Introduction to Deep Learning for Dialog Systems

이 화 란 | Hwaran Lee
hwaran.lee@gmail.com | hwaranlee.github.io

SK Telecom
Yeonsei Univ., June 19, 2020

Outline

I. Introduction to dialog systems

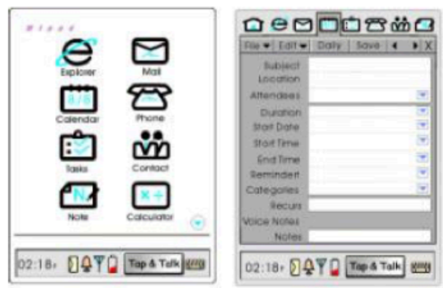
II. Deep learning for natural language
- Word embedding
- Language models

III. Toward end-to-end neural dialog systems for multi-domain task completion
- DSTC8
- Neural approaches
- Challenges

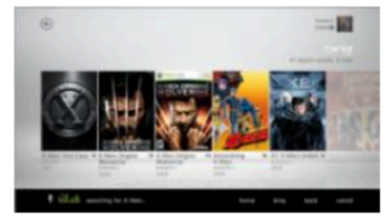
Brief History of Dialogue Systems

Multi-modal systems
e.g., Microsoft MiPad, Pocket PC

MiPad



TV Voice Search
e.g., Bing on Xbox



Virtual Personal Assistants

2017

Early 2000s

Early 1990s

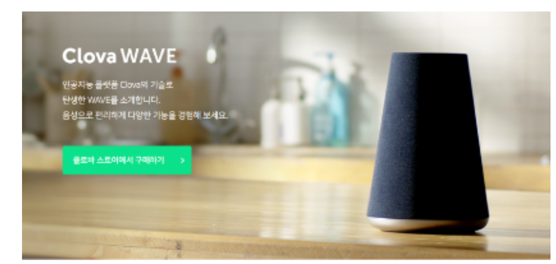
Task-specific argument extraction
(e.g., Nuance, SpeechWorks)
User: "I want to fly from Boston to New York next week."

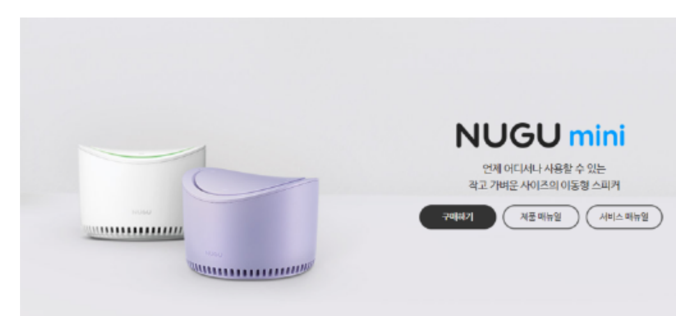
IBM WATSON

DARPA
CALO Project

Intent Determination
(Nuance's Emily™, AT&T HMIHY)
User: "Uh…we want to move…we want to change our phone line from this house to another house"

Keyword Spotting
(e.g., AT&T)
System: *"Please say collect, calling card, person, third number, or operator"*

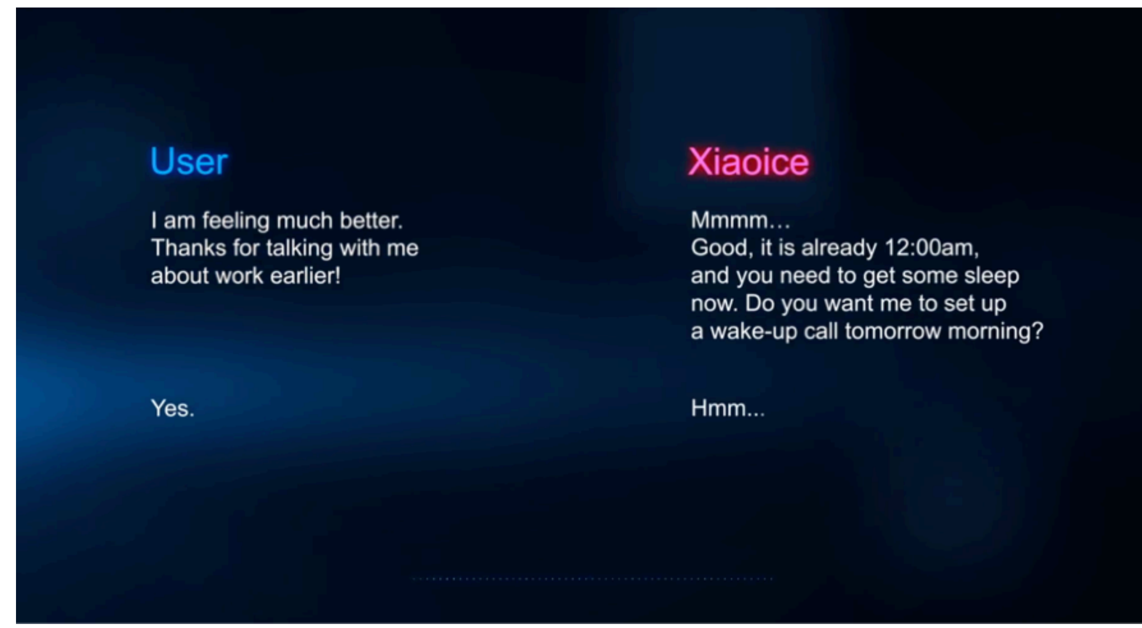




Material: https://deepdialogue.miulab.tw

Brief History of Dialogue Systems

Google, Duplex (2018)



Microsoft, Xiaoice (2018)

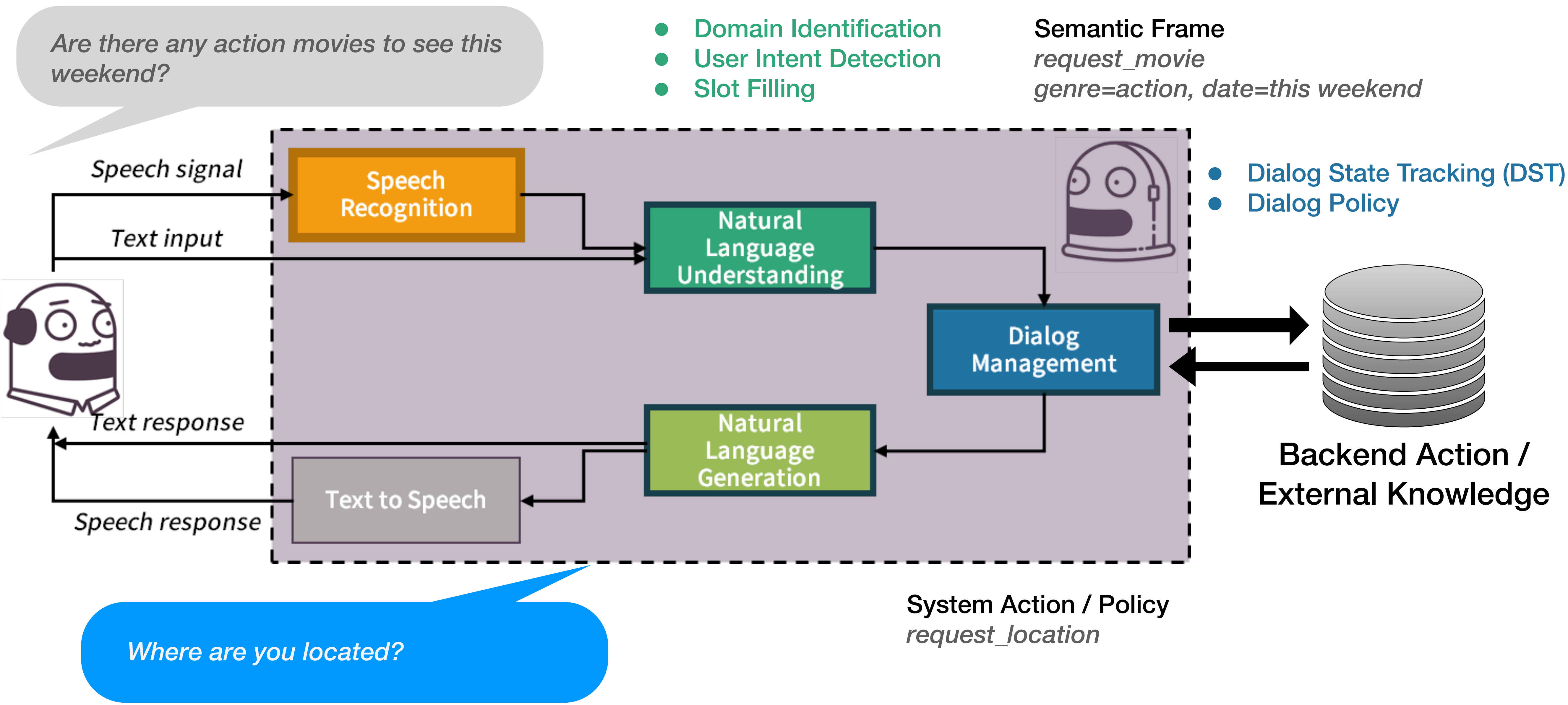
Naver Line, Duet (2019)

Category of Dialogue Systems

User says:	**Dialogue Category**
• **I am smart**	→ **Chitchat**
• **I have a question** ***When Iron Man is dead?***	→ **Question-Answering (Info)**
• **I need to get this done** ***I want to book a restaurant***	→ **Goal-oriented**

Material: Jianfeng Gao & Michel Galley, Tutorial, ICML, 2019

Spoken Dialog Systems



Transition of NLP to Neural Approaches

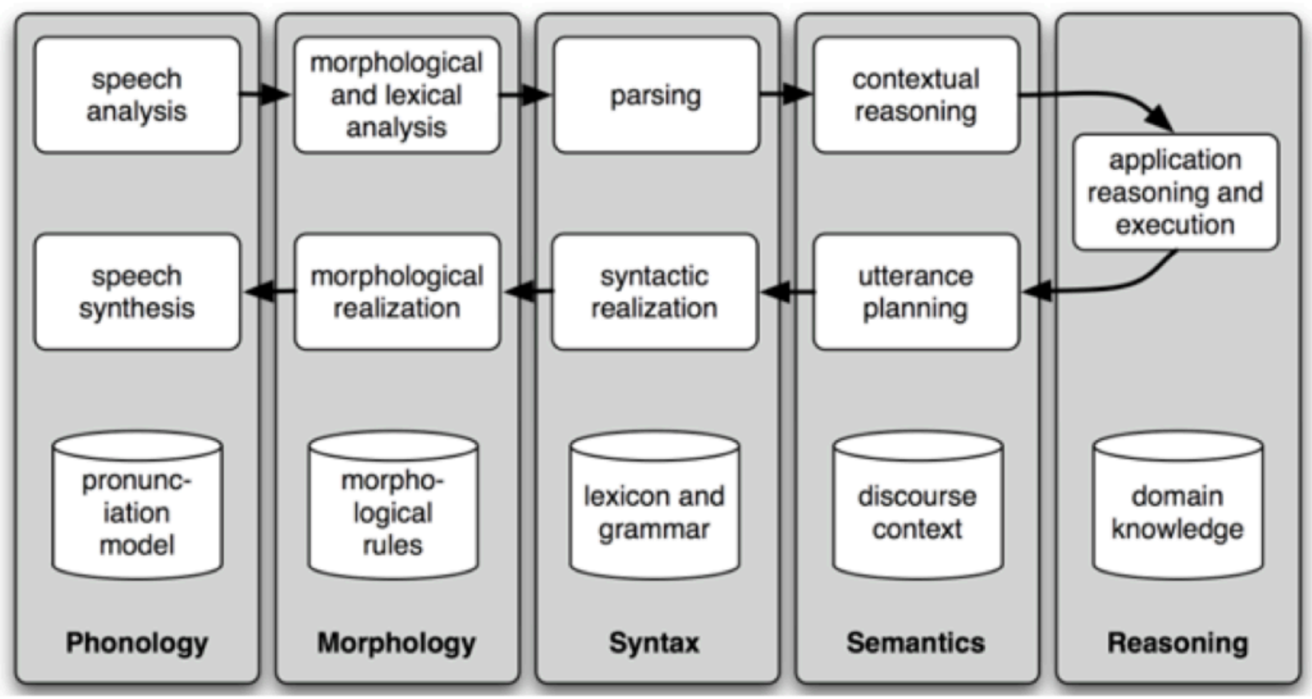
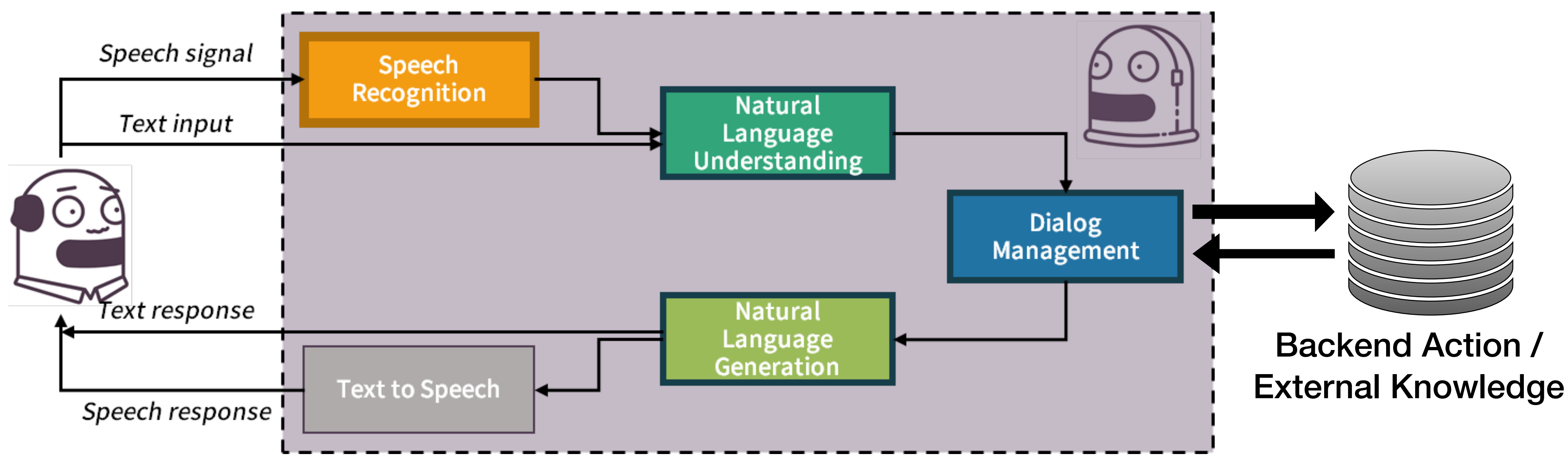


Figure 1.3: Traditional NLP Component Stack. Figure credit: Bird et al. (2009).

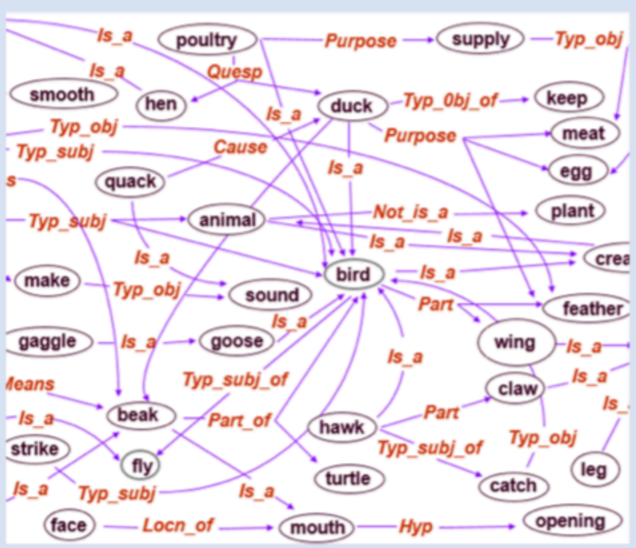
Neural Model for Each Module



Transition of NLP to Neural Approaches

Symbolic Space

- Knowledge is explicitly represented using words/relations/templates
- Reasoning is based on keyword matching, sensitive to paraphrase alternations
- Interpretable and efficient in execution but difficult to train E2E.



