Research Progress on Multimodal Summarization

Jiajun Zhang

Institute of Automation Chinese Academy of Sciences
jjzhang@nlpr.ia.ac.cn
www.nlpr.ia.ac.cn/cip/jjzhang.htm

Joint work with Haoran Li, Junnan Zhu, Yu Zhou, Chengqing Zong and etc.

Outline

- 多模态摘要背景介绍
- 异源多模态摘要方法
- 异源多模态摘要评价
- 总结和展望

多模态摘要背景介绍

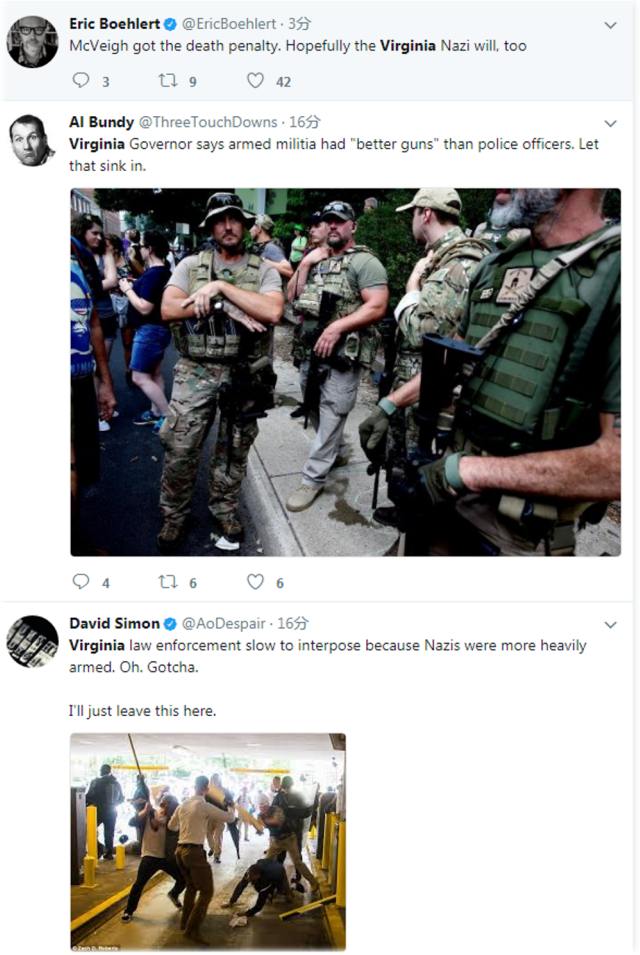
■ 文本自动摘要

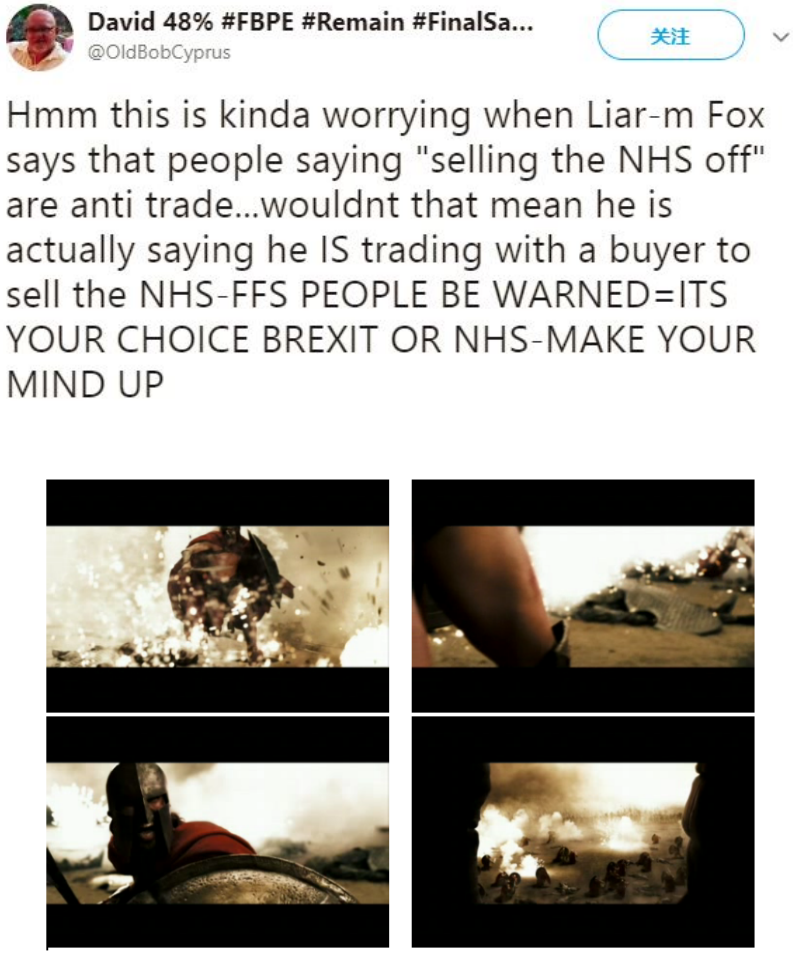
- 利用计算机对文本内容进行归纳并生成摘要

一个CNN调查表明，走私者以登船折扣为诱惑诱使非洲移民带来更多的登船者使原本拥挤的船只变得更加拥挤

CNN调查揭露非法移民内幕

多模态摘要背景介绍


Eric Boehlert @EricBoehlert · 3分
McVeigh got the death penalty. Hopefully the Virginia Nazi will, too
3 9 42
Al Bundy @ThreeTouchDowns · 16分
Virginia Governor says armed militia had "better guns" than police officers. Let that sink in.
4 6 6
David Simon @AoDespair · 16分
Virginia law enforcement slow to interpose because Nazis were more heavily armed. Oh. Gotcha.
I'll just leave this here.


David 48% #FBPE #Remain #FinalSa...
@OldBobCyprus
关注
Hmm this is kinda worrying when Liar-m Fox says that people saying "selling the NHS off" are anti trade...wouldnt that mean he is actually saying he IS trading with a buyer to sell the NHS-FFS PEOPLE BE WARNED=ITS YOUR CHOICE BREXIT OR NHS-MAKE YOUR MIND UP

多模态摘要背景介绍

■ 多模态自动摘要

➢ 利用计算机对多模态输入内容进行归纳并生成摘要

Twenty-four MSF doctors, nurses, logisticians and hygiene and sanitation experts are already in the country, while additional staff will strengthen the team in the coming days. With the help of the local community, MSF's

emergency teams focus on searching.

Ebola a serious disease that spreads rapidly through direct contact with infected people.

同步多模态摘要

■ 会议摘要

Audio
- sound localization output
- magnitude of the audio signal

Video
- luminance changes in a small window

Text
- IF-IDF

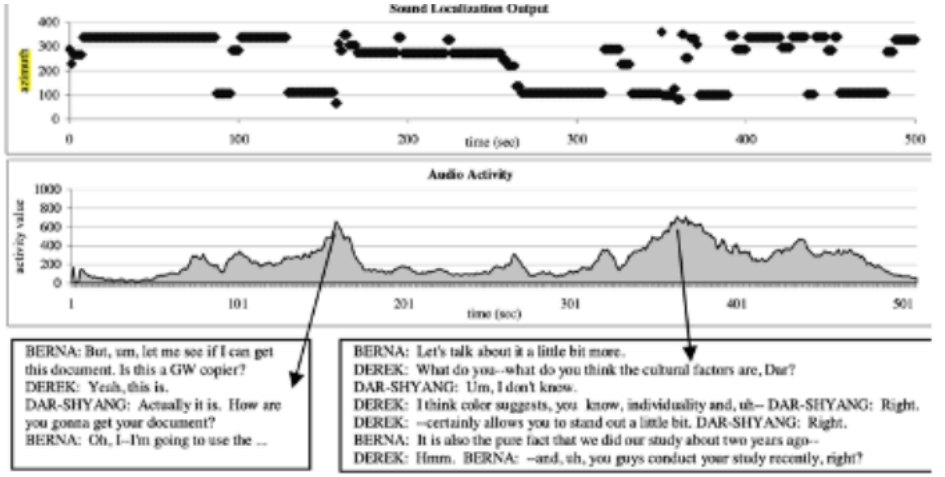
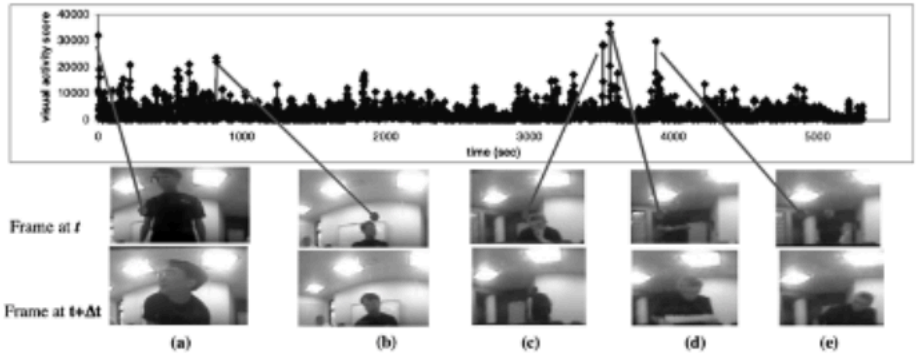


Figure 1. Sound localization and audio activity output for a staff meeting recording.

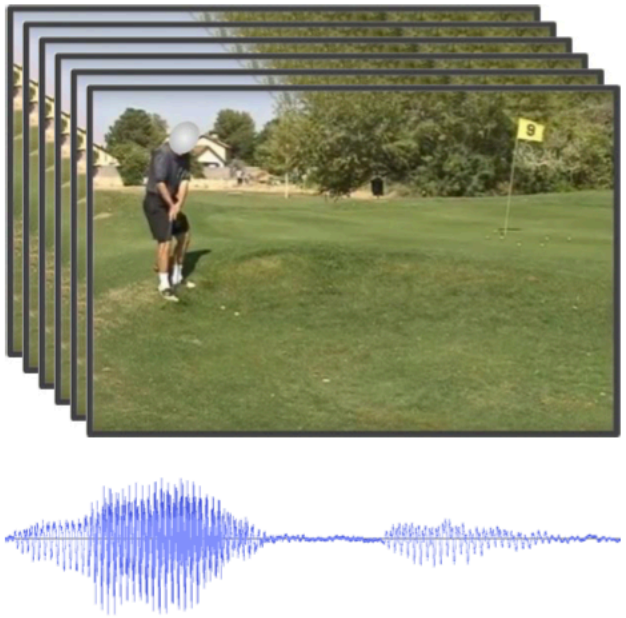


B. Erol, Dar-Shyang Lee, and J. Hull. Multimodal summarization of meeting recordings. ICME 2003.

Manling Li, L. Zhang, H. Ji, and R. Radke. Keep meeting summaries on topic: Abstractive multimodal meeting summarization. ACL 2019.

同步多模态摘要

■ 教学视频摘要



I'm very close to the green but I didn't get it on the green so now I'm in this grass bunker.

Eu estou muito perto do green, mas eu não pus a bola no green, então agora estou neste bunker de grama.

In golf, get the body low in order to get underneath the golf ball when chipping out of thick grass from a side hill lie.

Shruti Palaskar, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia,and F. Metze. How2: A large-scale dataset for multimodal language understanding. NeurIPS 2018.

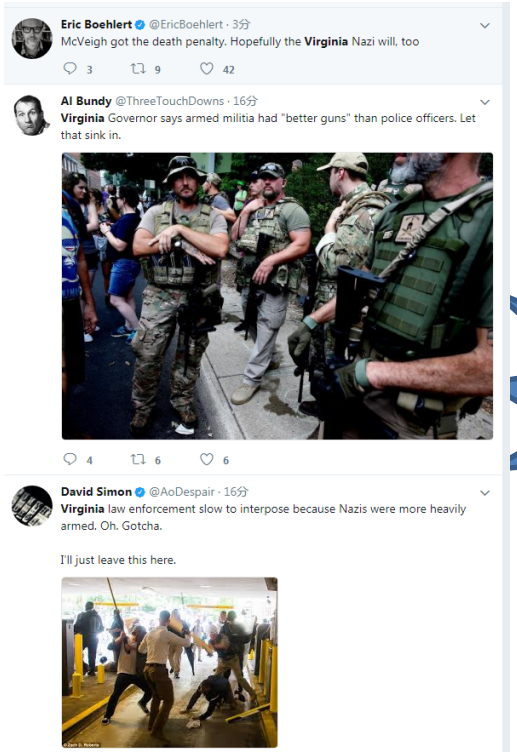
Shruti Palaskar, Jindřich Libovický, Spandana Gella, and F. Metze. Multimodal abstractive summarization for how2 videos. ACL 2019.

