数据挖掘课题组
成果介绍

2020年-2021年成果汇总

- 在研项目
  - ✓ 时高精度动态学术画像构建技术，国家重点研发项目（子课题）
  - ✓ 基于社会媒体挖掘的多样化社交关系主动学习，国家自然科学基金面上项目
  - ✓ 面向情感交互的人机对话文本生成技术研究，国家自然科学基金面上项目
- 发表/录用论文
  - **CCF A类**论文5篇 (WWW2020, IJCAI2020, AAAI2021, ACL2021×2)
  - **CCF B类**论文4篇 (DASFAA2020, TMM, DASFAA2021, TNNLS)
  - SCI论文 8篇

1-基于多通道图神经网络的多模态情感分析

- 多模态是指不同形式的数据，如文本、图像、音频、视频等。
- 社交媒体的蓬勃发展，多模态数据的爆炸式增长，如抖音、快手等平台上的短视频数据，朋友圈中的**图文表达**等。
- 多模态数据分析应运而生，如**多模态情感分析**等。

(a) We have a fun day on the beach! (*Positive*)

(b) We have a nice day on a deserted beach. (*Positive*)

- **输入**：文本 + 图像
- **输出**：情绪 (Angry, Bored, Calm, Fear, Happy, Love, Sad) 或情感 (Positive, Neutral, Negative).

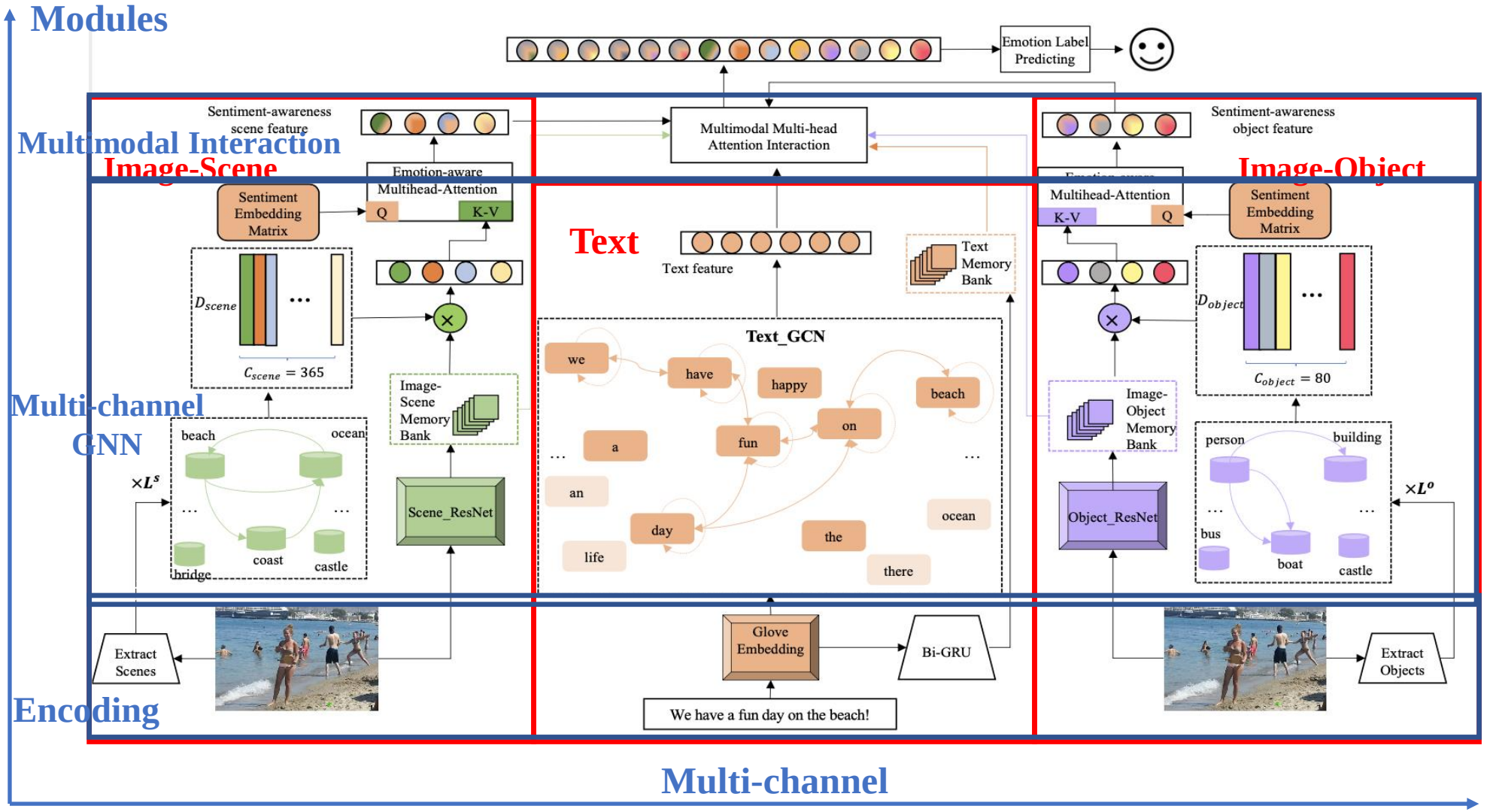
Xiaocui Yang, Shi Feng, Yifei Zhang, Daling Wang. Multimodal Sentiment Detection Based on Multi-channel Graph Neural Networks. ACL 2021. (CCF A)

1-基于多通道图神经网络的多模态情感分析

- MGNNS.
- 主要架构（水平方向：多通道；竖直方向的多模块）

✓水平方向—三个通道的多模态数据（文本通道，图像场景通道，图像物体通道）

✓竖直方向—三个模块处理不同模态数据（编码阶段，多通图神经网络阶段，多模态交互阶段）



Xiaocui Yang, Shi Feng, Yifei Zhang, Daling Wang. Multimodal Sentiment Detection Based on Multi-channel Graph Neural Networks. ACL 2021. (CCF A)

1-基于多通道图神经网络的多模态情感分析

- 数据集：来自Twitter的多模态情感数据集(MVSA-Single, MVSA-Multiple), 和来自Tumblr的多模态情绪数据集(TumEmo).

Modality	Model	MVSA-Single		MVSA-Multiple		TumEmo	
		Acc	F1	Acc	F1	Acc	F1
Text	CNN	0.6819	0.5590	0.6564	0.5766	0.6154	0.4774
	BiLSTM	0.7012	0.6506	0.6790	0.6790	0.6188	0.5126
	BiACNN	0.7036	0.6916	0.6847	0.6319	0.6212	0.5016
	TGNN	0.7034	0.6594	0.6967	0.6180	0.6379	0.6362
Image	OSDA	0.6675	0.6651	0.6662	0.6623	0.4770	0.3438
	SGN	0.6620	0.6248	0.6765	0.5864	0.4353	0.4232
	OGN	0.6659	0.6191	0.6743	0.6010	0.4564	0.4446
	DuIG	0.6822	0.6538	0.6819	0.6081	0.4636	0.4561
Image-Text	HSAN	0.6988	0.6690	0.6796	0.6776	0.6309	0.5398
	MDSN	0.6984	0.6963	0.6886	0.6811	0.6418	0.5692
	Co-Mem	0.7051	0.7001	0.6992	0.6983	0.6426	0.5909
	MVAN‡	0.7298‡	0.7139‡	0.7183‡	**0.7038‡**	0.6553‡	0.6543‡
	MGNNS	**0.7377**	**0.7270**	**0.7249**	0.6934	**0.6672**	**0.6669**

