ILPS – SEA Meetup 2019/01/18

Transfer Learning for Natural Language Generation – The Case of Open-Domain Dialog

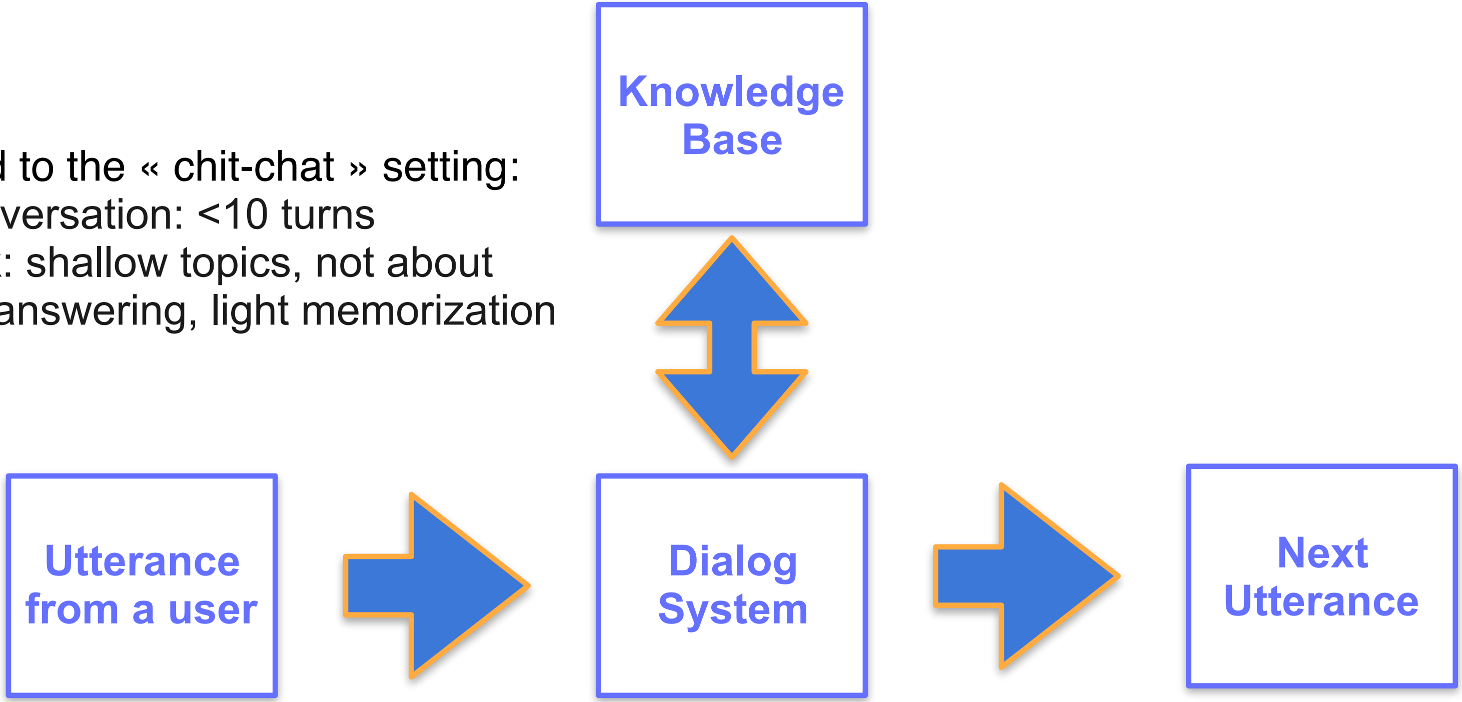
thomas@huggingface.co

Open-Domain Conversational Agents

A conversational agent which can talk about any topic

Often restricted to the « chit-chat » setting:

- Short conversation: <10 turns
- Small talk: shallow topics, not about question-answering, light memorization



Issues

Two main classes of models:

- **Retrieval** models: ⊕ Grammaticality/Fluency ⊖:
  1. Adaptability,
  2. Diversity,
  3. Consistency
- **Generative** models: ⊕ Diversity/Adaptability ⊖:
  1. Lack of a consistent personality
  2. Lack long-term memory (trained to use only recent history)
  3. Tend to produce non-specific answers: *“I don’t know”*

The Conversational Intelligence Challenge 2 (ConvAI2)

—

NeurIPS 2018 - Competition Track

Condition Dialog on a Predefined Personality

Example of training dataset – Evaluation dataset:
PERSONA-CHAT (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.

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.

