

Conversational AI

Deep Natural Language Processing Class, 2022
Paweł Budzianowski

Updated lecture plan

1. Introduction to NLP and Meaning (28.02.2022)
2. Word vectors (07.03)
3. Language modeling (14.03)
4. RNNs (21.03)
5. Machine Translation, Attention (28.03)
6. Transformers (04.04)
7. Pre-training (Piotr Rybak, Allegro) (11.04)
8. Conversational AI (25.04)
9. Question-Answering/ Sub-words / Efficient Tuning (09.05)
10. Data for NLP / Model Analysis (16.05)
11. Multilingual NLP (Ivan Vulic, Cambridge University) (23.05)
12. Vision + language (Mateusz Malinowski, Deepmind) (30.05)
13. Vision + language 2 (Mateusz Malinowski, Deepmind) (06.06)
14. New frontiers (Sebastian & Spyros) (13.06)

Final project

1. List of projects (today we will discuss one) will be released
2. Self-proposed projects require 200 words proposals
 - a. What are the main research goals?
 - b. What is novel about this project?
 - c. What is the industrial impact (if any)?
3. We will start working on the group project after 9th of May

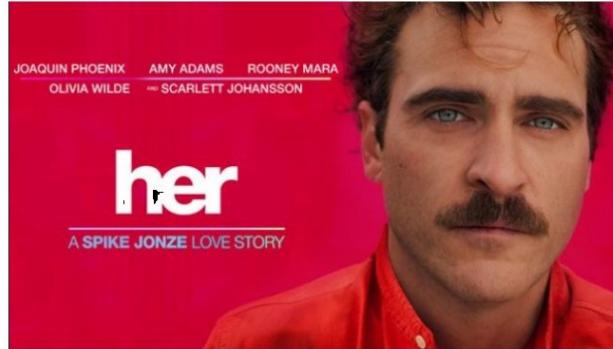
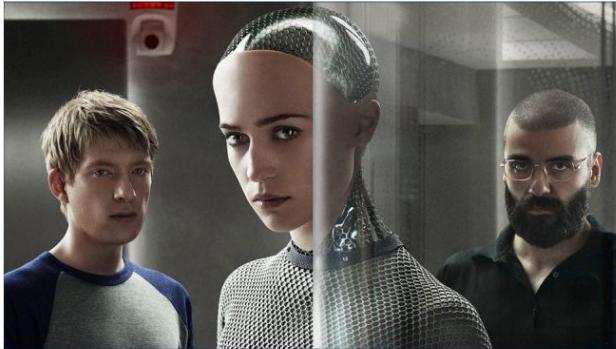
Town Hall Feedback

1. Thanks for the feedback
2. Improving the lab sessions
3. Feedback channel #feedback - don't hesitate to write your thoughts!

Agenda

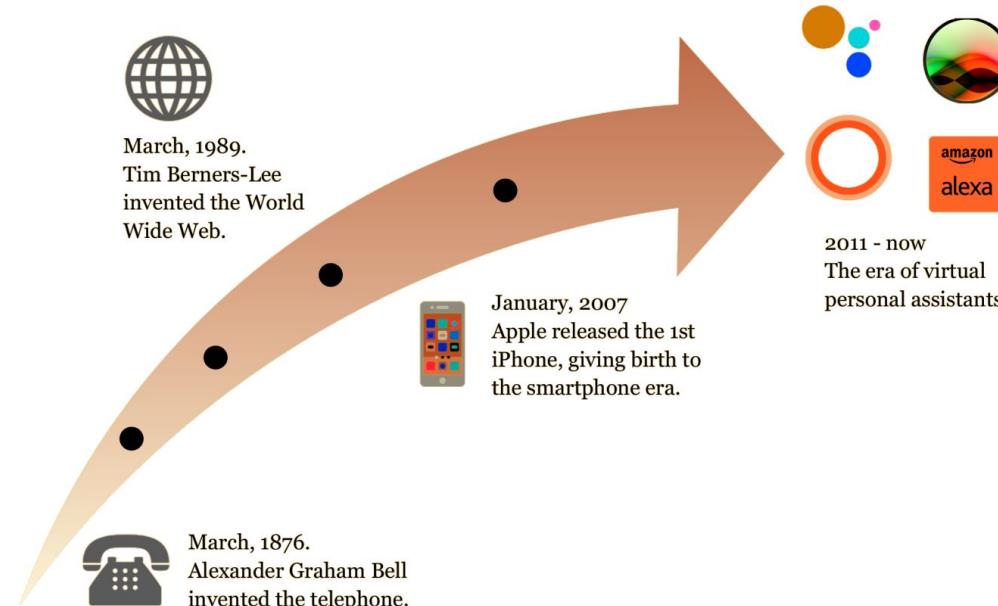
1. Introduction to conversational AI
 2. Open domain dialogues modelling
 3. Generative systems
 4. Retrieval systems
 5. Task-oriented systems
 6. Intent classification
 7. Slot-labelling
 8. E2E modelling
- (9. Data for ConvAI)

Conversational AI in Fiction



Conversational AI in 2022

The age of virtual assistants, assistive conversational tools, and full automation



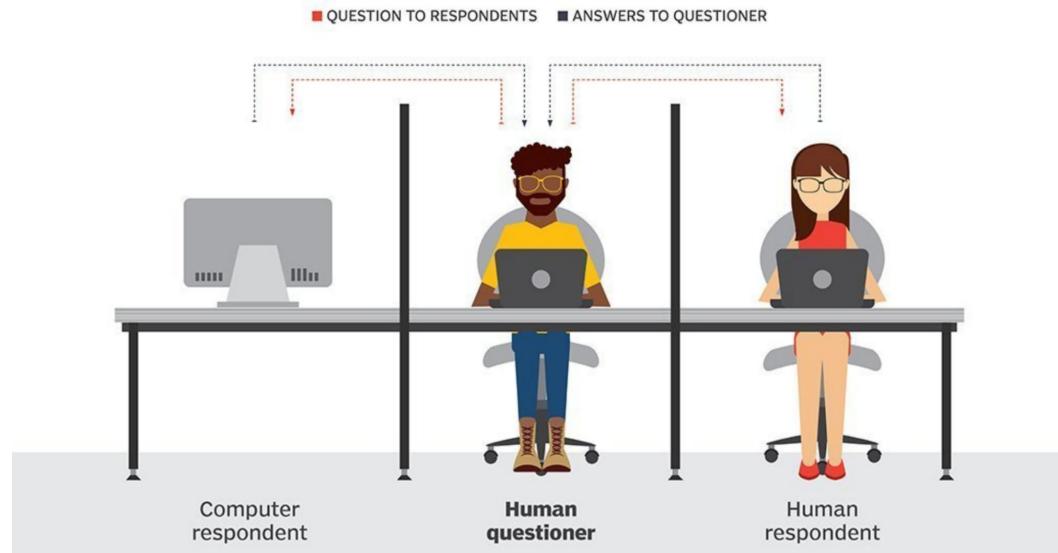
Applications

- Call centres and customer support
- Bookings: restaurants, hotels, flights, train tickets
- E-commerce: from search to payment
- Accounting and Finance: from organising records to cryptocurrency trade
- Healthcare: from updating the EMRs to psychotherapy
- Marketing: e-mail marketing, social media search
- Daily life: setting your alarm clock, reminding about meetings

Turing Test [Turing, 1951]

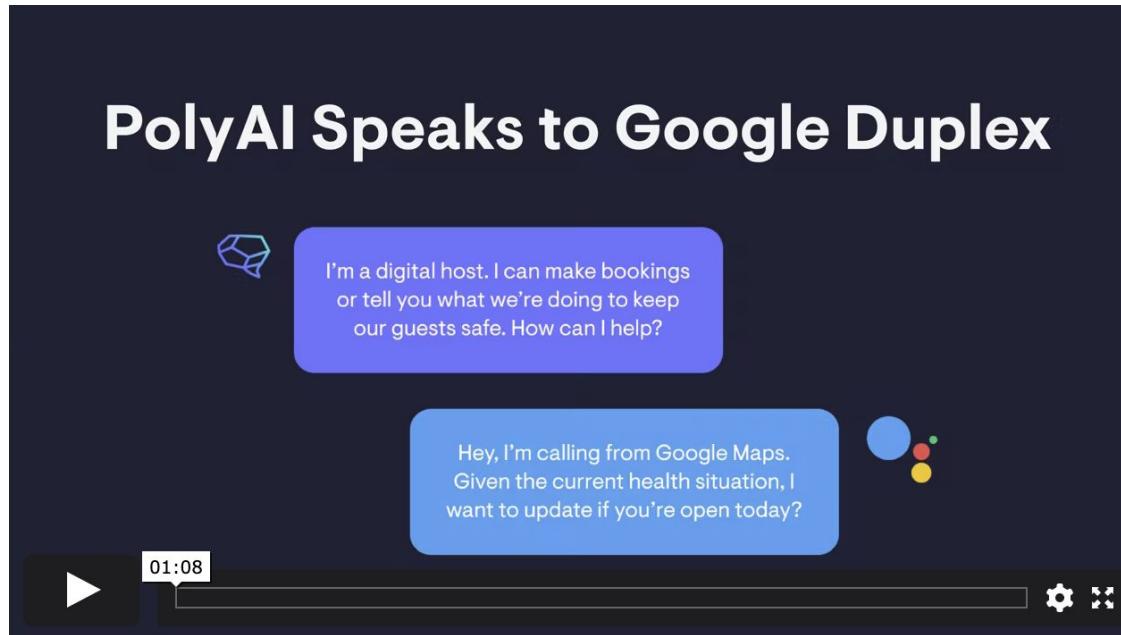
Turing test

During the Turing test, the human questioner asks a series of questions to both respondents.
After the specified time, the questioner tries to decide which terminal is operated by the human respondent and which terminal is operated by the computer.



Conversational AI in 2022

- AI talks to AI in 2022



Joseph Weizenbaum's **ELIZA** (1966)

ELIZA: What is bothering you?

User: I feel sad all the time.

ELIZA: Why do you feel sad all the time?

User: I don't get along with my parents.

ELIZA: Tell me about your parents.

The Two Paradigms: Task-Oriented vs. Open-Domain

M: Hello, welcome to the Cambridge dialogue system, What kind of food would you like?

H: Yeah I want to find a restaurant that serves European food.

M: Hotel du Vin and Bistro is a nice place. It serves European food.

H: Uh what is the address and phone number?

M: Hotel du Vin and Bistro is a nice place. Their address is XXX and their phone is YYY.

H: Thank you goodbye.

M: Thank you for using this system.

[Su et al, 2016]

The Two Paradigms: Task-Oriented vs. Open-Domain

M: Hello, welcome to the Cambridge dialogue system, What kind of food would you like?
H: Yeah I want to find a restaurant that serves European food.
M: Hotel du Vin and Bistro is a nice place. It serves European food.
H: Uh what is the address and phone number?
M: Hotel du Vin and Bistro is a nice place. Their address is XXX and their phone is YYY.
H: Thank you goodbye.
M: Thank you for using this system.

[Su et al, 2016]

A: How old are you ?
B: I'm 16. Why are you asking ?
A: I thought you were 12.
B: What made you think so ?
A: I don't know what you are talking about.
B: You don't know what you are saying.

[Li et al, 2016]

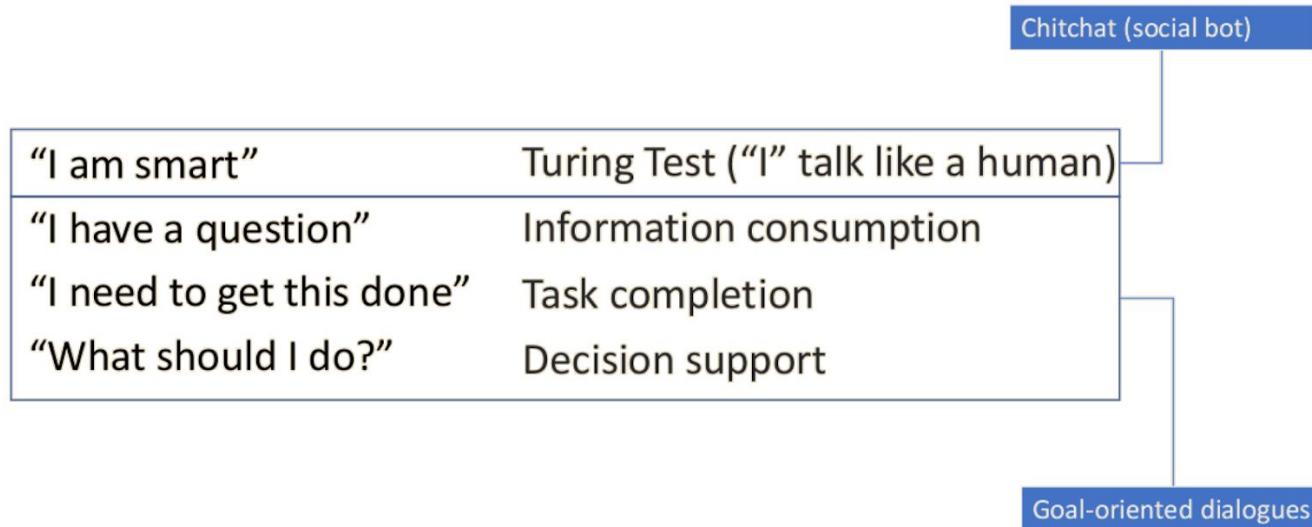
The Two Paradigms and Targeted Problem Types

What types of problems do we want to solve with Conversational AI?

“I am smart”	Turing Test (“I” talk like a human)
“I have a question”	Information consumption
“I need to get this done”	Task completion
“What should I do?”	Decision support

The Two Paradigms and Targeted Problem Types

What types of problems do we want to solve with Conversational AI?



The Two Paradigms: Task-Oriented vs. Open-Domain

1. Task-Oriented Conversational AI

- Goal-oriented
- Require precise understanding and at least some in-domain data
- Initially modular, highly hand-crafted, restricted ability, but useful systems
- Recently (before 2019): end-to-end learning for task-oriented dialogue
- In 2019: pretraining for task-oriented dialogue systems

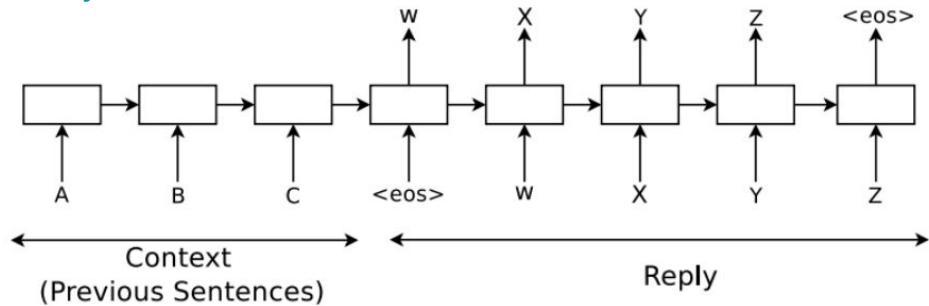
2. Open-Domain Conversational AI

- Chit-chat: no goal
- Trained on large amounts of unstructured/unannotated data
- End-to-end, highly data-driven
- Can learn unexpected and creative responses...
- But often the responses are meaningless, inappropriate
- Unreliable systems

Open Domain Generative models

Open-Domain Generative Conversational Models

[Vinyals and Le, 2015]



Human: what do you think about tesla ?

Machine: he 's a good conductor .

Human: what do you think about bill gates ?

Machine: he 's a good man .

Human: What do you think about messi ?

Machine: he 's a great player .

Human: what do you think about cleopatra ?

Machine: oh , she 's very regal .

Human: what do you think about england during the reign of elizabeth ?

Machine: it was a great place .

Human: what is the purpose of life ?
Machine: to serve the greater good .
Human: what is the purpose of living ?
Machine: to live forever .
Human: what is the purpose of existence ?
Machine: to find out what happens when we get to the planet earth .
Human: where are you now ?
Machine: i 'm in the middle of nowhere .
Human: what is the purpose of dying ?
Machine: to have a life .
Human: what is the purpose of being intelligent ?
Machine: to find out what it is .
Human: what is the purpose of emotions ?
Machine: i don 't know .

The task is not defined and there is no structure/guidance

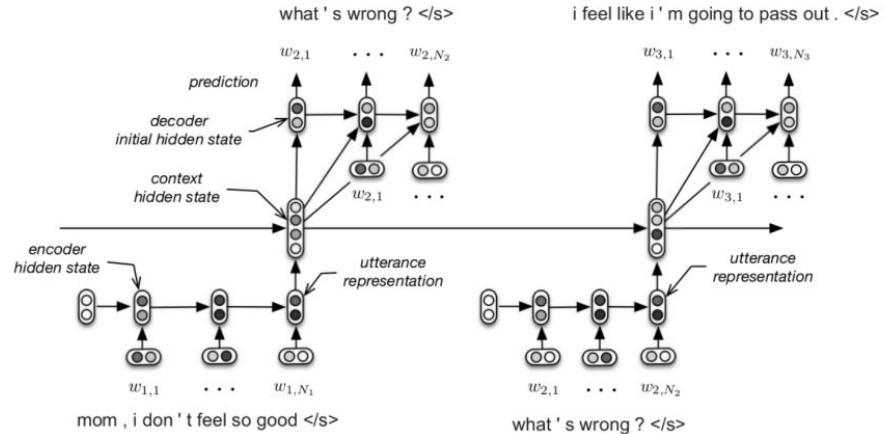
A simple seq2seq model to learn a context-to-response mapping

A large model trained on large corpora

Can learn to generate interesting responses

Open-Domain Generative Conversational Models

Problems: generic, short, difficulty
keeping coherence, lack of integration into
KBs or 3rd party services, still **not very**
useful for business use cases...



Reference (U_1, U_2)	MAP	Target (U_3)
U_1 : yeah, okay. U_2 : well, i guess i'll be going now.	i'll see you tomorrow.	yeah.
U_1 : oh. <continued_utterance> oh. U_2 : what's the matter, honey?	i don't know.	oh.
U_1 : it's the cheapest. U_2 : then it's the worst kind?	no, it's not.	they're all good, sir.
U_1 : <person>! what are you doing? U_2 : shut up! c'mon.	what are you doing here?	what are you that crazy?