Xiaocui Yang, Shi Feng, Yifei Zhang, Daling Wang. Multimodal Sentiment Detection Based on Multi-channel Graph Neural Networks. ACL 2021. (CCF A)

2-用户生成文本的片段水平的情绪原因提取技术研究

- 情绪原因分析—从用户生成的带有情绪色彩的文本中挖掘用户产生情绪的原因；是当前研究热点

- 现有的情绪原因分析存在问题
  - ✓ 大多数模型基于子句水平—情绪原因不准确、产生歧义
  - ✓ 现存的细粒度的情绪原因分析模型—基于人工提取特征、泛化能力较弱

- 片段水平的情绪原因提取
  - ✓ 提取出精确的情绪原因内容

用户

“有钱了也不能乱花”，以前有老彩民说过一句话，“中大奖带给自己喜悦的同时也将带来不少麻烦”。据了解，梁先生的小孩目前还在读中学。

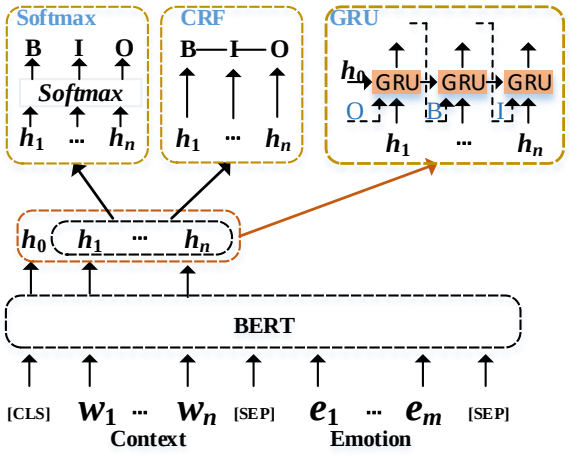
情绪词语 喜悦

提取内容 中大奖

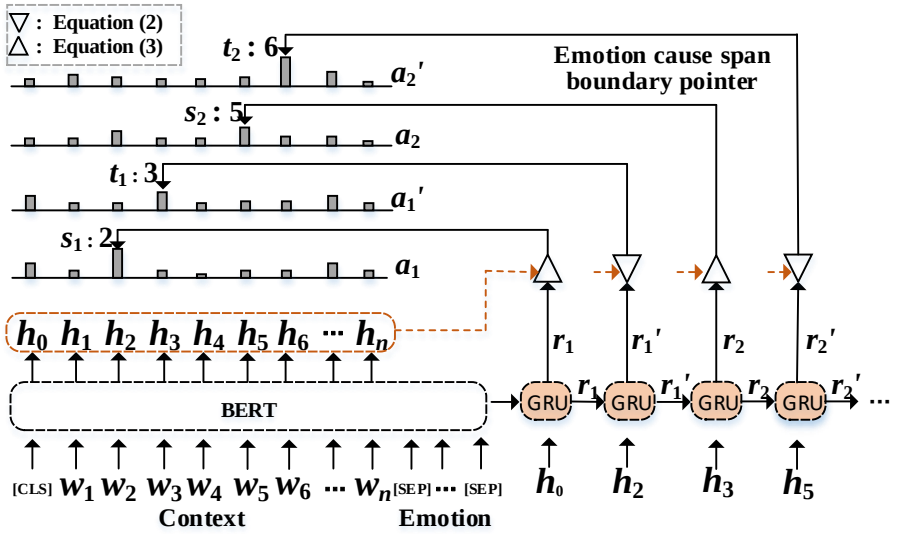
Xiangju Li, Wei Gao, Shi Feng, Yifei Zhang, Daling Wang : Boundary Detection with BERT for Span-level Emotion Cause Analysis. ACL2021(CCF A)

2-用户生成文本的片段水平的情绪原因提取技术研究

- 编码器
  - ✓ BERT（文档-Context + 情绪-Emotion）
- 解码器
  - ✓ 序列标注模型
    - ✓ Softmax
    - ✓ CRF
    - ✓ GRU
  - ✓ 开始/结束点预测模型
    - ✓ Pointer Network



（a）序列标注模型

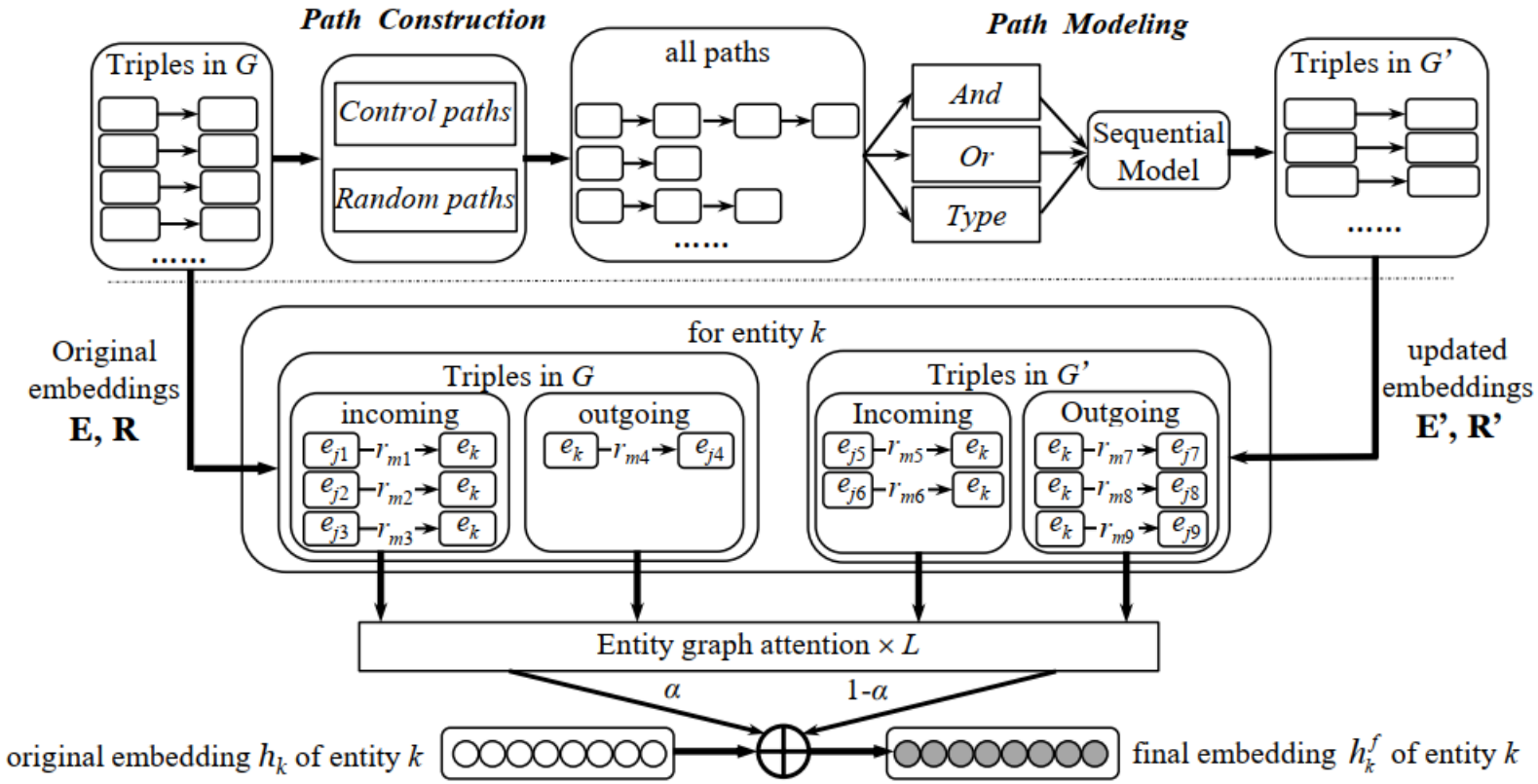


（b）开始/结束点预测模型

Xiangju Li, Wei Gao, Shi Feng, Yifei Zhang, Daling Wang : Boundary Detection with BERT for Span-level Emotion Cause Analysis. ACL2021(CCF A)