同步多模态摘要

■ 电影摘要

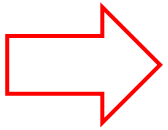
Evangelopoulos, Georgios, Athanasia Zlatintsi, Alexandros Potamianos, Petros Maragos, Konstantinos Rapantzikos, Georgios Skoumas, and Yannis Avrithis. "Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention." *IEEE Transactions on Multimedia* 2013.

异步多模态摘要



Twenty-four MSF doctors, nurses, logisticians and hygiene and sanitation experts are already in the country, while additional staff will strengthen the team in the coming days. With the help of the local community, MSF's

emergency teams focus on searching.



Ebola a serious disease that spreads rapidly through direct contact with infected people.

[Li et al., 2017; Li et al., 2018; Li et al., 2019; Li et al., 2020; Zhu et al., 2018; Zhu et al., 2020]

Outline

- 多模态摘要背景介绍
- 异源多模态摘要方法
  - 抽取式方法 [Li et al., 2017; Li et al., 2019]
  - 生成式方法
  - 多模态输出
- 异源多模态摘要评价
- 总结和展望

异步多模态摘要-抽取式方法

■ 图片信息特点：

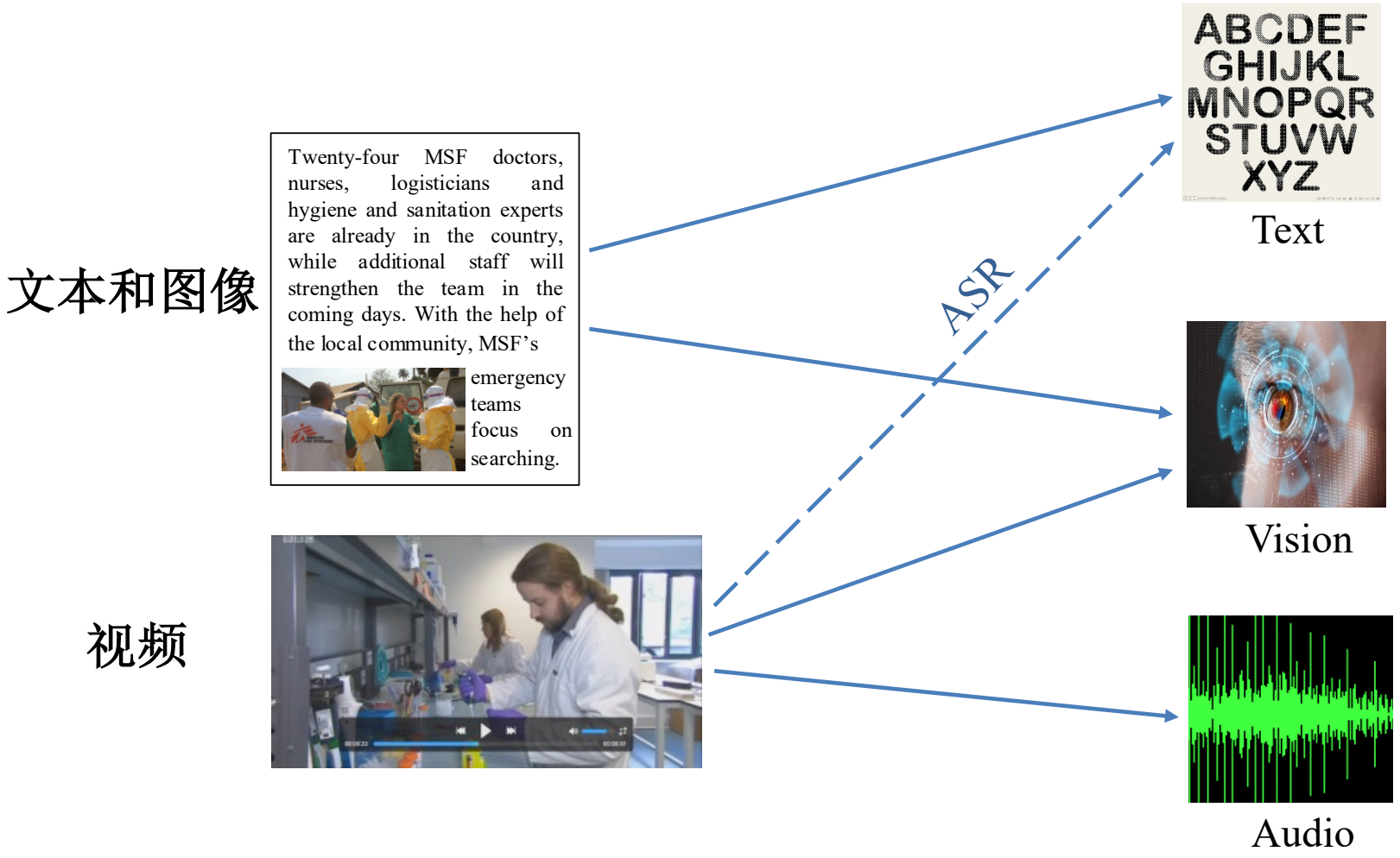
表：图文对应关系统计结果

类别	数量（篇）	占比（%）
图片强调人物	95	42.2
图片强调事件	120	53.3
图片强调地点	6	2.70
图片文本无关	4	1.80
共计	225	100

