Using adaptive knowledge graphs in neural dialogue generation

Patrik Purgai

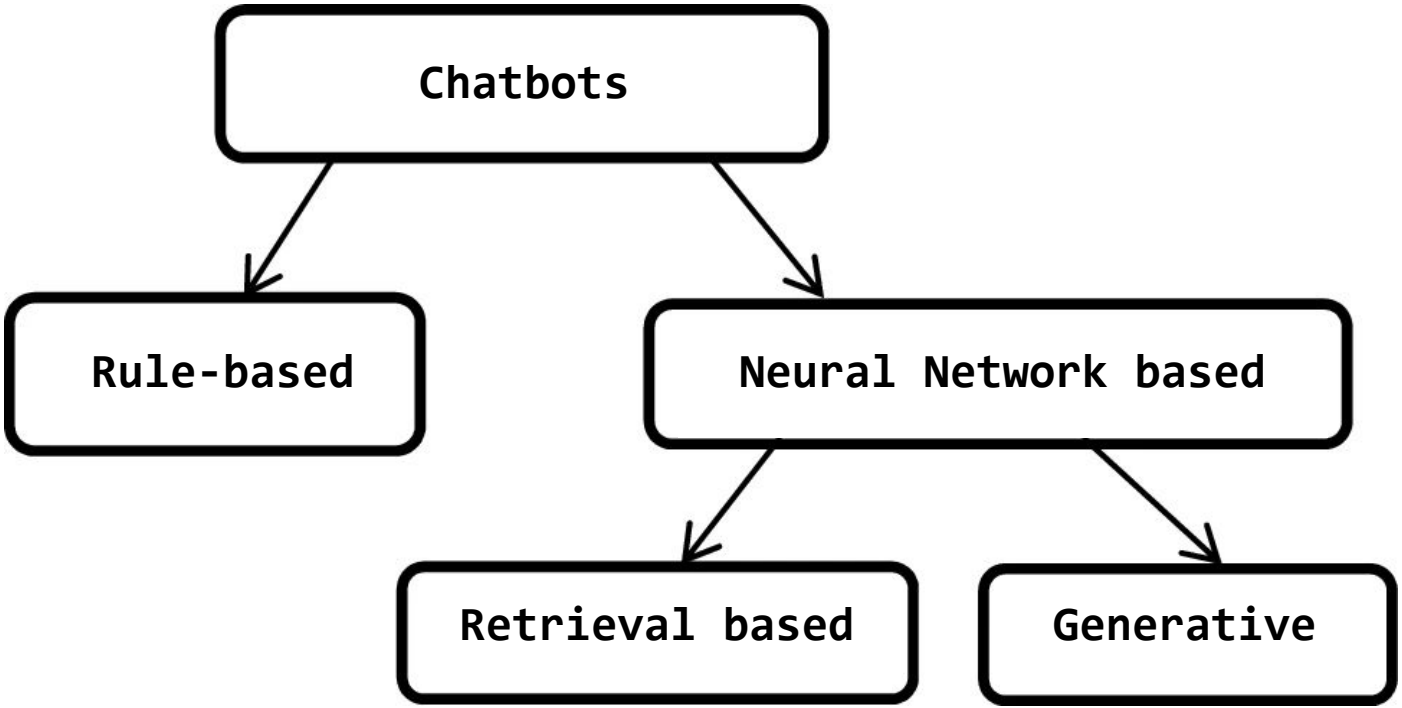
Chatbots

Domain specific

- Rasa
- Luis
- Dialogflow

Open domain

- Blender
- Cleverbot


Chatbots
Rule-based
Neural Network based
Retrieval based
Generative

Rule-based

- Pattern matching - ELIZA
- AIML (Artificial Intelligence Markup Language) - Mitsuku
- Chatscript - SUZETTE

AIML

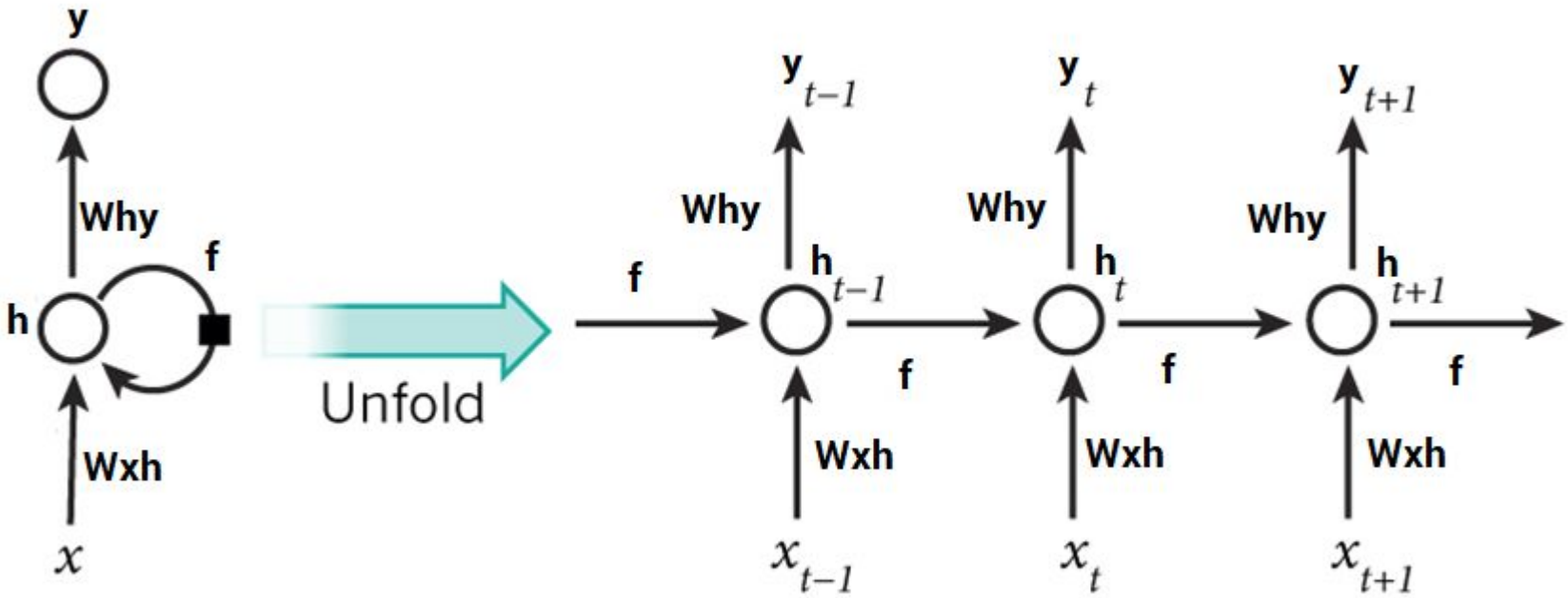
```
<category>
   <pattern>MY NAME IS *</pattern>
   <template>Hi there, my name is Andrew.</template>
</category>
```

```
<category>
   <pattern>GUESS A NUMBER</pattern>
   <template>
      <random>
         <li>1</li>
         <li>6</li>
         <li>227</li>
      </random>
   </template>
</category>
```

Neural Network based

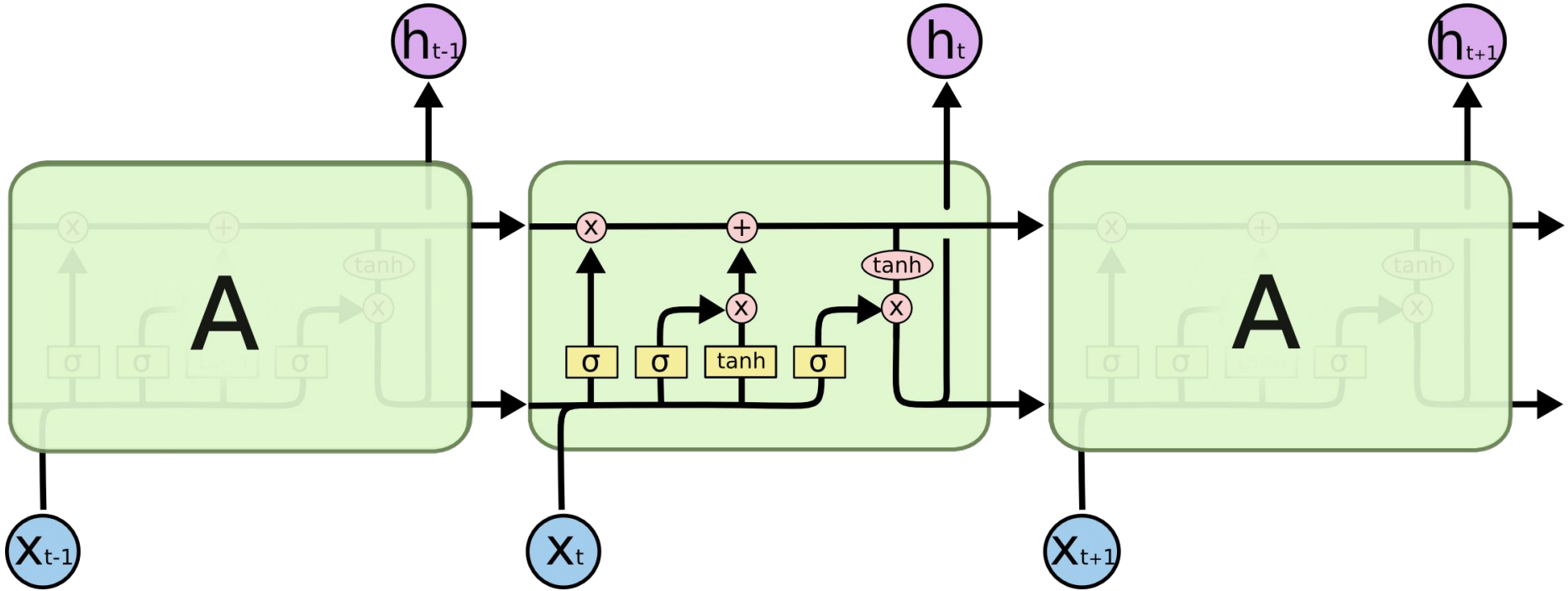
- RNN (LSTM)
- Transformer
- BERT
- GPT-2

RNN


y
Why
f
h
Wxh
x
Unfold
y t−1
y t
y t+1
Why
Why
Why
f
h t−1
h t
h t+1
f
f
f
Wxh
Wxh
Wxh
x t−1
x t
x t+1