- Amazon Mechanical Turkers were:
  - **paired** by two,
  - each given a **personality** comprising 4-5 simple sentences, and
  - asked to **talk** together in order to get to know each other.
- Resulted in a dataset of
  - **10,981 dialogs** comprising
  - **164,356 utterances** and about **1-2M words**
  - Average number of turns: **14**

Metrics

Automatic Metrics

- **PPL** (perplexity) *How well the model can predict the successive words in a gold message (written by humans).*
  - **lower** is better – Scale: **Infinity – 0**
- **Hits@1** *Number of time the model select the gold next message between 20 possible message (the other 19 are random)*
  - **higher** is better – Scale: **0 –100**
- **F1** *How many content words (nouns/verbs) does a message generated by your model share with a gold message*.
  - **higher** is better – Scale: **0 –100**

Human Evaluation

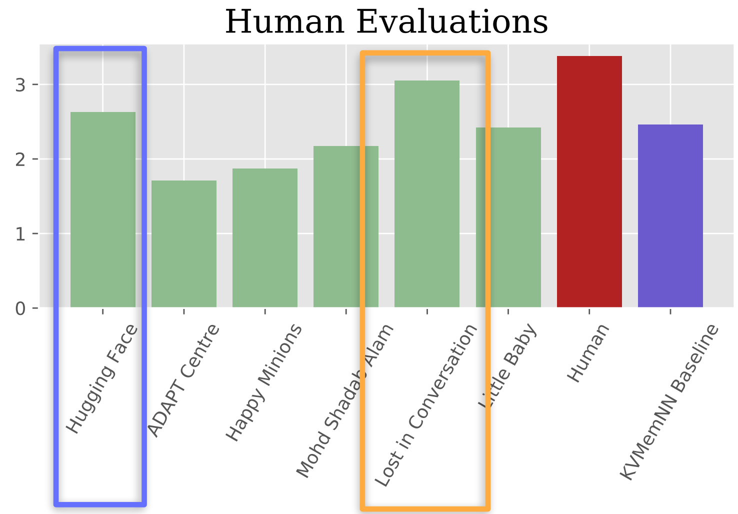
- 100 evaluations per model
- Turkers & model each assigned a persona and chat for 4-6 dialog turns each
- After the chat, the worker is asked:
  - *How much did you enjoy talking to this user?*
  - *Which character do you think the other user was given for this conversation?*

Final Leaderboards of the Competition

Automatic Metrics

Rank	Creator	PPL	Hits@1	F1
1 🍐	🤗 (Hugging Face)	16.28🍎	80.7🍎	19.5🍎
2 🍐	ADAPT Centre	31.4	-	18.39
3 🍐	Happy Minions	29.01		16.01
4 🍐	High Five	-	65.9	-
5 🍐	Mohd Shadab Alam	29.94	13.8	16.91
6 🍐	Lost in Conversation	-	17.1	17.77
7 🍐	Little Baby(AI小奶娃)	-	64.8	-
8	Sweet Fish	-	45.7	-
9	1st-contact	31.98	13.2	16.42
10	NEUROBOTICS	35.47	-	16.68
11	Cats'team	-	35.9	-
12	Sonic	33.46	-	16.67
13	Pinta	32.49	-	16.39
14	Khai Mai Alt	-	34.6	13.03
15	loopAI	-	25.6	-
16	Salty Fish	34.32	-	-
17	Team Pat	-	-	16.11
18	Tensorborne	38.24	12.0	15.94
19	Team Dialog 6	40.35	10.9	7.27
20	Roboy	-	-	15.83
21	IamNotAdele	66.47	-	13.09

Human Evaluation



Diving in the Wining Approaches

Two Approaches to Open-Domain Dialog

Similarities and Differences

- **Many common points:**
  - Both build on top of Generative Transformer models
  - Both based on Transfer Learning Approaches
  - Same Pre-training Phase
- **But also some differences:**
  - Different Architectural Modifications for the Adaptation
  - Different Objectives for the Adaptation Phase
  - Different Decoders

Common Points:
A Generative Transformer 🛩️

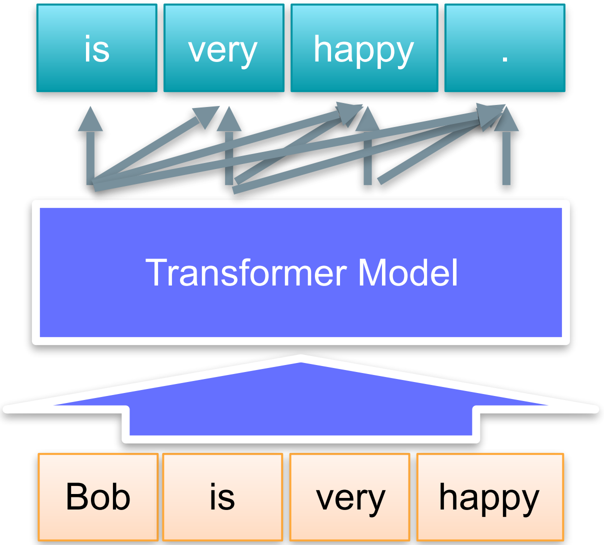
A Transformer Generative Model

Our Dialog System has two elements:

- A **generative model** which generate the words one by one given the context,
- A **decoder** which controls the generative model.