98.2%

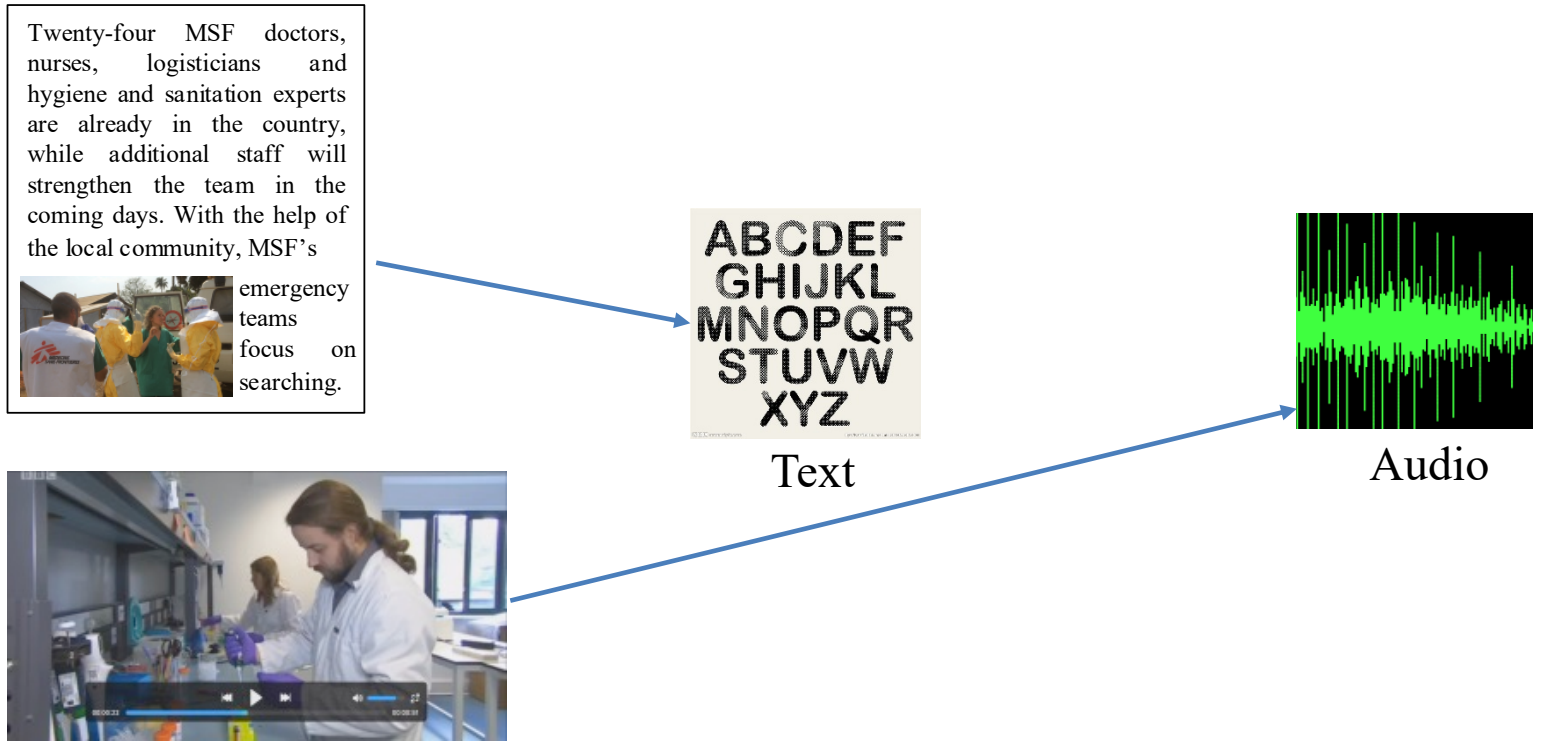
注：数据从英国每日邮报新闻门户网站随机抽取得到

异步多模态摘要-抽取式方法



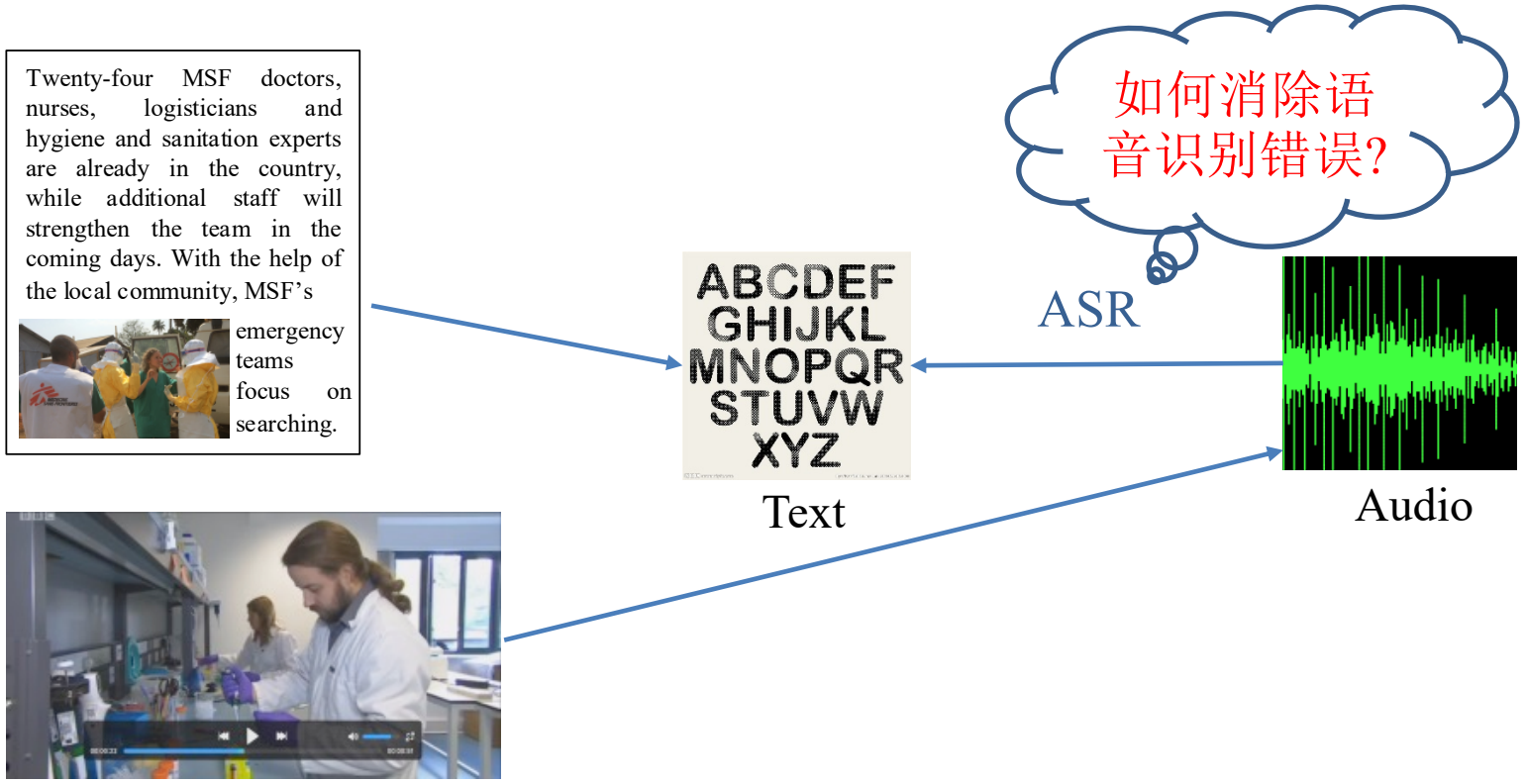
异步多模态摘要-抽取式方法

- 弥补多模态之间的语义鸿沟



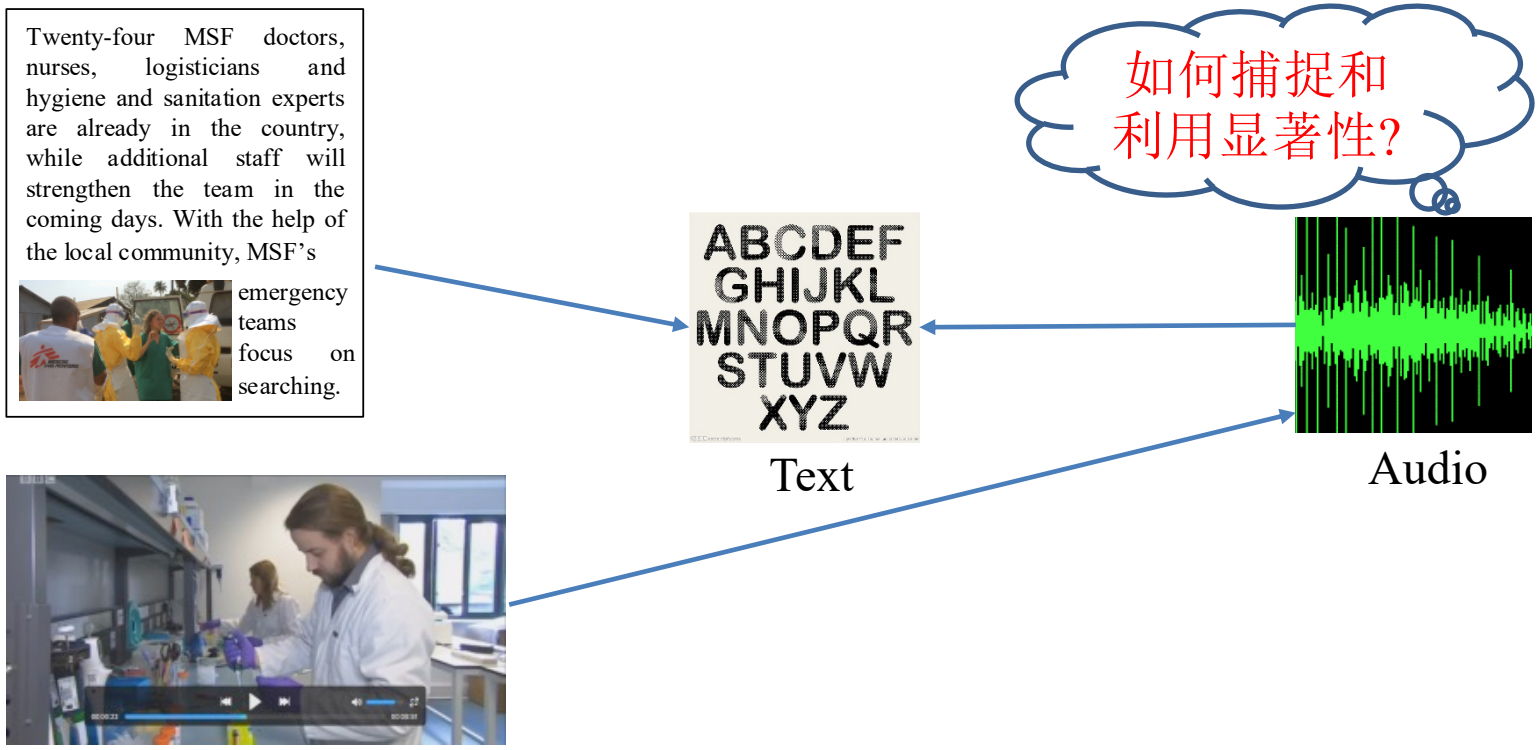
异步多模态摘要-抽取式方法

- 弥补多模态之间的语义鸿沟



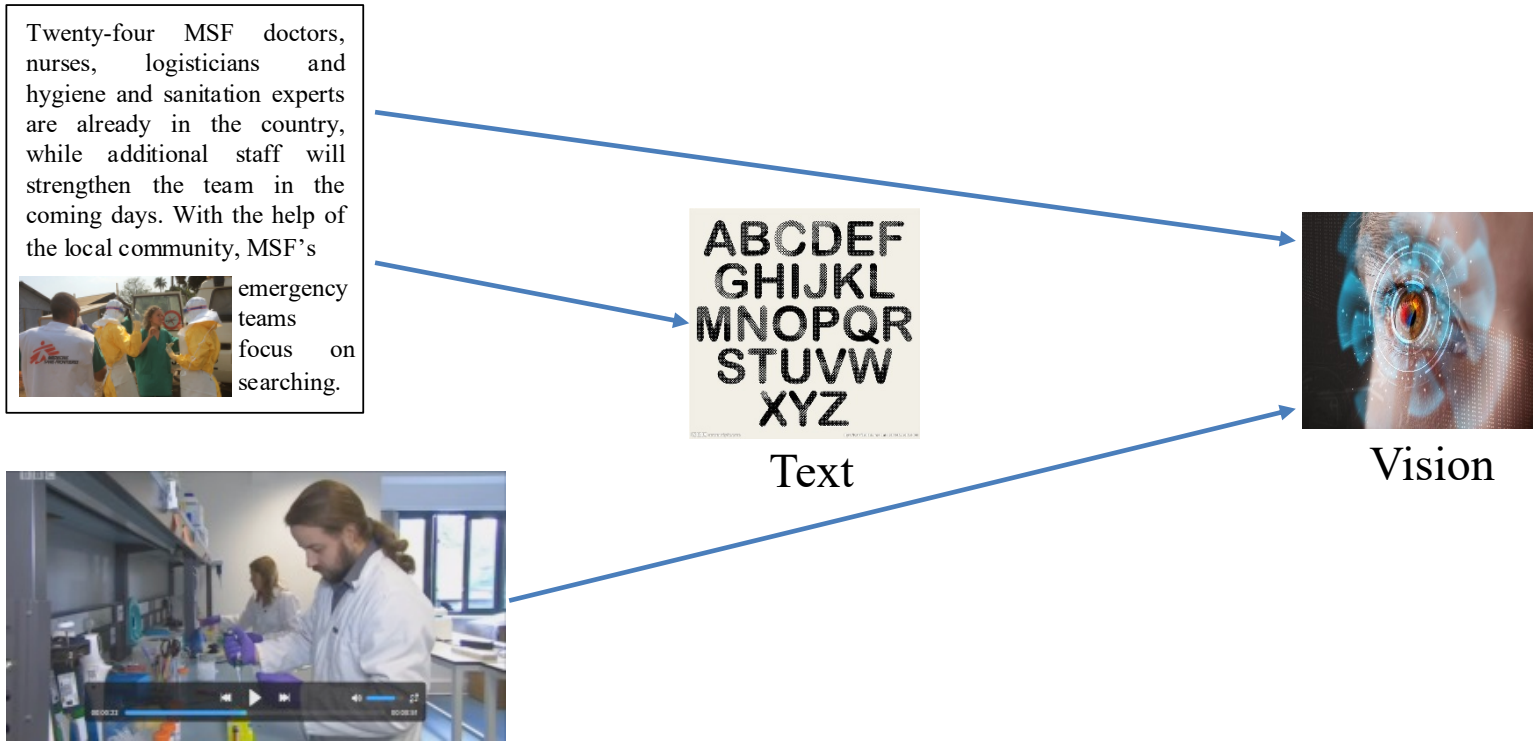
异步多模态摘要-抽取式方法

- 弥补多模态之间的语义鸿沟



异步多模态摘要-抽取式方法

- 弥补多模态之间的语义鸿沟

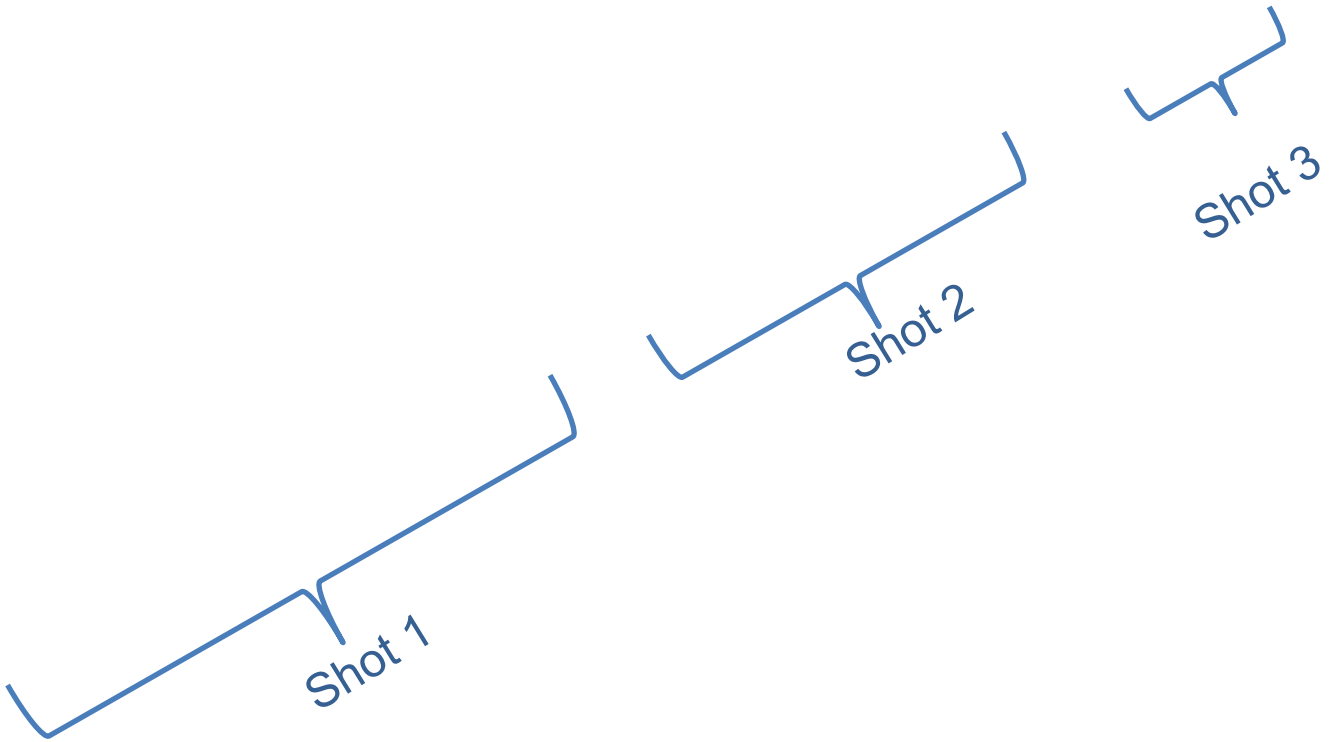


异步多模态摘要-抽取式方法

- 弥补多模态之间的语义鸿沟

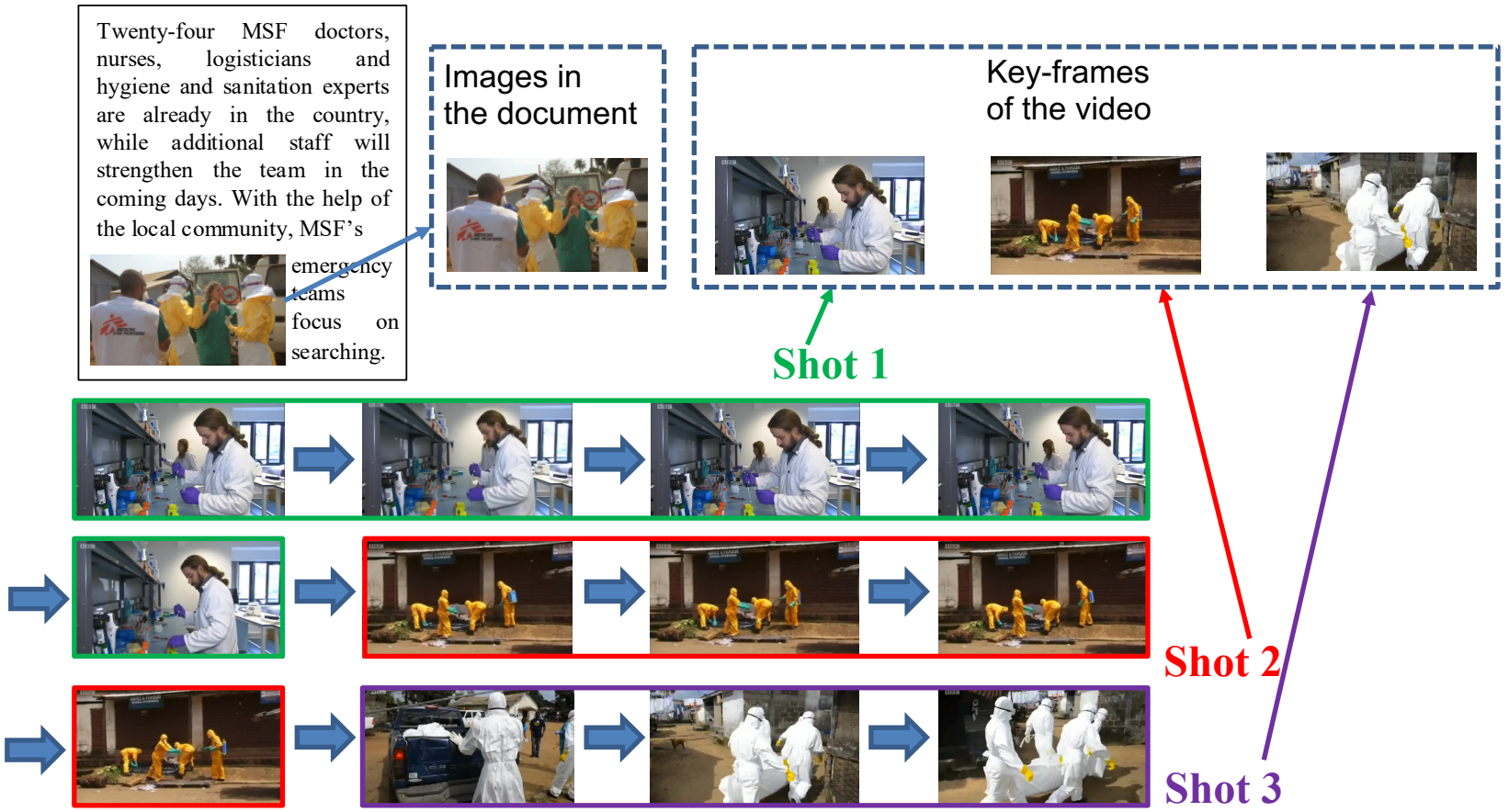
Twenty-four MSF doctors, nurses, logisticians and hygiene and sanitation experts are already in the country, while additional staff will strengthen the team in the coming days. With the help of the local community, MSF's

