第二届 THUNLP & HIT-SCIR学术联谊会

对话摘要最新进展简述

冯夏冲

2021-5-31

目录

01 对话摘要的过去

摘要

- 摘要旨在将输入数据转换为包含关键信息的简短文本。

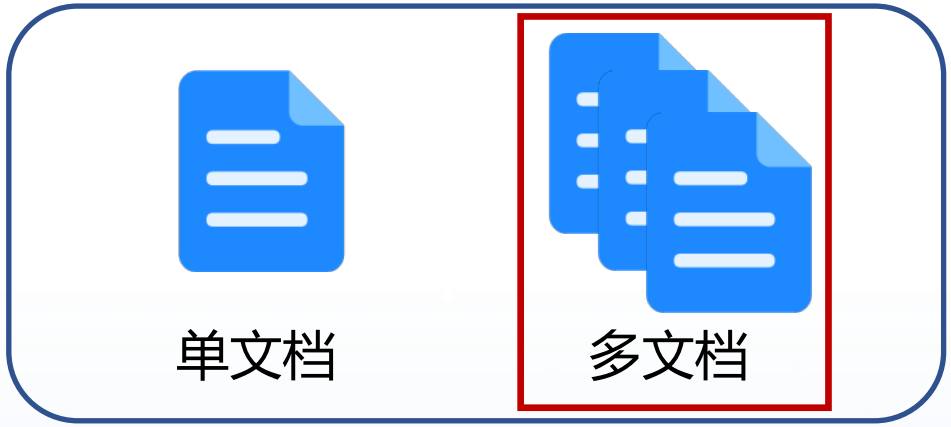
原文

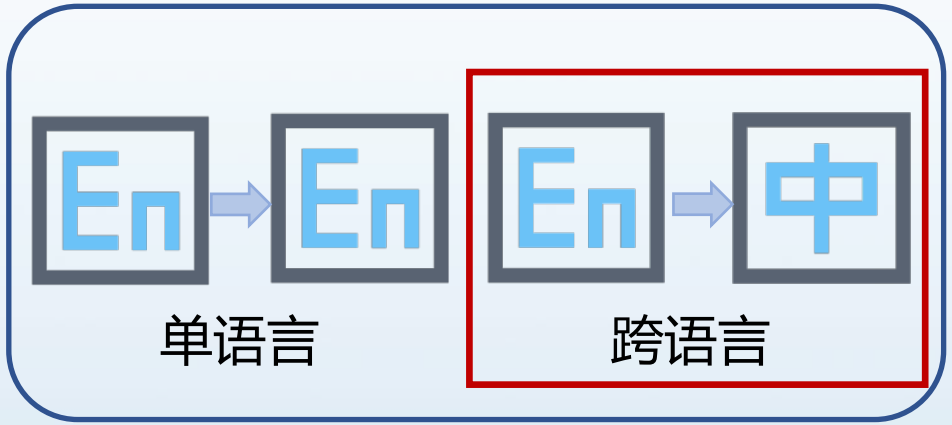
本次活动由清华THUNLP刘知远副教授和哈工大SCIR车万翔教授联合发起，哈工大SCIR博士生窦隆绪、施琦、冯夏冲、马龙轩和清华THUNLP博士生韩旭、肖朝军等同学负责了活动组织和现场主持等工作。上午9点整，联谊会在哈工大活动中心214会议室准时开始。刘挺教授首先作了开场致辞，对清华大学师生们的光临表示了热烈欢迎，回顾了哈工大与清华在自然语言处理领域过去20余年的多次合作，希望哈工大SCIR与清华THUNLP的学术交流活动能以本次活动为起点继续发扬光大，构建中国高校学术交流的新模式。接下来，清华THUNLP刘洋教授以《科研中的时间管理》为题，向在场的老师同学们介绍了科研中时间管理的重要性，以及培养时间管理能力方面的经验。随后，哈工大SCIR车万翔教授以《NLPer的核心竞争力是什么？》为题，从三个方面探讨了NLP从业者应该具备什么样的能力。在两位老师的报告后，秦兵、车万翔、刘洋、刘知远等四位老师进行了圆桌会议讨论并回答了同学们提出的问题，对预训练模型的未来、NLP未来的研究重点、应用落地和创业前景等多个问题发表了看法。下午，双方同学以现场海报展示以及口头报告的形式进行了科研成果交流。现场海报共展示了37篇双方在2019年的工作，负责海报讲解的同学们向双方同学介绍了各自的研究成果，现场讨论气氛十分热烈。随后，清华THUNLP的韩旭、高天宇、周界、张嘉成和哈工大SCIR的段俊文、姜天文、王少磊、刘元兴同学分别作了口头报告介绍自己的工作，现场老师和同学们针对报告内容进行了细致探讨。最后，孙茂松教授对本次活动作了总结。孙老师首先感谢了哈工大SCIR实验室举办本次活动，认为本次活动增加了双方同学的了解，进行了思想的碰撞，取得了预期的交流效果。孙老师对同学们的学术研究规划提出了建议并希望未来继续进行此类交流活动。

摘要

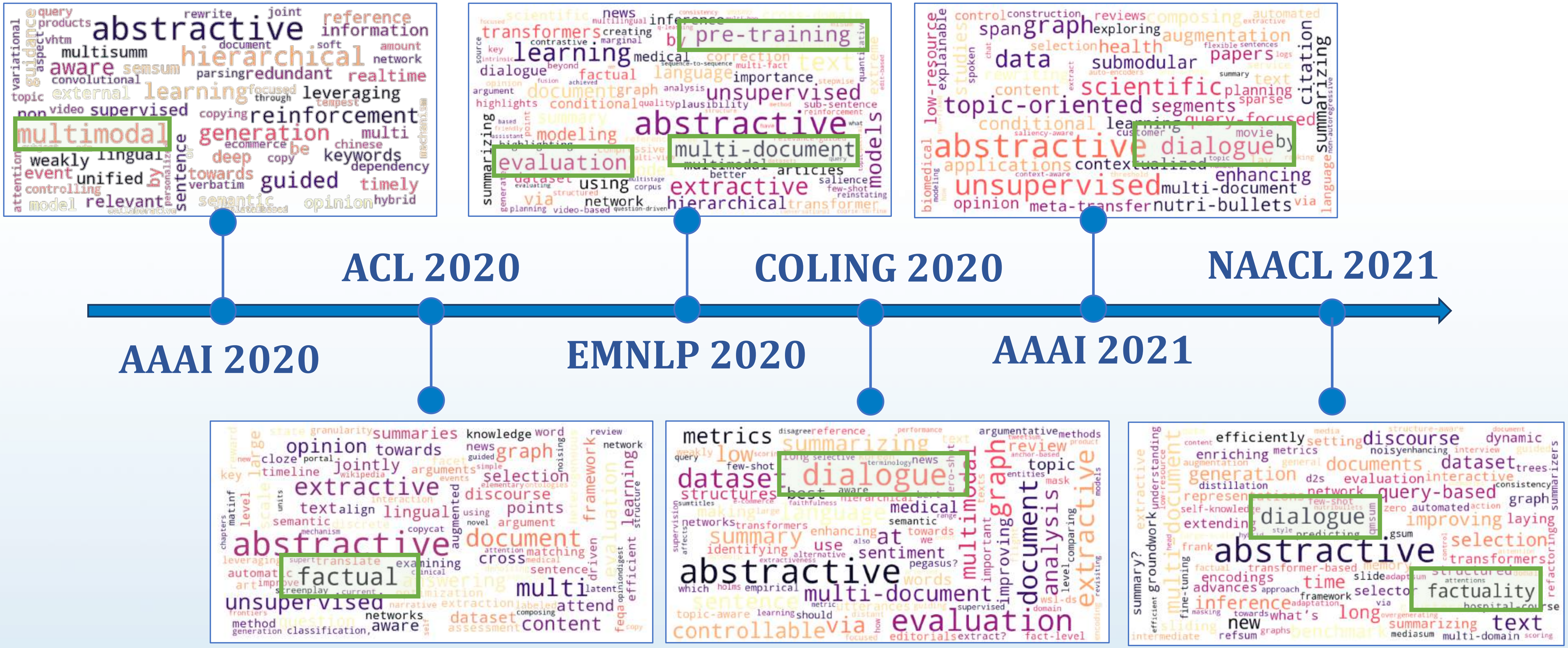
2020年1月12日，哈尔滨工业大学社会计算与信息检索研究中心（HIT-SCIR）与清华大学自然语言处理与社会人文计算实验室（THUNLP）首届学术联谊会于**哈尔滨**成功举办。清华THUNLP**孙茂松、刘洋、刘知远**以及哈工大SCIR**刘挺、秦兵、车万翔、刘铭、赵妍妍、丁效、冯骁骋、刘建伟**等老师出席本次学术联谊会，哈工大SCIR的**55位同学**和清华THUNLP的**27位同学**参与活动。

摘要的分类





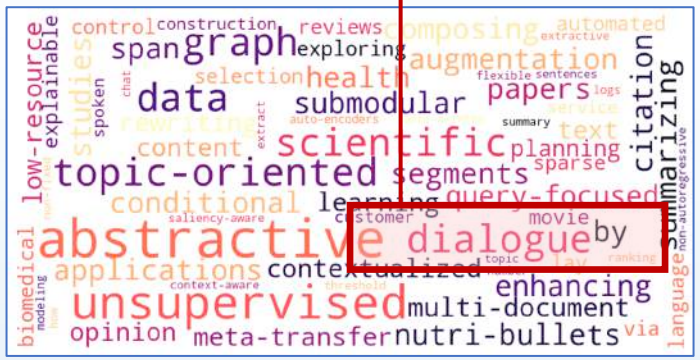
摘要近两年的进展



AAAI 2020
multimodal
ACL 2020
factual
EMNLP 2020
pre-training
multi-document
evaluation
COLING 2020
dialogue
AAAI 2021
dialogue
NAACL 2021
dialogue
factuality

对话摘要崭露头角

对话摘要



COLING 2020

AAAI 2021

NAACL 2021

对话摘要示例

- 对话摘要关注对话类文本
  - 会议，闲聊，邮件，客服对话，医患对话，辩论等

部分会议
工业设计师 ： 如果我们有电源支架呢? 界面设计师 ： 你可以为支架和遥控器设计一些简洁的小设计。 项目经理 ： 这会增加成本。 项目经理 ： 我们需要改变最终的成本。
标准摘要
工业设计师建议在设备中加入一个电源支架，但最终被决定这不是一个有用的功能。

