Neural Methods

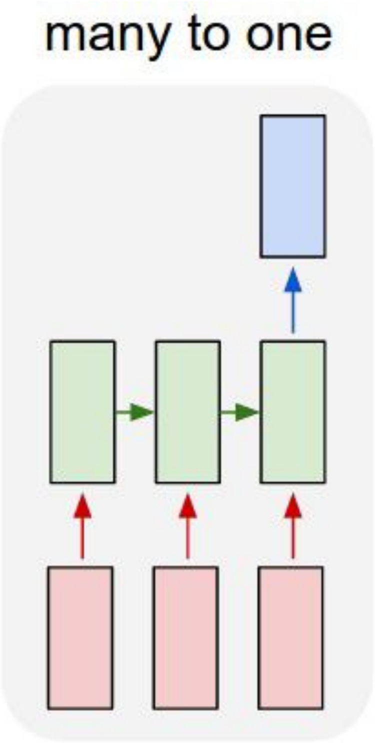
BIME 591

Review

Intent Classification Methods

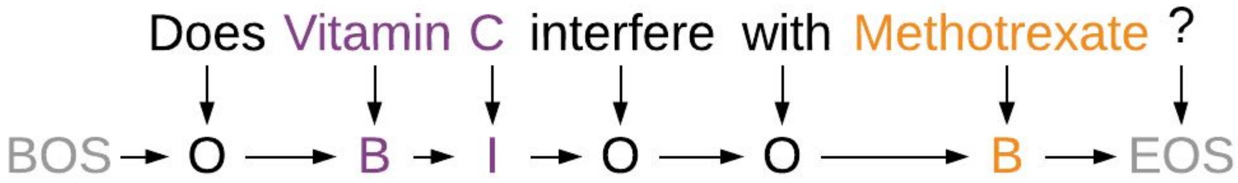
Input: Word Sequence **Output:** Intent Class

- Index all words to numerical representation, i.e. hospital □ 324
- Represent each word as a vector using a pre-trained embedding matrix.
- Use favorite classification model LR, CNN, LSTM, Transformer



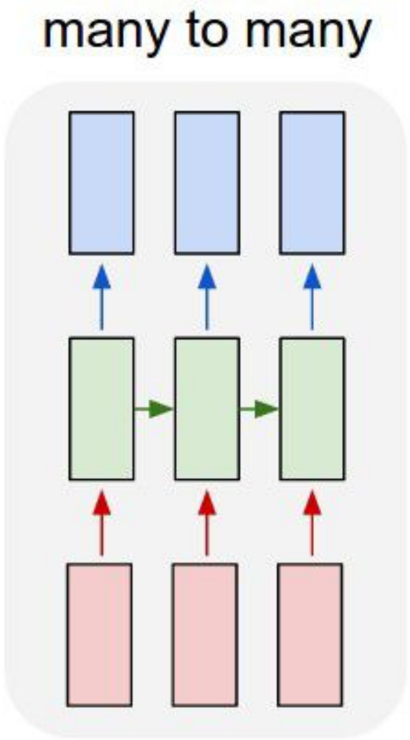
Slot Tagging

Input: Word Sequence **Output:** BIO Tags


Does Vitamin C interfere with Methotrexate ?
BOS → O → B → I → O → O → B → EOS

Choose favorite decoder: CRF, LSTM, Transformer.

Note: x-CRF Conditional Random Field (CRF) decoder with external knowledge.