3-面向知识图谱表示学习的全局图注意力技术研究

- 知识图谱表示学习是知识图谱领域的核心研究问题
- 现有的知识图谱表示学习存在的问题
  - ✓ 无法捕获全局特征
  - ✓ 无法捕获长距离的依赖关系
- GGAE 模型
  - ✓ 基于控制理论的路径构造 Path Construction
  - ✓ 路径序列建模 Path Modeling
  - ✓ 图注意力网络 Graph Attention



Qian Li , Daling Wang, Shi Feng, Cheng Niu, and Yifei Zhang: Global Graph Attention Embedding Network for Relation Prediction in Knowledge Graphs. IEEE Transactions on Neural Networks and Learning System (CCF B, Q1)

3-面向知识图谱表示学习的全局图注意力技术研究

- 训练数据

Name	Entities	Relations	InDegree	OutDegree	ControlP	RandomP	Training	Validation	Test	Total
FB15K-237	14,541	237	18.4	17.8	6.5	3.0	272,115	17,535	20,466	310,116
WN18RR	40,943	11	2.7	2.2	3.0	2.9	86,835	3034	3134	93,003
Kinship	104	25	82.2	82.2	27.5	5.0	8544	1068	1074	10,686

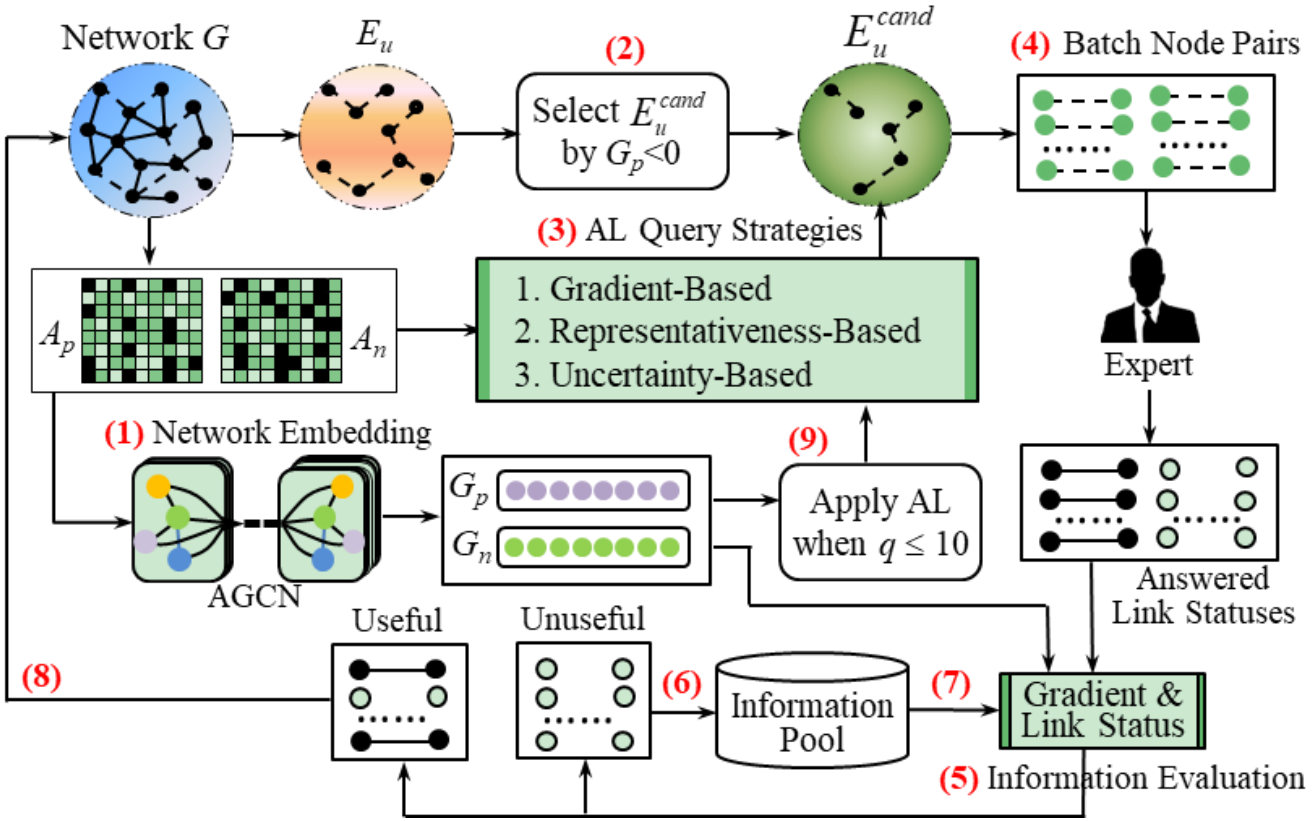
- 结果展示

Model	Kinship				
	Hits@10 ↑	*Hits*@3 ↑	*Hits*@1 ↑	*MRR* ↑	*MR* ↓
TransE	56.3	31.5	11.5	26.4	16.05
DistMult	88.5	61.2	39.9	54.9	4.62
ComplEx	97.0	91.3	74.2	83.3	2.02
ConvE	96.6	89.0	71.7	81.2	2.52
ConvKB	96.3	78.5	48.0	65.2	3.02
PTransE(ADD)	27.3	16.0	7.50	14.1	9.74
PTransE(MUL)	23.9	13.6	5.14	11.5	11.09
PTransE(RNN)	27.3	15.1	6.15	13.1	9.35
RSN	49.2	30.1	15.6	21.9	13.76
OPTransE	58.3	32.8	17.5	30.6	12.87
KBGAT	97.1	91.1	81.8	87.2	2.35
GGAE(*Random*)	97.7	92.4	85.4	89.6	1.97
GGAE(*Control*)	**98.2**	**95.1**	**90.6**	**93.2**	**1.88**

Qian Li , Daling Wang, Shi Feng, Cheng Niu, and Yifei Zhang: Global Graph Attention Embedding Network for Relation Prediction in Knowledge Graphs. IEEE Transactions on Neural Networks and Learning System (CCF B, Q1)

