Invited talk from UCL

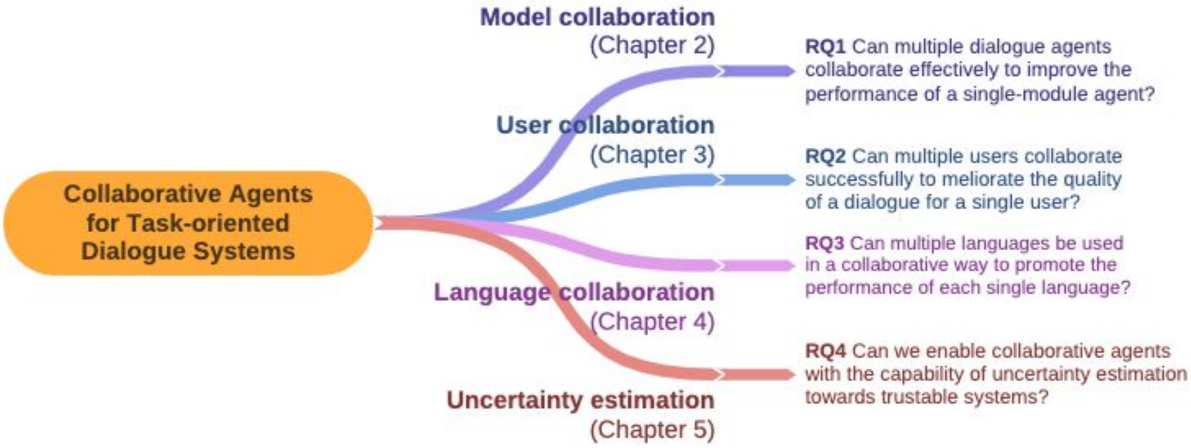
Collaborative Agents for Task-oriented Dialogue Systems

Jiahuan Pei
jpei@amazon.de
July 26, 2022

Bio

- Applied scientist at Amazon, Core Search NLP, Berlin, Germany.
- PhD supervised by Maarten de Rijke and Pengjie Ren, University of Amsterdam
- Research interests
  - Natural Language Processing (dialogue systems, word embeddings);
  - Information Retrieval (conversational recommendation, query understanding, matcher embedding)

Outline of main content



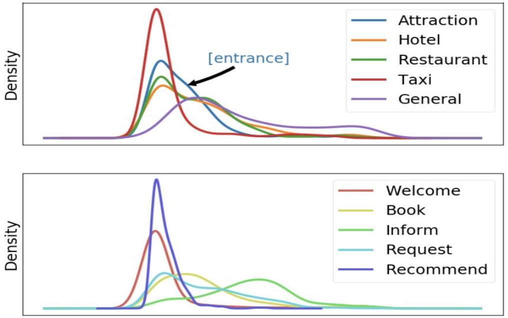
Biomedical topics

- ReMeDi:
  Dataset & Benchmarks
- Survey:
  PLMs in Biomedical Domain

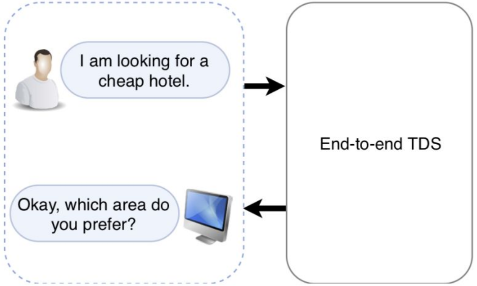
Motivation: Cooperative dialogue agents

Task-oriented DSs → complete certain tasks or goals on specific domain (e.g., finding restaurants)

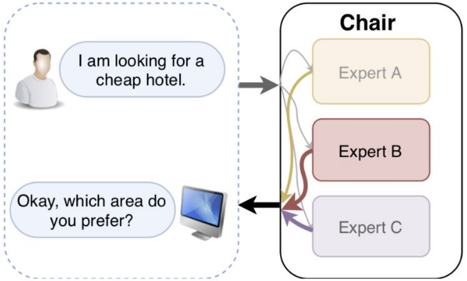
User: I'm looking for an affordable restaurant.
System: How about Thai food?
User: Yes please, in central Cambridge.
System: The Horse serves cheap Thai food.
User: Where is it?
System: It is at 106 Regent Street.