emergency teams focus on searching.



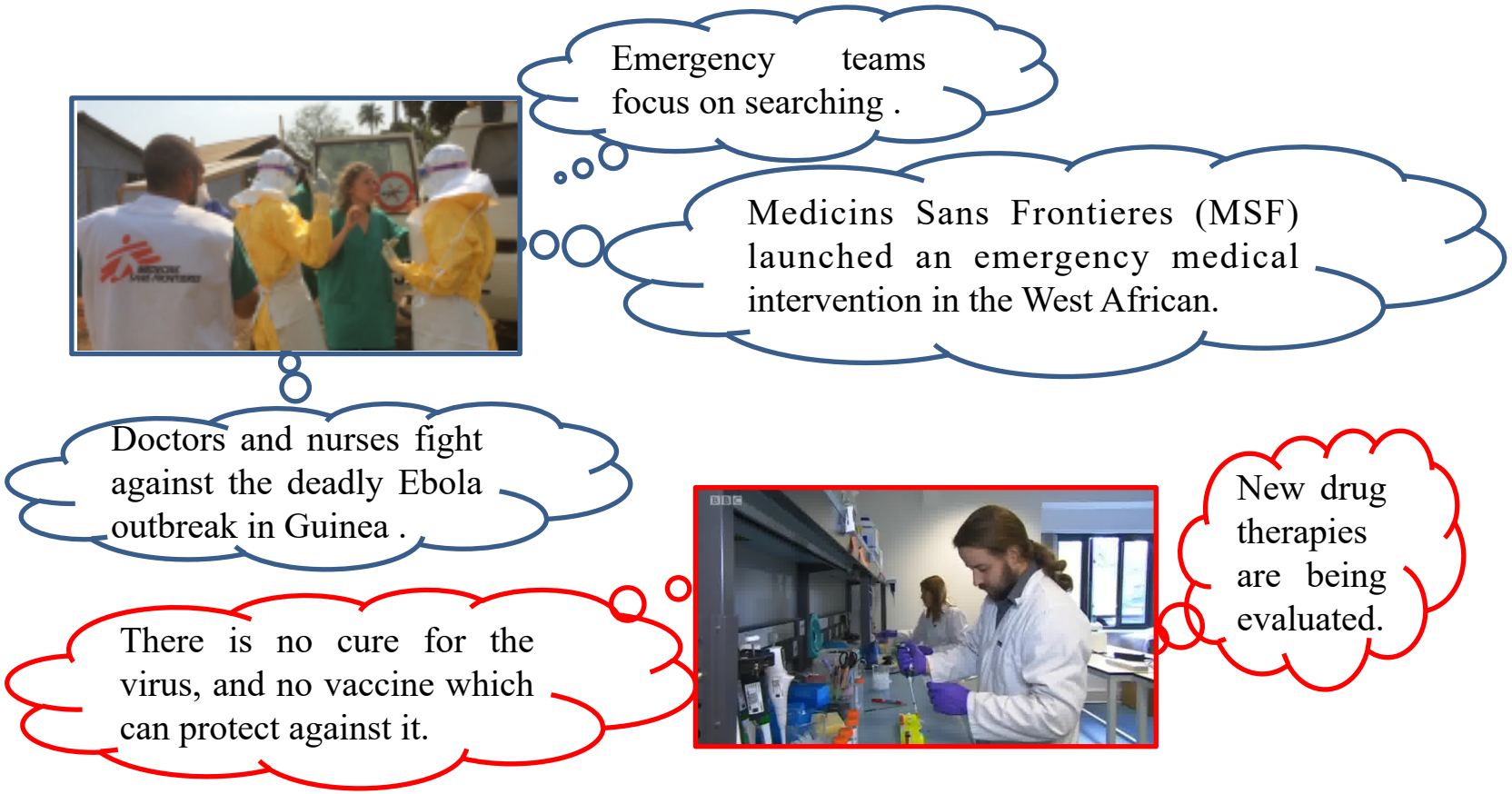
异步多模态摘要-抽取式方法

- 弥补多模态之间的语义鸿沟



异步多模态摘要-抽取式方法

- 弥补多模态之间的语义鸿沟



异步多模态摘要-抽取式方法

➢ 目标

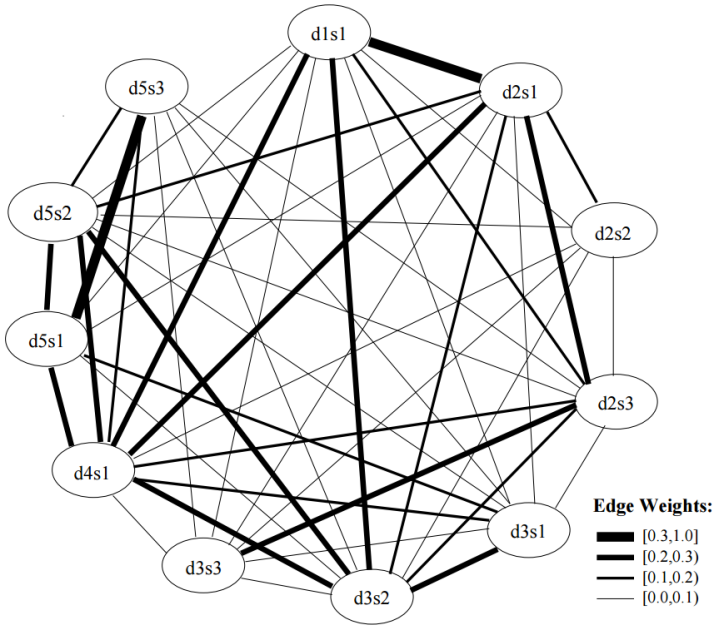
- 文本摘要: 显著性, 非冗余性
- 多模态摘要: 可读性, 视觉信息的覆盖度
  - 可读性: 消除语音识别错误
  - 视觉信息: 事件的重点线索

异步多模态摘要-抽取式方法

- 文本显著性 (包括文本中的句子和语音识别结果)
  - LexRank algorithm

$$Sa(t_i) = \mu \sum Sa(t_j) \cdot M_{ji} + \frac{1-\mu}{N}$$

$$M_{ji} = sim(t_j, t_i)$$



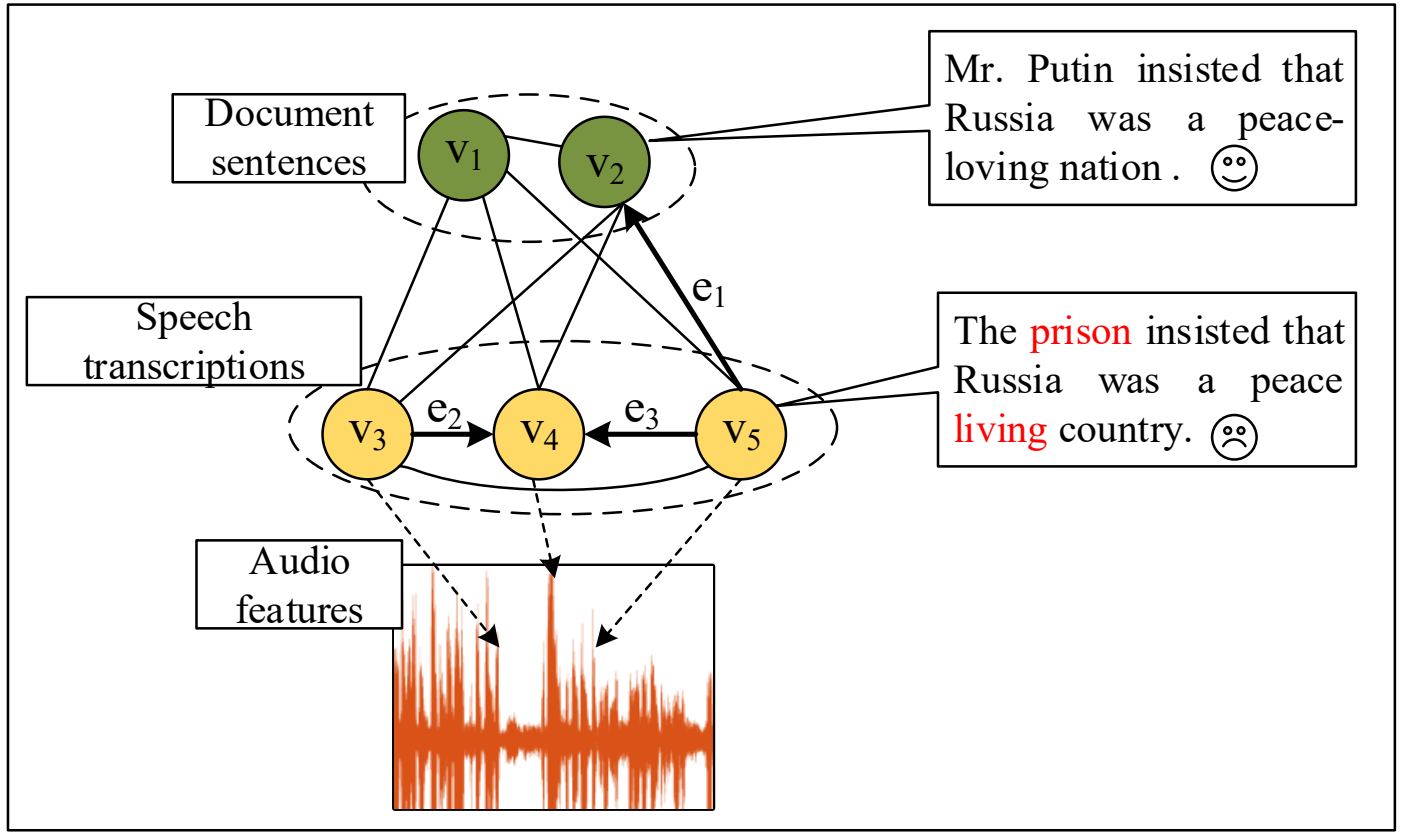
LexRank [Erkan and Radev JAIR 2004]

异步多模态摘要-抽取式方法

- 文本显著性
  - 有指导的LexRank图模型
    - 可读性指导原则: 语音识别结果单向连接文本信息
    - 音频信息指导原则: 音频强度、功率和置信度等信息作为显著性特征