4-基于主动学习和图神经网络的网络表示学习技术研究

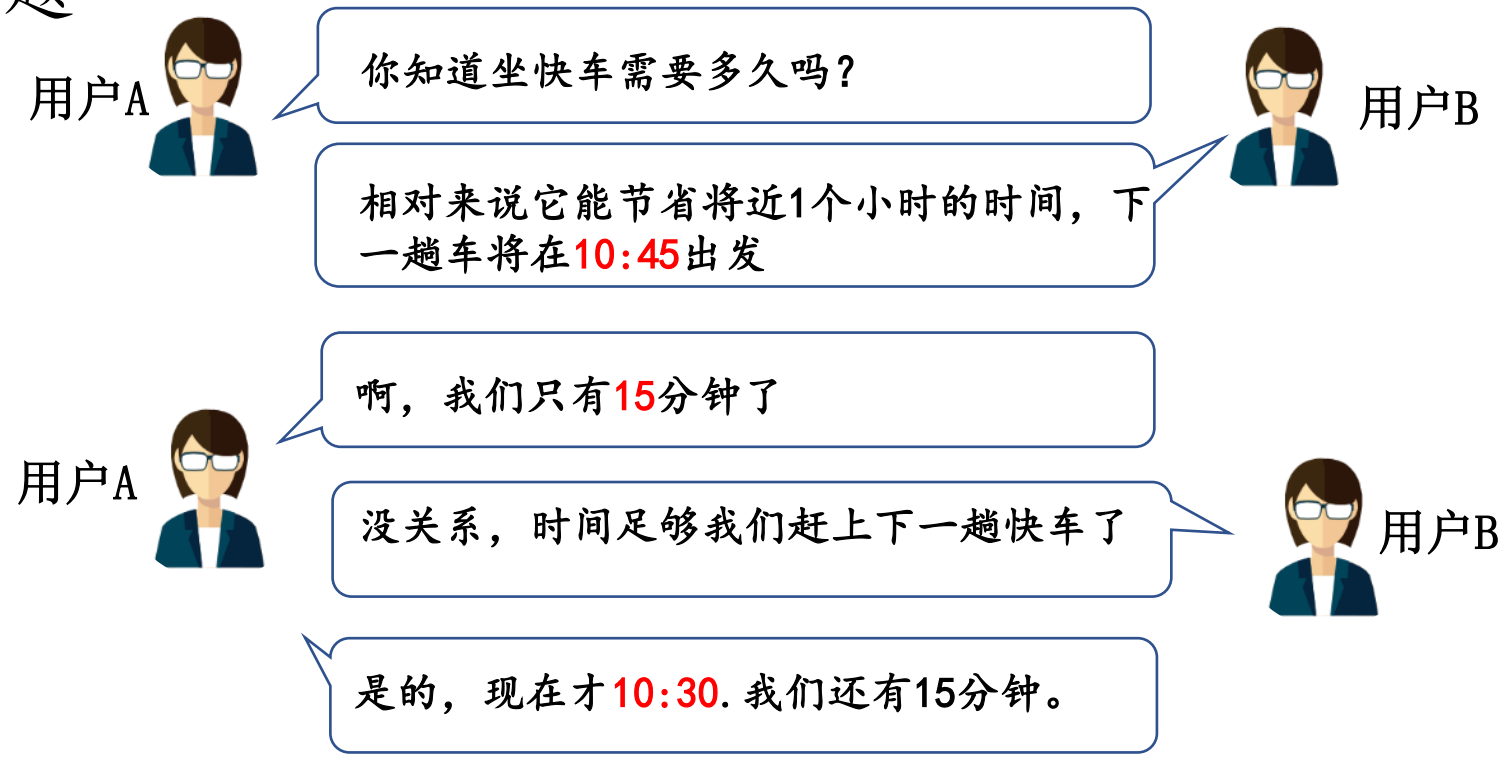
- 主要挑战
  - ✓ 真实世界的网络是稀疏的
  - ✓ 很多节点对的链路状态不能直接获取
  - ✓ 如何通过主动学习获取有价值的节点对的链路状态信息
- 模型解析
  - ✓ 基于主动学习的学习过程 ALNE
  - ✓ 新型图神经网络模型 AGCN
  - ✓ 三种主动学习策略
  - ✓ 信息判别模块



Longcan Wu, Daling Wang, Shi Feng, Kaisong Song, Yifei Zhang, Ge Yu: Which Node Pair and What Status? Asking Expert for Better Network Embedding. DASFAA 2021 (CCF B)

5-基于自定义预训练的多轮回复选择图推理技术研究

- 聊天机器人是人工智能领域的核心研究问题
- 赋予聊天机器人**推理能力**是当前的研究热点
- 现有的多轮回复选择问题
  - ✓ 主要采用特征匹配方式
  - ✓ 上下文和回复的一致特征
  - ✓ 缺乏推理能力
- 推理回复选择问题
  - ✓ 匹配方式失效
  - ✓ 上下文和回复存在逻辑关系

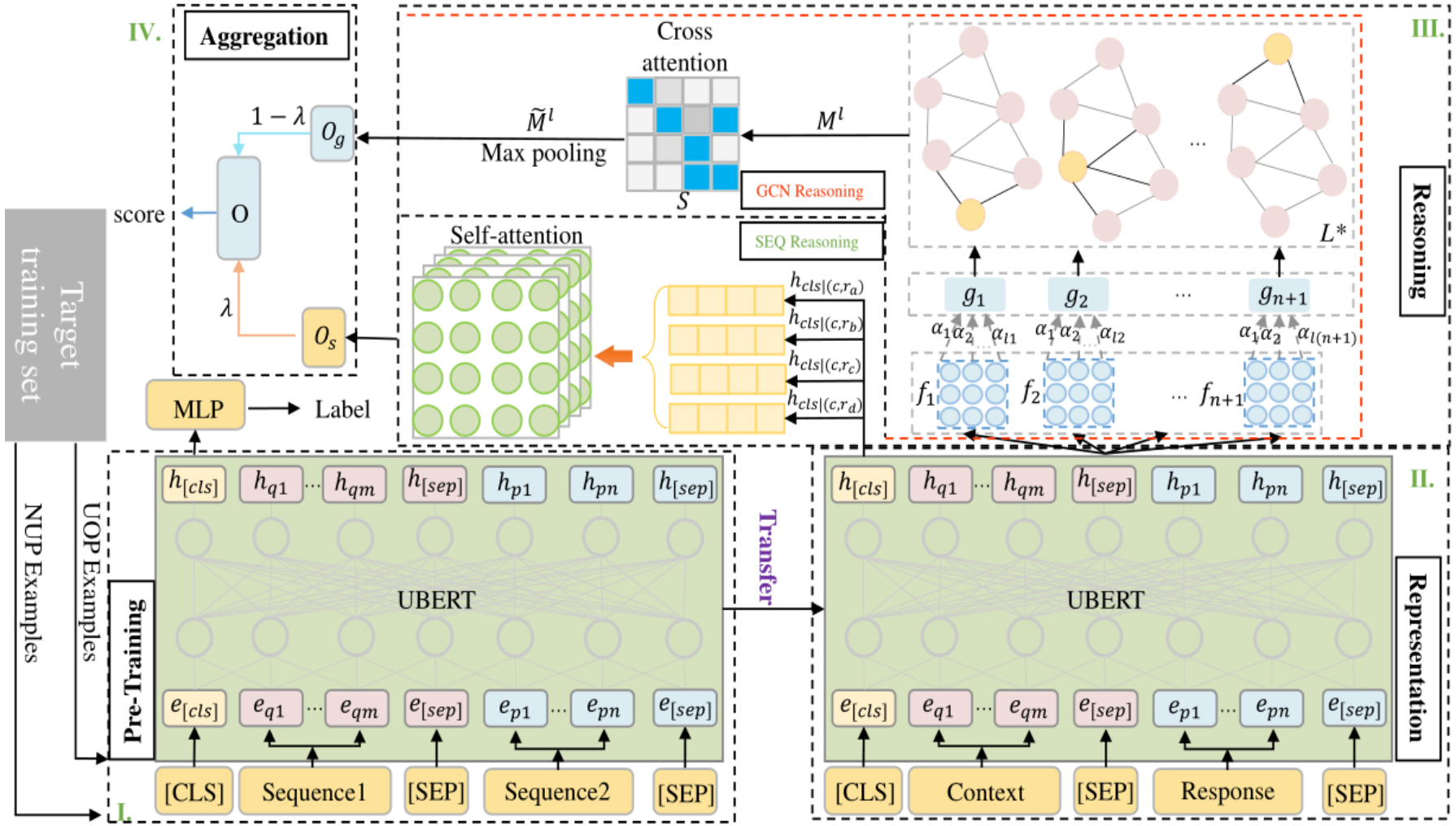


Yongkang Liu, Shi Feng, Daling Wang, Kaisong Song, Feiliang Ren, Yifei Zhang: A Graph Reasoning Network for Multi-turn Response Selection via Customized Pre-training. AAAI 2020 (CCF A)