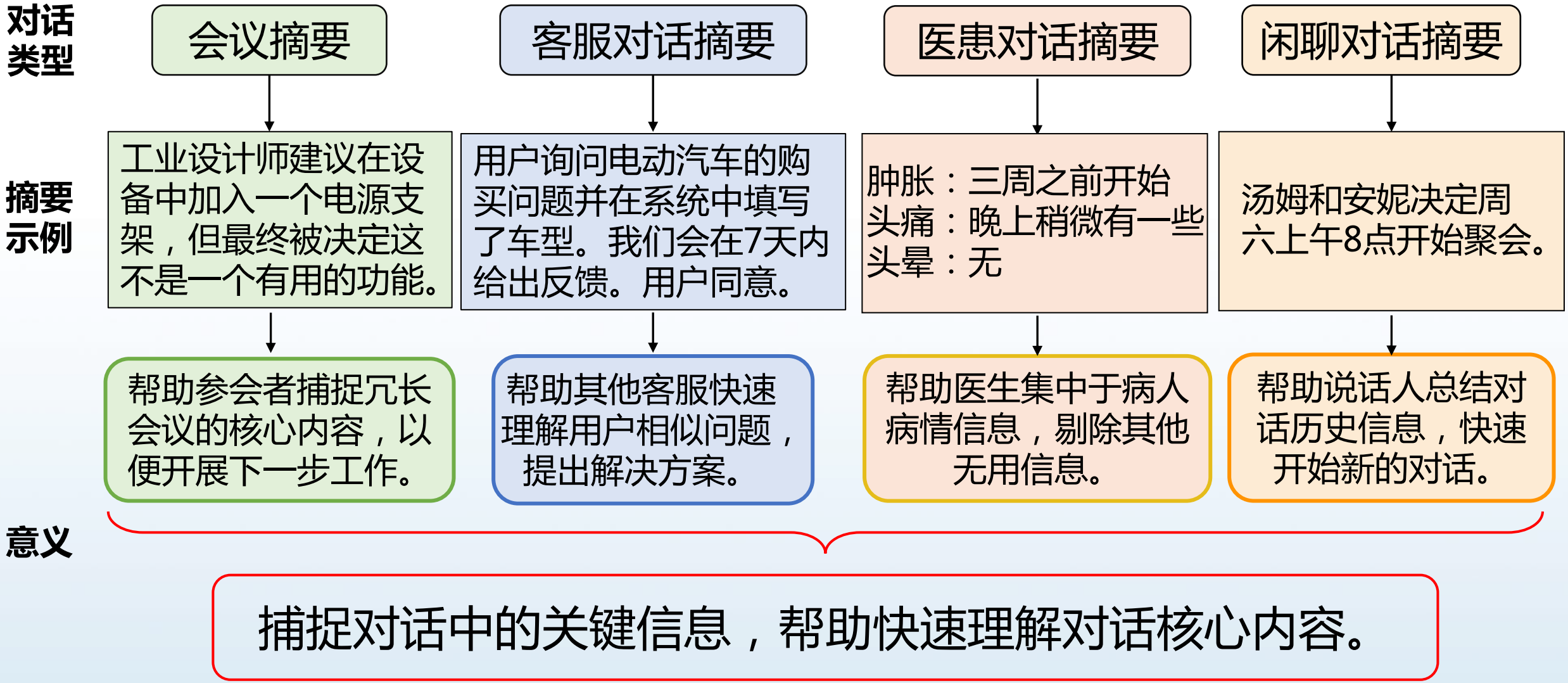
Meeting Minutes
会议纪要

闲聊对话
鲍勃 ： 老兄，你可以来接我一下吗? 汤姆 ： 你在哪里? 鲍勃 ： 在家，我的车坏了，我现在急需去上班，我需要你的帮助。 汤姆 ： 我现在出发，10分钟之内到。
标准摘要
鲍勃的车坏了，汤姆会在10分钟内让他搭便车，送他去上班。

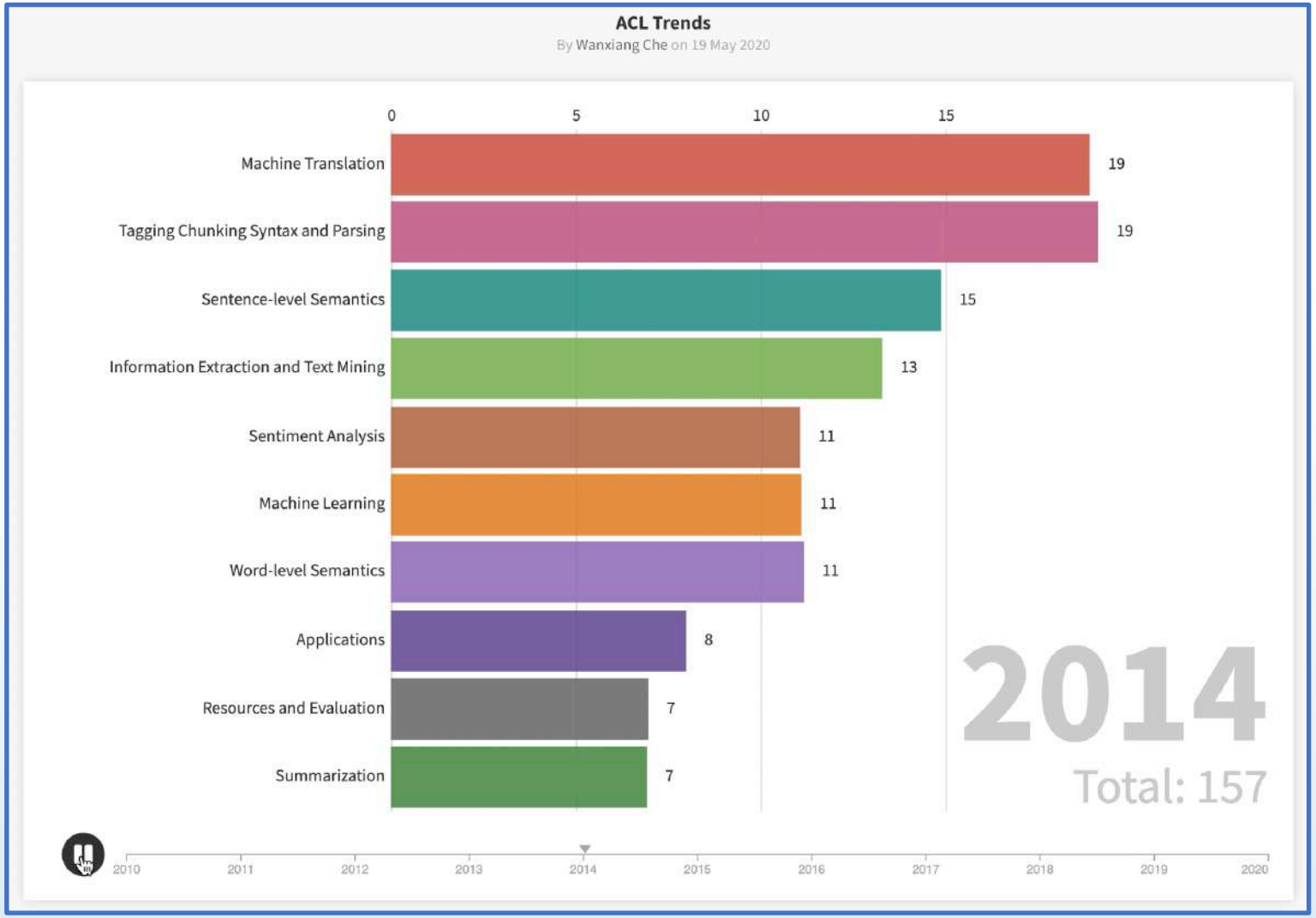
医患对话
医生 ： 你最近有肿胀吗? 患者 ： 时有时无。 医生 ： 我知道了，什么时候开始的? 患者 ： 大约在三周之前。
标准摘要
肿胀：大约三周之前开始，症状时有时无。

SOAP
主观描述、客观观察、医生诊断、治疗计划

对话摘要的意义

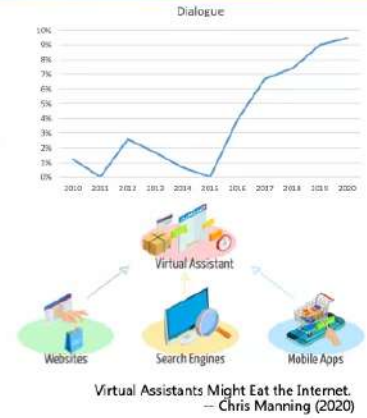

对话类型
会议摘要
客服对话摘要
医患对话摘要
闲聊对话摘要
摘要示例
工业设计师建议在设备中加入一个电源支架，但最终被决定这不是一个有用的功能。
用户询问电动汽车的购买问题并在系统中填写了车型。我们会在7天内给出反馈。用户同意。
肿胀：三周之前开始
头痛：晚上稍微有一些
头晕：无
汤姆和安妮决定周六上午8点开始聚会。
意义
帮助参会者捕捉冗长会议的核心内容，以便开展下一步工作。
帮助其他客服快速理解用户相似问题，提出解决方案。
帮助医生集中于病人病情信息，剔除其他无用信息。
帮助说话人总结对话历史信息，快速开始新的对话。
捕捉对话中的关键信息，帮助快速理解对话核心内容。

对话摘要的发展背景



人机对话异军突起

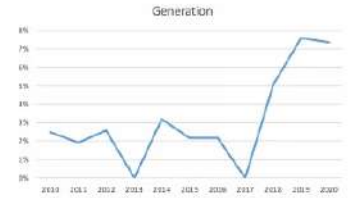
- 研究意义明确，应用价值巨大
  - 通过图灵测试
  - 新的人机交互模式
- 曾经主要在语音领域进行研究
  - 随着语音识别的突破，逐渐为NLPer所重视
- 研究重点
  - 任务型对话
    - NLU、DM、NLG
  - 开放域对话
    - 端到端
- 研究趋势
  - 任务型对话
    - 端到端
  - 开放域对话
    - 引入更多知识和步骤



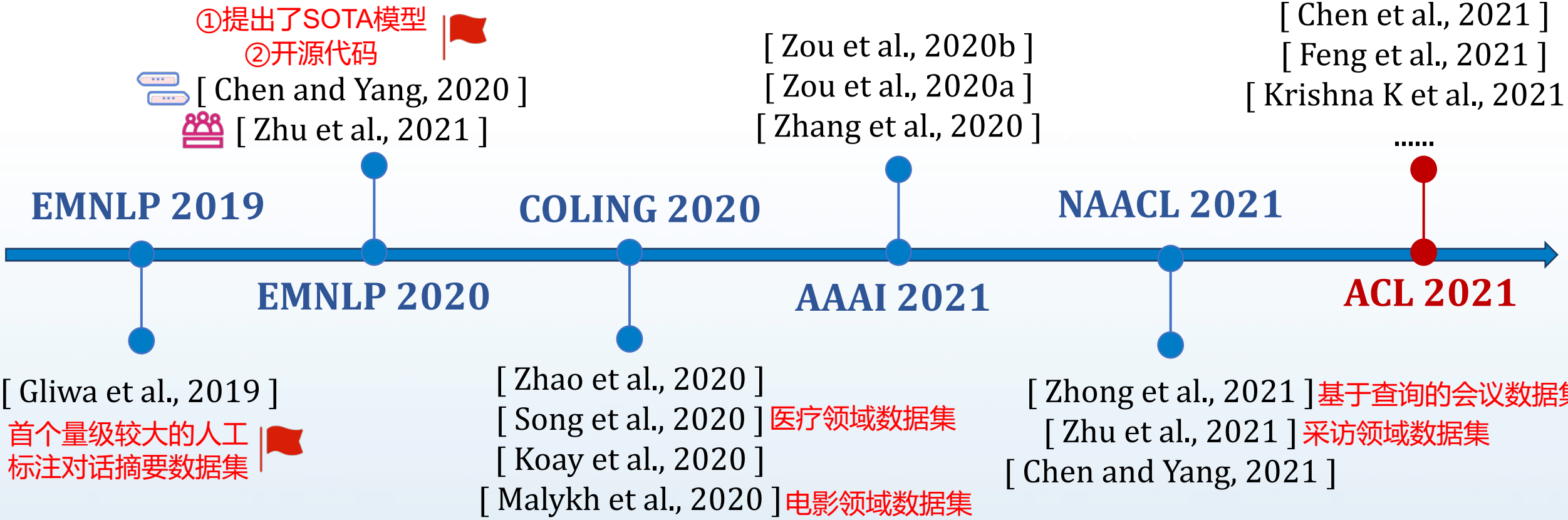
Virtual Assistants Might Eat the Internet.
-- Chris Manning (2020)

文本生成强势上升

- 深度学习使“无中生有”的生成为可能
- 与对话、文摘、机器翻译等关系紧密
- 研究热点
  - VAE vs. Pre-training
    - PPVAE , Coupled-VAE
    - BART
  - Non-autoregressive vs. Autoregressive
    - CoMMA
    - PMLM
  - 基于句法
    - Syn-QG , Paraphrase Generation , Systematic Assessment



