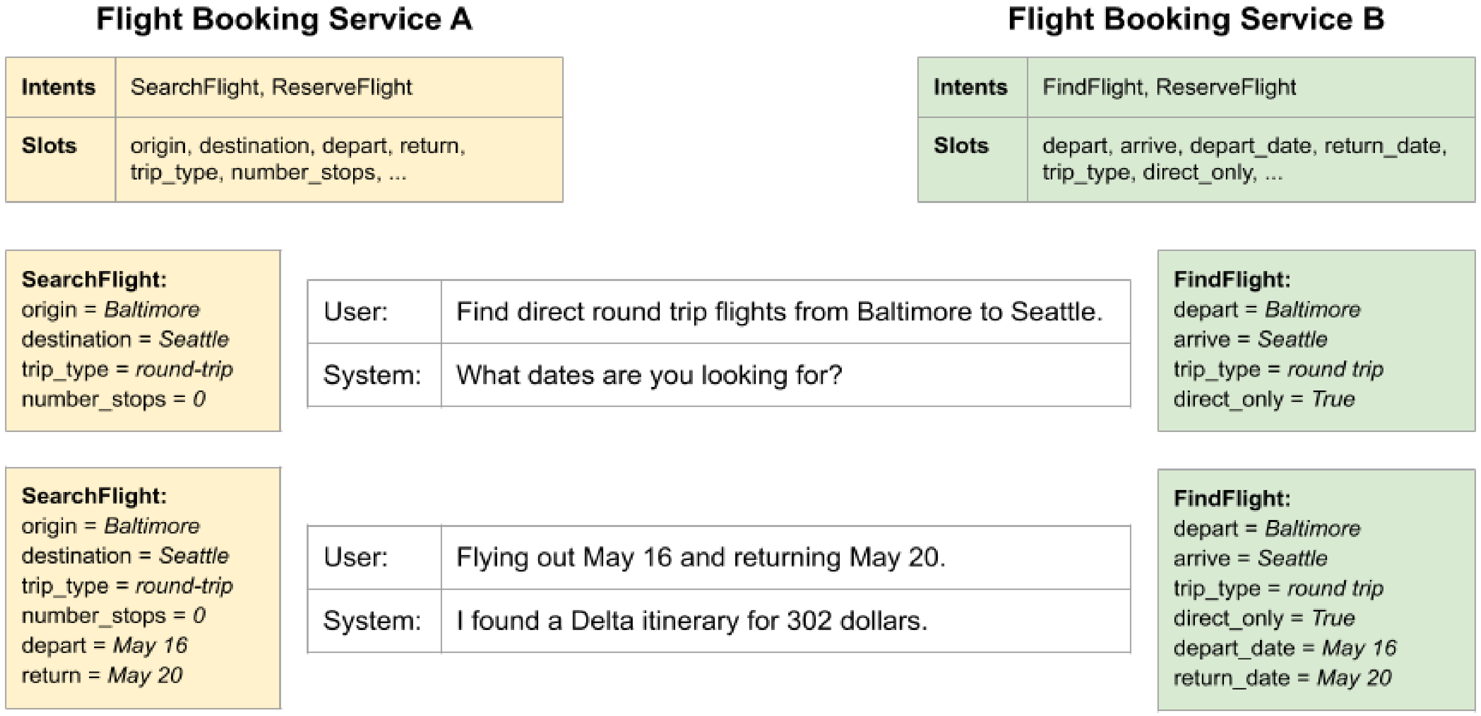
Rastogi A. et al. Schema-guided dialogue state tracking task at DSTC8 //arXiv preprint arXiv:2002.01359. – 2020.

Schema-guided dialog challenge

https://github.com/google-research-datasets/dstc8-schema-guided-dialogue

Main idea:

- New skill represented as schema
- The task is to predict dialog state (slots, intents) both for schemas seen in training set, and not seen
- Dataset consists of labelled dialogs and schema descriptions
- Training set: 20 schemas, ~1000 dialog per schema
- 5 new schema without any data in training set



Flight Booking Service A

Intents	SearchFlight, ReserveFlight
Slots	origin, destination, depart, return, trip_type, number_stops, ...

Flight Booking Service B

Intents	FindFlight, ReserveFlight
Slots	depart, arrive, depart_date, return_date, trip_type, direct_only, ...

SearchFlight:
origin = *Baltimore*
destination = *Seattle*
trip_type = *round-trip*
number_stops = *0*

User:	Find direct round trip flights from Baltimore to Seattle.
System:	What dates are you looking for?

FindFlight:
depart = *Baltimore*
arrive = *Seattle*
trip_type = *round trip*
direct_only = *True*

SearchFlight:
origin = *Baltimore*
destination = *Seattle*
trip_type = *round-trip*
number_stops = *0*
depart = *May 16*
return = *May 20*

User:	Flying out May 16 and returning May 20.
System:	I found a Delta itinerary for 302 dollars.

FindFlight:
depart = *Baltimore*
arrive = *Seattle*
trip_type = *round trip*
direct_only = *True*
depart_date = *May 16*
return_date = *May 20*

The predicted dialogue state for the first two user turns for an example dialogue, showing the active intent and the slot assignments. Note that the representation is conditioned on the schema under consideration.

Dialog State Tracking for new schemas

Task: classify intents (usually 1-4 intents per skill, no_intent is possible); one model supports all the skills, includ

Approach: binary classification of all possible pairs (utterance, intent description).

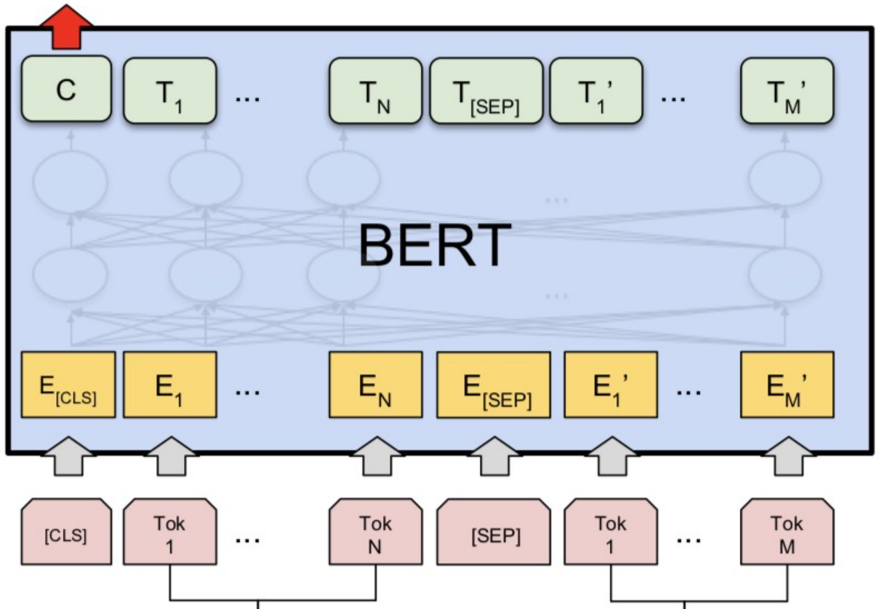
Model: BERT or other pre-trained Language Model

Best results: ~**98% intent classification accuracy**

Simplified setup: service is known

Drawback: needs separate pass for each class

Yes / No



Please find flight to Seattle for me

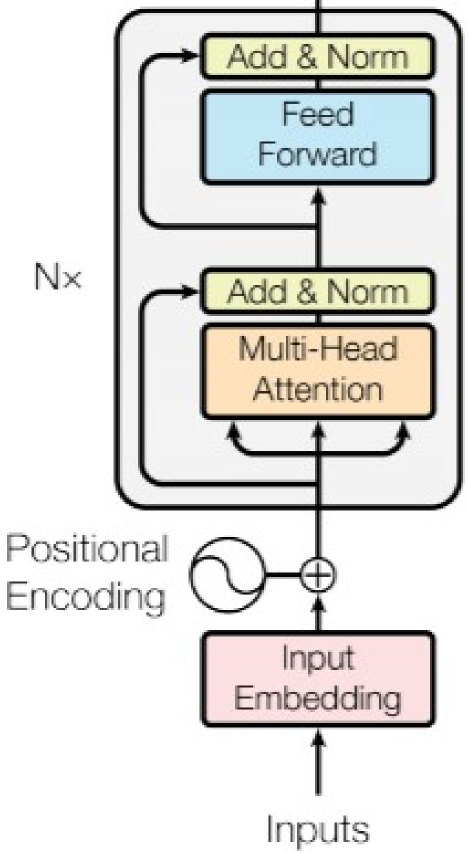
Flight Booking Service
Find flight

Pre-trained Language Mo

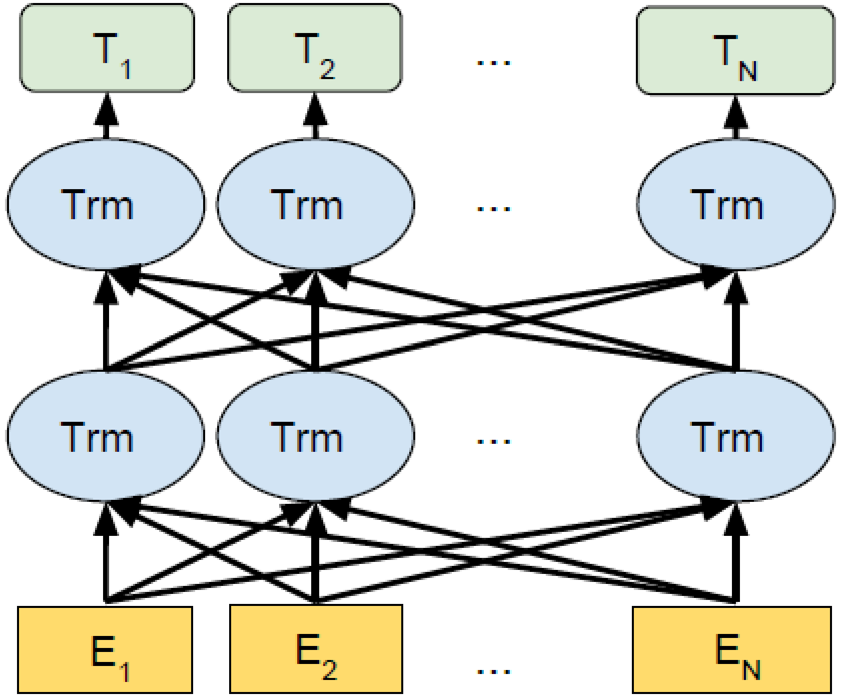
Transformer architecture

https://arxiv.org/abs/1706.03762

Attention Is All You Need Vaswani et al., Google Brain (2017)



