Dialogue Summarization

李昌群

目录

- 文档摘要
- 对话摘要挑战
- 总结

Summarization

- 摘要旨在将输入数据转换为包含关键信息的简短文本

阿什利杨据ESPN报道，根据消息源透露，曼联边锋阿什利一杨已与球队达成一致，准备续约3年。杨与曼联的合同今年将是最后一年，上赛季末时就已初步展开续约谈判，不过在7月时他表示自己还未与球队达成一致。据消息源透露，杨与球队将在本周稍晚时候正式签署合约。杨2011年从阿斯顿维拉转投曼联，虽然屡有球队对他有意，但他一直坚定不移地留在曼联。

→

据外媒报道，曼联边锋阿什利一杨已与球队达成一致续约3年，将于本周内正式宣布留守。

摘要的分类

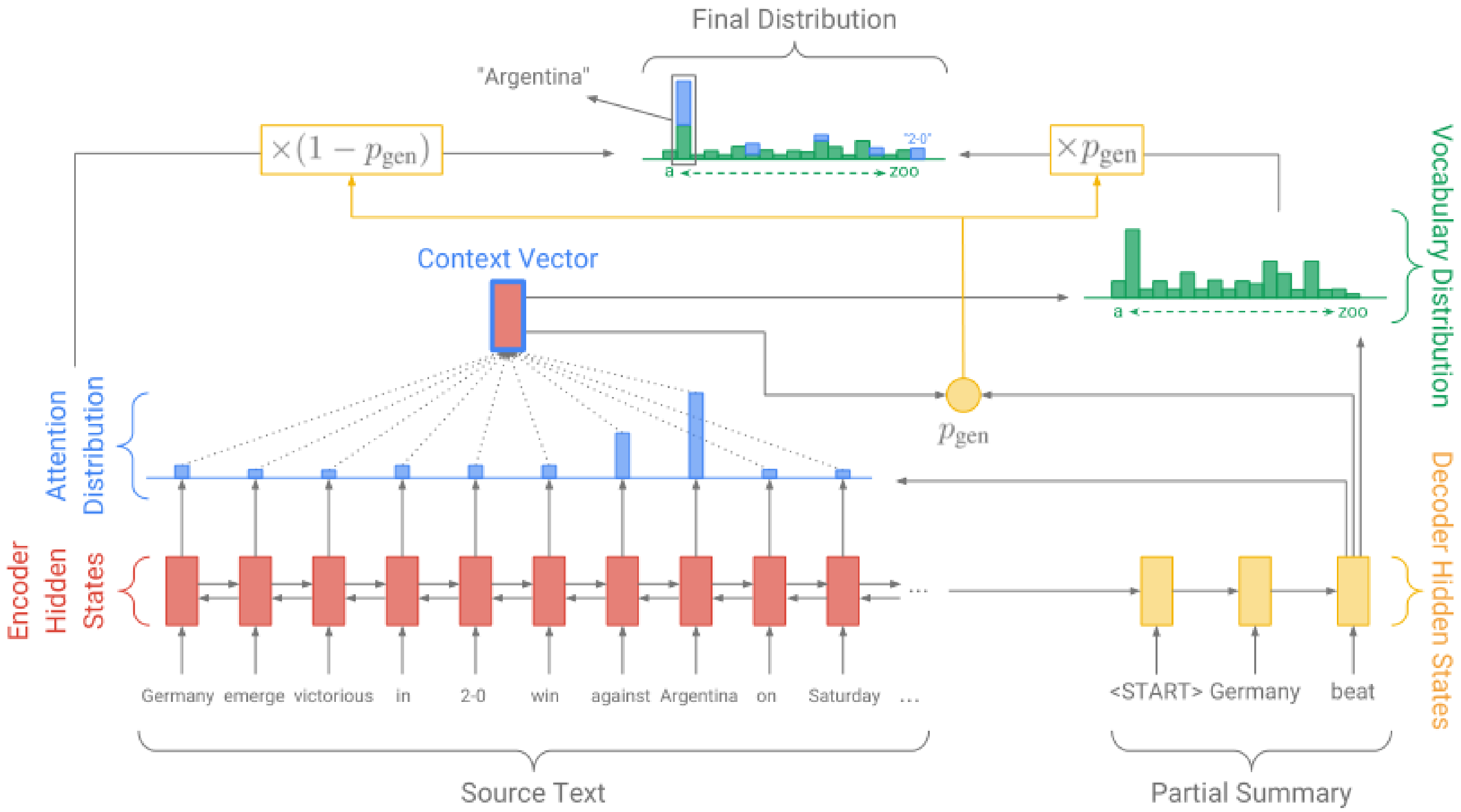
评价指标

- Rouge：计算机器摘要和标准摘要之间的n-gram重叠程度
  - Rouge-1/2，即比较两段摘要uni-gram, bi-gram的召回率
  - Rouge-L，计算两段摘要文本的最长公共子序列