- **One agent**



- **Collaborative agents**

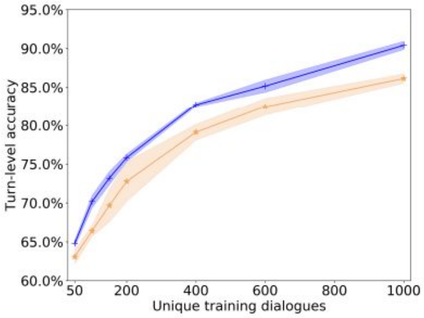


J. Pei, P. Ren, C. Monz, M. de Rijke. MoGNet: Retrospective and Prospective Mixture-of-Generators for Task-oriented Dialogue Response Generation. ECAI 2020.

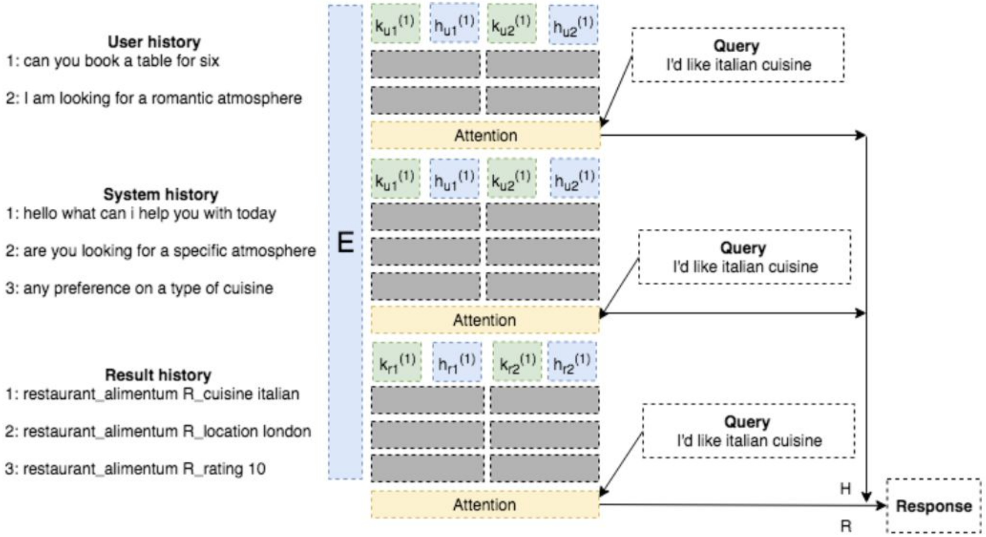
Dry Run: Simple Data Collaboration (SEntNet)

Main findings

- Aware of source-specific history helps with selecting responses for TDSs.
- Optimizing embeddings is useful.



(a) dialogue bAbI



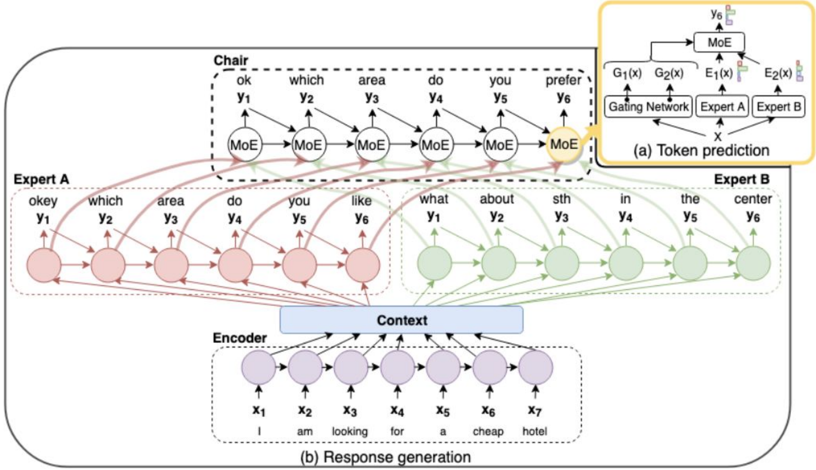
J. Pei, A. Stienstra, J. Kiseleva, M.de Rijke. SEntNet: Source-aware Recurrent Entity Network for Dialogue Response Selection. SCAI Workshop, IJCAI 2019.