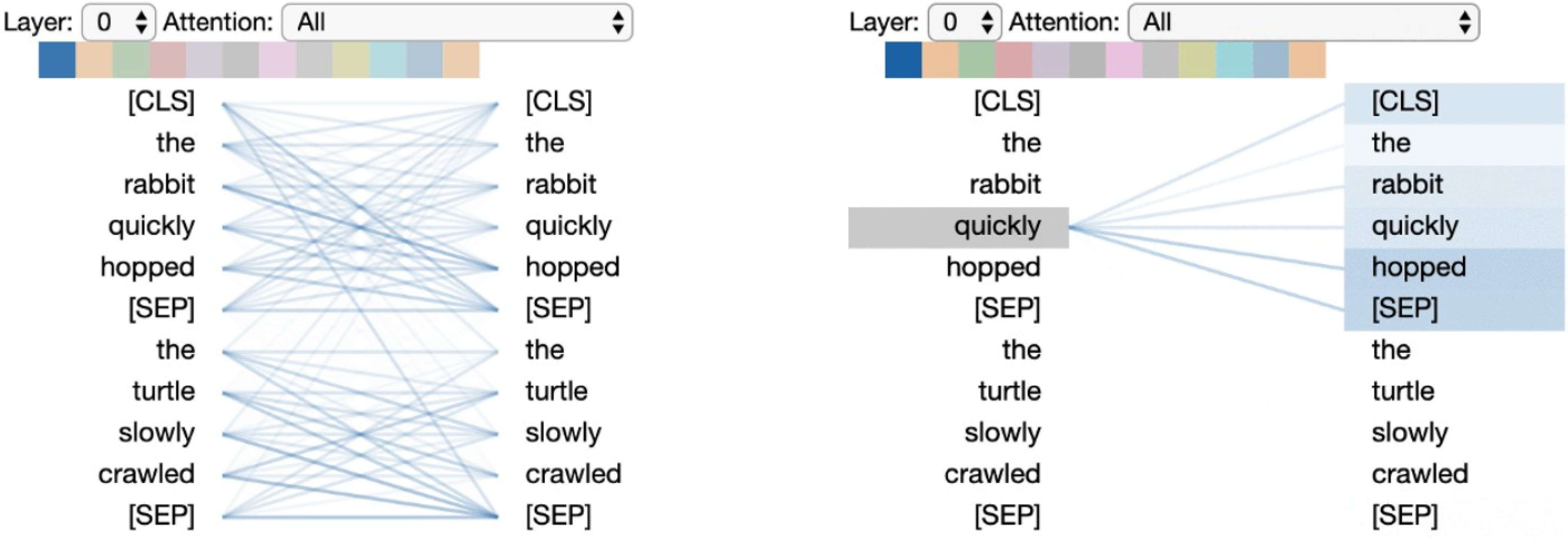
Attention mechanism



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

Attention is a way to obtain novel representation of tokens, based on combination of previous layer representations of tokens with interpretable weights

Attention



Left: visualization of attention between all words in the input. **Right:** visualization of attention from selected word only.

https://towardsdatascience.com/deconstructing-bert-part-2-visualizing-the-inner-workings-of-attention-60a16d86b5c1

BERT model

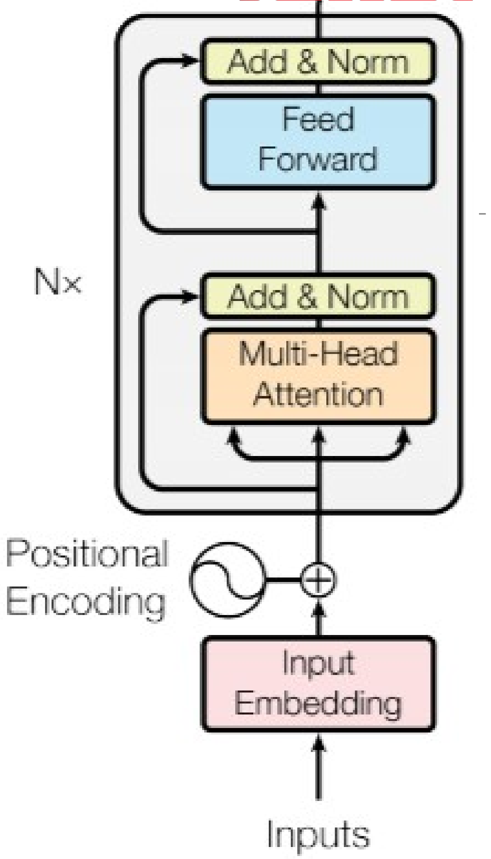
https://arxiv.org/abs/1706.03762

Attention Is All You Need Vaswani et al., Google Brain (2017)

Consists of the stack of the same multihead self-attention layers

Input: 1 vector for each token

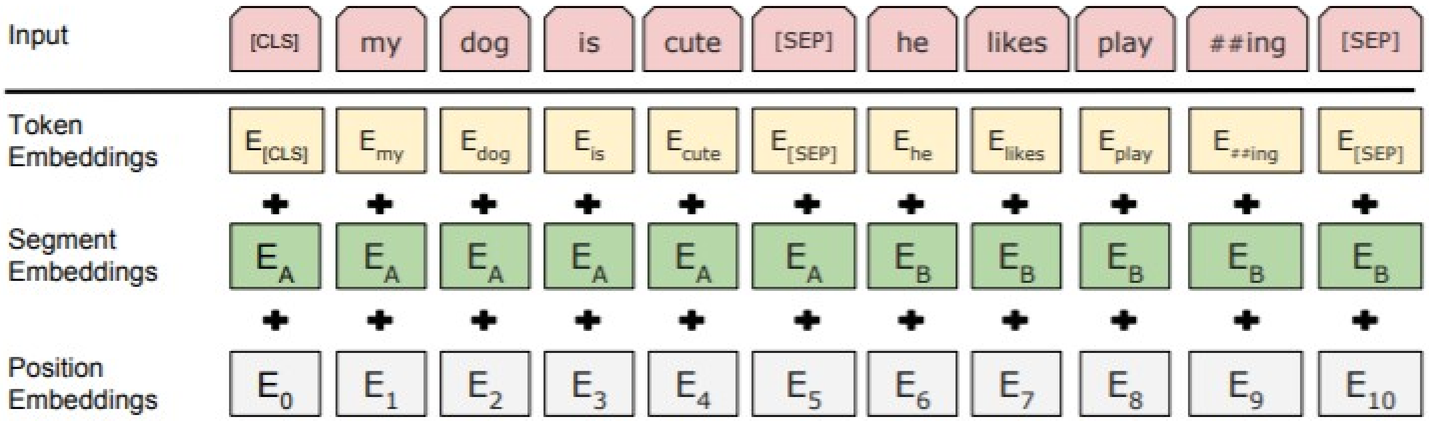
Output of attention block: 1 vector per token



BERT input representation

https://arxiv.org/abs/1810.04805

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (Devlin et al.)



Input
[CLS] my dog is cute [SEP] he likes play ##ing [SEP]
Token Embeddings
$E_{[CLS]}$ E_{my} E_{dog} E_{is} E_{cute} $E_{[SEP]}$ E_{he} E_{likes} E_{play} $E_{\#\#ing}$ $E_{[SEP]}$
Segment Embeddings
E_A E_A E_A E_A E_A E_A E_B E_B E_B E_B E_B
Position Embeddings
E_0 E_1 E_2 E_3 E_4 E_5 E_6 E_7 E_8 E_9 E_{10}

- Word: Jet makers feud over seat width with big orders at stake
- wordpieces: _J et _makers _fe ud _over _seat _width _with _big _orders _at _stake

Word-pies tokenization, dictionary of 30000 words: dictionary is created from Minimal Descriptive Length requirement fro the corpus
(longer pieces -> more frequent)

Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. Wu et al. https://arxiv.org/abs/1609.08144

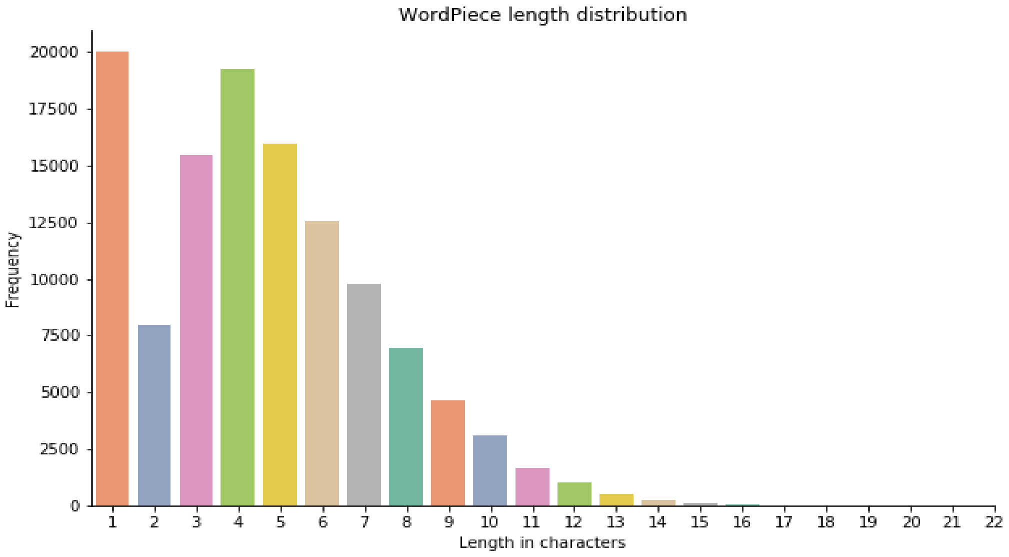
BERT input representation

https://arxiv.org/abs/1810.04805

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (Devlin et al.)