- BERTScore：语义相似性

- 自动评价与人工评价的统计相关性（判断评价指标的好坏）
  - Pearson相关系数、Spearman相关系数

Abstractive Summarization经典工作



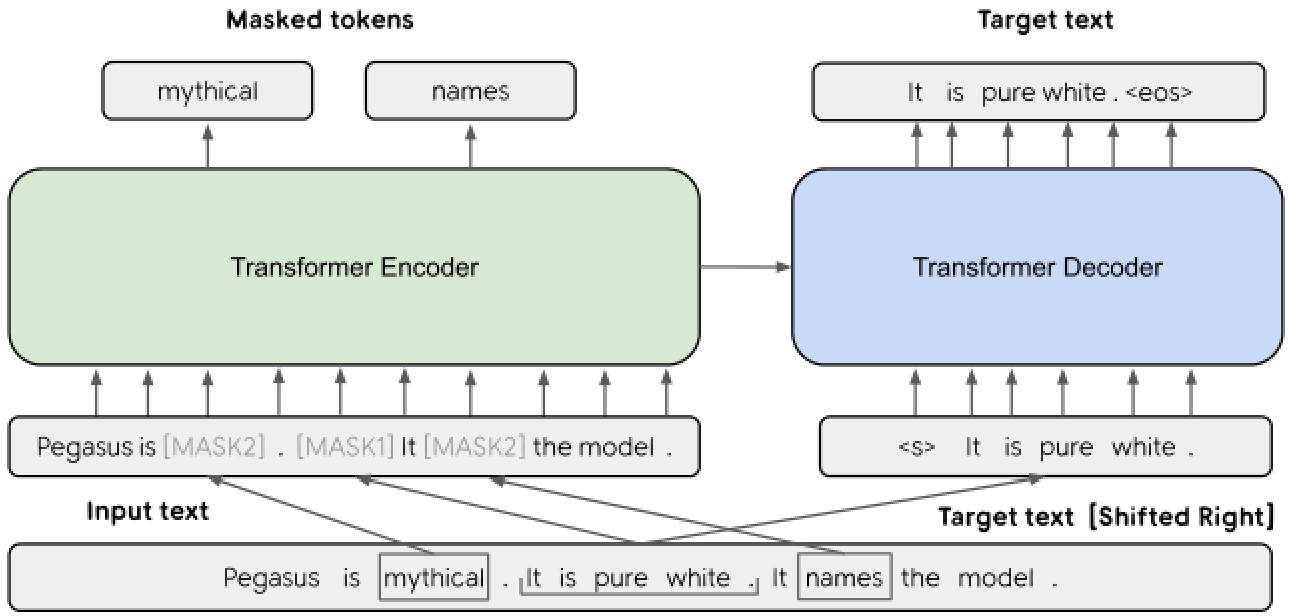
seq2seq存在的问题：1. 事实性错误 2. 无法处理词汇外(OOV)单词3. 生成重复的句子

Hybrid Pointer-Generator Network：1. 通过pointer从源文本复制单词，同时计算生成概率Pgen保持生成新单词的能力；2. 并提出了coverage机制来消除重复

ACL 2017 Get To The Point: Summarization with Pointer-Generator Networks

NIPS 2015 Pointer Networks

Task-Specific Pretraining



task-specific pretraining for summarization

论文提出新的预训练目标(Gap Sentences Generation)，即从文档中Mask重要句子并从文档的其余部分生成这些gap-sentences。

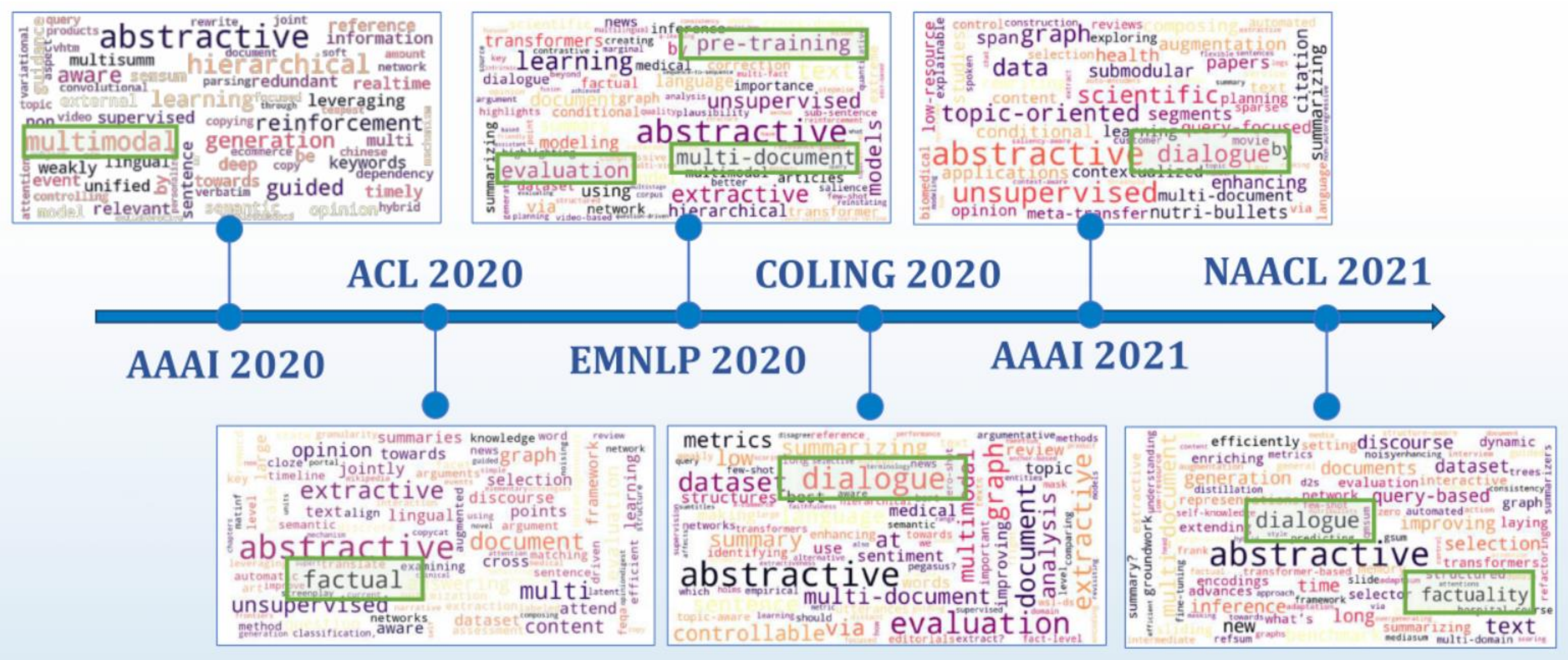
特点：Mask多个完整的句子，而不是较小的连续文本；并根据重要性来选择句子，而不是进行随机选择