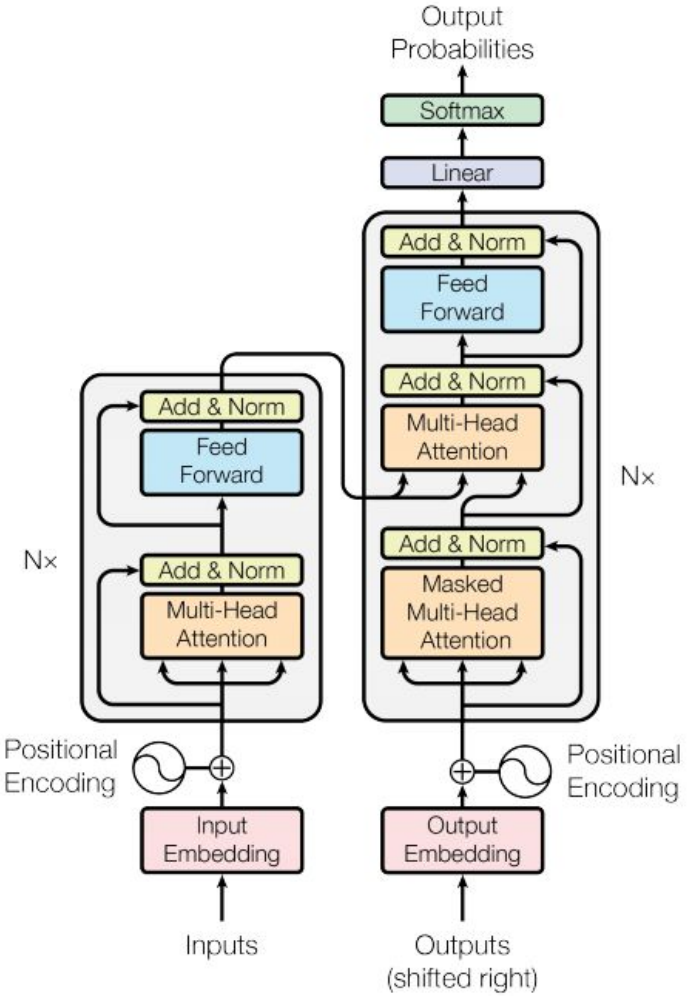
LSTM

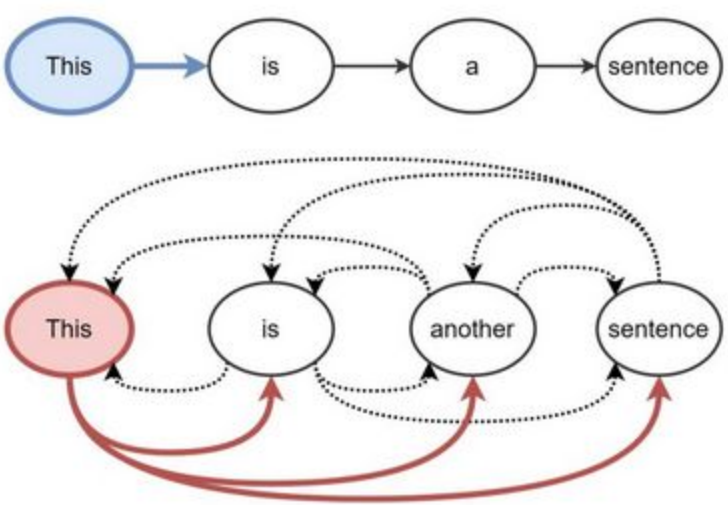
Hochreiter & Schmidhuber (1997)



Transformer

Vaswani et al. (2017)





RNN

Translation?

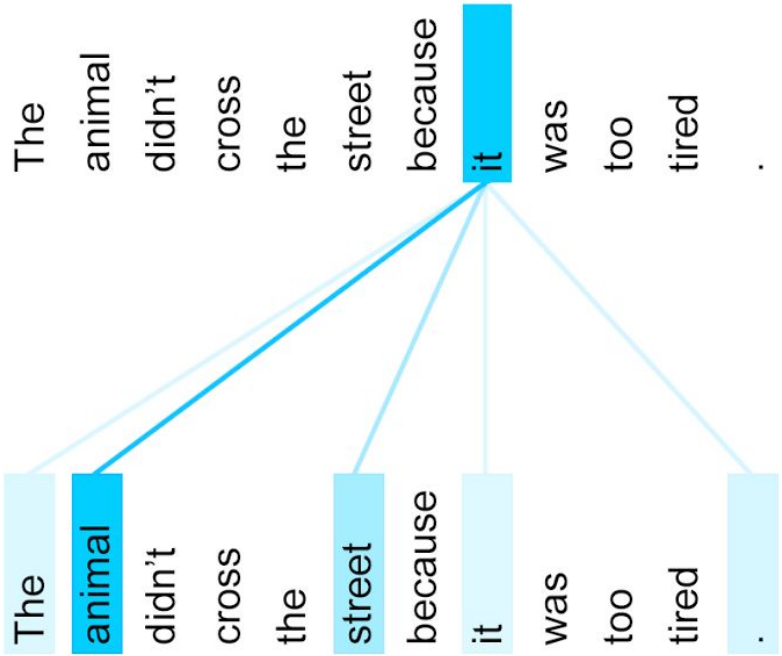
Sentiment?

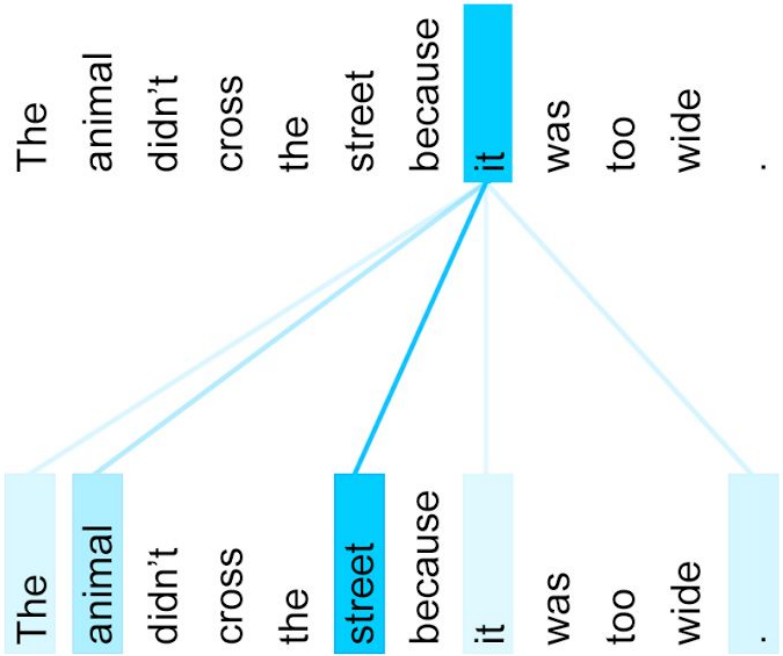
Next word?

Transf.

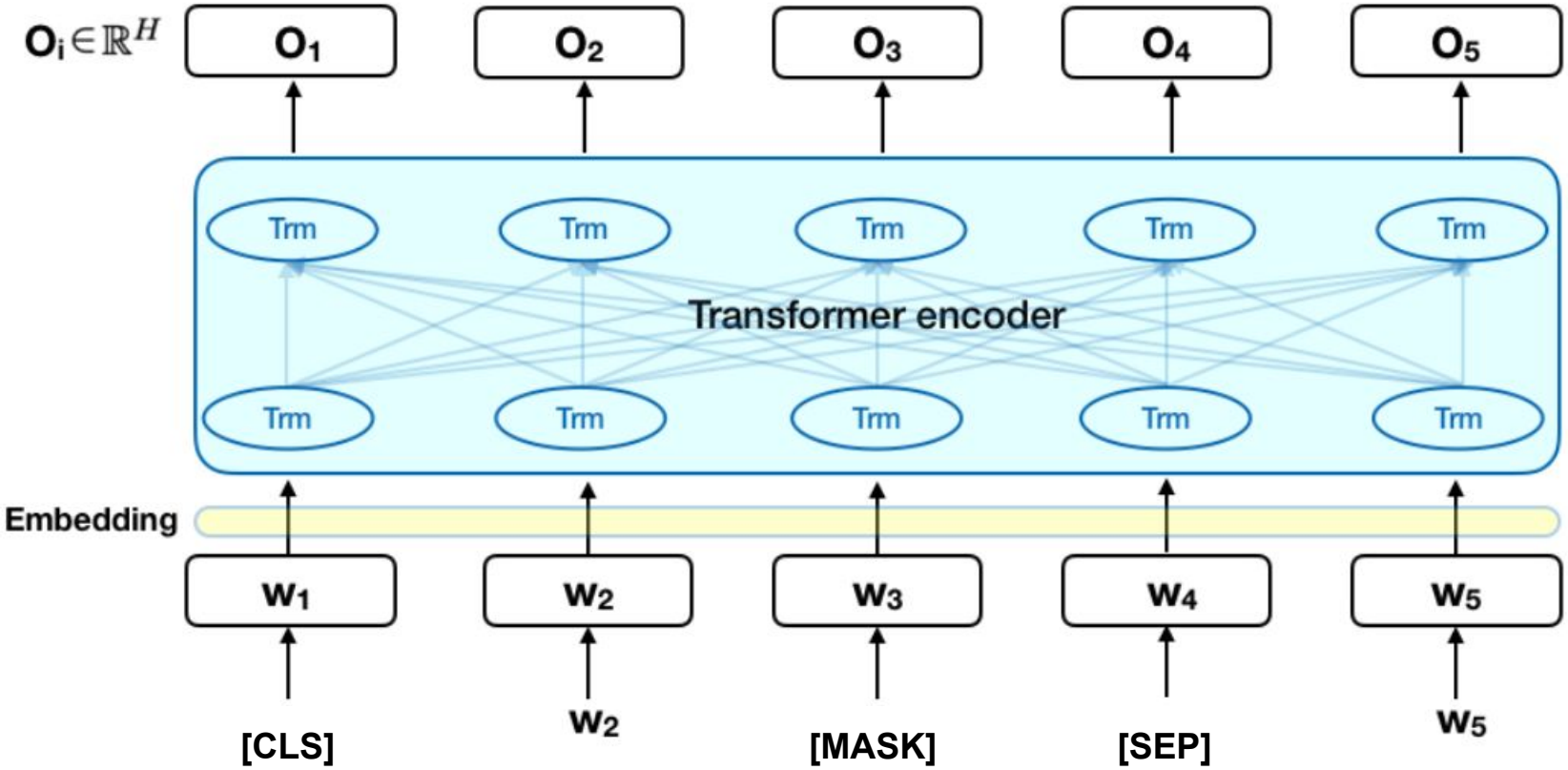
Part-of-speech tags?