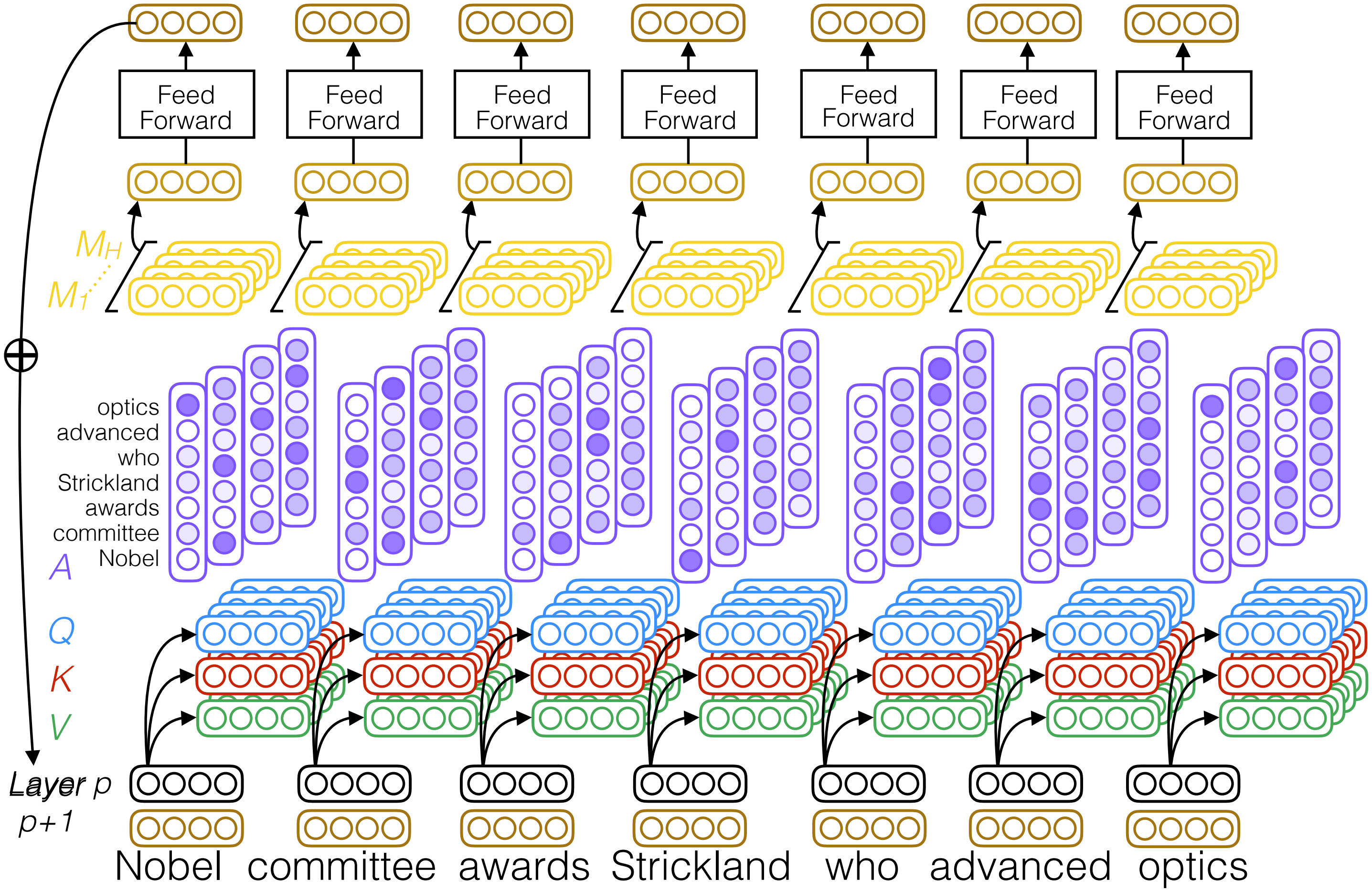
In both approaches, the **generative model** is based on the OpenAI GPT[1]:

- BPE vocabulary with 40000 tokens
- learned position embeddings with 512 positions
- 12 layers
- 12 attention head with 768 dimensional states
- position-wise feed-forward networks with 3072 dimensional inner states



1.Radford, A., Narasimhan, K., Salimans, T., Sutskever, I. (2018). Improving language understanding by generative pre-training.