对话摘要的发展脉络

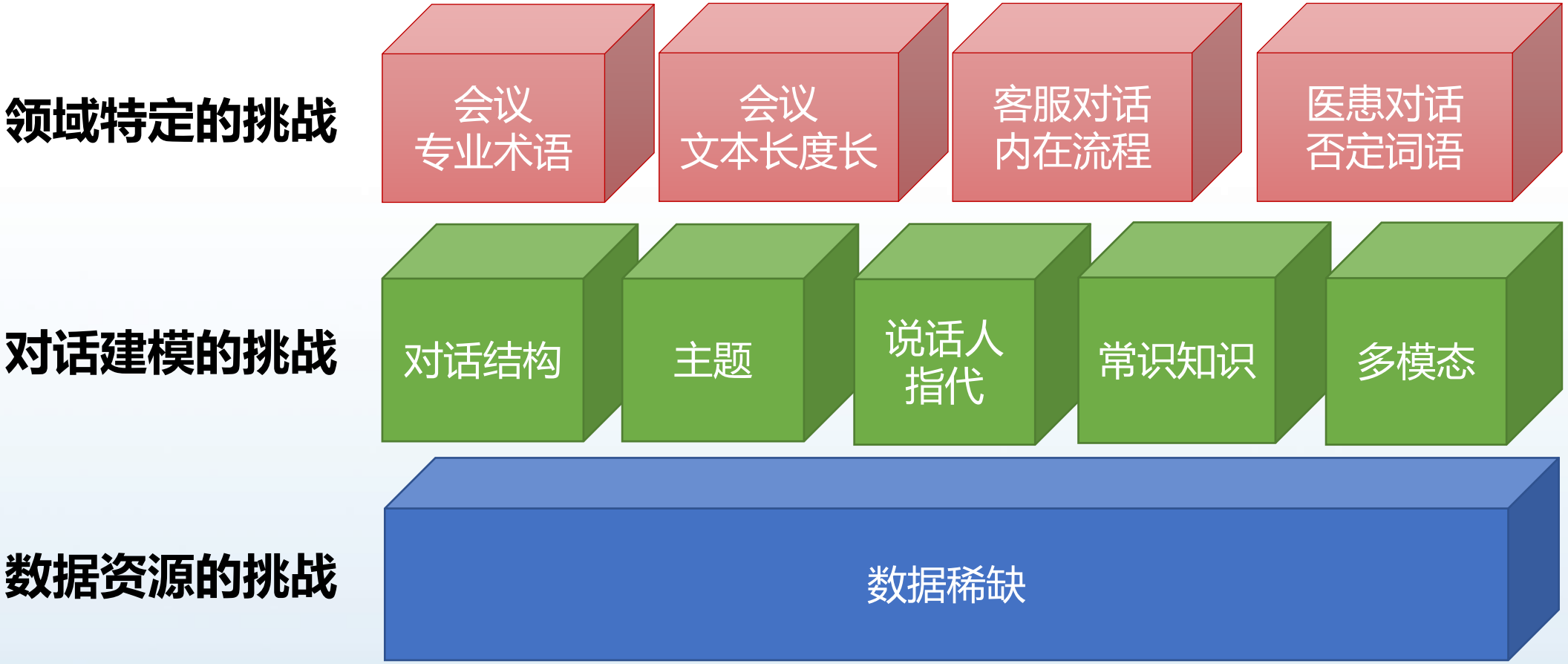

①提出了SOTA模型
②开源代码
[Chen and Yang, 2020]
[Zhu et al., 2021]
[Zou et al., 2020b]
[Zou et al., 2020a]
[Zhang et al., 2020]
口语领域数据集
[Chen et al., 2021]
[Feng et al., 2021]
[Krishna K et al., 2021]
.....
EMNLP 2019
EMNLP 2020
COLING 2020
AAAI 2021
NAACL 2021
ACL 2021
[Gliwa et al., 2019]
首个量级较大的人工
标注对话摘要数据集
[Zhao et al., 2020]
[Song et al., 2020] 医疗领域数据集
[Koay et al., 2020]
[Malykh et al., 2020] 电影领域数据集
[Zhong et al., 2021] 基于查询的会议数据集
[Zhu et al., 2021] 采访领域数据集
[Chen and Yang, 2021]

阶段性总结

- 摘要任务
  - 不同的分类标准有不同的摘要任务分类
  - 核心是选取输入关键信息
- 对话摘要出现的背景
  - 人机对话+文本生成
- 对话摘要发展的条件
  - 数据集的发展

02 对话摘要的现在

对话摘要任务的挑战


领域特定的挑战
会议
专业术语
会议
文本长度长
客服对话
内在流程
医患对话
否定词语
对话建模的挑战
对话结构
主题
说话人
指代
常识知识
多模态
数据资源的挑战
数据稀缺

数据资源的挑战

数据资源的挑战

数据稀缺
①新的数据集 ②借助预训练 ③无监督方法

新的数据集

ID	Dataset	# instances	# tokens (input)	# tokens (summary)	# speakers	Abstractive	Extractive	Domain
1	**AMI**	137	4757.0	322.0	4.0	√	√	Meetings
2	**ICSI**	59	10189.0	534.0	6.2	√	√	Meetings
3	**SAMSum**	16.4k	83.9	20.3	2.2	√		ChitChat
4	**MediaSum**	463.6k	1553.7	14.4	6.5	√		News Interviews
5	**QMSum**	1.8k	9069.8	69.6	9.2	√		Meetings
6	**SUMMSCREEN**	26.9k	6612.5	337.4	28.3	√		Television Series
7	**SumTitles**	21.4k	423.06	55.03	4.88	√		Movie
8	**DialoSum**	13.4k	131	13.8	-	√		Spoken
9	**GupShup**	16.4k	**83.9**	20.3	2.2	√		Cross-lingual
10	**LCSPIRT**	38500	684.3	75	2	√		Police

借助预训练：领域外数据

- 使用新闻摘要数据预训练

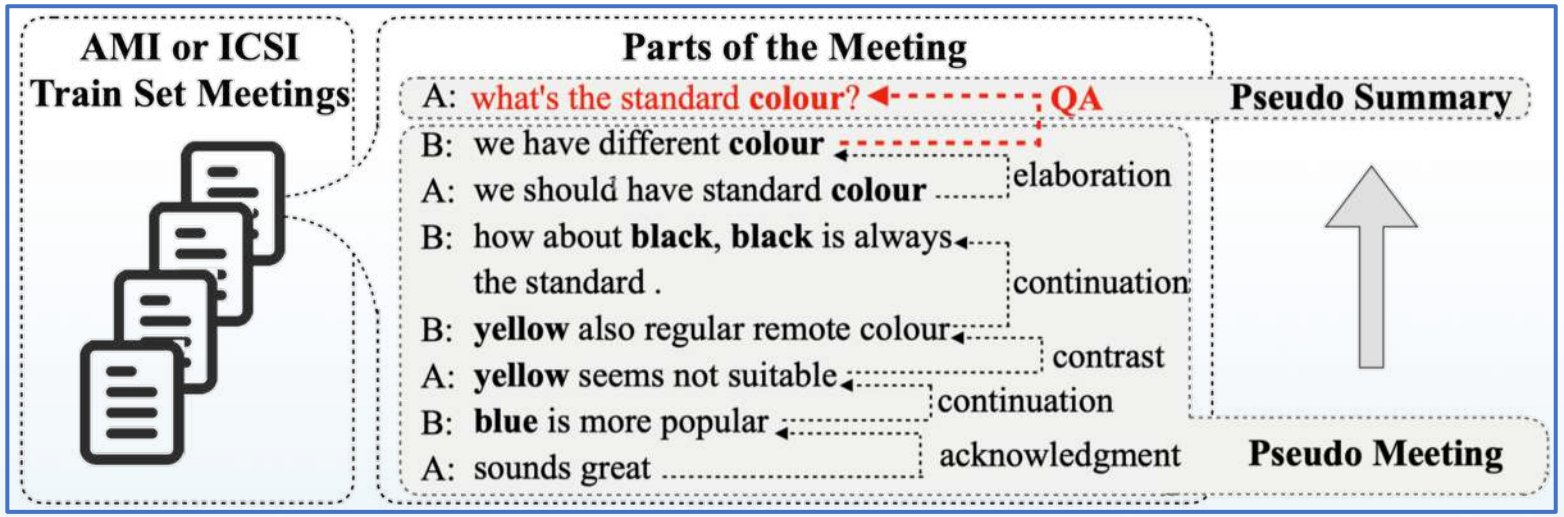
Model	ROUGE-1	R-2	R-SU4
AMI			
HMNet	**53.0**	**18.6**	**24.9**
−pretrain	48.7	18.4	23.5
−role vector	47.8	17.2	21.7
−hierarchy	45.1	15.9	20.5
ICSI			
HMNet	**46.3**	**10.6**	**19.1**
−pretrain	42.3	**10.6**	17.8
−role vector	44.0	9.6	18.2
−hierarchy	41.0	9.3	16.8

Table 3: Ablation study of HMNet.

A Hierarchical Network for Abstractive Meeting Summarization with Cross-Domain Pretraining Findings of EMNLP2020

借助预训练：领域内数据

- 构造伪造会议摘要数据集用于预训练
  - “问题”会引起“讨论”，“问题”包含了“讨论”的核心内容。



	AMI Pseudo Corpus	ICSI Pseudo Corpus
# of Original Data	97	53
# of Pseudo Data	1539	1877
Avg.Tokens	124.44	107.44
Avg.Sum	13.18	11.97

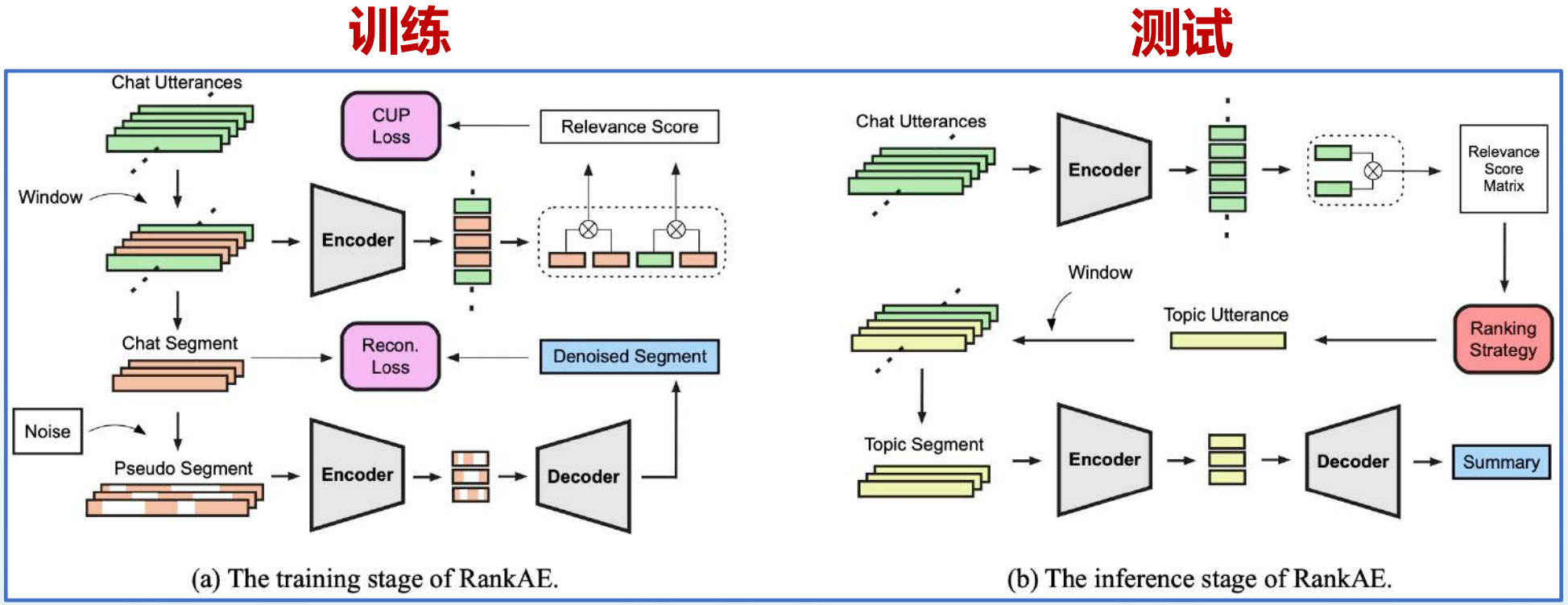
Dialogue Discourse-Aware Graph Model and Data Augmentation for Meeting Summarization IJCAI2021