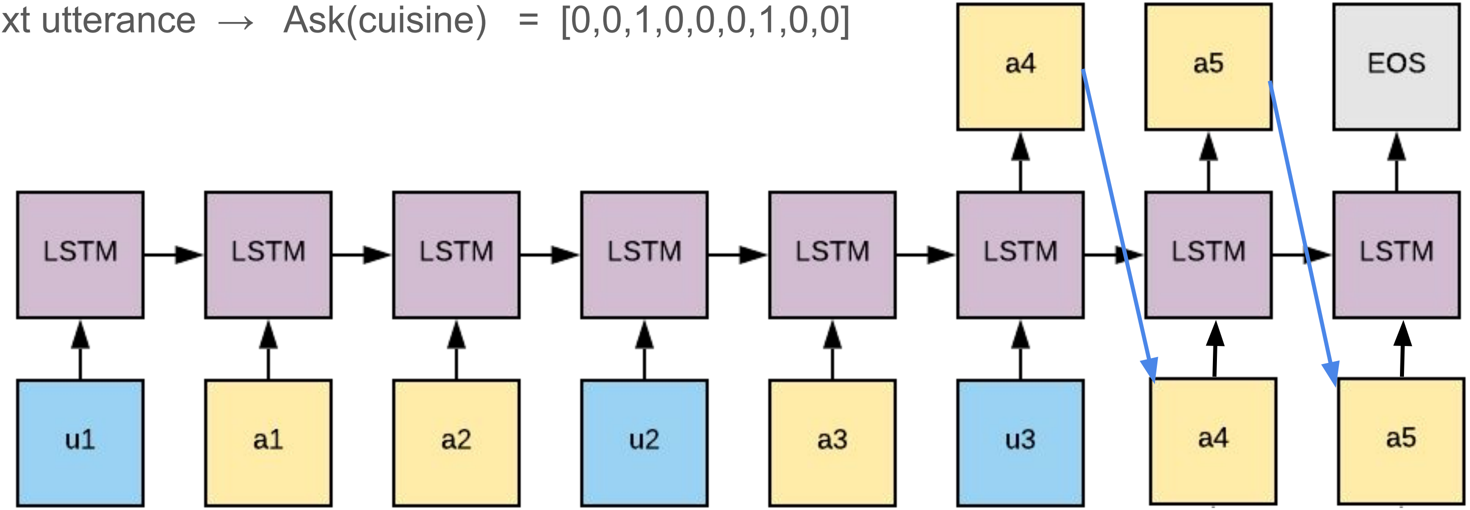


Dialog Management

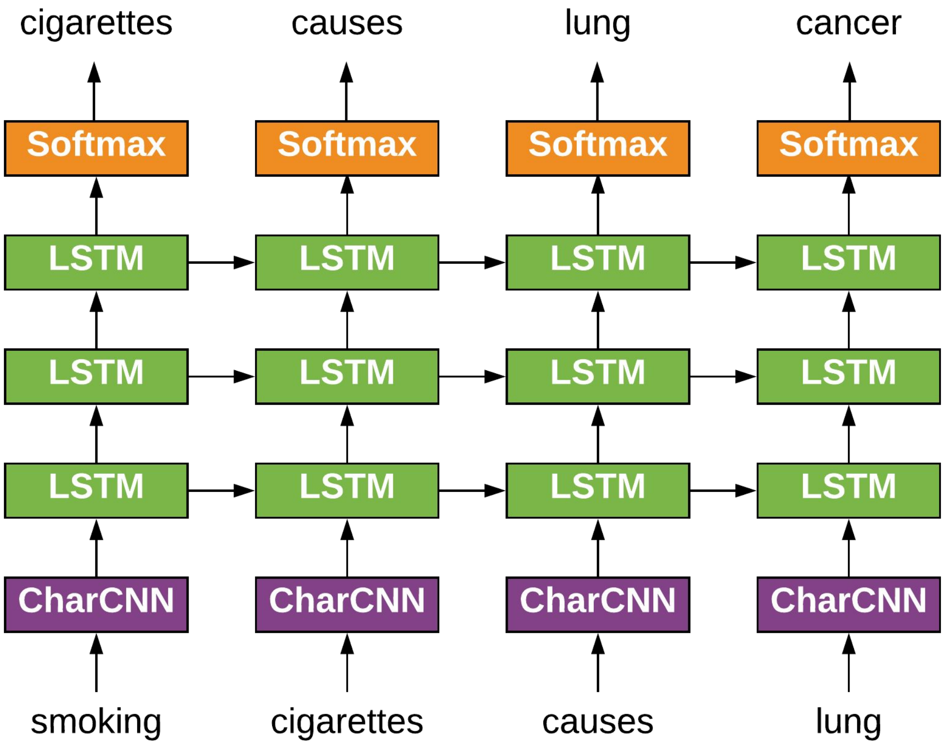
Task: Predict response based on utterance.