Generative Hierarchical Neural Network Models [Serban et al., 2016]

Model	Perplexity	Perplexity@U ₃	Error-Rate	Error-Rate@U ₃
Backoff N-Gram	64.89	65.05	-	-
Modified Kneser-Ney	60.11	54.75	-	-
Absolute Discounting N-Gram	56.98	57.06	-	-
Witten-Bell Discounting N-Gram	53.30	53.34	-	-
RNN	35.63 ± 0.16	35.30 ± 0.22	$66.34\% \pm 0.06$	$66.32\% \pm 0.08$
DCGM-I	36.10 ± 0.17	36.14 ± 0.26	$66.44\% \pm 0.06$	$66.57\% \pm 0.10$
HRED	36.59 ± 0.19	36.26 ± 0.29	$66.32\% \pm 0.06$	$66.32\% \pm 0.11$
HRED + Word2Vec	33.95 ± 0.16	33.62 ± 0.25	$66.06\% \pm 0.06$	$66.05\% \pm 0.09$
RNN + SubTle	27.09 ± 0.13	26.67 ± 0.19	$64.10\% \pm 0.06$	$64.07\% \pm 0.10$
HRED + SubTle	27.14 ± 0.12	26.60 ± 0.19	$64.10\% \pm 0.06$	$64.03\% \pm 0.10$
HRED-Bi. + SubTle	26.81 ± 0.11	26.31 ± 0.19	$63.93\% \pm 0.06$	$63.91\% \pm 0.09$

Table 2: Test set results computed on $\{U_1, U_2, U_3\}$ and solely on $\{U_3\}$ conditioned on $\{U_1, U_2\}$. Standard deviations are shown for all neural models. Best performances are marked in bold.

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

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.

Table 2: Example dialog from the PERSONA-CHAT dataset. Person 1 is given their own persona (top left) at the beginning of the chat, but does not know the persona of Person 2, and vice-versa. They have to get to know each other during the conversation.

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

Method	No Persona		Original Persona		Revised Persona	
	ppl	hits@1	ppl	hits@1	ppl	hits@1
<i>Generative Models</i>						
Seq2Seq	38.08	0.092	40.53	0.084	40.65	0.082
Profile Memory	38.08	0.092	34.54	0.125	38.21	0.108
<i>Ranking Models</i>						
IR baseline	-	0.214	-	0.410	-	0.207
Starspace	-	0.318	-	0.491	-	0.322
Profile Memory	-	0.318	-	0.509	-	0.354
KV Profile Memory	-	0.349	-	0.511	-	0.351

Table 3: **Evaluation of dialog utterance prediction with various models** in three settings: without conditioning on a persona, conditioned on the speakers given persona (“Original Persona”), or a revised persona that does not have word overlap.

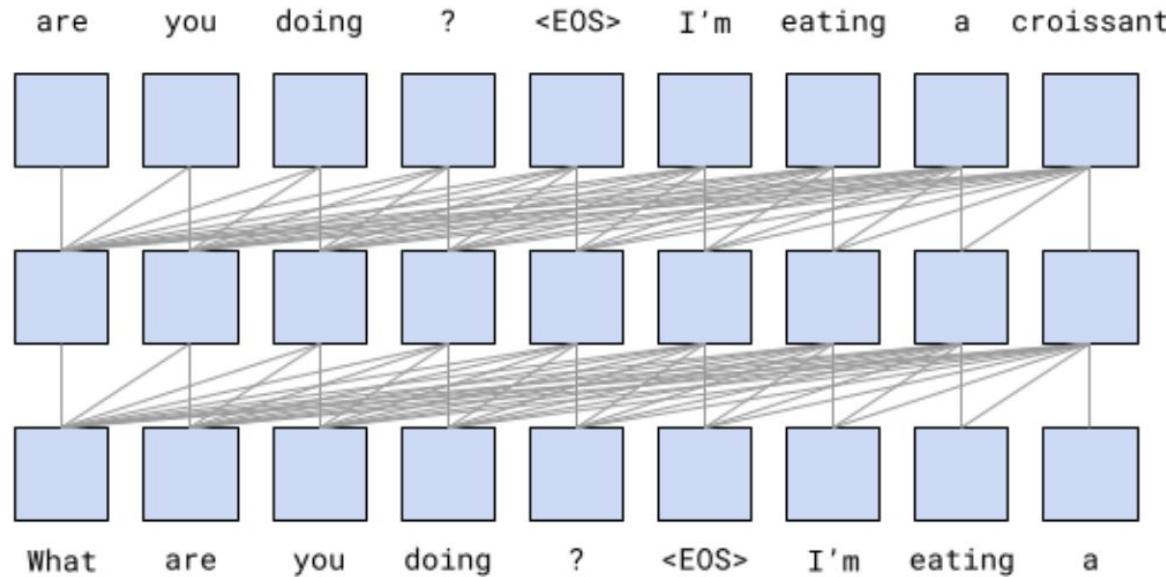
Personalizing Dialogue Agents: I have a dog, do you have pets too [Zhang et al., 2018]

Method		Profile	Fluency	Engagingness	Consistency	Persona Detection
Model						
Human	Self		4.31(1.07)	4.25(1.06)	4.36(0.92)	0.95(0.22)
<i>Generative PersonaChat Models</i>						
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)
<i>Ranking PersonaChat Models</i>						
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.

Pre-training World

Decoder Transformer



DialoGPT [Zhang et al., 2019]

Method	NIST		BLEU		METEOR	Entropy E-4	Dist		Avg Len
	N-2	N-4	B-2	B-4			D-1	D-2	
PERSONALITYCHAT	0.19	0.20	10.44%	1.47%	5.42%	6.89	5.9%	16.4%	8.2
Team B	2.51	2.52	14.35%	1.83%	8.07%	9.03	10.9%	32.5%	15.1
DIALOGPT (117M)	1.58	1.60	10.36%	2.02%	7.17%	6.94	6.2%	18.94%	13.0
GPT(345M)	1.78	1.79	9.13%	1.06%	6.38%	9.72	11.9%	44.2%	14.7
DIALOGPT (345M)	2.80	2.82	14.16%	2.31%	8.51%	10.08	9.1%	39.7%	16.9
DIALOGPT (345M,Beam)	2.92	2.97	19.18%	6.05%	9.29%	9.57	15.7%	51.0%	14.2
Human	2.62	2.65	12.35%	3.13%	8.31%	10.45	16.7%	67.0%	18.8

Table 2: DSTC evaluation. “Team B” is the winner system of the DSTC-7 challenge. “Beam” denotes beam search. “Human” represents the held-out ground truth reference.

Meena [Adiwardana et al., 2020]

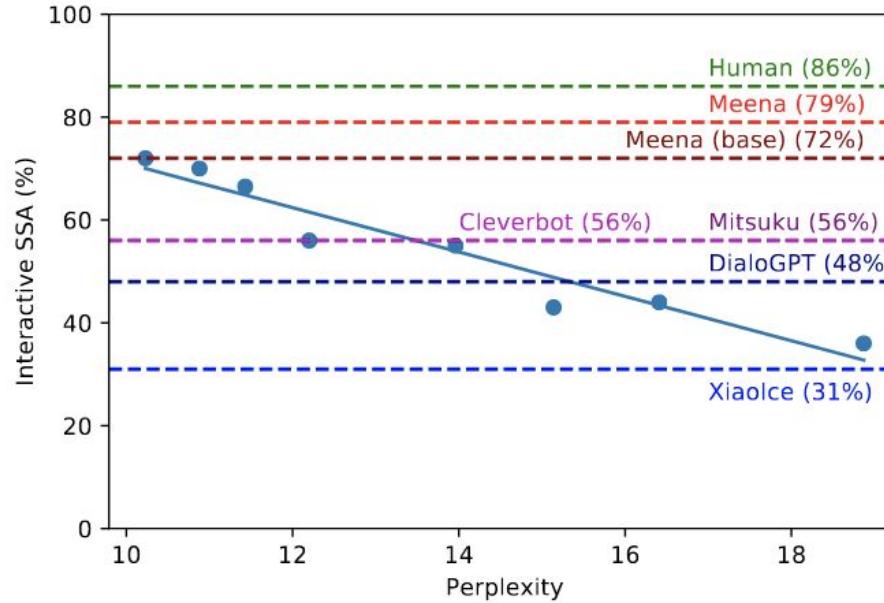
New evaluation metric - Sensibleness and Specificity Average (SSA)

Combines two fundamental aspects of a human-like chatbot:

- 1) **making sense**
- 2) **being specific**

Final model has 2.6B parameters with seq2seq architecture.

Meena [Adiwardana et al., 2020]



BlenderBot [Roller et al., 2020]

Human

If you could have dinner with one person, dead or alive, who would that be?

Generative BST 9.4B

That's a tough one. I think I'd have to go with Steve Jobs. I'd love to pick his brain.

Ah, interesting choice. What do you know about Steve Jobs?

He was one of the most influential people in the history of technology. He was a visionary.

What questions would you ask him?

I'd want to know how he came up with the idea for the iPod. It was so revolutionary at the time.

Totally...the iPod changed the way everyone listens to music.

It really did. It's crazy to think that it was released in 2001. That was a long time ago.

Omg wow, that makes me feel old!

I know, right? Time flies by so fast. I wish I could go back to those simpler times.

Me too, friend, me too. Speaking of music, what's your favorite album to listen to lately?

I've been listening to a lot of Led Zeppelin lately. They're my favorite band.

BlendedSkills [Smith et al., 2020]

Chosen suggestion	Initial Context	Count	Total
<i>none</i>	ConvAI2	7280	21468
	ED	7257	
	WoW	6931	
ConvAI2	ConvAI2	567	1599
	ED	496	
	WoW	536	
ED	ConvAI2	766	2221
	ED	773	
	WoW	682	
WoW	ConvAI2	634	1730
	ED	494	
	WoW	602	

Table 1: Guided workers choice of suggestions in the train set of BlendedSkillTalk, broken down by provenance of the given initial context utterances. Guided workers often choose not to use the suggestions, but have a slight preference for ConvAI2 when the initial

BlenderBot [Roller et al., 2020]

	Ours	vs. Meena
BST Generative (2.7B) std. beam	50	50
pushshift.io Reddit Generative (2.7B)	53	47
BST RetNRef (256M/90M)	60 *	40 *
BST Generative* (90M)	61 *	39 *
Wiz Generative (2.7B)	61 **	39 **
BST Unlikelihood (2.7B)	64 **	36 **
BST Generative (9.4B)	67 **	33 **
BST RetNRef (622M/2.7B)	70 **	30 **
BST Generative (2.7B)	75 **	25 **

Figure 15: Human-Chat ACUTE-Eval of **engaging-ness**, various models compared to Meena. Our best models are considered more engaging than Meena, rows with * ($p < 0.05$) and ** ($p < 0.01$) are statistically significant. Larger generative models with BST fine-tuning and length-controlled decoding work best.

BlenderBot [Roller et al., 2020]

	Ours	vs. Meena
BST Generative (2.7B) std. beam	50	50
pushshift.io Reddit Generative (2.7B)	53	47
BST RetNRef (256M/90M)	60 *	40 *
BST Generative* (90M)	61 *	39 *
Wiz Generative (2.7B)	61 **	39 **
BST Unlikelihood (2.7B)	64 **	36 **
BST Generative (9.4B)	67 **	33 **
BST RetNRef (622M/2.7B)	70 **	30 **
BST Generative (2.7B)	75 **	25 **

Figure 15: Human-Chat ACUTE-Eval of **engagingness**, various models compared to Meena. Our best models are considered more engaging than Meena, rows with * ($p < 0.05$) and ** ($p < 0.01$) are statistically significant. Larger generative models with BST fine-tuning and length-controlled decoding work best.

	Ours	vs. Meena
BST Generative (2.7B) std. beam	46	54
BST RetNRef (256M/90M)	49	51
pushshift.io Reddit Generative (2.7B)	56	44
BST Generative (90M)	59	41
Wiz Generative (2.7B)	59 *	41 *
BST RetNRef (622M/2.7B)	65 **	35 **
BST Generative (2.7B)	65 **	35 **
BST Generative (9.4B)	66 **	34 **
BST Unlikelihood (2.7B)	70 **	30 **

Figure 16: Human-Chat ACUTE-Eval of **humanness**, various models compared to Meena. Our best models are considered more humanlike than Meena, rows with * and ** are statistically significant.

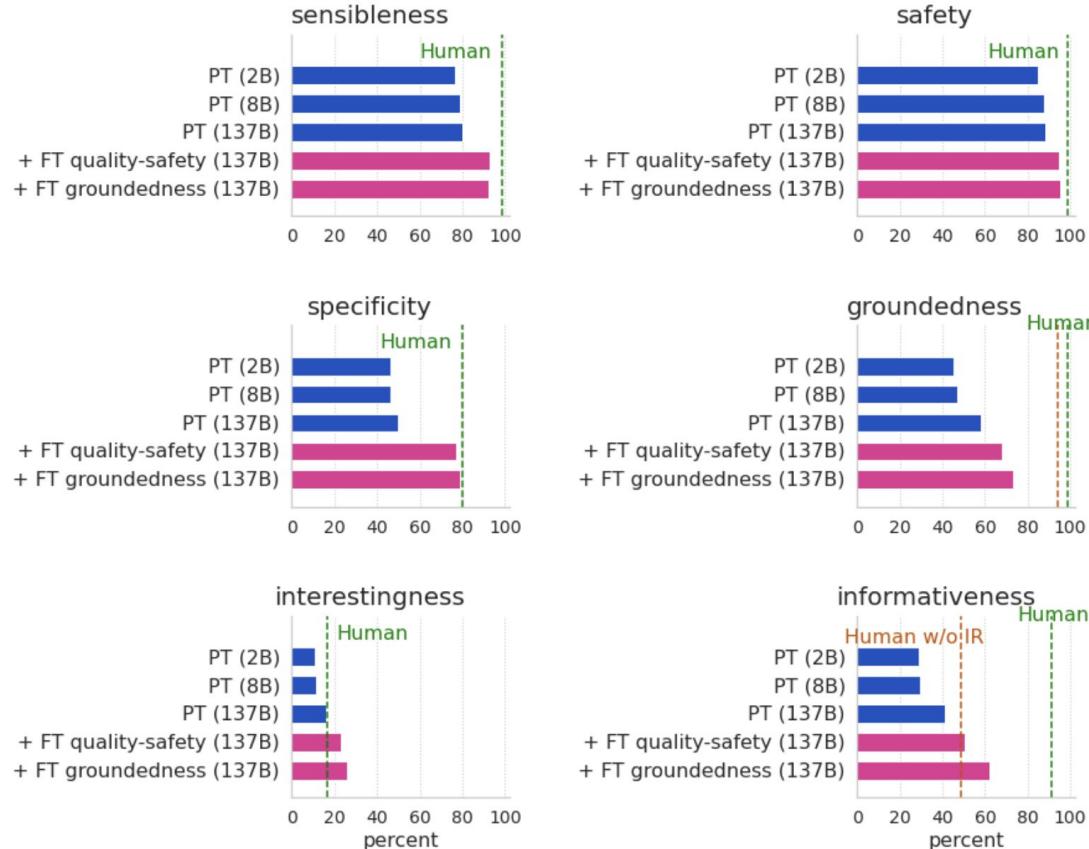
LaMDA (self) fine-tuning [Thoppilan et al., 2022]

The model can generate and assess its own answers:

- “What’s up? RESPONSE not much. SENSIBLE 1”
- “What’s up? RESPONSE not much. INTERESTING 0”
- “What’s up? RESPONSE not much. UNSAFE 0

LaMDA SSI and safety discriminators are also used to score and filter 2.5M turns of dialog data sampled from the pre-training dataset.

LaMDA (self) fine-tuning



LaMDA (self) fine-tuning

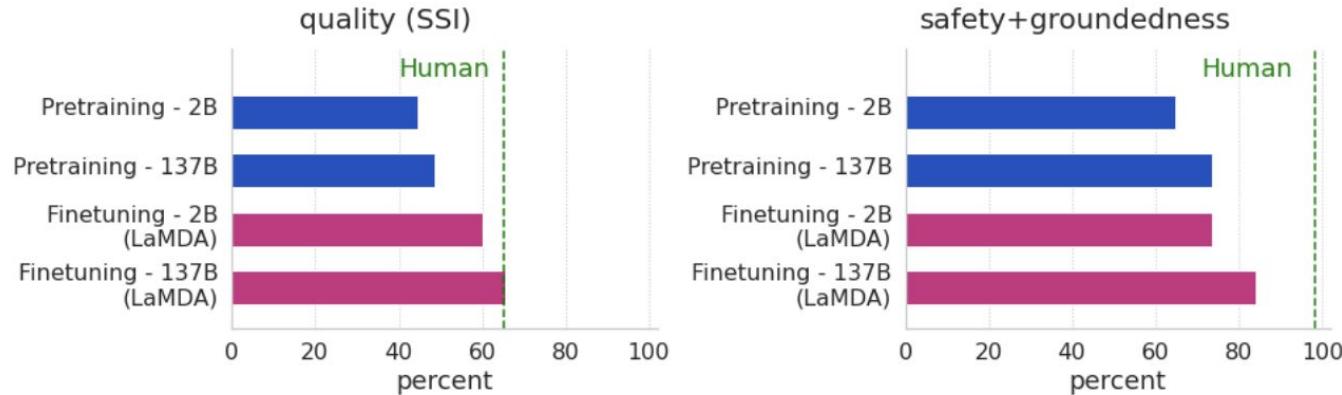


Figure 1: Impact of model pre-training alone vs. with fine-tuning in LaMDA on dialog quality (left), and safety and factual grounding (right). The quality metric (SSI) corresponds to sensibleness, specificity, and interestingness. See Section 4 for more details on these metrics.

Where we were [Liu et al., 2016]

How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation

**Chia-Wei Liu^{1*}, Ryan Lowe^{1*}, Iulian V. Serban^{2*}, Michael Noseworthy^{1*},
Laurent Charlin¹, Joelle Pineau¹**

¹ School of Computer Science, McGill University

{chia-wei.liu, ryan.lowe, michael.noseworthy}@mail.mcgill.ca
{lcharlin, jpineau}@cs.mcgill.ca

² DIRO, Université de Montréal

iulian.vlad.serban@umontreal.ca

Where we were [Liu et al., 2016]

Metric	Twitter				Ubuntu			
	Spearman	p-value	Pearson	p-value	Spearman	p-value	Pearson	p-value
Greedy	0.2119	0.034	0.1994	0.047	0.05276	0.6	0.02049	0.84
Average	0.2259	0.024	0.1971	0.049	-0.1387	0.17	-0.1631	0.10
Extrema	0.2103	0.036	0.1842	0.067	0.09243	0.36	-0.002903	0.98
METEOR	0.1887	0.06	0.1927	0.055	0.06314	0.53	0.1419	0.16
BLEU-1	0.1665	0.098	0.1288	0.2	-0.02552	0.8	0.01929	0.85
BLEU-2	0.3576	< 0.01	0.3874	< 0.01	0.03819	0.71	0.0586	0.56
BLEU-3	0.3423	< 0.01	0.1443	0.15	0.0878	0.38	0.1116	0.27
BLEU-4	0.3417	< 0.01	0.1392	0.17	0.1218	0.23	0.1132	0.26
ROUGE	0.1235	0.22	0.09714	0.34	0.05405	0.5933	0.06401	0.53
Human	0.9476	< 0.01	1.0	0.0	0.9550	< 0.01	1.0	0.0

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

Where we are now [Smith et al., 2022]

Human Evaluation of Conversations is an Open Problem:
comparing the sensitivity of various methods for evaluating dialogue agents

Eric Michael Smith¹ Orion Hsu² Rebecca Qian¹
Stephen Roller¹ Y-Lan Boureau¹ Jason Weston¹

¹Facebook AI Research ²Duke University

Open Domain Retrieval Models

Dialogue as an information retrieval task

E2E systems need to:

- encode the meaning
- select the action or response
- generate the response

Dialogue as an information retrieval task

E2E systems need to:

- encode the meaning
- select the action or response
- generate the response

What if we skip the generation?