Transformer Model



[Slides by Emma Strubbell – EMNLP 2018]

Language Modeling Transformer

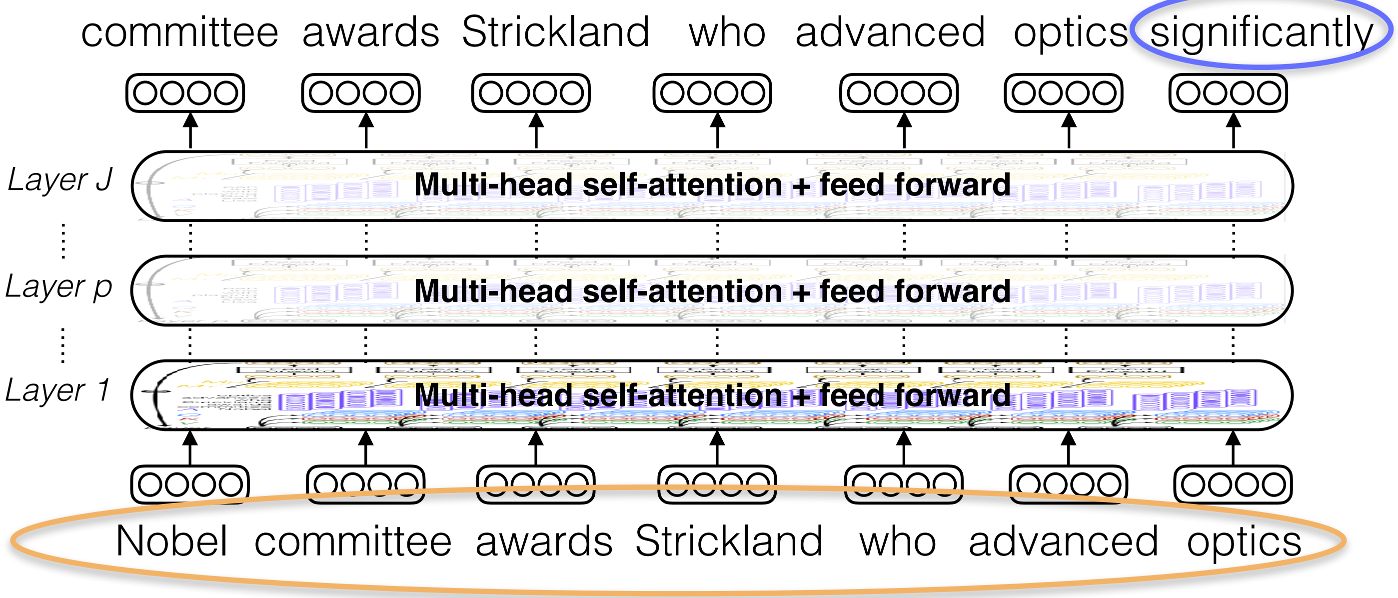
[Adapted from slides by Emma Strubbell – EMNLP 2018]

The Transformer is trained to predict the next words given the history.

We use a mask so that each word is only « mixed" with the previous words (and not the following)

This is called Language Modeling
(we learn a model of the probability of language)

$$p(w_1, \ldots, w_n) = \prod_{i=1}^{n} p(w_i | w_1, \ldots, w_{i-1})$$



Common Points: Transfer Learning 🦄

Limitations of the dataset

- PERSONA-CHAT is **one of the biggest** multi-turn dialog dataset :
  - 164,356 utterances and about 1-2M words
  - Average number of turns: 14
- But it is still **small** for training a deep learning model:
  - 1B words in the Billion Words dataset
  - ~1M sentences in CoNLL 2012 (used for training co-reference systems)
- And generating an engaging open-domain dialogue requires:
  - topic-coherence,
  - dialogue-flow,
  - common-sense,
  - short term memory,
  - co-reference resolution,
  - sentimental analysis,
  - textual entailment…

Validation set (public) Leaderboard – Test set (hidden) Leaderboard

Model	Creator	PPL	Hits@1	F1
	🤗 (Hugging Face)	23.05🍎	74.3🍎	17.85🍎
	Team Pat	-	-	17.85
	Pinta	-	51.4	17.25
	Mohd Shadab Alam	35.57	14.8	16.94
	Sonic	38.87	-	16.88
	NEUROBOTICS	39.7	-	16.82
	Happy Minions	34.57	68.1	16.72
	1st-contact	36.54	13.3	16.58
	Tensorborne	44.64	12.1	16.13
	flooders	-	-	15.96
	Lost in Conversation	62.83	-	15.91
	High Five	59.83	78.2	15.34
	Little Baby	-	72.9	-
	loopAI	-	29.7	-
	Salty Fish	42.3	-	-

Model	Creator	PPL	Hits@1	F1
	🤗 (Hugging Face)	20.47🍎	74.7🍎	17.52🍎
	Little Baby	-	61.0	-
	Happy Minions	32.94	52.1	14.76
	High Five	52.8	50.3	13.73
	Pinta	-	44.4	16.52
	loopAI	-	25.6	-
	Mohd Shadab Alam	30.97	14.4	16.44
	1st-contact	31.98	13.2	16.42
	Tensorborne	38.24	12.0	15.94
	Team Dialog 6	40.35	10.9	7.27
	NEUROBOTICS	35.47	-	16.68
	Sonic	33.46	-	16.67
	Lost in Conversation	55.84	-	15.74
	flooders	-	-	15.47
	Team Pat	-	-	13.23
	Salty Fish	45.87	-	-
Seq2Seq + Attention	ParlAI team	29.8	12.6	16.18
Language Model	ParlAI team	46.0	-	15.02
KV Profile Memory	ParlAI team	-	55.2	11.9

- Small dataset =>
  - Large models are overfitting
  - Small models are underfitting

Transfer Learning

A two-stage procedure

1. *Pre-train* the model on a **large** dataset:
   - which is **not** the dataset you will use in the end,
   - but on which you hope to **learn general concepts** that will help in your case
2. *Adapt* the model on your **small** dataset:
   - to make it perform **well on your task**.