异步多模态摘要-抽取式方法

- 文本显著性

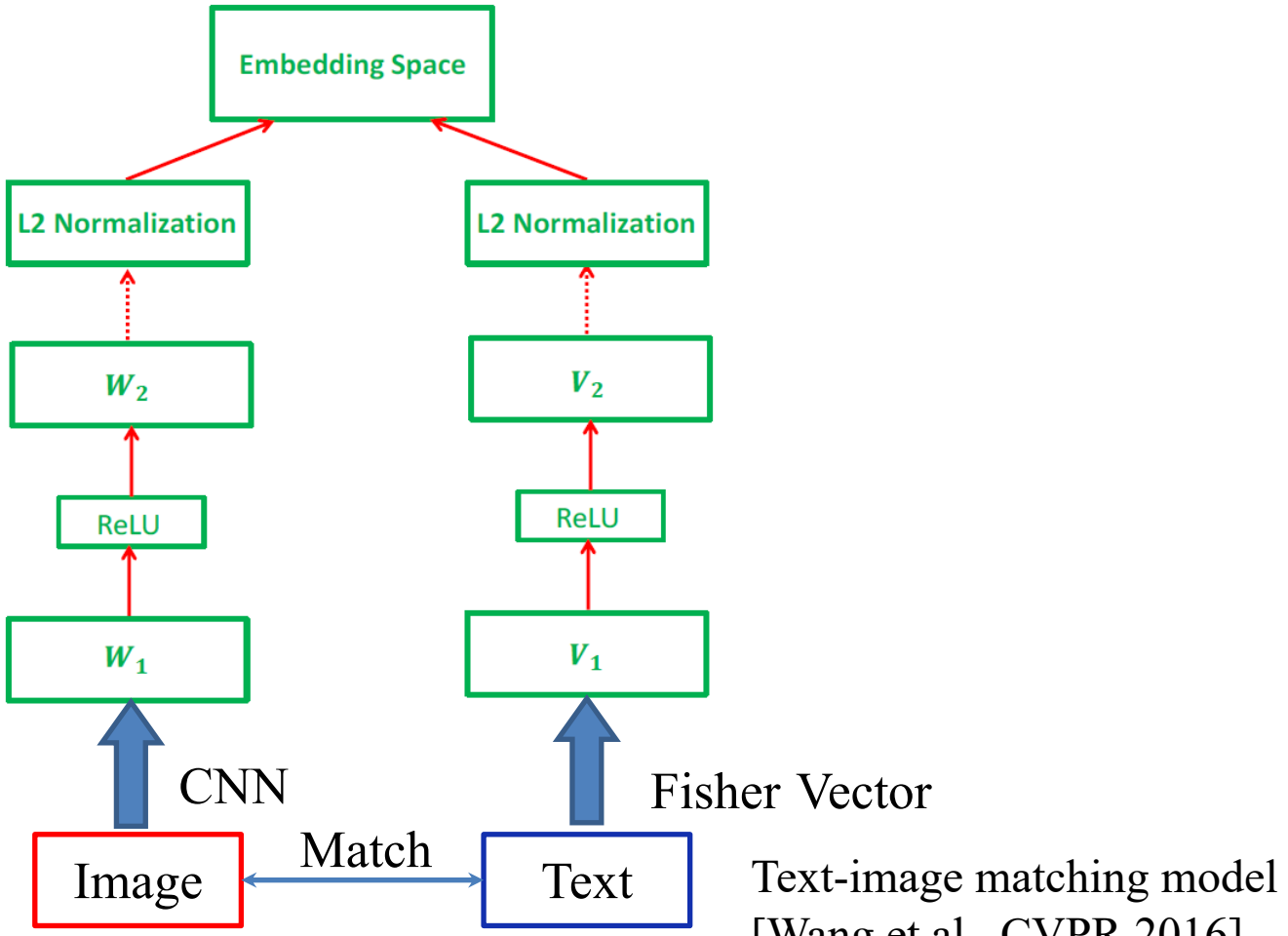


异步多模态摘要-抽取式方法

- 视觉信息覆盖度

异步多模态摘要-抽取式方法

- 视觉信息覆盖度



Text-image matching model
[Wang et al., CVPR 2016]

异步多模态摘要-抽取式方法

- 视觉信息覆盖度

>>A man in a tan jacket at the gas station pumping gas .
>>A man dressed in tan pumps gas .

Flickr30K and COCO Dataset

>>Whole streets and squares in the capital of more than 1 million people were covered in rubble .

Our Dataset

异步多模态摘要-抽取式方法

- 视觉信息覆盖度

>>A man in a tan jacket at the gas station pumping gas .
>>A man dressed in tan pumps gas .

Flickr30K and COCO Dataset

>>Whole streets and squares in the capital of more than 1 million people were covered in rubble .

Our Dataset

异步多模态摘要-抽取式方法

➢ 目标函数

- Salience for Text

$$\mathcal{F}_s(S) = \sum_{t_i \in S} Sa(t_i) - \frac{\lambda_s}{|S|} \sum_{t_i, t_j \in S} sim(t_i, t_j)$$

- Coverage for Visual

$$\mathcal{F}_c(S) = \sum_{p_i \in I} Im(p_i) b_i$$

- Considering all the modalities

$$\mathcal{F}_m(S) = \frac{1}{M_s} \sum_{t_i \in S} Sa(t_i) + \frac{1}{M_c} \sum_{p_i \in I} Im(p_i) b_i \ - \frac{\lambda_m}{|S|} \sum_{i,j \in S} sim(t_i, t_j)$$

异步多模态摘要-抽取式方法

➢ Dataset

- 50 news topics in the most recent five years, 25 in English and 25 in Chinese.
- 20 topics for each language as a test set, 5 as a development set.
- 20 documents and 5-10 videos for each topic.

	#Sentence	#Word	#Shot	Video Length
English	492.1	12,104.7	47.2	197s
Chinese	402.1	9,689.3	49.3	207s

Table 1: Corpus statistics.

异步多模态摘要-抽取式方法

- Dataset

English	(1) Nepal earthquake (2) Terror attack in Paris (3) Train derailment in India (4) Germanwings crash (5) Refugee crisis in Europe
Chinese	(6) “东方之星”客船翻沉 (“Oriental Star”passenger ship sinking) (7) 银川公交大火 (The bus fire in Yinchuan) (8) 香港占中 (Occupy Central in HONG KONG) (9) 李娜澳网夺冠 (Li Na wins Australian Open) (10) 抗议“萨德”反导系统 (Protest against “THAAD”anti-missile system)

Table 2: Examples of news topics.

异步多模态摘要-抽取式方法

Model	English			Chinese		
	R-1	R-2	R-SU4	R-1	R-2	R-SU4
Text only	0.41928	0.11342	0.16399	0.40052	0.10955	0.16233
Audio only	0.41283	0.08450	0.14332	0.25539	0.05016	0.08034
Text+audio	0.41773	0.11064	0.16217	0.39877	0.10686	0.15906
Text+audio+guide	0.41746	0.11349	0.16380	0.40095	0.10978	0.16253
Image caption	0.41470	0.10329	0.15946	0.36837	0.08259	0.13851
Image caption match	0.41496	0.10518	0.15893	0.37356	0.08327	0.13920
Image alignment	0.40729	0.08262	0.14072	0.28871	0.06870	0.10115
Image match sent	0.42234	0.11576	0.16706	0.39261	0.11187	0.15939
Image match frame	0.42594	0.13241	0.18088	0.39672	**0.13376**	0.17847
Image match chunk	0.43836	0.12627	0.18063	0.38625	0.12692	0.17367
Image match word	0.41863	0.10640	0.16249	0.37696	0.11470	0.16068
Image match frame+chunk	0.43995	**0.14301**	0.19357	0.38624	0.13011	0.17650
Image match frame+chunk+sent	0.43874	0.13909	0.18980	0.37485	0.11952	0.16437
Image match frame+chunk+sent+word	0.44073	0.13782	0.19085	0.38277	0.11835	0.16525
Image topic KL	0.43813	0.13387	0.18594	0.41066	0.12192	0.17294
Image topic IR	**0.44734**	0.14262	**0.19491**	**0.42540**	0.13232	**0.18513**

自动评测结果

异步多模态摘要-抽取式方法

	Method	Readability	Informativeness
English	Text only	3.72	3.28
	Text + audio	3.08	3.44
	Text + audio + guide	3.68	3.64
	Image match frame	3.67	3.83
	Image topic IR	3.80	4.10
	Reference	**4.52**	**4.36**
Chinese	Text only	3.64	3.40
	Text + audio	3.16	3.48
	Text + audio + guide	3.60	3.72
	Image match frame	3.62	3.92
	Image topic IR	3.73	4.00
	Reference	**4.88**	**4.84**

人工评测结果

异步多模态摘要-抽取式方法

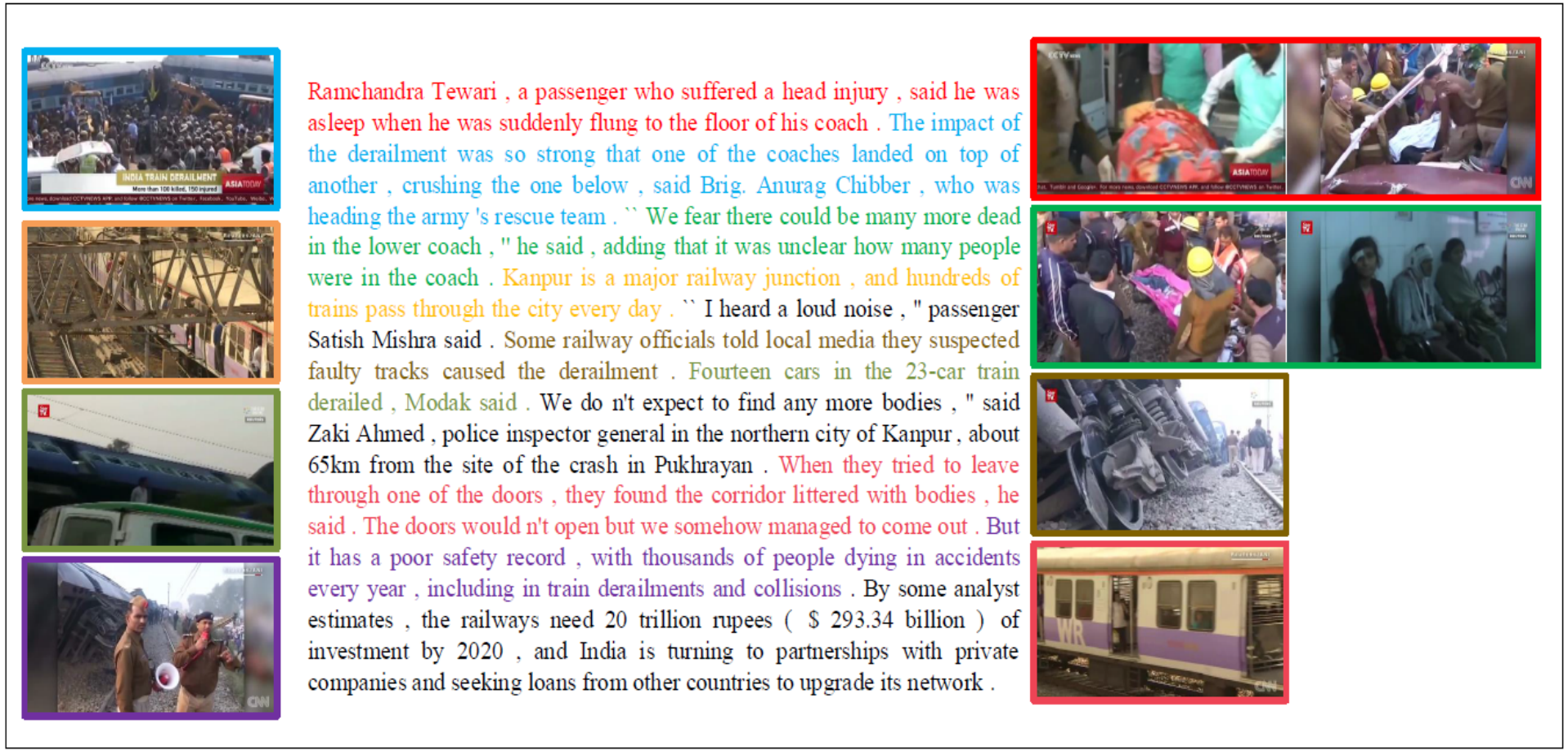


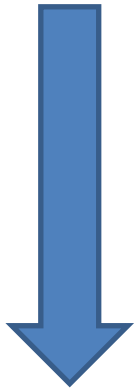
Figure: An example of generated summary for the news topic “India train derailment”.