1. task-agnostic pretraining （the left-to-right language modelling objective）
2. reconstructing the corrupted input text

ICML 2020 PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization

R1/R2/RL	XSum	CNN/DailyMail	Gigaword
BERTShare (Rothe et al., 2019)	38.52/16.12/31.13	39.25/18.09/36.45	38.13/19.81/35.62
MASS (Song et al., 2019)	39.75/17.24/31.95	42.12/19.50/39.01	38.73/19.71/35.96
UniLM (Dong et al., 2019)	-	43.33/20.21/40.51	38.45/19.45/35.75
BART (Lewis et al., 2019)	45.14/22.27/37.25	**44.16**/21.28/40.90	-
T5 (Raffel et al., 2019)	-	43.52/**21.55**/40.69	-
$\text{PEGASUS}_{\text{LARGE}}$ (C4)	45.20/22.06/36.99	43.90/21.20/40.76	38.75/**19.96/36.14**
$\text{PEGASUS}_{\text{LARGE}}$ (HugeNews)	**47.21/24.56/39.25**	**44.17/21.47/41.11**	**39.12/19.86/36.24**

摘要近两年的发展



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

对话摘要示例

- 对话摘要关注对话类文本
  - 会议(Meeting)，闲聊(Chat)、邮件(Email)、客服对话(Customer Service)、医患对话(Medical Dialogue)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。		