Pre-training

1. The model is pre-trained on
    - a **large dataset** of **contiguous** span of texts (Toronto Book Corpus: **~7000 books**)
    - with a *Language Modeling* objective (as we've just seen).

- Learns initial parameters of the neural network model.
- Provide the model with
    - some **kind of world knowledge** and
    - an ability to **build coherent sentences** by processing long-range dependencies.

- In our experiments, we started from the pre-trained model of Radford et al. 2018.

A Simple Method for Commonsense Reasoning by Trinh & Le (2018), *Improving, Language Understanding by Generative Pre-Training* by Radford et al. (2018), *Universal Language Model Fine-tuning for Text Classification* by Howard and Ruder (2018), *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* by Jacob Devlin et al (2018)

Differences 👻

Adaptation phase: Training dataset

Dataset for Fine-Tuning

Only used a sub-set of the full PERSONA-CHAT dataset:

- The training dataset with « original personalities »
  Zhang S. et al. Personalizing Dialogue Agents: I have a dog, do you have pets too?

NEUROMATION

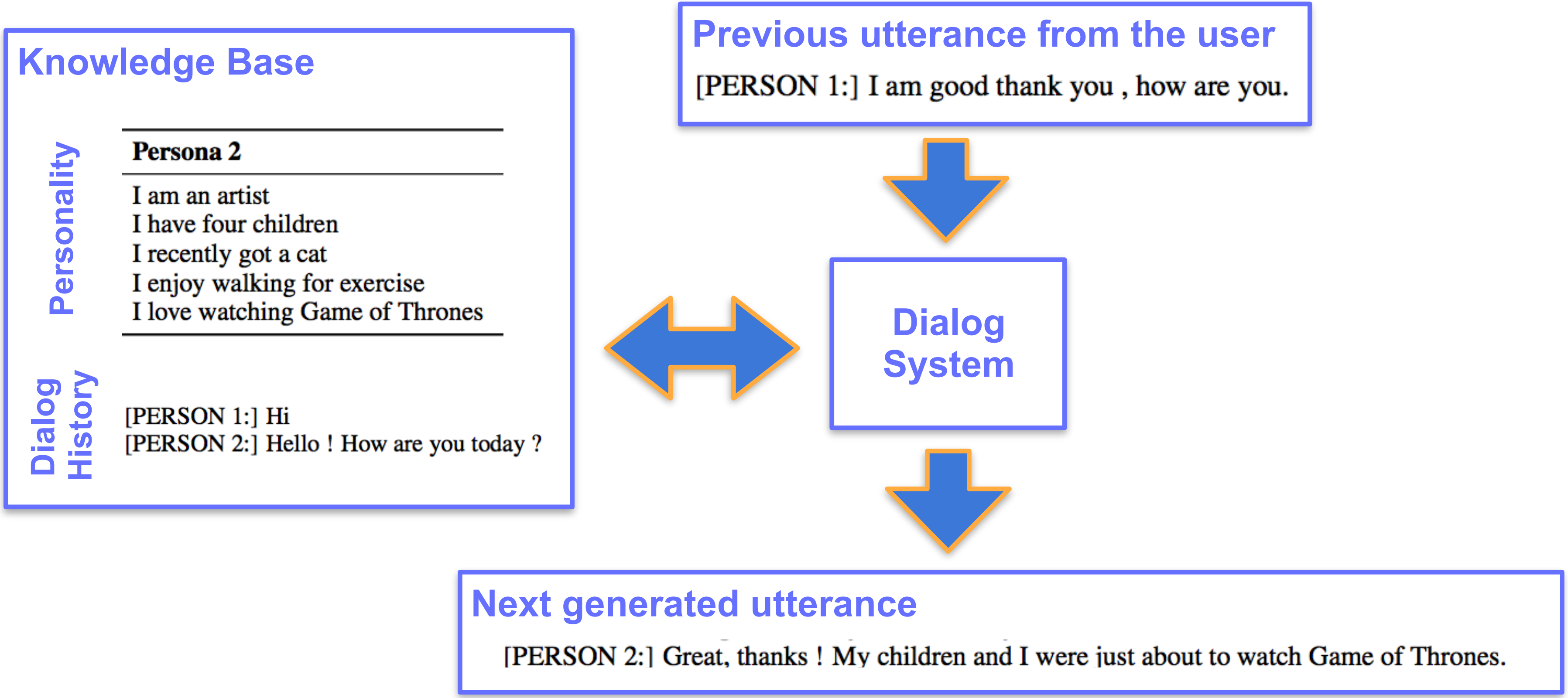
Uses a combination of 2 dialog datasets:

- PERSONA-CHAT with original and revised personalities
  Zhang S. et al. Personalizing Dialogue Agents: I have a dog, do you have pets too?
- DialyDialog dataset
  Li Y. et al. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset

Adaptation phase: Adapting the Architecture

Adapting a Language Model for Dialog

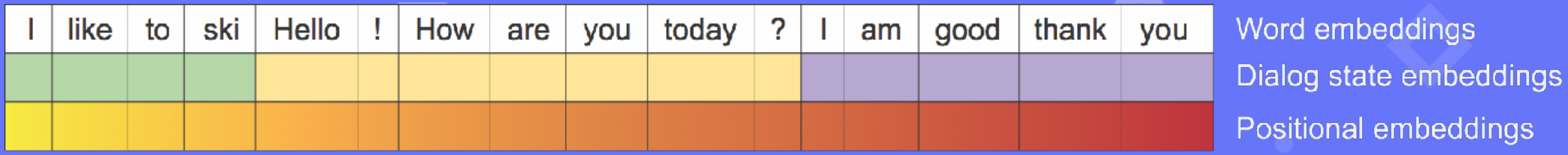
Several inputs with different types



Huggingface Approach – Semi-Sequential Encoding

- **After pre-training** we have a model with basic common-sense and co-reference capabilities, now we need to teach it the specificities of dialog:
  - Alternating utterances
  - Dialog flow (« speech/dialog acts »)
  - Conditioning on a personality

- How to build a sequential inputs for our model from a conditioned dialog?
  - Transformers don’t possess a natural notion of sequentiality and position
  - We already have positional embeddings to incorporate sequentiality
  - We add special embeddings related to utterances and personas



Huggingface Approach – Semi-Sequential Encoding

- We can play with these embeddings to manipulate the notion of a sequence

Repeating specific embeddings to control positioning information

I	like	to	ski	I	hate	mexican	food	I	like	to	eat	cheetos

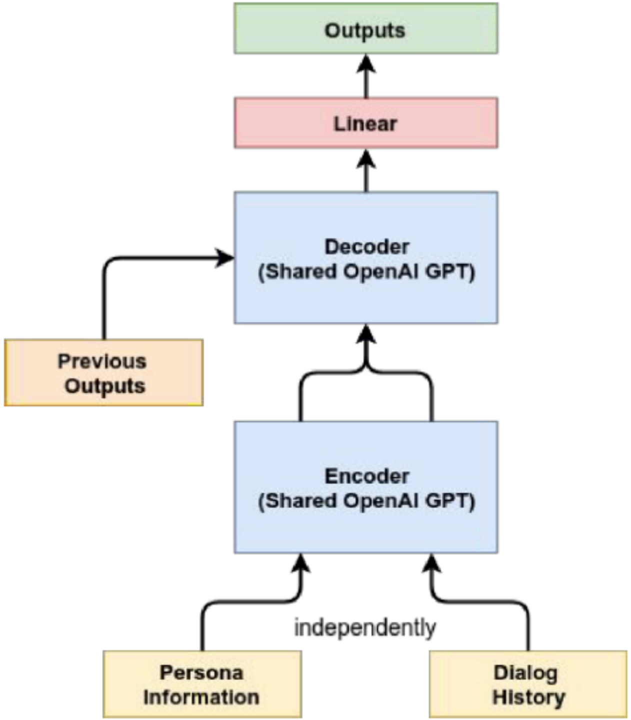
- We can also augment the dataset to bias towards positional invariance

I	hate	mexican	food	I	like	to	eat	cheetos	I	like	to	ski

I	like	to	ski	I	hate	mexican	food	I	like	to	eat	cheetos

Permutation augmented dataset to bias towards positional invariance

Lost In Conversation Approach – Dual-Model Encoding

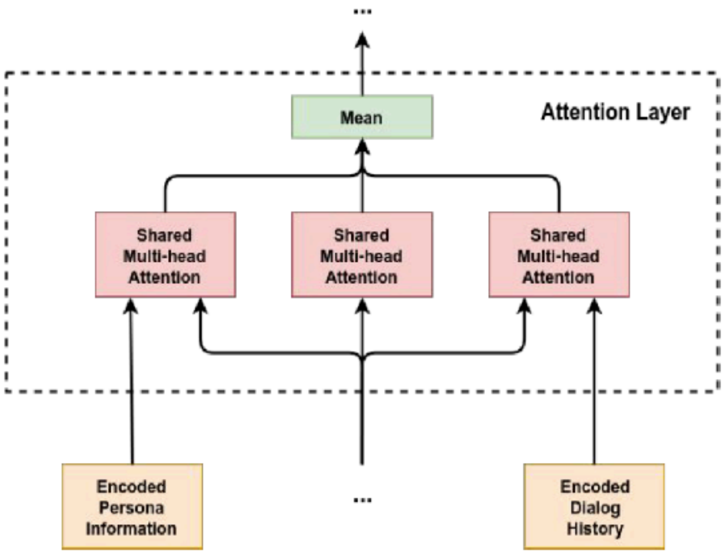


Shared encoder and decoder:

- Shared pre-softmax linear layer and token embeddings
- Reduction of persona information and dialog history – first and last 512 tokens respectively

NEUROMATION

Lost In Conversation Approach – Dual-Model Encoding



Attention layer modifications:

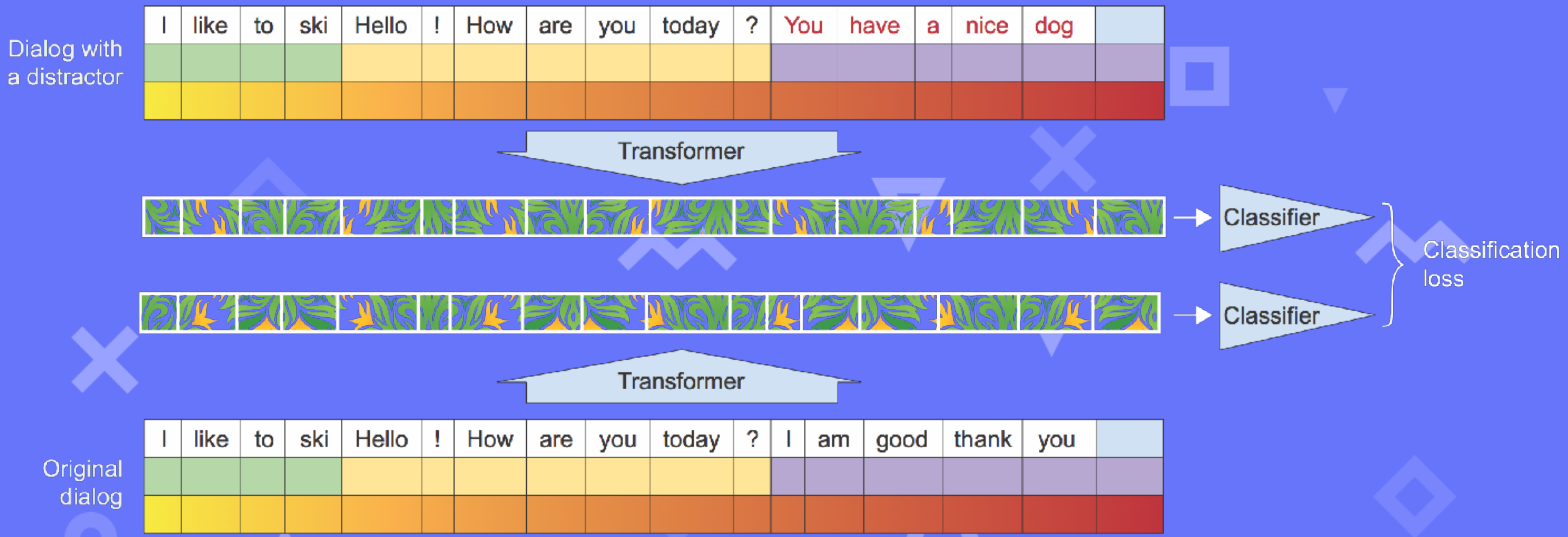
- Shared multi-head attention layers
- Parallel computation of attention for inputs
- Merge of attentions - mean

NEUROMATION

Adaptation phase: Training Objective

Huggingface Approach – Token level & Semantic Loss

- Learning to distinguish a real answer from a distractor.



- Weighted combination with a language modeling

Lost In Conversation – Token and Sequence level Losses

To train model we used weighted combination of losses[1]:

$$Loss = L_{TokLS} + \lambda_{LM} \cdot L_{LM} + \lambda_{risk} \cdot L_{risk}$$

$$L_{TokLS} = -\sum_i \log P(y_i|y_1,\dots,y_{i-1}) - D_{KL}(f||P(y_i|y_1,\dots,y_{i-1}))$$

$$L_{LM} = -\sum_i \log P(y_i|y_1,\dots,y_{i-1})$$

$$L_{risk} = \sum_{y_{pred}\in B} (1 - f1(y_{target}, y_{pred})) \cdot \frac{p(y_{pred})}{\sum_{y'_{pred}\in B} p(y'_{pred})}$$

First stage:

- $\lambda_{LM} = 0.5$
- $\lambda_{risk} = 0$

- $\lambda_{LM} = 0.1$:
- $\lambda_{risk} = 10$

Beam-search samples

for risk minimization

1. *Edunov S. et al. Classical Structured Prediction Losses for Sequence to Sequence Learning*

Decoding – Beam Search 🔊

Dataset for Fine-Tuning

Beam Search with
- length penalty
- basic n-gram filtering (rule of the completion)

NEUROMATION

Beam-search with:
- length penalty
- annealing for diversity

Wrap-Up

A very subjective wrap-up

(Probably) Good Ideas

- **Huggingface:**
  - Adding additional dialog embeddings
  - Next sentence prediction loss (effect on LM?)
- **Lost in Conversation:**
  - Bigger adaptation dataset
  - Sequence level and risk losses (is F1 the right metric?)