Outline

- 多模态摘要背景介绍
- 异源多模态摘要方法
  - 抽取式方法
  - 生成式方法 [Li et al., 2018; Li et al., 2020]
  - 多模态输出
- 异源多模态摘要评价
- 总结和展望

异步多模态摘要-生成式方法

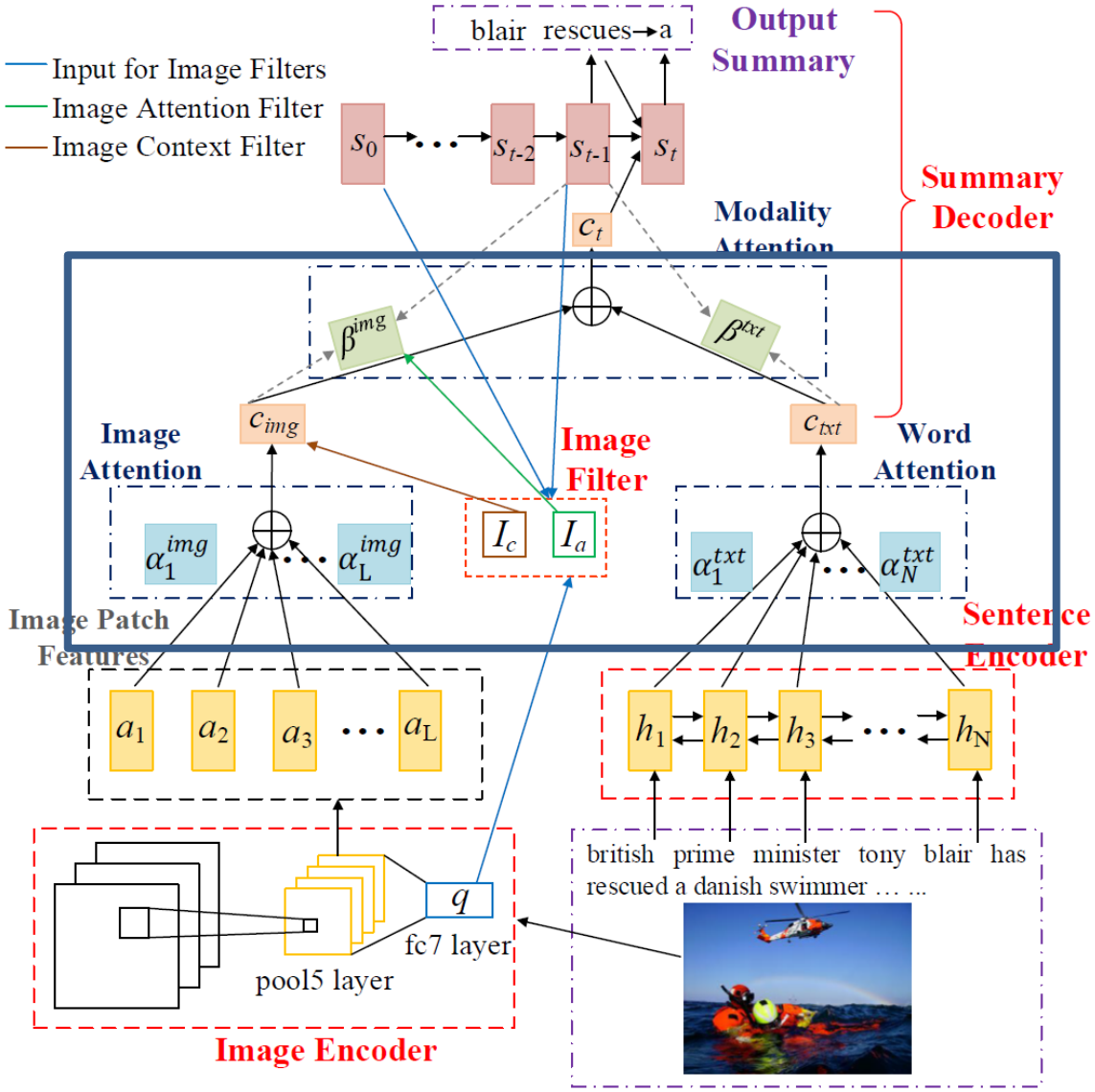
British prime minister Tony Blair has rescued a Danish swimmer from the shark infested sea during his holiday in Seychelles in Africa , a government spokesman said Friday .



Blair rescues a swimmer from sea .

异步多模态摘要-生成式方法

- 层次编码



异步多模态摘要-生成式方法

➢ Image Filtering

✓ Image Attention Filter

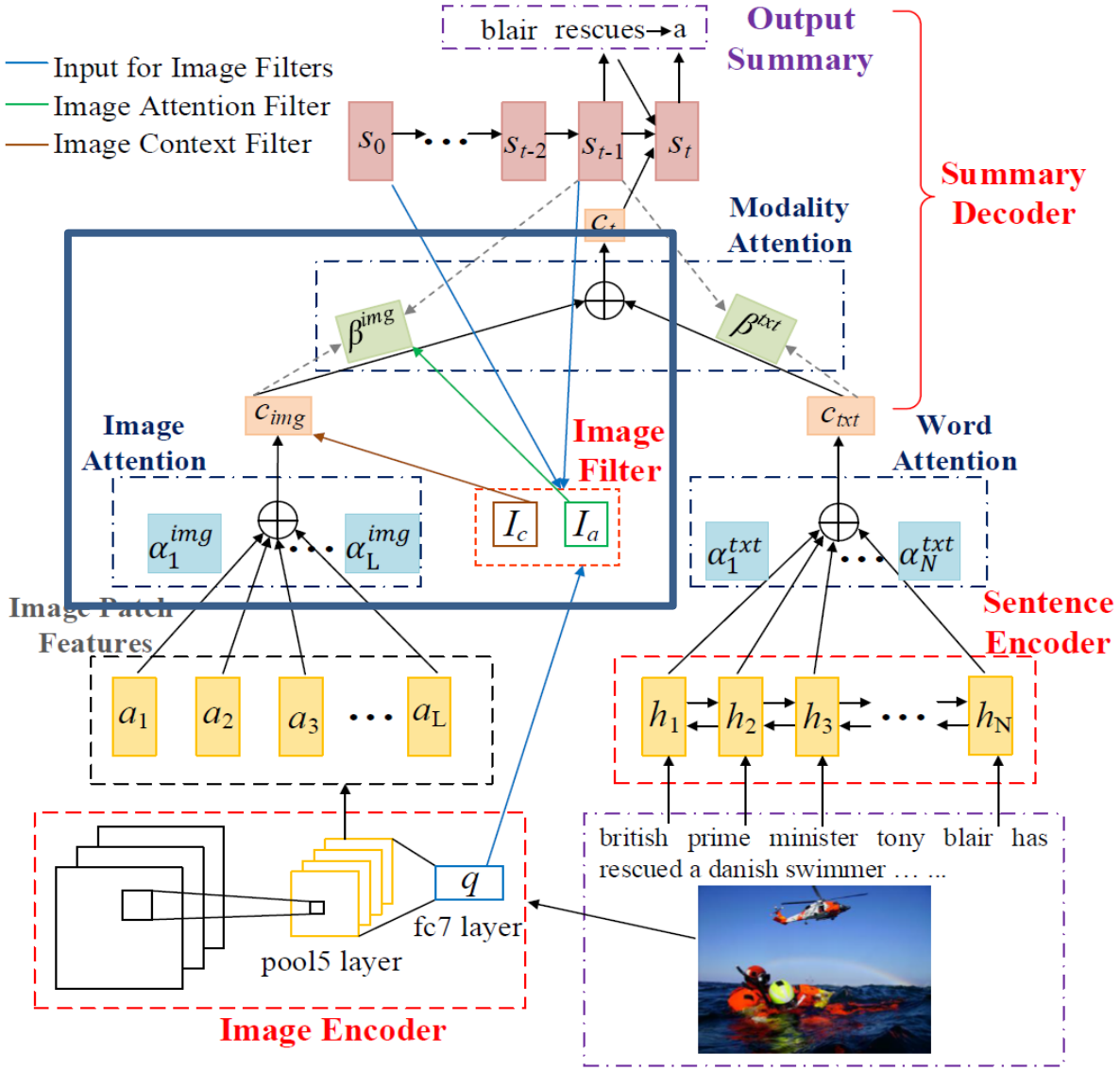
$$I_a = \sigma(v_s^T s_0 + v_q^T q + v_r^T s_{t-1})$$

$$\beta_t^{img} = I_a \cdot \beta_t^{img}$$

✓ Image Context Filter

$$I_c = \sigma(\mathbf{W}_s s_0 + \mathbf{W}_q q + \mathbf{W}_r s_{t-1})$$

$$c_{img} = I_c \odot c_{img}$$



异步多模态摘要-生成式方法

- Main results

Model		R-1	R-2	R-L
Lead		33.64	13.40	31.84
Compress [Clarke and Lapata, 2008]		31.56	11.02	28.87
ABS [Rush *et al.*, 2015]		35.95	18.21	31.89
SEASS [Zhou *et al.*, 2017]		44.86	23.03	41.92
Multi-Source [Libovický and Helcl, 2017]		39.67	19.11	38.03
Doubly-Attentive [Calixto *et al.*, 2017]		41.11	21.75	39.92
Our text only model		44.58	22.68	41.91
Multi-modal model without image filter		44.88	23.20	42.11
Multi-modal model with attention filter	Decoder_{text}	45.17	23.39	42.20
	Decoder_{fc}	45.02	23.06	42.24
	Decoder_{aconv}	45.21	23.82	42.50
	Decoder_{wconv}	45.78	23.45	43.16
	+text coverage	**47.28**	**24.85**	**44.48**
	+image coverage	47.16	24.43	44.23
Multi-modal model with context filter	Decoder_{text}	45.43	23.47	42.67
	Decoder_{fc}	45.12	23.29	42.34
	Decoder_{aconv}	45.99	23.86	43.19
	Decoder_{wconv}	46.08	24.00	43.29
	+text coverage	46.84	24.25	43.76
	+image coverage	46.56	24.34	43.59

异步多模态摘要-生成式方法

- 图片可以帮助生成更加精准的摘要

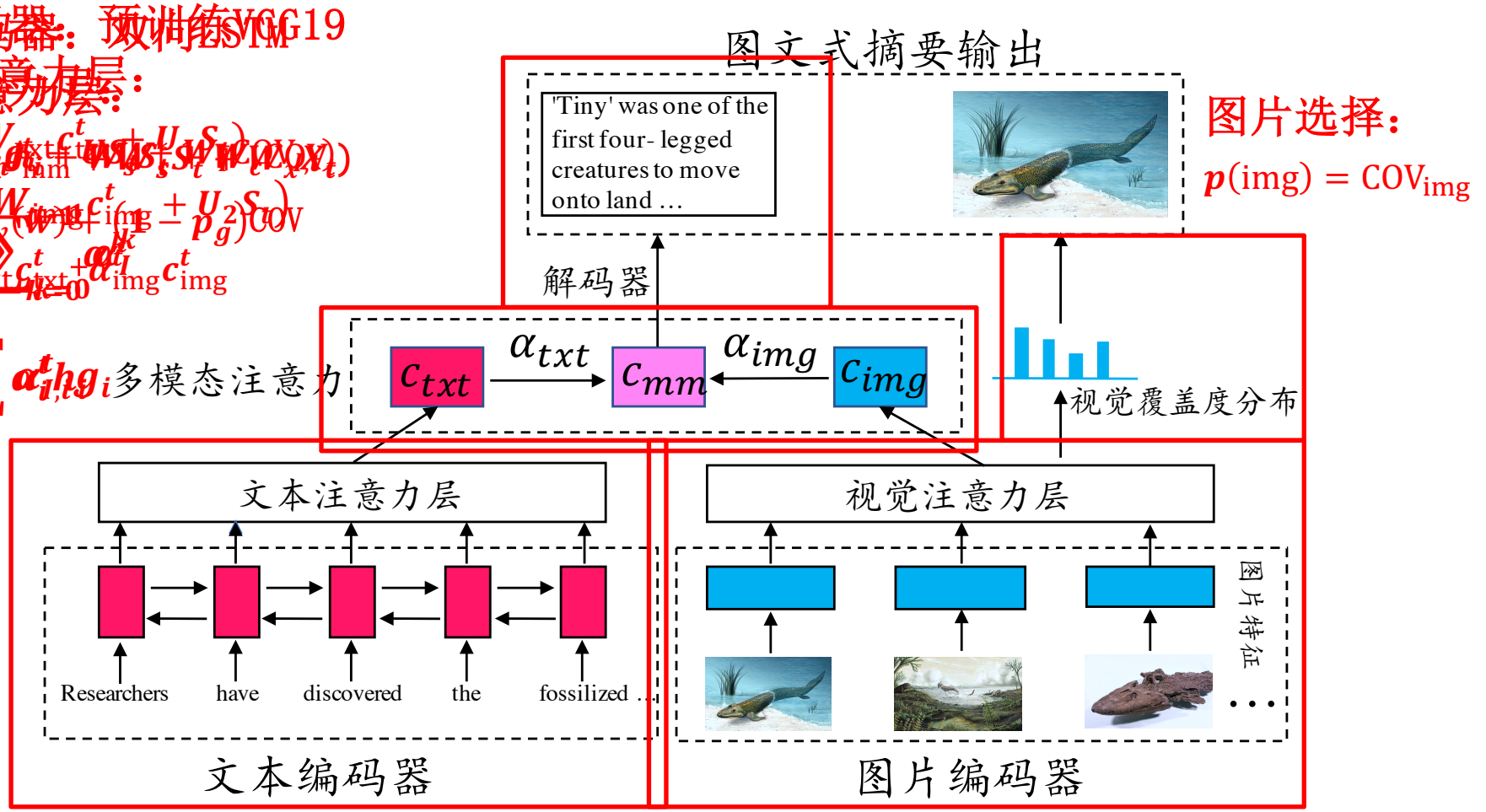
Source sentence: at least ## people were killed and ## injured when a passenger bus plunged into a deep ravine in north ethiopia , police said wednesday .
Reference summary: *bus accident* kills at least ## in north ethiopia
Text-only model: ## killed as bus plunges into ravine
Multi-modal model: ## killed in ethiopian *bus accident*