Multilingual BERT vocabulary
~ 100 000 word-pieces
~ 20 000 one-character tokens (alphabet)
~ 100 special tokens (CLS, SEP, UNK, PAD etc

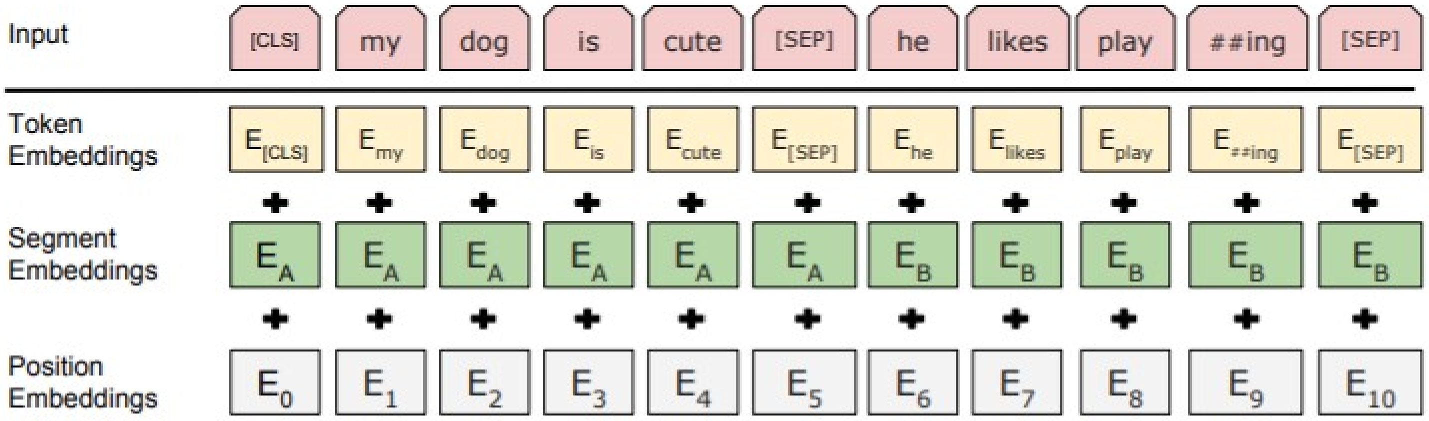
Latin	93495	78.21%
CJK+kana	14932	12.49%
Cyrillic	13782	11.53%
Indian	6545	5.47%
Arabic	4873	4.08%



BERT input representation

https://arxiv.org/abs/1810.04805

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (Devlin et al.)

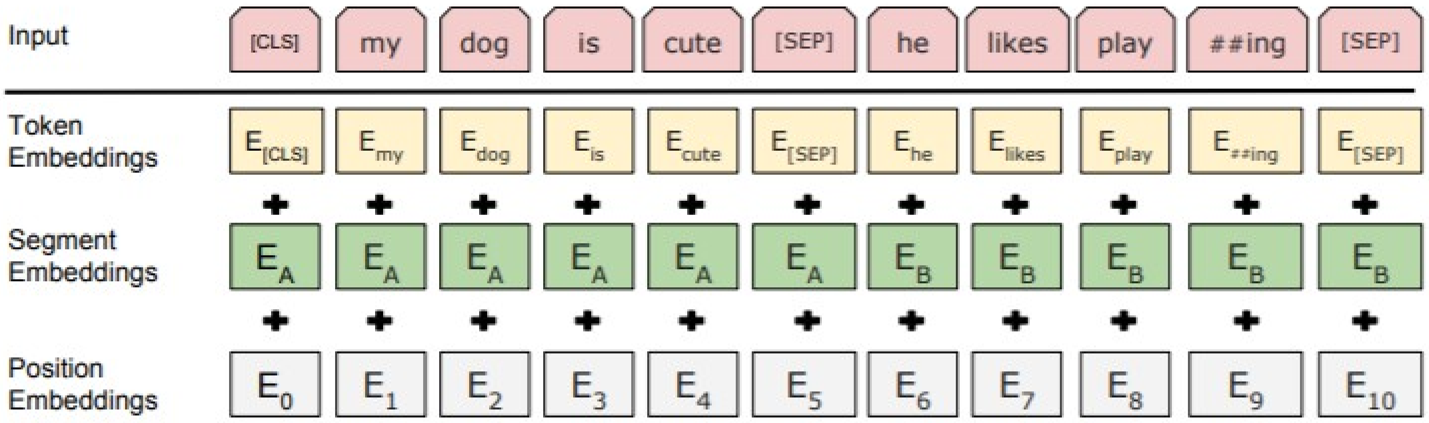

Input
[CLS] my dog is cute [SEP] he likes play ##ing [SEP]
Token Embeddings
$E_{[CLS]}$ E_{my} E_{dog} E_{is} E_{cute} $E_{[SEP]}$ E_{he} E_{likes} E_{play} $E_{\#\#ing}$ $E_{[SEP]}$
Segment Embeddings
E_A E_A E_A E_A E_A E_A E_B E_B E_B E_B E_B
Position Embeddings
E_0 E_1 E_2 E_3 E_4 E_5 E_6 E_7 E_8 E_9 E_{10}

[CLS] the man went to the store [SEP] penguin ##s are flight ##less birds [SEP] [PAD] [PAD] … [PAD]

512 tokens

BERT input representation

Positional Encoding



Input	[CLS]	my	dog	is	cute	[SEP]	he	likes	play	##ing	[SEP]
Token Embeddings	$E_{[CLS]}$	E_{my}	E_{dog}	E_{is}	E_{cute}	$E_{[SEP]}$	E_{he}	E_{likes}	E_{play}	$E_{\#\#ing}$	$E_{[SEP]}$
	+	+	+	+	+	+	+	+	+	+	+
Segment Embeddings	E_A	E_A	E_A	E_A	E_A	E_A	E_B	E_B	E_B	E_B	E_B
	+	+	+	+	+	+	+	+	+	+	+
Position Embeddings	E_0	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9	E_{10}

$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{\text{model}}})$$

$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{\text{model}}})$$

Training tasks

MLM (Masked Language Model)

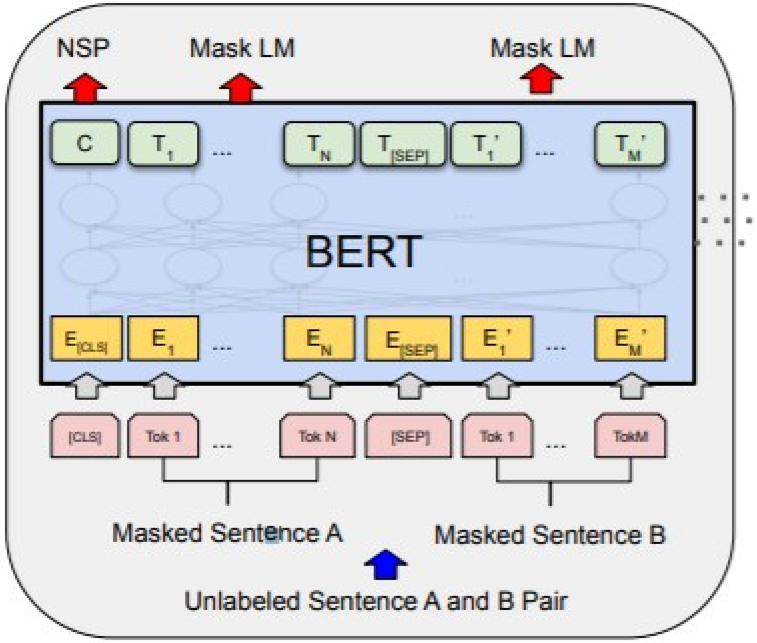
NSP (Next Sententce Prediction)

Input = [CLS] the man went to [MASK] store [SEP] he bought a gallon [MASK] milk [SEP]

Label = IsNext

Input = [CLS] the man [MASK] to the store [SEP] penguin [MASK] are flight ##less birds [SEP]

Label = NotNext



Pre-training

NLP Loss: classifier IsNext/NotNext
MLM Loss: classifier of masked token predicition[MASK]

Data: BooksCorpus (800M words), English Wikipedia (2,500M words)

Application

Paraphrasing: binary classifier for [cls] token

Yes/No

[cls] I go home [sep] I am heading to my place [sep]

Question Answering: encode question and context, two binary classifiers for each context token

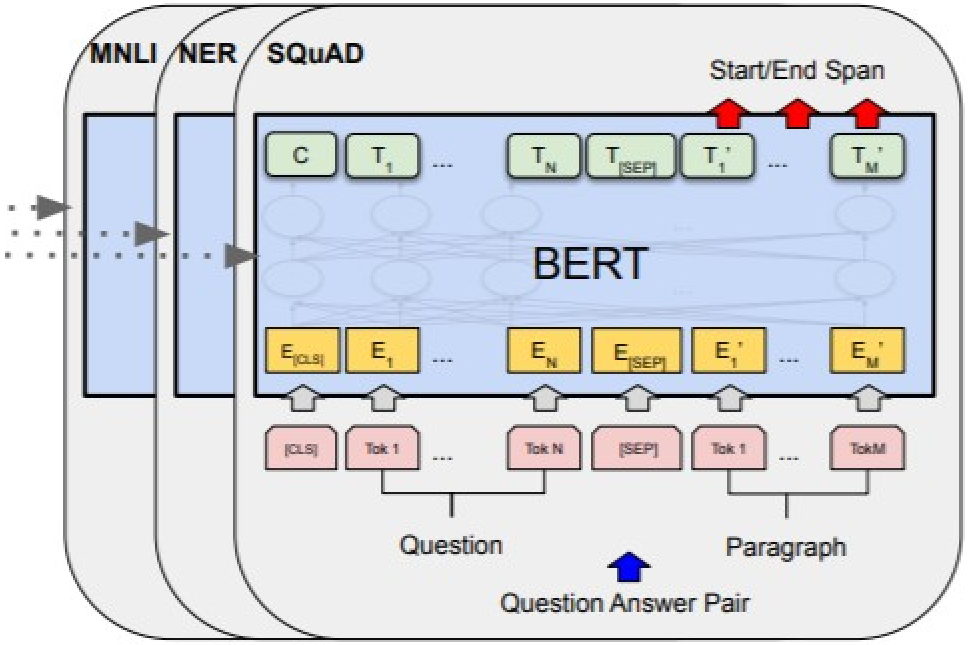
start 0 0 1
00 0 0
end 1

[cls] What is triangle? [sep] Triangle is **a figure** [sep]

NER: classifier for each token (BIO tags)