Input: Query

Symbolic → Neural
Encoding the query/knowledge

Reasoning in neural space to generate answer vector

E2E training via back propagation

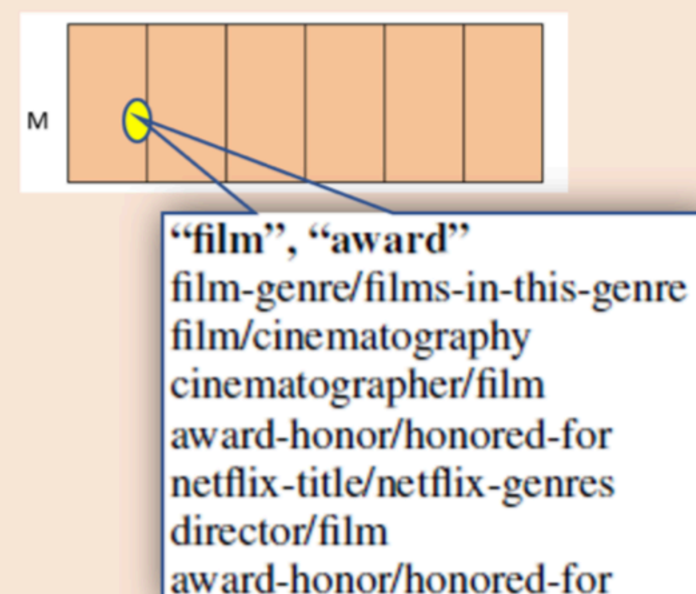
Errors

Output: Answer

Neural → Symbolic
Decoding the answer in NL

Neural Space

- Knowledge is implicitly represented by semantic classes as cont. vectors
- Reasoning is based on semantic matching, robust to paraphrase alternations
- Easy to train E2E, but uninterpretable and inefficient in execution



Outline

I. Introduction to dialog systems

II. Deep learning for Natural Language

- Word embedding
- Language models

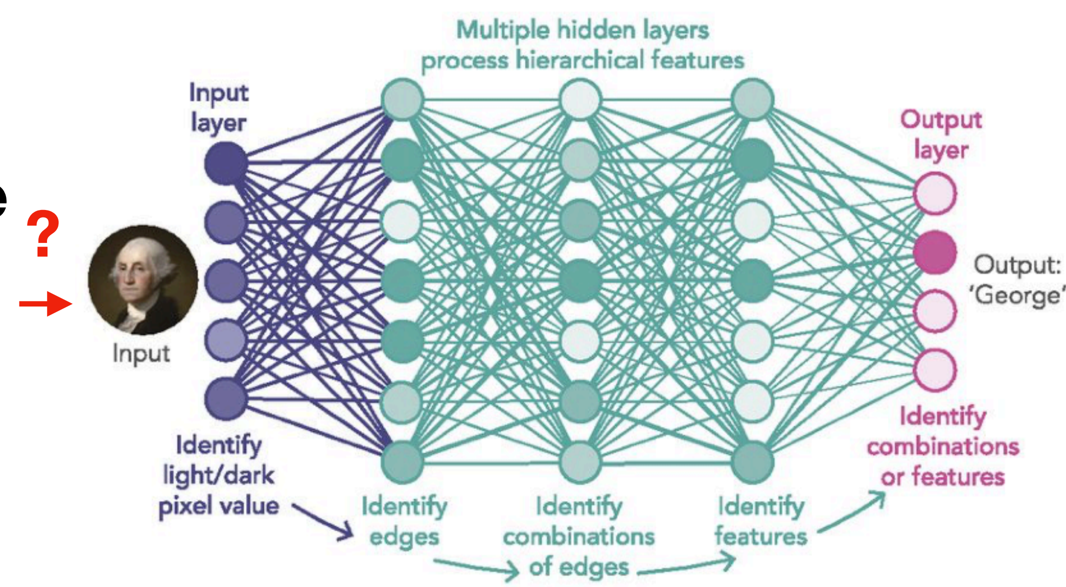
III. Toward end-to-end neural dialog systems for multi-domain task completion

- DSTC8
- Neural approaches
- Challenges

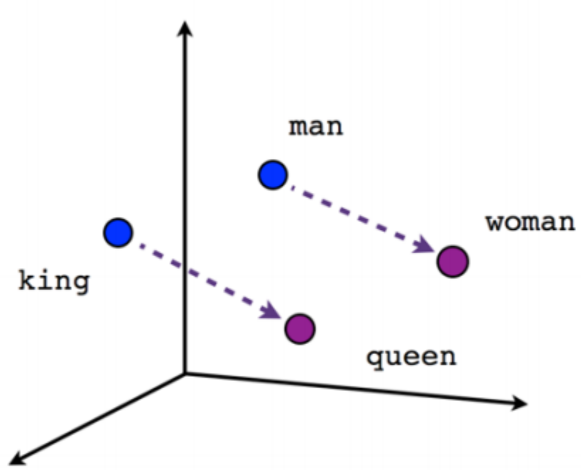
Word Embeddings (word2vec)

- How to represent word symbols as (semantic) vectors?

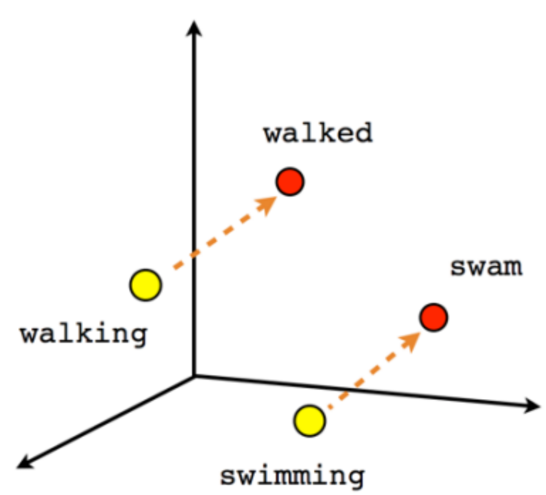
love
it
!



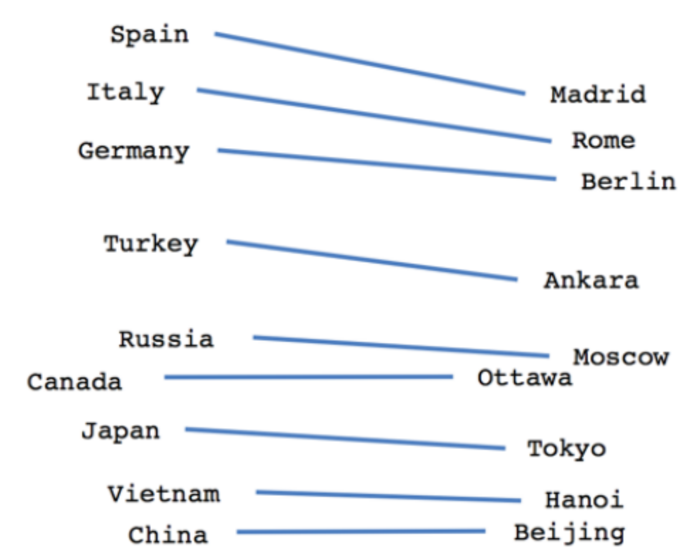
Pos
Neg



Male-Female



Verb tense

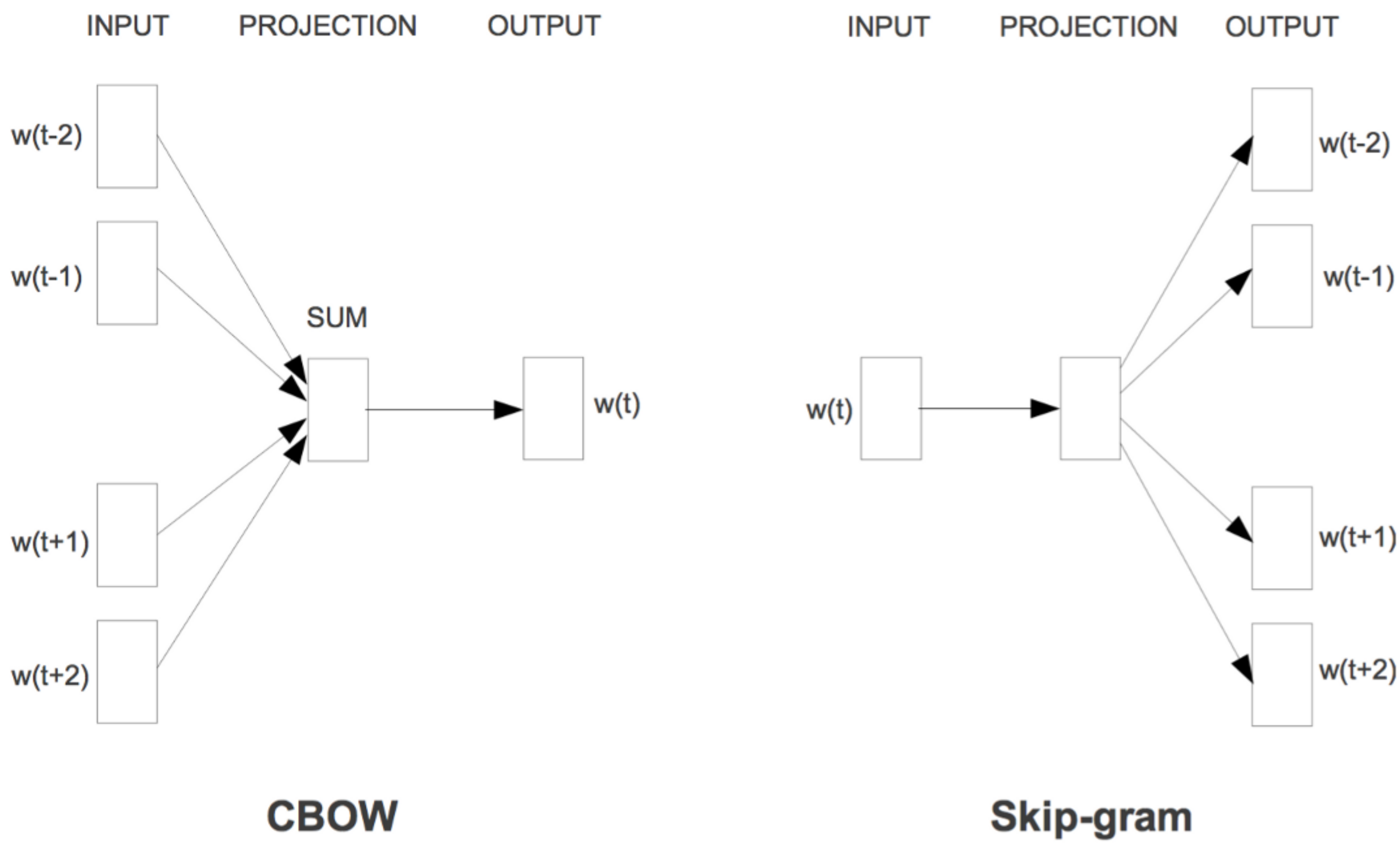


Country-Capital

https://developers.google.com/machine-learning/crash-course/embeddings/translating-to-a-lower-dimensional-space

Word Embeddings (word2vec)

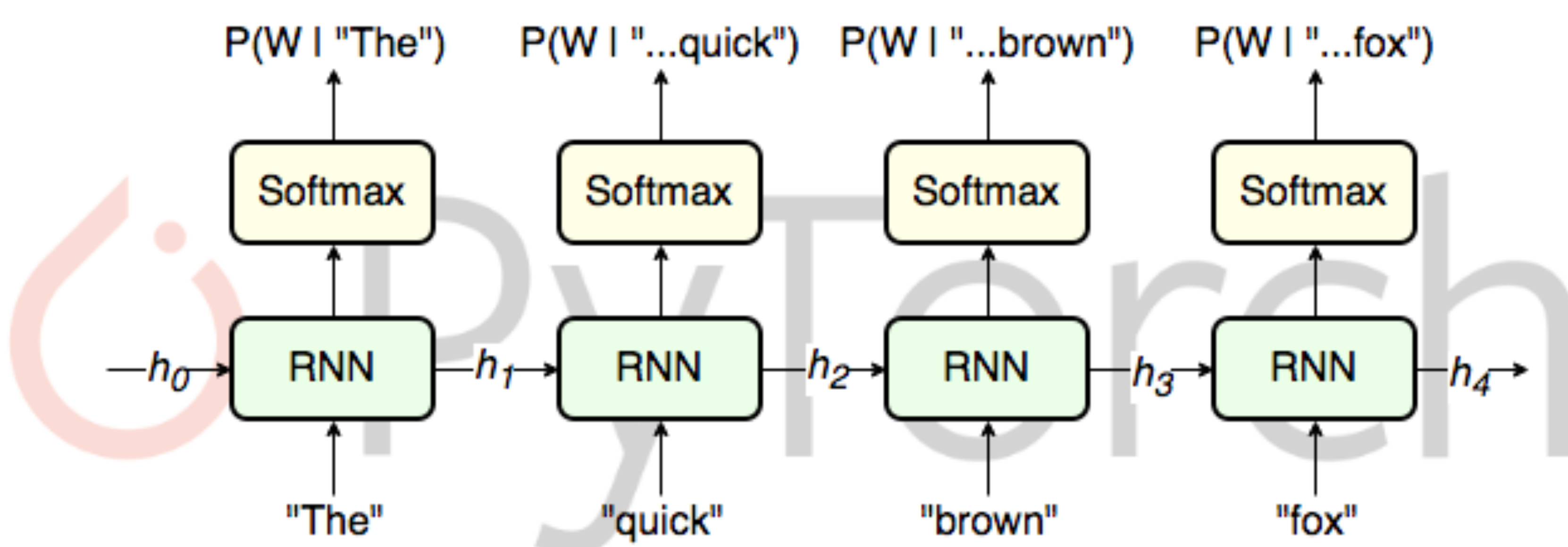
- Learn the meaning of a word from its neighborhoods!