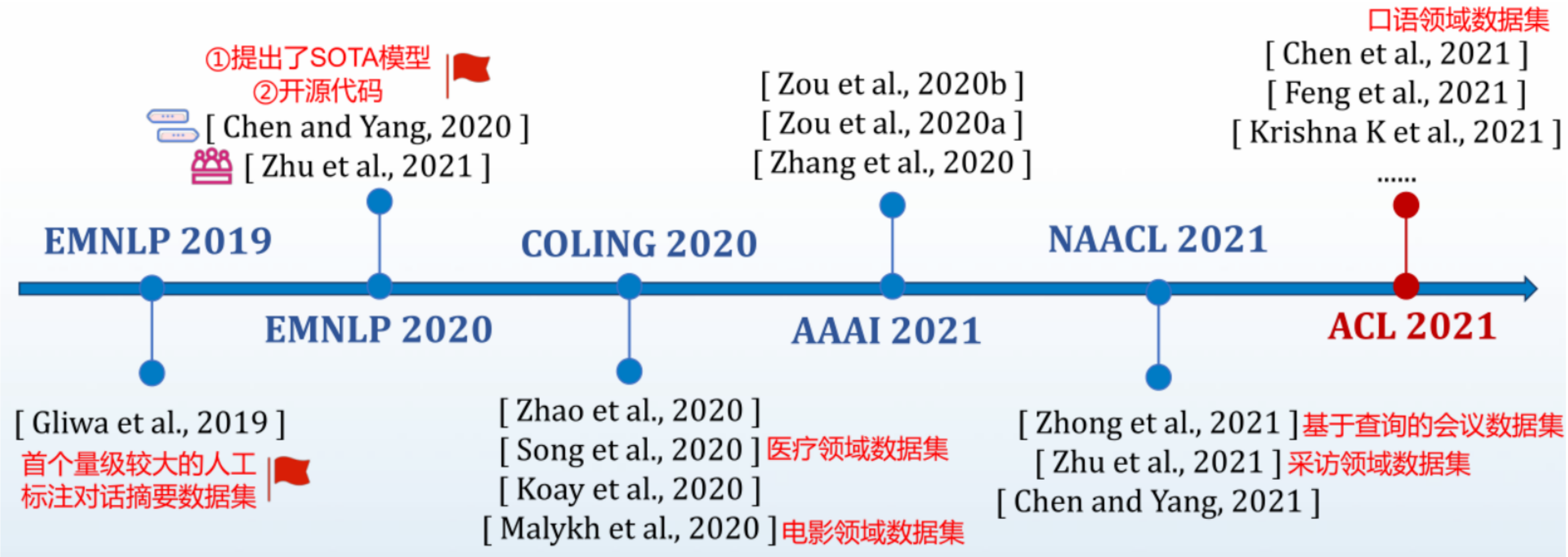
Meeting Minutes
会议纪要

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。		

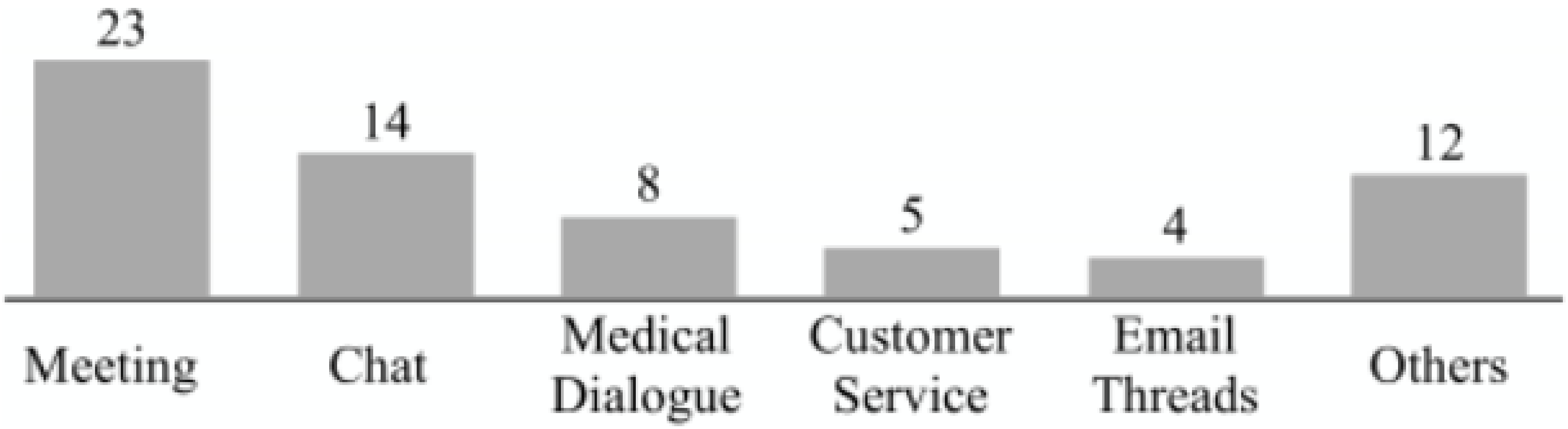
医患对话		
医生	：	你最近有肿胀吗?
患者	：	时有时无。
医生	：	我知道了，什么时候开始的?
患者	：	大约在三周之前。
标准摘要		
肿胀：大约三周之前开始，症状时有时无。		

SOAP
主观描述、客观观察、医生诊断、治疗计划

对话摘要的发展脉络

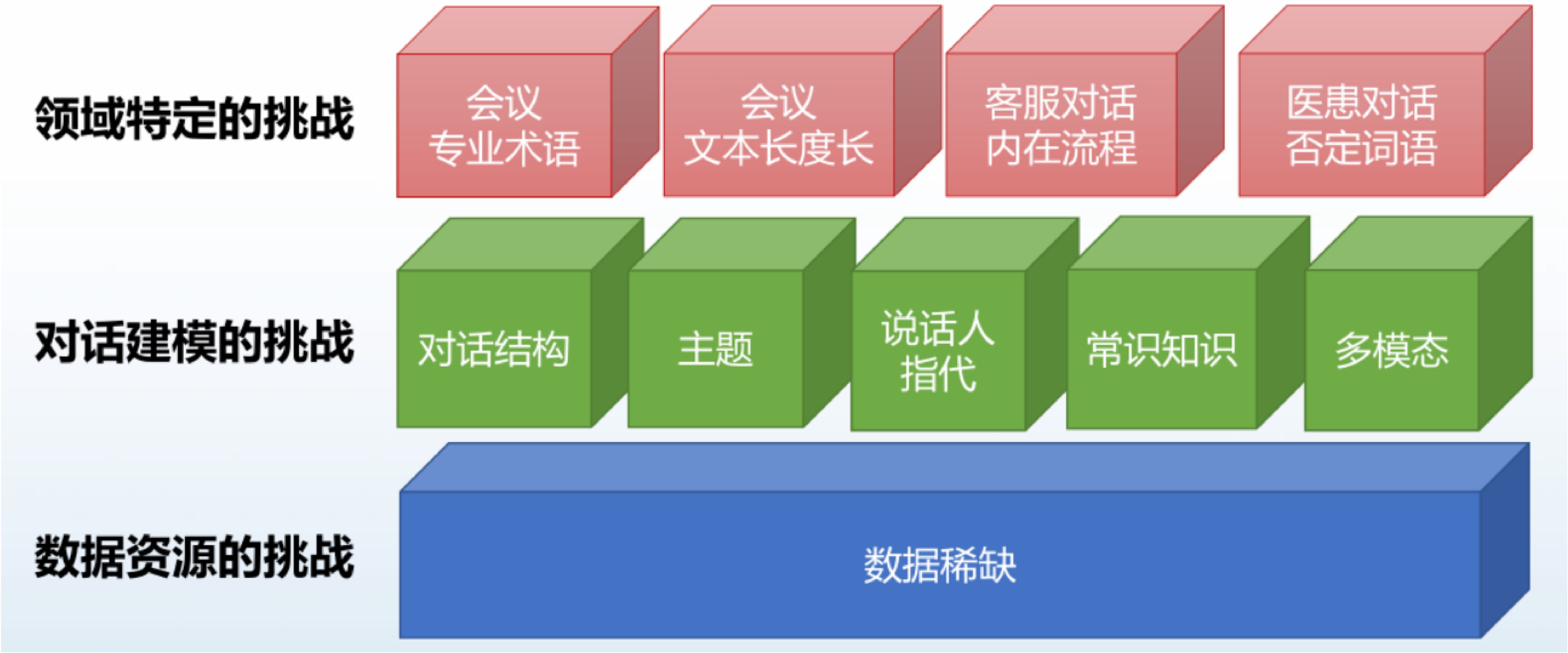

①提出了SOTA模型
②开源代码
[Chen and Yang, 2020]
[Zhu et al., 2021]
EMNLP 2019
[Gliwa et al., 2019]
首个量级较大的人工
标注对话摘要数据集
EMNLP 2020
COLING 2020
[Zhao et al., 2020]
[Song et al., 2020] 医疗领域数据集
[Koay et al., 2020]
[Malykh et al., 2020] 电影领域数据集
[Zou et al., 2020b]
[Zou et al., 2020a]
[Zhang et al., 2020]
AAAI 2021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