Represent utterance → Inform(place) = [0,1,0,0,0,0,0,1,0]

Predict next utterance → Ask(cuisine) = [0,0,1,0,0,0,1,0,0]

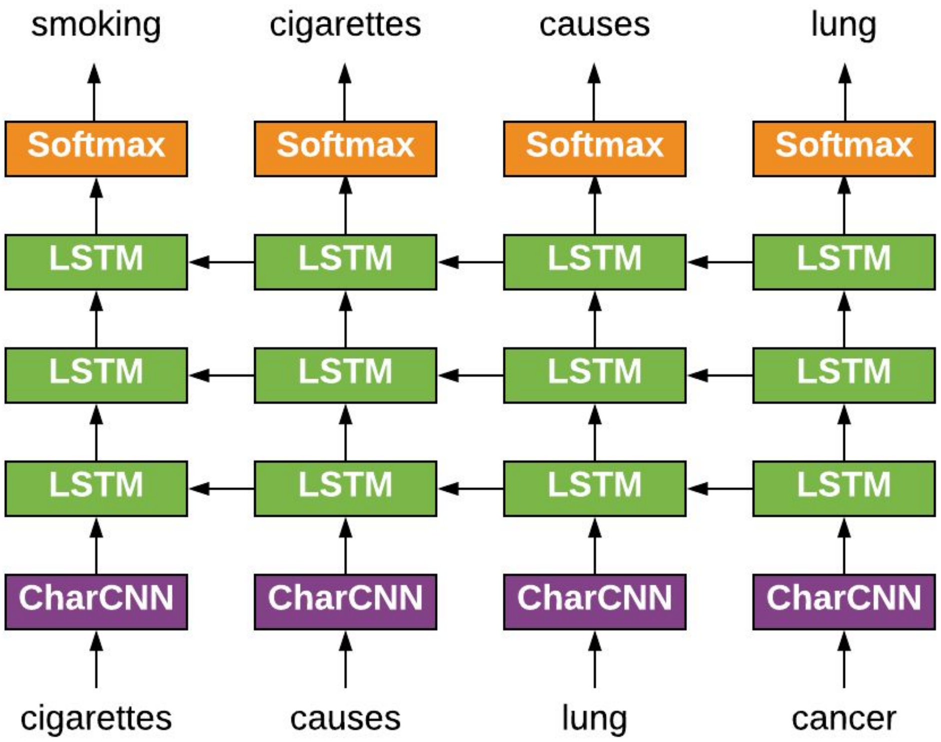


Neural Language Model



$$p(x_1, \ldots, x_T) = \prod_{t=1}^{T} p(x_t | x_{1:t-1})$$

Neural Language Model



$$p(x_1, \dots, x_T) = \prod_{t=1}^{T} p(x_t | x_{t+1:T})$$

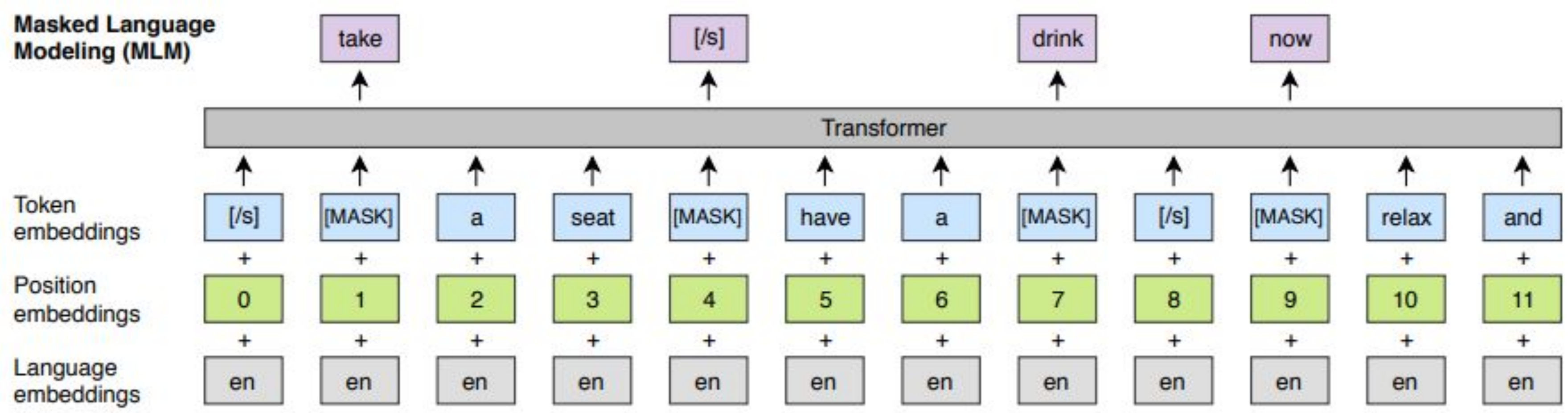
Transformer

Difference from LSTM models

- Use self-attention as primary computation
- Improved computational performance
- Improved accuracy

https://ai.googleblog.com/2017/08/transformer-novel-neural-network.html

Masked Language Model



Lample G, Conneau A. Cross-lingual Language Model Pretraining. 2019. http://arxiv.org/abs/1901.07291.