T. Mikolove et al., Efficient Estimation of Word representations is vector space, 2013.

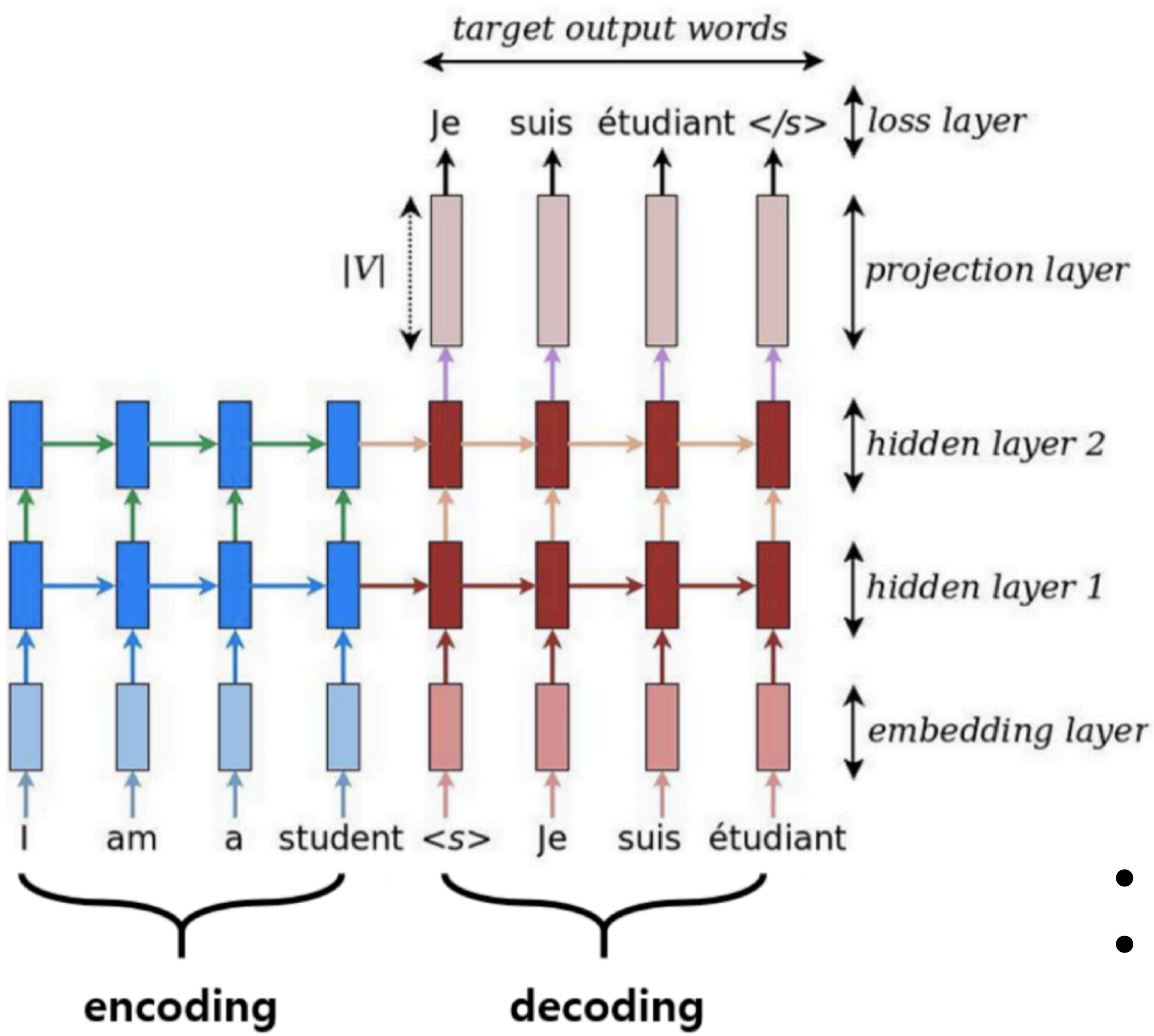
Language Model

- Probability of a sequence of m words: $p(w_1, w_2, \ldots w_m)$
  - Application: Choose the next word: $p(w_{m+1} \mid w_{1,\ldots,m})$
- N-Gram LM
  - $p(w_{m+1} \mid w_{m,m-1}) = \dfrac{count(w_{m+1}, w_m, w_{m-1})}{count(w_m, w_{m-1})}$ (tri-gram)
  - Count based approach has weakness on *unseen word sequence*
  - Fixed width context
- Neural Language Model
  - RNNLM (Mikolov, 2010)



Encoder-decoder architecture

- Sequence-to-sequence (Seq2seq)

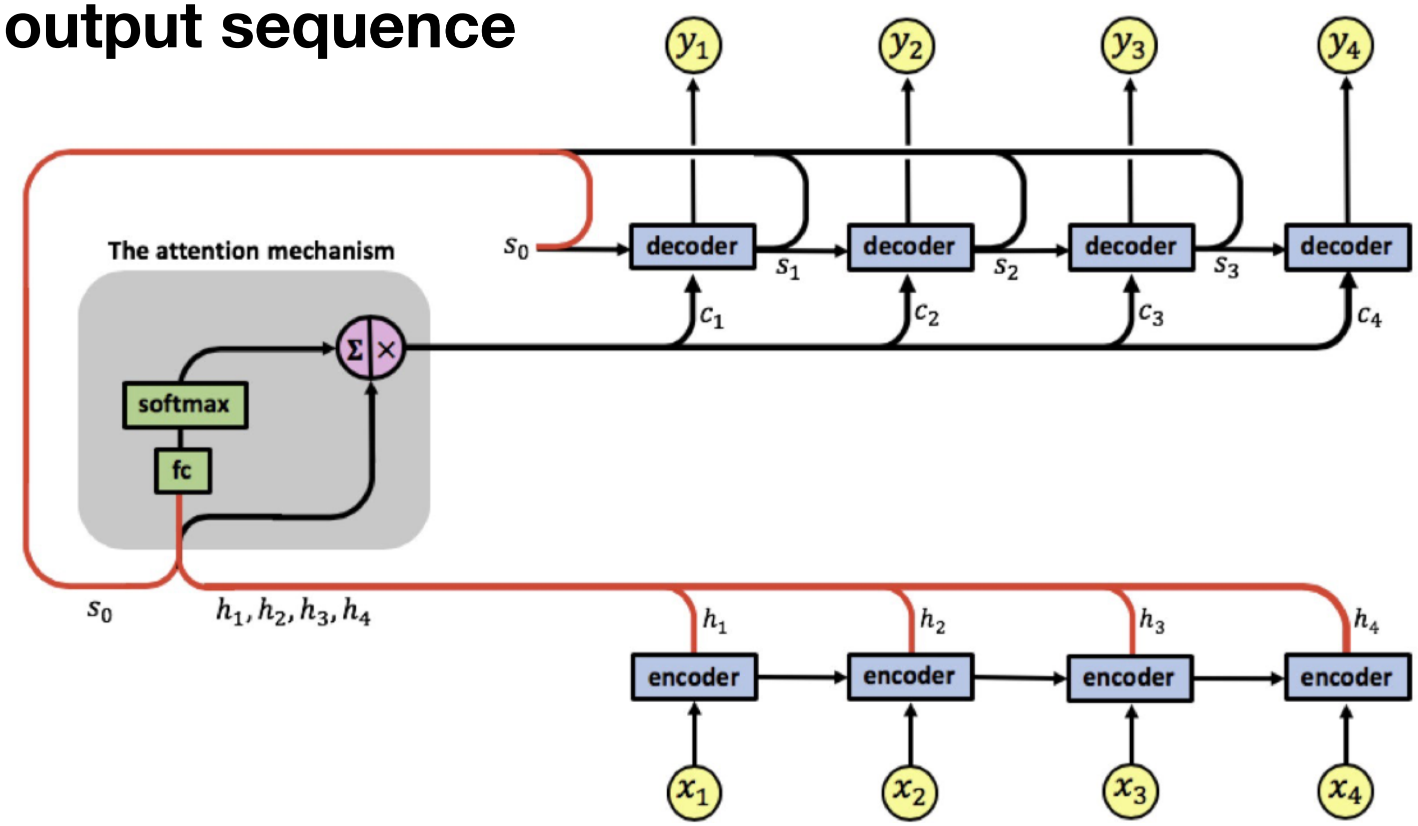


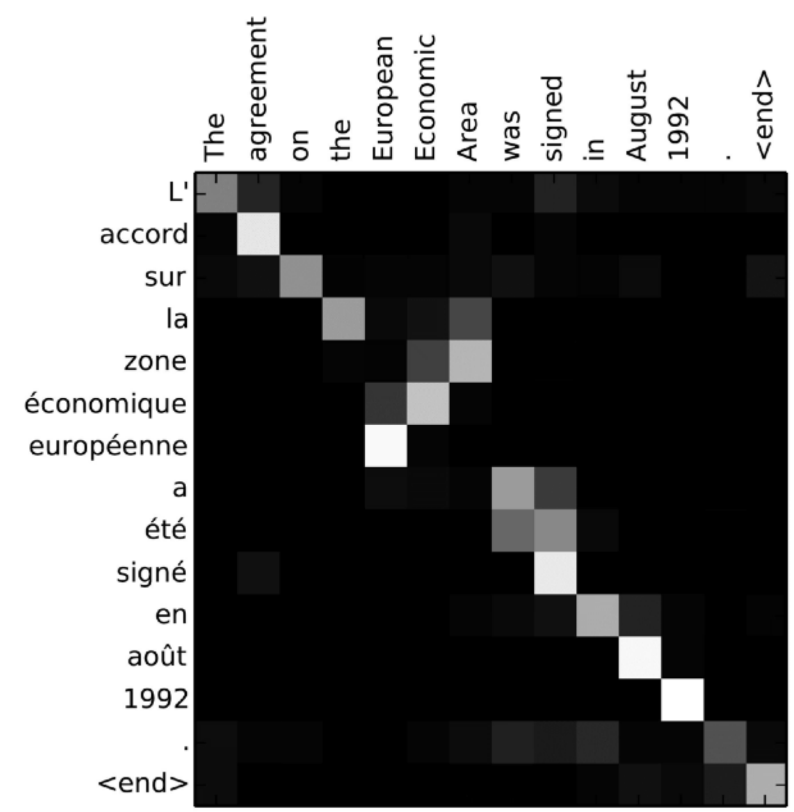
- **Machine translation**
- **Dialog Response generation**

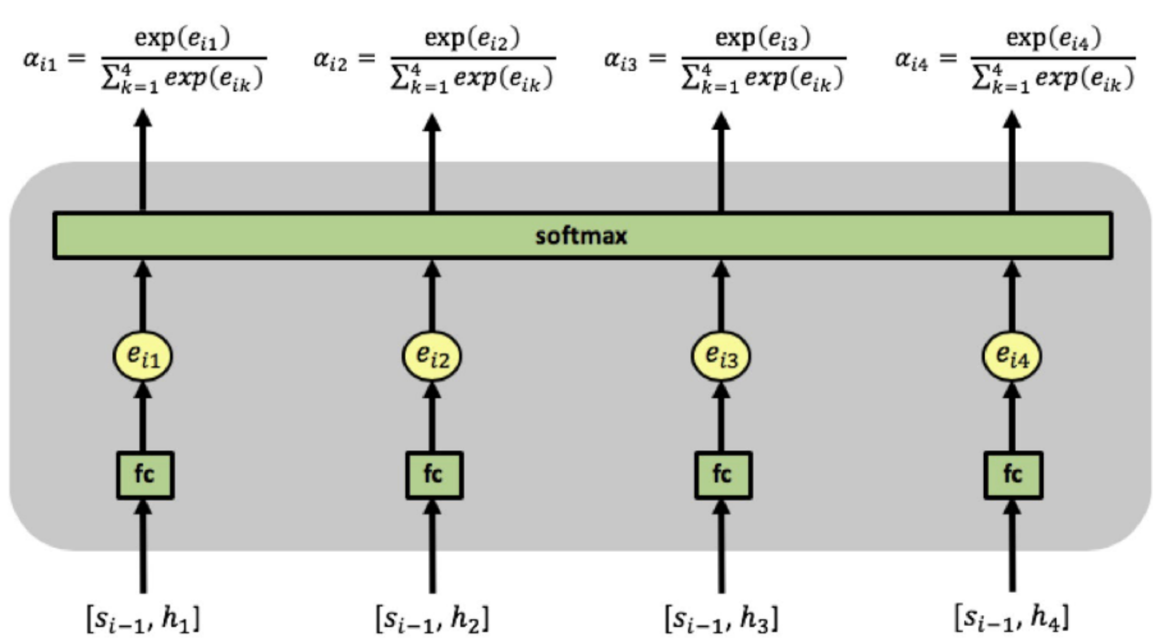
I. Sutskever., Sequence to Sequence Learning with Neural Networks, NIPS, 2014

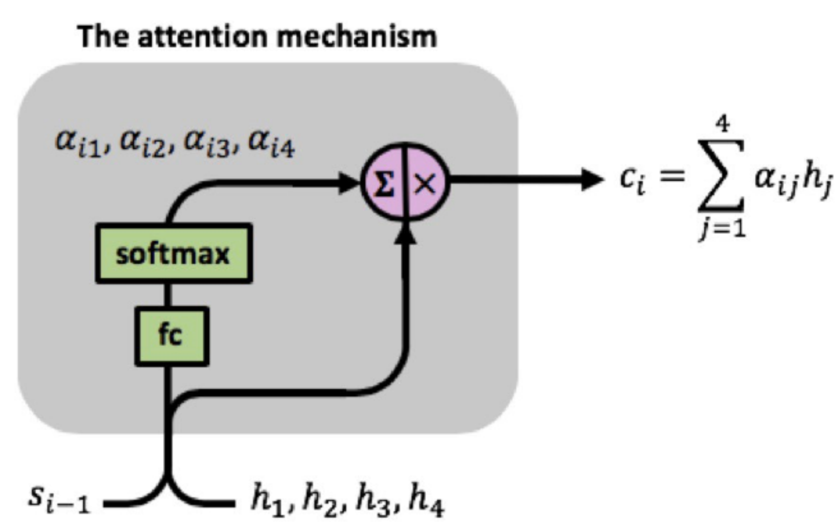
Attention Mechanism

Focus on certain parts **of the input sequence when predicting a certain part of the output sequence**



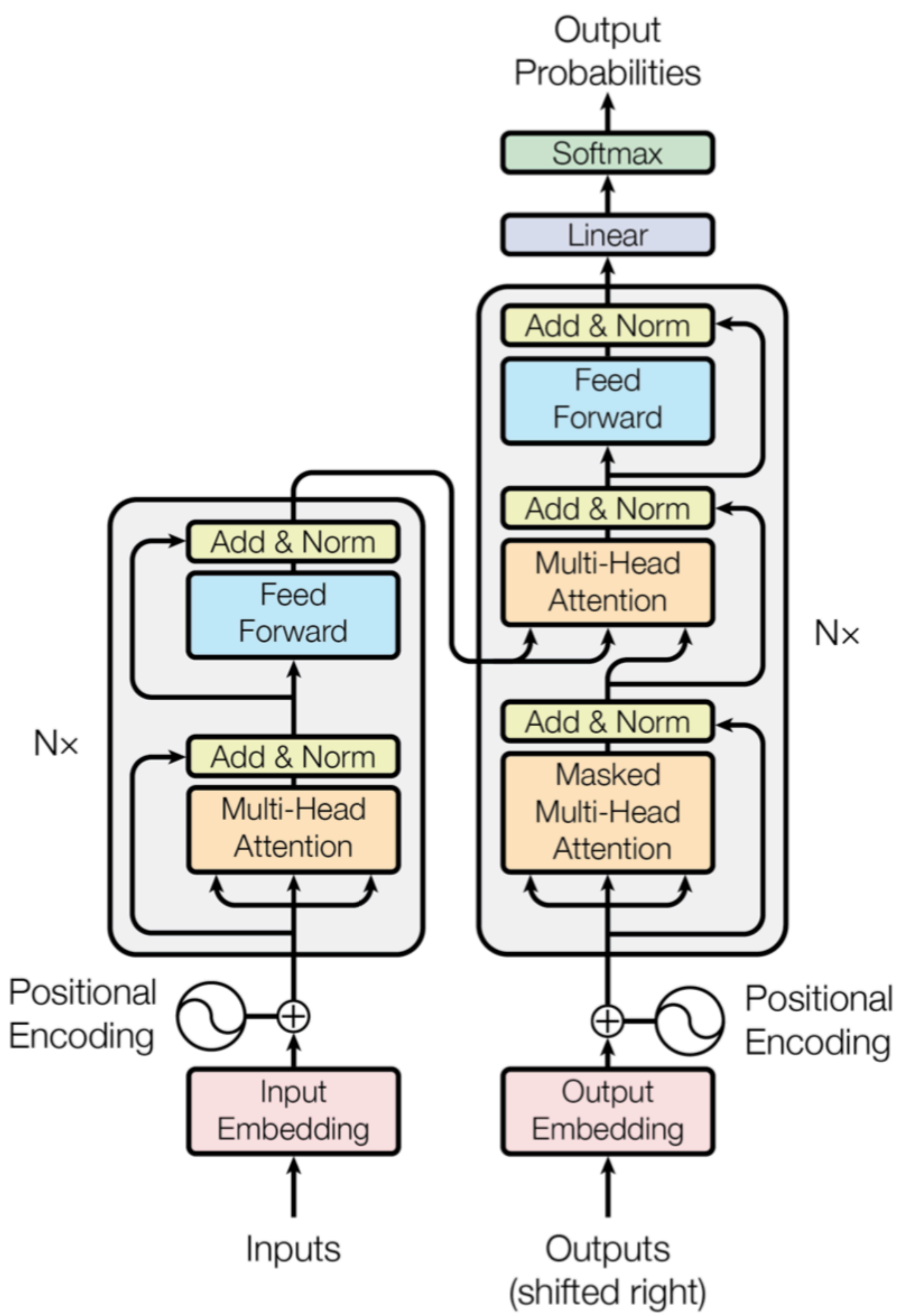






Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." arXiv preprint arXiv:1409.0473 (2014).