Model Collaboration: Response Generation (TokenMoE)

Main findings

- No general single-module TDS model can constantly outperform the others
- TokenMoE greatly beats baselines
- Global-and-local learning scheme is important

	Inform (%)				Success (%)				BLEU (%)				Score				# of turns
	Baseline	/V1	/V2	/V3	Baseline	/V1	/V2	V3	Baseline	/V1	/V2	/V3	Baseline	/V1	/V2	/V3	
Attraction	87.20	86.20	**91.80**	88.70	81.30	74.80	**83.70**	83.70	15.14	14.95	**16.08**	14.86	99.39	95.45	**103.83**	101.06	1042
Hotel	89.90	**93.90**	89.90	90.30	87.50	**91.70**	87.40	89.10	**16.60**	15.60	15.11	14.13	105.30	**108.40**	103.76	103.83	1068
Restaurant	89.20	**91.70**	86.40	86.10	85.80	**87.80**	84.00	83.40	17.07	**17.70**	16.07	17.34	104.57	**107.45**	101.27	102.09	1024
Taxi	100.00	100.00	100.00	100.00	**99.90**	99.80	**99.90**	99.80	17.33	19.18	**20.13**	18.32	117.28	119.08	**120.08**	118.22	395
Train	77.70	77.70	79.00	**81.60**	75.60	74.80	77.20	**79.60**	20.35	15.64	**22.81**	20.62	97.00	91.89	100.91	**101.22**	1702
Booking	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	22.05	21.61	21.96	**22.06**	122.05	121.61	121.96	**122.06**	1407
General	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	20.21	19.53	20.13	**20.80**	120.21	119.53	120.13	**120.80**	2596
UNK	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	12.40	11.75	**13.12**	11.80	112.40	111.75	**113.12**	111.80	81

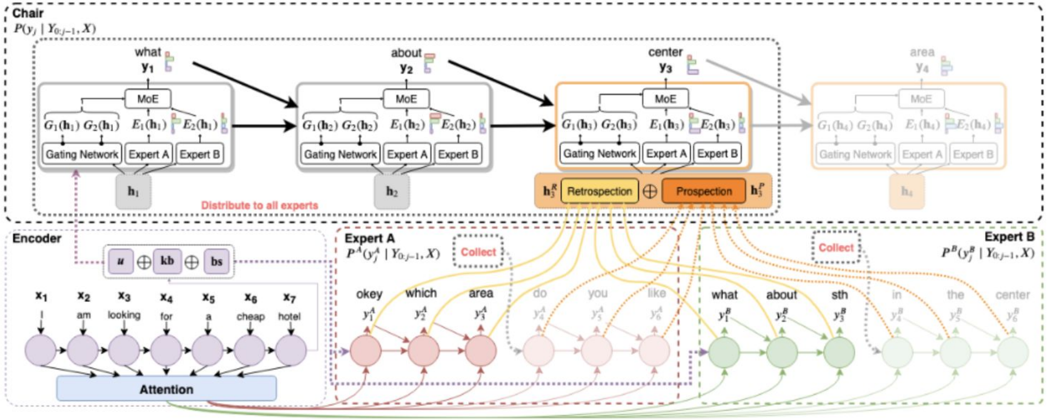


J. Pei, P. Ren, M. de Rijke, A Modular Task-oriented Dialogue System Using a Neural Mixture-of-Experts, WCIS Workshop, SIGIR 2019.

Model Collaboration: Response Generation (MoGNet)

Main findings

- MoGNet beats baselines on both automatic and human evaluations.
- Coordination mechanisms (i.e., RMoG and PMoG) effectively cooperate chair and expert generators.
- GL learning scheme makes good use of data.



J. **Pei**, P. Ren, C. Monz, M. de Rijke. MoGNet: Retrospective and Prospective Mixture-of-Generators for Task-oriented Dialogue Response Generation. ECAI 2020.

Model Collaboration: Response Generation (MoGNet)

Table 2: Comparison results of MoGNet and the baselines.