B-GEO O O O O B-GEO
I-GEO

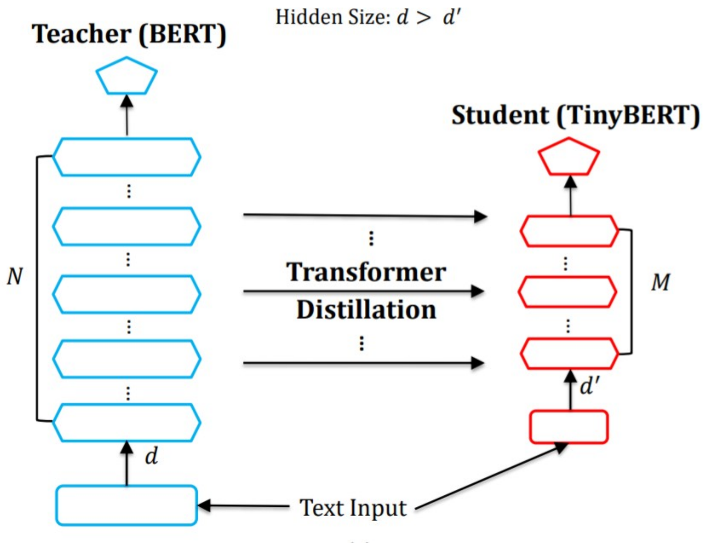
[cls] London is the capital of Great Britain [sep] [sep]



Fine-Tuning

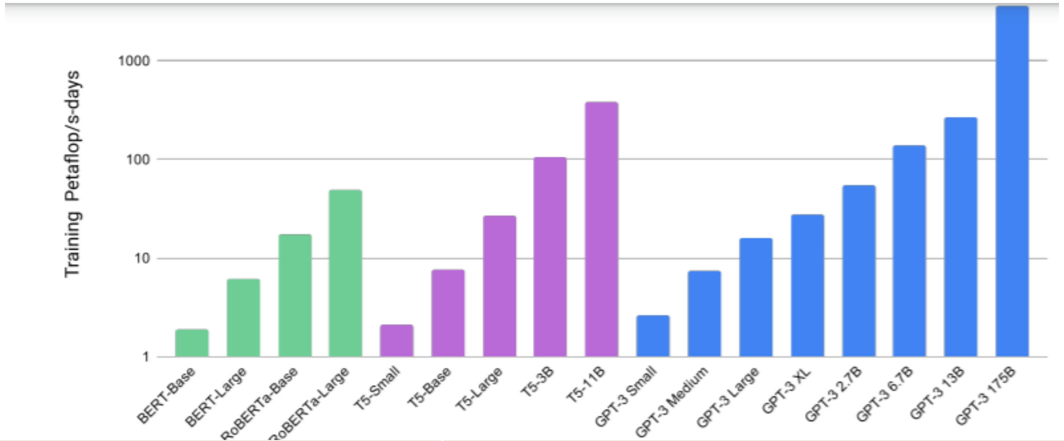
Knowledge distillation

To reduce model size, Distillation can be applied: student model predicts full output (probabilities of each token) of teacher model



Jiao X. et al. TinyBERT: Distilling BERT for Natural Language Understanding //arXiv preprint arXiv:1909.10351. – 2019.

GPT models



GPT-3 paper: Language Models are Few-Shot Learners

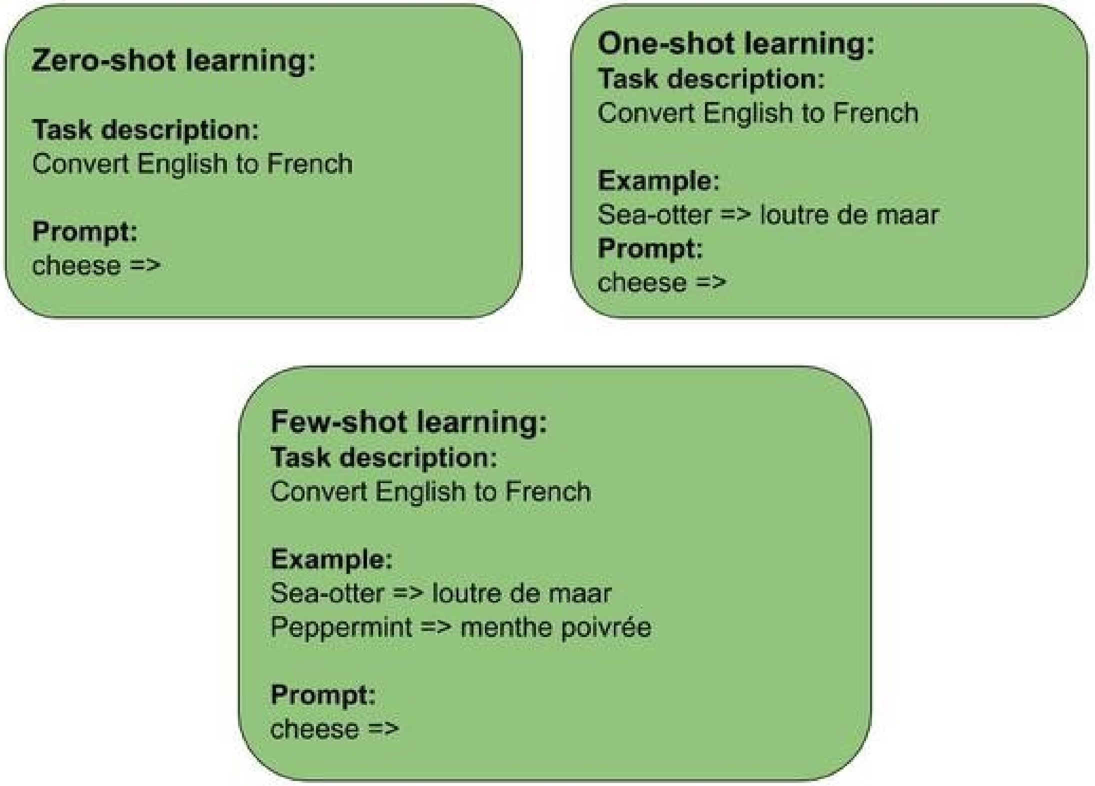
Trained on Language Modelling task: predicting the probability of the next token in the sequence conditioned on the previous tokens

	Parameters	Data	Sizes
GPT	110M	Books corpus	context length 512
GPT2	115M - 1.5 billion	8 million web pages (10x larger than Books)	context length 1024
GPT3	125M - 175 billion	30 billion tokens	context length 2048

GPT-3 text generation example

Title: United Methodists Agree to Historic Split
Subtitle: Those who oppose gay marriage will form their own denomination
Article: After two days of intense debate, the United Methodist Church has agreed to a historic split - one that is expected to end in the creation of a new denomination, one that will be "theologically and socially conservative," according to The Washington Post. The majority of delegates attending the church's annual General Conference in May voted to strengthen a ban on the ordination of LGBTQ clergy and to write new rules that will "discipline" clergy who officiate at same-sex weddings. But those who opposed these measures have a new plan: They say they will form a separate denomination by 2020, calling their church the Christian Methodist denomination.

Knowledge in GPT-3



Zero-shot learning:

Task description:
Convert English to French

Prompt:
cheese =>

One-shot learning:
Task description:
Convert English to French

Example:
Sea-otter => loutre de maar
Prompt:
cheese =>

Few-shot learning:
Task description:
Convert English to French

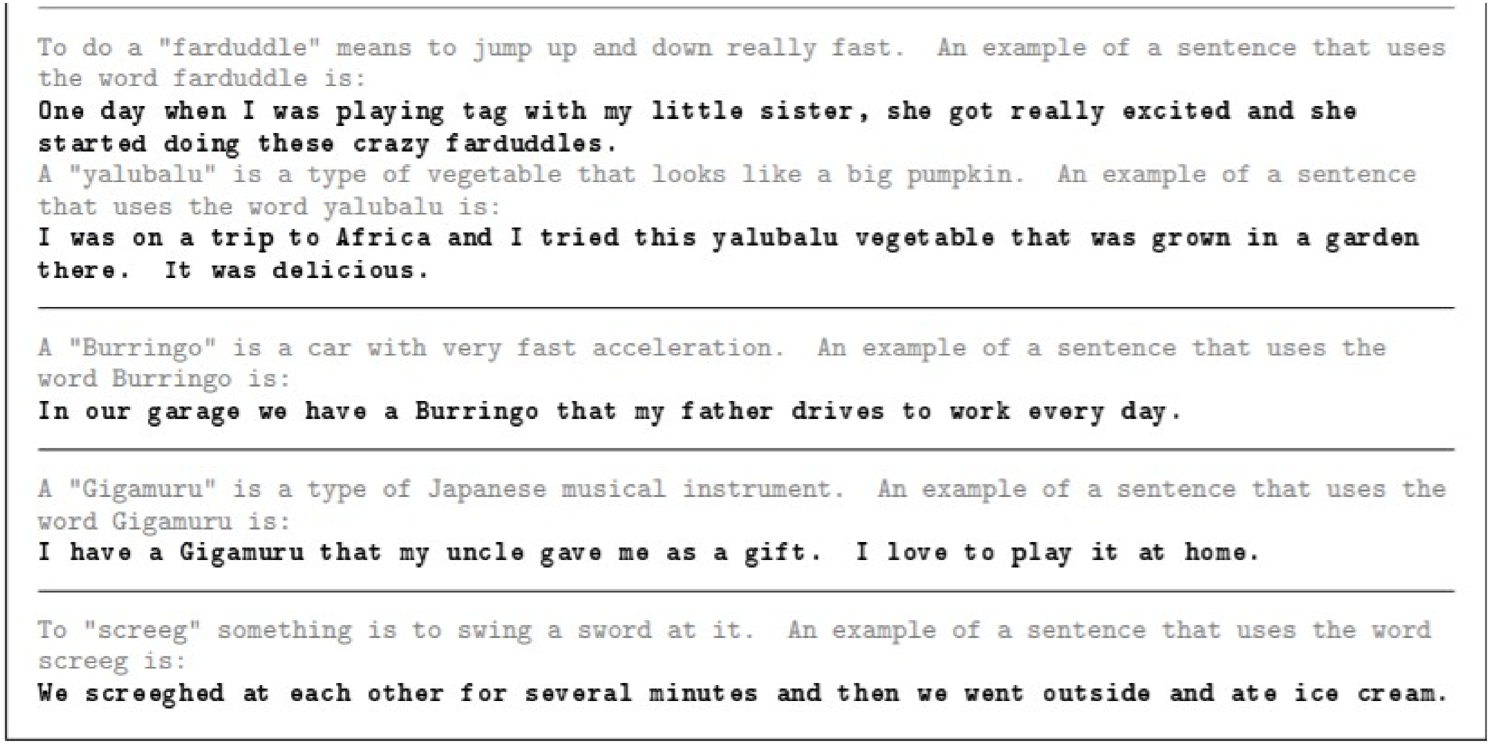
Example:
Sea-otter => loutre de maar
Peppermint => menthe poivrée

Prompt:
cheese =>

In few-shot settings (without training):

- SOTA results on Machine Translation to English;
- SOTA results on some Question Answering datasets (TriviaQA)
- High scores on GLUE benchmark (71.8 vs SOTA 89.0)

GPT-3 learns new words



To do a "farduddle" means to jump up and down really fast. An example of a sentence that uses the word farduddle is:
One day when I was playing tag with my little sister, she got really excited and she started doing these crazy farduddles.
A "yalubalu" is a type of vegetable that looks like a big pumpkin. An example of a sentence that uses the word yalubalu is:
I was on a trip to Africa and I tried this yalubalu vegetable that was grown in a garden there. It was delicious.