无监督方法

基于相似度选择主题句+降噪自编码器

①
训练句子相似度计算模型

②
训练降噪自编码器



③
根据句子相似度，使用MMR算法选择主题句

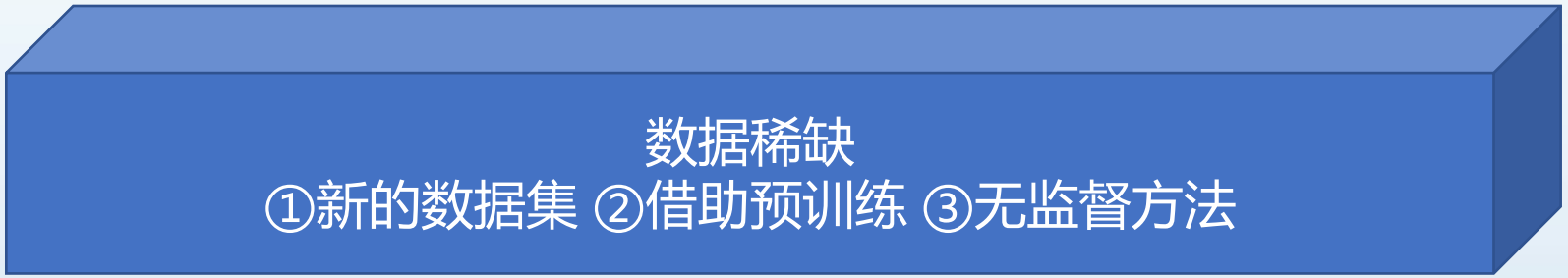
④
生成摘要

Unsupervised Summarization for Chat Logs with Topic-Oriented Ranking and Context-Aware Auto-Encoders AAAI2020

对话建模的挑战

对话建模的挑战

数据资源的挑战

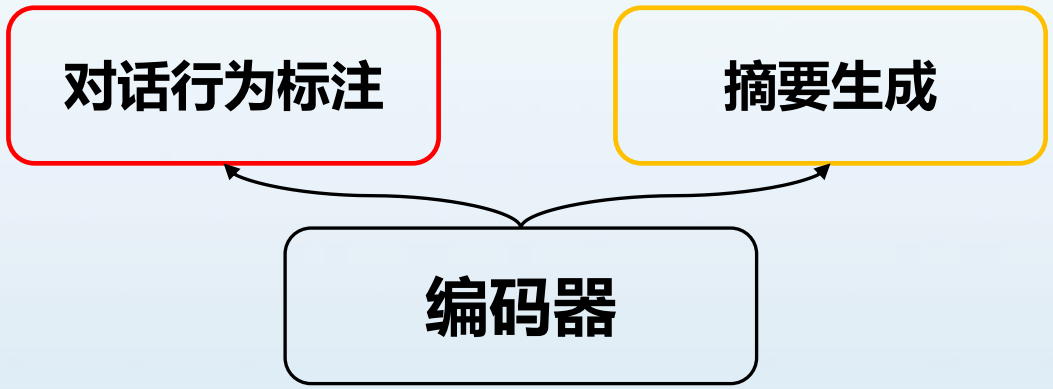


对话结构：对话行为

- 对话行为（Dialogue Act）指示了句子在对话中的作用与影响

Multi-Party Dialogue	Dialogue Act
A: mm-hmm .	Backchannel
B: mm-hmm .	Backchannel
C: then , these are some of the remotes which are different in shape and colour , but they have many buttons .	Inform
C: so uh sometimes the user finds it very difficult to recognise which button is for what function and all that .	Inform
D: so you can design an interface which is very simple , and which is user-friendly .	Inform
D: even a kid can use that .	Inform
A: so can you got on t t uh to the next slide .	Suggest
Summary: alternative interface options	

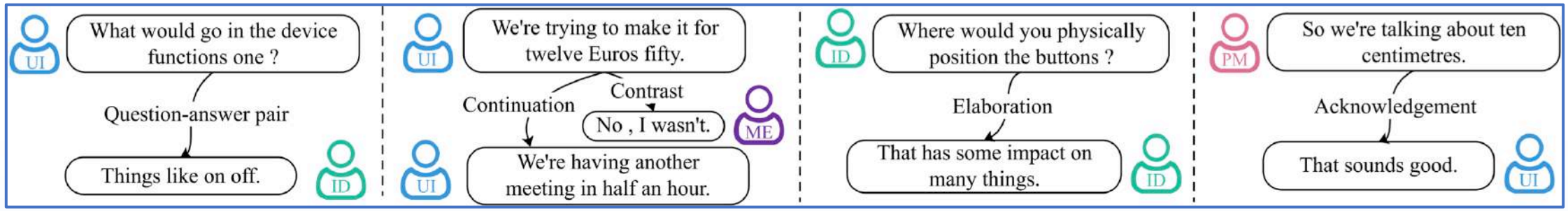
- 模型：多任务学习



Abstractive Dialogue Summarization with Sentence-Gated Modeling Optimized by Dialogue Acts SLT2018

对话结构：对话篇章结构

- 对话篇章结构指示了句子之间的交互关系



Parts of the Meeting
$\mathcal{A}$: What if we have a battery charger? $\mathcal{B}$: You can have neat design for it. $\mathcal{C}$: It would increase the cost. $\mathcal{C}$: We have to change the end cost.   QA, Contrast   Continuation
Summary
$\mathcal{A}$ **asked** whether to include a battery charger. $\mathcal{B}$ **answered his question**. However, $\mathcal{C}$ **disagrees with** $\mathcal{A}$ **since it would increase the final cost.**

Dialogue Discourse-Aware Graph Model and Data Augmentation for Meeting Summarization IJCAI2021

闲聊中的主题信息

- 主题漂移（Topic Drift）是对话中的一种常见现象

对话级别

句子级别

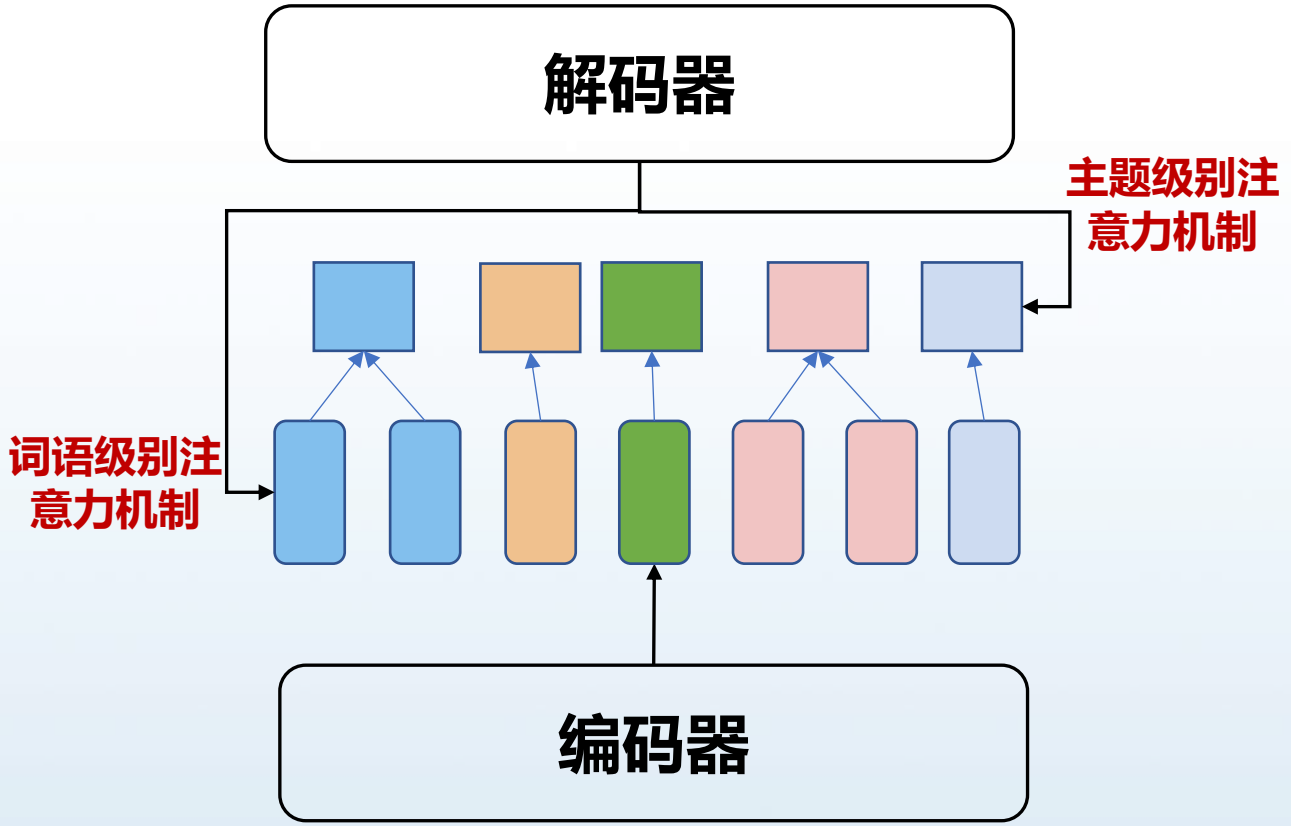
Conversation		Topic View	Stage View
James:	Hey! I have been thinking about you :)	Greetings	Openings
Hannah:	Oh, that's nice ;)		
James:	What are you up to?	Today's plan	Intention
Hannah:	I'm about to sleep		
James:	I miss u. I was hoping to see you		
Hannah:	Have to get up early for work tomorrow	Plan for tomorrow	Discussion
James:	What about tomorrow?		
Hannah:	To be honest I have plans for tomorrow evening		
James:	Oh ok. What about Sat then?	Plan for Saturday	
Hannah:	Yeah. Sure I am available on Sat		
James:	I'll pick you up at 8?	Pick up time	
Hannah:	Sounds good. See you then.		Conclusion
Summary	James misses Hannah. They agree for James to pick Hannah up on Saturday at 8.		

Table 1: Example conversation from SAMSum (Gliwa et al., 2019) with its topic view and stage view (extracted by our methods), and the human annotated summary.

Multi-View Sequence-to-Sequence Models with Conversational Structure for Abstractive Dialogue Summarization EMNLP2020

医患对话中的主题信息

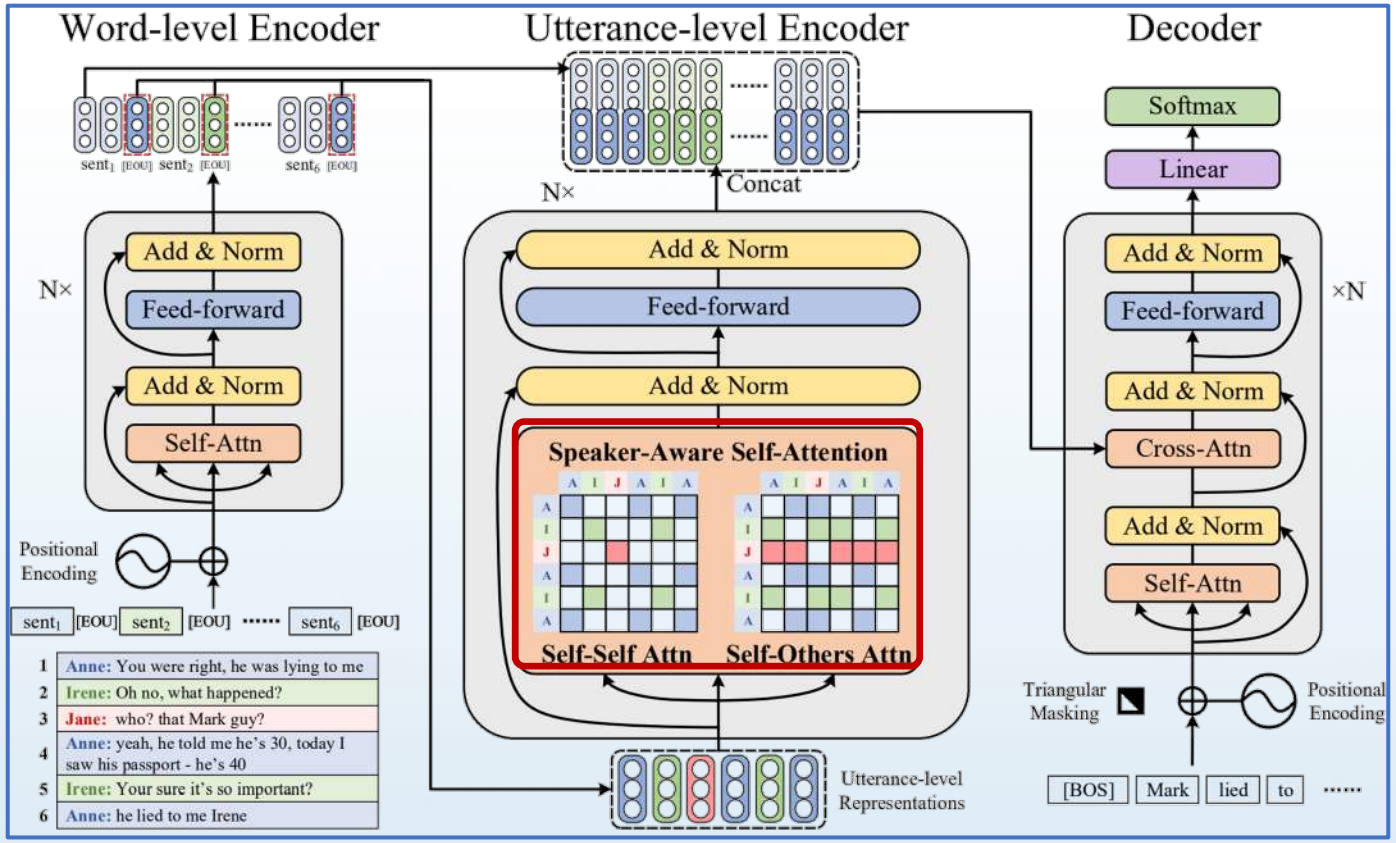
- 医生针对不同的症状进行询问



Topic-aware Pointer-Generator Networks for Summarizing Spoken Conversations ASRU2019