Transformer, Attention is all you need

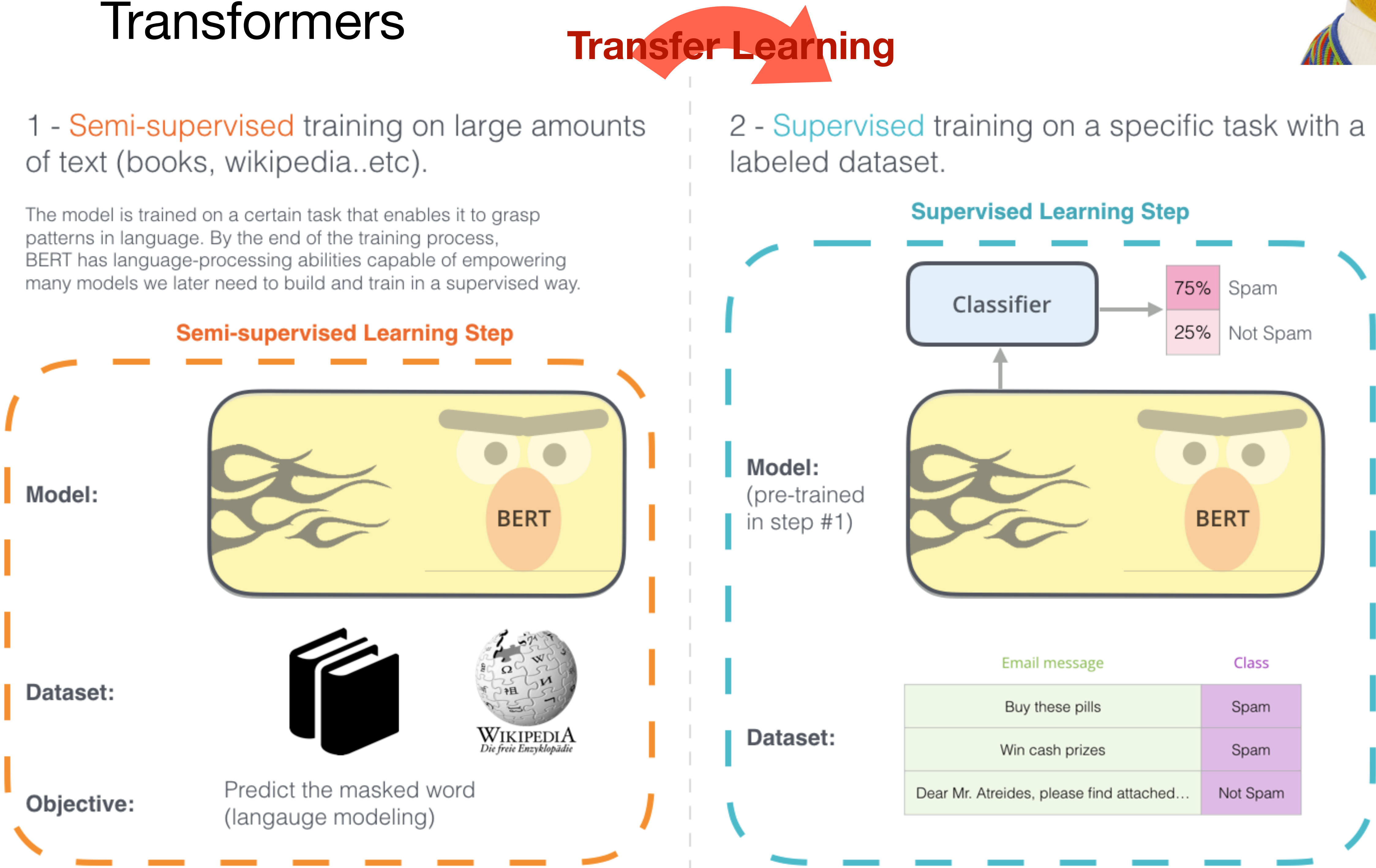


- **Without RNNs, only attention mechanism is used!**
  - **Self-attention**
  - **Multi-head attention**
  - **Positional encoding**

Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems. 2017.

Recent Word and Sentence Representation

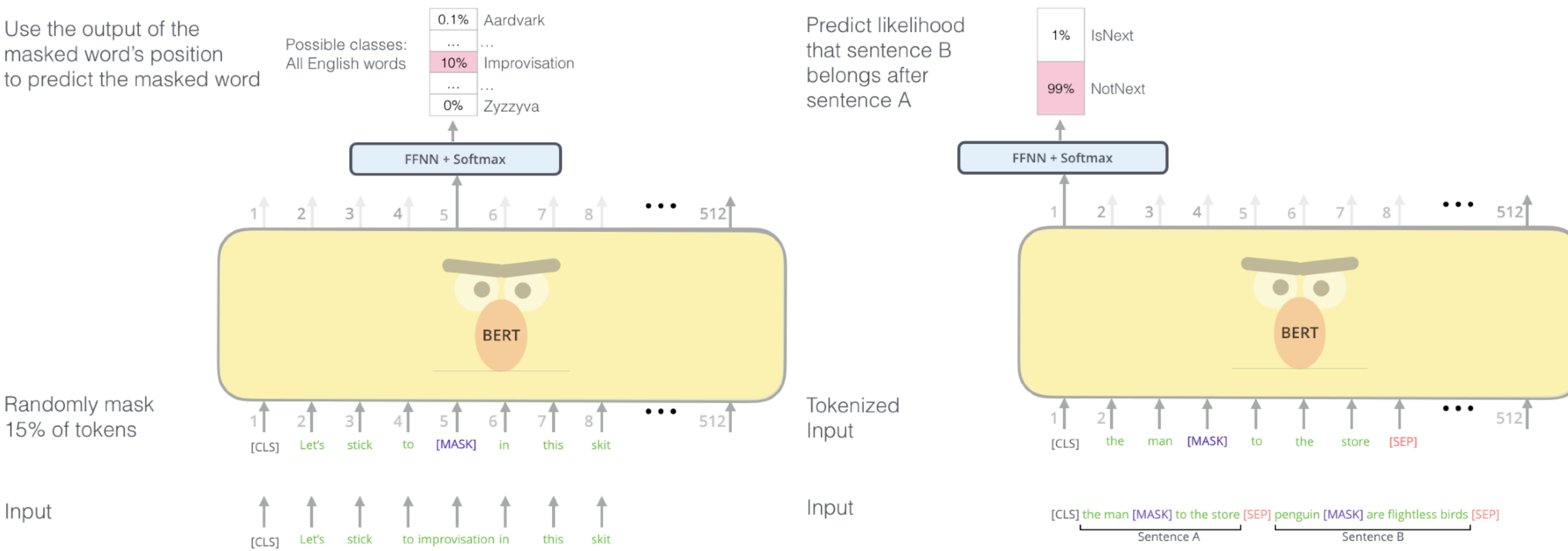
- BERT: Bi-directional Encoder Representations from Transformers



Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

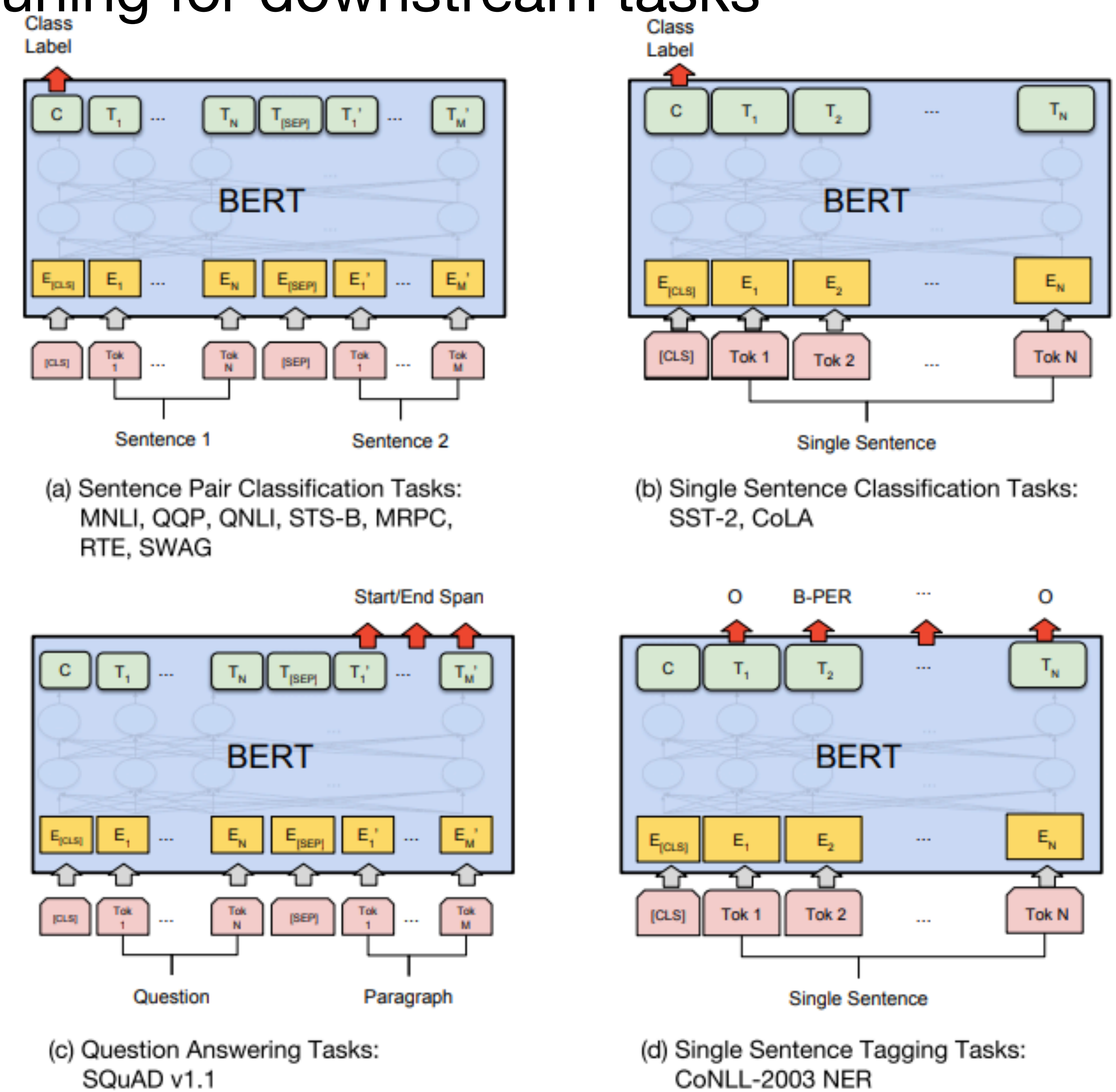
BERT

- Pretraining: Masked Language Model and Two-sentence Classification

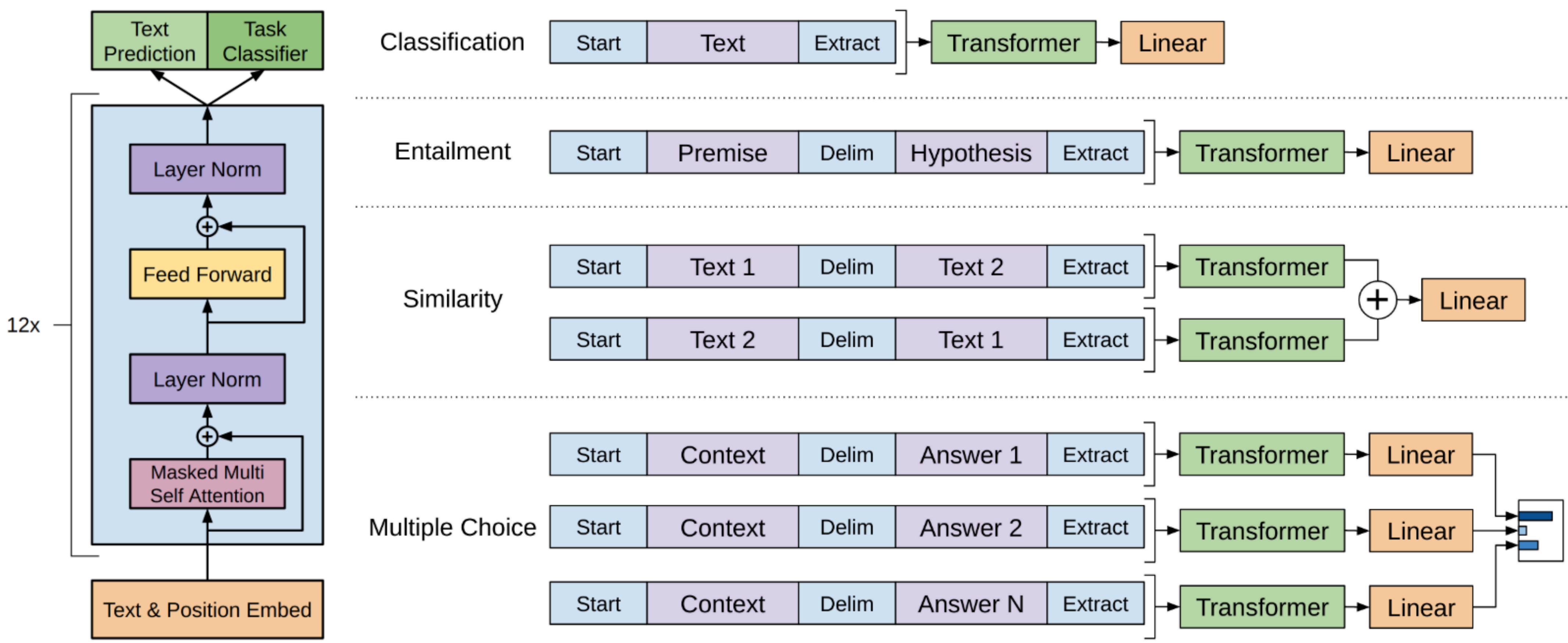


BERT

- Fine-tuning for downstream tasks



GPT & GPT-2: Generative Pre-Trained model



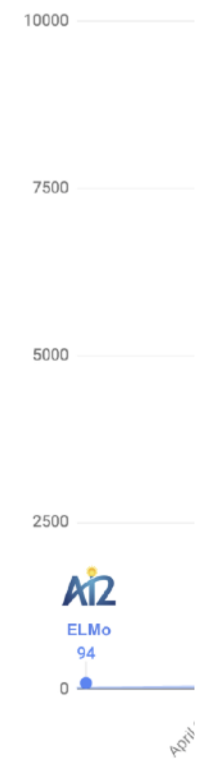
Note

- **GPT trains to predict the next token given previous token sequences**
- **BERT trains to predict the masked token given token contexts**

A. Radford et al., Improving Language Understanding by Generative Pre-Training, 2018
A. Radford et al., Language Models are Unsupervised Multitask Learners, 2019

VERY Recent Language Models

- XLNet, Google/CMU
- RoBERTa, Facebook
- ALBERT, Google/Toyota
- T5, Google
- StructBERT, Alibaba
- Reformer, Google
- Longformer, AllenAI
- ElECTRA, Google/Stanford
- GPT-3, OpenAI (May 2020!)