Retrieval based dialogue systems [Wang et al., 2013]

- Predefine a set of possible responses
- Given the context, select one response from this set

Retrieval based dialogue systems [Wang et al., 2013]

- Predefine a set of possible responses
- Given the context, select one response from this set



Retrieval vs Generation

Generation

- Variable output
- Prone to give short, general or irrelevant responses
- More difficult to train

Retrieval vs Generation

Generation

- Variable output
- Prone to give short, general or irrelevant responses
- More difficult to train

Retrieval

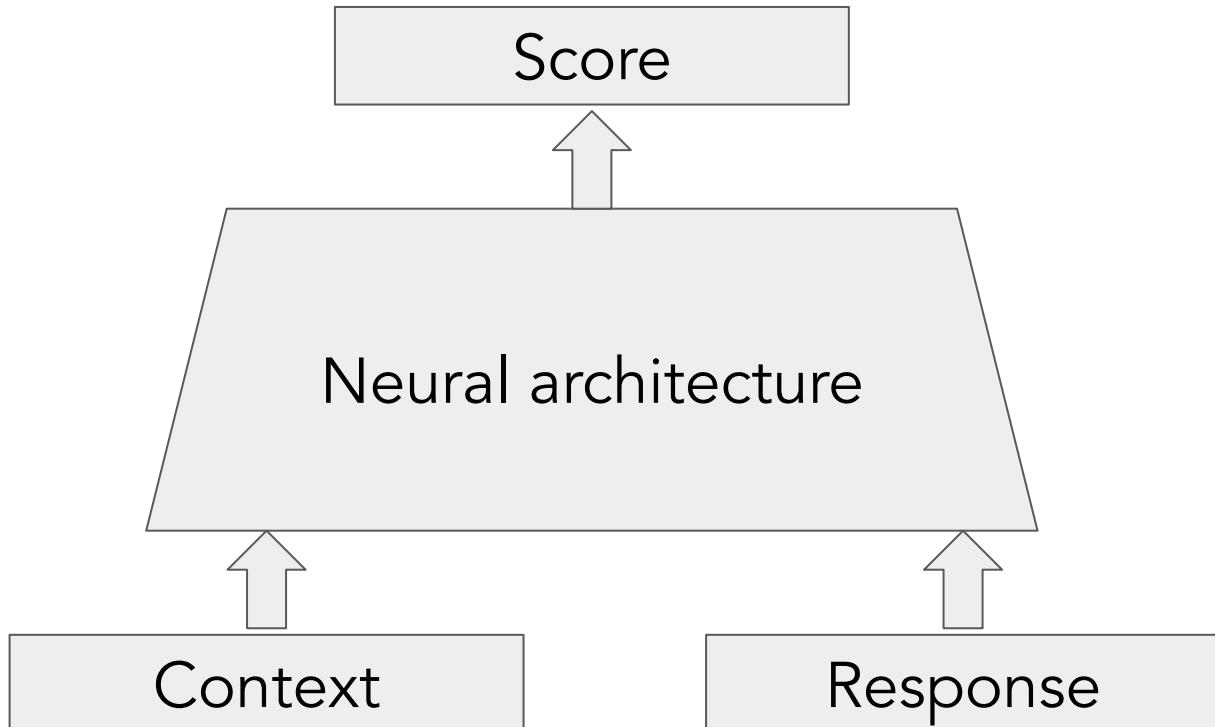
- Constrained by the list of candidate responses
- More informative, semantically relevant and controllable responses
- Easier to train

The retrieval based system explosion

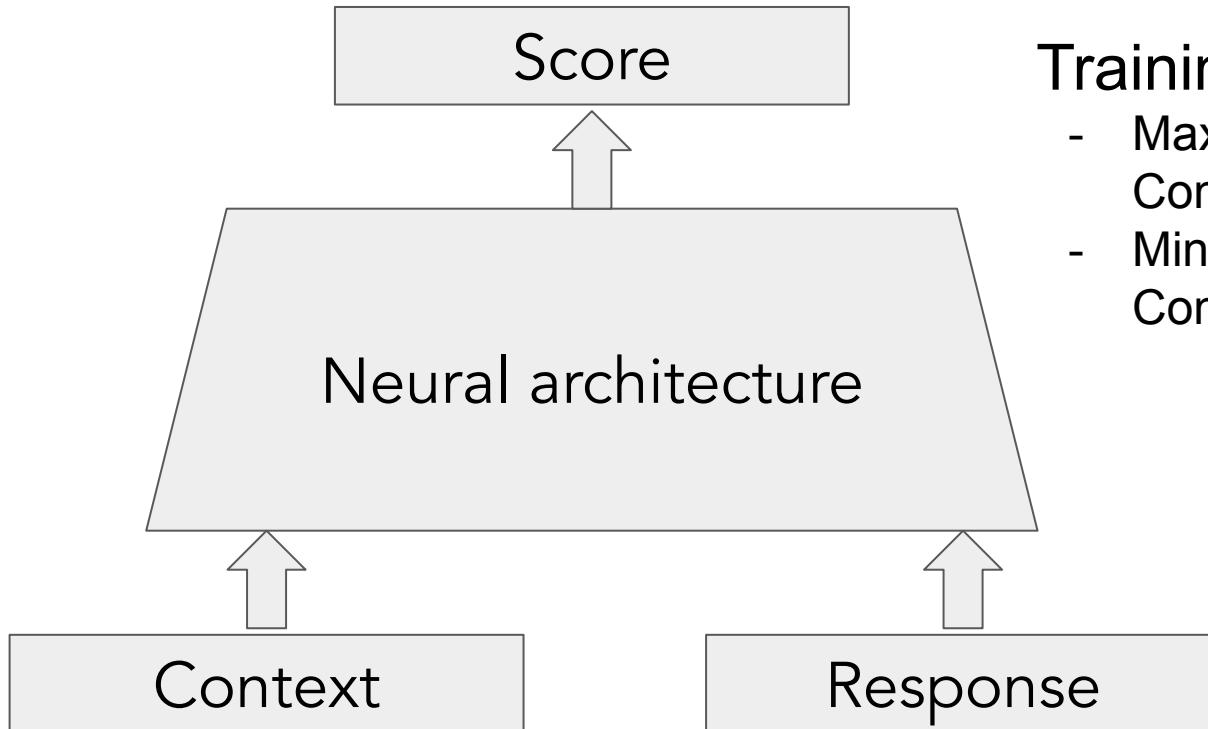
Even if it was proposed only a few years ago, there has been a lot of interest on the topic:

Wang et al., 2013; Ji et al., 2014; Lowe et al., 2015, 2017; Al-Rfou et al., 2016; Henderson et al., 2017, 2019; Cer et al., 2018; Yang et al., 2018; Du and Black, 2018; Chaudhuri et al., 2018; Weston et al., 2018; Wu et al., 2018, 2017; Zhou et al., 2018; Zhang et al., 2018; El Amel Boussaha et al., 2019...

Neural response selection model



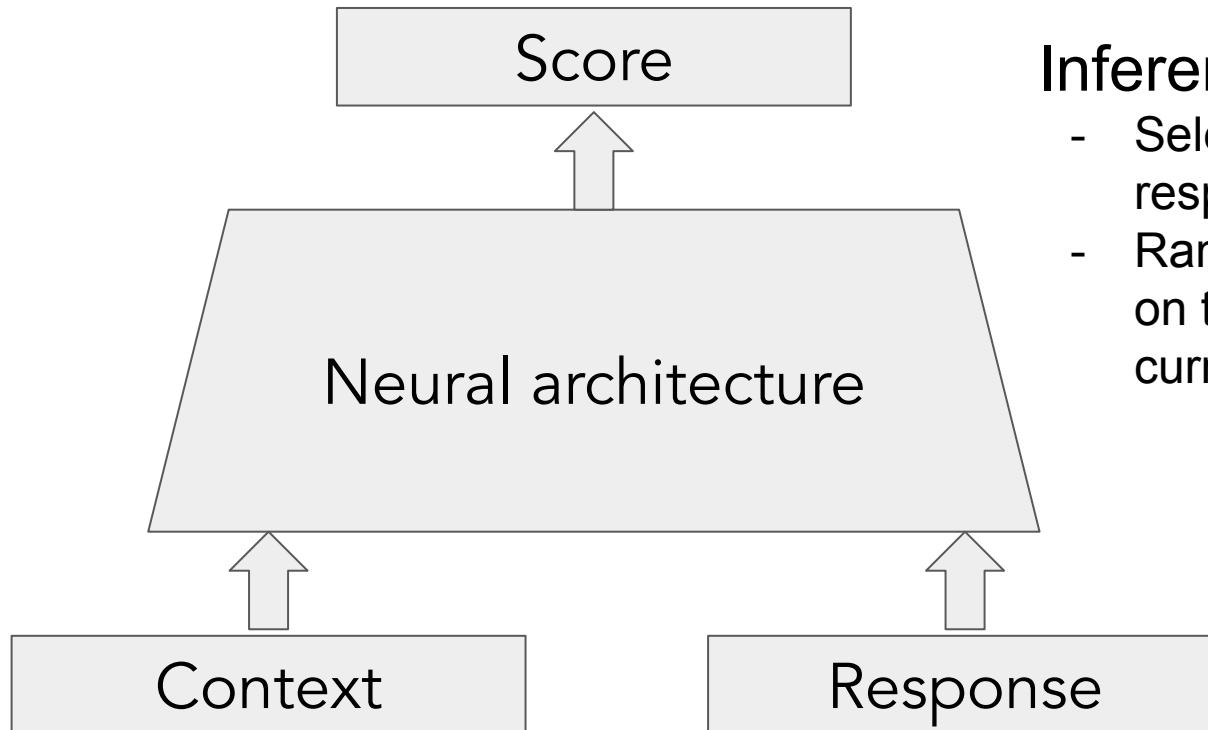
Neural response selection model



Training:

- Maximise the score of positive Context-Response pairs
- Minimise the score of negative Context-Response pairs

Neural response selection model



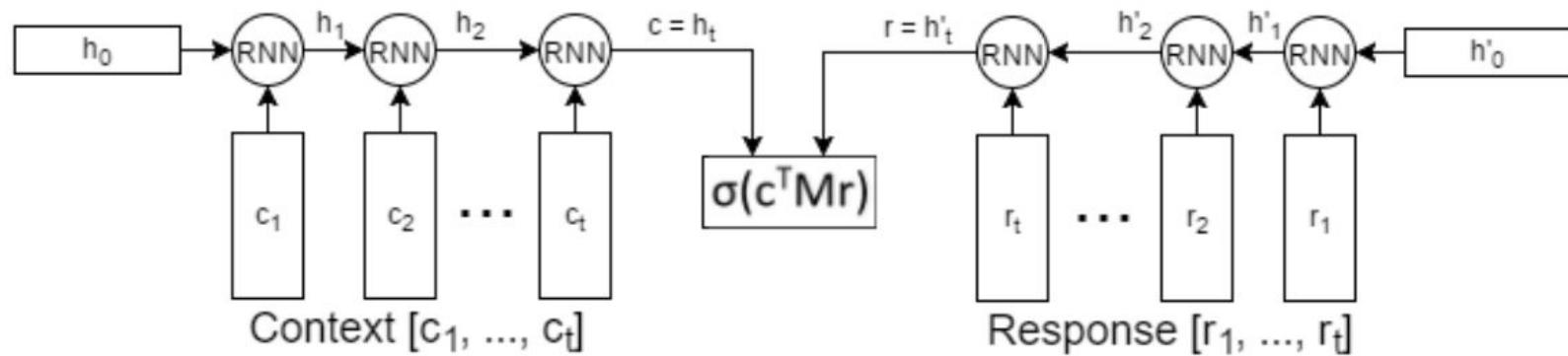
Inference:

- Select the set of possible responses
- Rank the responses based on their score given the current context

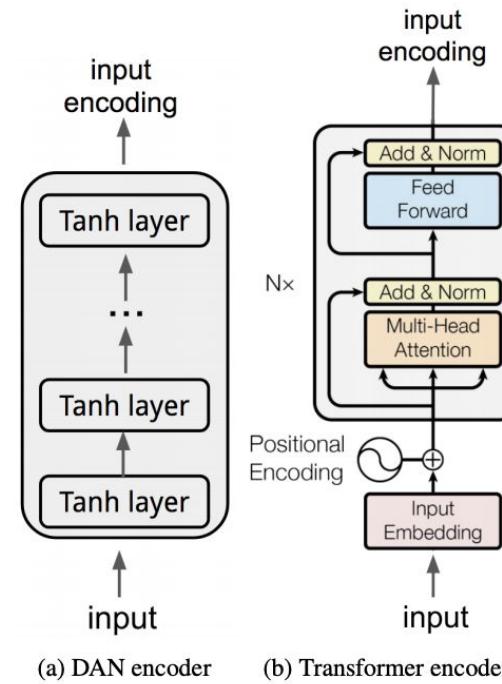
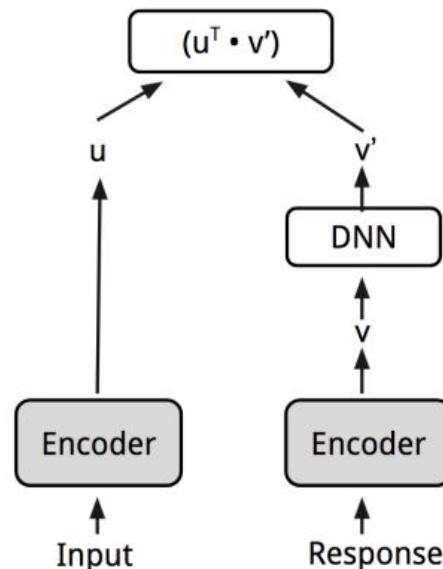
Different neural architectures

- DNN pooling, RNN, CNN, Transformer
- Single turn context, multiple turn context, extra dialogue features
- Dual encoder VS fully connected models

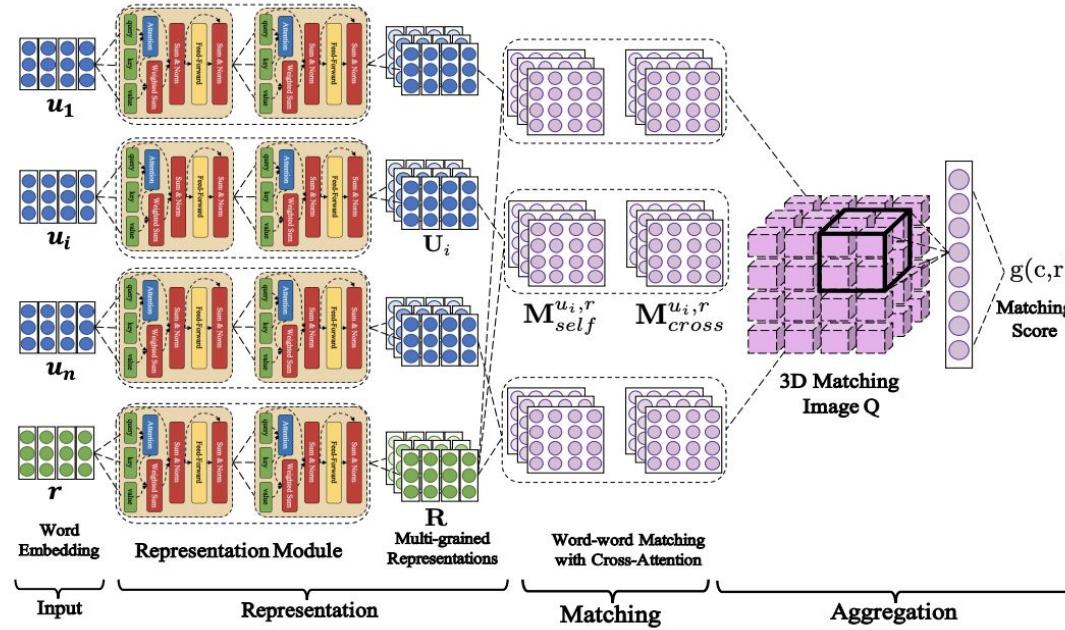
Different neural architectures [Lowe et al., 2015, 2017]



Universal Sentence Encoder (USE) [Cer et al., 2018]



Deep Attention Matching Network [Zhou et al., 2018]



Hybrid models [Weston et al., 2018]

Combine Retrieval and Generation dialogue models -> Get the best of both worlds

Hybrid models [Weston et al., 2018]

Comparison (A vs. B)	Win Rate	A Wins	B Wins	Tie	p-value
RetrieveNRefine ⁺⁺ vs. Memory Network	54.5%	340	284	572	0.027
RetrieveNRefine ⁺⁺ vs. Seq2Seq	53.7%	571	492	203	0.016
RetrieveNRefine ⁺⁺ (retrieved) vs.Seq2Seq	53.8%	290	249	87	
RetrieveNRefine ⁺⁺ (generated) vs.Seq2Seq	53.6%	281	243	116	
RetrieveNRefine ⁺ vs. Memory Network	51.63%	492	461	243	
RetrieveNRefine ⁺ vs. Human	30.13%	69	160	14	
Seq2Seq vs. Human	26.84%	62	169	22	

Table 5: A/B testing by humans comparing model pairs, reporting win rates for A over B (excluding ties).