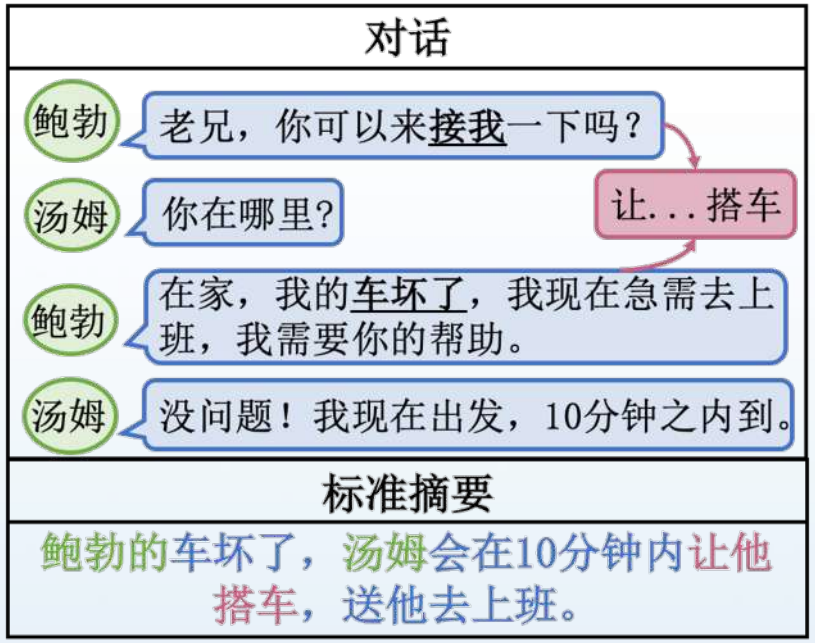
说话人建模

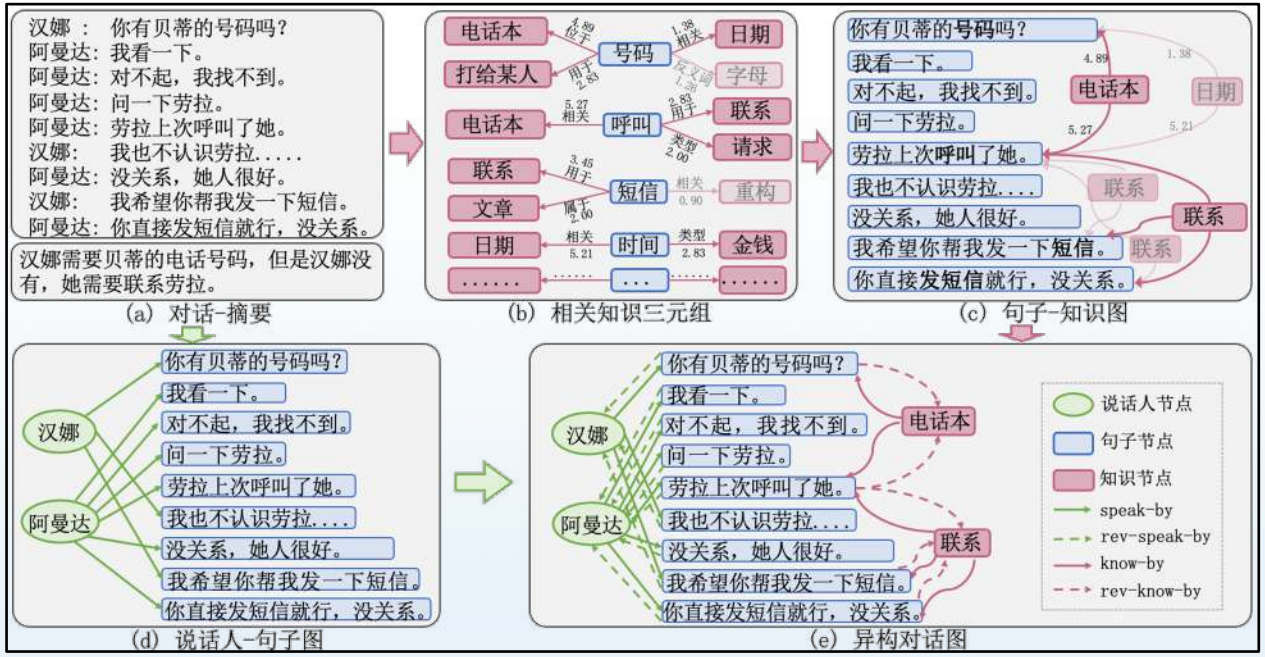
- 同一说话人之间的注意力机制（Self-Self Attn）
- 不同说话人之间的注意力机制（Self-Others Attn）



Hierarchical Speaker-Aware Sequence-to-Sequence Model for Dialogue Summarization ICASSP2021

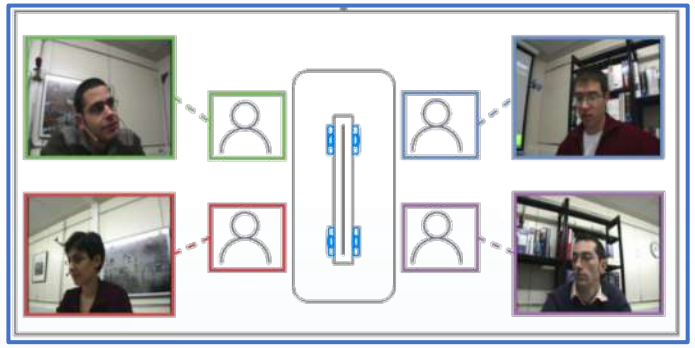
- 对话参与者通过自己的常识知识理解对话内容，做出回复

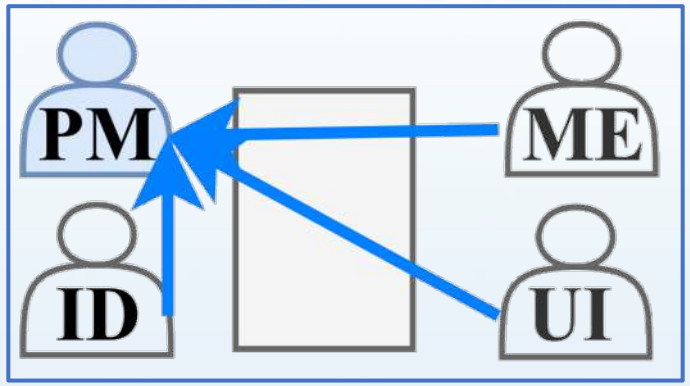


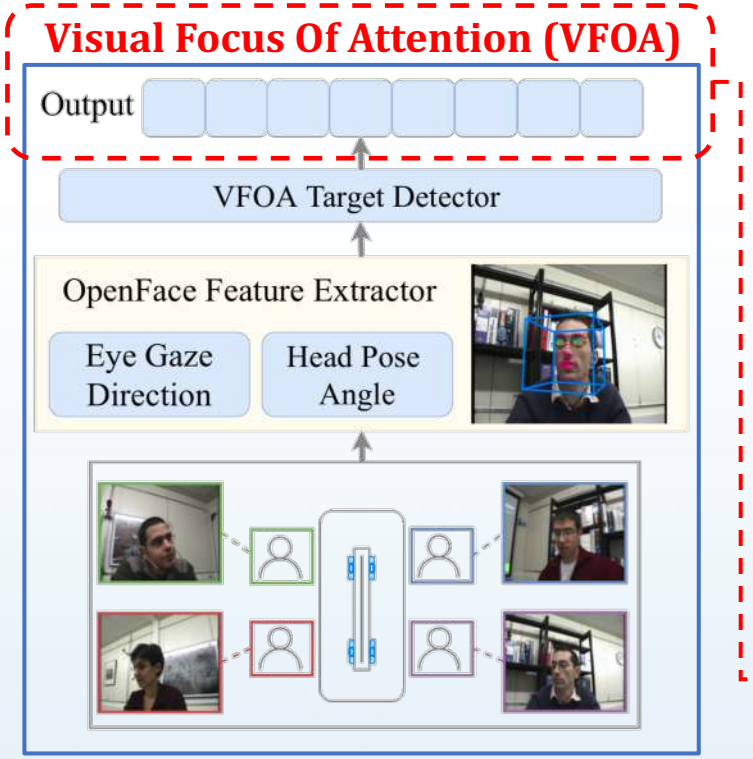


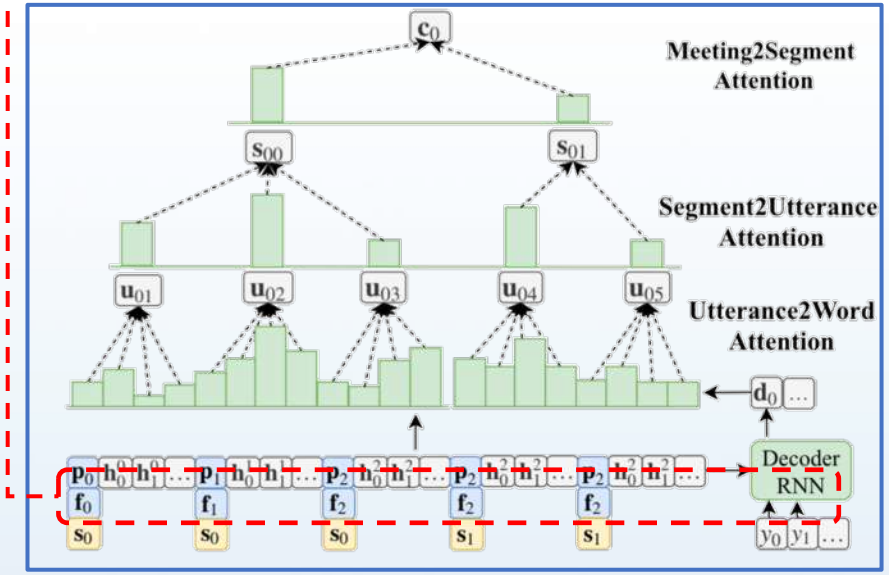
Incorporating Commonsense Knowledge into Abstractive Dialogue Summarization via Heterogeneous Graph Networks CCL2021