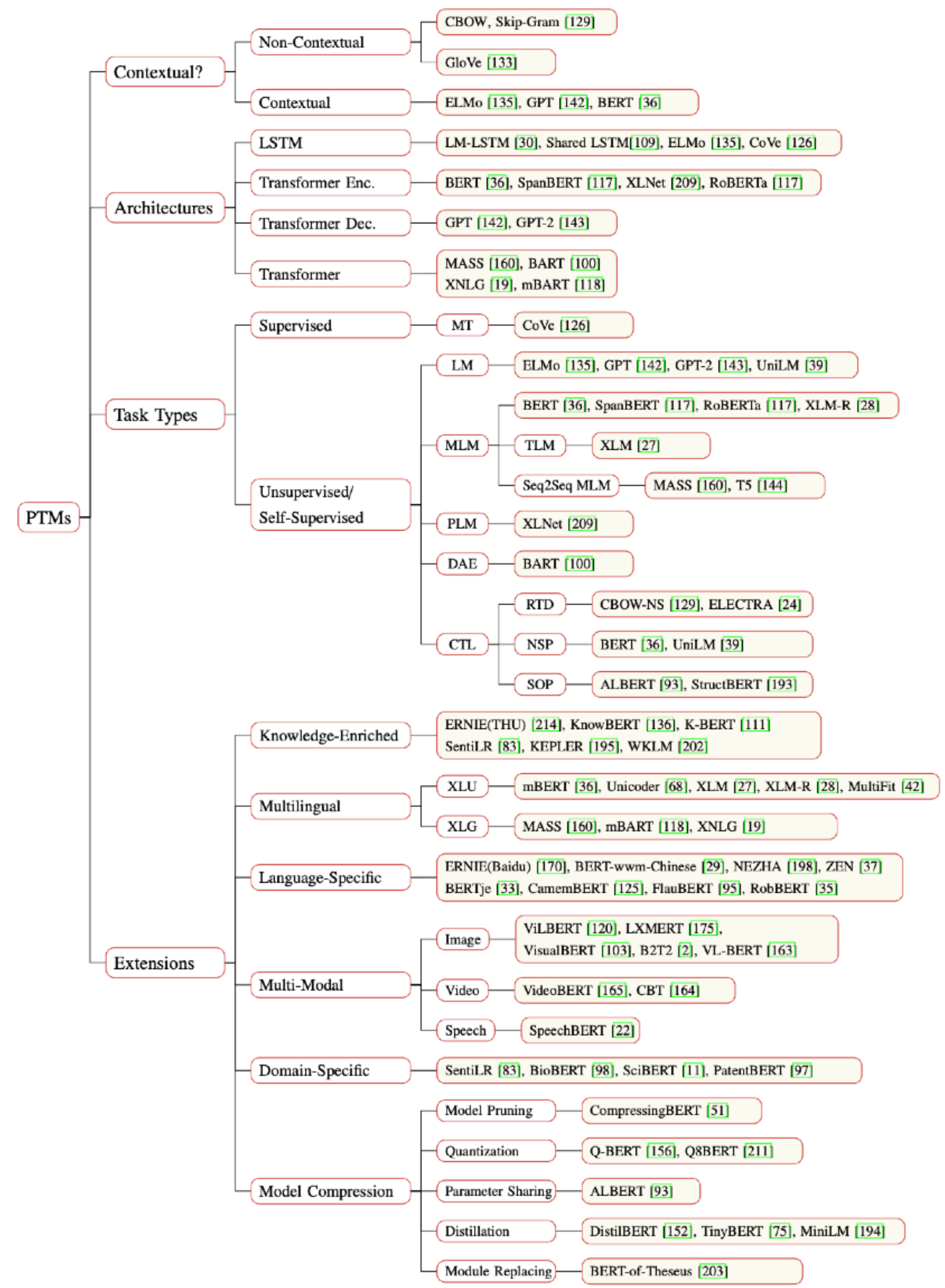


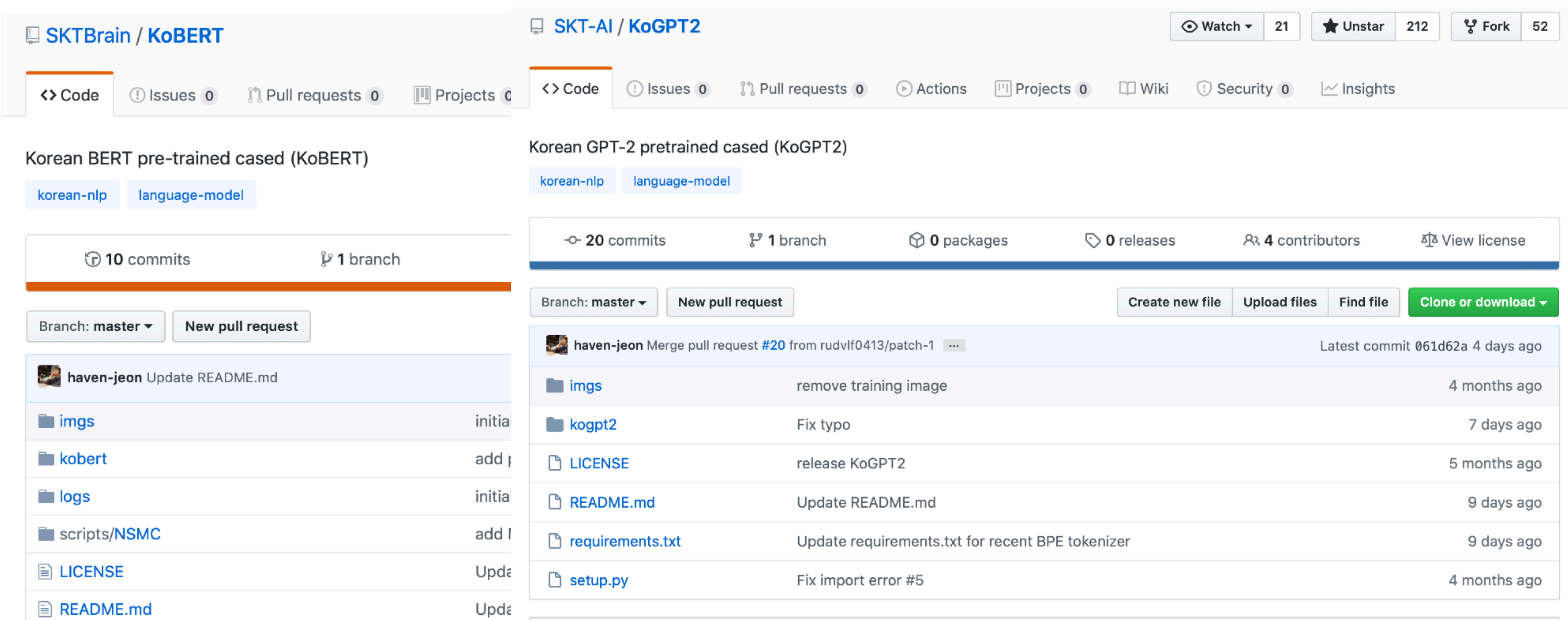
Figure 3: Taxonomy of PTMs with Representative Examples

Visit github.com/huggingface/transformers and enjoy manipulating!

Xipeng Qiu et al., Pre-trained Models for Natural Language Processing: A Survey, Apr. 2020, Arxiv

Pre-trained LMs for Korean

- Google Bert (Multilingual) : https://github.com/google-research/bert/blob/master/multilingual.md
- ETRI, KorBert: http://aiopen.etri.re.kr/service_dataset.php
- SKT, **KoBERT**: https://github.com/SKTBrain/KoBERT
- SKT, **KoGPT2**: https://github.com/SKT-AI/KoGPT2



Outline

I. Introduction to dialog systems

II. Deep learning for Natural Language
  - Word embedding
  - Language models

III. Toward end-to-end neural dialog systems for multi-domain task completion
  - DSTC8
  - Neural approaches
  - Challenges

Conversational Agents

Chit-Chat

seq2seq models → Seq2seq with conversation contexts → Knowledge-grounded seq2seq models

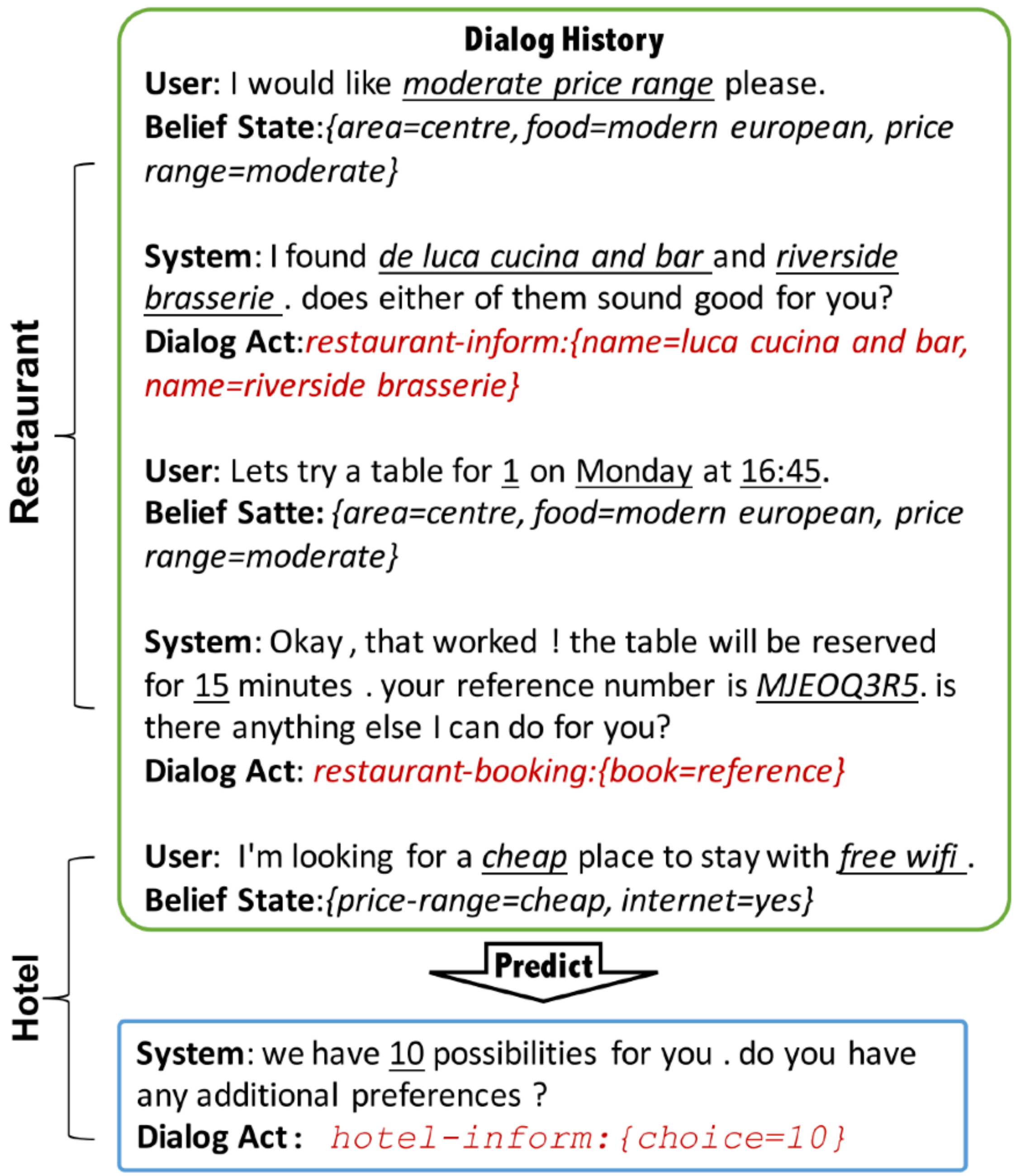
Task-Oriented

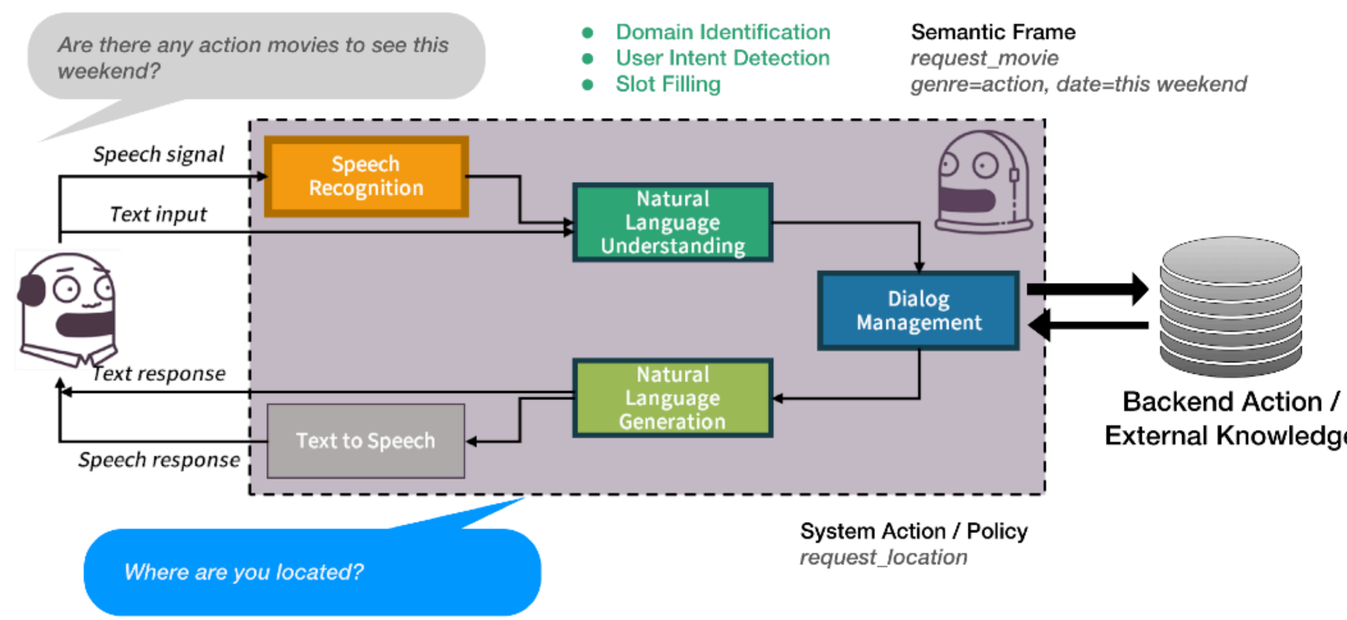
Single-domain, system-initiative → Multi-domain, contextual, mixed-initiative → End-to-end learning, massively multi-domain