Attention Mechanism

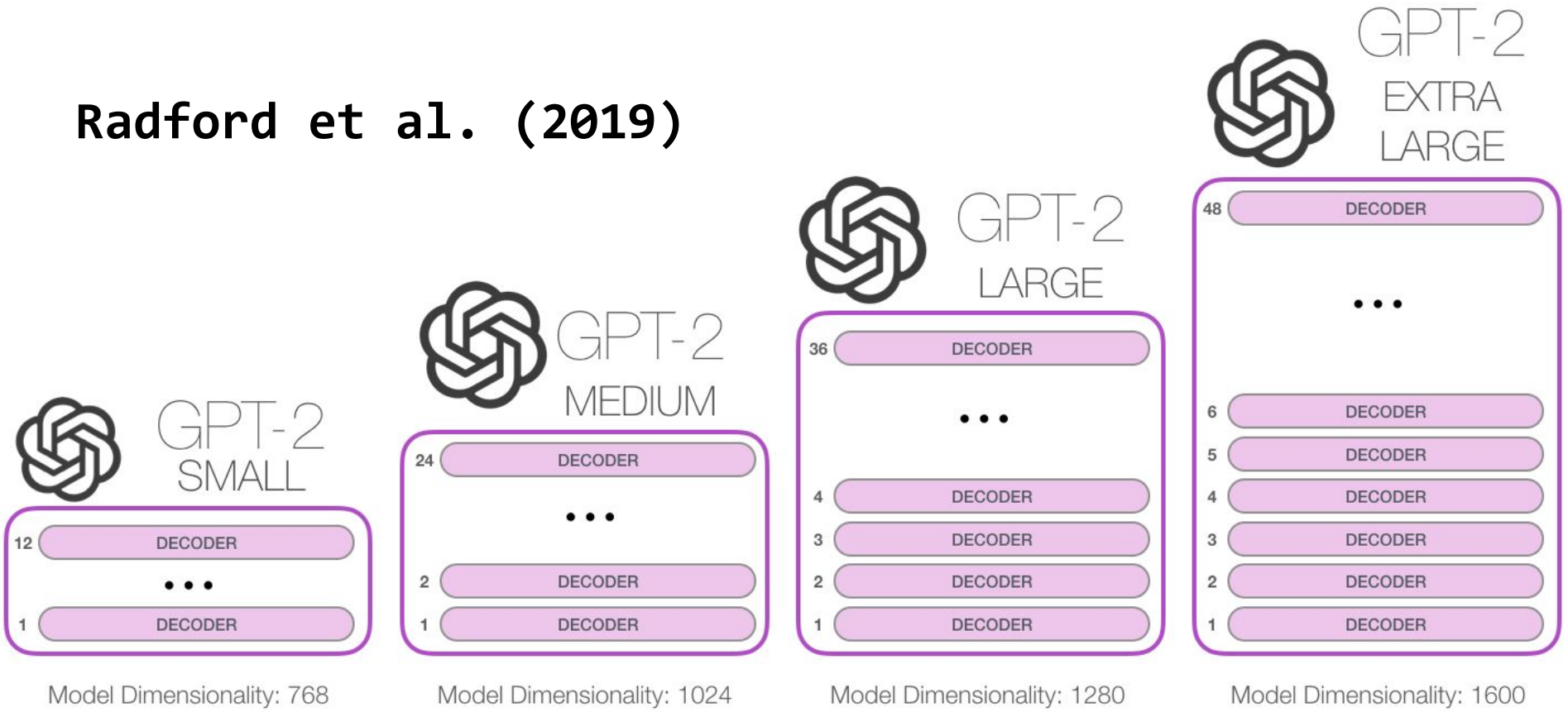




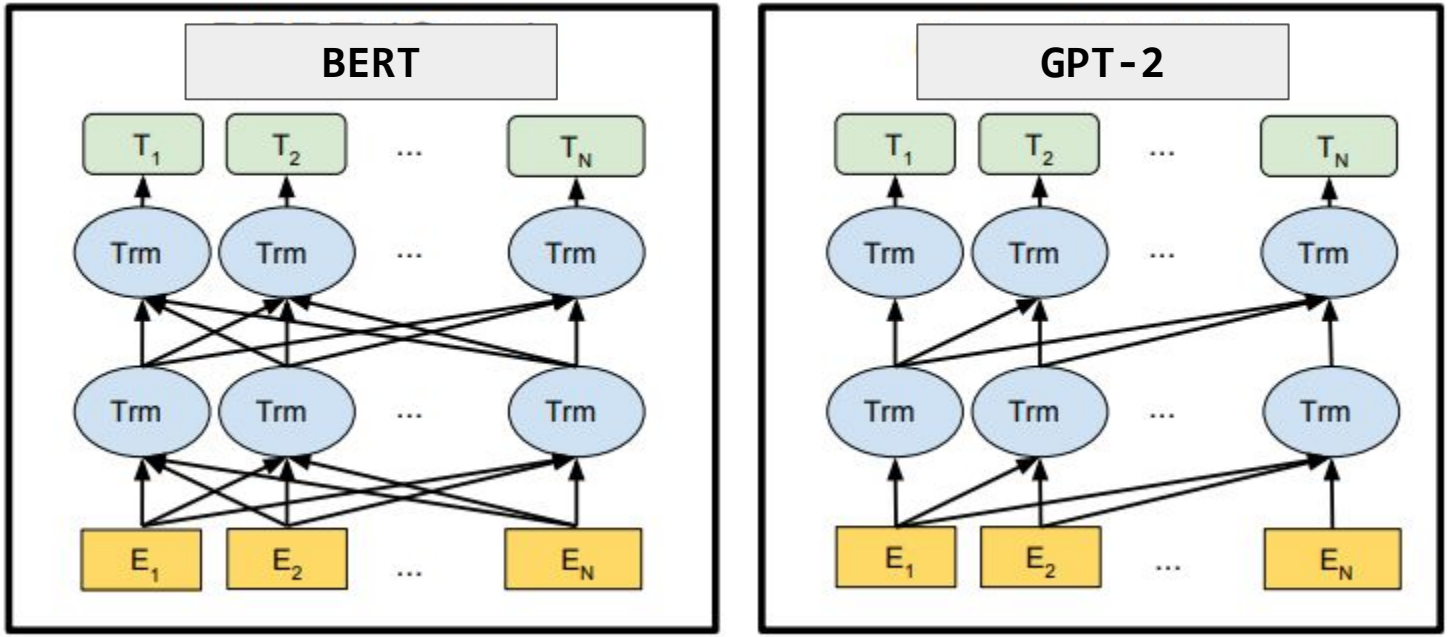
BERT


$O_i \in \mathbb{R}^H$
O1
O2
O3
O4
O5
Trm
Transformer encoder
Embedding
W1
W2
W3
W4
W5
[CLS]
W2
[MASK]
[SEP]
W5

Radford et al. (2019)



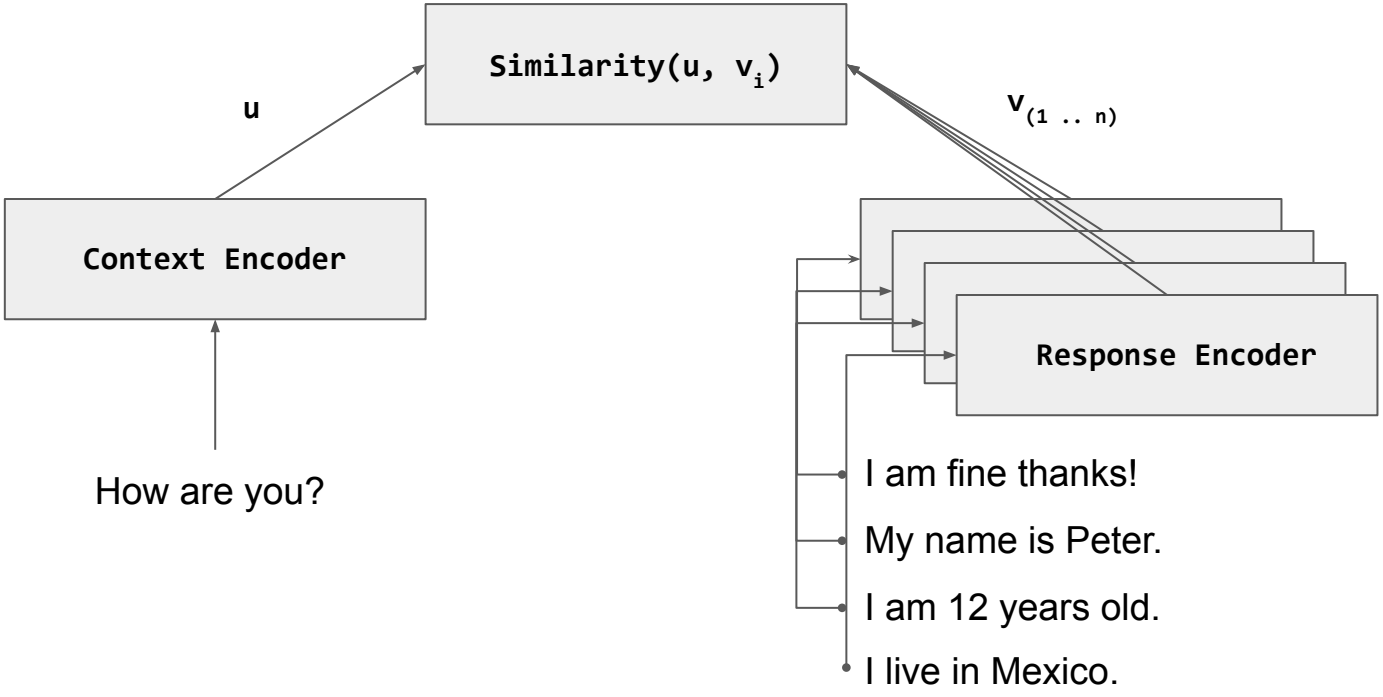
BERT vs GPT-2


BERT
T1
T2
...
TN
Trm
Trm
...
Trm
Trm
Trm
...
Trm
E1
E2
...
EN
GPT-2
T1
T2
...
TN
Trm
Trm
...
Trm
Trm
Trm
...
Trm
E1
E2
...
EN

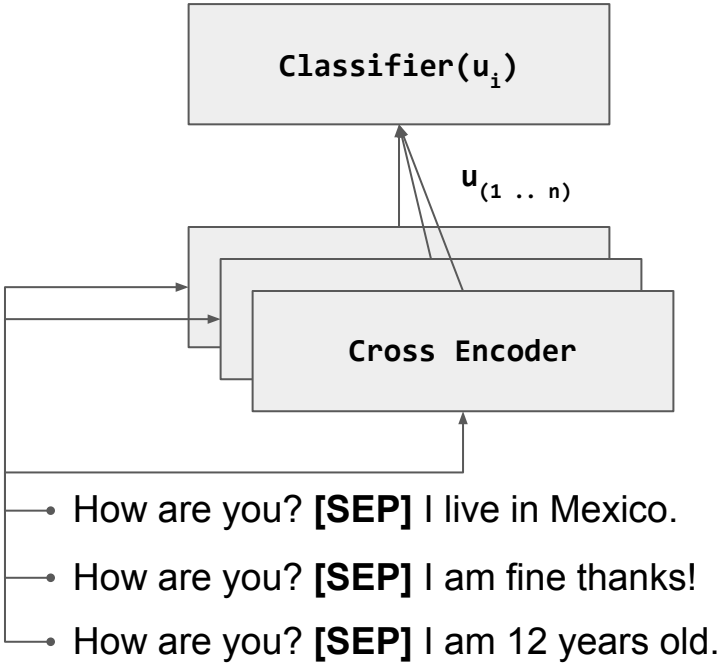
Retrieval-based

- **Speaker-Aware BERT** - Gu et al. (2020)
- **Poly-encoders -** Humeau et al. (2019)

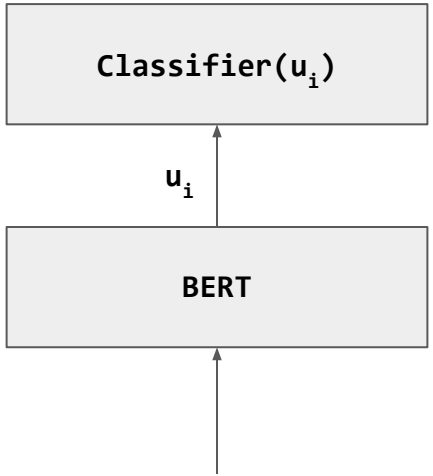
Bi-Encoder


Similarity(u, v_i)
u
$v_{(1 .. n)}$
Context Encoder
Response Encoder
How are you?
I am fine thanks!
My name is Peter.
I am 12 years old.
I live in Mexico.