近5年不同类型论文数量



A Survey on Dialogue Summarization: Recent Advances and New Frontiers

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

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

CNN-DailyMail：311k

Xsum：227k

新的数据集

Name	Domain	Language
ICSI [Janin *et al.*, 2003]	Meeting	English
AMI [Carletta *et al.*, 2005]	Meeting	English
QMSum [Zhong *et al.*, 2021]	Meeting	English
SUMMSCREEN [Chen *et al.*, 2021a]	TV Show	English
CRD3 [Rameshkumar and Bailey, 2020]	TV Show	English
SAMSum [Gliwa *et al.*, 2019]	Chat	English
GupShup [Mehnaz *et al.*, 2021]	Chat	Hindi-English
ADSC [Misra *et al.*, 2015]	Debate	English
[Song *et al.*, 2020]	Medical	Chinese
SumTitles [Malykh *et al.*, 2020]	Movie	English
LCSPIRT [Xi *et al.*, 2020]	Police	Chinese
MEDIASUM [Zhu *et al.*, 2021]	Interview	English
DIALOGSUM [Chen *et al.*, 2021b]	Spoken	English
EMAILSUM [Zhang *et al.*, 2020b]	Email	English
ConvoSumm [Fabbri *et al.*, 2021]	Mix	English

A Survey on Dialogue Summarization: Recent Advances and New Frontiers

借助预训练：领域外数据

- Using news summarization datasets
  - CNN/DailyMail(311k)、NYT(104k)、Xsum(227k)

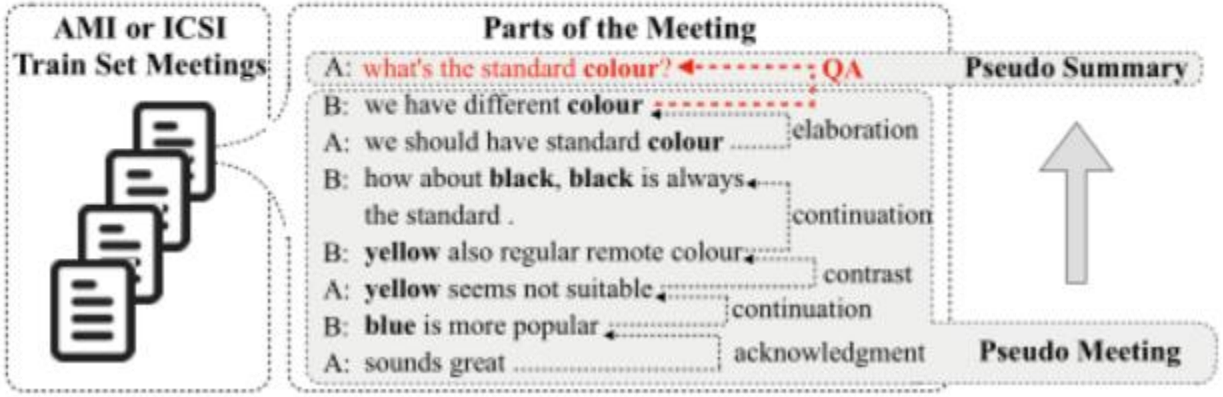
Model	ROUGE-1	R-2	R-SU4
AMI			
HMNet	**53.0**	**18.6**	**24.9**
−pretrain	48.7	18.4	23.5
−role vector	47.8	17.2	21.7
−hierarchy	45.1	15.9	20.5
ICSI			
HMNet	**46.3**	**10.6**	**19.1**
−pretrain	42.3	**10.6**	17.8
−role vector	44.0	9.6	18.2
−hierarchy	41.0	9.3	16.8

Table 3: Ablation study of HMNet.

A Hierarchical Network for Abstractive Meeting Summarization with Cross-Domain Pretraining Findings of EMNLP 2020

借助预训练：领域内数据

- Pseudo-summarization Corpus Construction
  - “问题”会引起“讨论”，“问题”包含了“讨论”得核心内容



	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 IJCAI 2021

	Model	AMI R-1	AMI R-2	AMI R-L	ICSI R-1	ICSI R-2	ICSI R-L
Extractive	TextRank [Mihalcea and Tarau, 2004]	35.19	6.13	15.70	30.72	4.69	12.97
	SummaRunner [Nallapati *et al.*, 2017]	30.98	5.54	13.91	27.60	3.70	12.52
Abstractive	UNS [Shang *et al.*, 2018]	37.86	7.84	13.72	31.73	5.14	14.50
	Pointer-Generator [See *et al.*, 2017]	42.60	14.01	22.62	35.89	6.92	15.67
	HRED [Serban *et al.*, 2016]	49.75	18.36	23.90	39.15	7.86	16.25
	Sentence-Gated [Goo and Chen, 2018]	49.29	19.31	24.82	39.37	9.57	17.17
	TopicSeg [Li *et al.*, 2019]	51.53	12.23	25.47	-	-	-
	HMNet [Zhu *et al.*, 2020]	52.36	18.63	24.00	**45.97**	10.14	18.54
Ours	DDAMS	51.42	20.99	24.89	39.66	10.09	17.53
	DDAMS + DDADA	**53.15**	**22.32**	**25.67**	40.41	**11.02**	**19.18**
	DDAMS + DDADA (w/o fine-tune)	28.35	4.67	14.92	25.94	4.18	13.92