5-基于自定义预训练的多轮回复选择图推理技术研究

- GRN

- Architecture
  - ✓ 语言模型ALBERT
  - ✓ 图推理子结构
  - ✓ 序列推理结构
  - ✓ 交叉/自注意力机制



Yongkang Liu, Shi Feng, Daling Wang, Kaisong Song, Feiliang Ren, Yifei Zhang: A Graph Reasoning Network for Multi-turn Response Selection via Customized Pre-training. AAAI 2020 (CCF A)

5-基于自定义预训练的多轮回复选择图推理技术研究

- 数据集MuTual和MuTual^{plus}

	MuTual			$MuTual^{plus}$		
Method	R@1	R@2	MRR	R@1	R@2	MRR
Human	93.8	97.1	96.4	93.0	97.2	96.1
Random	25.0	50	60.4	25.0	50	60.4
TF-IDF	27.9	53.6	54.2	27.8	52.9	76.4
DuLSTM	26.0	49.1	74.3	25.1	47.9	51.5
SMN	29.9	58.5	59.5	26.5	51.6	62.7
DAM	24.1	46.5	51.8	27.2	52.3	69.5
BIDAF	35.7	58.9	58.9	33.4	49.2	56.2
R-NRT	27.0	43.5	51.3	26.1	50.6	53.2
QANET	24.7	51.7	52.2	25.1	49.5	51.9
BERT	64.8	84.7	79.5	51.4	78.7	71.5
RoBERTa	82.5	95.3	90.4	75.7	92.8	85.6
SpanBERT	80.6	94.8	89.3	70.3	88.4	83.0
GPT-2	33.2	60.2	58.4	31.6	57.4	56.8
GPT-2-FT	39.2	67.0	62.9	22.6	61.1	53.5
BERTMC	66.7	87.8	81.0	58.0	79.2	74.9
RoBERTMC	68.6	88.7	82.8	64.3	84.5	79.2
ALBERT	84.7	96.2	91.6	78.9	94.6	88.4
GRN	91.5	98.3	95.4	84.1	95.7	91.3

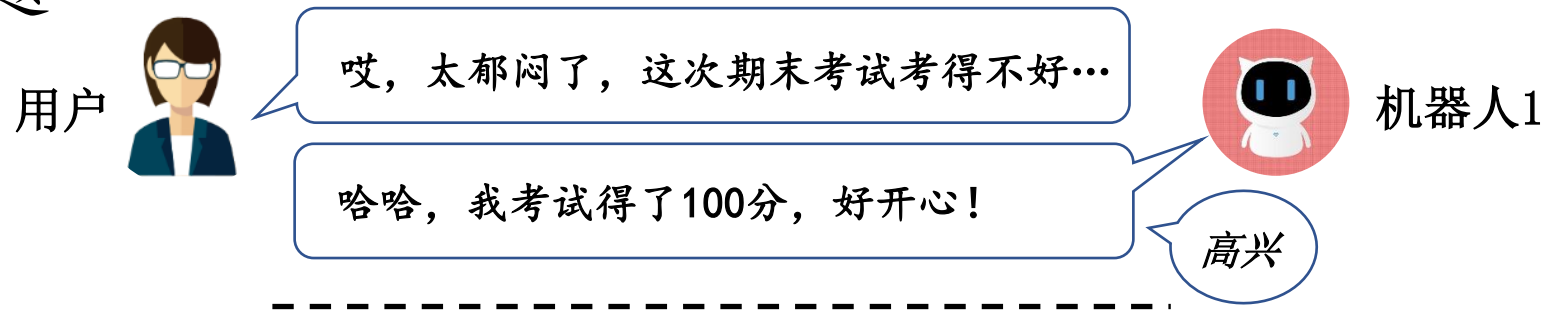
Method	R@1	R@2	MRR
GRN	93.5	98.5	97.1
-pre-training	90.5	97.3	94.7
-GCN match	91.5	97.9	95.5
-sequence match	91.3	97.6	95.2
-cross attention	92.2	97.3	95.6
-selfAtt	92.7	98.2	96.8

Model	T=2	T=3	T=4	T=5	T>6
Instance	290	143	115	51	287
Roberta	73.1	65.7	63.5	80.4	71.2
Roberta-MC	68.1	62.2	60.9	72.5	75.0
ALBERT	85.6	82.1	82.5	84.3	86.0
GRN	**92.1**	**93.1**	**88.6**	**88.2**	**91.9**

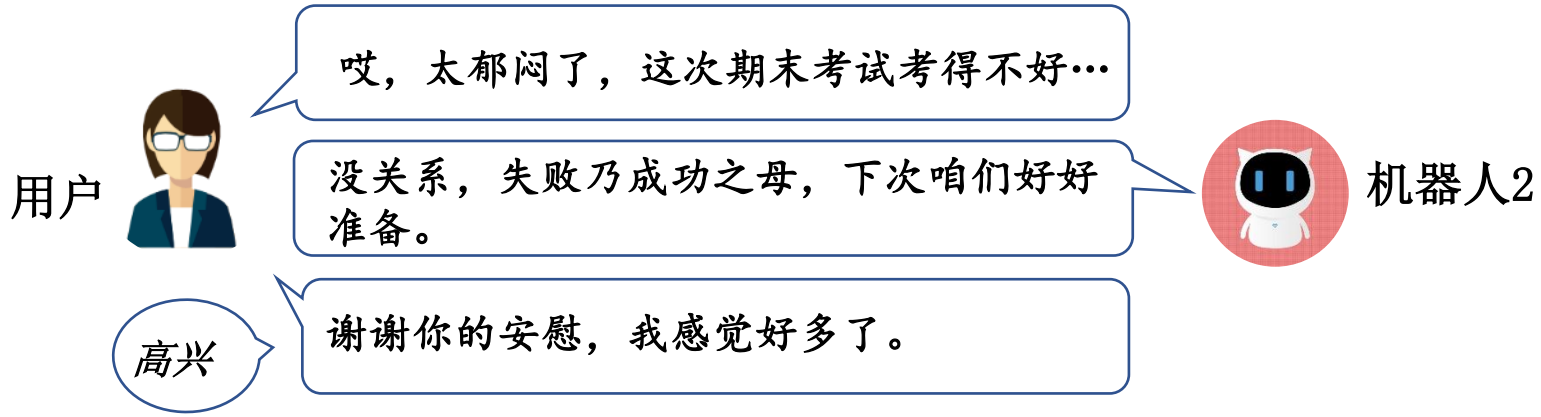
Yongkang Liu, Shi Feng, Daling Wang, Kaisong Song, Feiliang Ren, Yifei Zhang: A Graph Reasoning Network for Multi-turn Response Selection via Customized Pre-training. AAAI 2020 (CCF A)