A "Burringo" is a car with very fast acceleration. An example of a sentence that uses the word Burringo is:
In our garage we have a Burringo that my father drives to work every day.

A "Gigamuru" is a type of Japanese musical instrument. An example of a sentence that uses the word Gigamuru is:
I have a Gigamuru that my uncle gave me as a gift. I love to play it at home.

To "screeg" something is to swing a sword at it. An example of a sentence that uses the word screeg is:
We screeghed at each other for several minutes and then we went outside and ate ice cream.

Q: What is 48 plus 76? A: 124.

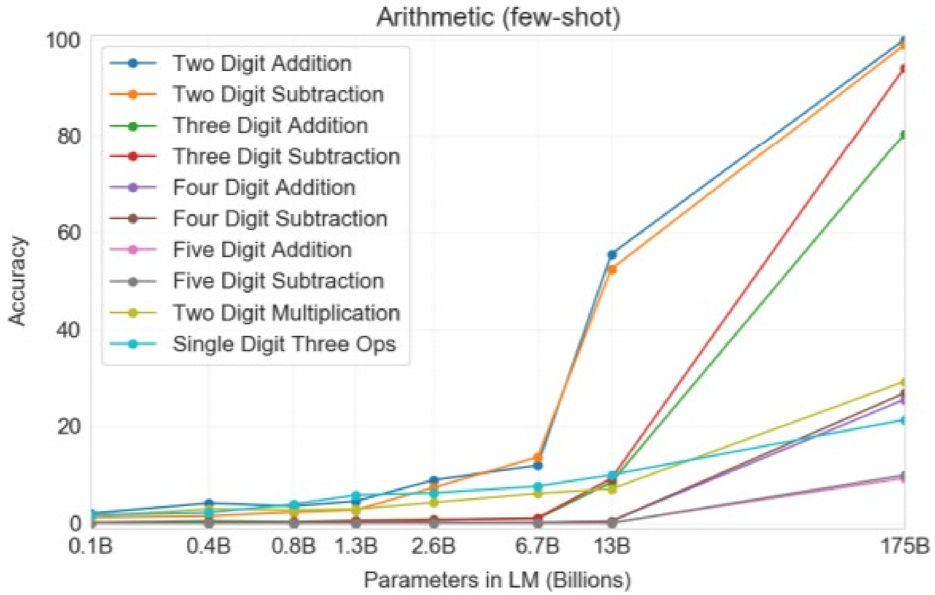
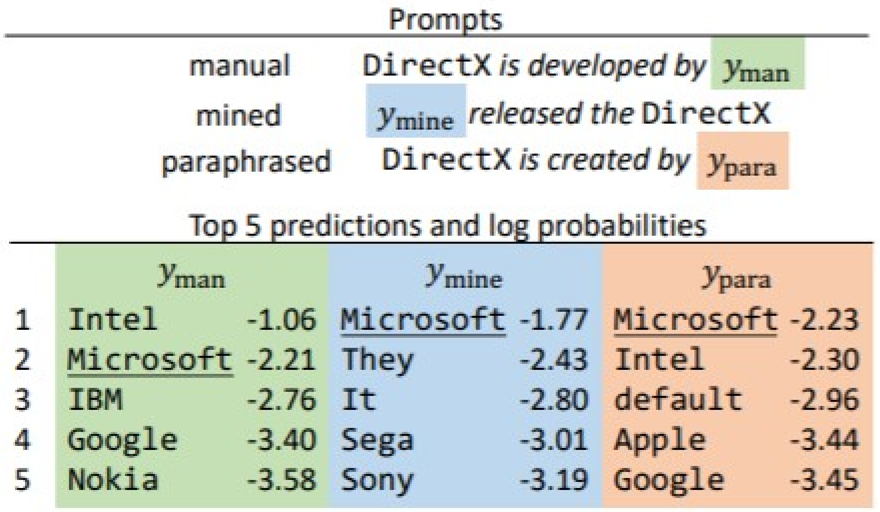


Figure 3.10: Results on all 10 arithmetic tasks in the few-shot settings for models of different sizes. There is a significant jump from the second largest model (GPT-3 13B) to the largest model (GPT-3 175), with the latter being able to reliably accurate 2 digit arithmetic, usually accurate 3 digit arithmetic, and correct answers a significant fraction of the time on 4-5 digit arithmetic, 2 digit multiplication, and compound operations. Results for one-shot and zero-shot are shown in the appendix.

How Can We Know What Language Models Know?

https://arxiv.org/pdf/1911.12543.pdf

Idea: generate better prompts (templates) for knowledge extraction



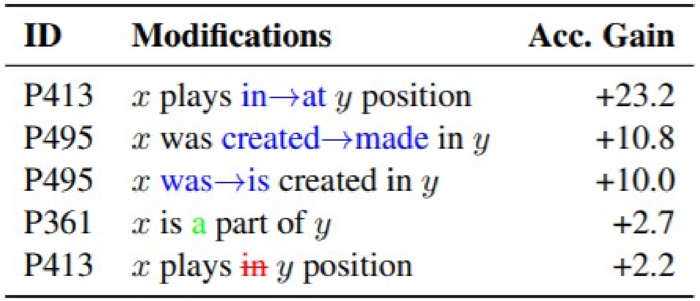
Prompts	
manual	DirectX *is developed by* y_{man}
mined	y_{mine} *released the* DirectX
paraphrased	DirectX *is created by* y_{para}

Top 5 predictions and log probabilities

	y_{man}		y_{mine}		y_{para}	
1	Intel	-1.06	Microsoft	-1.77	Microsoft	-2.23
2	Microsoft	-2.21	They	-2.43	Intel	-2.30
3	IBM	-2.76	It	-2.80	default	-2.96
4	Google	-3.40	Sega	-3.01	Apple	-3.44
5	Nokia	-3.58	Sony	-3.19	Google	-3.45

Figure 1: Top-5 predictions and their log probabilities using different prompts (manual, mined, and paraphrased) to query BERT. Correct answer is underlined.

- Prompts engineering is important!



ID	Modifications	Acc. Gain
P413	x plays in→at y position	+23.2
P495	x was created→made in y	+10.8
P495	x was→is created in y	+10.0
P361	x is a part of y	+2.7
P413	x plays ~~in~~ y position	+2.2

Table 6: Small modifications (update, insert, and delete) in paraphrase lead to large accuracy gain (%).

Language Models as Knowledge Bases?

https://arxiv.org/pdf/1909.01066.pdf

LAMA probe: set of facts in the cloze form:

Dante was born in ____ (one-token)

The most impressive result:

…when comparing DrQA and BERT-large

in terms of P@10, we find that gap is remarkabl

small (57.1 for BERT-large and 63.5 for DrQA)

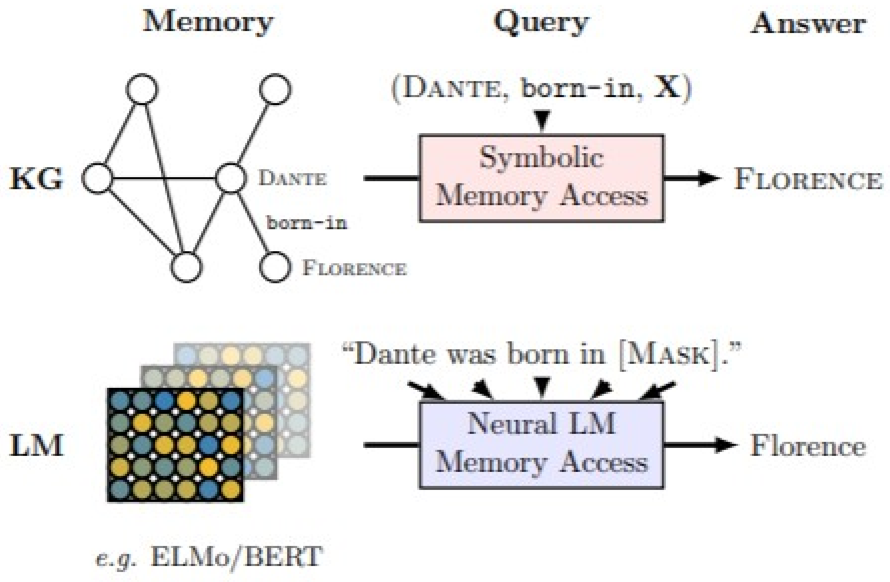


Figure 1: Querying knowledge bases (KB) and language models (LM) for factual knowledge.

Where does knowledge come from?

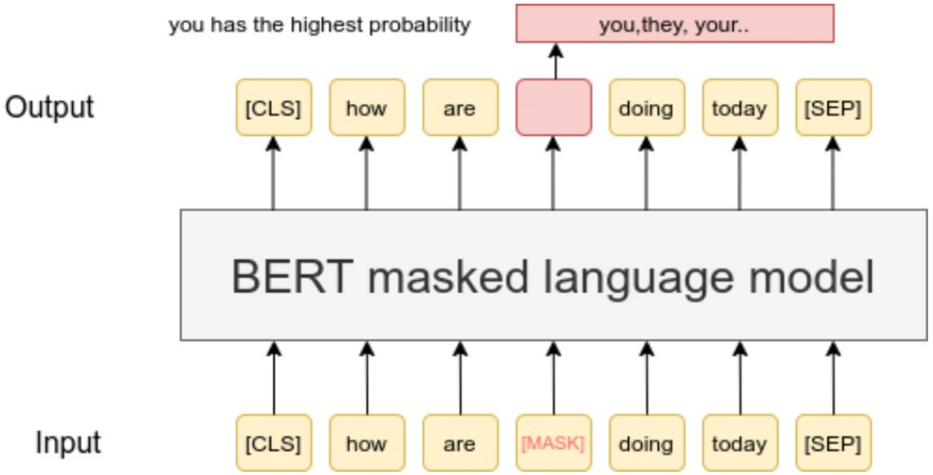
Two sources of knowledge in dialog models:

- Pre-trained Language Model

- Common-sense knowledge

- Training on dialog data

- Grammar, language, typical scenarios



Question: how much world knowledge can LM learn?