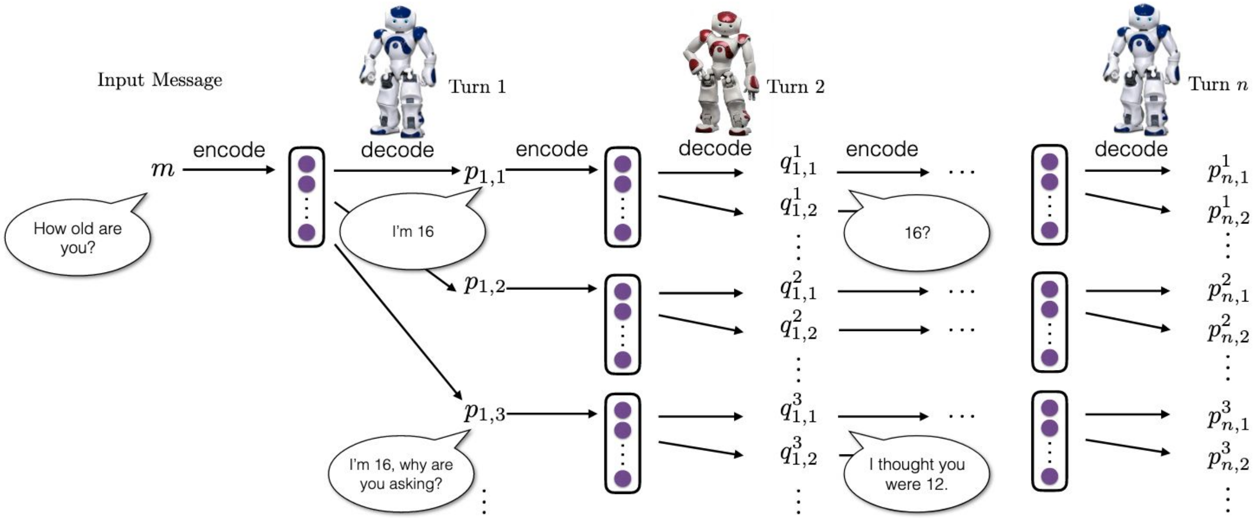
End-to-End Models

Reinforcement Learning for E2E Dialog

LSTM encoder-decoder

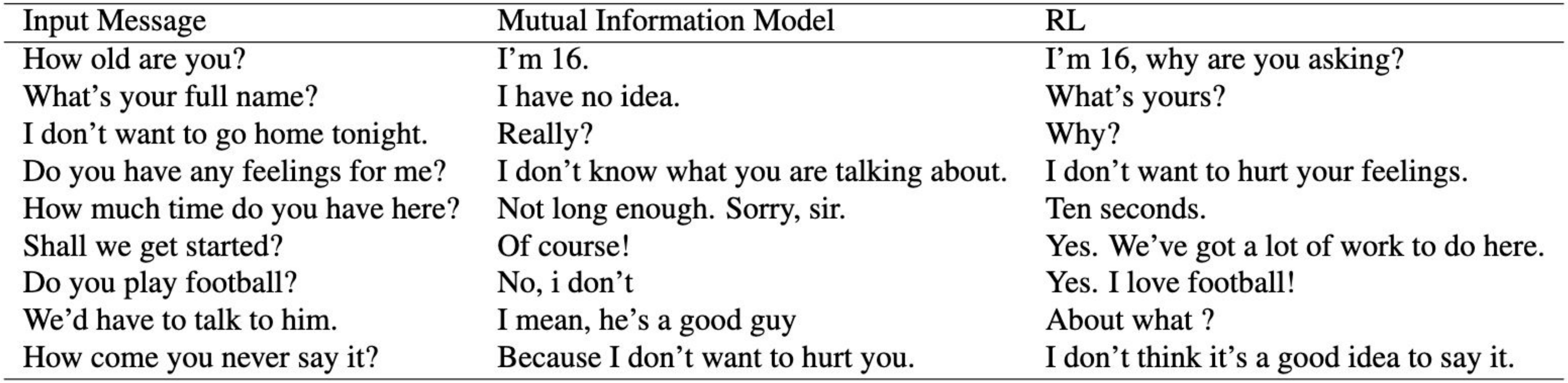
Reward based on:

- Suitability of a dull response to the selected response
- New information added
- Increase semantic coherence



Li J, Monroe W, Ritter A, Galley M, Gao J, Jurafsky D. Deep Reinforcement Learning for Dialogue Generation. 2016;(4). doi:10.1103/PhysRevB.94.020410

Reinforcement Learning for E2E Dialog



Input Message	Mutual Information Model	RL
How old are you?	I’m 16.	I’m 16, why are you asking?
What’s your full name?	I have no idea.	What’s yours?
I don’t want to go home tonight.	Really?	Why?
Do you have any feelings for me?	I don’t know what you are talking about.	I don’t want to hurt your feelings.
How much time do you have here?	Not long enough. Sorry, sir.	Ten seconds.
Shall we get started?	Of course!	Yes. We’ve got a lot of work to do here.
Do you play football?	No, i don’t	Yes. I love football!
We’d have to talk to him.	I mean, he’s a good guy	About what ?
How come you never say it?	Because I don’t want to hurt you.	I don’t think it’s a good idea to say it.

Li J, Monroe W, Ritter A, Galley M, Gao J, Jurafsky D. Deep Reinforcement Learning for Dialogue Generation. 2016;(4). doi:10.1103/PhysRevB.94.020410

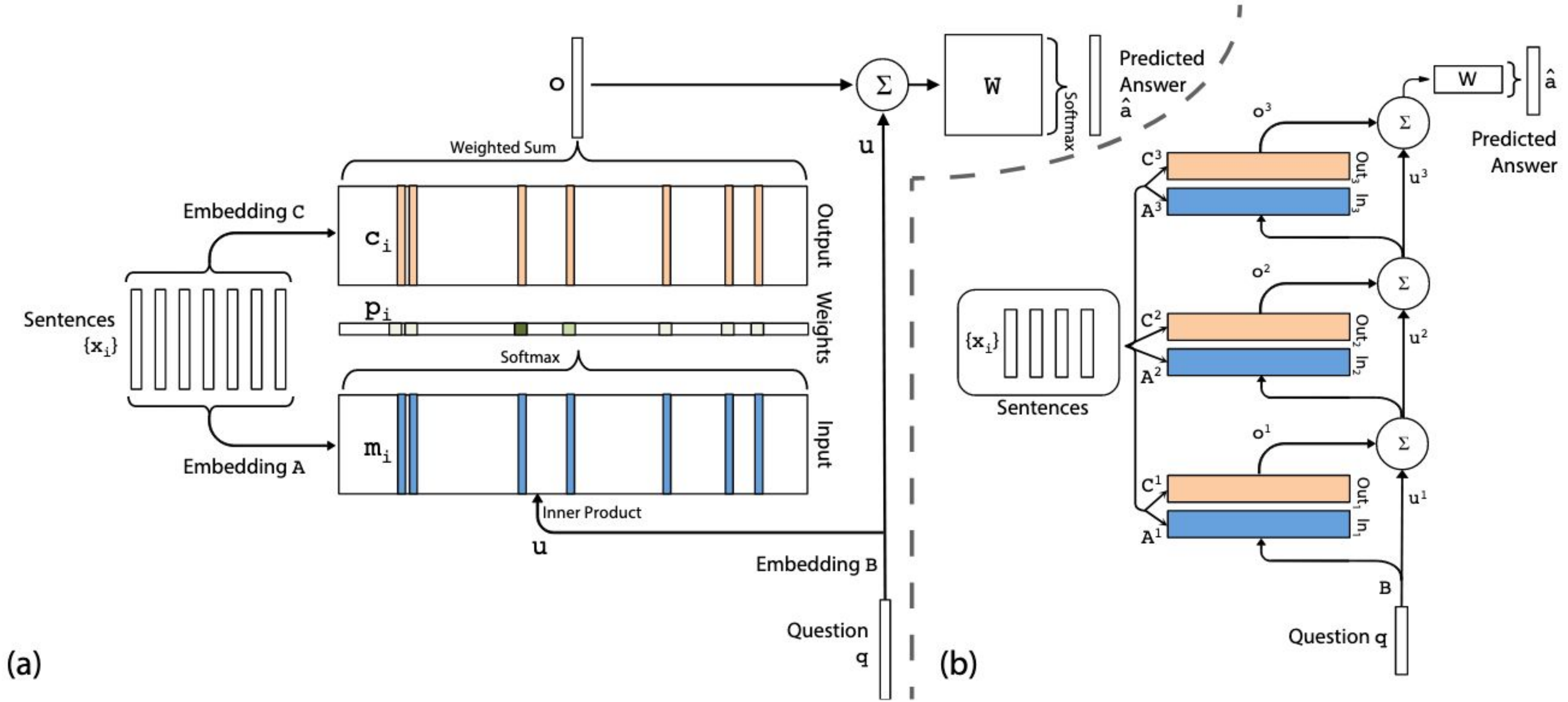
Retrieval Models

Retrieval-Based Models

- No NLG
  - Control over responses
  - Grammatically correct
  - Content is correct, e.g. clinical recommendations

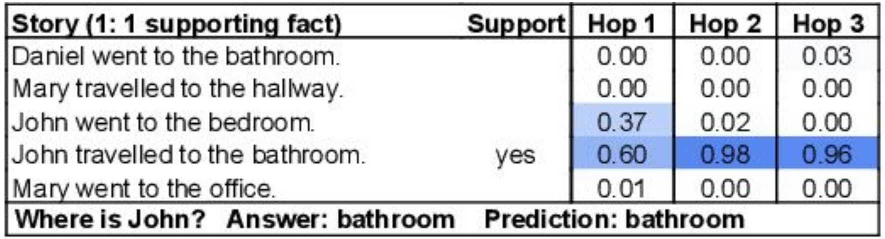
- No task-specific ontology
  - NLU is combined with response selection (a single action)

End-to-End Memory Neural Networks

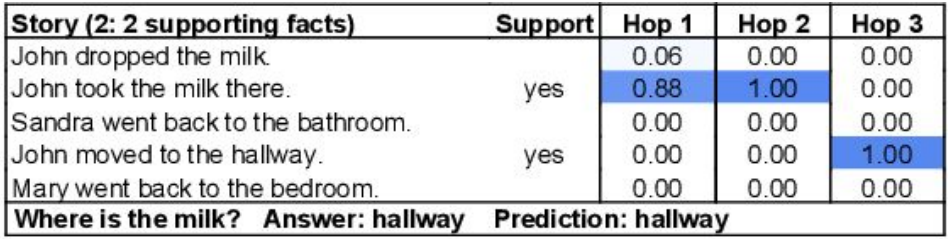


Sukhbaatar S, Szlam A, Weston J, Fergus R. End-To-End Memory Networks. 2015.

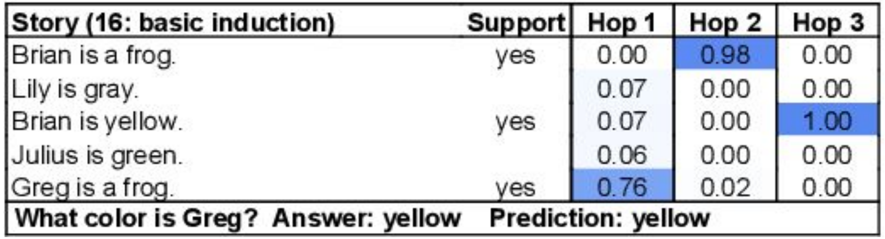
End-to-End Memory Neural Networks



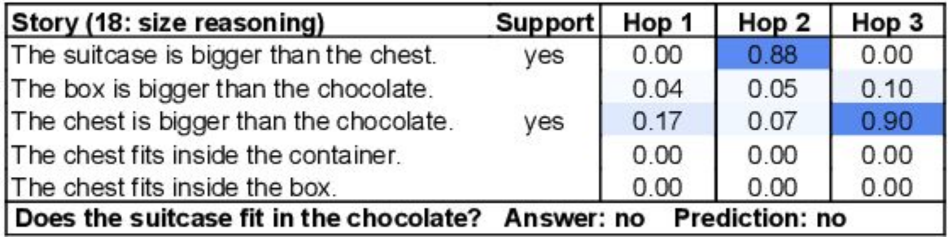
Story (1: 1 supporting fact)	Support	Hop 1	Hop 2	Hop 3
Daniel went to the bathroom.		0.00	0.00	0.03
Mary travelled to the hallway.		0.00	0.00	0.00
John went to the bedroom.		0.37	0.02	0.00
John travelled to the bathroom.	yes	0.60	0.98	0.96
Mary went to the office.		0.01	0.00	0.00
Where is John? Answer: bathroom Prediction: bathroom				