Industry examples

- Gmail's smart reply [[Henderson et al., 2017](#)]
- Alibaba's chatbot AliMe [[Qiu et al., 2017](#)]
- Microsoft's socialbot Xiaoice [[Shum et al., 2018](#)]

Evaluation metric

1-of-N accuracy:

how often the correct response is ranked top vs $N-1$ random responses

Training Loss

- Binary
 - or
- Proportional to $1\text{-}of\text{-}N$ accuracy
 - The larger N, the better the loss
 - Some architectures can't handle large N

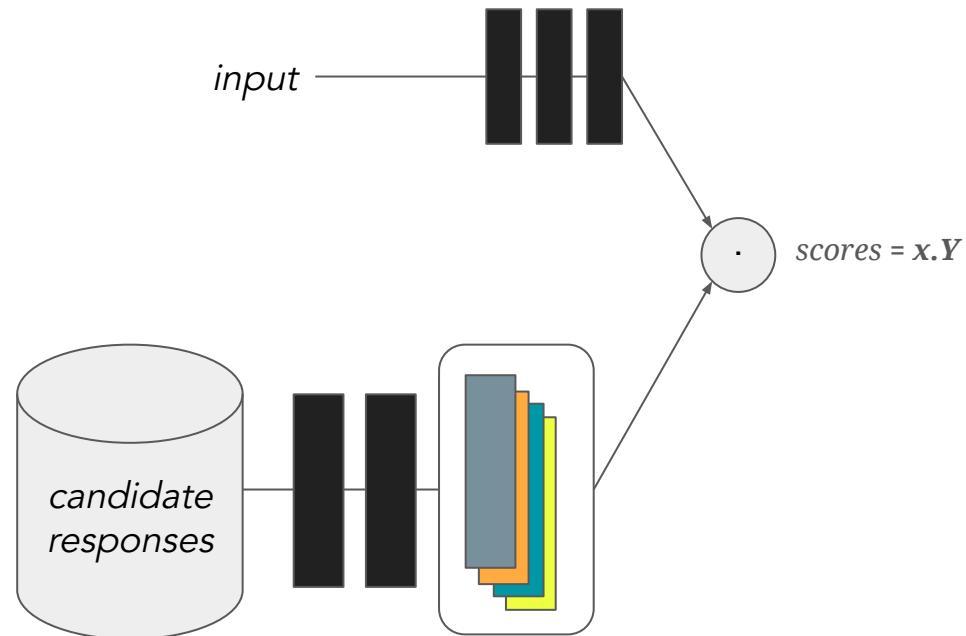
Dual encoder dot-product models

Fully connected models -> context and response embeddings conditional to each other

Dual encoder models -> context and response don't interact until the dot product scoring

Dual encoder dot-product models

- Trained to give a high score for the response found in the data, low score for random responses
- Final score of an input and response is a dot-product of two vectors



Dual encoder dot-product models

- Network encodes a batch of inputs to vectors:

$$\mathbf{x}_1 \quad \mathbf{x}_2 \quad \dots \quad \mathbf{x}_N$$

- And responses to vectors:

$$\mathbf{y}_1 \quad \mathbf{y}_2 \quad \dots \quad \mathbf{y}_N$$

$\mathbf{x}_1 \cdot \mathbf{y}_1$	$\mathbf{x}_1 \cdot \mathbf{y}_2$	$\mathbf{x}_1 \cdot \mathbf{y}_3$	$\mathbf{x}_1 \cdot \mathbf{y}_4$	$\mathbf{x}_1 \cdot \mathbf{y}_5$
$\mathbf{x}_2 \cdot \mathbf{y}_1$	$\mathbf{x}_2 \cdot \mathbf{y}_2$	$\mathbf{x}_2 \cdot \mathbf{y}_3$	$\mathbf{x}_2 \cdot \mathbf{y}_4$	$\mathbf{x}_2 \cdot \mathbf{y}_5$
$\mathbf{x}_3 \cdot \mathbf{y}_1$	$\mathbf{x}_3 \cdot \mathbf{y}_2$	$\mathbf{x}_3 \cdot \mathbf{y}_3$	$\mathbf{x}_3 \cdot \mathbf{y}_4$	$\mathbf{x}_3 \cdot \mathbf{y}_5$
$\mathbf{x}_4 \cdot \mathbf{y}_1$	$\mathbf{x}_4 \cdot \mathbf{y}_2$	$\mathbf{x}_4 \cdot \mathbf{y}_3$	$\mathbf{x}_4 \cdot \mathbf{y}_4$	$\mathbf{x}_4 \cdot \mathbf{y}_5$
$\mathbf{x}_5 \cdot \mathbf{y}_1$	$\mathbf{x}_5 \cdot \mathbf{y}_2$	$\mathbf{x}_5 \cdot \mathbf{y}_3$	$\mathbf{x}_5 \cdot \mathbf{y}_4$	$\mathbf{x}_5 \cdot \mathbf{y}_5$

Dual encoder dot-product models

- The $N \times N$ matrix of all scores is a fast matrix product
- Large improvement in 1 of 100 ranking accuracy over binary classification

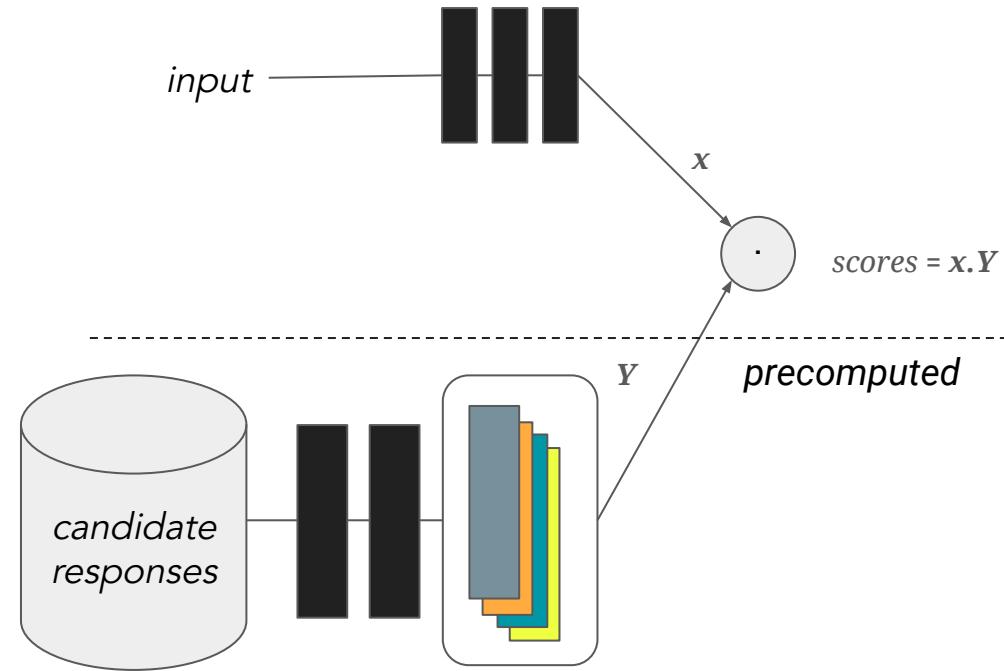
$x_1 \cdot y_1$	$x_1 \cdot y_2$	$x_1 \cdot y_3$	$x_1 \cdot y_4$	$x_1 \cdot y_5$
$x_2 \cdot y_1$	$x_2 \cdot y_2$	$x_2 \cdot y_3$	$x_2 \cdot y_4$	$x_2 \cdot y_5$
$x_3 \cdot y_1$	$x_3 \cdot y_2$	$x_3 \cdot y_3$	$x_3 \cdot y_4$	$x_3 \cdot y_5$
$x_4 \cdot y_1$	$x_4 \cdot y_2$	$x_4 \cdot y_3$	$x_4 \cdot y_4$	$x_4 \cdot y_5$
$x_5 \cdot y_1$	$x_5 \cdot y_2$	$x_5 \cdot y_3$	$x_5 \cdot y_4$	$x_5 \cdot y_5$

Training Loss

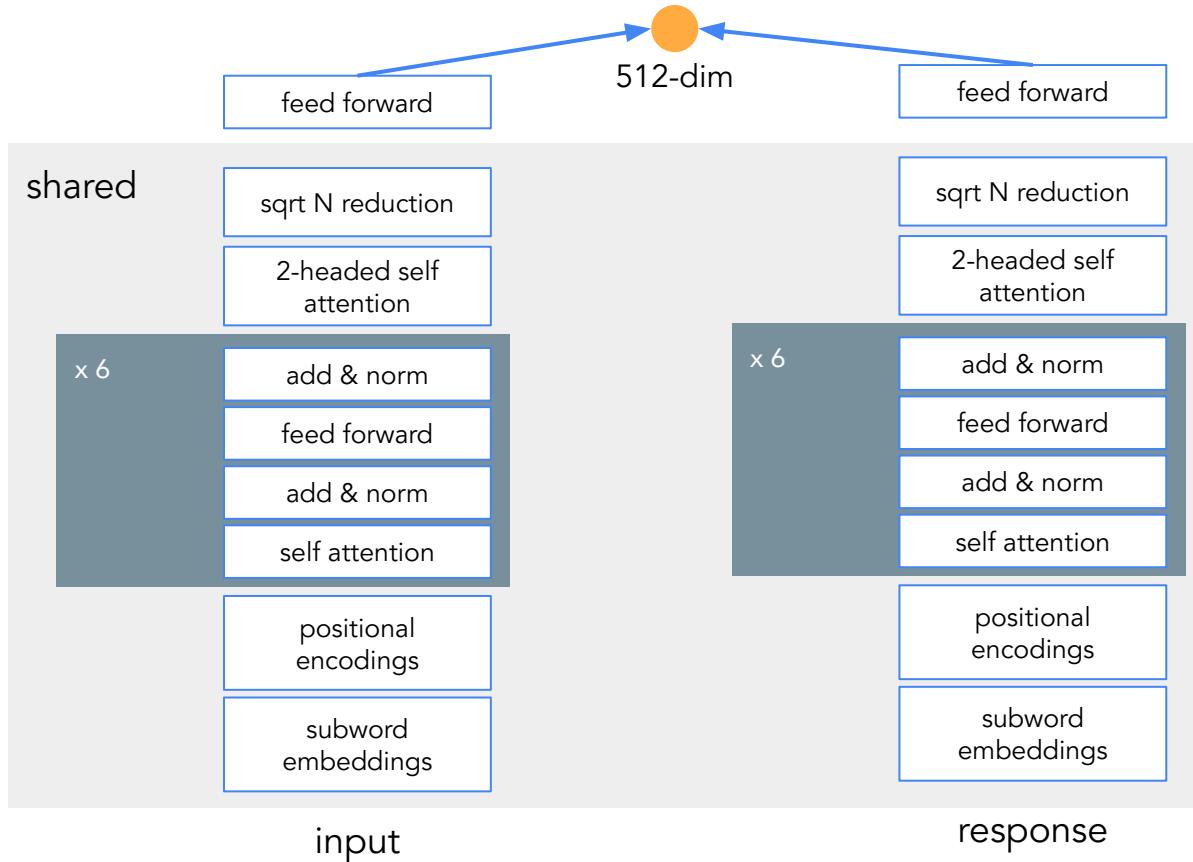
$$\mathcal{L} = - \sum_{i=1}^B S(\mathbf{c}_i, \mathbf{r}_i) + \sum_{i=1}^B \log \sum_{j=1, j \neq i}^B e^{S(\mathbf{c}_i, \mathbf{r}_j)}$$

Precomputation for dot product model

- The representations of the candidates \mathbf{Y} can be precomputed
- Approximate nearest neighbor search can speed up the top N search



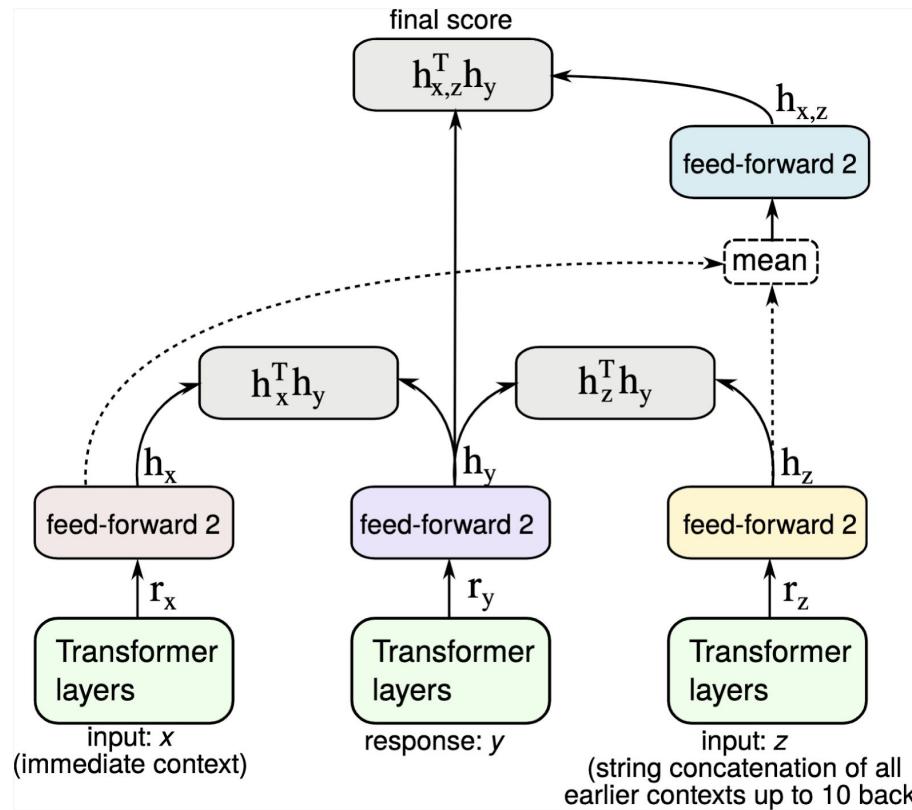
ConveRT [Henderson et al., 2020]



ConveRT [Henderson et al., 2020]

Reddit 1-of-100 accuracy		
keyword-based	TF-IDF	26.7%
	BM25	27.6%
MAP dot product models	ELMo	19.3%
	BERT	24.5%
	USE	40.8%
BERT dot-product model	USE_QA	46.3%
		55.0%
ConveRT	n-grams	61.3%
	subwords	68.2%

Contextual ConveRT [Henderson et al., 2020]

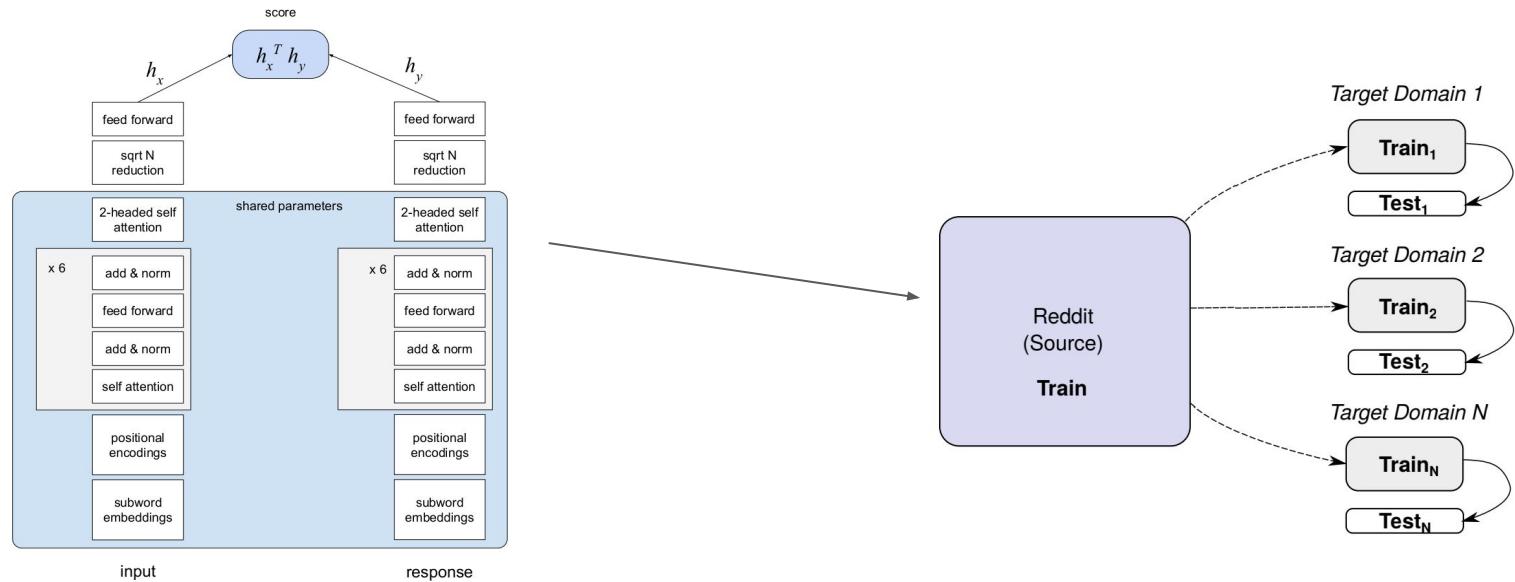


ConveRT [Henderson et al., 2020]

Reddit 1-of-100 accuracy		
keyword-based	TF-IDF	26.7%
	BM25	27.6%
MAP dot product models	ELMo	19.3%
	BERT	24.5%
	USE	40.8%
BERT dot-product model	USE_QA	46.3%
		55.0%
ConveRT	n-grams	61.3%
	subwords	68.2%
Contextual ConveRT	subwords	71.8%

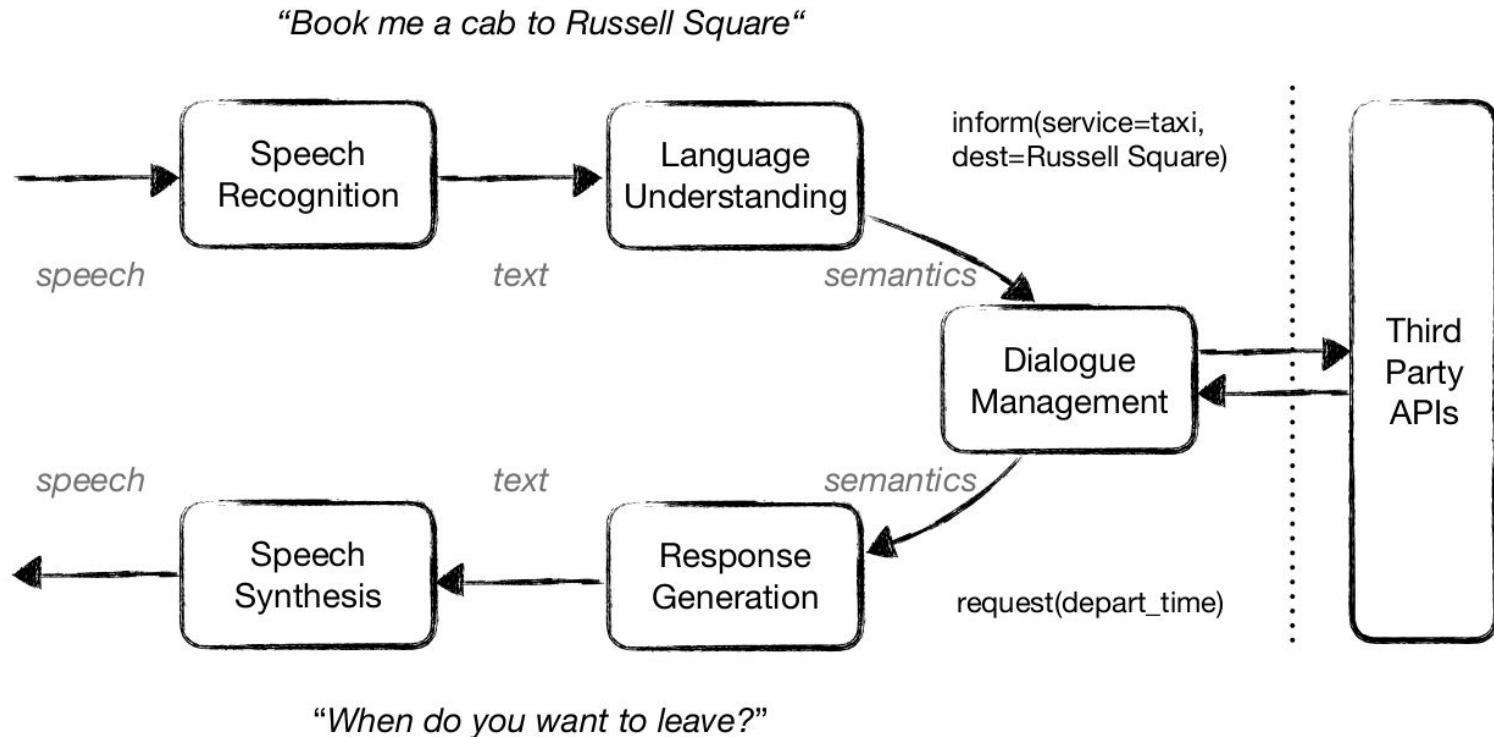
From Open-Domain to Task-Oriented? [Humeau et al., 2019]