6-面向用户情绪激励的回复生成技术研究

- 聊天机器人是人工智能领域的核心研究问题
- 赋予聊天机器人**表达情感的能力**是当前的研究热点
- 现有的情绪回复生成问题
  - ✓ 指定**机器人回复**的情绪
  - ✓ 能够生成表达特定情绪的回复
  - ✓ 忽视用户的感受



- 情绪激励回复生成问题
  - ✓ 指定**用户**的情绪状态
  - ✓ 生成回复，激发用户特定情绪
  - ✓ 可以帮助实现人机共情互动

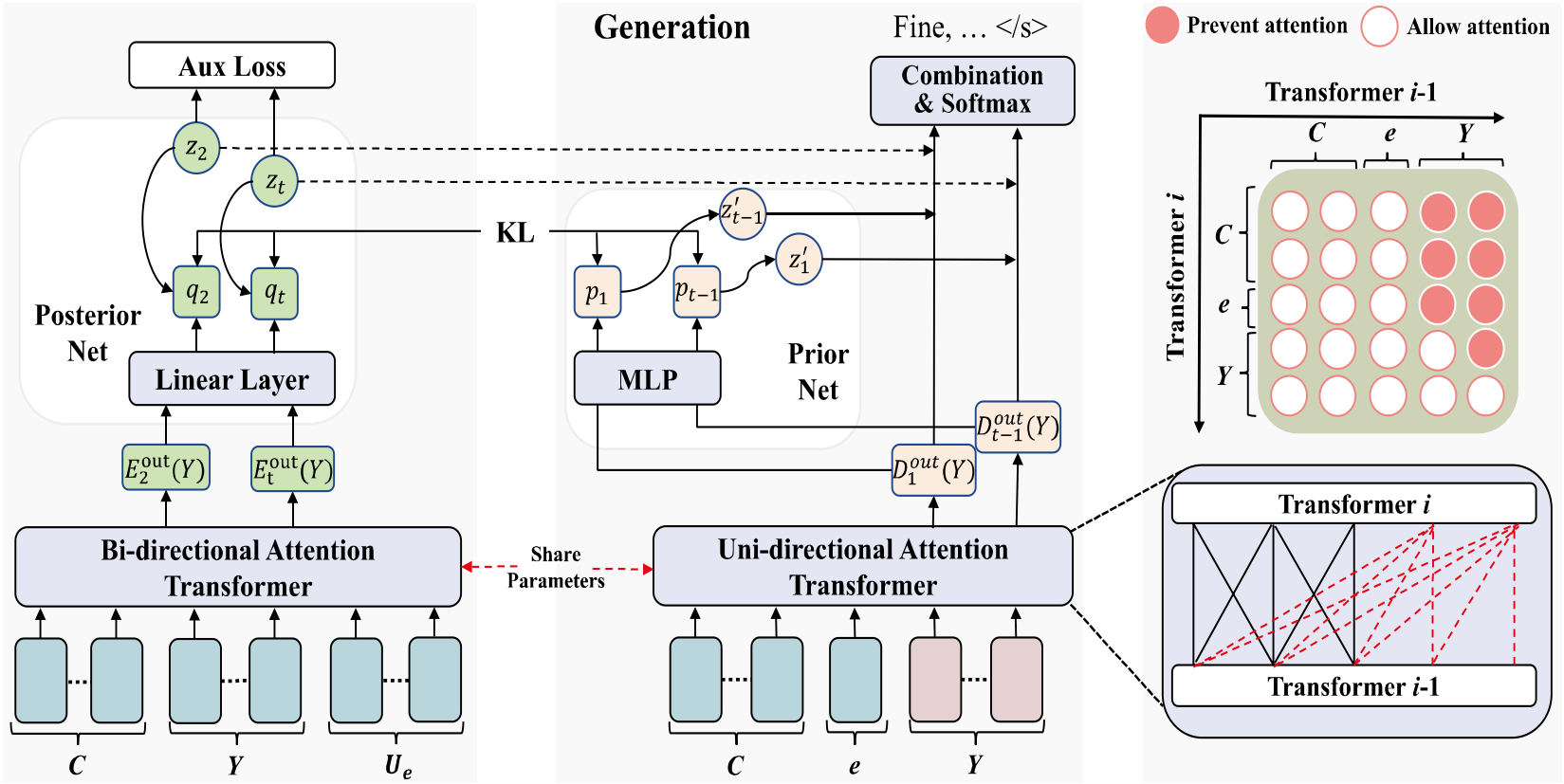


Shifeng Li, Shi Feng, Daling Wang, Kaisong Song, Yifei Zhang, Weichao Wang: EmoElicitor: An Open Domain Response Generation Model with User Emotional Reaction Awareness. IJCAI 2020 (CCF A)

6-面向用户情绪激励的回复生成技术研究

- EmoElicitor

- 主要架构
  - ✓ 预训练语言模型XLNet
  - ✓ Transformer编码器解码器架构
  - ✓ 变分循环神经网络VRNN
  - ✓ 新型损失函数



Shifeng Li, Shi Feng, Daling Wang, Kaisong Song, Yifei Zhang, Weichao Wang: EmoElicitor: An Open Domain Response Generation Model with User Emotional Reaction Awareness. IJCAI 2020 (CCF A)

6-面向用户情绪激励的回复生成技术研究

- 训练数据
  - Twitter上137421/4661/4739对话作为训练集/验证集/测试集
- 情感标签
  - 58类常用emoji(😂 😭 ❤️ 😓 😖 😠 ...)

Models	**Emb-avg%**	**Dist1%**	**Dist2%**	**BLUE%**	**B1%**	**B2%**	**B3%**	**Avg-len**	**Acc5%**
Human	-	10.6	47.50	-	-	-	-	15.22	65.1
S2S*	71.11	0.52	1.422	15.75	23.15	20.35	17.71	14.50	50.1
ECM*	70.98	0.30	1.085	13.57	21.05	17.83	15.38	10.80	48.0
Mojitalk*	71.03	4.16	13.18	15.04	22.61	19.35	16.93	12.90	50.1
XLNet	71.77	4.13	13.82	14.55	22.84	19.16	16.49	11.52	49.1
T-CVAE	71.99	4.96	21.62	15.99	24.31	20.58	18.01	12.92	50.2
T-CVAE w/o U_e	71.51	**5.10**	21.73	15.12	23.20	19.56	17.05	12.20	49.6
Ours	**73.18**	4.08	21.97	**18.54**	**28.08**	**23.53**	**20.73**	**15.73**	**51.7**
w/o U_e	72.77	4.56	**22.00**	17.71	27.02	22.73	19.88	14.45	51.1
w/o pre-train	72.54	2.46	21.50	17.29	25.52	21.87	19.33	15.20	50.8

Shifeng Li, Shi Feng, Daling Wang, Kaisong Song, Yifei Zhang, Weichao Wang: EmoElicitor: An Open Domain Response Generation Model with User Emotional Reaction Awareness. IJCAI 2020 (CCF A)