Outline

- 多模态摘要背景介绍
- 异源多模态摘要方法
  - 抽取式方法
  - 生成式方法
  - 多模态输出 [Zhu et al., 2018; Zhu et al., 2020]
- 异源多模态摘要评价
- 总结和展望

异步多模态摘要-多模态输出

- 融入多模态注意力机制的图文式摘要生成方法



异步多模态摘要-多模态输出

- 实验设置
  - ✓ 数据集构造（英国每日邮报新闻门户网站）

Argentina cancel final World Cup warm-up friendly against Israel in Jerusalem after widespread protests

- Argentina have cancelled their final World Cup warm-up match against Israel
- The friendly clash was expected to be played at the Teddy Stadium in Jerusalem
- Palestinian FA praised the decision to cancel the game as moral victory for sport

By REUTERS REPORTER
PUBLISHED: 01:33 BST, 6 June 2018 | **UPDATED:** 08:09 BST, 6 June 2018

Share 70 shares 10 View comments

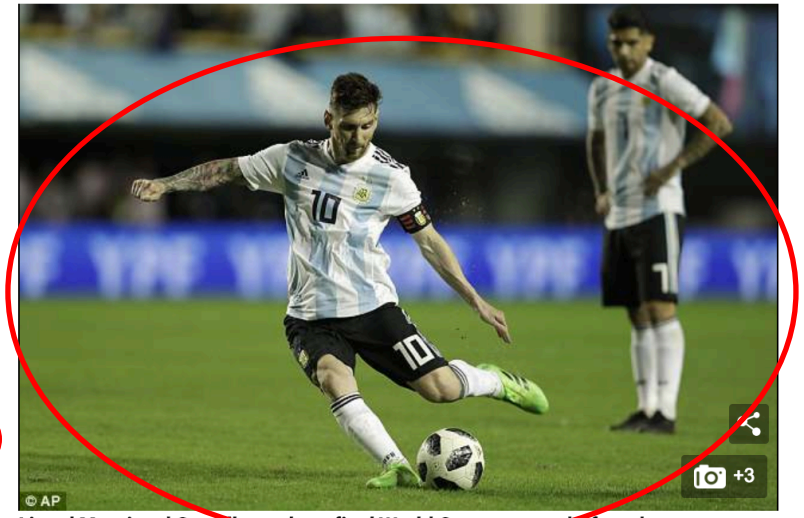
Argentina have cancelled their final World Cup warm-up match against **Israel**, striker Gonzalo Higuain said on Tuesday, as political pressure grew ahead of Saturday's scheduled fixture in Jerusalem.

'They've finally done the right thing,' Higuain told ESPN, confirming reports the game had been cancelled.

The match at Jerusalem's Teddy Stadium was to be Argentina's last before they kick off their World Cup campaign in **Russia** on June 16.

文本摘要

The match at Jerusalem's Teddy Stadium was to be Argentina's last before they kick off their World Cup campaign in **Russia** on June 16.



Lionel Messi and Co will not play a final World Cup warm-up before the tournament starts

文本输入

图片–文本描述匹配对

异步多模态摘要-多模态输出

- 实验设置
  - ✓ 实验数据：

图文式自动摘要数据集（MSMO）统计信息

	训练集	验证集	测试集
文档数	293,965	10,355	10,261
图片标题数	1,928,356	68,520	71,509
原文平均词数	720.87	766.08	730.80
参考摘要平均词数	70.12	70.02	72.16
图片标题平均词数	22.07	22.64	22.34
新闻平均图片数	6.56	6.62	6.97

异步多模态摘要-多模态输出

■ 实验结果

➢ 人工评价-图文式摘要得分

模型	质量	模型	质量
ATG	3.45	MOF_D^R	3.67
ATL	3.39	MOF_E^R	3.52
HAN	3.35	MOF_D^O	3.62
GR	3.30	MOF_E^O	3.56

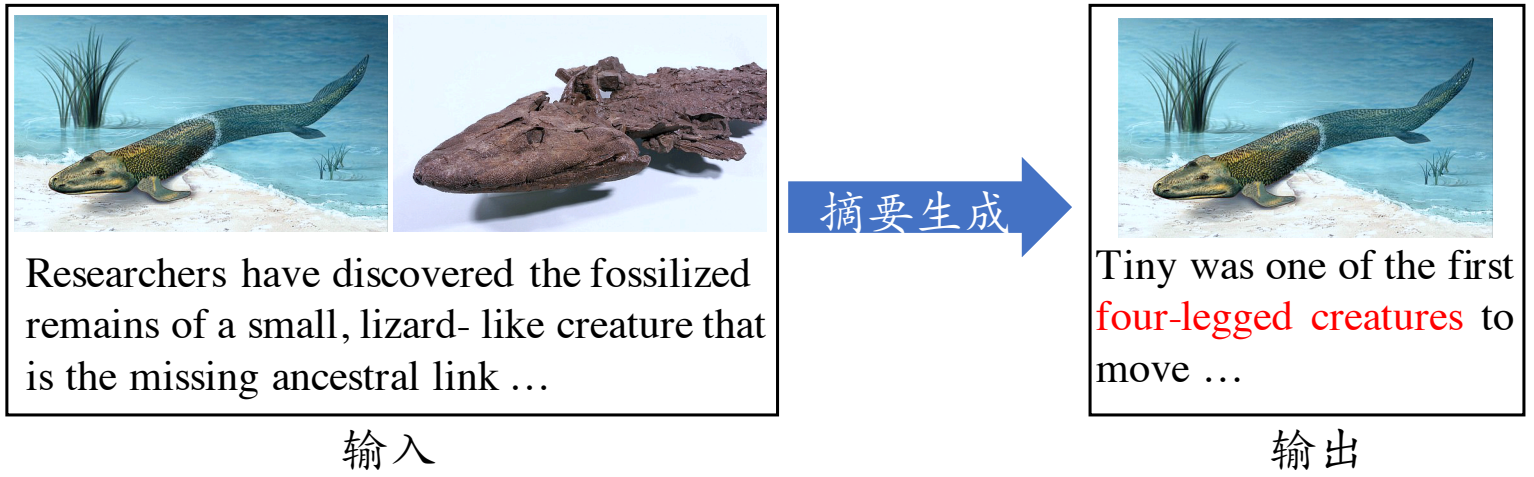
- 分数说明：0-5分，分数越高，图文式摘要质量越高
- **所提方法能够有效提升图文式摘要质量**

Outline

- 多模态摘要背景介绍
- 异源多模态摘要方法
- 异源多模态摘要评价 [Zhu et al., 2018; Zhu et al., 2020]
- 总结和展望

异步多模态摘要-自动评价

- 图文式自动摘要任务



如何自动评估图文式输出的质量？

异步多模态摘要-自动评价

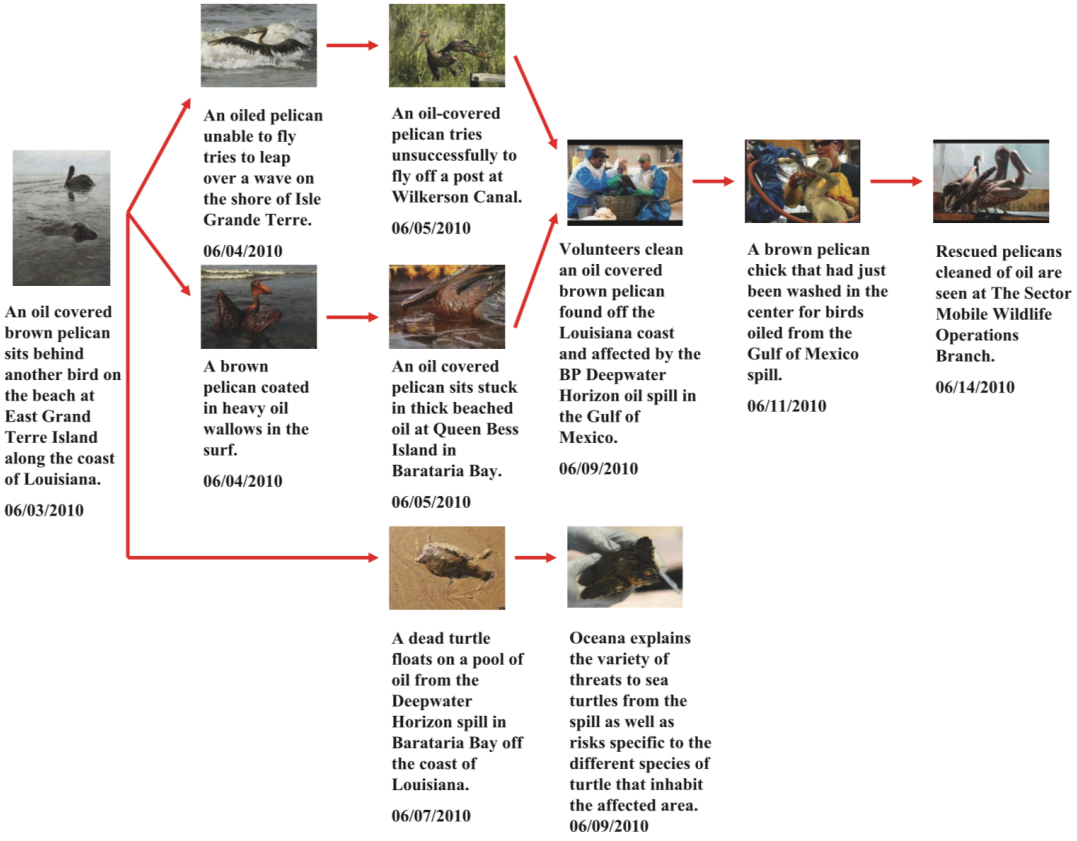
- 现有评价方法

✓ ROUGE （文本摘要）

$$ROUGE-\mathrm{N} = \frac{\sum_{S\in\{Ref\}}\sum_{ngram} Count_{match}(ngram)}{\sum_{S\in\{Ref\}}\sum_{ngram} Count(ngram)}$$