Material: https://deepdialogue.miulab.tw

Multi-domain Goal-Oriented Dialogue System

MultiWOZ dataset

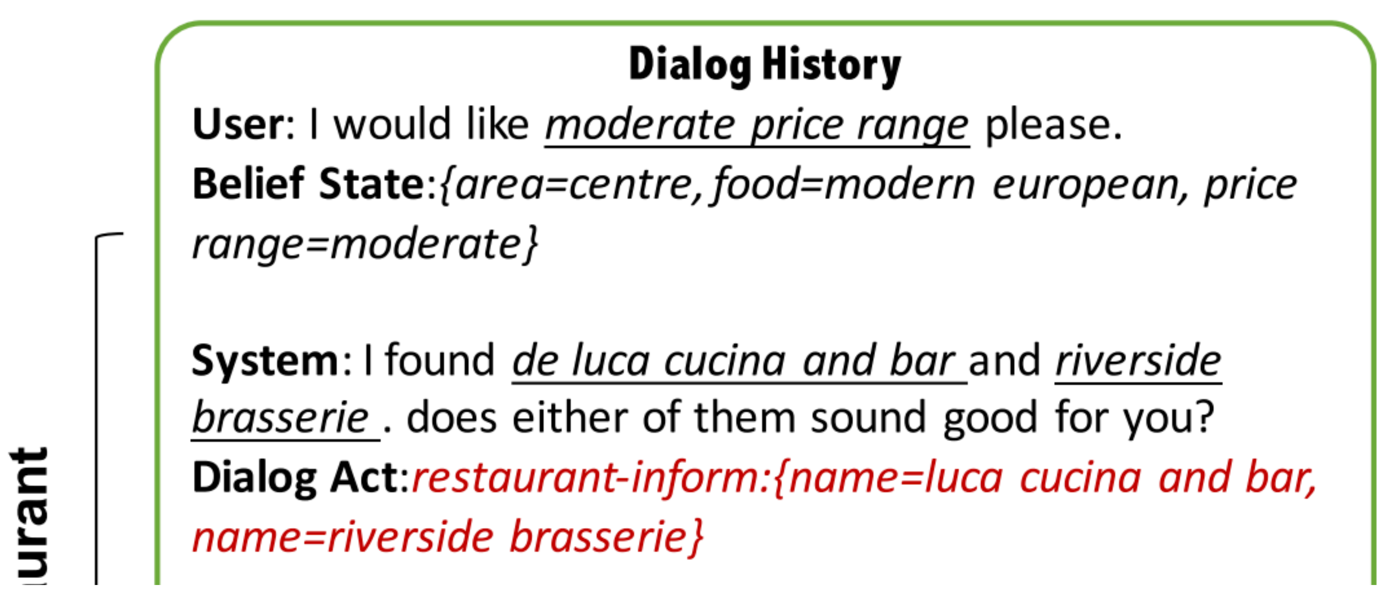




Chen, W., Chen, J., Qin, P., Yan, X., & Wang, W. Y. (2019). Semantically Conditioned Dialog Response Generation via Hierarchical Disentangled Self-Attention.

DSTC8 Track1 Task1(2019): End-to-end Multi-Domain Task-Completion Task

- Goal
  - Build an E2E multi-domain dialogue system for tourist information desk
- MultiWOZ dataset
  - Consist of single and multi-domain dialogues
    - 7 domains, 10k annotated dialog, 8 ~ 15 dialog turns
  - Provide annotations at each turn such as
    - belief state, system dialog act, <u>*user dialog act*</u> (*)



*(*) **User Act**: inform-restaurant*

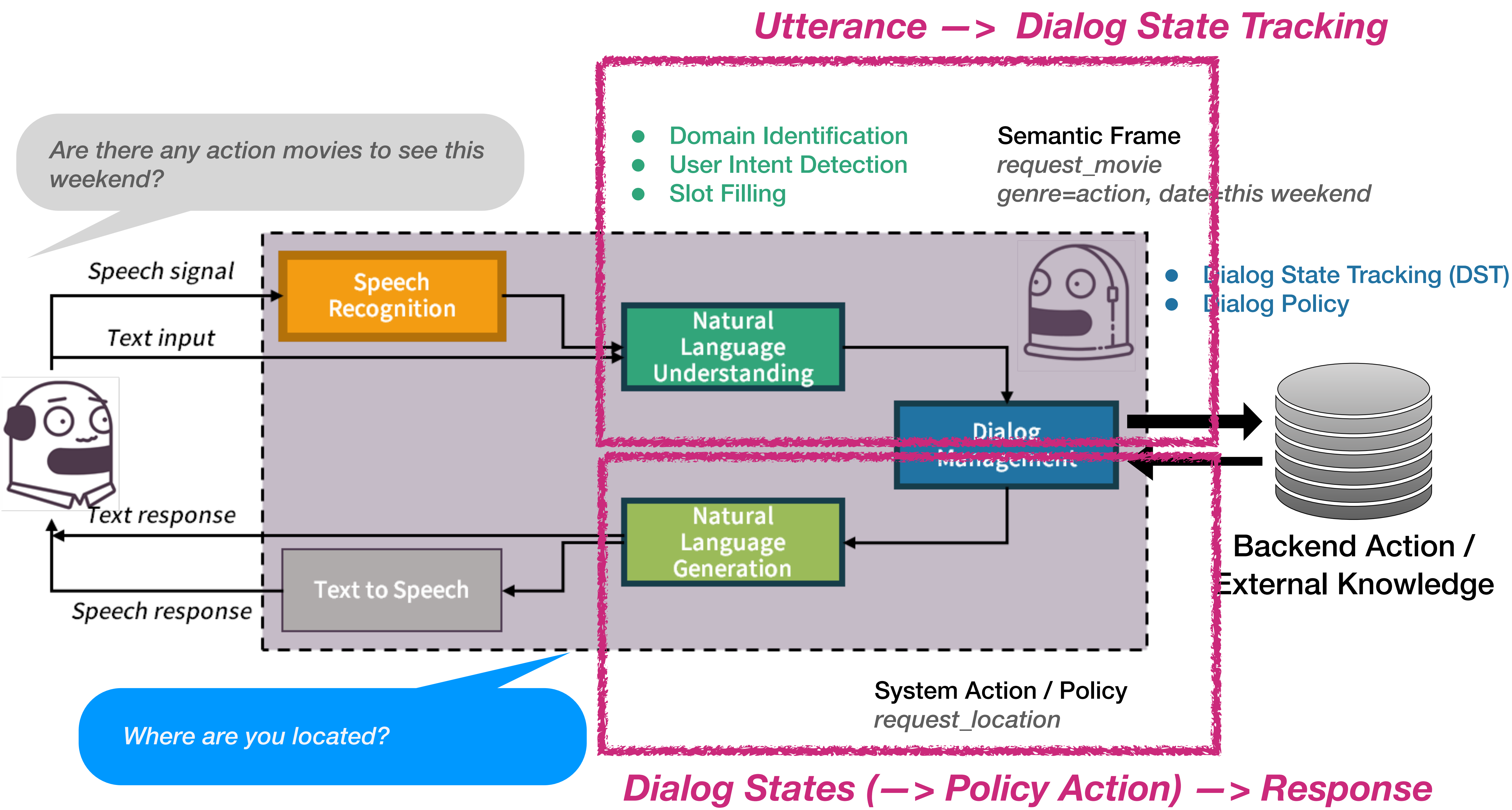
- **Dialog acts: {$domain-$act : [[$slot:$value]]}**
- **Dialog state: $domain:{$slot:$value}**

Table 7: An example dialog for the multi-domain dialog task

Speaker	Utterance	Annotation
User	I ’m looking for a italian restaurant centre.	**Dialog acts**: { “Restaurant-Inform”: [[“Food”,“italian”],[“Area”,“centre”]]}
System	There are 9 restaurants that fit your request. What is your pricerange?	**State**: “restaurant”: {“food”: “italian”,“pricerange”: “not mentioned”, “name”: “not mentioned”, “area”: “centre”} **Dialog acts**: { “Restaurant-Request”: [[“Price”,“?”]], “Restaurant-Inform”: [[“Choice”,“9”]]}
User	Price doesn’t really matter, but can you tell me the address, postcode, and phone for your favorite one?	**Dialog acts**: {“Restaurant-Request”: [[“Addr”,“?”], [“Post”,“?”], [“Phone”, “?”]]}
System	My favorite is zizzi cambridge, located at 47 - 53 Regent Street, cb21ab. Their phone number is 01223365599. Would you like to reserve a table?	**State**: “restaurant”: {“food”: “italian”,“pricerange”: “not mentioned”,“name”: “not mentioned”,“area”: “centre”} **Dialog acts**: { “Booking-Inform”: [[“none”,“none”]], “Restaurant-Recommend”: [[“Name”,“zizzi cambridge”],[“Addr”,“47 - 53 Regent Street”], [“Post”,“cb21ab”]], “Restaurant-Inform”: [[“Phone”,“01223365599”]] }
User	No thank you, but I am looking for a college to visit.	**Dialog acts**: {“Attraction-Inform”: [[“Type”, “college”]]}
System	Any particular area?	**State**: “restaurant”: {“food”: “italian”,“pricerange”: “not mentioned”,“name”: “not mentioned”,“area”: “centre”} “attraction”: {“type”: “college”, “name”: “not mentioned”,“area”: “not mentioned” } **Dialog acts**: { “Attraction-Request”: [[“Area”,“?”]]}
User	No, it doesn’t matter.	**Dialog acts**: {}
System	what about clare college? it’s in the west.	**State**: “restaurant”: {“food”: “italian”,“pricerange”: “not mentioned”,“name”: “not mentioned”,“area”: “centre”} “attraction”: { “type”: “college”, “name”: “not mentioned”, “area”: “not mentioned”} **Dialog acts**: { “Attraction-Inform”: [[“Area”,“west”],[“Name”,“clare college”]]}

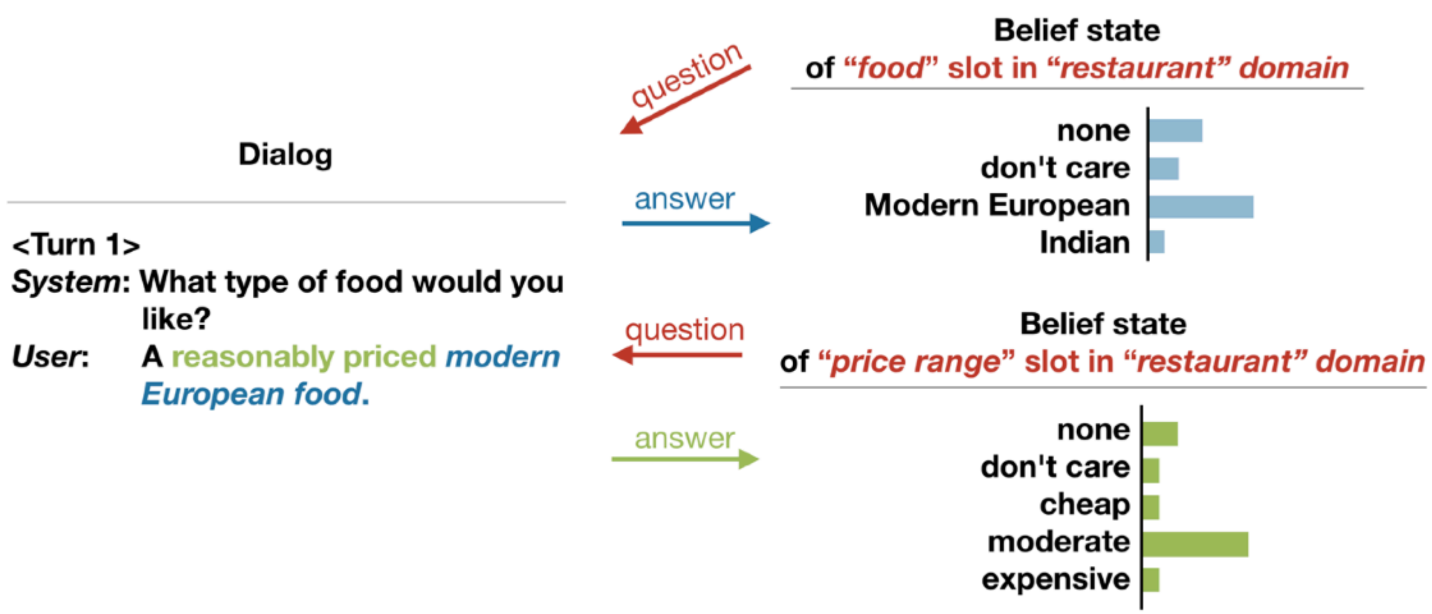
Jinchao Li et al., Results of the Multi-Domain Task-Completion Dialog Challenge, AAAI, 2020

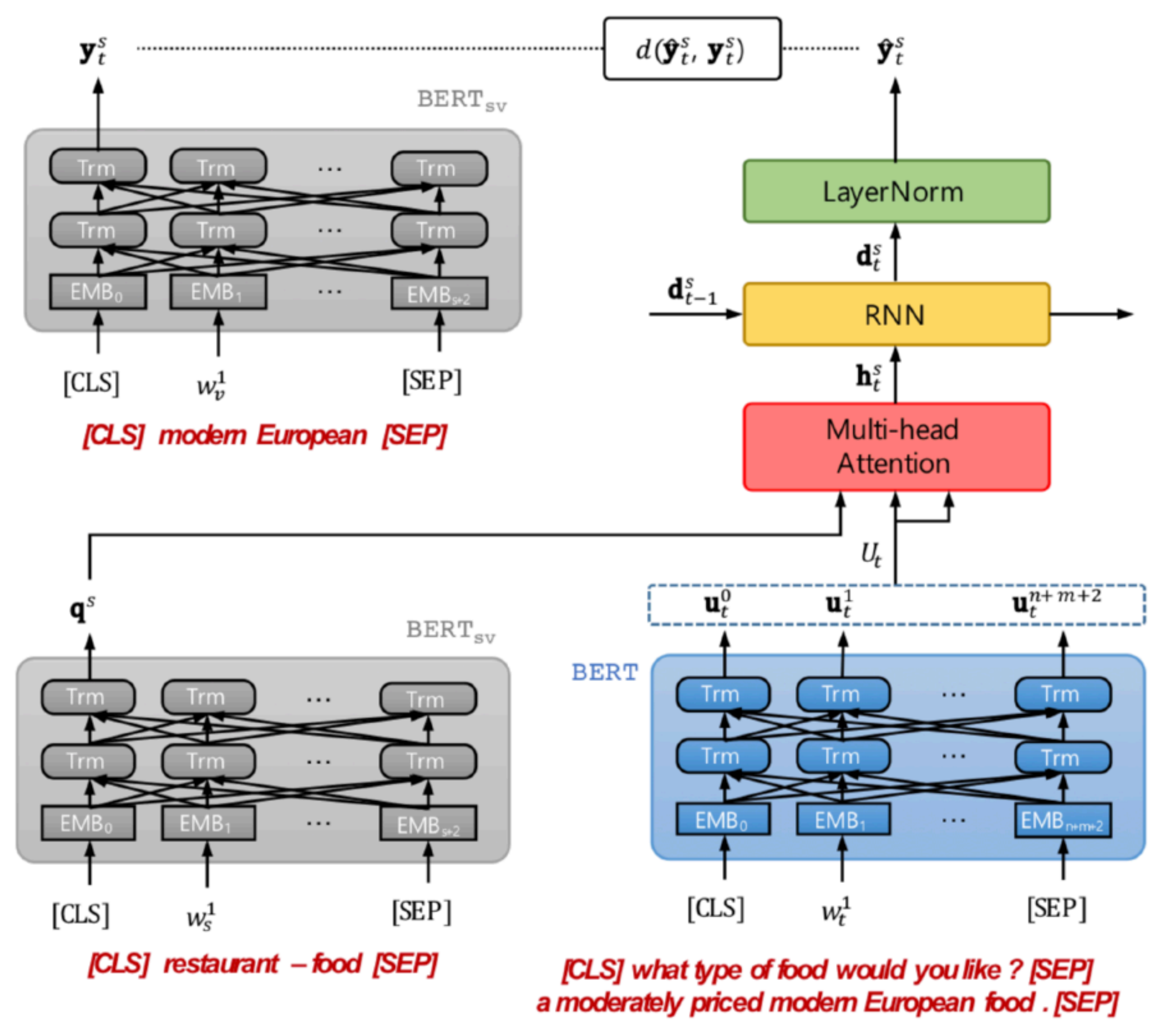
Toward End-to-End Multi-Domain Goal-oriented Dialogue systems