Story (2: 2 supporting facts)	Support	Hop 1	Hop 2	Hop 3
John dropped the milk.		0.06	0.00	0.00
John took the milk there.	yes	0.88	1.00	0.00
Sandra went back to the bathroom.		0.00	0.00	0.00
John moved to the hallway.	yes	0.00	0.00	1.00
Mary went back to the bedroom.		0.00	0.00	0.00
Where is the milk? Answer: hallway Prediction: hallway				



Story (16: basic induction)	Support	Hop 1	Hop 2	Hop 3
Brian is a frog.	yes	0.00	0.98	0.00
Lily is gray.		0.07	0.00	0.00
Brian is yellow.	yes	0.07	0.00	1.00
Julius is green.		0.06	0.00	0.00
Greg is a frog.	yes	0.76	0.02	0.00
What color is Greg? Answer: yellow Prediction: yellow				



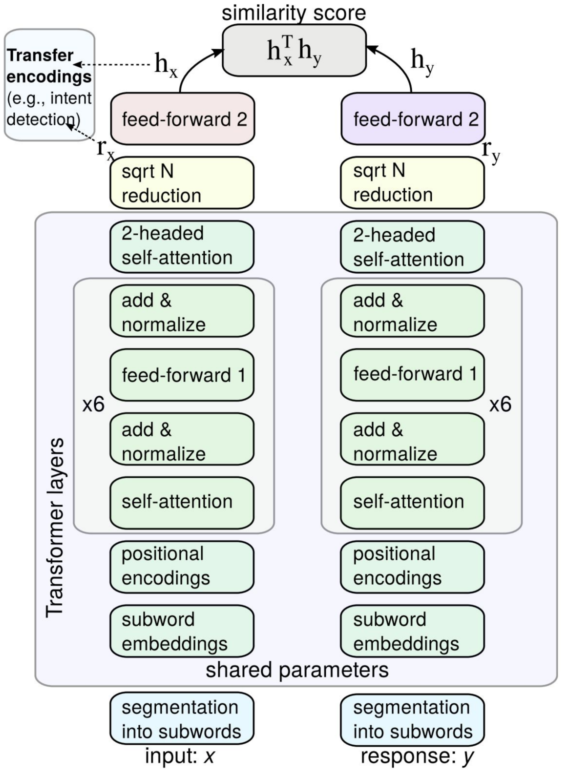
Story (18: size reasoning)	Support	Hop 1	Hop 2	Hop 3
The suitcase is bigger than the chest.	yes	0.00	0.88	0.00
The box is bigger than the chocolate.		0.04	0.05	0.10
The chest is bigger than the chocolate.	yes	0.17	0.07	0.90
The chest fits inside the container.		0.00	0.00	0.00
The chest fits inside the box.		0.00	0.00	0.00
Does the suitcase fit in the chocolate? Answer: no Prediction: no				

Sukhbaatar S, Szlam A, Weston J, Fergus R. End-To-End Memory Networks. 2015.

ConveRT

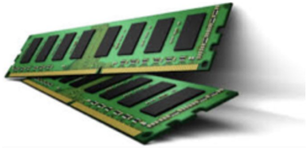
A transfer learning approach to improve domain-specific dialog models

- Shared semantic space for both query and response
- Uses a simplified transformer-based sister network architecture
- Pre-train: Reddit corpus
- Test: 5 domain datasets



Henderson M, Casanueva I, Mrkšić N, Su P-H, Tsung-Hsien, Vulić I. ConveRT: Efficient and Accurate Conversational Representations from Transformers. November 2019. http://arxiv.org/abs/1911.03688.

Model Size Reduction

1  **= 100** ✖✚

Knowledge Distillation (Hinton 2015):

- Train a simpler model to predict the logits of the more complex model prior to activation function

Quantization (Han 2016)

- Convert the 32-bit float values used during training to 8-bit integers for inference

Hinton G, Vinyals O, Dean J. Distilling the Knowledge in a Neural Network. 2015:1-9. http://arxiv.org/abs/1503.02531.