✓ 准确率（整体）

只适用于图片和文本对齐的场景

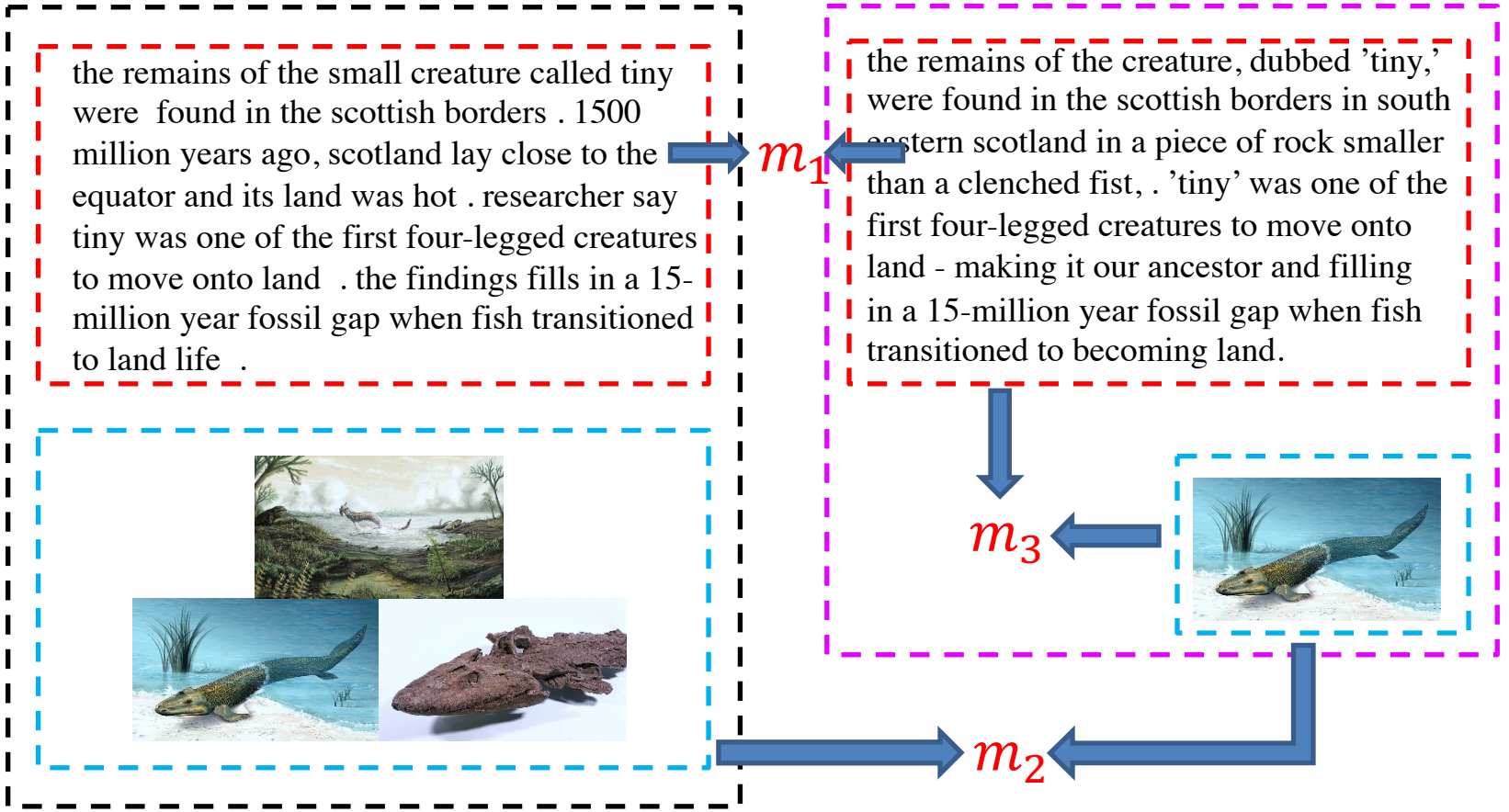


异步多模态摘要-自动评价

- 多模态自动摘要评价方法（MMAE）
  - ✓ 定义：$y = f(m_1, m_2, m_3)$
    - m_1: 文本重要性度量
    - m_2: 图片重要性度量
    - m_3: 图片和文本相关性度量
    - $f(\cdot)$: 映射函数
    - y: 自动评估分数

异步多模态摘要-自动评价

- 多模态自动摘要评价方法（MMAE）
  - 参考答案
  - 待评价摘要



异步多模态摘要-自动评价

- 多模态自动摘要评价方法（**MMAE**）
  - ✓ 文本重要性度量指标
    - ROUGE-1、ROUGE-2、ROUGE-L
    - BLEU
  - ✓ 图片重要性度量指标
    - 图片准确率 **(Image Precision, IP):** $$\mathrm{IP} = \frac{|\{\mathrm{ref}_{\mathrm{img}}\} \cap \{\mathrm{sys}_{\mathrm{img}}\}|}{|\{\mathrm{sys}_{\mathrm{img}}\}|}$$
    - 图片相似度：
      - I-I: 图片特征向量的余弦相似度
      - Hist: 图片像素分布直方图之间的Bhattacharyya距离
      - Temp: 基于傅里叶分析的模板匹配相似度

异步多模态摘要-自动评价

➢ 多模态自动摘要评价方法（MMAE）

✓ 图片和文本相关性度量指标

- 动机：图文式摘要中图片和文本应该具有一定的相关性

北京时间6月4日，NBA总决赛第二战如期而至。第三节的一个回合，裁判的判罚再惹争议，当时骑士从后场发动长传进攻，勒布朗在前场跳起接球，这时水花兄弟贴在一起防守勒布朗，身体都有一定的接触，落地之后勒布朗趔趔趄趄运球倒地，在这个过程中，倒地的库里疑似有一个伸腿的动作。

较差

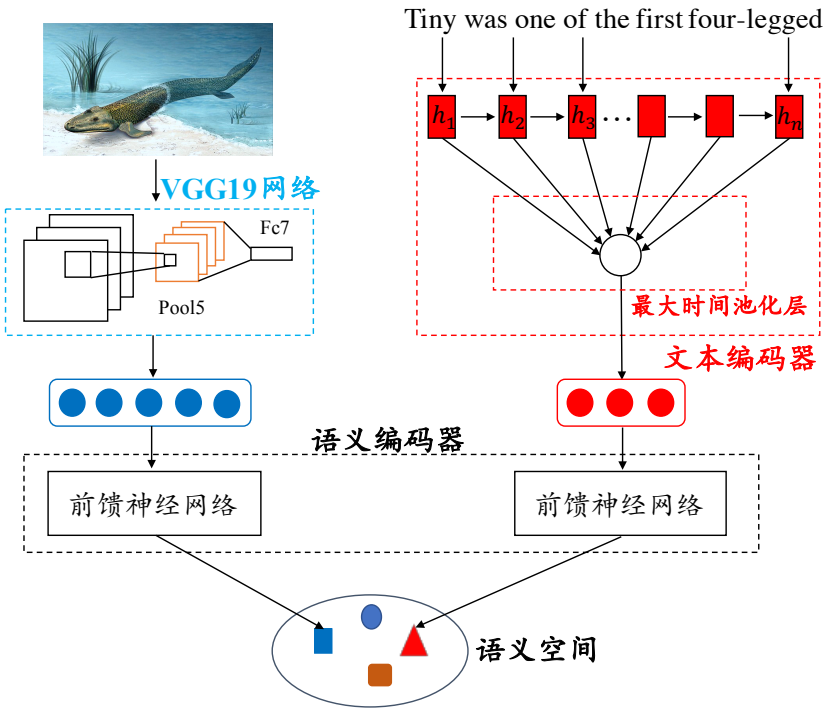
较好！

异步多模态摘要-自动评价

➢ 多模态自动摘要评价方法（MMAE）

✓ 图片和文本相关性度量指标

● 视觉语义向量模型（Visual-Semantic Embedding）



- 最大间隔损失

$$L = \sum_{\hat{c}} \max(\beta - s(i,c) + s(i,\hat{c}), 0) + \sum_{\hat{i}} \max(\beta - s(i,c) + s(\hat{i},c), 0)$$

第一项

第二项

第一项: 用图片检索文本的损失

第二项: 用文本检索图片的损失

异步多模态摘要-自动评价

- 实验设置
  - ✓ 数据集构造（英国每日邮报新闻门户网站）

Argentina cancel final World Cup warm-up friendly against Israel in Jerusalem after widespread protests

- Argentina have cancelled their final World Cup warm-up match against Israel
- The friendly clash was expected to be played at the Teddy Stadium in Jerusalem
- Palestinian FA praised the decision to cancel the game as moral victory for sport

By REUTERS REPORTER
PUBLISHED: 01:33 BST, 6 June 2018 | **UPDATED:** 08:09 BST, 6 June 2018

Share 70 shares 10 View comments

Argentina have cancelled their final World Cup warm-up match against **Israel**, striker Gonzalo Higuain said on Tuesday, as political pressure grew ahead of Saturday's scheduled fixture in Jerusalem.

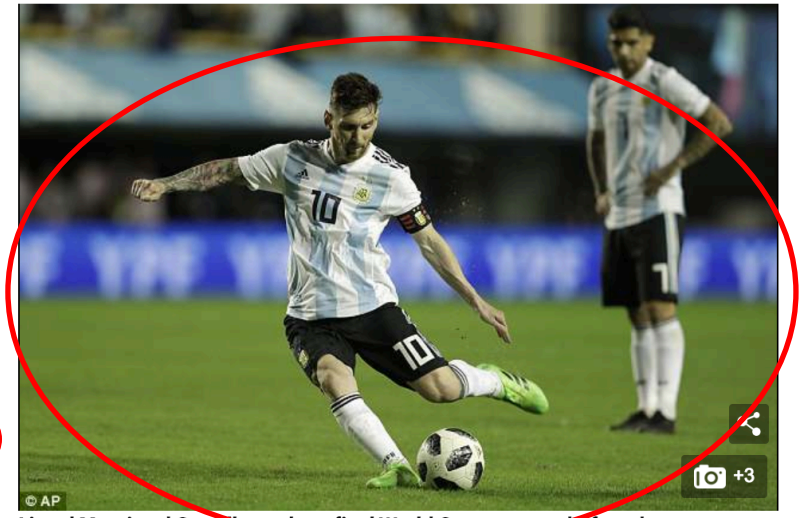
'They've finally done the right thing,' Higuain told ESPN, confirming reports the game had been cancelled.

The match at Jerusalem's Teddy Stadium was to be Argentina's last before they kick off their World Cup campaign in **Russia** on June 16.

文本输入

文本摘要

The match at Jerusalem's Teddy Stadium was to be Argentina's last before they kick off their World Cup campaign in **Russia** on June 16.



Lionel Messi and Co will not play a final World Cup warm-up before the tournament starts

图片-文本描述匹配对

异步多模态摘要-自动评价

➢ 实验设置

✓ 数据集构造（数据集统计结果）

表 图文式自动摘要数据集统计信息

	训练集	验证集	测试集
文档数	293,965	10,355	10,261
图片标题数	1,928,356	68,520	71,509
原文平均词数	720.87	766.08	730.80
参考摘要平均词数	70.12	70.02	72.16
图片标题平均词数	22.07	22.64	22.34
新闻平均图片数	6.56	6.62	6.97

异步多模态摘要-自动评价

- 实验（多模态自动摘要评价方法相关性实验）
  - ✓ 人工评价：
    - 将600样本划分为450（训练）和150（测试）
  - ✓ 映射函数
    - 线性回归（Linear Regression，LR）
    - 多层感知机（Multilayer Perceptron，MLP）
    - 逻辑斯蒂回归（Logistic Regression，Logis）

异步多模态摘要-自动评价

➢ 实验（多模态自动摘要评价方法相关性实验）

表 测试集（150样本）相关性结果

指标	皮尔逊系数	斯皮尔曼系数	肯德尔系数
ROUGE-L	0.3488	0.3554	0.2669
MAX_{sim}	0.2541	0.2339	0.1773
IP	0.5982	0.5966	0.5485
MMAE-LR	**0.6646**	0.6644	0.5265
MMAE-MLP	0.6632	0.6646	0.5265
MMAE-Logis	0.6630	**0.6653**	**0.6277**

- 多模态自动摘要评价方法显著优于单模态或跨模态的单一评价指标

总结与展望

- 多模态摘要方法变迁：
  - 抽取式方法→生成式方法→图文生成式方法
- 面向图文摘要的自动评价：
  - 综合考虑文本和图片重要性以及文本图片相关性
  - 与人工打分更加相关
- 未来方向：
  - 多模态语义表示和对齐
  - 扩展至更多的实际应用场景

文献与数据资源

Haoran Li, Junnan Zhu, Cong Ma, Jiajun Zhang and Chengqing Zong. 2017. Multi-modal Summarization for Asynchronous Collection of Text, Image, Audio and Video. ***In Proc. of EMNLP 2017***.

Haoran Li, Junnan Zhu, Cong Ma, Jiajun Zhang and Chengqing Zong. 2019. Read, Watch, Listen and Summarize: Multi-modal Summarization for Asynchronous Text, Image, Audio and Video. ***IEEE TKDE 2019***.

Haoran Li, Junnan Zhu, Tianshang Liu, Jiajun Zhang, and Chengqing Zong. 2018. Multi-modal Sentence Summarization with Modality Attention and Image Filtering. ***In Proc. of IJCAI 2018***.

Haoran Li, Junnan Zhu, Jiajun Zhang, Xiaodong He and Chengqing Zong. 2020. Multimodal Sentence Summarization via Multimodal Selective Encoding. ***In Proc. of COLING 2020***.

Junnan Zhu, Haoran Li, Tianshang Liu, Yu Zhou, Jiajun Zhang and Chengqing Zong. 2018. MSMO: Multimodal Summarization with Multimodal Output. ***In Proc. of EMNLP 2018***.

Junnan Zhu, Yu Zhou, Jiajun Zhang, Haoran Li, Chengqing Zong and Changliang Li. 2020. Multimodal Summarization with Guidance of Multimodal Reference. ***In Proc. of AAAI 2020***.

所有多模态数据资源：http://www.nlpr.ia.ac.cn/cip/dataset.htm


NLPR
LABORATORY OF PATTERN RECOGNITION
中国科学院自动化研究所
CASIA

谢谢！

Thanks!