Dialogue structure： Dialogue Act

- 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	

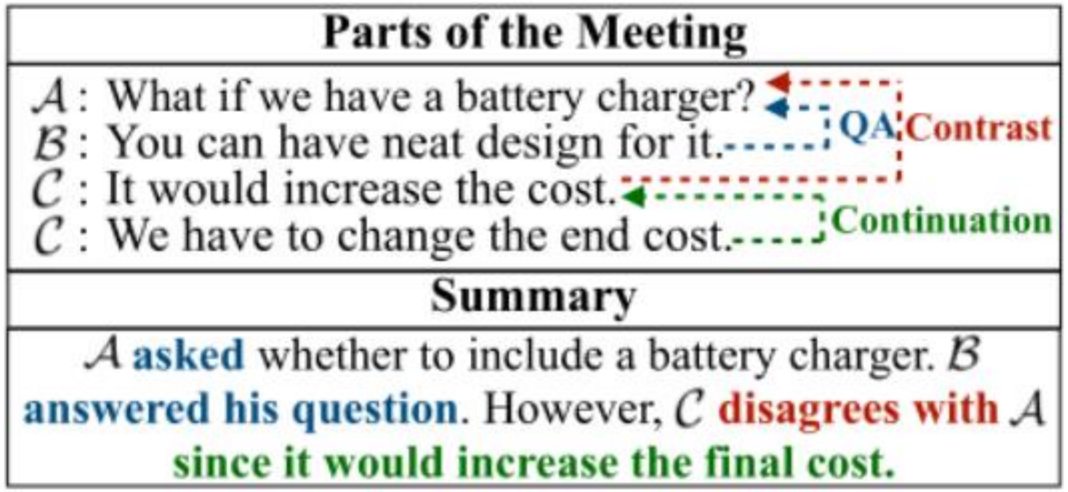
Fig. 1. A dialogue instance in the dataset built from the AMI meeting corpus.

- Multi-task learning

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

Dialogue structure： Dialogue Discourse

- 对话篇章结构(Dialogue Discourse)指示了句子之间的交互关系

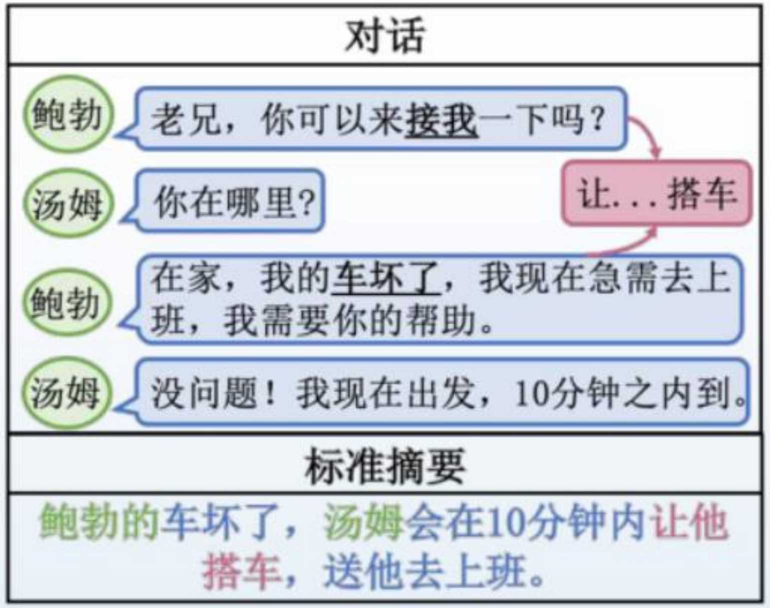


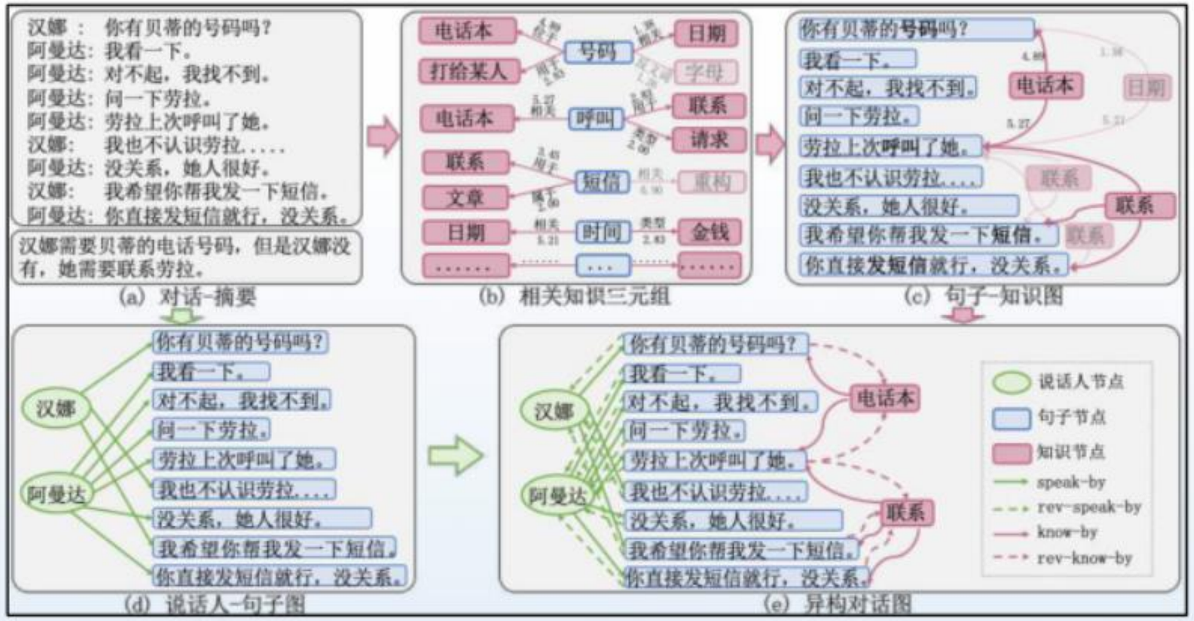
Dialogue discourse-aware graph model and data augmentation for meeting summarization IJCAI 2021