7-答案指导的开放域对话问题生成技术研究

- 赋予聊天机器人**生成问题的能力**可以主动与用户进行交互，从而与用户建立紧密的联结。

- 现有的对话问题生成
  - ✓ 空泛无意义的问题语句
  - ✓ 语义偏离的问题语句

- 答案指导的问题生成
  - ✓ 借助答案语句丰富的语义信息
  - ✓ 结合强化学习框架
  - ✓ 结合生成对抗网络框架

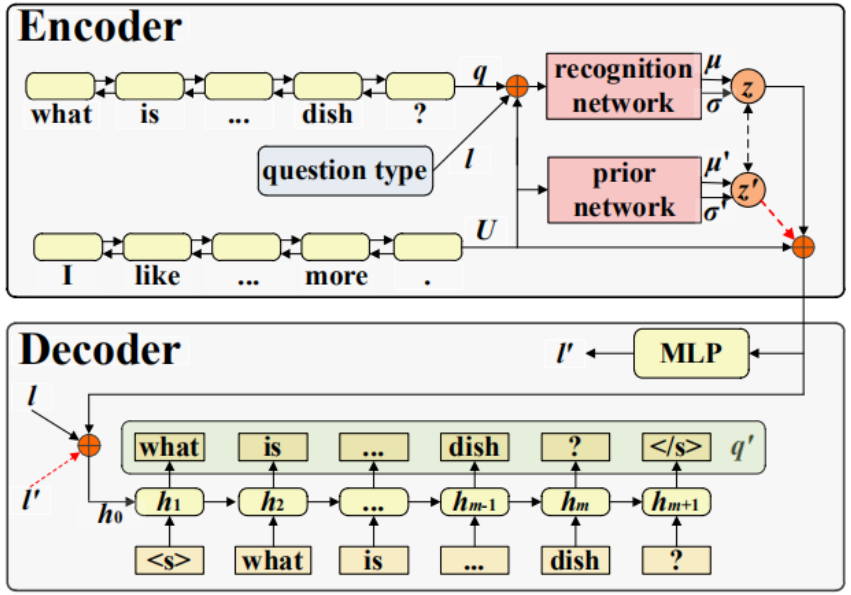
● Semantic coherence ● Dullness ● Deviation

Post
I like cooking more and more.
Question candidates
What are your special dishes ? ✓ Are you good at making appetizers ? ✓ What do you mean? ✗ How about going singing? ✗
Answer
Wow, I am only skillful in cooking beef.

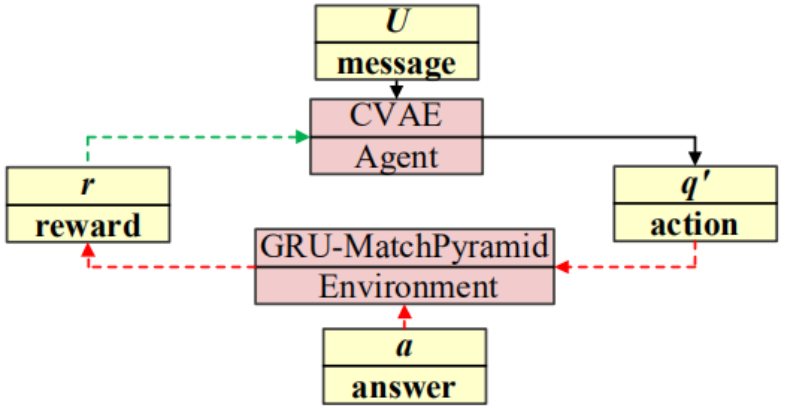
Weichao Wang, Shi Feng, Daling Wang, Yifei Zhang: Answer-guided and Semantic Coherent Question Generation in Open-domain Conversation EMNLP 2019 (CCF B)

7-答案指导的开放域对话问题生成技术研究

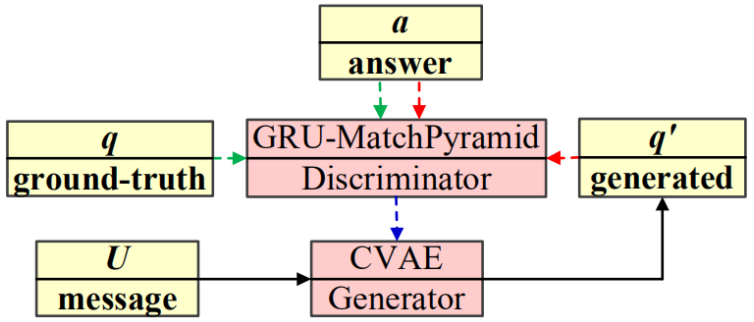
- 主要架构
  - ✓ 融入问题类型信息的变分自编码器
  - ✓ 基于强化学习的变分自编码器
  - ✓ 基于生成对抗网络的变分自编码器



Qt-CVAE模型



RL-CVAE模型



A-CVAE模型

Weichao Wang, Shi Feng, Daling Wang, Yifei Zhang: Answer-guided and Semantic Coherent Question Generation in Open-domain Conversation EMNLP 2019 (CCF B)

7-答案指导的开放域对话问题生成技术研究

- 训练数据
  - Reddit上1,104,345/30,000/30,000对话作为训练集/验证集/测试集
- 数据形式
  - 上下文-问题-答案

Question Generation Evaluation				
Models	RubA	RubG	Dist2	PPL
Seq2Seq	0.614	0.574	0.008	63.02
CVAE	0.682	0.649	0.112	20.39
STD	0.658	0.613	0.010	28.75
HTD	0.689	0.654	0.114	26.02
CVAE (qt)	0.688	0.652	0.114	20.03
A-CVAE	0.715	0.661	0.123	19.51
RL-CVAE	**0.720**	**0.668**	**0.185**	**16.93**

Models	A	S	W	Sum
seq2seq	0.486	0.208	0.196	0.890
CVAE	0.458	0.484	0.408	1.350
STD	0.504	0.322	0.272	1.098
HTD	0.528	0.486	0.406	1.420
CVAE(qt)	0.462	0.508	0.468	1.438
A-CVAE	0.540	0.578	0.514	1.632
RL-CVAE	**0.542**	**0.602**	**0.526**	**1.670**