视觉信息







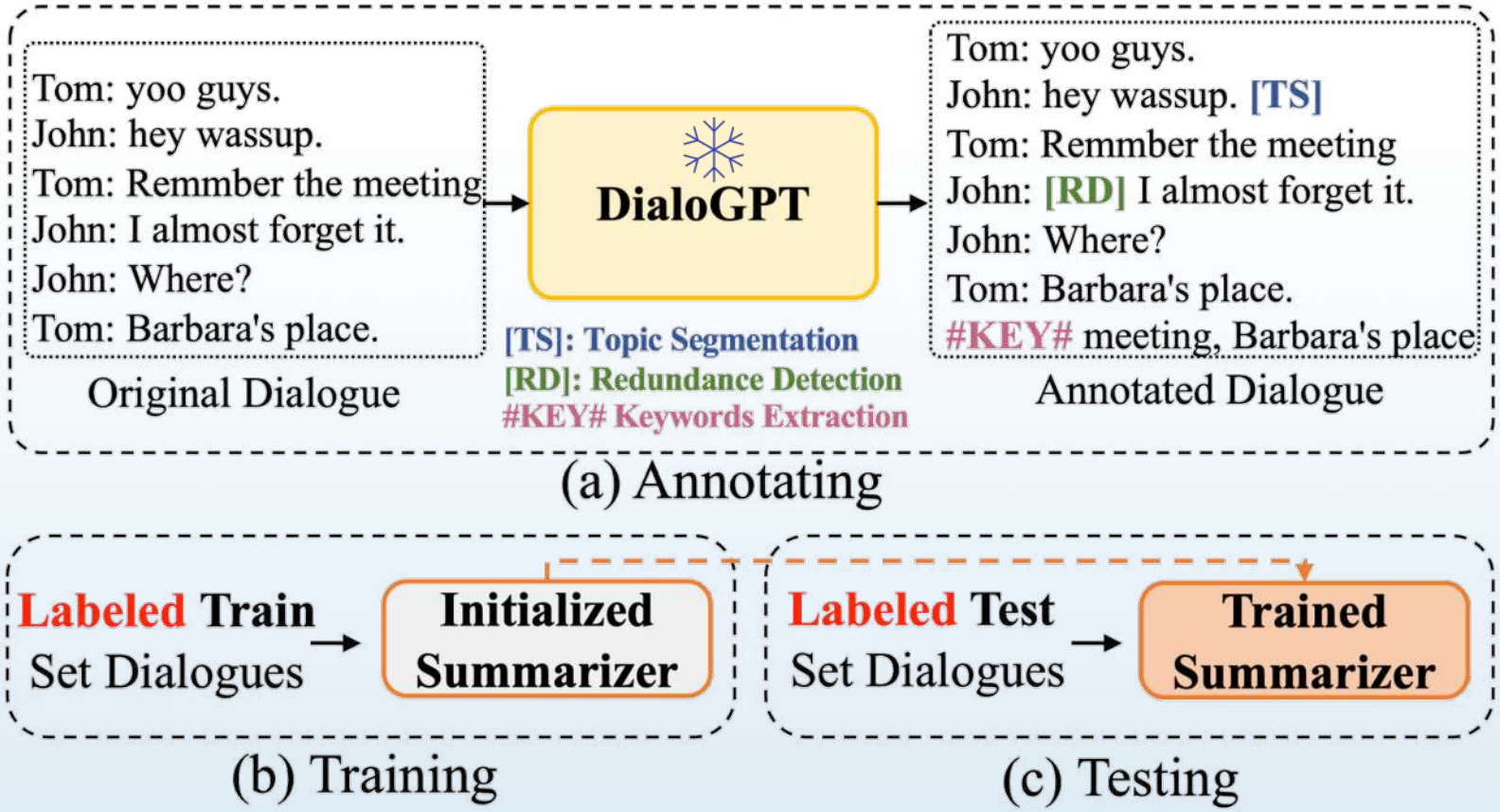


说话人被其他参与者注视的时间越长，该说话者的信息越重要。

Keep Meeting Summaries on Topic: Abstractive Multi-Modal Meeting Summarization ACL2019

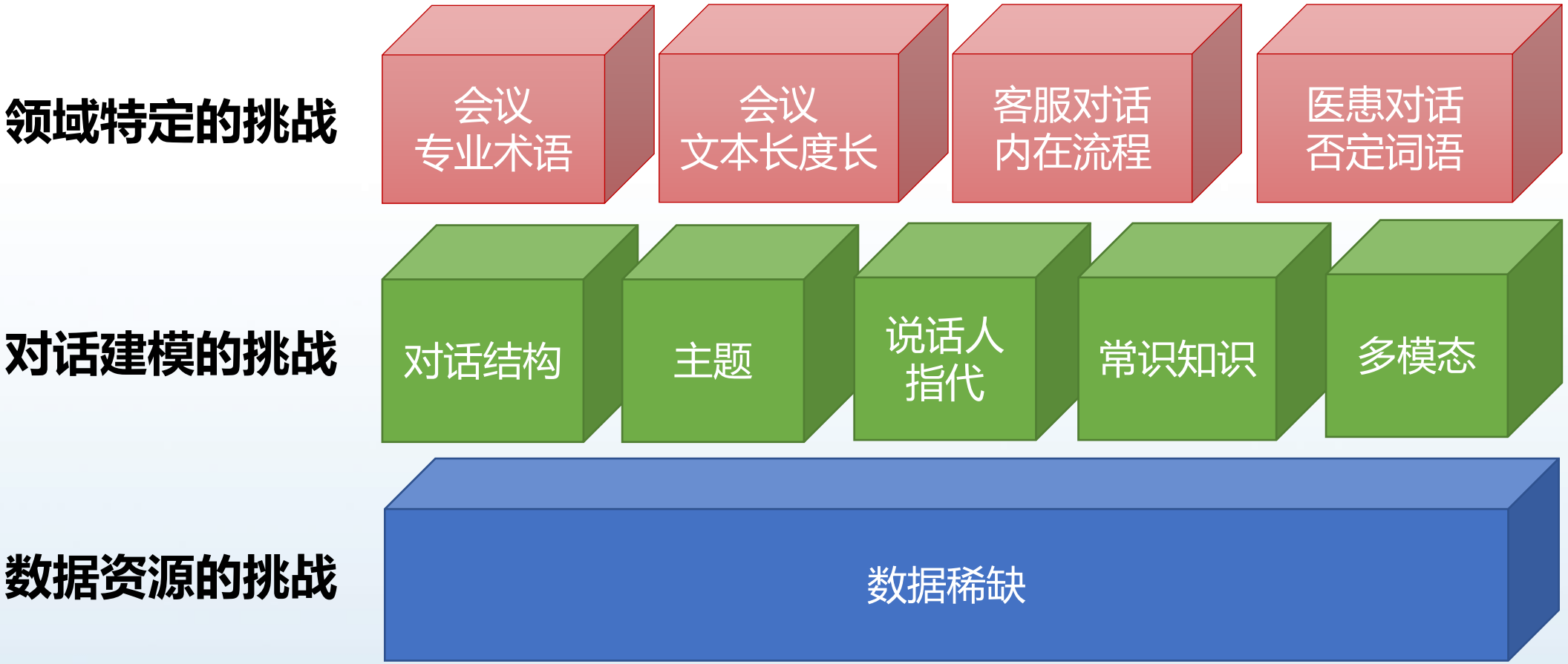
将预训练语言模型作为无监督标注器

- 关键词抽取、冗余句检测、主题分割



Language Model as an Annotator: Exploring DialoGPT for Dialogue Summarization ACL2021

领域特定的挑战


领域特定的挑战
会议
专业术语
会议
文本长度长
客服对话
内在流程
医患对话
否定词语
对话建模的挑战
对话结构
主题
说话人
指代
常识知识
多模态
数据资源的挑战
数据稀缺

会议中的领域术语

- 会议中常常出现领域的专业术语，专业术语往往是稀有词语

Start	End	Spoken Utterance
247.255	252.672	with Andreas' help um Andreas put together a sort of no frills recognizer which is uh
252.672	258.837	gender-dependent but like no adaptation, **no cross-word models, no trigrams - a bigram recognizer**
258.837	262.221	and that's trained on **Switchboard** which is telephone conversations.
263.983	267.154	and thanks to Don's help wh- who - Don took
267.154	270.431	the first meeting that Jane had transcribed
270.431	277.520	and um you know separated - used the individual channels we segmented it in- into the segments that Jane had used
277.520	279.952	and uh Don sampled that so -
281.374	289.611	um and then we ran up to I guess the first twenty minutes, up to **synch time** of one two zero zero so is that - that's twenty minutes or so?
289.611	296.601	Um yeah because I guess there's some, and Don can talk to Jane about this, there's some bug in **the actual synch time file** that

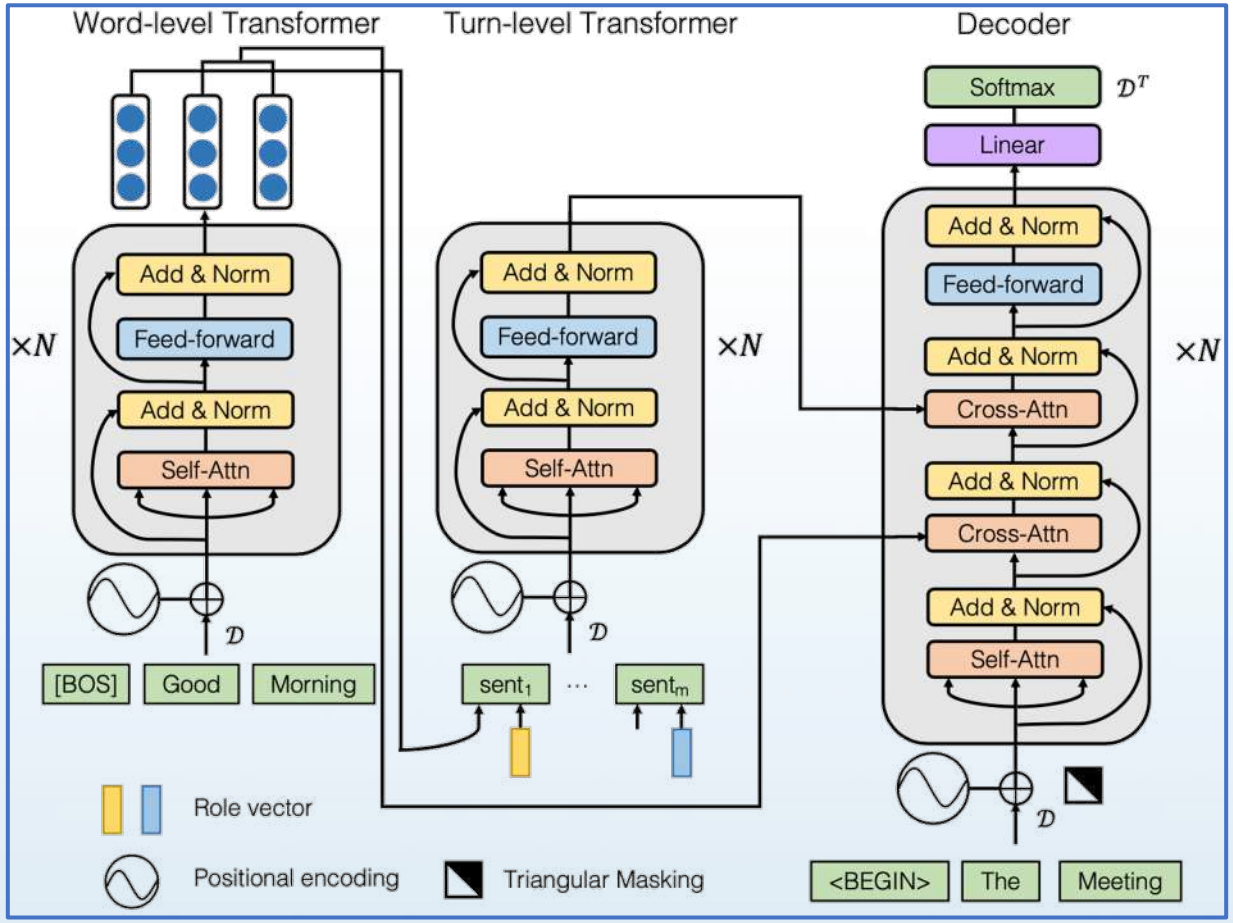
Table 1: A snippet of a human transcript that contains spoken utterances and their start/end times. Domain terminology is in bold.

Utterance with Jargon
she wanted to display the **stylized F_ zeroes,** I think they're called?
Utterance without Jargon
she wanted to display the **[MASK]** I think they're called?