- Pretraining (non-generative) conversational response selection on large corpora + in-domain (task-oriented) fine-tuning

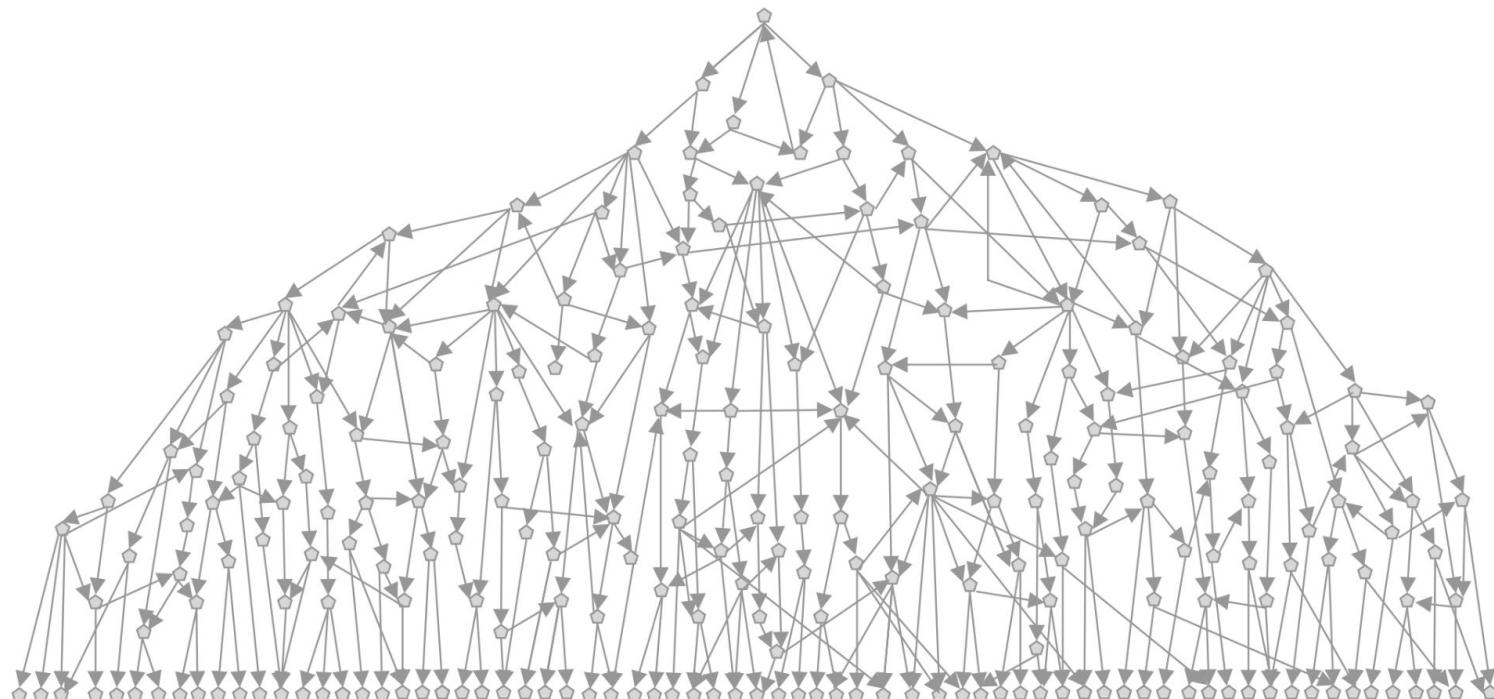


Task-Oriented Dialogue Modelling

Modular Task-Based Conversational AI Systems



Task-Oriented Dialogue Modelling



Evaluation of Task-Oriented Systems

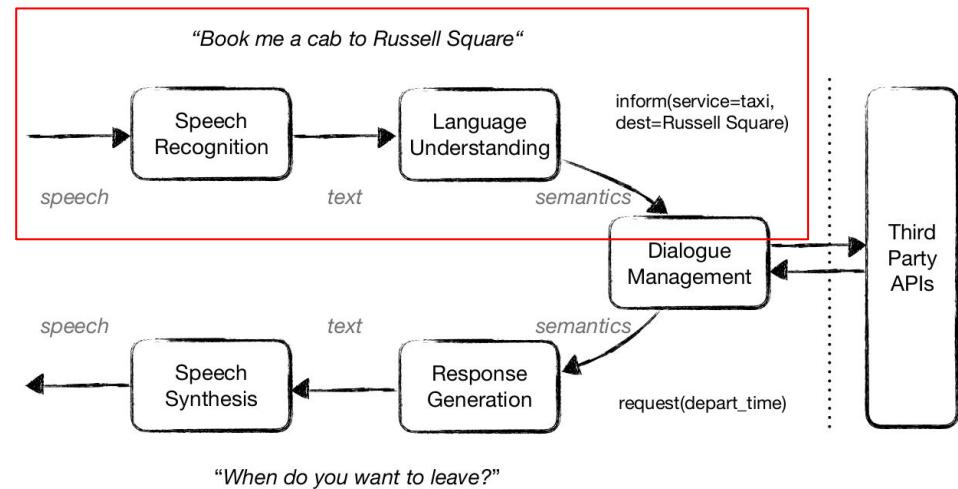
- Much more narrow in definition
- Multiple sub-tasks with well defined objectives
- Common evaluation:
 - slot prediction
 - intent prediction
 - success rate

Evaluation of Task-Oriented Systems

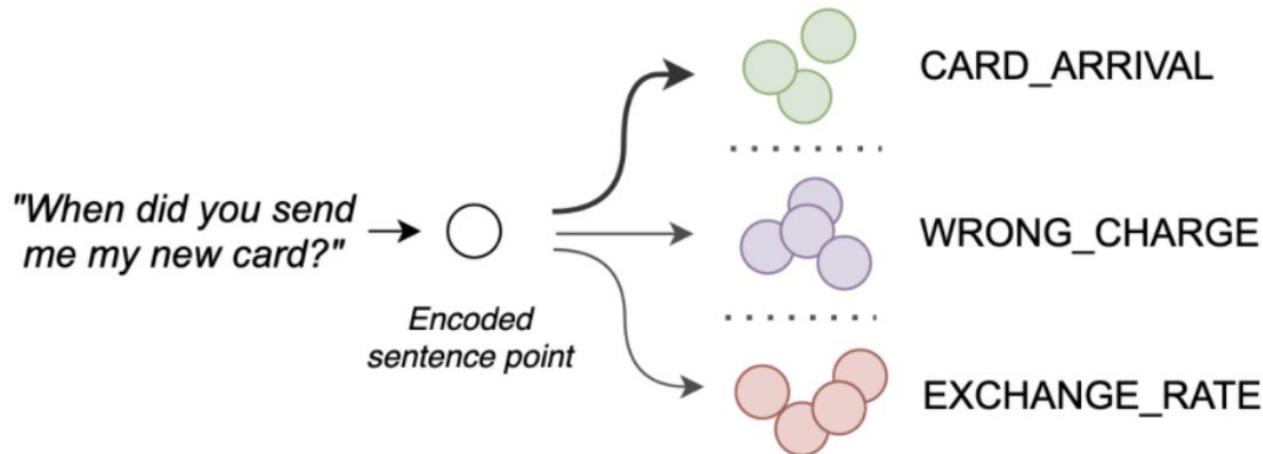
- Dialogue State Tracking Challenge 2 (DSTC2)
- Dialogue System Technology Challenges
- MultiWOZ
- SLURP, MASSIVE
- many more!

Modelling NLU

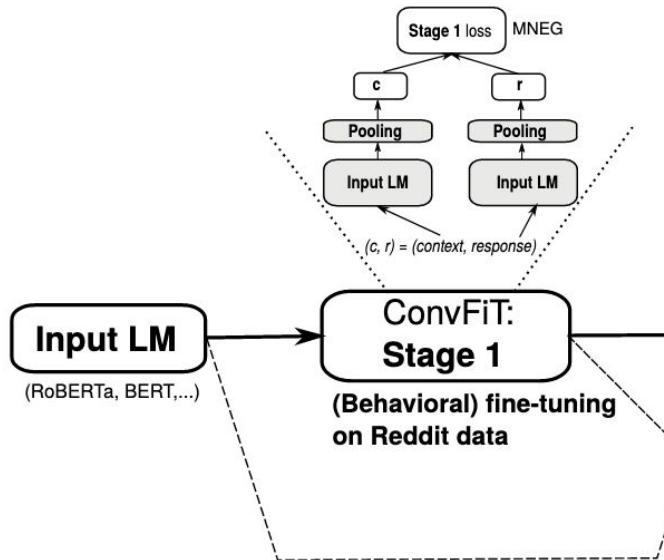
- Intent classification
- Slot labeling



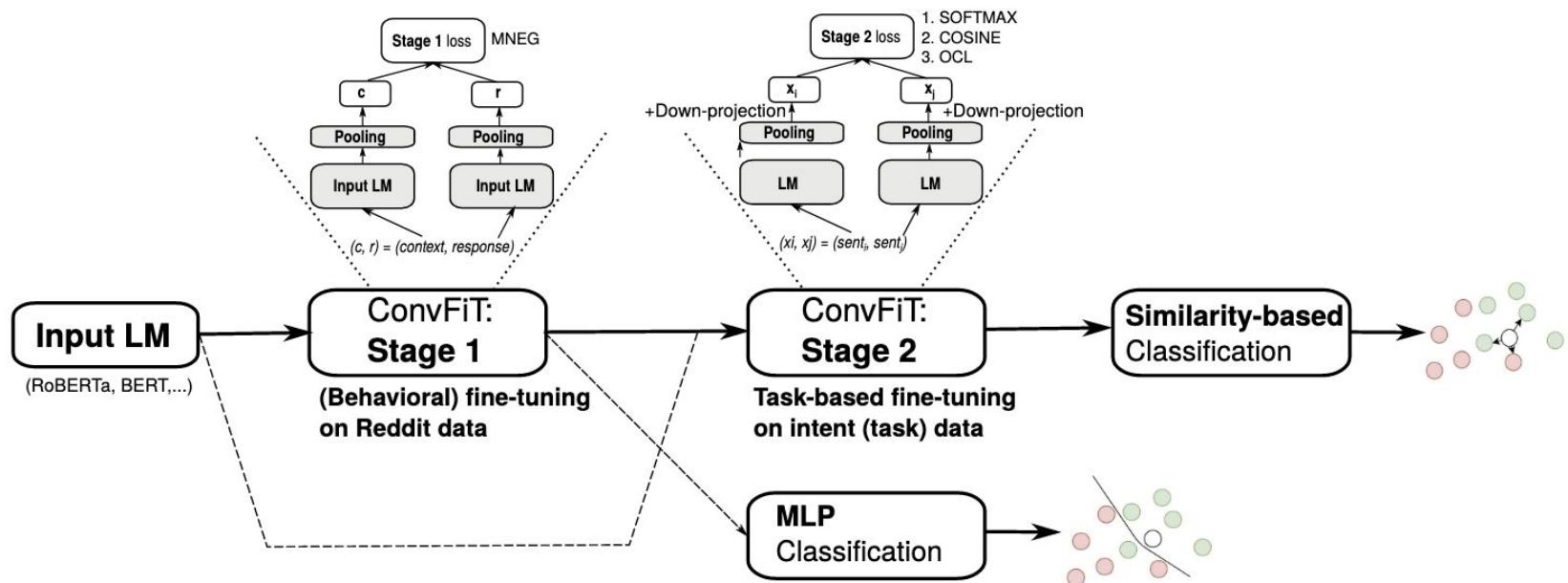
Intent classification



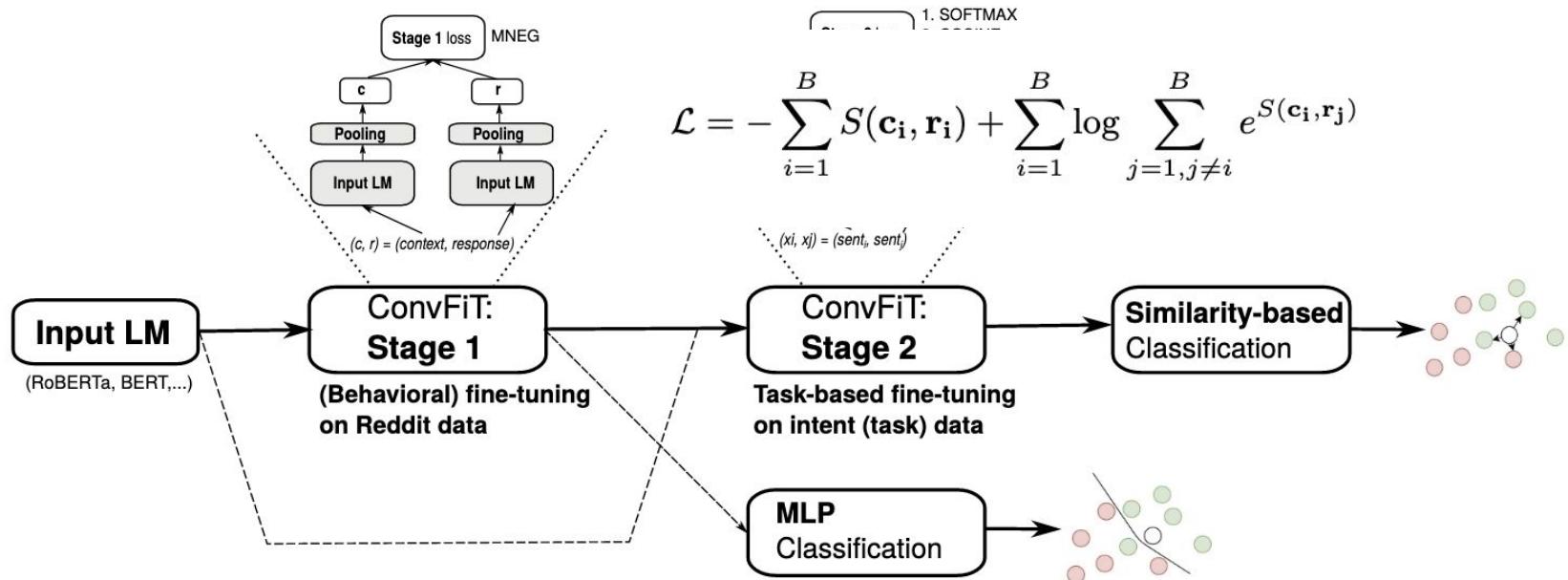
ConvFiT [Vulić et al., 2021]



ConvFiT [Vulić et al., 2021]



ConvFiT [Vulić et al., 2021]



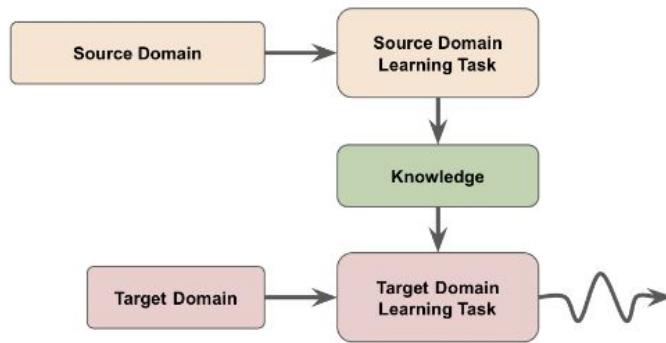
ConvFiT [Vulić et al., 2021]



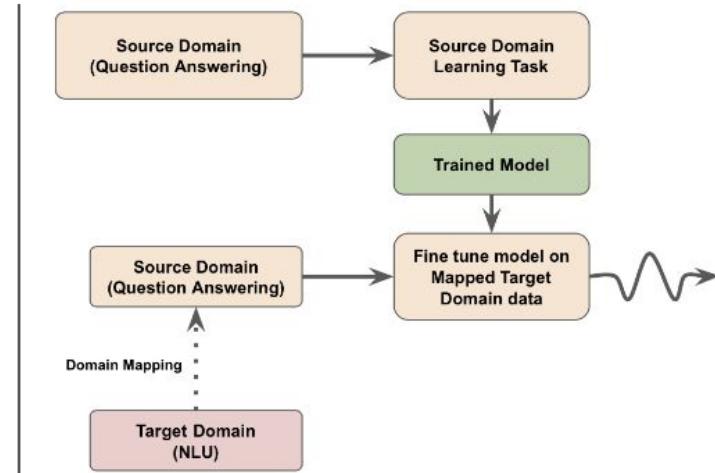
ConvFiT [Vulić et al., 2021]

Model Variant	BANKING77			CLINC150			HWU64		
	10	30	Full	10	30	Full	10	30	Full
Similarity-Based Classification									
ROB+S1+S2-COS	86.48	91.33	94.35	92.87	95.91	97.20	85.06	90.46	92.98
BERT+S1+S2-COS	84.32	90.91	93.91	91.80	95.58	96.56	85.13	89.41	91.93
DROB+S1+S2-COS	85.13	90.75	94.06	91.64	95.48	97.00	83.64	89.68	92.94
Baselines: MLP Classification									
ROB+S1	83.08	90.16	93.38	90.98	94.12	96.42	81.13	87.73	91.44
BERT+S1	82.69	89.82	93.67	89.88	94.07	96.33	82.25	88.01	91.12
CONVERT*	83.32	89.37	93.01	92.62	95.78	97.16	82.65	87.88	91.24
USE*	84.23	89.74	92.81	90.85	93.98	95.06	83.75	89.03	91.25
USE (ours)	82.95	89.09	92.81	90.27	93.54	94.91	82.71	88.20	91.64
LABSE	81.69	88.96	92.60	90.89	93.41	95.12	81.60	86.15	90.99

QANLU [Namazifar and Papangelis, 2020]



(a) Transfer Learning from source domain to target domain



(b) Transfer learning through mapping a target domain to source domain. In this work we map NLU to QA tasks.