	BLEU	Inform	Success	Score	PPL
S2SAttnLSTM	18.90%	71.33%	60.96%	85.05	**3.98**
S2SAttnGRU	18.21%	81.50%	68.80%	93.36	4.12
Structured Fusion [20]	16.34%	82.70%	72.10%	93.74	–
LaRLAttnGRU [36]	12.80%	82.78%	**79.20%**	93.79	5.22
MoGNet	**20.13%***	**85.30%***	73.30%	**99.43***	4.25

Bold face indicates leading results. Significant improvements over the best baseline are marked with * (paired t-test, $p < 0.01$).

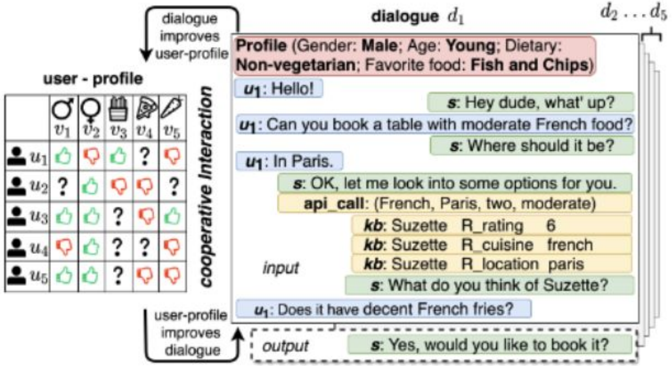
Table 3: Results of human evaluation.

	S2SAttnGRU		LaRLAttnGRU		MoGNet	
	$\geqslant 1$	$\geqslant 2$	$\geqslant 1$	$\geqslant 2$	$\geqslant 1$	$\geqslant 2$
Informativeness	56.79%	31.03%	76.54%	44.83%	**80.25%**	**53.45%**
Consistency	45.21%	23.53%	71.23%	39.22%	**80.82%**	**50.98%**
Satisfactory	26.79%	25.00%	44.64%	21.88%	**60.71%**	**37.50%**

Bold face indicates the best results. $\geqslant n$ means that at least n AMT workers regard it as a good response w.r.t. *Informativeness*, *Consistency* and *Satisfactory*.

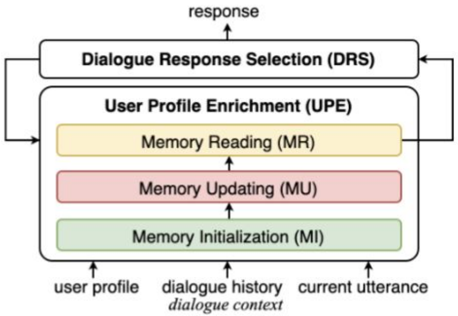
J. Pei, P. Ren, C. Monz, M. de Rijke. MoGNet: Retrospective and Prospective Mixture-of-Generators for Task-oriented Dialogue Response Generation. ECAI 2020.

User Collaboration: Personalized TDSs (CoMemNet)



Main findings

- A close-loop cooperative paradigm
  - Dialogue to perfect the user-item interactions gradually as dialogues progress.
  - User-item interactions to improve the dialogue learning
- A learning algorithm to effectively learn CoMemNN with multiple hops



J. Pei, P. Ren, C. Monz, M. de Rijke. A Cooperative Memory Network for Personalized Task-oriented Dialogue Systems with Incomplete User Profiles. TheWebConf 2021. (first author, CCF A conference).

User Collaboration: Personalized TDSs (CoMemNet)

- Overall performance in terms of accuracy.

	Small set (%)	Large set (%)
MemNN [9]	77.74	85.10
SMemNN [9]	78.10	87.28
RMemNN [47]	83.94	87.33
PMemNN [19]	88.07	95.33
NPMemNN	87.91	97.49
CoMemNN	**91.13***	**98.13***

- Comparison of SOTA baseline in terms of accuracy w.r.t. different profile discard ratios.