How Domain Terminology Affects Meeting Summarization Performance COLING2020

会议文本长度长

层次化结构

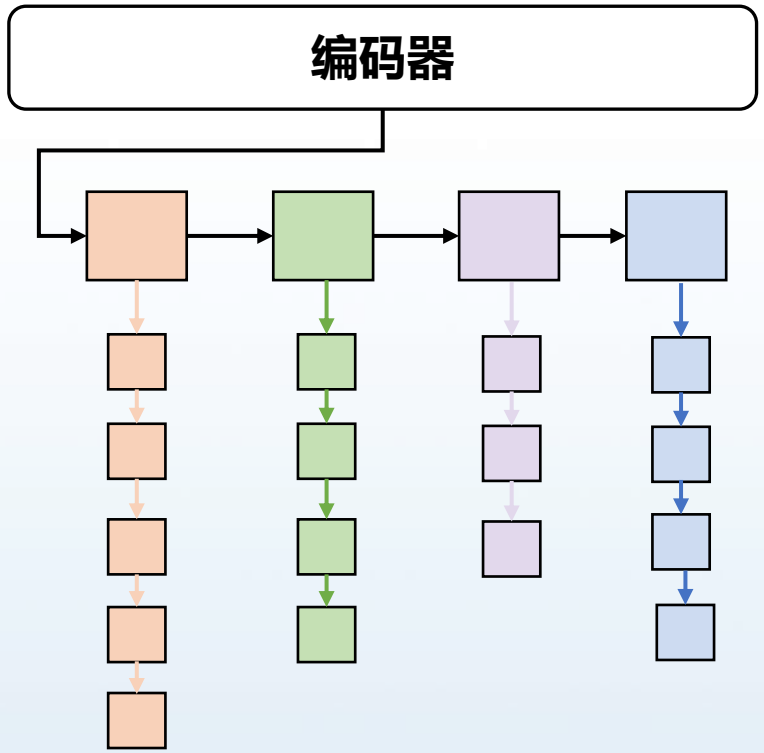


A Hierarchical Network for Abstractive Meeting Summarization with Cross-Domain Pretraining Findings of EMNLP2020

客服对话的内在结构

客服对话存在隐式的流程结构

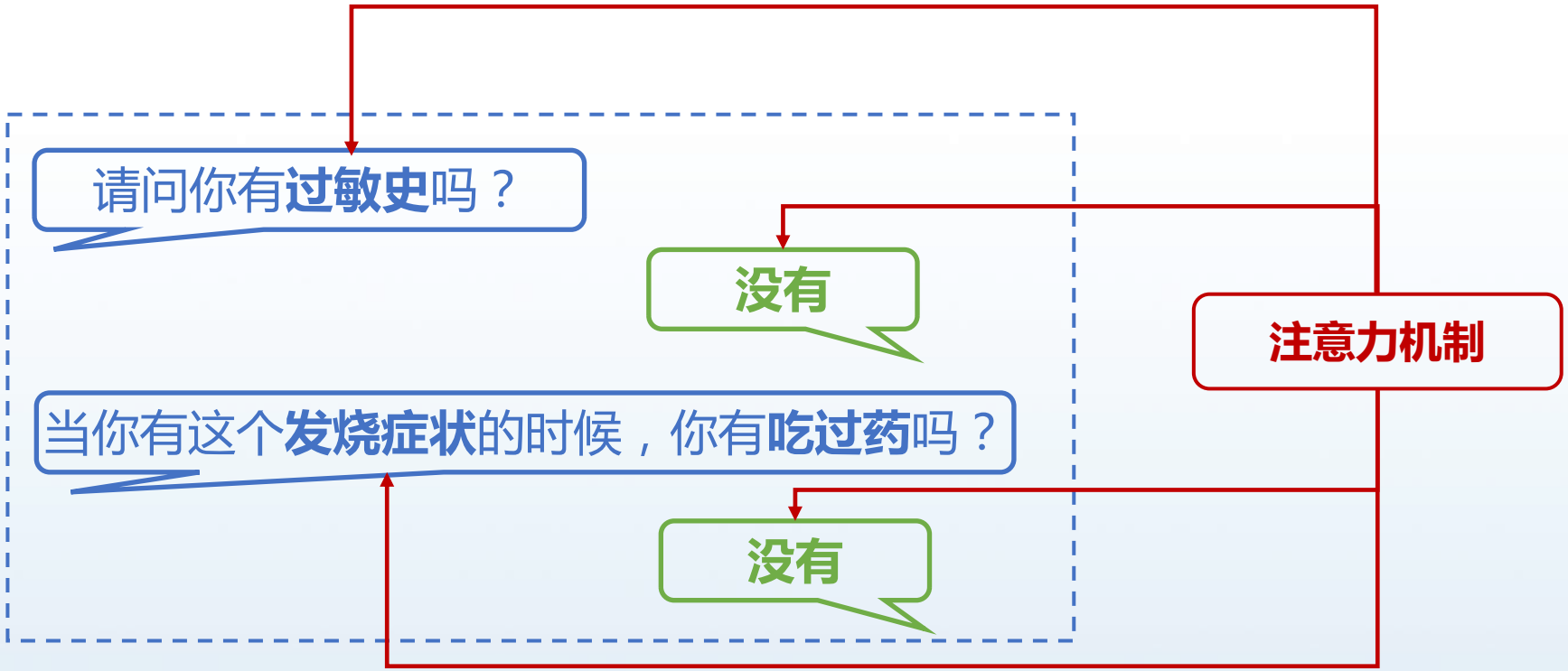
Dialogue
AGENT: Hello, what can I do for you? *USER*: What's the standard of electric vehicles for the Express. *AGENT*: Do you have a car? *AGENT*: Or are you going to buy a car? *USER*: I am hesitating which car to buy. One is Jianghuai EV Seven, the other is BYD YUAN. *AGENT*: OK, you can fulfill the table in this link (link info) with the type of vehicle you wish to check. We will give you feedback in seven days. *USER*: I have not bought yet. *USER*: Can you check it now? *AGENT*: I am quite sorry for that. A specialist on this issue will check it and call you back. *AGENT*: They will give a precise answer for your question. *USER*: OK. *AGENT*: Thanks for your understanding. What else can I do for you? *USER*: Nothing, thanks. Bye. *AGENT*: Thank you. Have a nice day.
Summary
The user's question was about the standard of EV car for the Express. He asked the standard to decide which car to buy. I told the user to fill in the type of the cars in our system and we would give feedback in seven days. The user approved the result. The user hung up.
Key point sequence
Question description ⟶ Solution ⟶ User approval ⟶ End



Automatic Dialogue Summary Generation for Customer Service KDD2019

医患对话中的否定用语

- 医患对话中的否定回答需要额外注意



Dr.Summarize: Global Summarization of Medical Dialogue by Exploiting Local Structures Findings of EMNLP2020

阶段性总结

领域特定的挑战

- 会议 专业术语
- 会议 文本长度长
- 客服对话 内在流程
- 医患对话 否定词语

对话建模的挑战

- 对话结构
- 主题
- 说话人 指代
- 常识知识
- 多模态

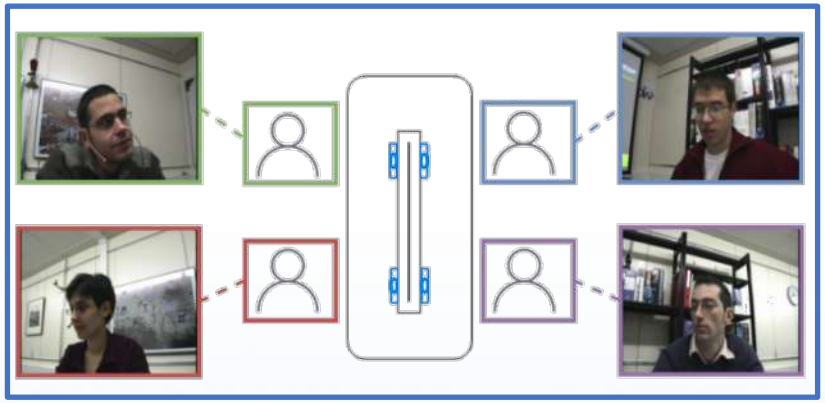
数据资源的挑战

- 数据稀缺

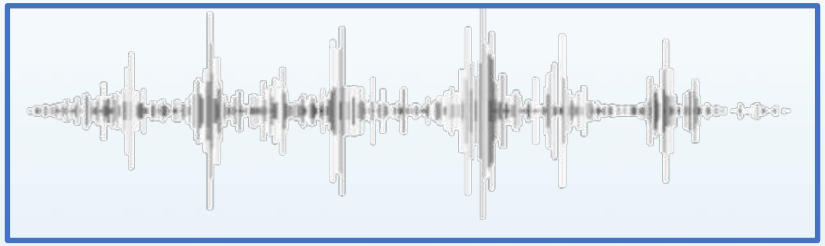
03 对话摘要的未来

未来趋势：多模态对话摘要

同步的多模态



视频



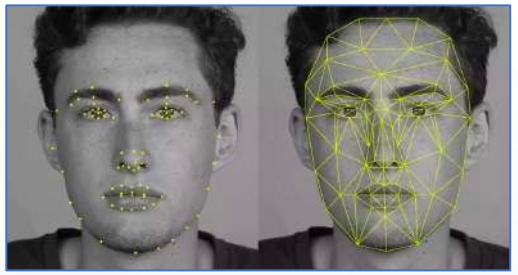
音频

We're developing a remote control which you probably already know.

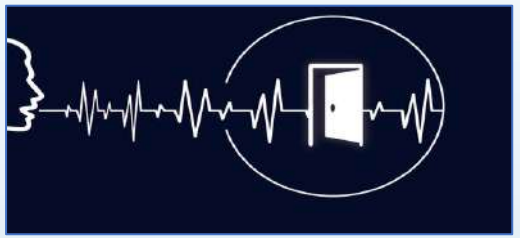
文本

潜在问题

数据隐私性



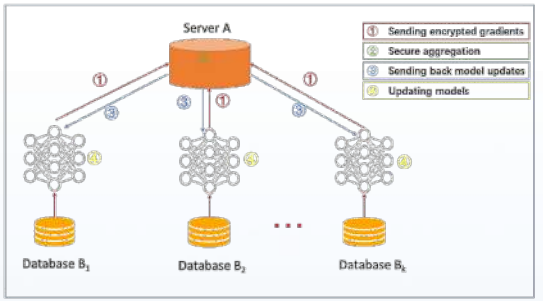
面部特征



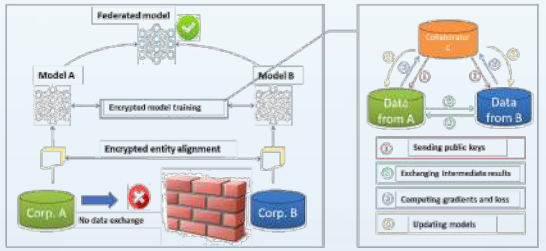
声纹特征

可行方案

联邦学习



横向联邦



纵向联邦

未来趋势：多模态对话摘要

- 异步的多模态
  - 文本
  - 图片（静态）
  - 表情包（静态+动态）
  - 视频（动态）
  - 语音

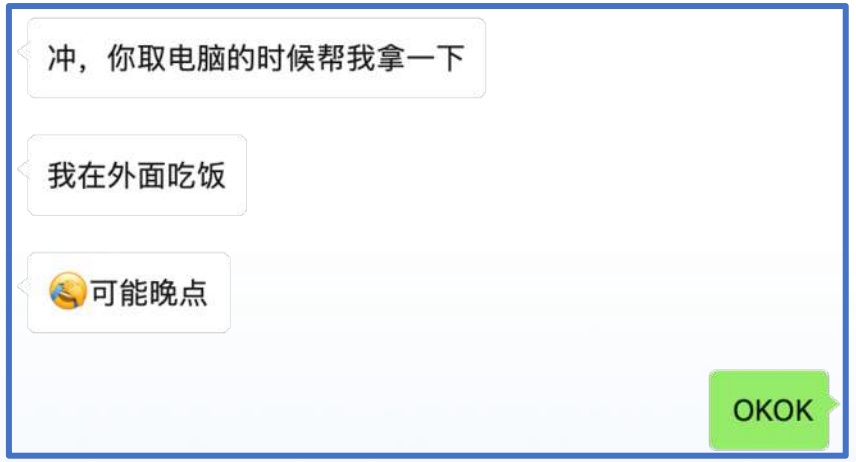
- 相关方向延伸
  - 情感分析

未来趋势：多领域对话摘要

- 对话数据形式不一，各有特点。

电影对话

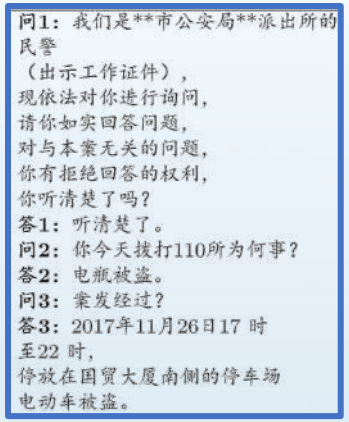
采访



闲聊

邮件

会议



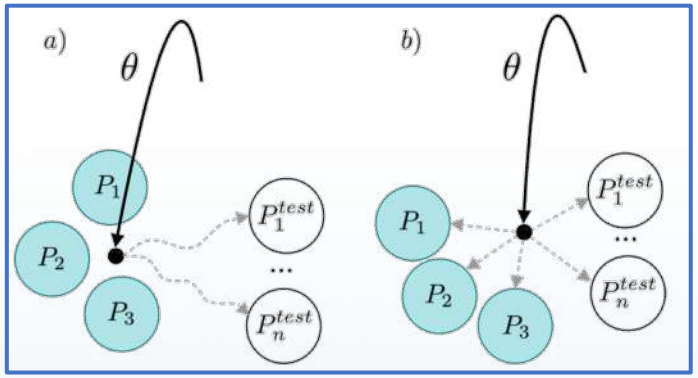
派出所报警

未来趋势：多领域对话摘要

- 现有数据集规模不一，领域多样

ID	Dataset	Language	Domain	Train	Valid	Test
1	SAMSum	English	chat	14732	818	819
2	SumTitle	English	movie	21469	-	290
3	AMI	English	meeting	97	20	20
4	MediaSum	English	interview	463596	-	-
5	BC3	English	email	40	-	-
6	CRD3	English	TV show	34243	-	-
7	LCSPIRT	Chinese	police	38500	-	-