	Model	AMI R-1	AMI R-2	AMI R-L	ICSI R-1	ICSI R-2	ICSI R-L
Extractive	TextRank [Mihalcea and Tarau, 2004]	35.19	6.13	15.70	30.72	4.69	12.97
	SummaRunner [Nallapati *et al.*, 2017]	30.98	5.54	13.91	27.60	3.70	12.52
Abstractive	UNS [Shang *et al.*, 2018]	37.86	7.84	13.72	31.73	5.14	14.50
	Pointer-Generator [See *et al.*, 2017]	42.60	14.01	22.62	35.89	6.92	15.67
	HRED [Serban *et al.*, 2016]	49.75	18.36	23.90	39.15	7.86	16.25
	Sentence-Gated [Goo and Chen, 2018]	49.29	19.31	24.82	39.37	9.57	17.17
	TopicSeg [Li *et al.*, 2019]	51.53	12.23	25.47	-	-	-
	HMNet [Zhu *et al.*, 2020]	52.36	18.63	24.00	**45.97**	10.14	18.54
Ours	DDAMS	51.42	20.99	24.89	39.66	10.09	17.53
	DDAMS + DDADA	**53.15**	**22.32**	**25.67**	40.41	**11.02**	**19.18**
	DDAMS + DDADA (w/o fine-tune)	28.35	4.67	14.92	25.94	4.18	13.92

Commonsense Knowledge

- 对话参与者通过自己的常识知识理解对话内容，做出回复





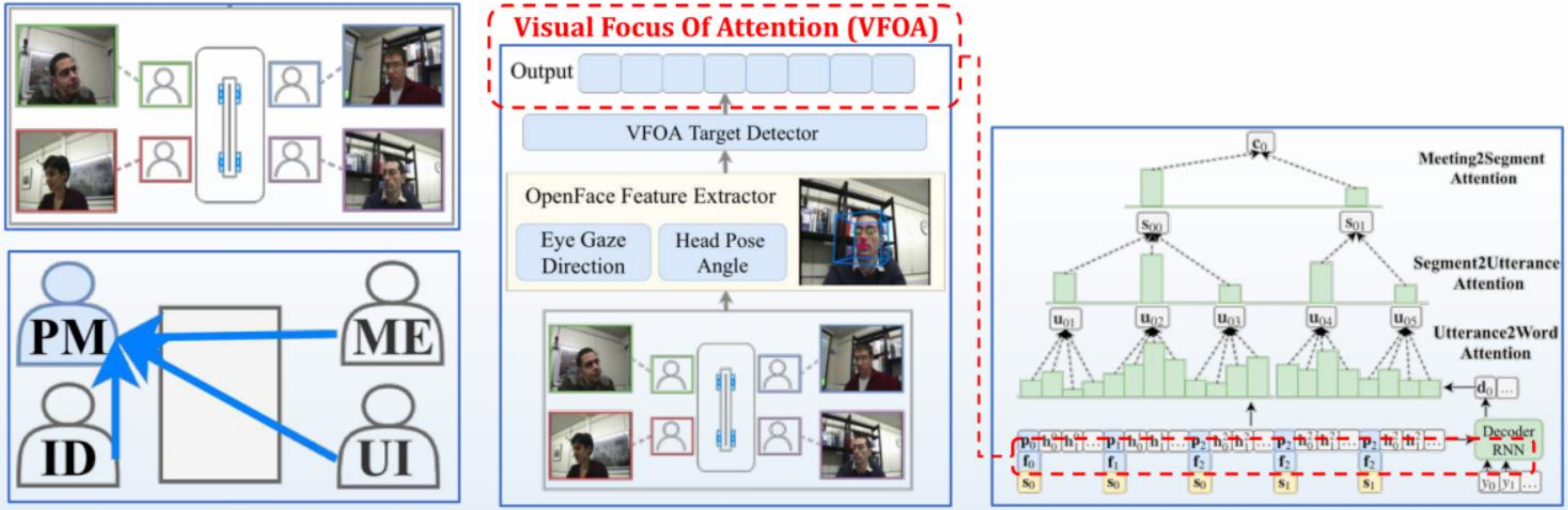
Incorporating Commonsense Knowledge into Abstractive Dialogue Summarization via Heterogeneous Graph Networks CCL 2021

Model	R-1	R-2	R-L
Extractive Methods			
LONGEST-3	32.46	10.27	29.92
TextRank [Mihalcea and Tarau, 2004]	29.27	8.02	28.78
Abstractive Methods			
DynamicConv [Wu *et al.*, 2019]	33.69	10.88	30.93
Transformer [Vaswani *et al.*, 2017]	36.62	11.18	33.06
PGN [See *et al.*, 2017]	40.08	15.28	36.63
Fast Abs RL [Chen and Bansal, 2018]	41.95	18.06	39.23
D-HGN [Feng *et al.*, 2020b]	42.03	18.07	39.56
TGDGA [Zhao *et al.*, 2020]	43.11	19.15	40.49
Pre-trained Language Model-based Methods			
DialoGPT [Zhang *et al.*, 2020d]	39.77	16.58	38.42
UniLM [Dong *et al.*, 2019]	47.85	24.23	46.67
PEGASUS [Zhang *et al.*, 2020a]	50.50	27.23	49.32
BART [Lewis *et al.*, 2020]	52.98	27.67	49.06
S-BART [Chen and Yang, 2021]	50.70	25.50	48.08
FROST [Narayan *et al.*, 2021]	51.86	27.67	47.52
CODS [Wu *et al.*, 2021]	52.65	27.84	50.79
MV-BART [Chen and Yang, 2020]	53.42	27.98	49.97
BART($\mathcal{D}_{\text{ALL}}$) [Feng *et al.*, 2021]	53.70	28.79	50.81

SAMSum (Chat)