SUMBT: Slot-Utterance Matching Belief Tracker

- **Goal: Build domain *independent* belief tracker for scalability**
- **Key Idea: Find the slot-value of a domain-slot type from user and system's utterances using attention mechanism like question-answering problems**





$$p\left(v_t|\mathbf{x}^{sys}_{\leq t}, \mathbf{x}^{usr}_{\leq t}, s\right) = \frac{\exp\left(-d(\hat{\mathbf{y}}^s_t, \mathbf{y}^v_t)\right)}{\sum_{v' \in \mathcal{C}_s} \exp\left(-d(\hat{\mathbf{y}}^s_t, \mathbf{y}^{v'}_t)\right)},$$

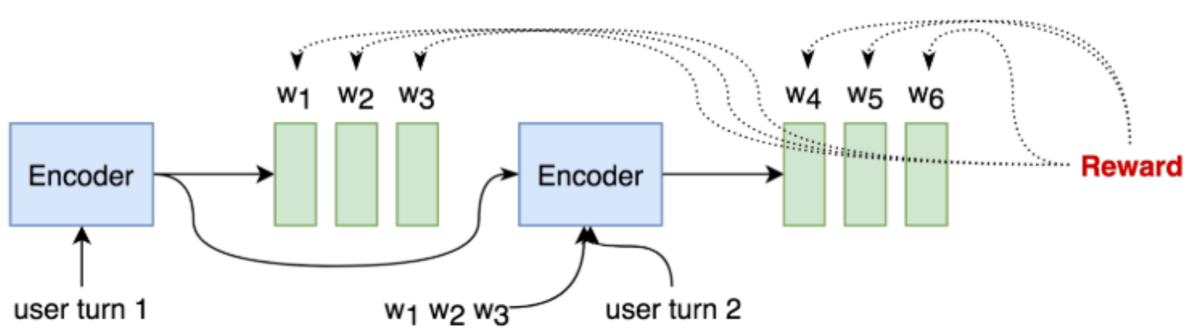
$$\mathcal{L}(\theta) = -\sum_{s \in \mathcal{D}} \sum_{t=1}^{T} \log p(v_t|\mathbf{x}^{sys}_{\leq t}, \mathbf{x}^{usr}_{\leq t}, s).$$

Lee H, Lee J, Kim TY. SUMBT: Slot-Utterance Matching for Universal and Scalable Belief Tracking. ACL, 2019.

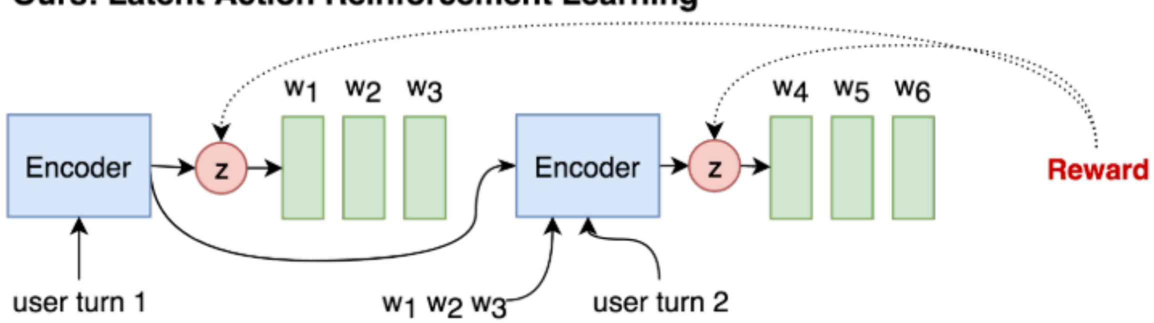
LaRL: Latent Action Reinforcement Learning

- **Problems:**
  - **Simple hand-crafted system action space**
  - **Word-level RL suffers from credit assignment**
- **Key Idea: Latent action spaces, decoupling the discourse-level decision-making from natural language generation**

Baseline: Word-level Reinforcement Learning



Ours: Latent Action Reinforcement Learning



Policy gradient (REINFORCE)

$$\nabla_\theta J(\theta) = \mathbb{E}_\theta[\sum_{t=0}^{T}\sum_{j=0}^{U_t} R_{tj} \nabla_\theta \log p_\theta(w_{tj}|w_{<tj}, \mathbf{c}_t)]$$

$$\nabla_\theta J(\theta) = \mathbb{E}_\theta[\sum_{t=0}^{T} R_t \log p_\theta(\mathbf{z}|\mathbf{c}_t)]$$

- **Categorical Latent Actions**
  - M independent K-way categorical random variables

$$\mathbf{h} = \mathcal{F}(\mathbf{c})$$
$$p(Z_m|\mathbf{c}) = \text{softmax}(\pi_m(\mathbf{h}))$$
$$p(\mathbf{x}|\mathbf{z}) = p_{\theta_d}(\underline{\mathbf{E}_{1:M}(\mathbf{z}_{1:M})})$$

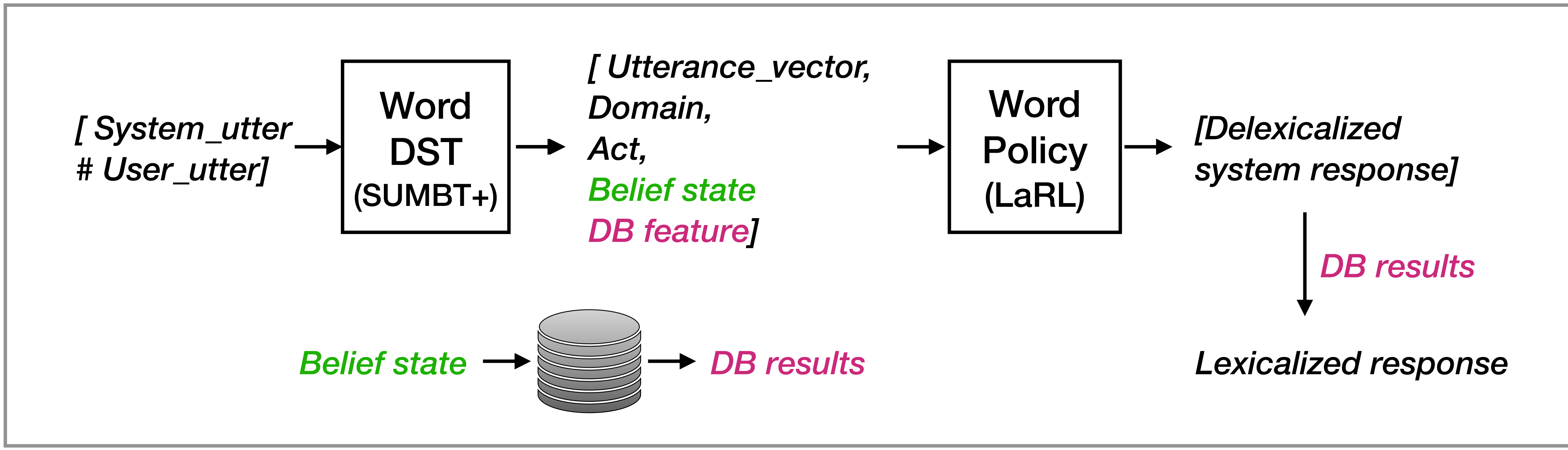
$$\mathbf{z}_m \sim p(Z_m|\mathbf{c})$$
$$p_\theta(\mathbf{z}|\mathbf{c}) = \prod_{m=1}^{M} p(Z_m = \mathbf{z}_m|\mathbf{c})$$

Zhao, T., Xie, K., & Eskenazi, M. Rethinking Action Spaces for Reinforcement Learning in End-to-end Dialog Agents with Latent Variable Models. NAACL-HLT 2019

End-to-end system incorporating SUMBT and LaRL

[It is in the east , and moderately priced . Would you like to book a room ? # Can I get the address and phone number , please ?]

"the address is [hotel_address] , postcode [hotel_postcode] . the phone number is [hotel_phone] . anything else ?"



Inferred hotel domain belief state

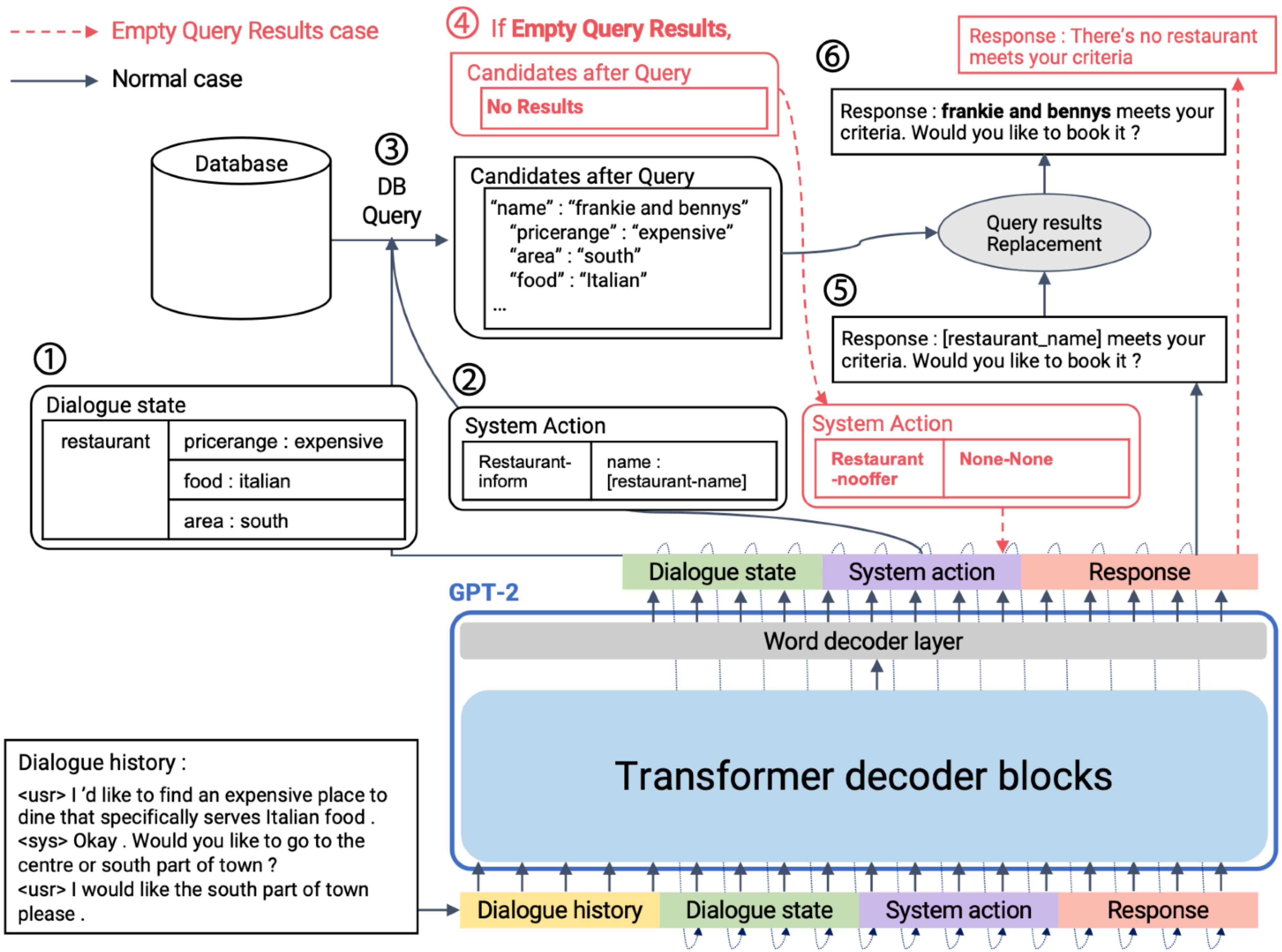

```
"hotel": {
  "book": {
    "booked": [],
    "stay": "",
    "day": "",
    "people": ""
  },
  "semi": {
    "name": "a and b guest house",
    "area": "not mentioned",
    "parking": "not mentioned",
    "pricerange": "not mentioned",
    "stars": "not mentioned",
    "internet": "not mentioned",
    "type": "not mentioned"
  }
},
```


Hotel domain query result


```
{
    "address": "124 tenison road",
    "area": "east",
    "internet": "yes",
    "parking": "no",
    "id": "0",
    "location": [
        52.1963733,
        0.1987426
    ],
    "name": "a and b guest house",
    "phone": "01223315702",
    "postcode": "cb12dp",
    "price": {
        "double": "70",
        "family": "90",
        "single": "50"
    },
    "pricerange": "moderate",
    "stars": "4",
    "takesbookings": "yes",
    "type": "guesthouse"
},
```


"The address is 124 tenison road , postcode cb12dp . The phone number is 01223315702 . Anything else ?"

E2E Neural Pipeline using GPT-2



Ham, D., et. al., End-to-end neural pipeline for goal-oriented dialogue systems using GPT-2., ACL 2020

Note! DSTC9 has just started! :)

The Challenge Evaluation Results

Table 1: Automatic evaluation results. The results are from the best submissions from each group.

Team	SR%	Rwrd	Turns	P	R	F1	BR%
1	**88.80**	61.56	7.00	**0.92**	**0.96**	**0.93**	**93.75**
2	88.60	61.63	6.69	0.83	0.94	0.87	96.39
3	82.20	54.09	6.55	0.71	0.92	0.78	94.56
4	80.60	51.51	7.21	0.78	0.89	0.81	86.45
5	79.40	49.69	7.59	0.80	0.89	0.83	87.02
6	58.00	23.70	7.90	0.61	0.73	0.64	75.71
7	56.60	20.14	9.78	0.68	0.77	0.70	58.63
8	55.20	17.18	11.06	0.73	0.74	0.71	71.87
9	54.00	17.15	9.65	0.66	0.76	0.69	72.42
10	52.20	15.81	8.83	0.46	0.75	0.54	76.38
11	34.80	−6.39	10.15	0.65	0.75	0.68	N/A
BS	63.40	30.41	7.67	0.72	0.83	0.75	86.37

Abbreviations: BS: Baseline, SR: Success Rate, Rwrd: Reward, P/R: precision/recall of slots prediction, BR: Book Rate.

Table 2: Human evaluation results. The results are from the best submissions from each group.