可行方案



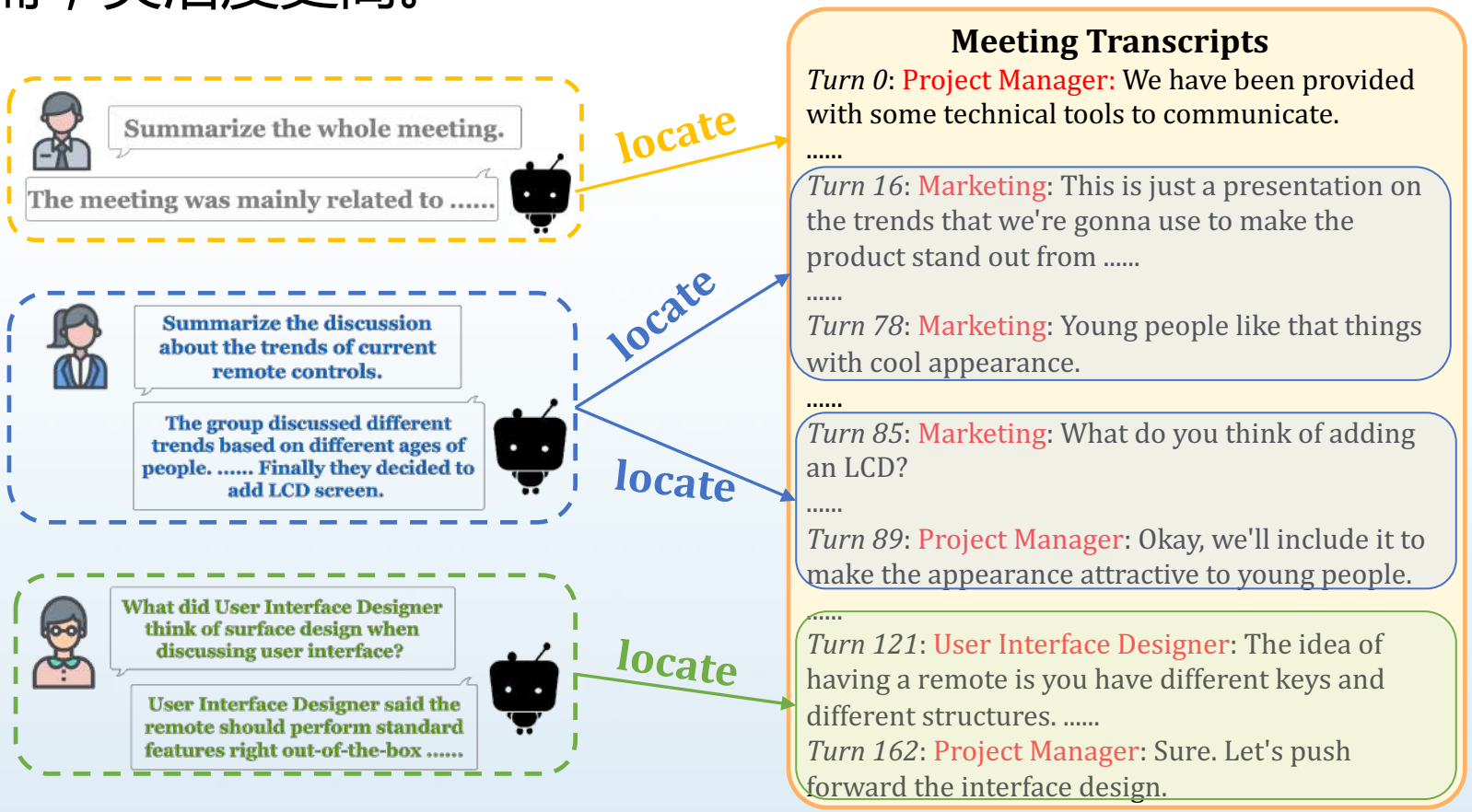
元学习

PLMs

预训练语言模型

未来趋势：新的任务形式（1）

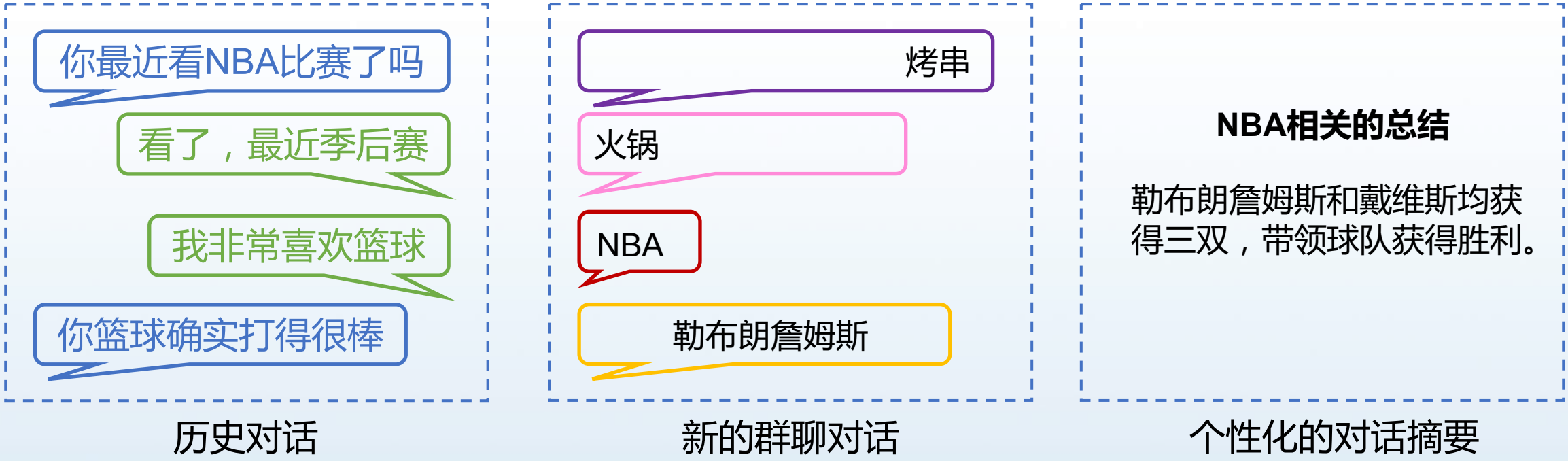
- 基于查询的对话摘要（Query-based dialogue summarization）
  - 各取所需，灵活度更高。



QMSum: A New Benchmark for Query-based Multi-domain Meeting Summarization. NAACL21

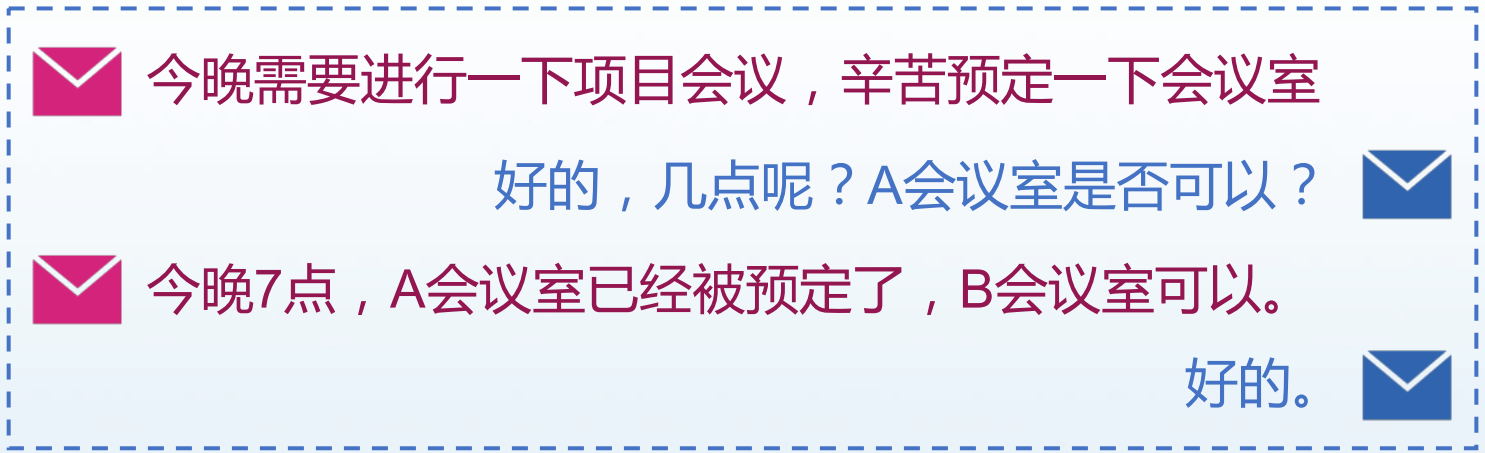
未来趋势：新的任务形式（2）

- 个性化的对话摘要（Personalized dialogue summarization）
  - 自动推测感兴趣的主题。

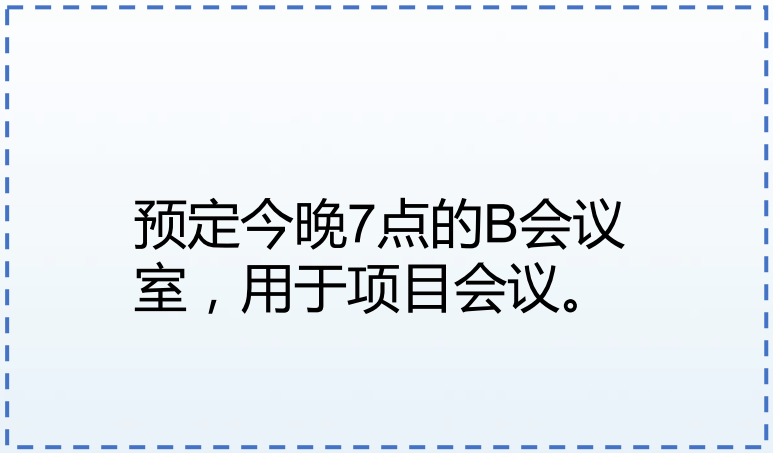


未来趋势：新的任务形式（3）

- 目标特定的对话摘要（Task-specific dialogue summarization）
  - 邮件中的TODO、医疗对话中的诊断……



邮件交流



目标特定的摘要

04 总结

总结

- 对话摘要在对话系统和文本生成技术发展的基础之上崭露头角。
- 众多数据集的提出进一步推动了对话摘要领域的发展。
- 对话摘要的未来发展需要更加落地的任务作为基础。
- 多模态对话摘要和多领域对话摘要可能成为下一个研究趋势。
- 摘要论文列表： https://github.com/xcfcode/Summarization-Papers

谢谢！