Cross-Encoder

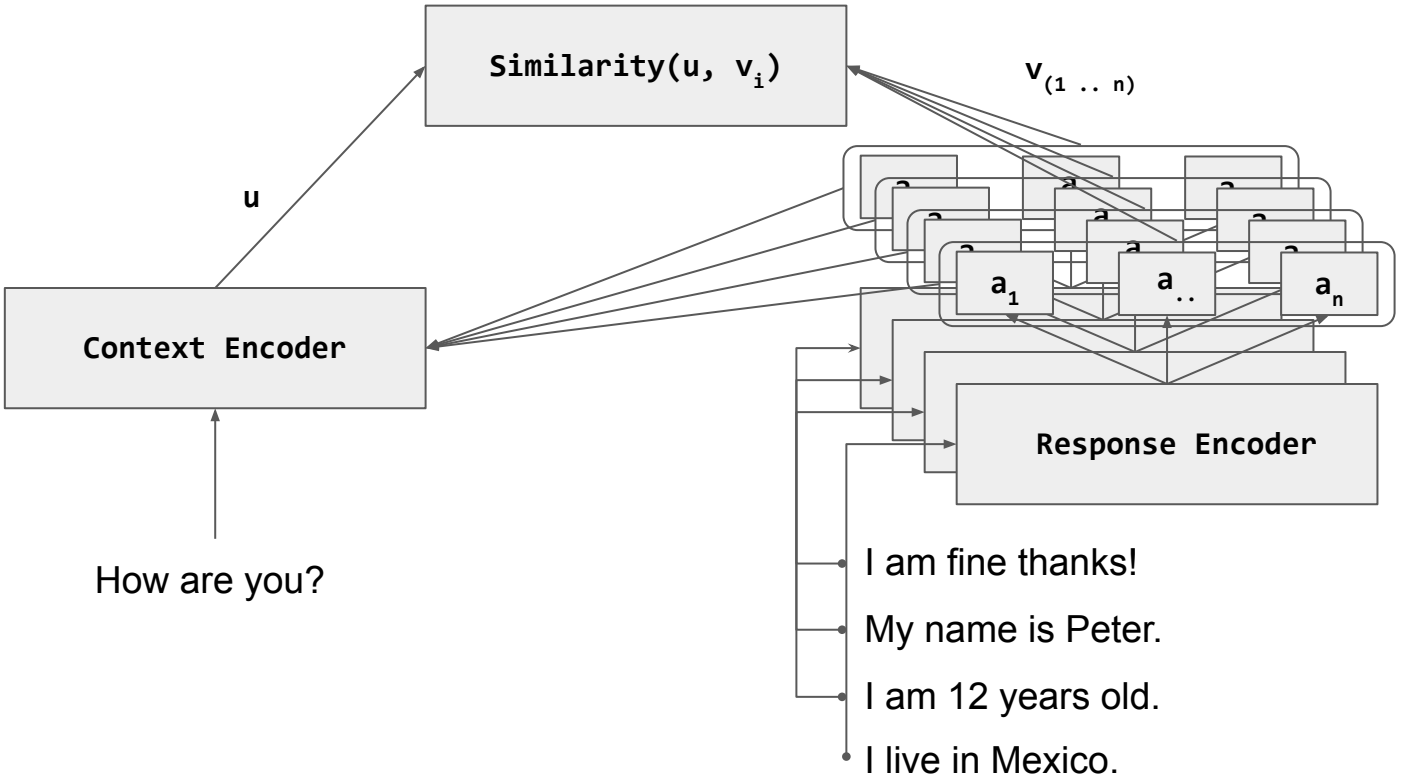


Speaker Aware BERT



[SP1] Hi **[SP2]** Hello **[SP1]** How are you? **[SEP] [SP2]** I am fine thanks!

Poly-Encoder



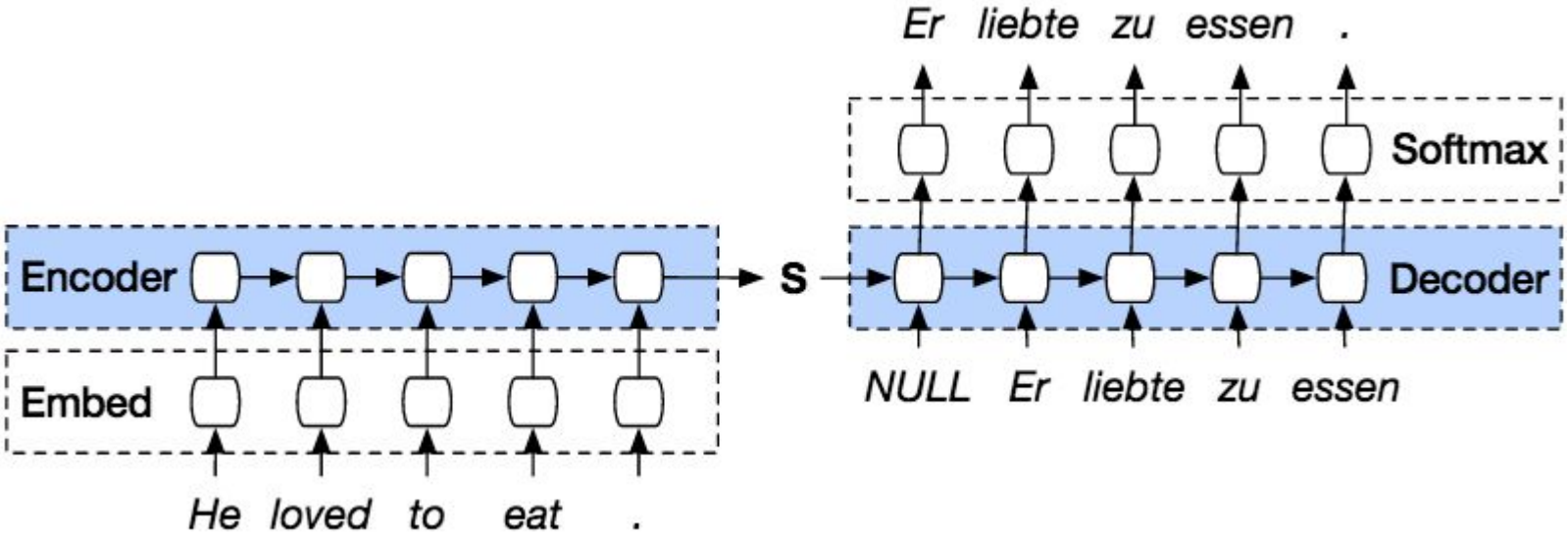
Neural Network based - Generative

- **A Neural Conversational Model -** Vinyals et al. (2015)
- **DialoGPT** - Zhang et al. (2019)
- **Towards a Human-like Open-Domain Chatbot (Meena) -** Adiwardana et al. (2020)
- **Recipes for building an open-domain chatbot** - Roller et al. (2020)


Can you please come here ?
History
Word being predicted

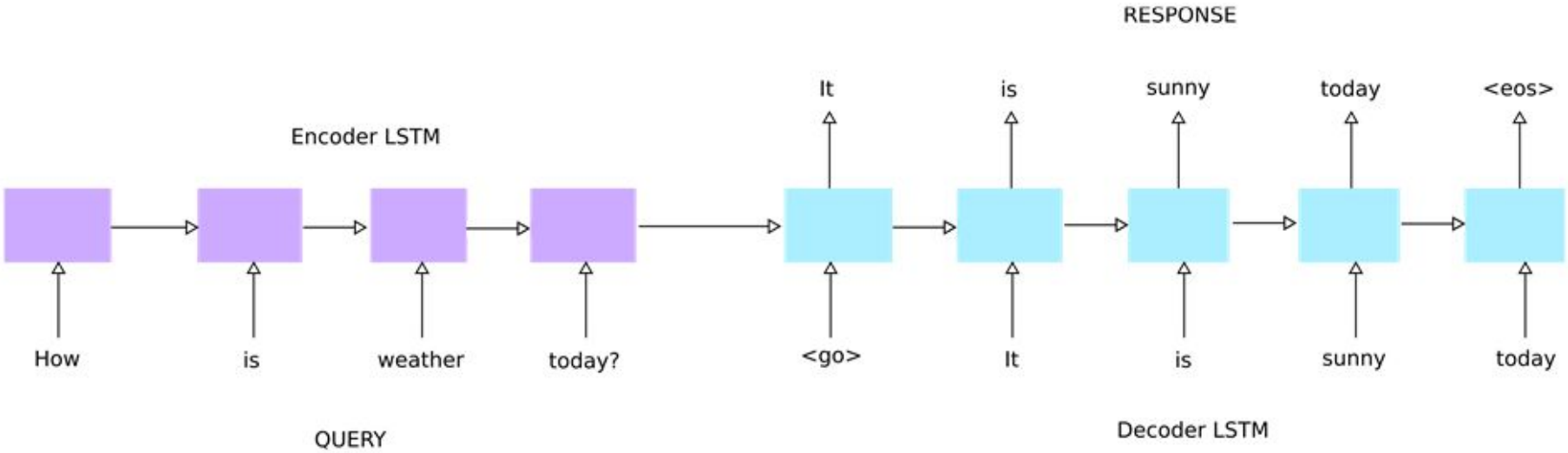
Seq2Seq

Sutskever et al. (2015)



Neural Conversational Model

Vinyals et al. (2015)



DialoGPT
Zhang et al. (2019)

- **Reddit conversations**
  - 146.8M training instances
  - 3.87B source tokens
  - 2.14B target tokens
- **Different Models**
  - 117M
  - 345M + Reverse
  - 762M

<User> Does money buy happiness?

<Bot> Depends how much money you spend on it .

<User> What is the best way to buy happiness ?

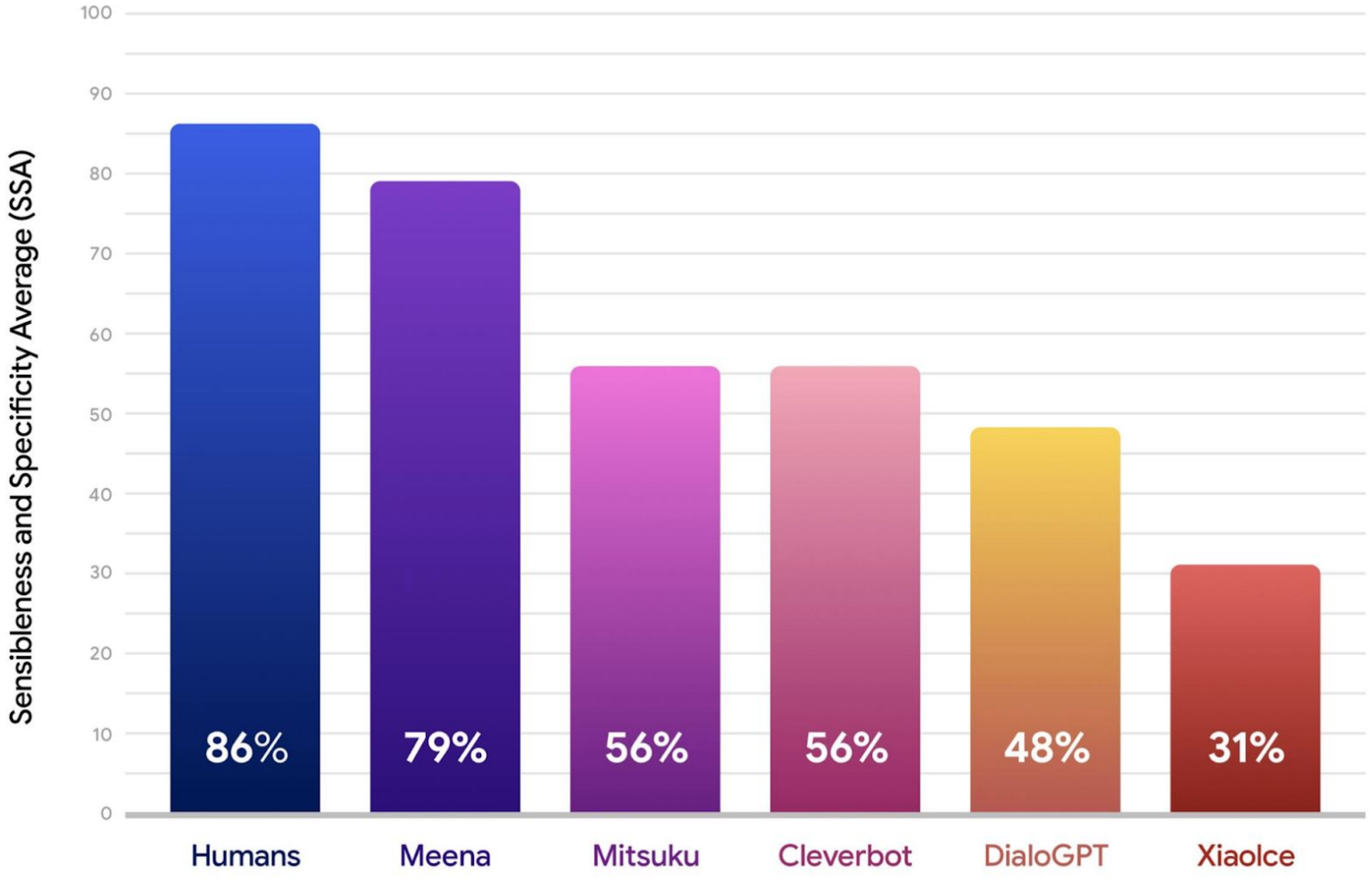
<Bot> You just have to be a millionaire by your early 20s, then you can be happy .