Han S, Mao H, Dally WJ. Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding. *4th Int Conf Learn Represent ICLR 2016 - Conf Track Proc*. 2016:1-14.

Exercise

Understanding the ConveRT Model

Appendix

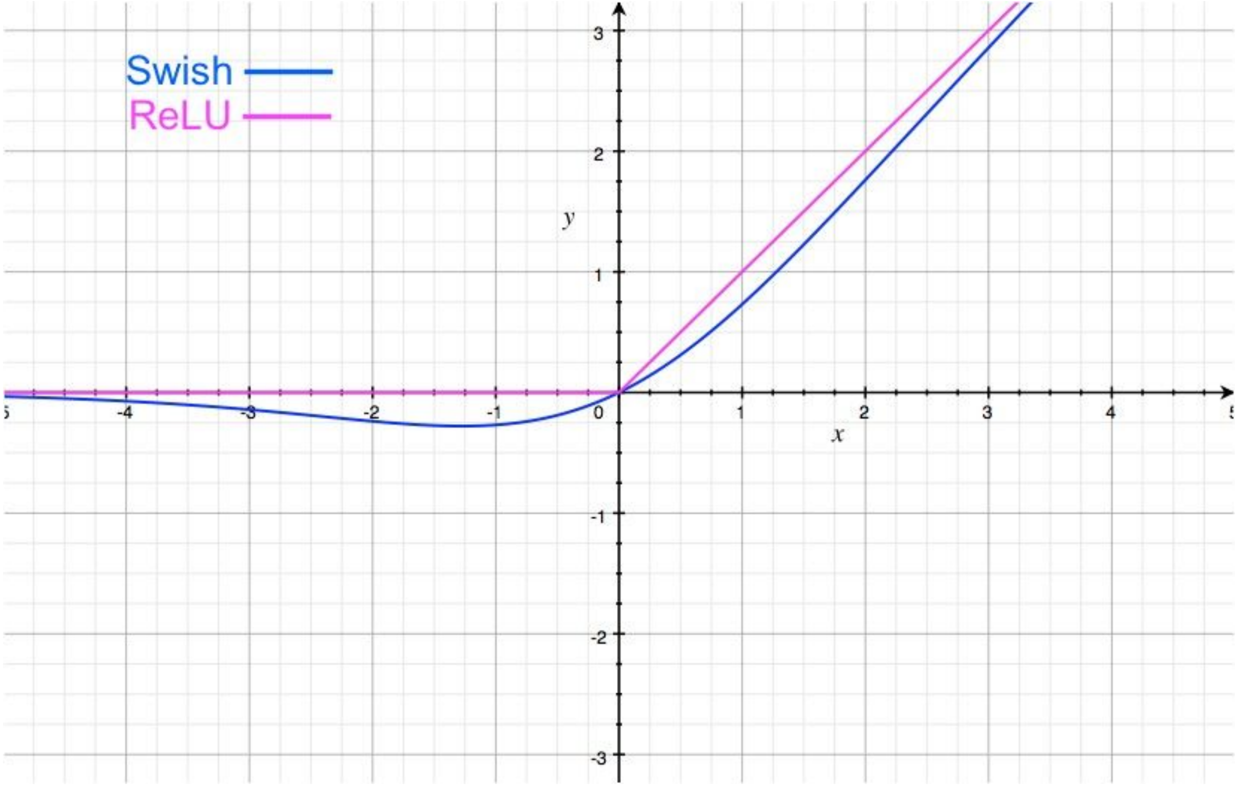
Activation Function

Relu

f(x) = max(x,0)

Swish

f(x) = x·sigmoid(βx)



Zoph B, Le Q V. Searching for activation functions. *6th Int Conf Learn Represent ICLR 2018 - Work Track Proc*. 2018:1-13.

Unigram and Bigram Features

I need to set an appointment for with my doctor.

Unigrams

I
need
to
set
an
appointment
with
my
doctor
.

Bigrams

I need
need to
to set
set an
an appointment
appointment with
with my
my doctor
doctor .

Subword Features

When do I need to go see my doctor?

Subword (ws=2)

wh
he
en
do
I
ne
ee
ed
to
go
..

When am I going to see my doctor?

Subword (ws=2)

wh
he
en
am
I
go
oi
in
ng
to
..