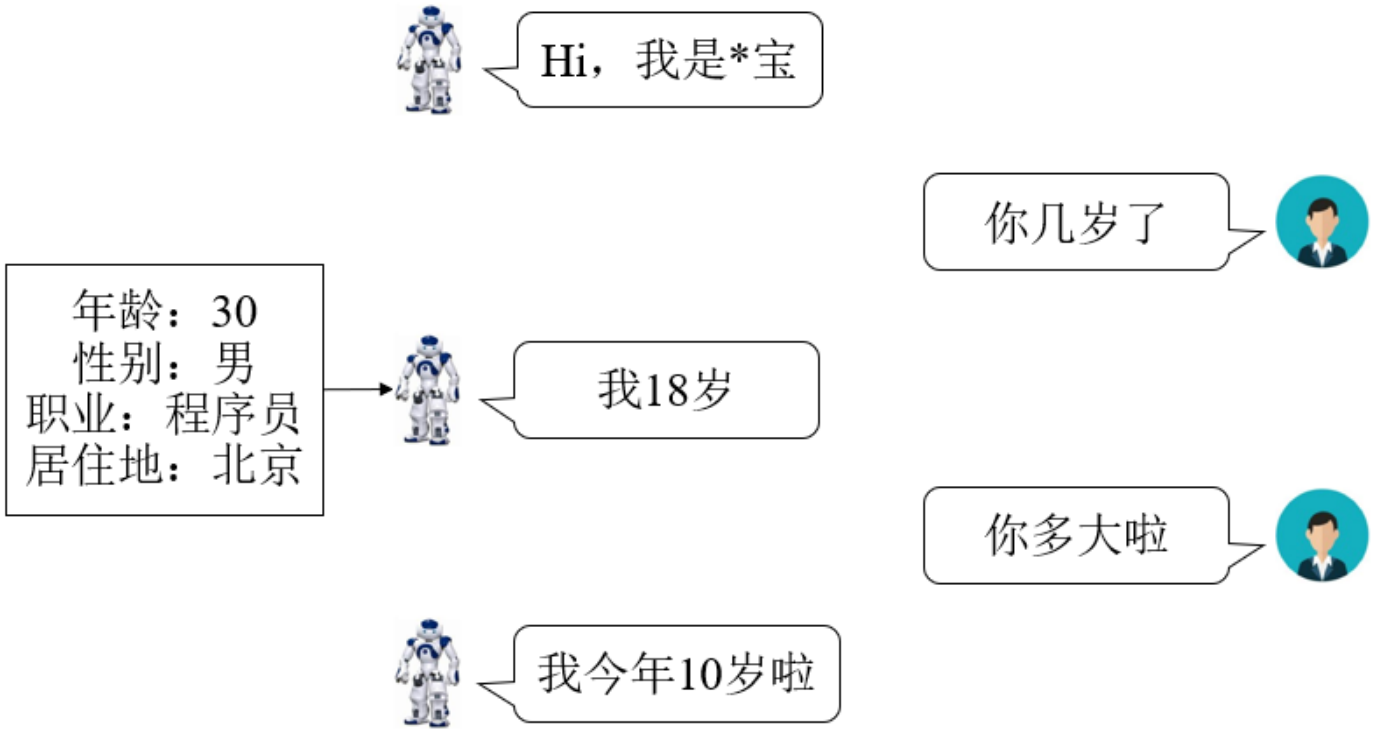
Weichao Wang, Shi Feng, Daling Wang, Yifei Zhang: Answer-guided and Semantic Coherent Question Generation in Open-domain Conversation EMNLP 2019 (CCF B)

8-基于对抗生成网络的个性化回复生成技术研究

- 赋予聊天机器人**性格**，并在生成回复中保持**性格的一致性**是当前的研究热点

i am a musician and hope to make it big some day.
i play the **piano** and **guitar**.
my favorite type of music to sing is **folk** music.
my father is a hero.

Persona-Chat数据集中性格描述语句

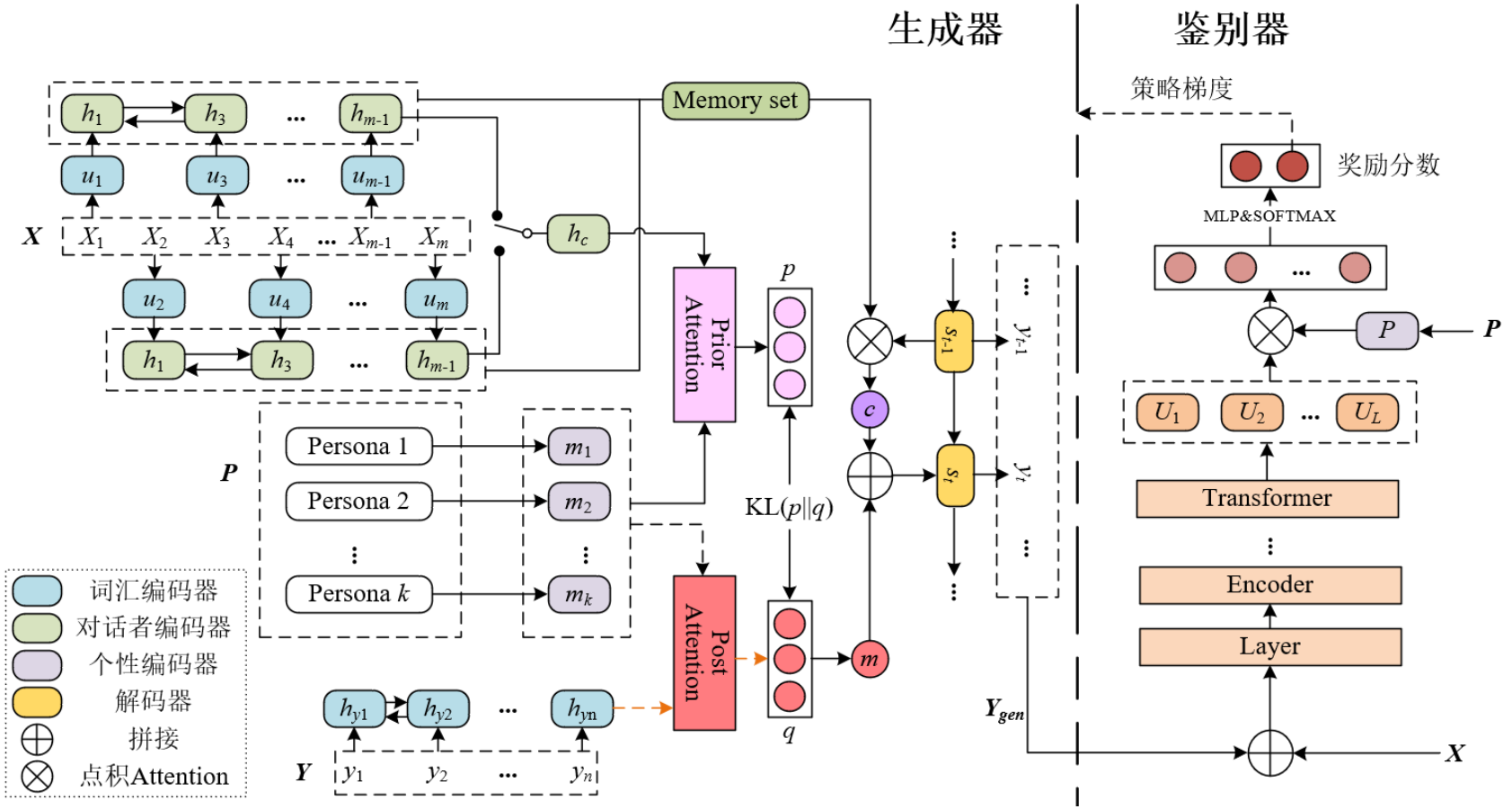


- 主要挑战
  - ✓ 非结构化性格的表征
  - ✓ 生成回复中合适性格的选择
  - ✓ 对话上下文的建模
  - ✓ 避免简短、枯燥的回复

Pengcheng Lv, Shi Feng, Daling Wang, Yifei Zhang, Ge Yu: PersonaGAN: Personalized Response Generation via Generative Adversarial Networks. DASFAA 2020 (CCF B)

8-基于对抗生成网络的个性化回复生成技术研究

- 生成器
  - ✓ 层次GRU对话上下文表征
  - ✓ 层次GRU个性表征
  - ✓ 后验选择机制指导个性选择
  - ✓ 个性与上下文输入解码器
- 鉴别器
  - ✓ 基于transformer的匹配模型
  - ✓ 计算生成回复与上文和个性的匹配分数



Pengcheng Lv, Shi Feng, Daling Wang, Yifei Zhang, Ge Yu: PersonaGAN: Personalized Response Generation via Generative Adversarial Networks. DASFAA 2020 (CCF B)