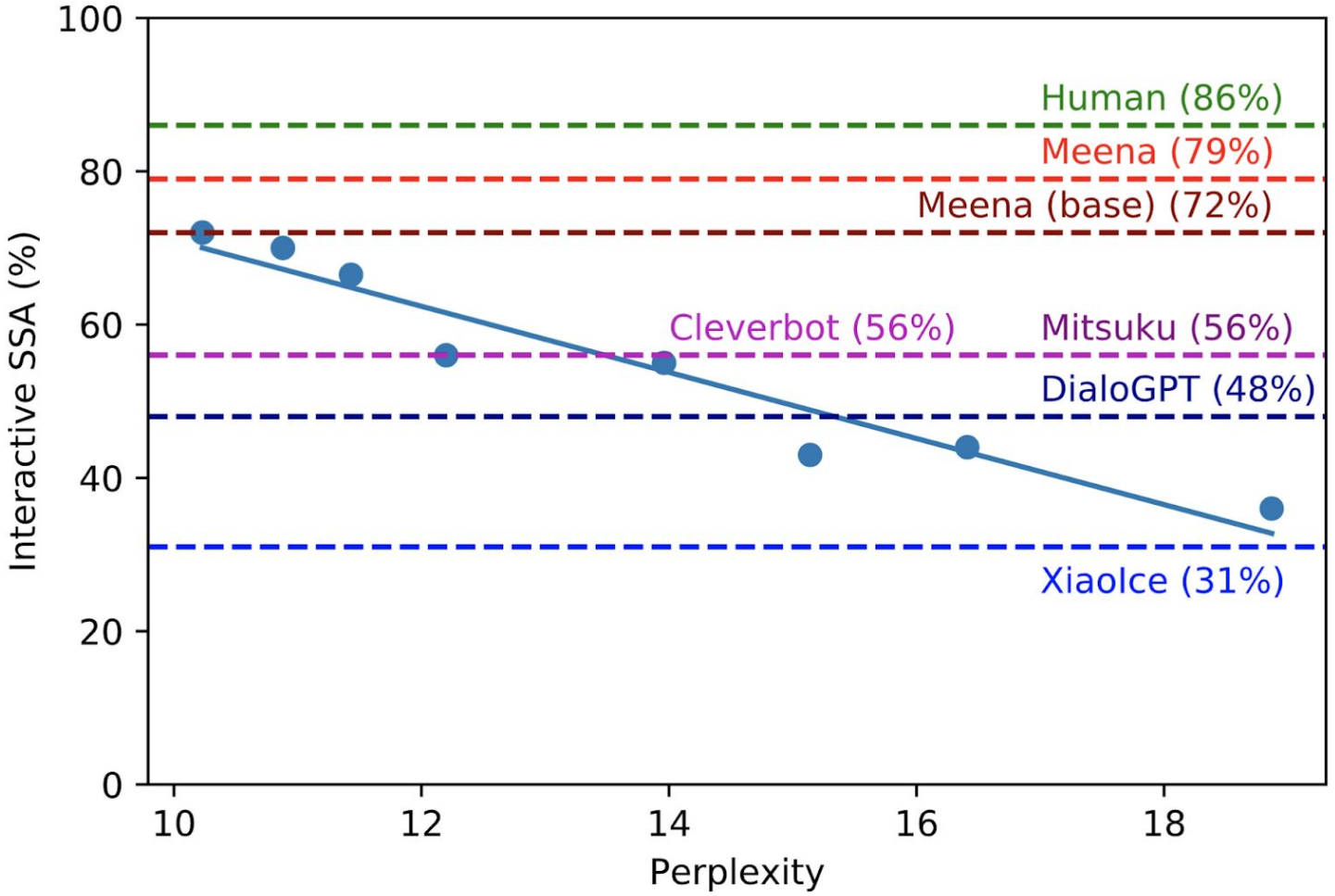
<User> This is so difficult !

<Bot> You have no idea how hard it is to be a millionaire and happy . There is a reason the rich have a lot of money

Meena

Adiwardana et al. (2020)


Sensibleness and Specificity Average (SSA)
100
90
80
70
60
50
40
30
20
10
0
86%
79%
56%
56%
48%
31%
Humans
Meena
Mitsuku
Cleverbot
DialoGPT
Xiaolce


100
80
60
40
20
0
Interactive SSA (%)
10
12
14
16
18
Perplexity
Human (86%)
Meena (79%)
Meena (base) (72%)
Cleverbot (56%)
Mitsuku (56%)
DialoGPT (48%)
Xiaolce (31%)

Blender
Roller et al. (2020)

- **Blending skills**
  - Engaging use of personality (PersonaChat)
  - Engaging use of knowledge (Wizard of Wikipedia)
  - Display of empathy (Empathetic Dialogues)
  - Ability to blend all three seamlessly (BST)
- **Retrieve and refine**

Generative Decoding

- **Deterministic**
  - Greedy Decoding
  - Beam Search
- **Sampling**
  - Top-k
  - Nucleus
- **Similarity search (Continuous output)**

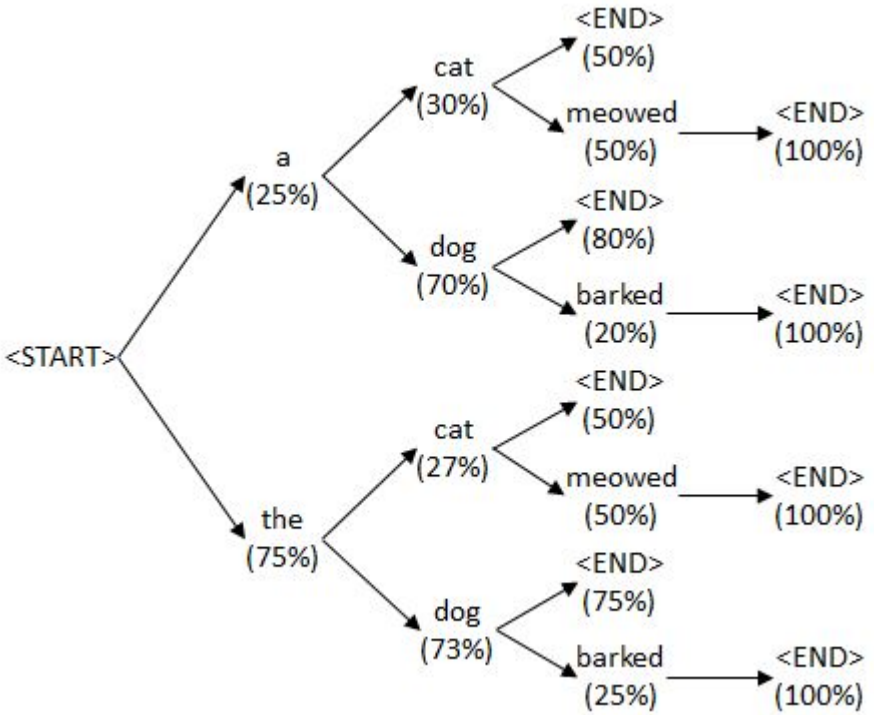
Beam search

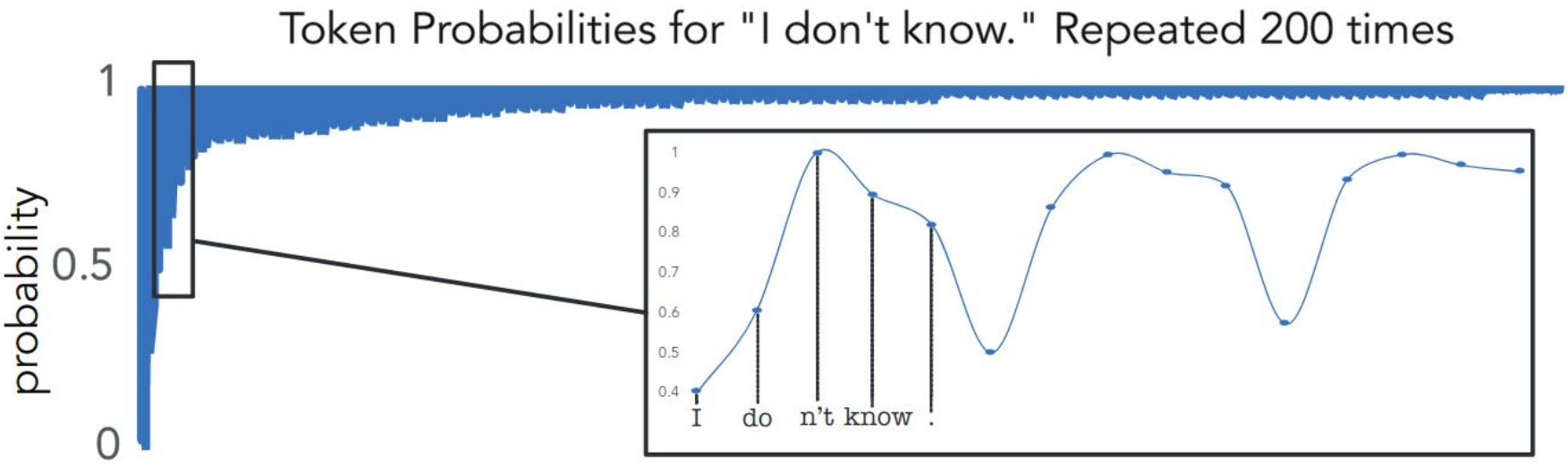
Hyperparameter:

K (beam width)

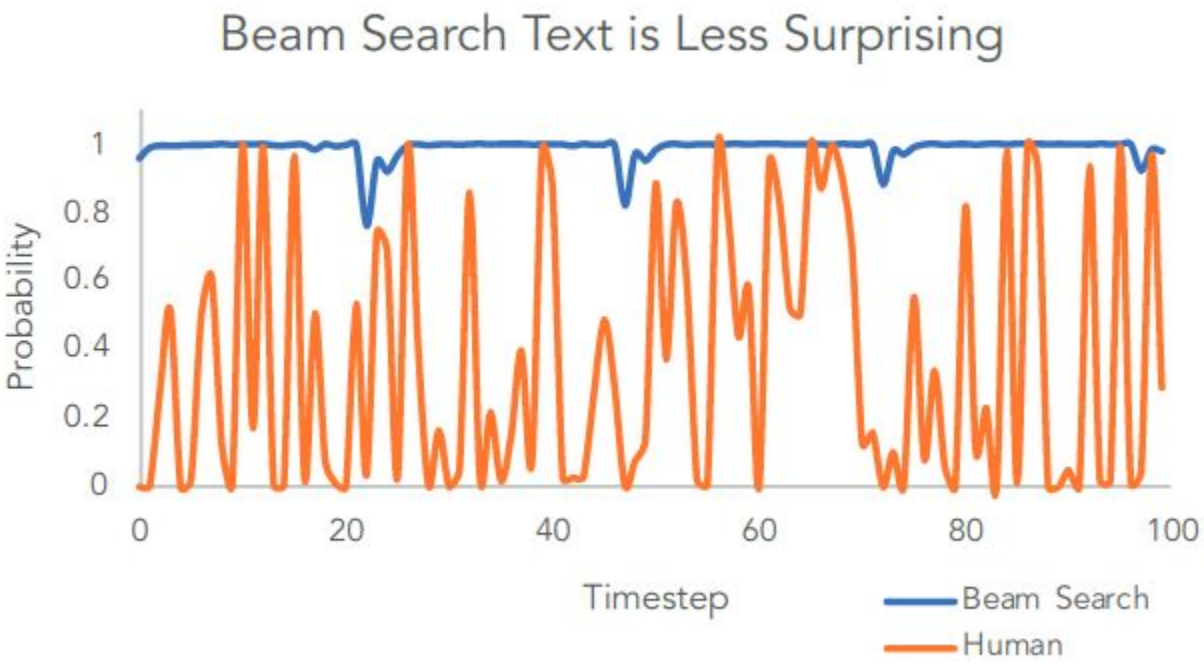
Greedy decoding

K = 1




Token Probabilities for "I don't know." Repeated 200 times
probability
1
0.5
0
1
0.9
0.8
0.7
0.6
0.5
0.4
I
do
n't
know
.