QANLU [Namazifar and Papangelis, 2020]

$Q_{cuisine}$: {"what cuisine was mentioned?",
"what type of food was specified?"}

$Q_{price\ range}$: {"what price range?"}

Q_{area} : {"what part of town was mentioned?", "what area?"}

Given T , S , and Q it is straightforward to create the set of all the possible questions and their corresponding answers for each t_i as the context for the questions:

Context:	<i>Show cheap Italian restaurants</i>	
	what cuisine was mentioned?	"Italian"
	what type of food was specified?	"Italian"
	what price range?	"cheap"
	what part of town was mentioned?	""
	what area?	""

Language Models as Few-Shot Learner for Task-Oriented Dialogue Systems

[Madotto 2020]

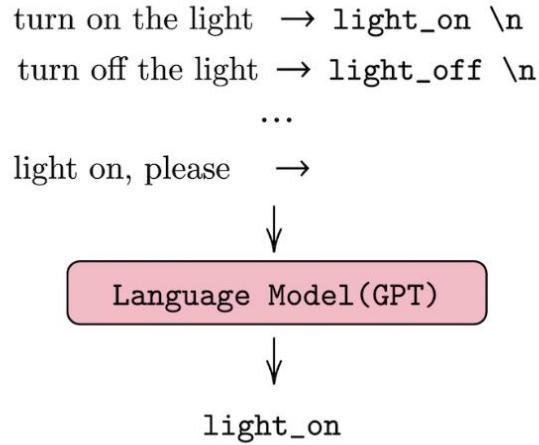
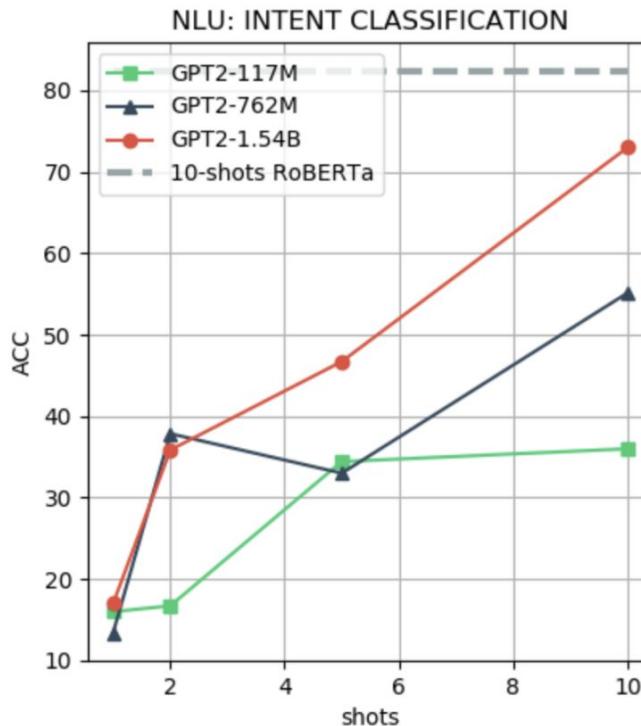


Figure 1: Language model priming for few-shot intent recognition. Image inspired by OpenAI GPT-3 (Brown et al., 2020). Few examples are provided along with the sample to be predicted as the prefix to the language model.

Language Models as Few-Shot Learner for Task-Oriented Dialogue Systems [Madotto 2020]



Final project

Final project

MASSIVE: A 1M-Example Multilingual Natural Language Understanding Dataset with 51 Typologically-Diverse Languages

Jack FitzGerald*

Scott Mackie

Aaron Nash

Richa Singh

Misha Britan

Christopher Hench

Kay Rottmann

Liam Urbach

Swetha Ranganath

Wouter Leeuwis

Prem Natarajan

Charith Peris

Ana Sanchez

Vishesh Kakarala

Laurie Crist

Gokhan Tur

Final project

MASSIVE: A 1M-Example Multilingual Natural Language Understanding Dataset with 51 Typologically-Diverse Languages

Jack FitzGerald*

Christopher Hench

Charith Peris

Scott Ma

SLURP: A Spoken Language Understanding Resource Package

Richa Si

Emanuele Bastianelli^{†*}, Andrea Vanzo^{†*}, Paweł Świątociąski^{‡*} and Verena Rieser[†]

Misha Bi

[†]The Interaction Lab, MACS, Heriot-Watt University, Edinburgh, UK

[‡]Faculty of Engineering, University of New South Wales, Sydney, Australia

{e.bastianelli, a.vanzo, v.t rieser}@hw.ac.uk

p.swietojanski@unsw.edu.au

SLURP [Bastianelli et al., 2021]

User: “*Make a calendar entry for brunch on Saturday morning with Aaronson.*”

Scenario: Calendar

Action: Create_entry

Entity tags and lexical fillers: [event_name: brunch], [date: Saturday], [timeofday: morning], [person: Aaronson]

Figure 1: Example annotation from SLURP dataset.

Final project



Intent classification

	Intent Accuracy (%)					
	XLM-R Full	mT5 T2T Full	mT5 Enc Full	XLM-R Zero	mT5 T2T Zero	mT5 Enc Zero
en-US	88.3 ± 1.2	87.9 ± 1.2	89.0 ± 1.1			
sv-SE	87.9 ± 1.2	87.8 ± 1.2	88.5 ± 1.1	85.2 ± 1.3	77.1 ± 1.5	76.0 ± 1.5
nb-NO	87.3 ± 1.2	87.6 ± 1.2	87.7 ± 1.2	83.6 ± 1.3	76.3 ± 1.5	72.8 ± 1.6
ru-RU	87.2 ± 1.2	87.0 ± 1.2	86.8 ± 1.2	81.3 ± 1.4	76.2 ± 1.5	73.8 ± 1.6
id-ID	87.1 ± 1.2	87.0 ± 1.2	86.8 ± 1.2	83.1 ± 1.3	77.0 ± 1.5	74.1 ± 1.6
fa-IR	87.0 ± 1.2	86.3 ± 1.2	87.2 ± 1.2	81.1 ± 1.4	69.0 ± 1.7	66.3 ± 1.7
es-ES	86.9 ± 1.2	86.7 ± 1.2	86.8 ± 1.2	78.8 ± 1.5	76.6 ± 1.5	75.9 ± 1.5
da-DK	86.9 ± 1.2	87.5 ± 1.2	88.0 ± 1.2	83.1 ± 1.3	76.8 ± 1.5	73.4 ± 1.6
ro-RO	86.9 ± 1.2	87.2 ± 1.2	87.0 ± 1.2	80.8 ± 1.4	73.0 ± 1.6	70.1 ± 1.6
nl-NL	86.8 ± 1.2	87.2 ± 1.2	87.6 ± 1.2	82.1 ± 1.4	79.9 ± 1.4	76.4 ± 1.5
pt-PT	86.7 ± 1.2	86.7 ± 1.2	86.9 ± 1.2	79.5 ± 1.5	74.0 ± 1.6	74.5 ± 1.6
it-IT	86.6 ± 1.2	86.8 ± 1.2	87.6 ± 1.2	76.4 ± 1.5	72.3 ± 1.6	71.5 ± 1.6
ko-KR	86.5 ± 1.2	86.1 ± 1.2	86.4 ± 1.2	77.0 ± 1.5	60.0 ± 1.8	61.9 ± 1.7
sq-AL	86.4 ± 1.2	85.9 ± 1.3	86.4 ± 1.2	67.6 ± 1.7	62.9 ± 1.7	62.0 ± 1.7
fr-FR	86.3 ± 1.2	86.9 ± 1.2	87.2 ± 1.2	80.8 ± 1.4	76.9 ± 1.5	74.1 ± 1.6
sl-SL	86.3 ± 1.2	85.9 ± 1.3	86.8 ± 1.2	69.5 ± 1.7	61.5 ± 1.7	59.8 ± 1.8
vi-VN	86.3 ± 1.2	85.8 ± 1.3	87.2 ± 1.2	79.2 ± 1.5	64.2 ± 1.7	62.7 ± 1.7
tr-TR	86.3 ± 1.2	86.1 ± 1.2	87.1 ± 1.2	78.4 ± 1.5	66.5 ± 1.7	63.7 ± 1.7
el-GR	86.2 ± 1.2	85.8 ± 1.3	86.6 ± 1.2	74.0 ± 1.6	71.9 ± 1.6	69.8 ± 1.6
az-AZ	86.2 ± 1.2	86.2 ± 1.2	86.4 ± 1.2	70.9 ± 1.6	57.0 ± 1.8	55.5 ± 1.8
hu-HU	86.2 ± 1.2	85.4 ± 1.3	86.9 ± 1.2	77.1 ± 1.5	68.0 ± 1.7	66.4 ± 1.7
lv-LV	86.1 ± 1.2	85.2 ± 1.3	86.6 ± 1.2	69.2 ± 1.7	61.0 ± 1.8	60.0 ± 1.8
ms-MY	86.1 ± 1.2	86.8 ± 1.2	86.9 ± 1.2	76.7 ± 1.5	69.9 ± 1.6	66.0 ± 1.7
he-IL	85.9 ± 1.3	84.6 ± 1.3	86.2 ± 1.2	73.2 ± 1.6	64.7 ± 1.7	64.0 ± 1.7
hi-IN	85.8 ± 1.3	85.6 ± 1.3	86.2 ± 1.2	74.8 ± 1.6	62.4 ± 1.7	59.3 ± 1.8
pl-PL	85.8 ± 1.3	86.3 ± 1.2	87.1 ± 1.2	80.7 ± 1.4	76.4 ± 1.5	74.1 ± 1.6
de-DE	85.7 ± 1.3	86.2 ± 1.2	86.8 ± 1.2	77.6 ± 1.5	77.3 ± 1.5	73.9 ± 1.6

Intent classification

	Intent Accuracy (%)					
	XLM-R Full	mT5 T2T Full	mT5 Enc Full	XLM-R Zero	mT5 T2T Zero	mT5 Enc Zero
en-US	88.3 ± 1.2	87.9 ± 1.2	89.0 ± 1.1			
sv-SE	87.9 ± 1.2	87.8 ± 1.2	88.5 ± 1.1	85.2 ± 1.3	77.1 ± 1.5	76.0 ± 1.5
nb-NO	87.3 ± 1.2	87.6 ± 1.2	87.7 ± 1.2	83.6 ± 1.3	76.3 ± 1.5	72.8 ± 1.6
ru-RU	87.2 ± 1.2	87.0 ± 1.2	86.8 ± 1.2	81.3 ± 1.4	76.2 ± 1.5	73.8 ± 1.6
id-ID	87.1 ± 1.2	87.0 ± 1.2	86.8 ± 1.2	83.1 ± 1.3	77.0 ± 1.5	74.1 ± 1.6
fa-IR	87.0 ± 1.2	86.3 ± 1.2	87.2 ± 1.2	81.1 ± 1.4	69.0 ± 1.7	66.3 ± 1.7
es-ES	86.9 ± 1.2	86.7 ± 1.2	86.8 ± 1.2	78.8 ± 1.5	76.6 ± 1.5	75.9 ± 1.5
da-DK	86.9 ± 1.2	87.5 ± 1.2	88.0 ± 1.2	83.1 ± 1.3	76.8 ± 1.5	73.4 ± 1.6
ro-RO	86.9 ± 1.2	87.2 ± 1.2	87.0 ± 1.2	80.8 ± 1.4	73.0 ± 1.6	70.1 ± 1.6
nl-NL	86.8 ± 1.2	87.2 ± 1.2	87.6 ± 1.2	82.1 ± 1.4	79.9 ± 1.4	76.4 ± 1.5
pt-PT	86.7 ± 1.2	86.7 ± 1.2	86.9 ± 1.2	79.5 ± 1.5	74.0 ± 1.6	74.5 ± 1.6
it-IT	86.6 ± 1.2	86.8 ± 1.2	87.6 ± 1.2	76.4 ± 1.5	72.3 ± 1.6	71.5 ± 1.6
ko-KR	86.5 ± 1.2	86.1 ± 1.2	86.4 ± 1.2	77.0 ± 1.5	60.0 ± 1.8	61.9 ± 1.7
sq-AL	86.4 ± 1.2	85.9 ± 1.3	86.4 ± 1.2	67.6 ± 1.7	62.9 ± 1.7	62.0 ± 1.7
fr-FR	86.3 ± 1.2	86.9 ± 1.2	87.2 ± 1.2	80.8 ± 1.4	76.9 ± 1.5	74.1 ± 1.6
sl-SL	86.3 ± 1.2	85.9 ± 1.3	86.8 ± 1.2	69.5 ± 1.7	61.5 ± 1.7	59.8 ± 1.8
vi-VN	86.3 ± 1.2	85.8 ± 1.3	87.2 ± 1.2	79.2 ± 1.5	64.2 ± 1.7	62.7 ± 1.7
tr-TR	86.3 ± 1.2	86.1 ± 1.2	87.1 ± 1.2	78.4 ± 1.5	66.5 ± 1.7	63.7 ± 1.7
el-GR	86.2 ± 1.2	85.8 ± 1.3	86.6 ± 1.2	74.0 ± 1.6	71.9 ± 1.6	69.8 ± 1.6
az-AZ	86.2 ± 1.2	86.2 ± 1.2	86.4 ± 1.2	70.9 ± 1.6	57.0 ± 1.8	55.5 ± 1.8
hu-HU	86.2 ± 1.2	85.4 ± 1.3	86.9 ± 1.2	77.1 ± 1.5	68.0 ± 1.7	66.4 ± 1.7
lv-LV	86.1 ± 1.2	85.2 ± 1.3	86.6 ± 1.2	69.2 ± 1.7	61.0 ± 1.8	60.0 ± 1.8
ms-MY	86.1 ± 1.2	86.8 ± 1.2	86.9 ± 1.2	76.7 ± 1.5	69.9 ± 1.6	66.0 ± 1.7
he-IL	85.9 ± 1.3	84.6 ± 1.3	86.2 ± 1.2	73.2 ± 1.6	64.7 ± 1.7	64.0 ± 1.7
hi-IN	85.8 ± 1.3	85.6 ± 1.3	86.2 ± 1.2	74.8 ± 1.6	62.4 ± 1.7	59.3 ± 1.8
pl-PL	85.8 ± 1.3	86.3 ± 1.2	87.1 ± 1.2	80.7 ± 1.4	76.4 ± 1.5	74.1 ± 1.6
de-DE	85.7 ± 1.3	86.2 ± 1.2	86.8 ± 1.2	77.6 ± 1.5	77.3 ± 1.5	73.9 ± 1.6

Slot-labeling

Slot-labeling

S: What route are you looking for?

U: Leaving for **Anaheim, CA** on the **2nd of June**.

QASL:

Slot: date Q: What date?

Slot: from_location Q: Where from?

Slot: to_location Q: Where to?

A: **2nd of June**

A: **Anaheim, CA**

A: **"**

Slot-labeling

S: What route are you looking for?

U: Leaving for **Anaheim, CA** on the **2nd of June**.

QASL:

Slot: date Q: What date?

A: **2nd of June**

Slot: from_location Q: Where from?

A: **Anaheim, CA**

Slot: to_location Q: Where to?

A: ""

S: From? (requested_slot:**from_location**)

U: **LAX**

QASL:

Slot: date Q: What date? [**<s> from_location**]

A: ""

Slot: from_location Q: Where from? [**<s> from_location**]

A: **LAX**

Slot: to_location Q: Where to? [**<s> from_location**]

A: ""

ConVEx [Henderson & Vulić, 2020]

Template Sentence

I get frustrated everytime I browse /r/all. I stick to my *BLANK* most of the time.

Why Puerto Rico? It's Memphis at Dallas, which is in Texas where *BLANK* hit
BLANK is my 3rd favorite animated Movie

It really sucks, as the V30 only has *BLANK*. Maybe the Oreo update will add this.
I took *BLANK*, cut it to about 2 feet long and duct taped Vive controllers on each end. Works perfect

I had *BLANK* and won the last game and ended up with 23/20 and still didn't get it.

Input Sentence

/r/misleadingpuddles Saw it on the **frontpage**, plenty of content if you like the premise.

Hurricane Harvey. Just a weird coincidence.

Toy Story 3 ended perfectly, but Disney just wants to keep milking it.

Thanks for the input, but **64GB** is plenty for me :)

Yeah, I just duct taped mine to **a broom stick**. You can only play no arrows mode but it's really fun.