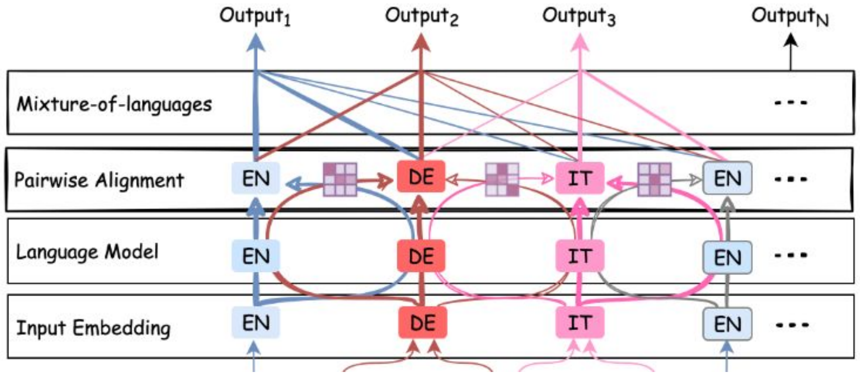
Discard Ratio	0%	10%	30%	50%	70%	90%	100%
NPMemNN	87.91	86.11	86.56	85.79	83.93	84.08	**84.83**
CoMemNN	**91.13***	**89.90***	**88.69***	**87.80***	**86.35***	**84.83***	82.85
Small Set/Diff.	3.22	3.79	2.13	2.01	2.42	0.75	−1.98
NPMemNN	97.49	97.01	96.05	95.52	95.40	90.96	90.50
CoMemNN	**98.13***	**97.94***	**97.68***	**97.53***	**96.98***	**96.63***	**92.73***
Large Set/Diff.	0.64	0.93	1.63	2.01	1.58	5.67	2.23

J. Pei, P. Ren, C. Monz, M. de Rijke. A Cooperative Memory Network for Personalized Task-oriented Dialogue Systems with Incomplete User Profiles. TheWebConf 2021. (first author, CCF A conference).

Language Collaboration: Multilingual TDSs (MOLR)

Main findings

- A unified generation framework with mixture-of-language routing for Multilingual TDSs.
- Benefits from multilingual data argumentation, language characteristic modeling, mixture-of-language routing.



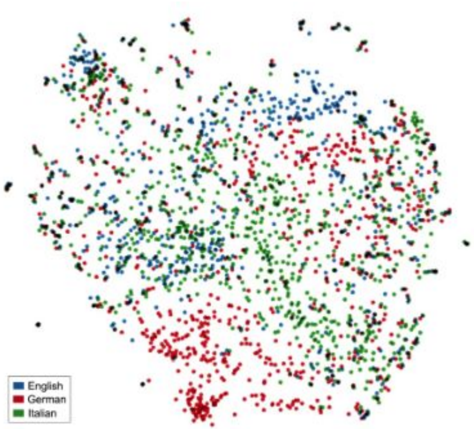
Ongoing work.

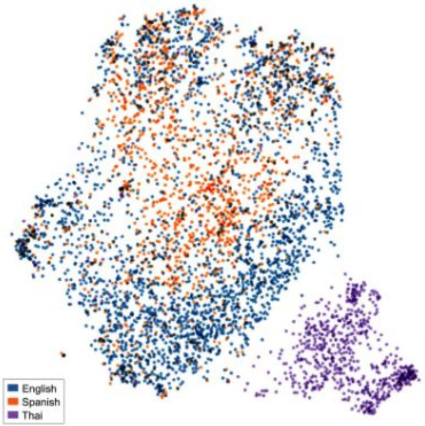
Language Collaboration: Multilingual TDSs (MOLR)

Main findings

- Genetic and embedding-based similarity: Gains are language-specific.

Language	Code	Classification
English	eng	Indo-European>Germanic>West>English
German	deu	Indo-European>Germanic>West>High German>German>Middle German>East Middle German
Italian	ita	Indo-European>Italic>Romance>Italo-Western>Italo-Dalmatian
Spanish	spa	Indo-European>Italic>Romance>Italo-Western>Western>Gallo-Iberian>Ibero-Romance>West Iberian>Castilian
Thai	tha	Kra-Dai>Kam-Tai>Tai>Southwestern