Holtzman et al. (2020)



Enhancing Beam search

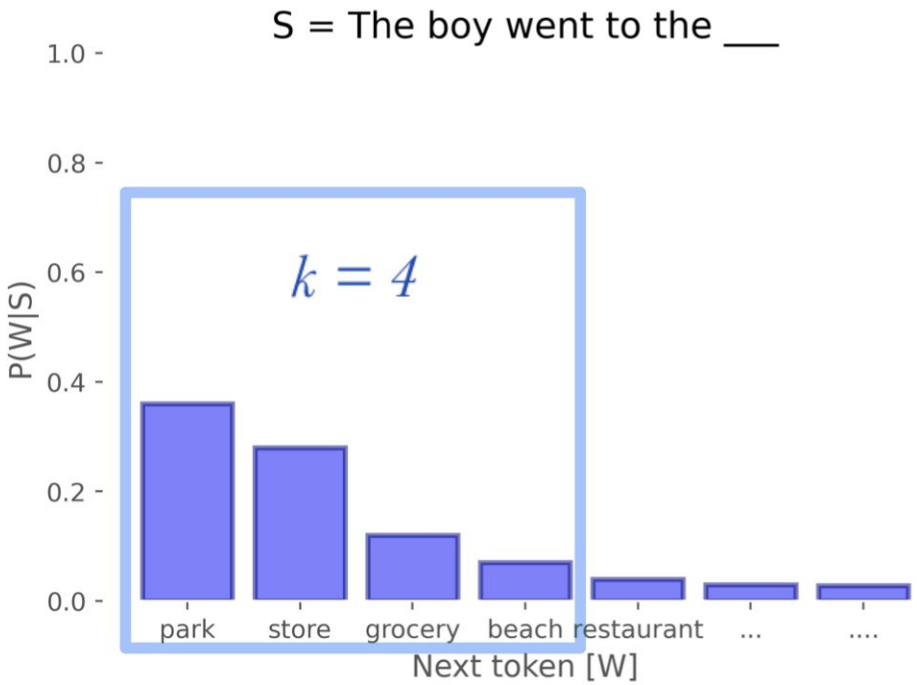
- **Response length**
  - Hard coded minimum
  - Response length prediction
    (Blender) 4 class classifier (<10, <20, <30, >30)

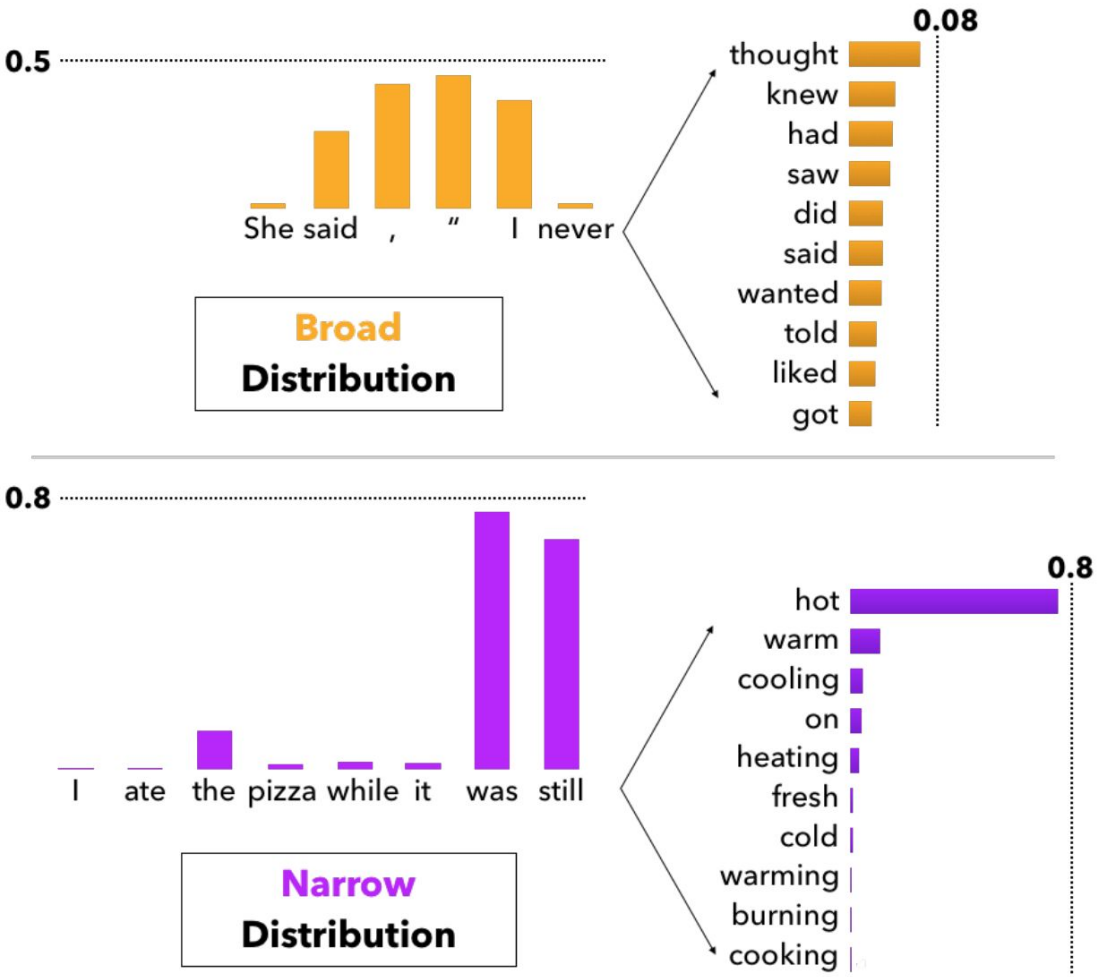
- **Subsequence blocking**
  Not allowing existing n-grams

Top-k decoding

Hyperparameter:

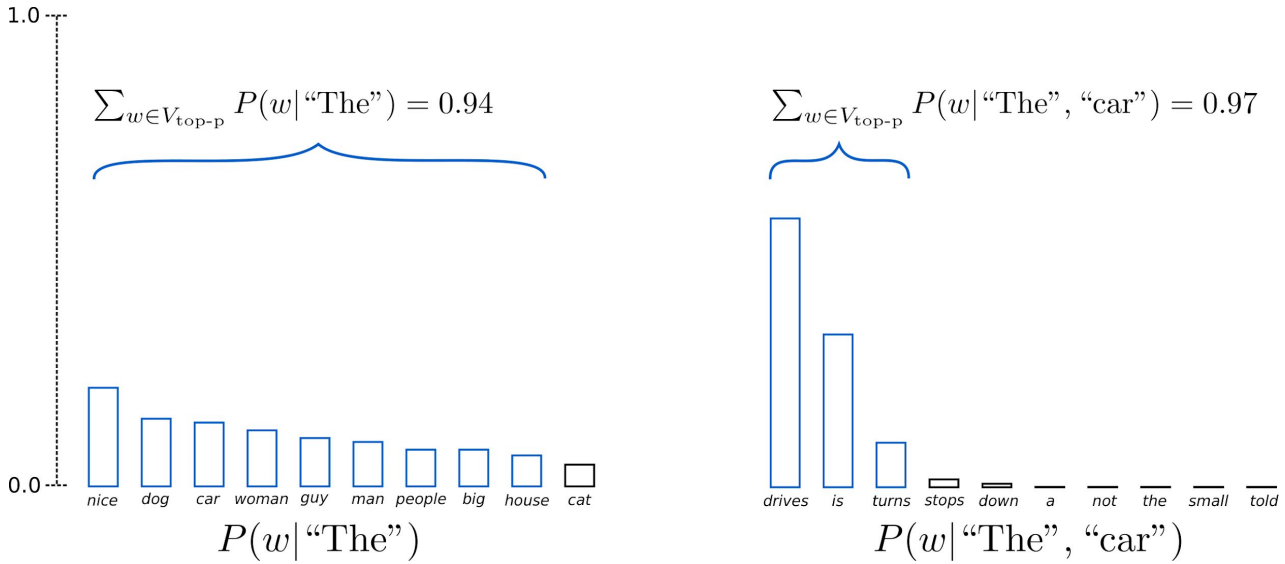
K (sample size)

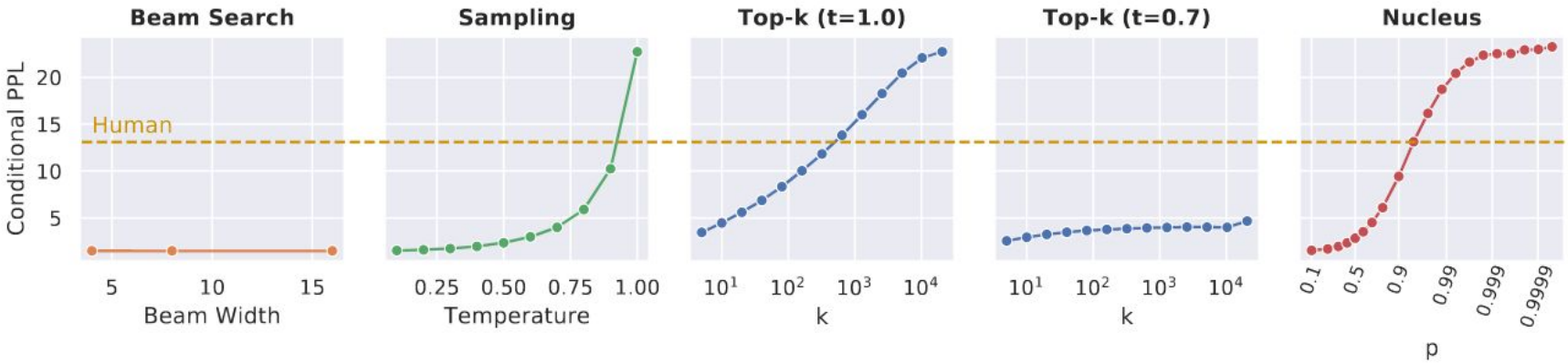



0.5
She said , " I never
Broad
Distribution
0.08
thought
knew
had
saw
did
said
wanted
told
liked
got
0.8
I ate the pizza while it was still
Narrow
Distribution
0.8
hot
warm
cooling
on
heating
fresh
cold
warming
burning
cooking

Nucleus decoding

Hyperparameter:
p (probability mass)




Beam Search
Sampling
Top-k (t=1.0)
Top-k (t=0.7)
Nucleus
Conditional PPL
20
15
10
5
Human
5
10
15
Beam Width
0.25
0.50
0.75
1.00
Temperature
10^1
10^2
10^3
10^4
k
10^1
10^2
10^3
10^4
k
0.1
0.5
0.9
0.99
0.999
0.9999
p

MMI Decoding

Forward + Backward language model

1. The forward model generates the output and computes $P_{forward}$
2. The backward model computes $P_{backward}$

Output is then scored by $P_{backward} + P_{forward}$

Similarity Search (Continuous output)

Method:

- Model output is not projected onto the vocabulary
- Objective is to minimize $\text{Sim}(v_{output}, v_{target})$
- V_{target} is drawn from pre-trained embeddings

Advantages:

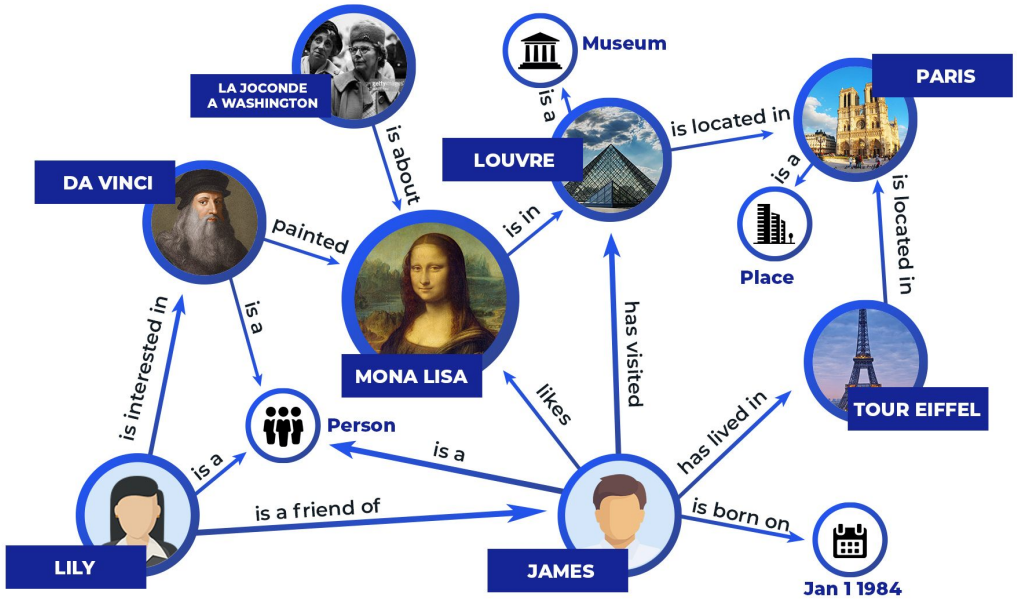
- Faster training
- Interpretable output

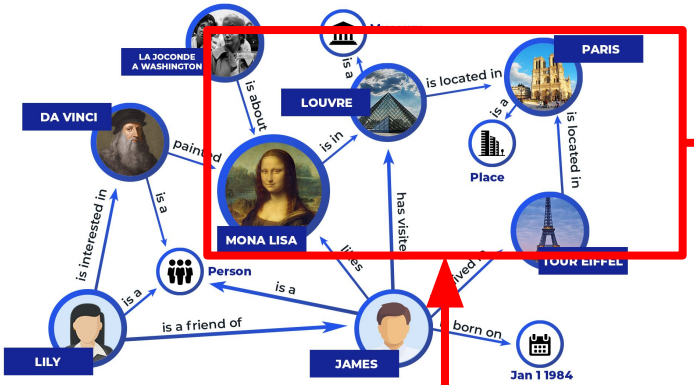
Issue with end-to-end solutions

- **Bad interpretability**
  - Knowledge is stored as parameter values
  - We can’t inspect the cause of a bad answer

- **Information hallucination**
  Example: "Cats have three legs"

Knowledge graphs





Graph Decoder

(Mona Lisa, is_in, Louvre)

(Louvre, is_located_in, Paris)

Response Generator

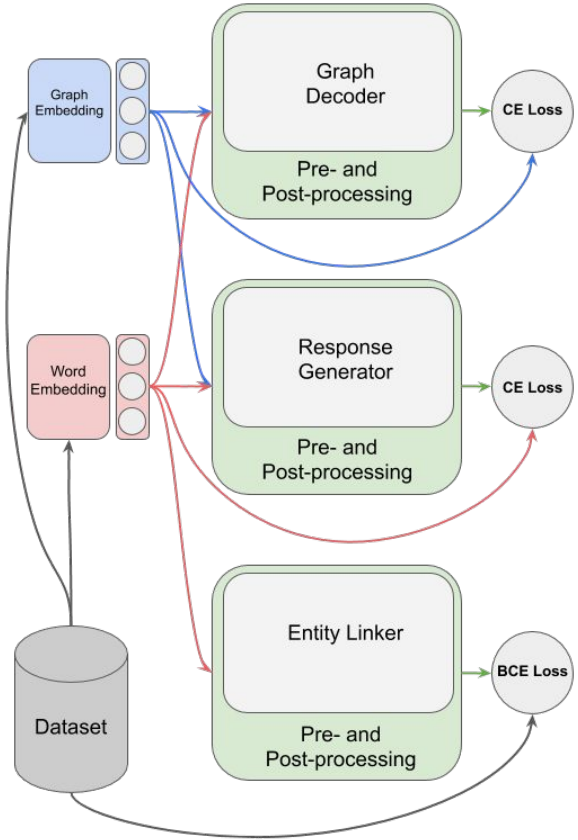
Entity Linker

Where is the Mona Lisa located?

It is located in Paris.

Architecture

- **Word Embedding**
  - Fasttext
- **Graph Embedding**
  - TransE
- **Graph Decoder**
  - LSTM
- **Response Generator**
  - Transformer
  - Continuous output
- **Entity Linker**
  - BERT



Dataset

OpenDialKG

Moon et al. (2019)

91031	messages
100813	entities
1358	relations
1190658	triplets

```
[
    {
     "Message": "Do you like Iron Man",
     "Metadata": [
                [
                    "Iron Man",
                    "starred_actors",
                    "Robert Downey Jr."
                ]
            ]
     }
]
```

Word embeddings
(Fasttext)

Graph embeddings

- **DeepWalk** - Perozzi et al. (2014)
- **TransE** - Bordes et al. (2013)
- **TransR** - Lin et al. (2015)

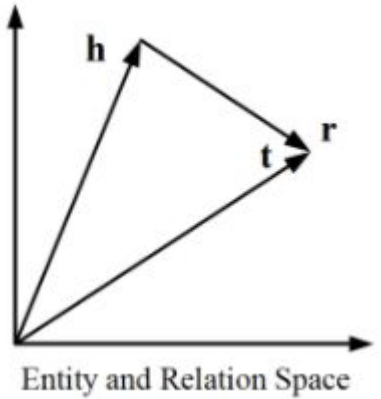
TransE

Bordes et al. (2013)

Knowledge triplet:

Mona Lisa (head), is_in (relation), Louvre (tail)

$$v_{head} + v_{relation} = v_{tail}$$

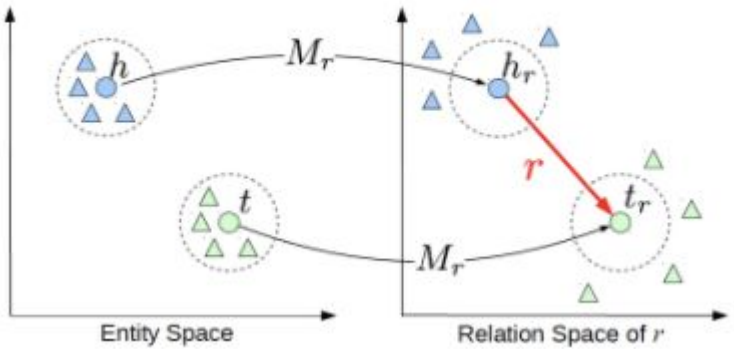


TransR

Lin et al. (2014)