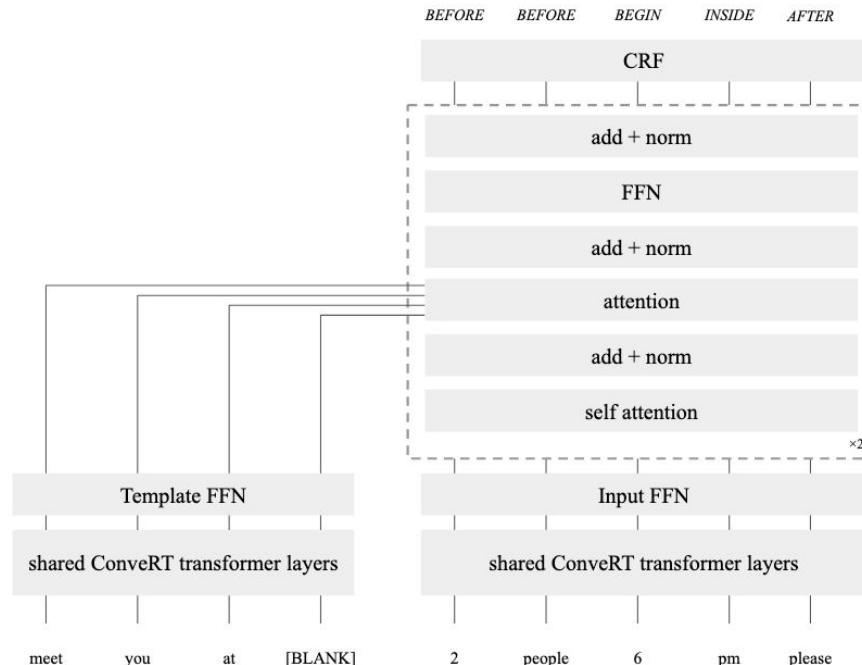
I know how you feel my friend and I got **19/20** on the tournament today

ConVEx [Henderson & Vulić, 2020]

Total Reddit comments	3,680,746,776
Comments filtered by length	1,993,294,538
Extracted keyphrases	3,296,519,827
Training set size	1,172,174,919
Test set size	61,696,649
Mean number of words per keyphrase	1.3

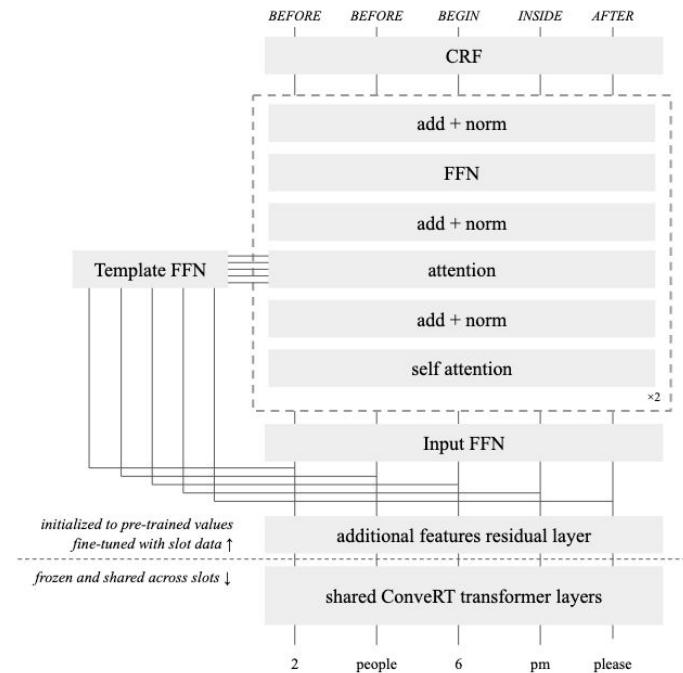
Table 2: Statistics of the pairwise cloze training data.

Pre-training Phase



(a) ConVEx: Pretraining

Finetuning Phase

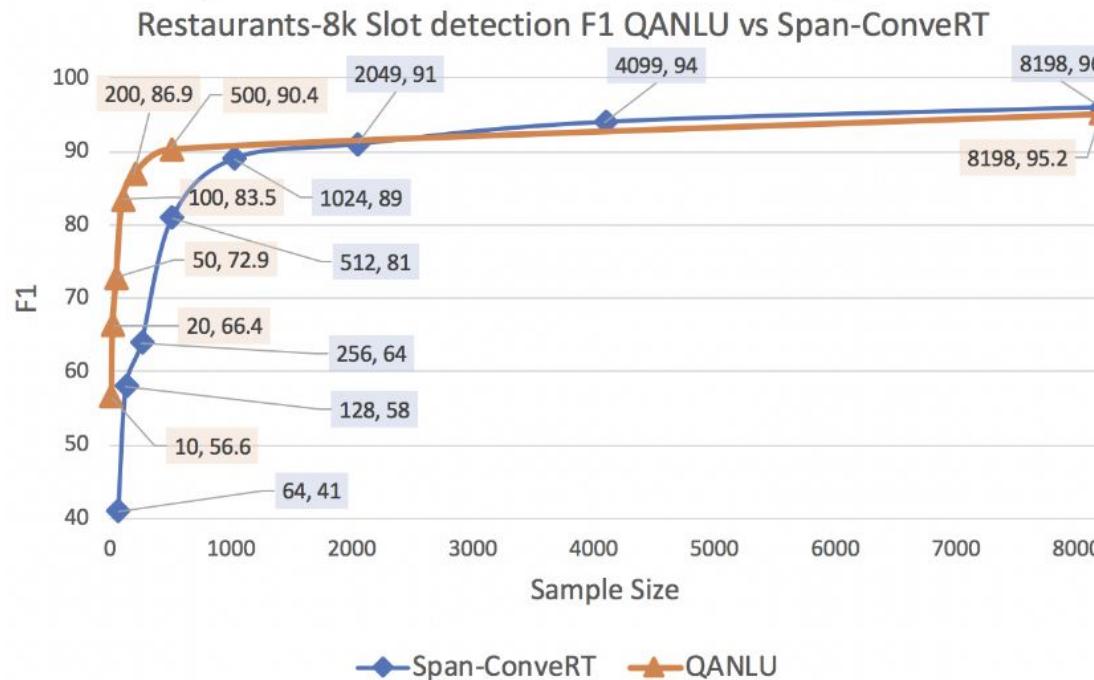


(b) ConVEx: Fine-tuning

Specialized Pre-training

	Weather	Music	Playlist	Book	Search	Restaurant	Creative	Average
<i>Hou et al. (2020)</i>								
TransferBERT	59.4	42.0	46.1	20.7	28.2	67.8	58.6	46.1
SimBERT	53.5	54.1	42.8	75.5	57.1	55.3	32.4	52.9
WPZ+BERT	67.8	56.0	46.0	72.2	73.6	60.2	66.9	63.2
TapNet	53.0	49.8	54.9	83.4	63.1	59.8	67.0	61.6
TapNet+CDT	66.5	66.4	68.2	85.8	73.6	64.2	68.5	70.4
L-WPZ+CDT	74.7	56.7	52.2	78.8	80.6	69.6	67.5	68.6
L-TapNet+CDT	71.6	67.2	75.9	84.4	82.6	70.1	73.4	75.0
<i>This work</i>								
ConVEx (with aux)	71.5	77.6	79.0	84.5	84.0	73.8	67.4	76.8

QANLU [Namazifar and Papangelis, 2020]



Dialogue State Tracking

DSTC 2 Example

hello I am looking for a cheap place in the east
> inform(pricerange=cheap, area=east)

sure, what type of food?
> request(food)

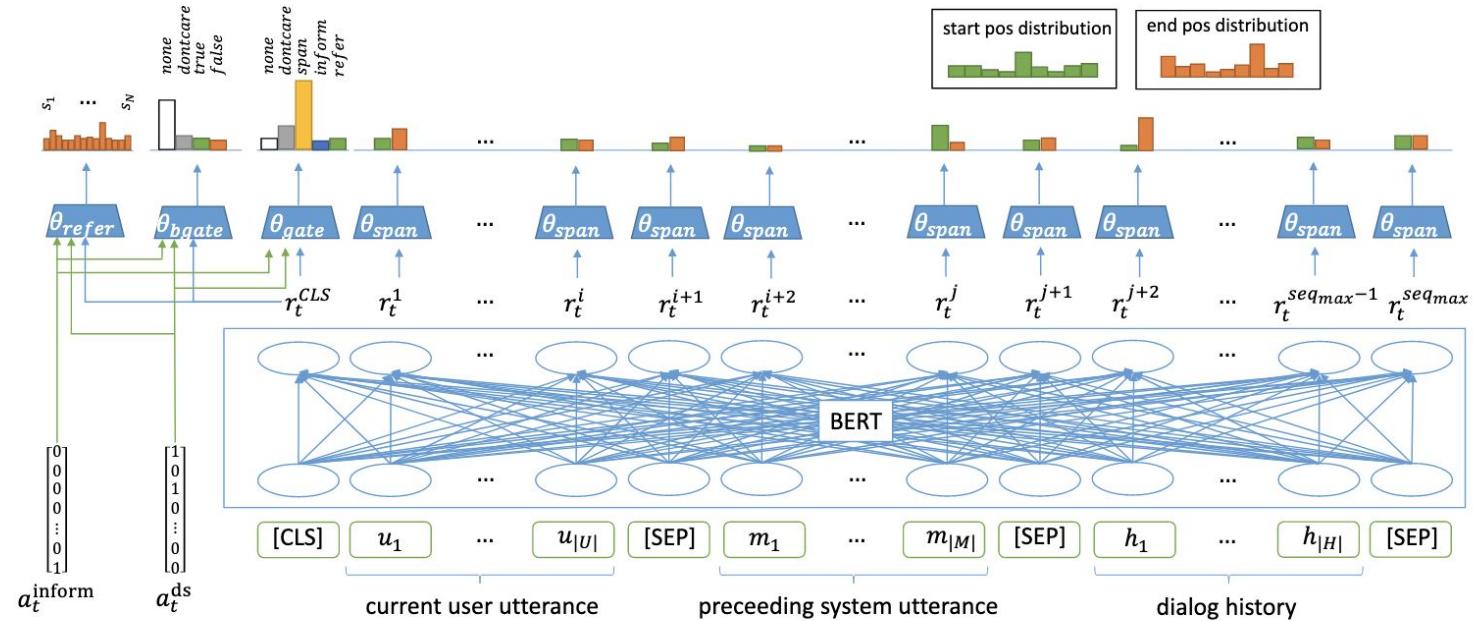
i want gastropub food
> inform(food=gastropub)

there are no cheap places serving gastropub in the east.
> inform(name=none, area=east, pricerange=cheap)

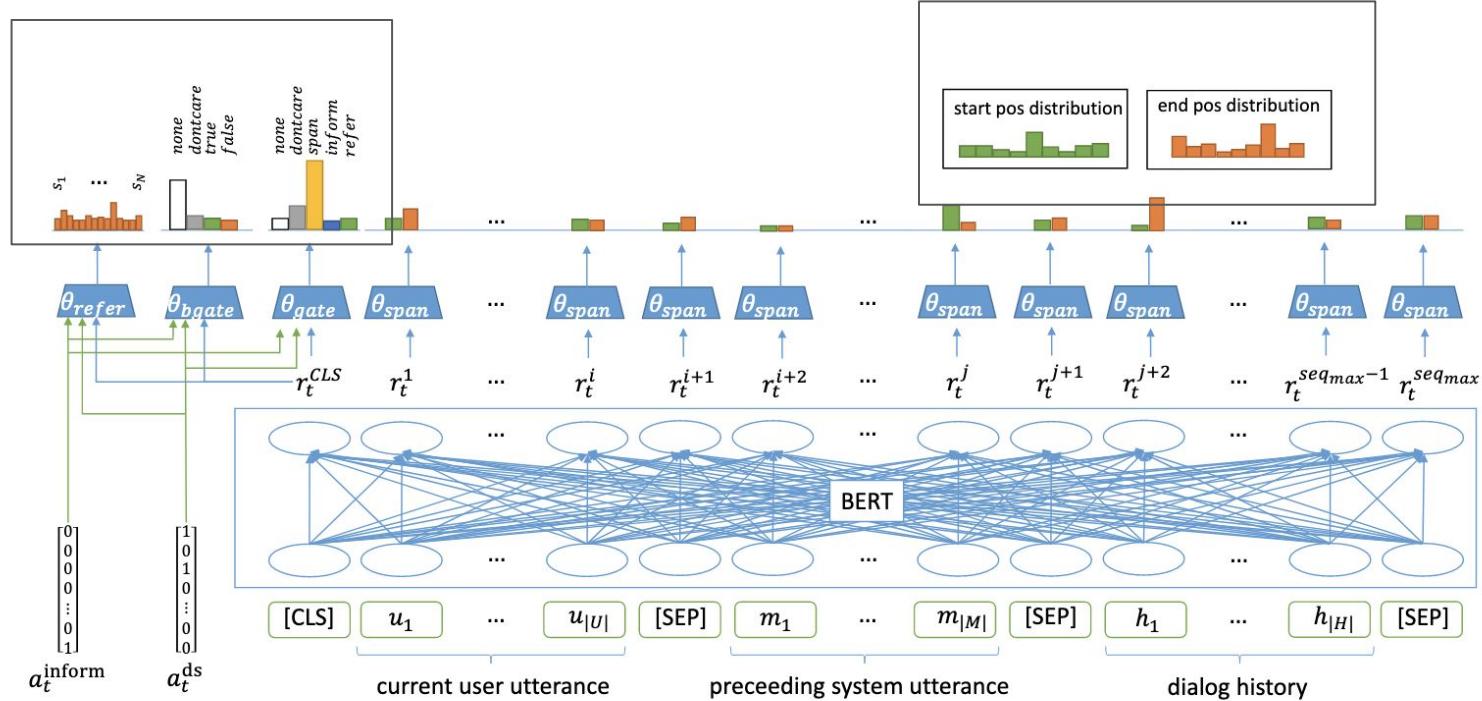
how about any pricerange? and i need to know if they have wifi.
> inform(pricerange=dontcare) request(has_wifi)

The King's Arms is a nice place in the east of town serving gastropub food. It has wifi.
> offer(name="The King's Arms", area=east, food=gastropub, has_wifi=true)

TripPy [Heck et al., 2020]



TripPy [Heck et al., 2020]



TripPy [Heck et al., 2020]

Models	MultiWOZ 2.1
DST-reader (2019)	36.40%
DST-span (2019)	40.39%
SUMBT (2019)	42.40%**
TRADE (2019)	45.60%
MA-DST (2020)	51.04%
DS-DST (2019)	51.21%
SOM-DST (2019)	52.57%
DST-picklist (2019)	53.30%
TripPy	55.29±0.28%

Table 1: DST Results on MultiWOZ 2.1 in JGA (\pm denotes the standard deviation. ** MultiWOZ 2.0 result.

E2E Dialogue Modelling

E2E Dialogue Modelling

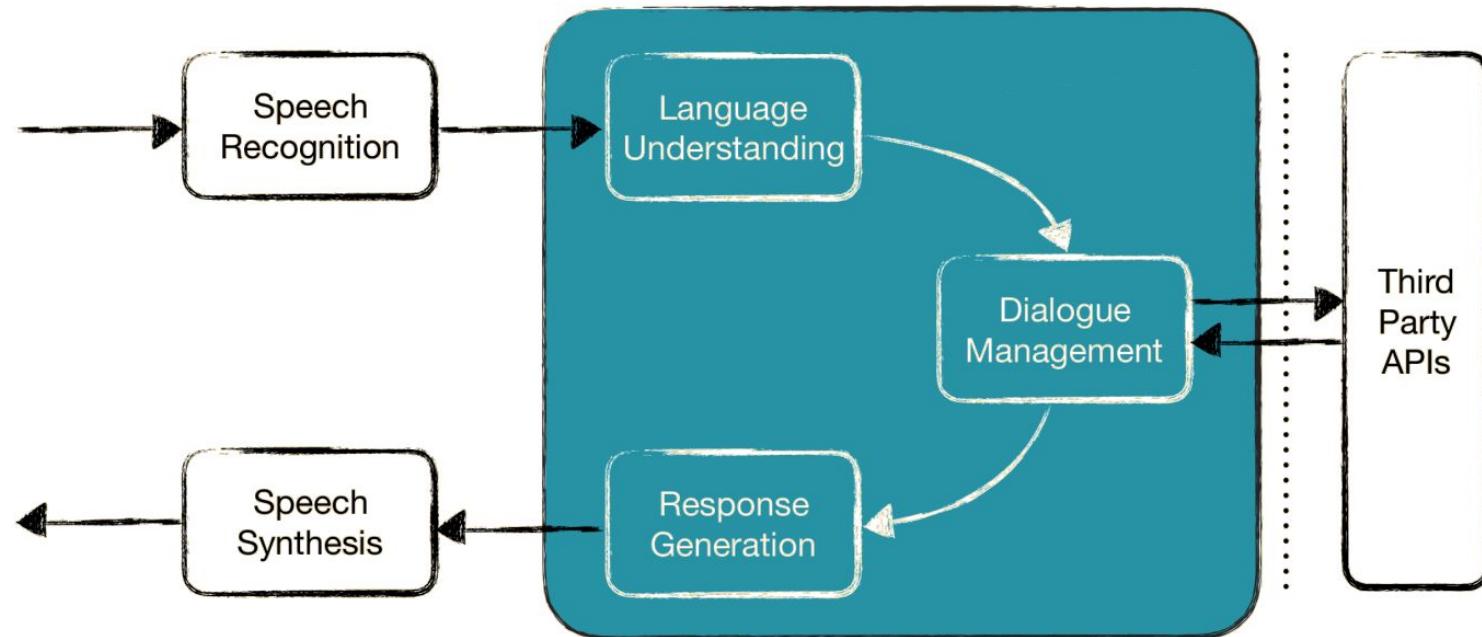
Pros:

- Faster re-training cycle - easier productization
- Natural framework for knowledge sharing across modules
- (Theoretically) Less sophisticated annotation schema
- (Theoretically) Taking advantage of online conversational resources

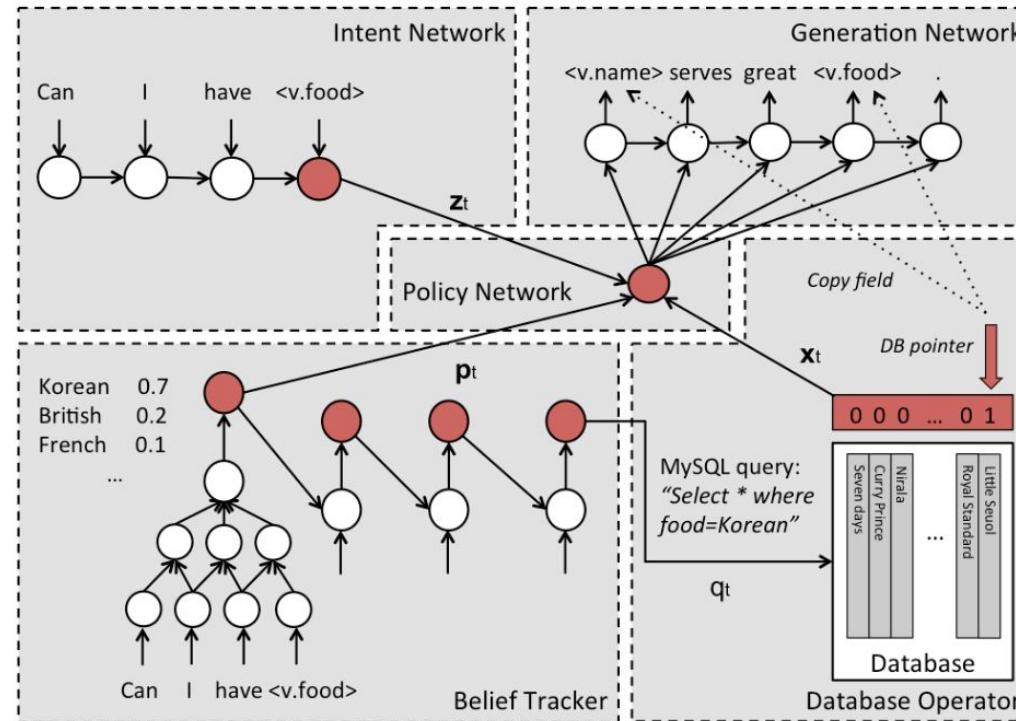
Cons:

- Let's not think about it for now

E2E Task-oriented modelling



Network-based End-to-End Task Oriented Dialogue System [Wen et al., 2017]