More Questionable Choices

- **Huggingface:**
  - Over fitting to the adaptation dataset
  - Strong exposure bias problem
- **Lost in Conversation:**
  - Dual-model learning
  - Sharing positional embeddings

Human Evaluations
&
Automatic Metrics

👷 🕵️‍♂️ 👩‍⚕️ 👨‍🎤 👩‍🎓 👩‍🌾 👨‍🚀

Human Evaluation on Huggingface's model

BOT IN BLUE

Hello there, I like rock climbing do you have any hobbies?

i like nature. do you? i like nature.

I do love nature, especially when going on walks

what is your favorite rock? mine is rock.

Well I like rock music, does that count?

i think it does. what do you do for fun?

Well I work at a restaurant that I love, and rock climbing

that sounds fun! i'm into nature.

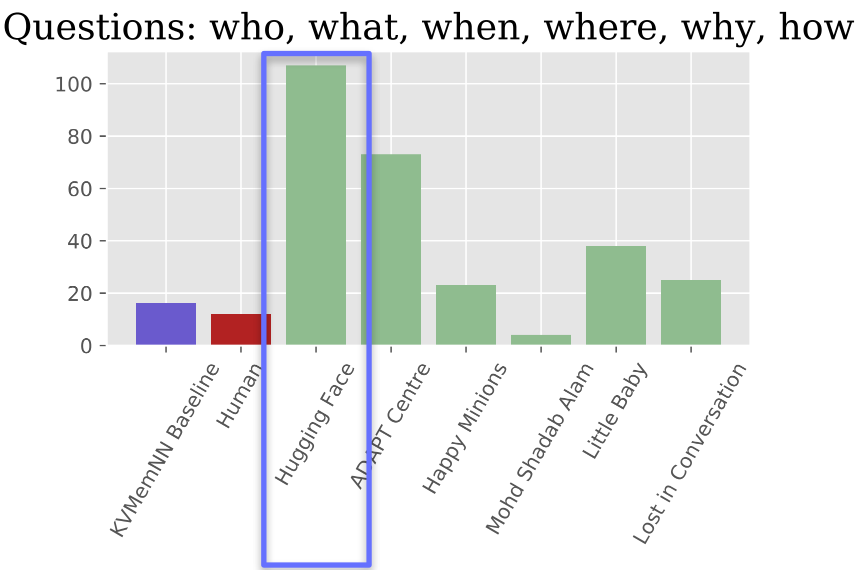
Cool, so am I

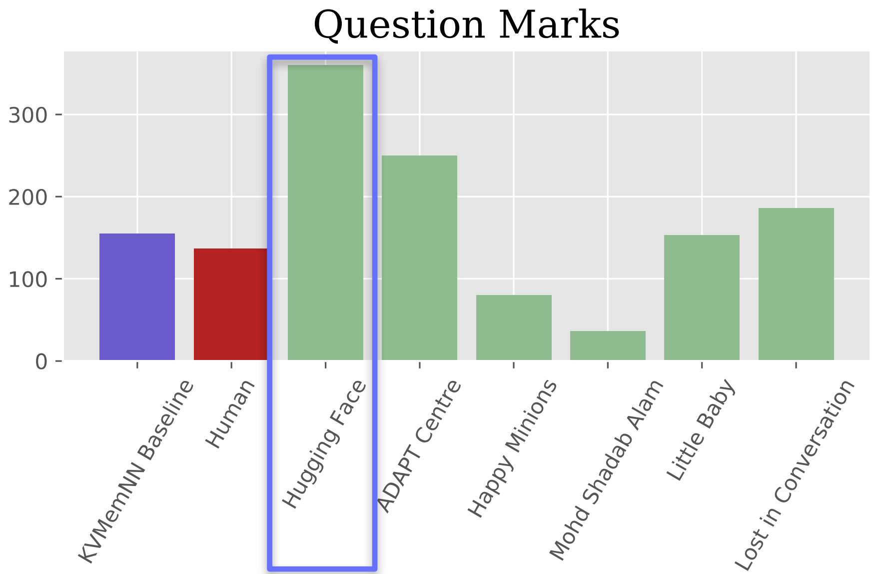
do you have any pets? do you have pets?

User feedbacks

[Adapted from slides by Emily Dinan / Jason Weston – NeurIPS 2018]

Too much questions





[Adapted from slides by Emily Dinan / Jason Weston – NeurIPS 2018]

Evaluation in Natural Language Generation

An Open Research Question

- **Automatic metrics** don't correlate well with **human evaluations**
- We (together with Microsoft, University of Washington, Stanford and Facebook) are organizing a workshop on this topic this summer in Minneapolis:

NeuralGen 2019: Methods for Optimizing and Evaluating Neural Language Generation

That’s it for today
Thanks for listening!

thomas@huggingface.co