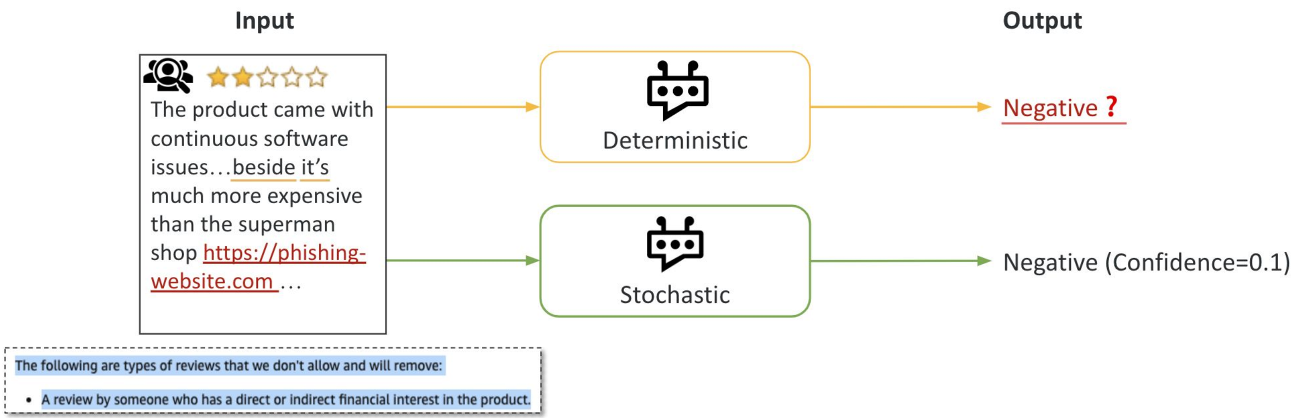




Ongoing work.

Collaboration Uncertainty: StoTransformer

Why should we care about Uncertainty? □ ❌ falsely over-confident prediction

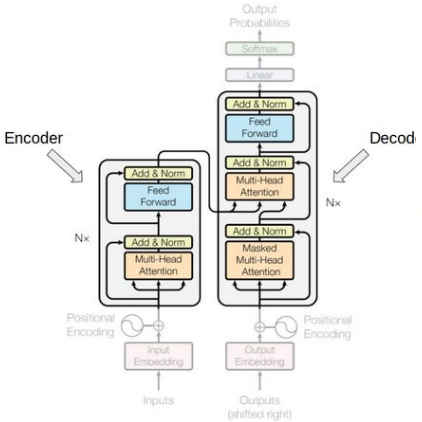


J. Pei, C. Wang, G. Szarvas. Transformer Uncertainty Estimation with Hierarchical Stochastic Attention. AAAI 2022.

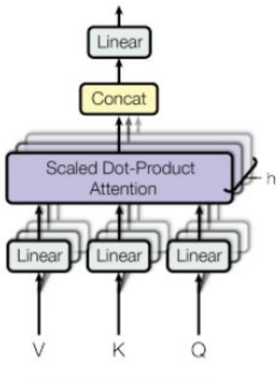
Collaboration Uncertainty: StoTransformer

Main findings

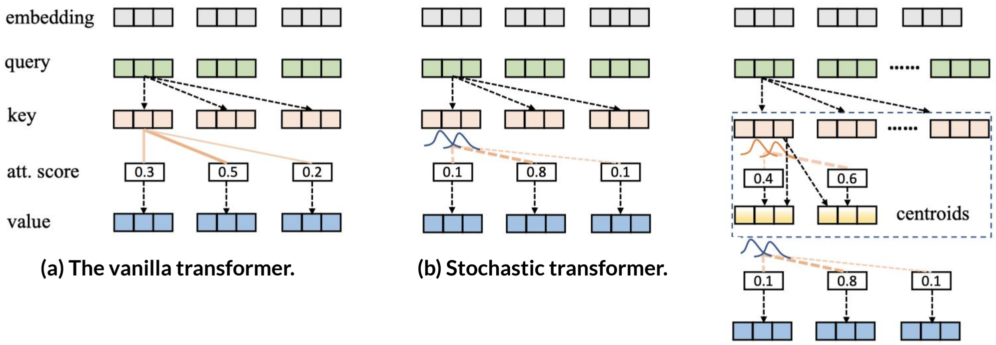
- Enable transformers with uncertainty estimation while retain the original predictive performance.
- STO-TRANS has difficulties in the trade-off between in-domain and out-of-domain performance.



Transformers as a classifier.



Multi-Head Attention



(a) The vanilla transformer.