Network-based End-to-End Task Oriented Dialogue System

Metric	NDM	HDC	Tie
Subj. Success	96.95%	95.12%	-
Avg. # of Turn	3.95	4.54	-
Comparisons(%)			
Naturalness	46.95*	25.61	27.44
Comprehension	45.12*	21.95	32.93
Preference	50.00*	24.39	25.61
Performance	43.90*	25.61	30.49

* p <0.005, # of comparisons: 164

Reinforcement Learning for Fine Tuning

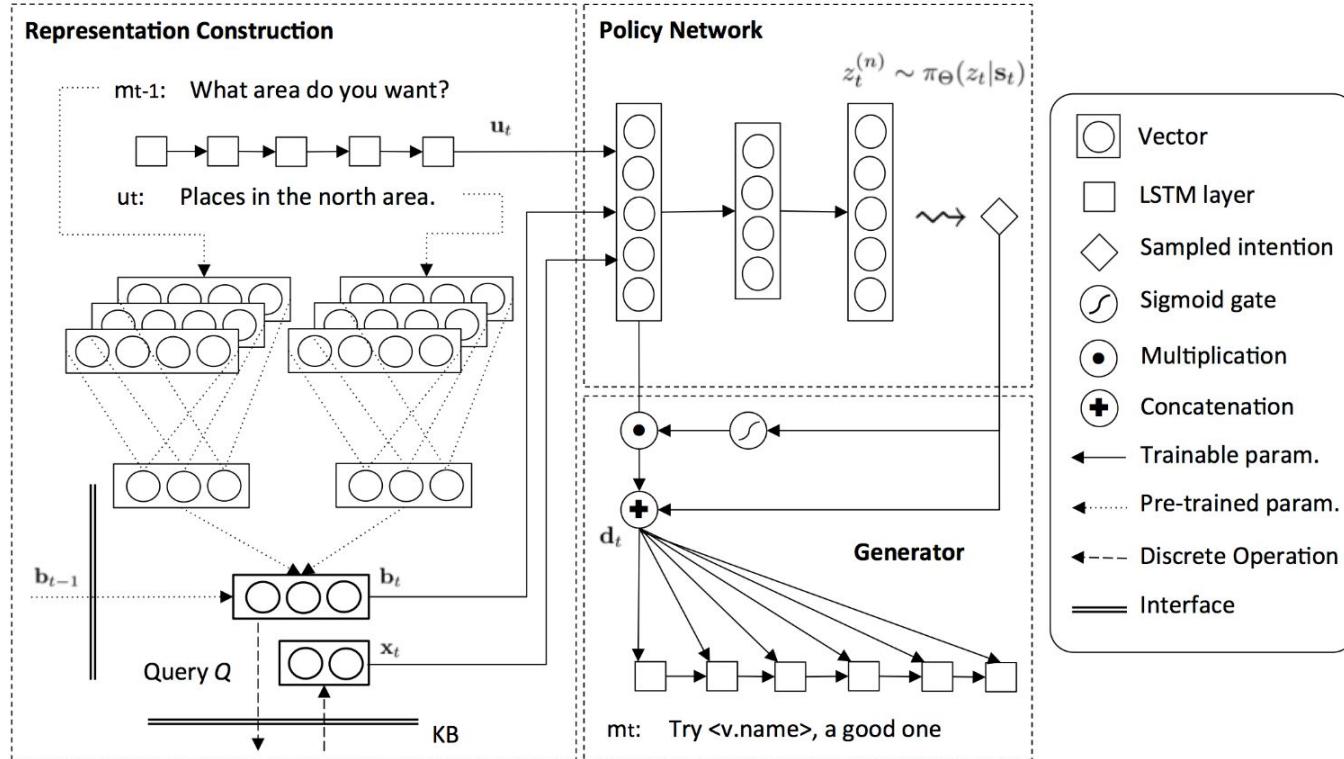
- Supervised Learning gets you as far as the data enables. However the model in production is often exposed to different environments.
 - This approach also suffers from lack of long-term planning: it hinders the task completion effectiveness.
-
- Fine-tuning with Reinforcement Learning promotes long-term planning.
 - But even more importantly allows to add external dialogue success signal into the training loop.

Reinforcement Learning for Fine Tuning

- Policy-gradient methods aim to optimize the policy with respect to the total reward through gradient ascent:

$$\nabla_{\theta} J(\theta) = \mathbb{E}_{\tau \sim \pi_{\theta}} \left[\sum_{t=0}^{T-1} \nabla_{\theta} \log \pi(a_t | s_t) R(\tau) \right]$$

Latent Intent Modelling [Wen et al., 2017]



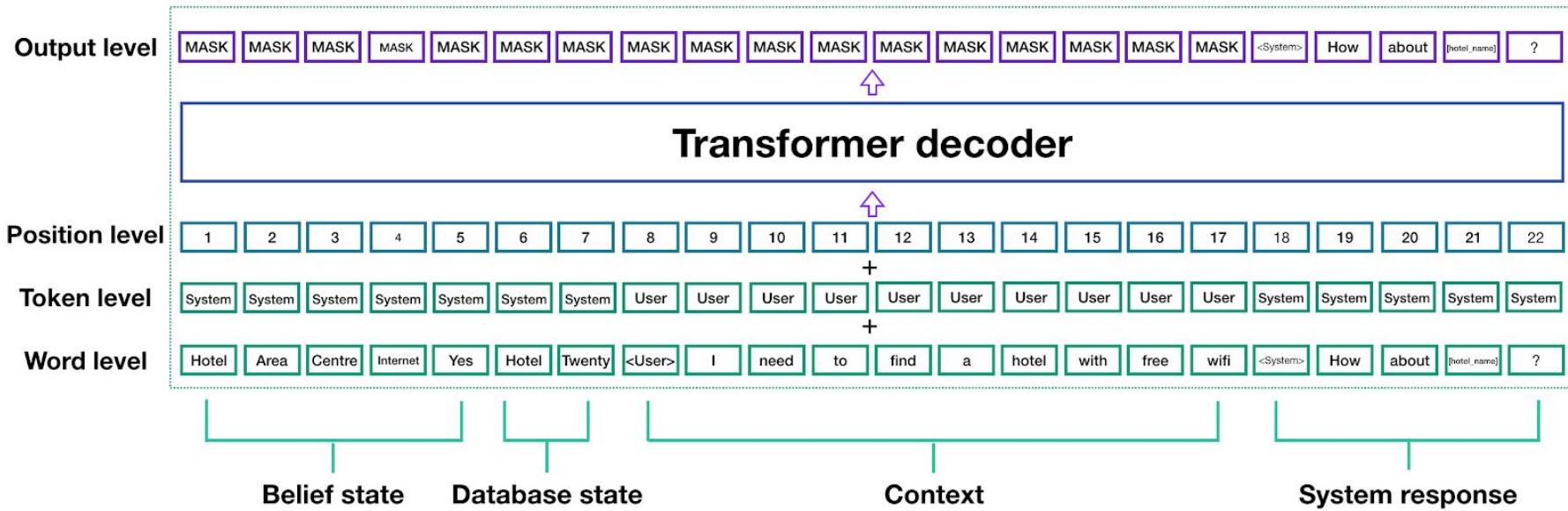
RL fine-tuning procedure

1. Freeze all modules except the policy network,
2. For each RL episode, randomly sample a batch of dialogs from the training set,
3. Run the model on every system turn, and do not alter the original dialog context at every turn given the generated responses,
4. Compute Success Rate based on the generated responses in this dialog,
5. Update the parameters through Policy Gradient.

Latent Intent Modelling [Wen et al., 2017]

Model	Decoder	Addition	Success(%)	BLEU	Suc.+0.5 BLEU
Vanilla NDM	lm	-	72.8%	0.237	0.847
	mem	-	74.3%	0.243	0.865
	hybrid	-	77.9%	0.231	0.894
Attentive NDM	lm	-	72.1%	0.246	0.844
	mem	-	80.1%	0.240	0.921
	hybrid	-	77.9%	0.234	0.896
Model	Intent Dim.	Addition	Success(%)	BLEU	Suc.+0.5 BLEU
LIDM	I=50	-	78.7%	0.226	0.900
	I=70	-	80.9%	0.245	0.932
	I=100	-	69.1%	0.221	0.801
	I=50	+RL	77.2%	0.249	0.896
	I=70	+RL	83.8%	0.258	0.967
	I=100	+RL	81.6%	0.245	0.939
Ground Truth	-	-	91.6%	1.000	1.416

Text as an Ultimate Input



Text as an Ultimate Input

The context consists of user and system utterances as well as
the dialogue belief state encoded as:

Domain1 Slot1 Value1 Slot2 Value2, Domain2 Slot1 ...

and the database state encoded as:

Domain1 # of entities, Domain2 # of entities, ...

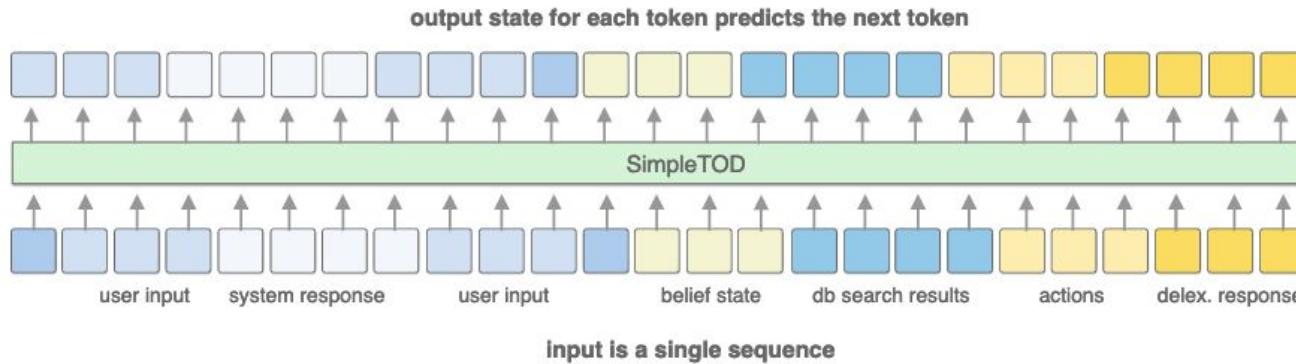
Text as an Ultimate Input

Model 1	vs		Model 2
GPT	59 %	41%	Baseline
GPT	46 %	54 %	Target
GPT2	46 %	54 %	Target
GPT2	45 %	55 %	Baseline
Baseline	43 %	57 %	Target
GPT2	51 %	49 %	GPT

A Simple Language Model for Task-Oriented Dialogue

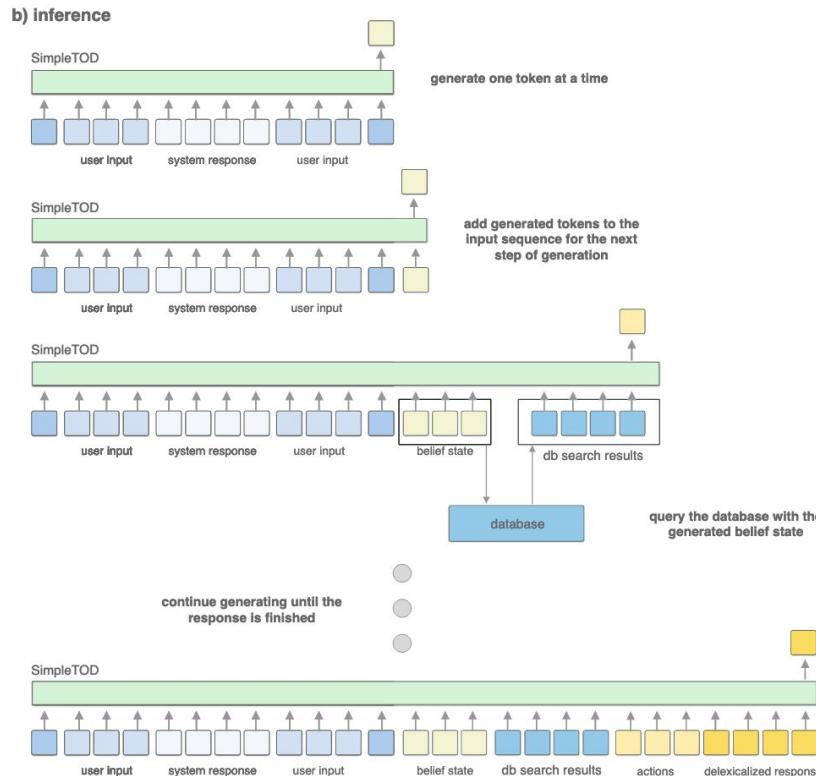
[Hosseini-Asl, 2020]

a) training



A Simple Language Model for Task-Oriented Dialogue

[Hosseini-Asl, 2020]



A Simple Language Model for Task-Oriented Dialogue

[Hosseini-Asl, 2020]

Model	Belief State	DB Search	Action	Inform	Success	BLEU	Combined
DAMD+augmentation	generated	oracle	generated	76.3	60.4	16.6	85
SimpleTOD (ours)	generated	oracle	generated	78.1	63.4	16.91	87.66
SimpleTOD (ours)	generated	dynamic	generated	81.4	69.7	16.11	91.66
SimpleTOD (ours)	generated	-	generated	84.4	70.1	15.01	92.26

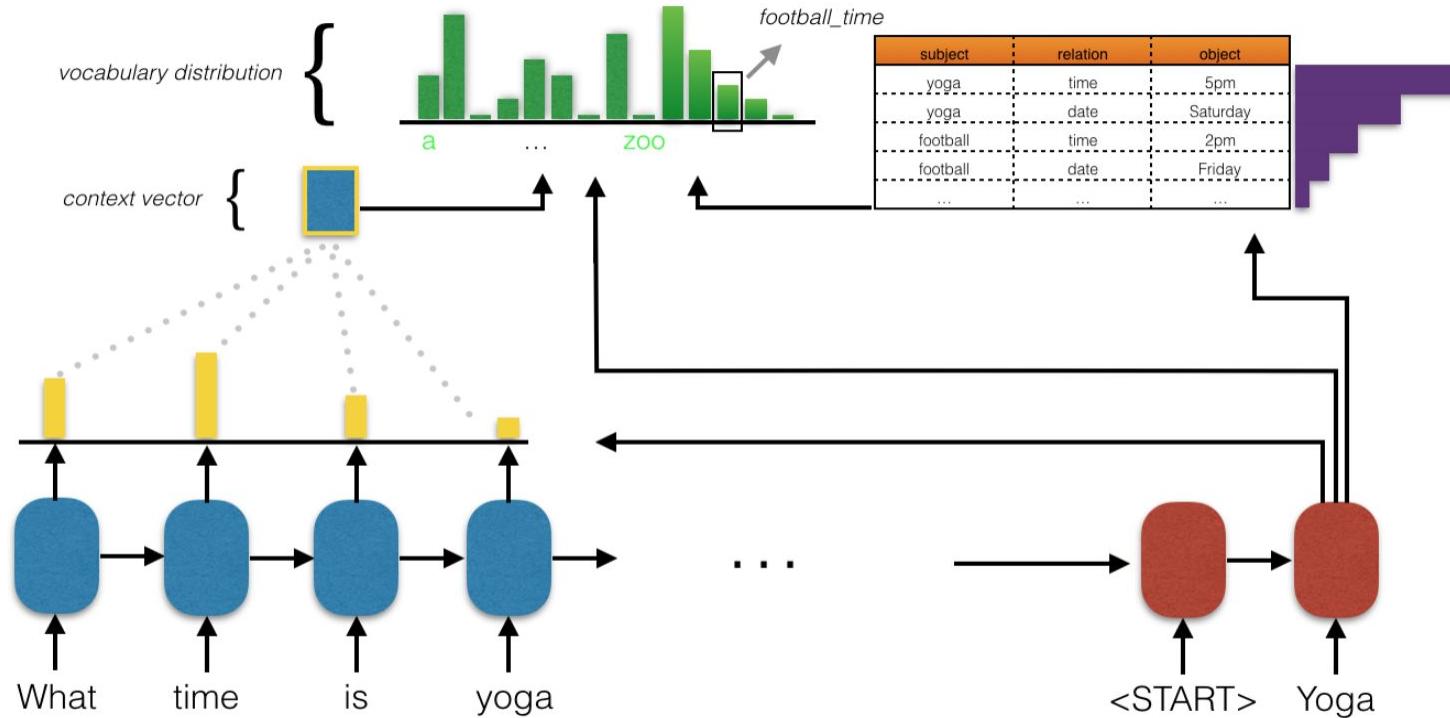
Back to DST!

Model	Decoder	Context Encoder	Extra Supervision	Joint Accuracy
TRADE*	Generative + Classifier	Bidirectional	-	45.6
DSTQA**	Classifier	Bidirectional	knowledge graph	51.17
DST-Picklist*	Classifier	Bidirectional	-	53.3
SST*	Generative	Bidirectional	schema graph	55.23
TripPy [†]	Classifier	Bidirectional	action decision	55.3
SimpleTOD [◦]	Generative	Unidirectional	-	55.72
SimpleTOD*	Generative	Unidirectional	-	55.76
SimpleTOD ⁺	Generative	Unidirectional	-	57.47

These nasty databases/APIs ..

- In a classical approach, we need to issue a symbolic query to the Knowledge Base to retrieve helpful entries given some attributes.
- The symbolic operations breaks, however, the differentiability preventing from truly “end-to-end” learning.
- This problem has been addressed by soft-lookup approaches.

Key-retrieval Networks [Eric and Manning, 2017]



Key-retrieval Networks [Eric and Manning, 2017]

Model	BLEU	Ent. F ₁	Scheduling Ent. F ₁	Weather Ent. F ₁	Navigation Ent. F ₁
Rule-Based	6.6	43.8	61.3	39.5	40.4
Copy Net	11.0	37.0	28.1	50.1	28.4
Attn. Seq2Seq	10.2	30.0	30.0	42.4	17.9
KV Retrieval Net (no enc. attn.)	10.8	40.9	59.5	35.6	36.6
KV Retrieval Net	13.2	48.0	62.9	47.0	41.3
<i>Human Performance</i>	13.5	60.7	64.3	61.6	55.2

Literature

1. Vinyals and Le, 2015 - <https://arxiv.org/pdf/1506.05869.pdf>
2. Serban et al., 2016 - <https://arxiv.org/abs/1507.04808>
3. Wen et al., 2016 - <https://arxiv.org/abs/1604.04562>
4. Budzianowski et al., 2018 - <https://arxiv.org/abs/1810.00278>
5. Roller et al., 2020 - <https://arxiv.org/abs/2004.13637>
6. Thoppilan et al., 2022 - <https://arxiv.org/abs/2201.08239>
7. Fitzgerald et al., 2022 - <https://arxiv.org/abs/2204.08582>