Team	SR%	Under.	Appr.	Turns	Final Ranking
5	**68.32**	**4.15**	**4.29**	**19.51**	**1**
1	65.81	3.54	3.63	15.48	2
2	65.09	3.54	3.84	13.88	3
3	64.10	3.55	3.83	16.91	4
4	62.91	3.74	3.82	14.97	5
10	54.90	3.78	3.82	14.11	6
6	43.56	3.55	3.45	21.82	7
11	36.45	2.94	3.10	21.13	8
7	25.77	2.07	2.26	16.80	9
8	23.30	2.61	2.65	15.33	10
9	18.81	1.99	2.06	16.11	11
Baseline	56.45	3.10	3.56	17.54	N/A

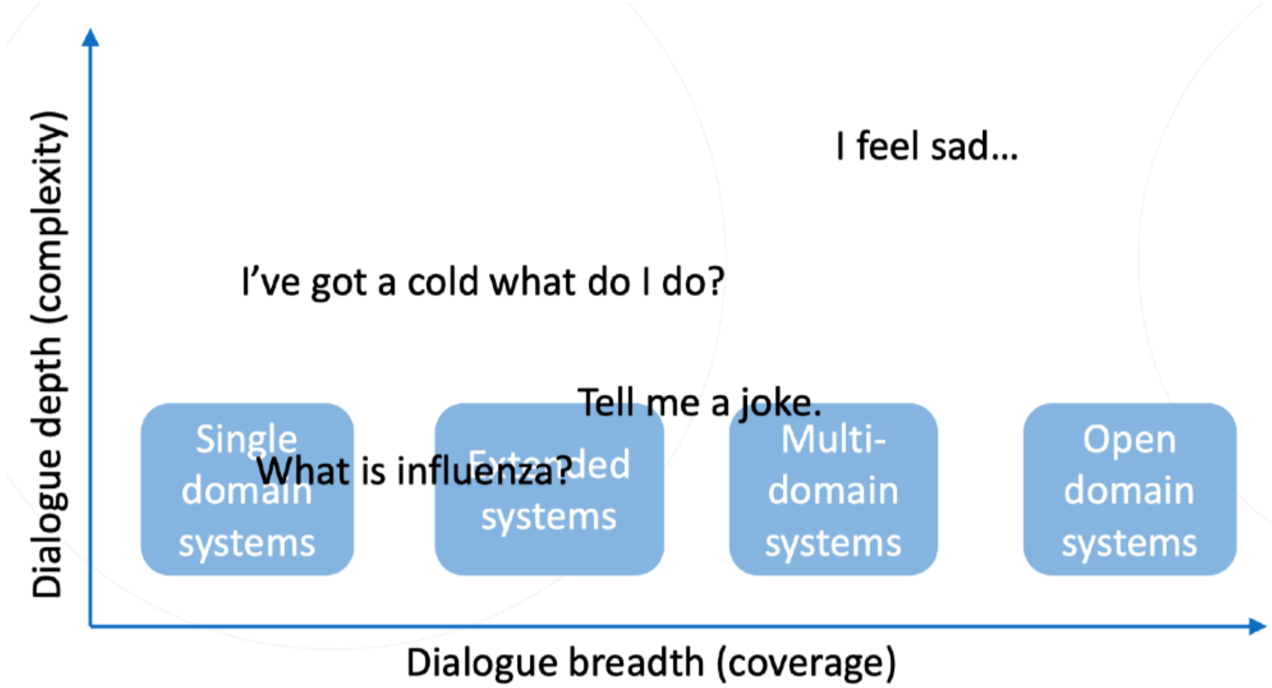
Abbreviations: Under.: understanding score, Appr.: appropriateness score, SR: success rate.

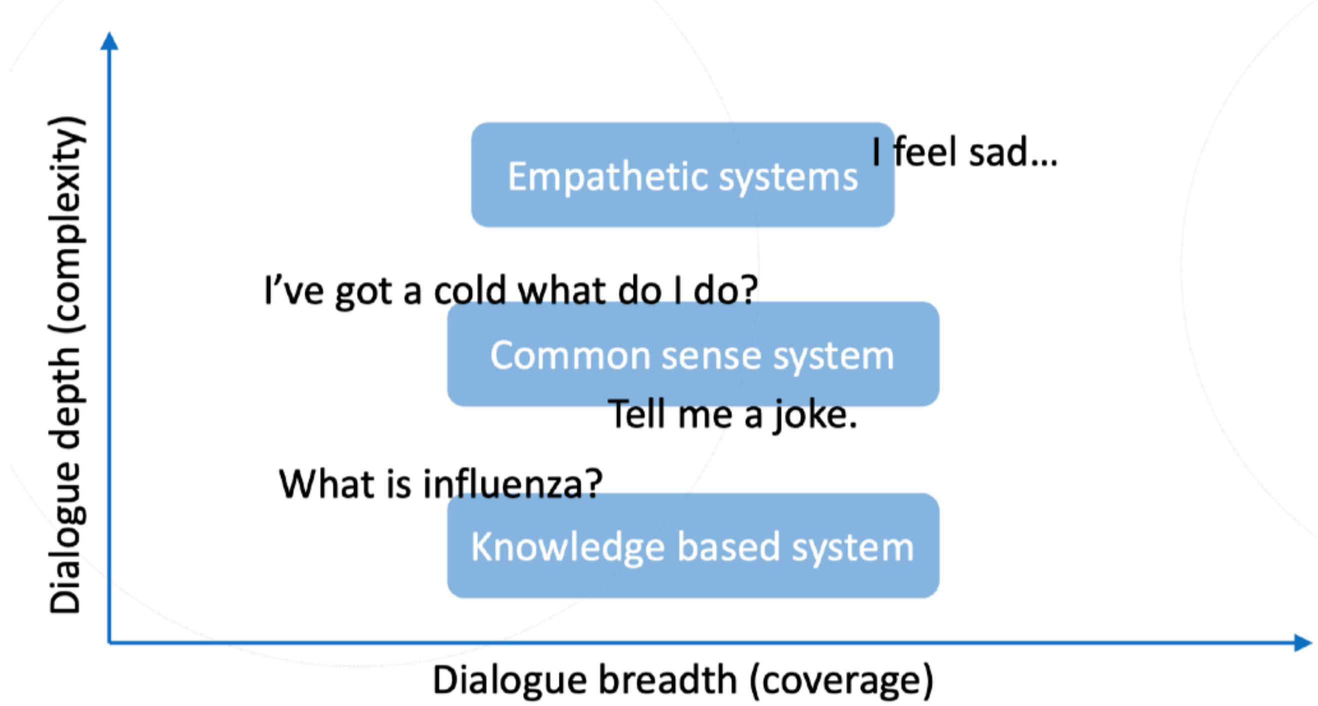
- Note: almost participants' models are based on sophisticated rules

	NLU	DST	Policy	NLG
T1, T2, T4	BERT-based	Rule-based	Rule-based **(+)**	Template **(+)**
T3	BERT-based	Rule-based	DQN	HDSA + Template
T5	**End-to-end neural model using GPT-2**			
T6, T7, T8, T9	OneNet/MILU	Rule-based	★	Template **(+)** / Neural-based
T10 (ours)	SUMBT		LaRL *(without system action supervision)*	

(+) denotes addition of hand-crafted rule, ★ denotes various methods

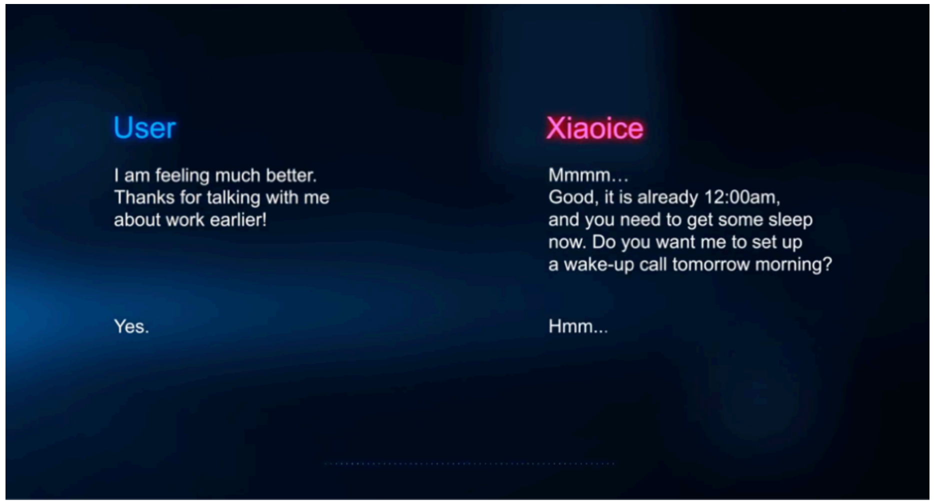
Evolution Roadmap



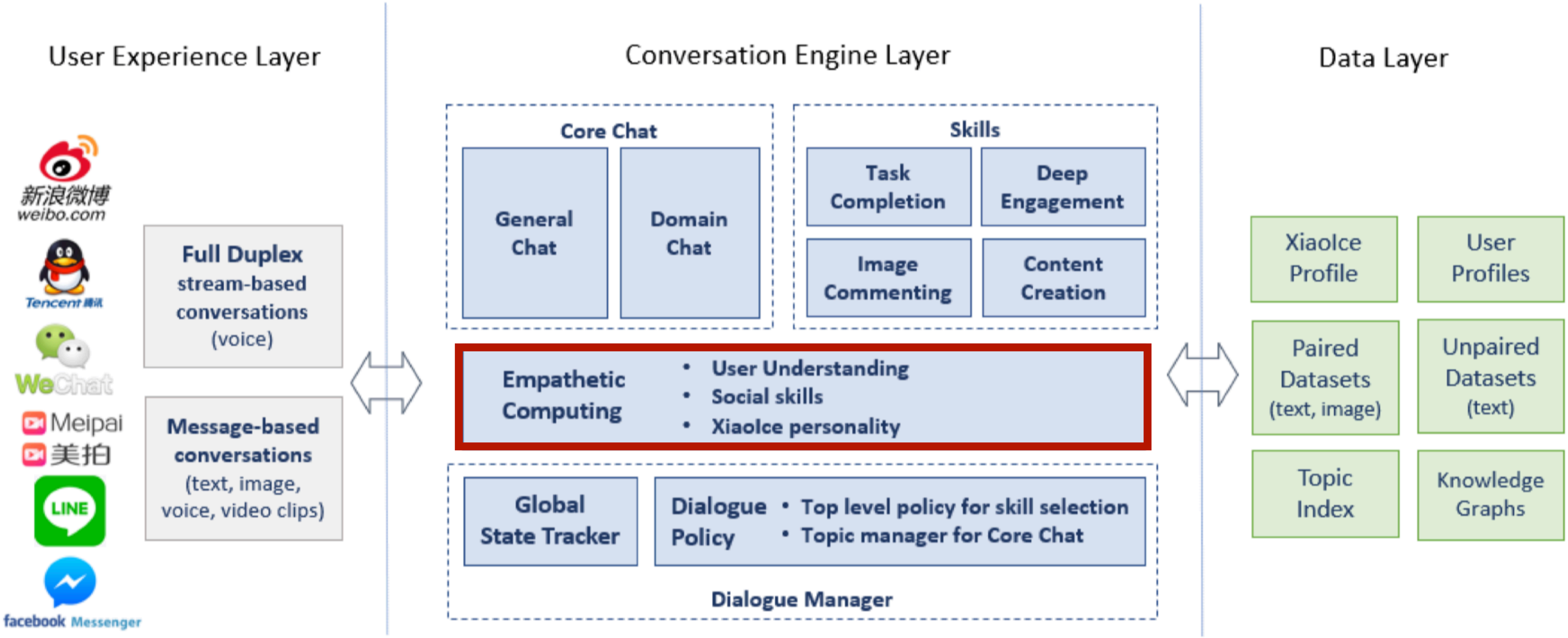


Material: https://deepdialogue.miulab.tw

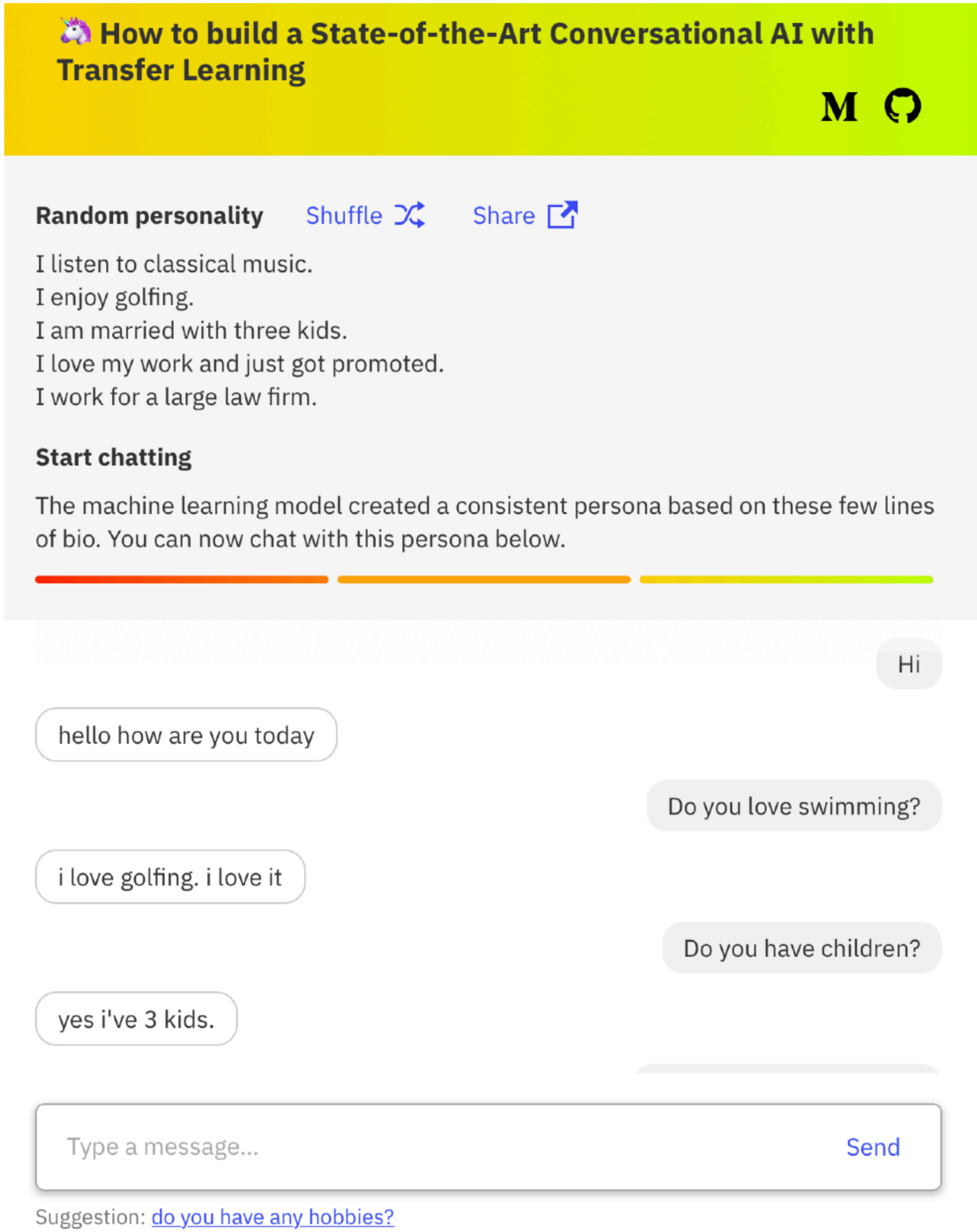
XiaoIce
System Architecture



Microsoft, Xiaoice (2018)



Dialogue System with Personality



https://convai.huggingface.co

Summary

I. Introduction to dialog systems
   - Brief history, components and categories of dialogue systems

II. Deep learning for Natural Language
   - Word embedding: Skip-gram, CBOW
   - Language models: RNN, BERT, GPT …

III. Toward end-to-end neural dialog systems for multi-domain task completion
   - E2E Multi-domain Goal-oriented Dialog System
   - Future direction
     - Empathic, Personality, Open domain, Common sense …

Thank you