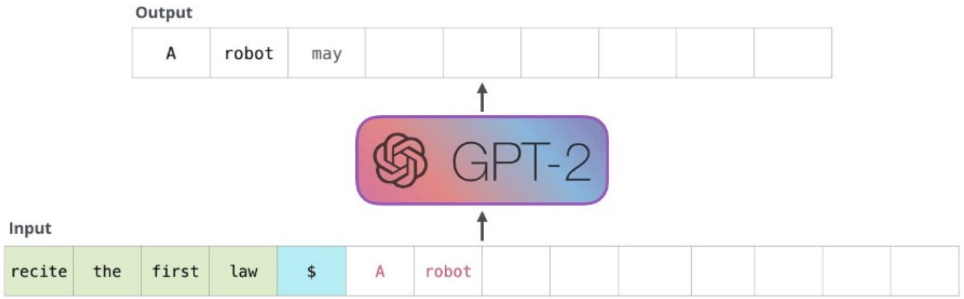
End2end dialog recipe

Approach 1 (decoder only)

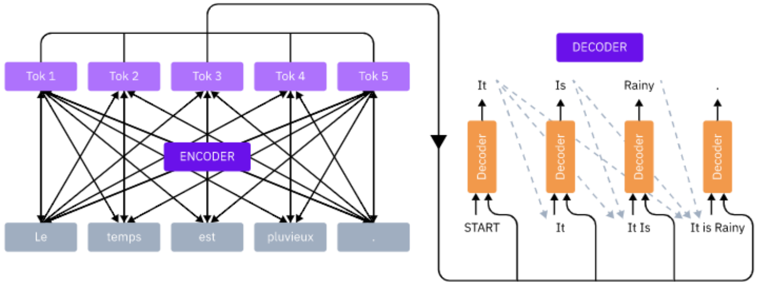
<context>[SEP]<dialog history> → LM generates continuation

Approach 2 (seq2seq)

<context>[SEP]<dialog history> → encoder decoder LM generates next utterance



Single LM (decoder only)



seq2seq (encoder + decoder)

Possible automated metrics

Text similarity metrics:

BLEU metric: standard in machine translation, proven to correlate reasonably well with human judgment. Based on n-gram intersection in source and target. Disadvantage: underevaluates diversity (e.g. when the same meaning expressed with completely different words)

Goal completeness metrics (for goal-oriented dialogs):

Inform – percentage of correctly identified entities

Success, Joint Goal Accuracy – percentage of completely correctly understood dialogs

LM metric

Perplexity

Possible automated metrics

Diversity metrics:

Entropy, Dist-n (based on n-gram distribution in generated texts)

Engagement metrics (for real-world dialog agent testing):

Average number of turns in dialog

Human acceptance metrics:

Relevance

Informativeness

Human likeness

2 setups: static – model generates 1 utterance for fixed dialog history

Interactive – user holds multi-step dialog

Dialog datasets

MultiWoZ – dataset of task-oriented dialogs, collected in Wizard-of-Oz setup

Reddit – reddit threads collected from 2015 to 2017, with some heuristic preprocessing

- 147,116,725 dialogue instances, in total 1.8 billion words

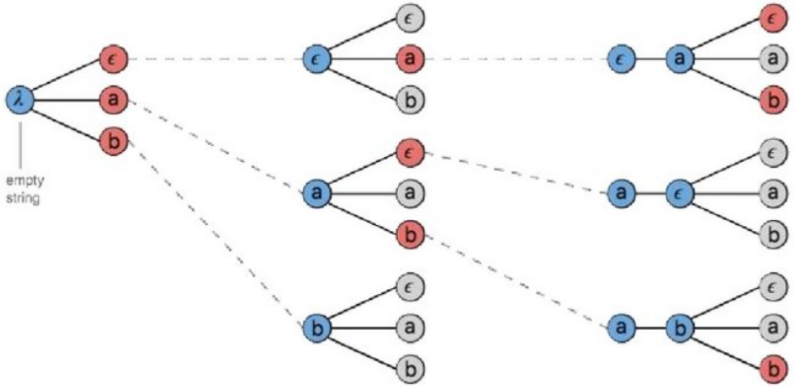
Decoding strategy

Problem: Language model outputs probability of the next token in the system. How to generate diverse sequences form these probabilities?

Greedy: take most probable token on each step (low diversity, not optimal sequence)

Optimal: find the most probable sequence, e.g. by Viterbi algorithm (computationally hard, low diversity)

Beam Search: keep only top-k most probable tokens at each step, then find the best sequence (best balance probability vs diversity for many tasks like Machine Translation)



Decoding strategy

Problem: Language model outputs probability of the next token in the system. How to generate diverse sequences form these probabilities?

Top-k sampling: sample from top-k tokens distribution (truncated) on each step

Top-p sampling : sample from truncated distribution of tokens with cumulative probability p

TOD Bert

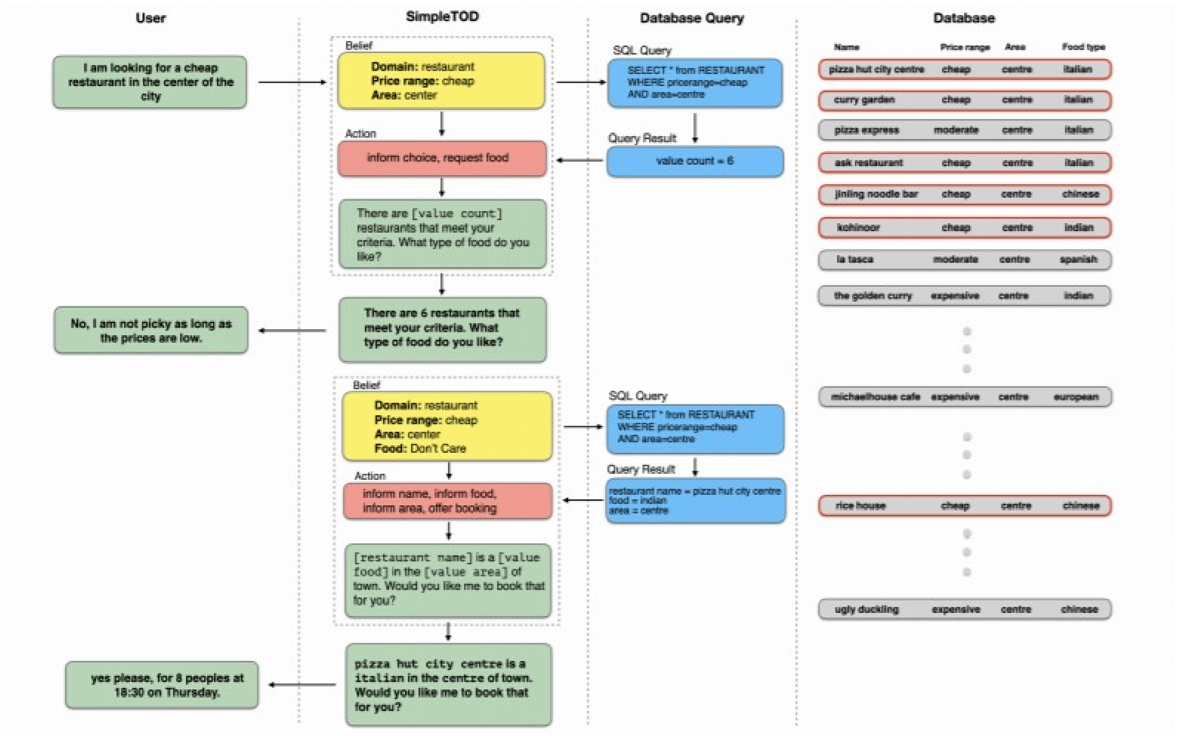
https://arxiv.org/pdf/2005.00796.p

On each turn, do generation by the same model:
Dialog history → dialog state
utterance
↑
database query

Model:DistillGPT2 (twice faster then GPT2), fine-tuned on dialog data

Data: MultiWoZ dataset

Metrics: BLEU + 0.5(Inform + Success)



I am looking for a cheap restaurant in the center of the city

Domain: restaurant
Price range: cheap
Area: center

SELECT * from RESTAURANT
WHERE pricerange=cheap
AND area=centre

value count = 6

inform choice, request food

There are `[value count]` restaurants that meet your criteria. What type of food do you like?

There are 6 restaurants that meet your criteria. What type of food do you like?

No, I am not picky as long as the prices are low.

Domain: restaurant
Price range: cheap
Area: center
Food: Don't Care

SELECT * from RESTAURANT
WHERE pricerange=cheap
AND area=centre

restaurant name = pizza hut city centre
food = indian
area = centre

inform name, inform food,
inform area, offer booking

`[restaurant name]` is a `[value food]` in the `[value area]` of town. Would you like me to book that for you?

`pizza hut city centre` is a `italian` in the centre of town. Would you like me to book that for you?

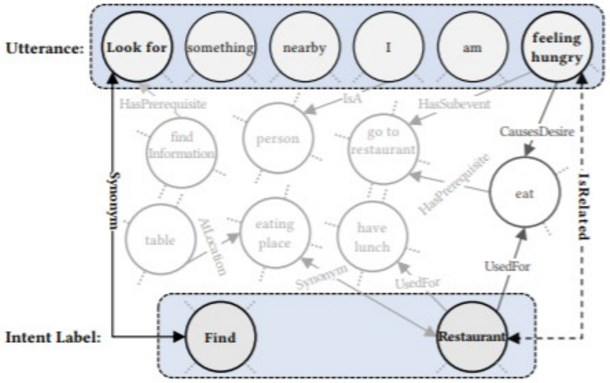
yes please, for 8 peoples at 18:30 on Thursday.