Multi-Modal

- Visual Information



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。

Domain-specific challenges--Meeting

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

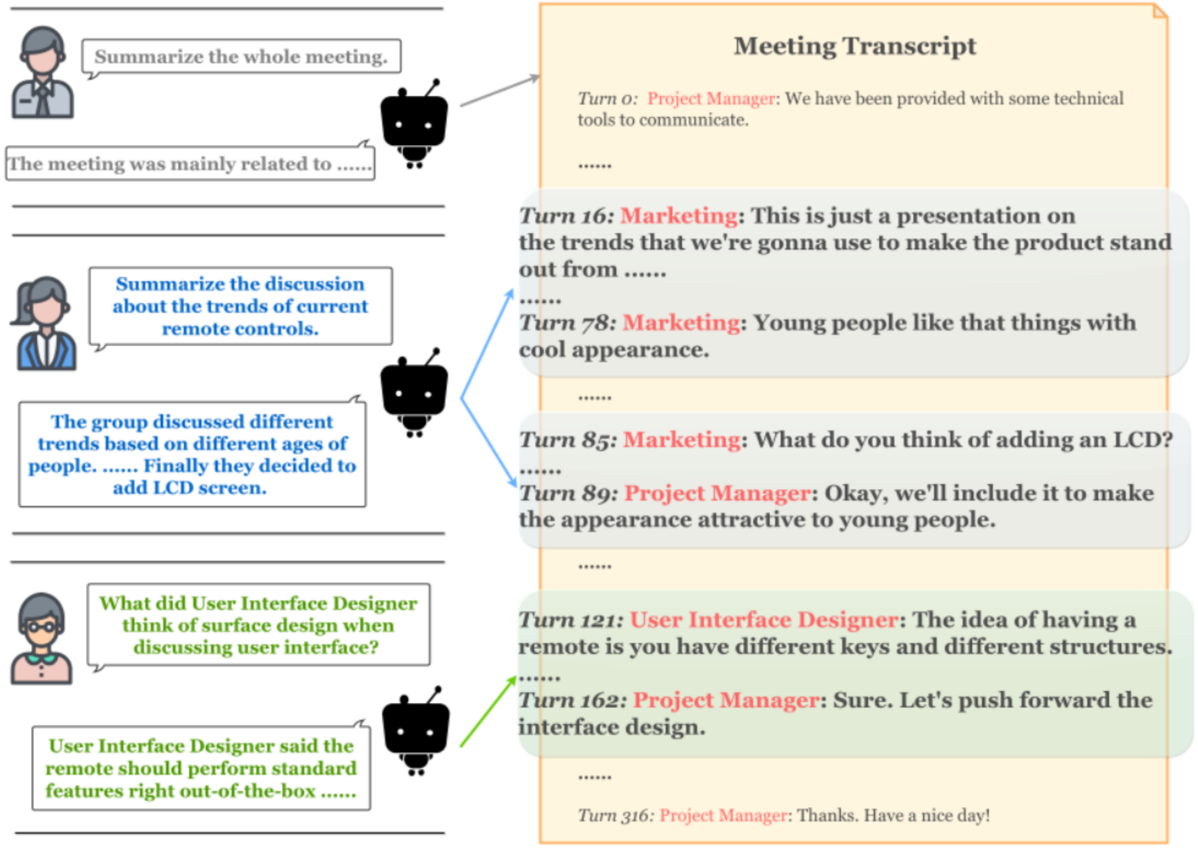
- 利用层次结构[1]
- 动态滑动窗口策略[2]
- 使用Longformer[3]

[1] A hierarchical network for abstractive meeting summarization with cross-domain pretraining EMNLP 2020
[2] Dynamic Sliding Window for Meeting Summarization
[3] ConvoSumm: Conversation Summarization Benchmark and Improved Abstractive Summarization with Argument Mining ACL 2021

Model	AMI			ICSI		
	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
Extractive Methods						
TextRank [Mihalcea and Tarau, 2004]	35.19	6.13	15.70	30.72	4.69	12.97
SummaRunner [Nallapati *et al.*, 2017]	30.98	5.54	13.91	27.60	3.70	12.52
Abstractive Methods						
UNS [Shang *et al.*, 2018]	37.86	7.84	13.72	31.73	5.14	14.50
PGN [See *et al.*, 2017]	42.60	14.01	22.62	35.89	6.92	15.67
Sentence-Gated [Goo and Chen, 2018]	49.29	19.31	24.82	39.37	9.57	17.17
TopicSeg [Li *et al.*, 2019a]	51.53	12.23	25.47	-	-	-
TopicSeg+VFOA [Li *et al.*, 2019a]	53.29	13.51	26.90	-	-	-
HMNet [Zhu *et al.*, 2020]	52.36	18.63	24.00	45.97	10.14	18.54
PGN($\mathcal{D}_{\text{ALL}}$) [Feng *et al.*, 2021]	50.91	17.75	24.59	-	-	-
DDAMS [Feng *et al.*, 2020a]	51.42	20.99	24.89	39.66	10.09	17.53
DDAMS+DDADA [Feng *et al.*, 2020a]	53.15	22.32	25.67	40.41	11.02	19.18
Pre-trained Language Model-based Methods						
Longformer-BART [Fabbri *et al.*, 2021]	54.81	20.83	25.98	43.40	12.19	19.29
Longformer-BART-arg [Fabbri *et al.*, 2021]	55.27	20.89	24.94	44.51	11.80	19.19

Meeting domain New dataset--QMSum

- Query-based dialogue summarization
  - 各取所需，灵活度更高



QMSum: A New Benchmark for Query-based Multi-domain Meeting Summarization. NAACL 2021

总结

- 数据资源挑战
  - 提出新的数据集、借助预训练
- 利用外部信息
  - Dialogue Act、Dialogue Discourse、
  - Commonsense Knowledge、Multi-modal等
- 领域特定挑战
  - 文本过长
  - 医疗对话的否定词等

Thanks