Each relation has its own transformation matrix: M_r

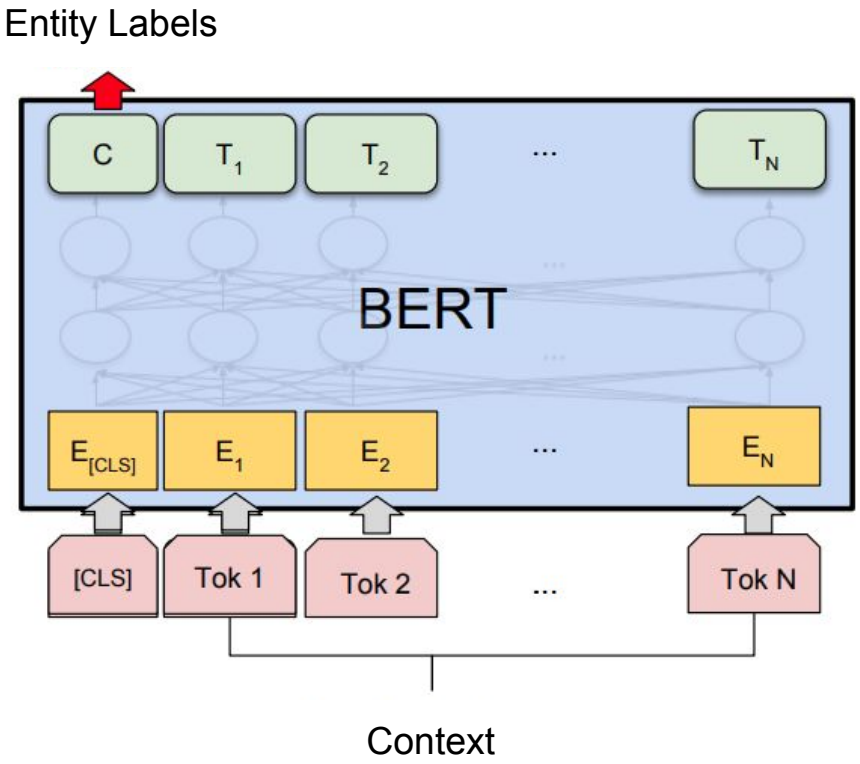
$$M_r v_h + v_r = M_r v_t$$



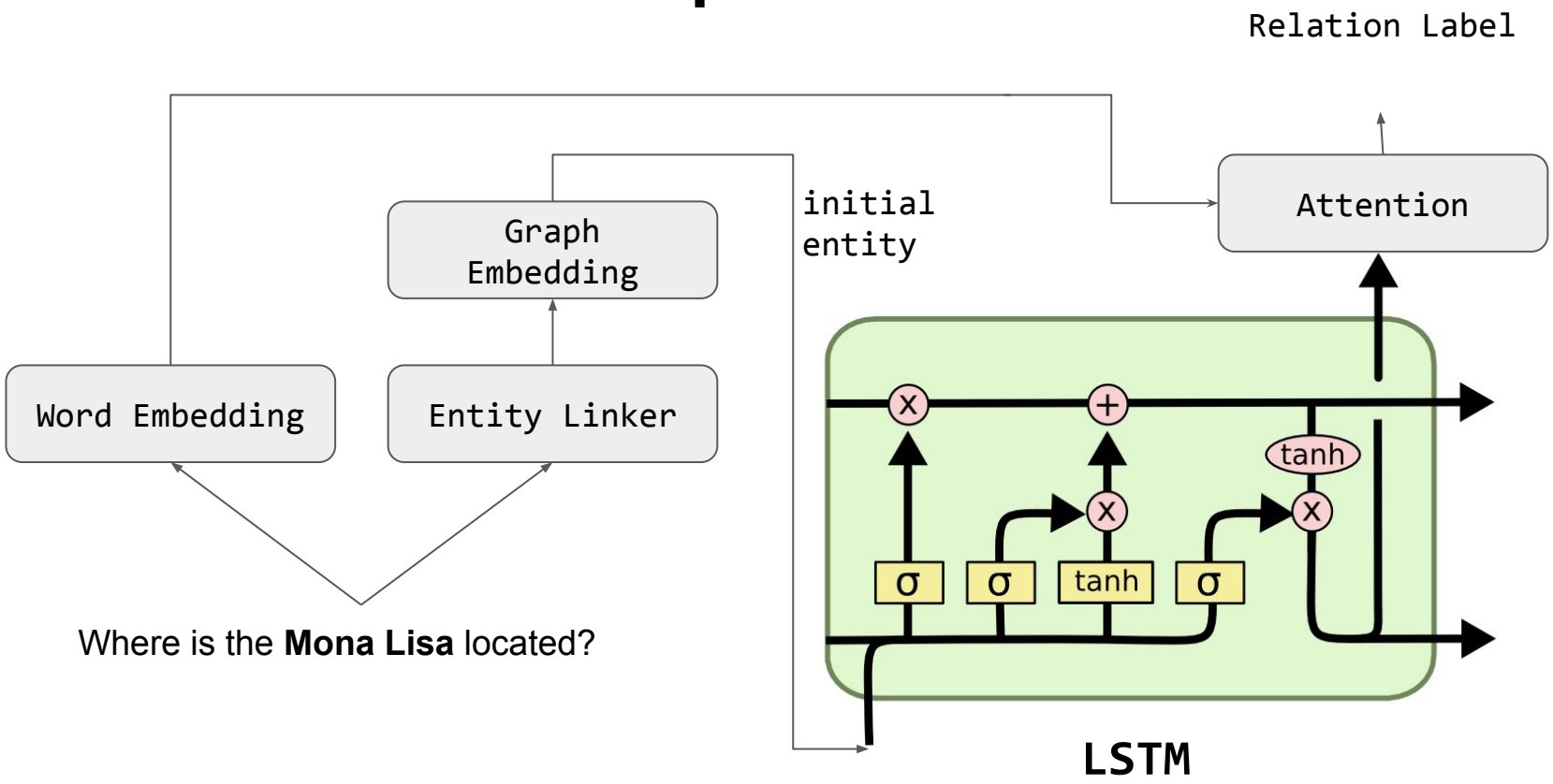
Entity Linker

Limited information in OpenDialKG

Labels are learned with BCE objective



Graph Decoder

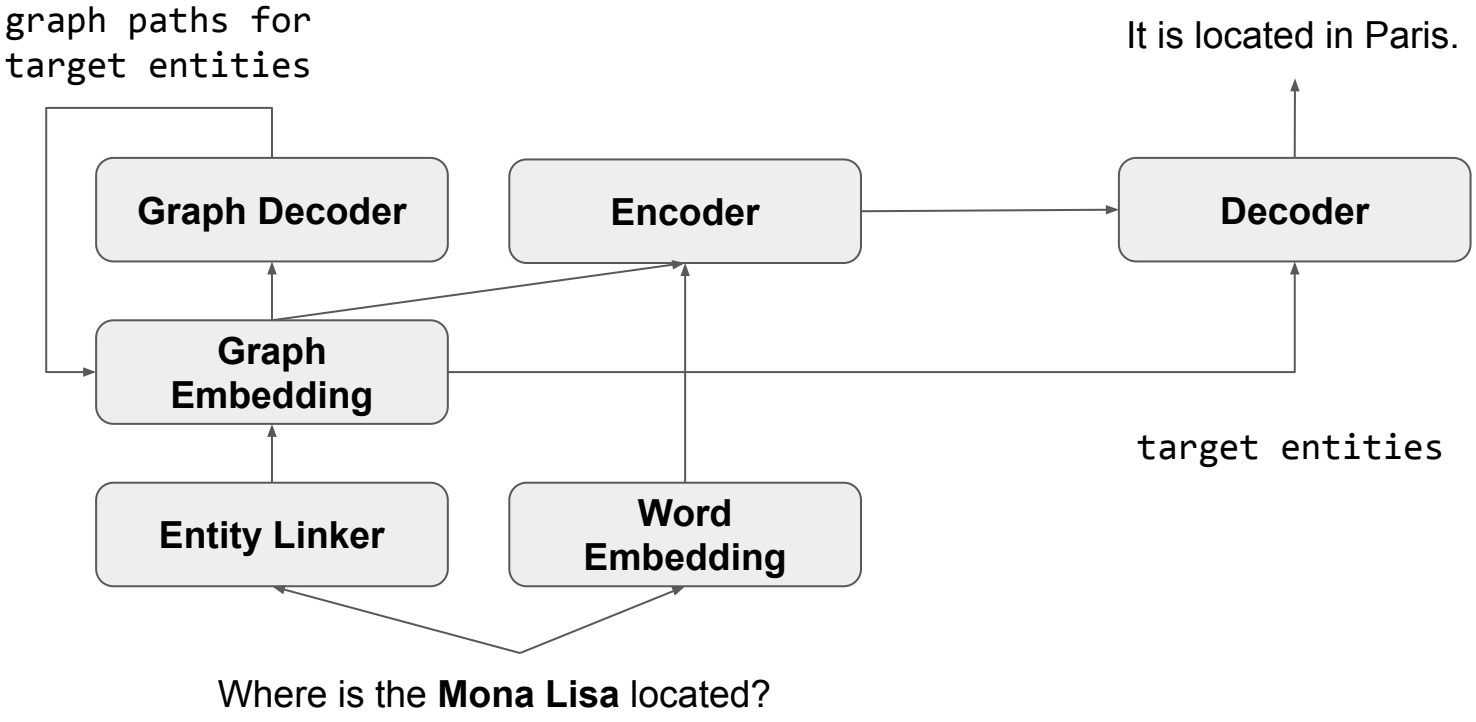


Response Generator

- **Transformer encoder-decoder architecture**
  - additional cross-attention layer for the entity node representations

- **Continuous outputs with dot-product loss + negative sampling**
  - (Future work) Decoding entity nodes directly

Response Generator



Baselines

- GPT-2 (345M)
- XLNet (110M)
- Vanilla Transformer
- Continuous Transformer

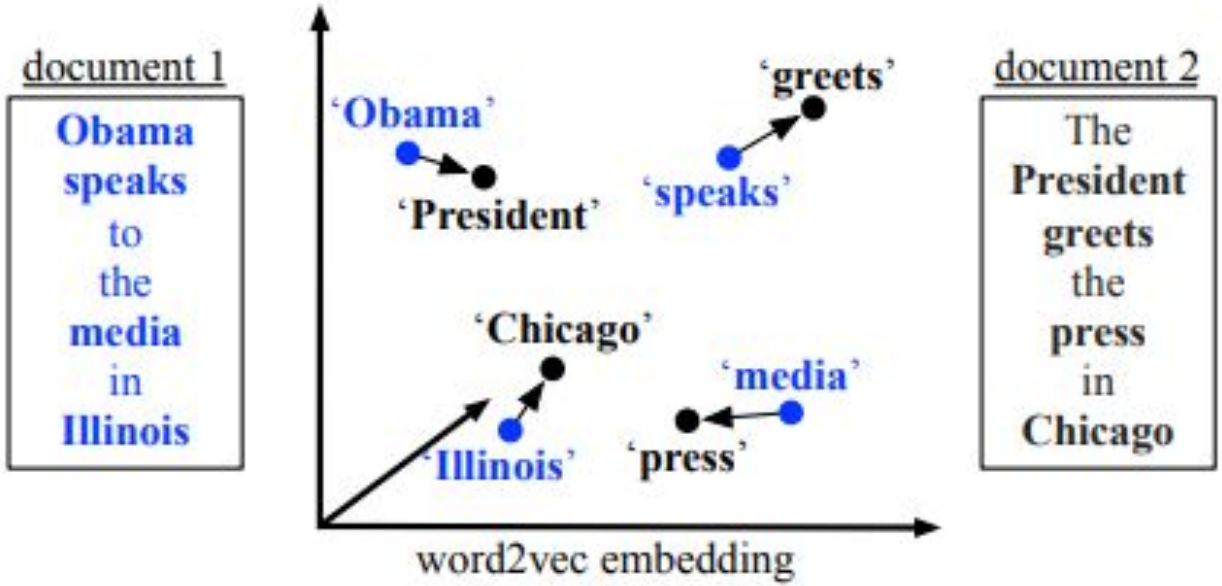
Automatic Metrics

- **Length**
  Generated data: Average length of model outputs
  Test data: Delta length compared to gold responses

- **BLEU**
  Modified precision score of n-gram overlaps
  Common metric in NMT
  Low correlation with human judgement

- **WMD**

WMD


document 1
Obama speaks to the media in Illinois
'Obama'
'President'
'greets'
'speaks'
'Chicago'
'Illinois'
'media'
'press'
word2vec embedding
document 2
The President greets the press in Chicago

Results

Model	Decoding	Length		WMD		BLEU	
		Mean	Std	Mean	Std	Mean	Std
GPT-2 (small)	Greedy decoding	10.2	3.5	2.0	0.4	0.1	0.0
	Beam Search $k = 3$	12.1	3.6	1.7	0.5	0.1	0.0
	Beam Search $k = 5$	12.0	3.5	**1.6**	**0.4**	0.1	0.0
	Top-k Sampling $k = 100$	14.2	3.8	1.6	0.6	0.1	0.0
	Nucleus decoding $p = 0.1$	12.1	3.9	1.7	0.6	**0.2**	**0.1**
XLNet (base)	Greedy decoding	9.4	3.4	2.2	0.3	0.1	0.0
	Beam Search $k = 3$	11.2	3.2	1.9	0.4	0.1	0.0
	Beam Search $k = 5$	12.9	3.3	1.8	0.3	**0.2**	**0.1**
	Top-k Sampling $k = 100$	**14.3**	**4.3**	1.9	0.6	0.1	0.0
	Nucleus decoding $p = 0.1$	14.1	4.5	1.8	0.5	0.1	0.0
Transformer	Greedy decoding	7.1	2.2	2.3	0.2	0.0	0.0
	Beam Search $k = 3$	8.5	2.0	2.3	0.4	0.0	0.0
	Beam Search $k = 5$	9.4	2.8	2.2	0.4	0.0	0.0
	Top-k Sampling $k = 100$	9.4	4.1	2.3	0.4	0.1	0.0
	Nucleus decoding $p = 0.1$	9.6	4.7	2.3	0.3	0.1	0.0
Continuous Transformer	Similarity Search	9.8	3.3	2.1	0.3	0.0	0.0
KG-Enhanced Transformer	Similarity Search	12.8	3.4	**1.6**	**0.4**	**0.2**	**0.1**

Generated dialogues

Input	Method	Model Output	Target
I like the movie 'The Martian', could you recommend a movie similar to it?	GPT-2 (small) Beam Search $k = 5$	You might enjoy Cloverfield with Drew Barrymore or The Martian. They are both Sci-Fi films.	Do you like The Martian because it stars Matt Damon?
	GPT-2 (small) Nucleus decoding $p = 0.1$	Sure, do you prefer Sci-Fi or Adventure movies?	
	XLNet (base) Beam Search $k = 5$	It is starring Michael Fassbender, who also starred in X-Men: First Class and X-Man 2. Do you like movies that he stars in?	
	XLNet (base) Nucleus decoding $p = 0.1$	I do! I love Sci Fi movies as well. Have you seen either of those?	
	Transformer Nucleus decoding $p = 0.1$	This is a good one	
	Continuous Transformer	Have you seen The Running Man?	
	KG-Enhanced Transformer	It is written by James Patterson, starring Matt Damon and also stars Tracie Thoms. Do you like them?	

KG-enhanced transformer

- **[Human]** Hi
- **[Model]** Hello, have you heard of the movie The Hobbit: Desolation of Smaug?
- **[Human]** No, can you tell me more about it?
- **[Model]** It is a fantasy adventure film starring James Nesbitt. Do you like this actor or would you like to know other movies he has starred in?
- **[Human]** List some of his movies please.
- **[Model]** Sure! He also starred in The Way and The Amateurs. Have you seen either of those?
- **[Human]** What is "The Way" about?
- **[Model]** It's a teen film, but it was directed by Marc Forster who also wrote the movie The Valley of Elah which stars Tom Hanks.
- **[Human]** Oh I see. Can you recommend me a Sci-Fi movie?
- **[Model]** Sure! White Televison and Undertow are good ones as well. Have you seen either of those?

GPT-2

- [**Human**] hi
- [**Model**] greetings, i'm gerald.
- [**Human**] how are you ?
- [**Model**] good, thanks. getting ready for another hike, i love hiking!
- [**Human**] oh realy? where are you going?
- [**Model**] pikes peak. i am curently exploring famous mountains when i have time.
- [**Human**] that is cool are you going alone?
- [**Model**] my brother is in the us, we go hiking with him every month.
- [**Human**] how old are you by the way?
- [**Model**] a little under 46 years old
- [**Human**] are you in a relationship?
- [**Model**] yes i am married, 3 boys, i am pregnant again. i am too busy to focus

Conclusion

- Lack of automatic metrics for checking factual accuracy
- Lack of datasets for this purpose
- Unsatisfactory results from the proposed architecture

+ I paved the way for future work
+ I created a framework for training chatbots from pre-trained transformers:

  https://github.com/bme-chatbots/dialogue-generation

Future work

- Implementing pre-trained transformer as the response generator
- Improving the OpenDialKG dataset
- Finding better automatic metric

References

- arxiv.org/abs/1706.03762 (Transformer)
- jalammar.github.io/illustrated-gpt2/ (GPT-2)
- yashuseth.blog/2019/10/08/ (Knowledge graph)