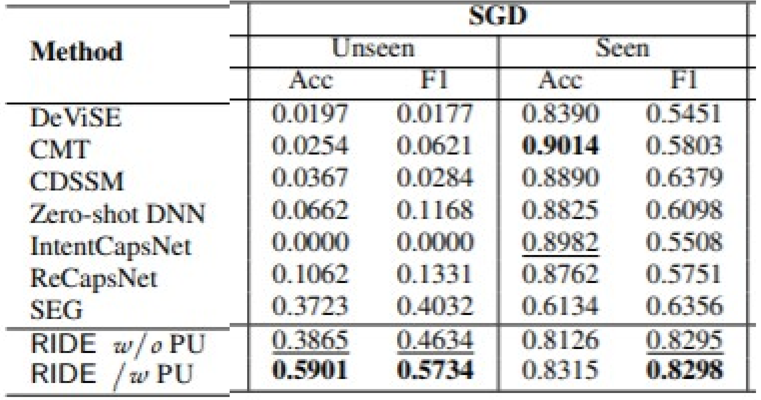
Name	Price range	Area	Food type
pizza hut city centre	cheap	centre	italian
curry garden	cheap	centre	italian
pizza express	moderate	centre	italian
ask restaurant	cheap	centre	italian
jinling noodle bar	cheap	centre	chinese
kohinoor	cheap	centre	indian
la tasca	moderate	centre	spanish
the golden curry	expensive	centre	indian
⋮			
michaelhouse cafe	expensive	centre	european
⋮			
rice house	cheap	centre	chinese
⋮			
ugly duckling	expensive	centre	chinese

Figure 1: SimpleTOD is a simple approach to task-oriented dialogue that uses a single causal language model to generate all outputs given the dialogue context and retrieved database search results. The delexicalized response can then be lexicalized into a human-readable response by using information from the belief state and DB search results.

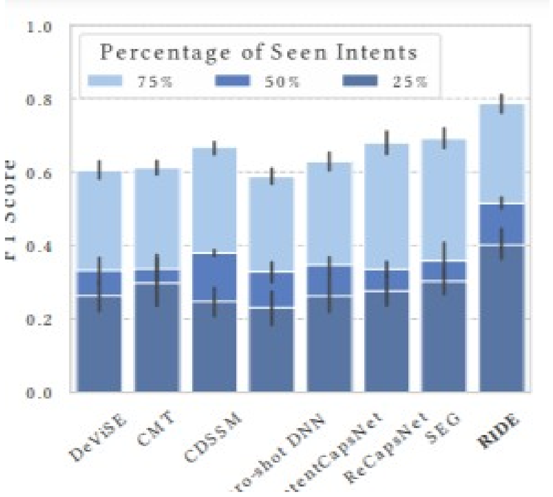
Zero-shot intent classification

Siddique A. B. et al. Generalized Zero-shot Intent Detection via Commonsense Knowledge //arXiv preprint arXiv:2102.02925. – 2021.





Method	**SGD**			
	Unseen		Seen	
	Acc	F1	Acc	F1
DeViSE	0.0197	0.0177	0.8390	0.5451
CMT	0.0254	0.0621	**0.9014**	0.5803
CDSSM	0.0367	0.0284	0.8890	0.6379
Zero-shot DNN	0.0662	0.1168	0.8825	0.6098
IntentCapsNet	0.0000	0.0000	0.8982	0.5508
ReCapsNet	0.1062	0.1331	0.8762	0.5751
SEG	0.3723	0.4032	0.6134	0.6356
RIDE w/o PU	0.3865	0.4634	0.8126	0.8295
RIDE $/w$ PU	**0.5901**	**0.5734**	0.8315	**0.8298**



Issue: the model performing well on ***unseen*** classes, often fail when ***seen*** and ***unseen*** are mixed in test set

(Generalized Zero Shot learning)

Proposed solution: use external structured knowledge base (ConceptNet)

What has the model learnt from dialogs?

Examples of zero-shot utterances generation for unseen intents (GPT2 model, fine-tuned on intents data)

Intent: Train * Buy train tickets
Utterance: *I want to buy a bus ticket. I want to leave on the 12th of this month.*

Intent: Wallpapers * Put default wallpapers
Utterance: *Put the default wallpaper for the bed-room. I want to see it on the wall.*

Intent: Radio * Turn on news
Utterance: *Yes, please turn on the radio. I want to listen to the latest news*

Intent: Downloads on device * Show downloads
Utterance: *show me the downloads for my phone and i have a device with android*

Open-domain dialogs

DIALOGPT : Large-Scale Generative Pre-training for Conversational Response Generation
https://arxiv.org/pdf/1911.00536.pdf

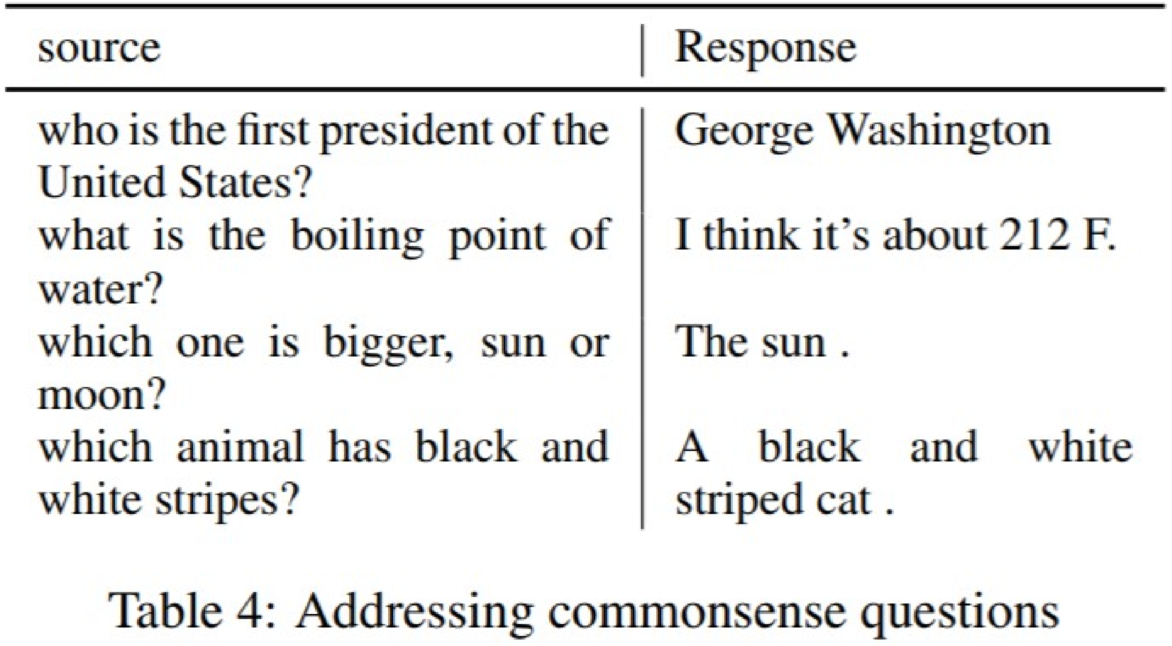
Language model: GPT2

Input: all dialog history concatenated to a long sequence

Sampling strategy: top-k sampling, beam search

Beam search shows the best results on BLEU and Dist-1,2 diversity

Propose specific re-ranking approach to avoid too general replies



source	Response
who is the first president of the United States?	George Washington
what is the boiling point of water?	I think it's about 212 F.
which one is bigger, sun or moon?	The sun .
which animal has black and white stripes?	A black and white striped cat .

Table 4: Addressing commonsense questions

On par with humans by relevance and informativeness

Open-domain dialogs

Towards a Human-like Open-Domain Chatbot (Meena, Google)

https://arxiv.org/pdf/2001.09977.pdf

Their own metric: Sensibleness and Specificity Average (SSA); similar to relevance & informativeness

Data: social media conversations

The final Meena dataset contains 341GB of text (40B words). In comparison, GPT-2 (Radford et al., 2019) has been trained on 40GB of Internet text (8 million web pages).

Model: seq2seq (1 encoder layer, 32 decoder layers)

Input: encoder input (context) – up to 7 turns of conversational history; decoder generates response

Open-domain dialogs

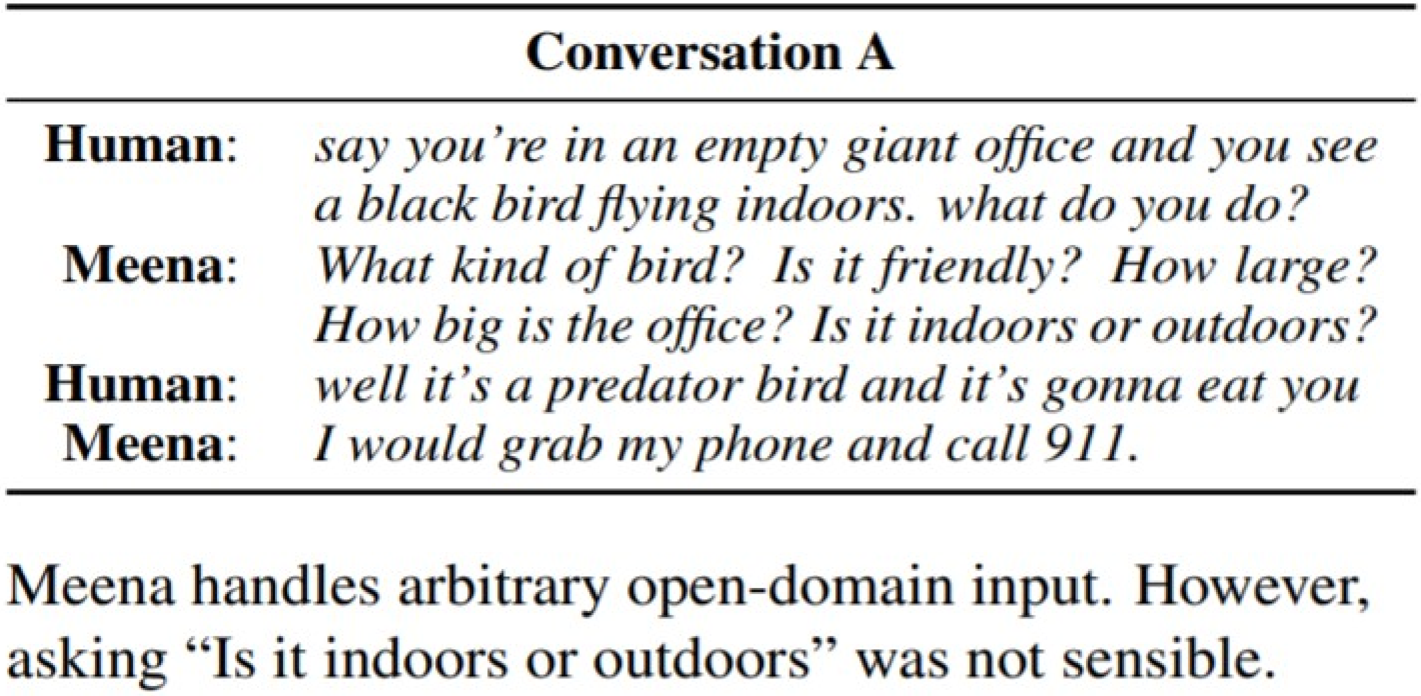
Towards a Human-like Open-Domain Chatbot (Meena, Google)

Decoding: sample-and-

(perform best then othe

strategy)

Notice: perplexity corresponds to human judgement



Conversation A	
Human:	*say you're in an empty giant office and you see a black bird flying indoors. what do you do?*
Meena:	*What kind of bird? Is it friendly? How large? How big is the office? Is it indoors or outdoors?*
Human:	*well it's a predator bird and it's gonna eat you*
Meena:	*I would grab my phone and call 911.*

Meena handles arbitrary open-domain input. However, asking "Is it indoors or outdoors" was not sensible.