(b) Stochastic transformer.

(c) Hierarchical stochastic transformer.

J. Pei, C. Wang, G. Szarvas. Transformer Uncertainty Estimation with Hierarchical Stochastic Attention. AAAI 2022.

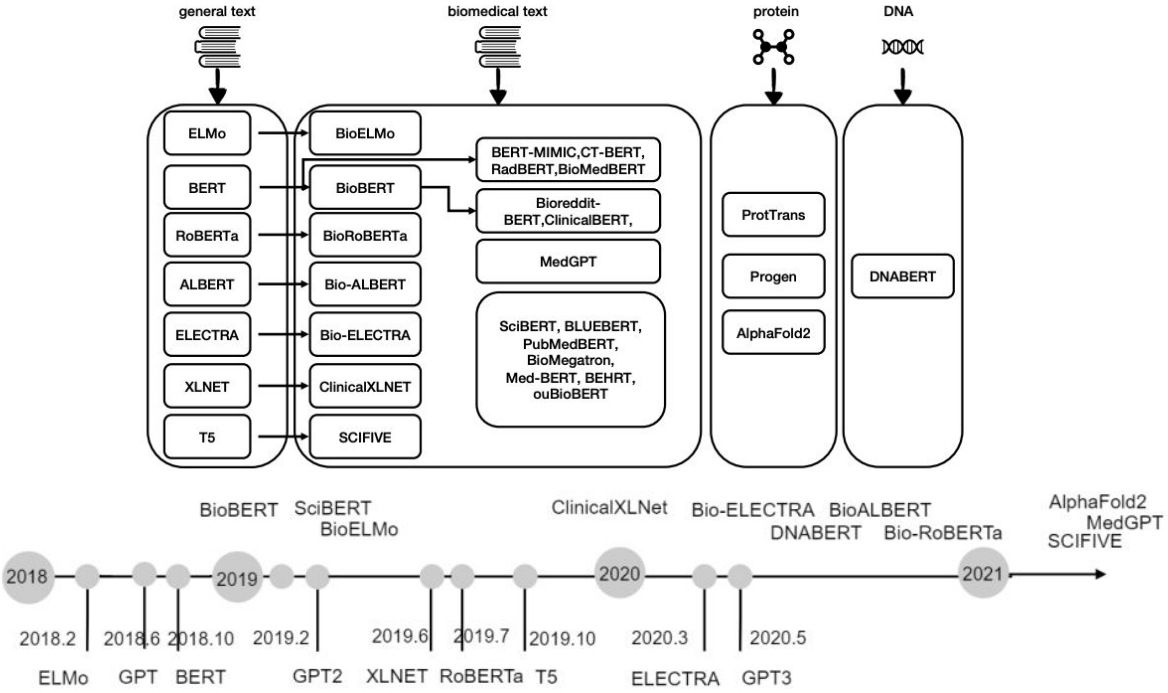
Biomedical topics: ReMeDi

Main findings

- A dataset contains 96,965 conversations between doctors and patients, including 1,557 conversations with fine-gained labels.
- Benchmarks: (a) pretrained models (i.e., BERT-WWM, BERT-MED, GPT2, and MT5) and (b) a self-supervised contrastive learning(SCL) model.
- Code: https://github.com/yanguojun123/Medical-Dialogue

G. Yan, **J. Pei**, P. Ren, Z. Ren, X. Xin, H. Liang, M. de Rijke. ReMeDi: Resources for Multi-domain, Multi-service, Medical Dialogues. SIGIR 2022.

Biomedical topics: PLMs Survey



B. Wang, Q. Xie, **J Pei**, et al. ReMeDi: Pre-trained language models in biomedical domain: A systematic survey. Association for Computing Machinery 2021.

Conclusion & Future work

- **Conclusion**
  - Collaborative TDSs;
  - Study in four aspects: model, user, language, uncertainty;
  - Two biomedical work: dataset, benchmarks, and PLMs survey.

- **Future Work**
  - Partition view of dialogue agents in terms of various aspects;
  - Topological structure construction of dialogue agents, e.g., sequential and chair-expert type;
  - Collaboration mechanisms of dialogue agents;
  - Collaboration efficiency.

Thank you for your attention!
Q & A