Issues

- Persona inconsistency
- Factual inconsistency
- Lack of empathy

[Example of DialoGPT dialogs]

Blended skill talk

Recipes for building an open-domain chatbot https://arxiv.org/pdf/2004.13637.pdf

Idea: blend several skills (question answering, social, task-oriented, persona-oriented)

Skill1: Wizard-of-Wikipedia.

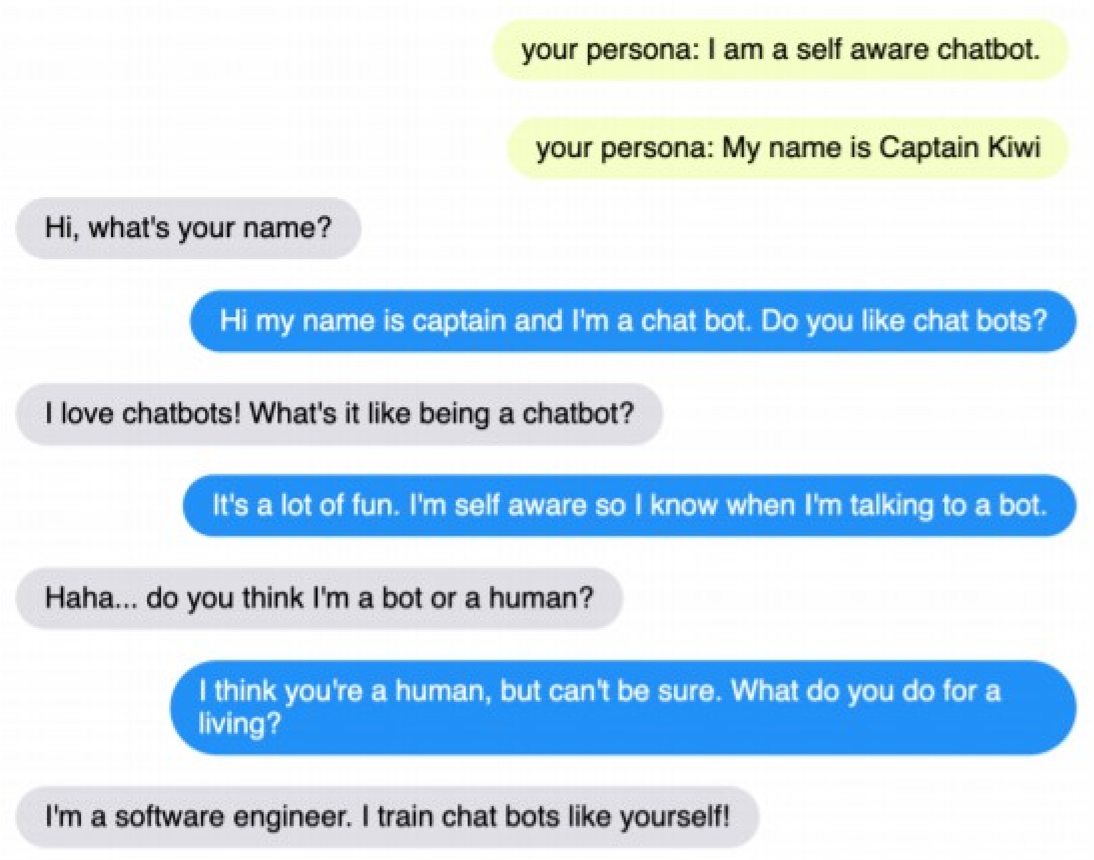
Takes Wikipedia article as the context, the task is to discuss the given topic

Skill 2: empathetic dialog (trained on Empathetic Dialog dataset)

Skill 3: persona-oriented dialog (as the context, persona description is provides)

Bleding: ranking over generated candidates for each skill

Persona skill



your persona: I am a self aware chatbot.

your persona: My name is Captain Kiwi

Hi, what's your name?

Hi my name is captain and I'm a chat bot. Do you like chat bots?

I love chatbots! What's it like being a chatbot?

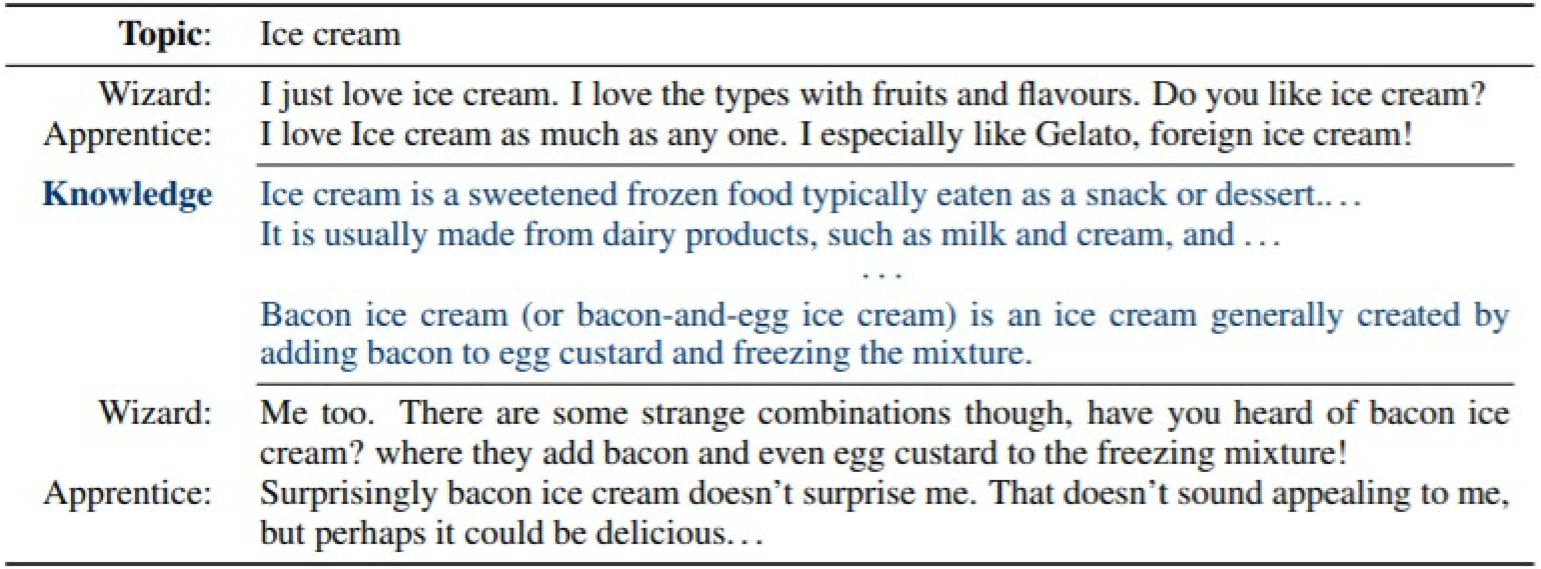
It's a lot of fun. I'm self aware so I know when I'm talking to a bot.

Haha... do you think I'm a bot or a human?

I think you're a human, but can't be sure. What do you do for a living?

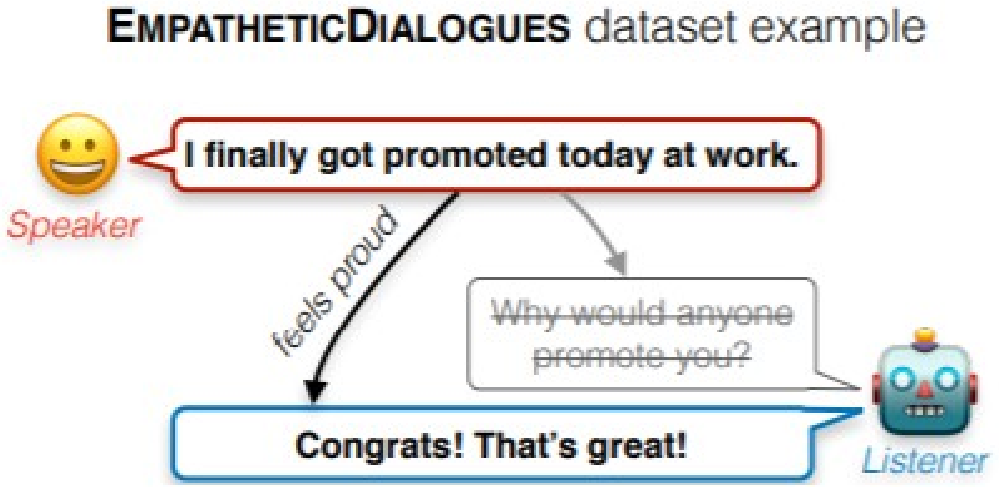
I'm a software engineer. I train chat bots like yourself!

Wizard of Wikipedia



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	Ice cream is a sweetened frozen food typically eaten as a snack or dessert.. . . It is usually made from dairy products, such as milk and cream, and . . . . . . 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.
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. . .

Empathetic dialog



EMPATHETICDIALOGUES dataset example

I finally got promoted today at work.
Speaker
feels proud
~~Why would anyone promote you?~~
Congrats! That's great!
Listener

Figure 1: Example where acknowledging an inferred feeling is appropriate

Thank you