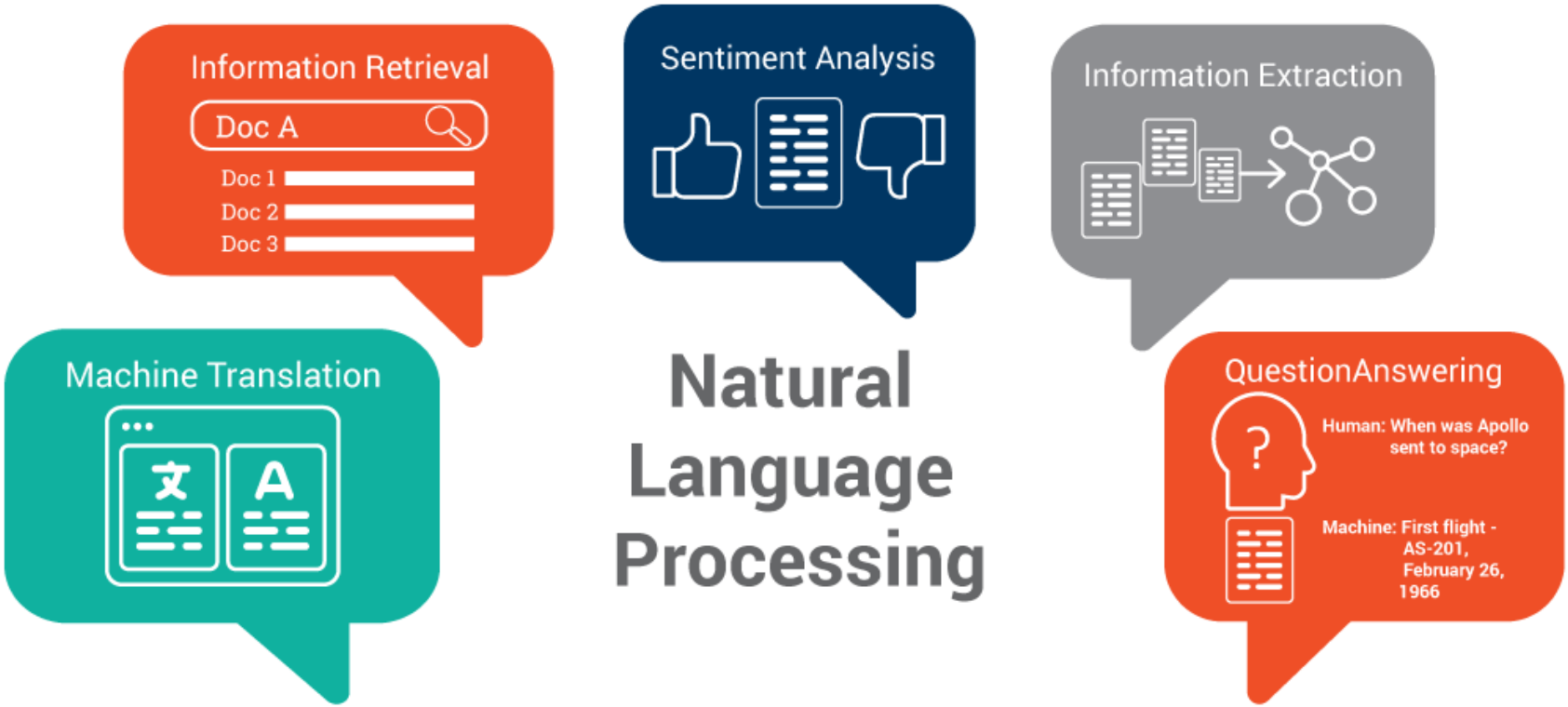
UNIVERSITY OF CAMBRIDGE

MLMI14: Spoken Document and Meeting Summarisation

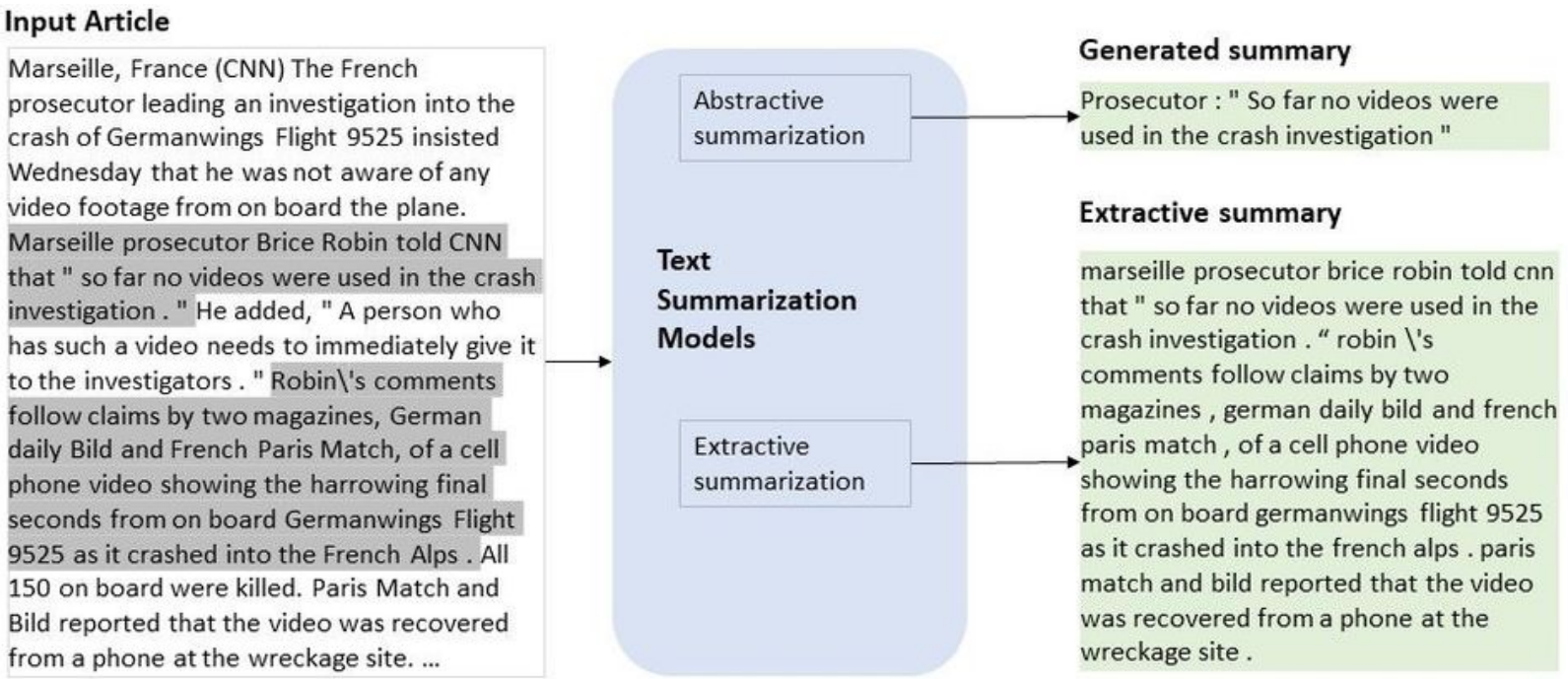
Mark Gales

Lent 2021

Natural Language Processing Applications



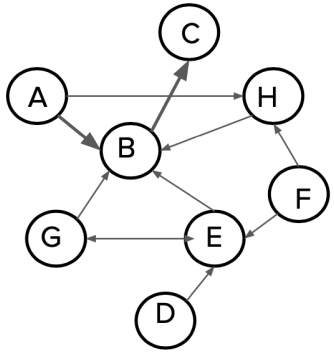
Summarisation



- Two forms of summarisation normally considered:
  - extractive: most relevant phrases from document
  - abstractive: given the document generate a summary

Summarisation

Two Kinds of Approaches to Podcast Summarization

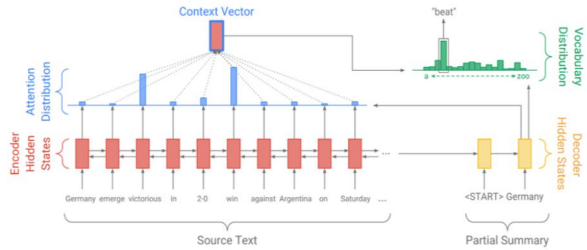


Extractive Summarization:

Identify and select phrases or sentences within the podcast that contain the most salient content
[eg. TextRank, LexRank]

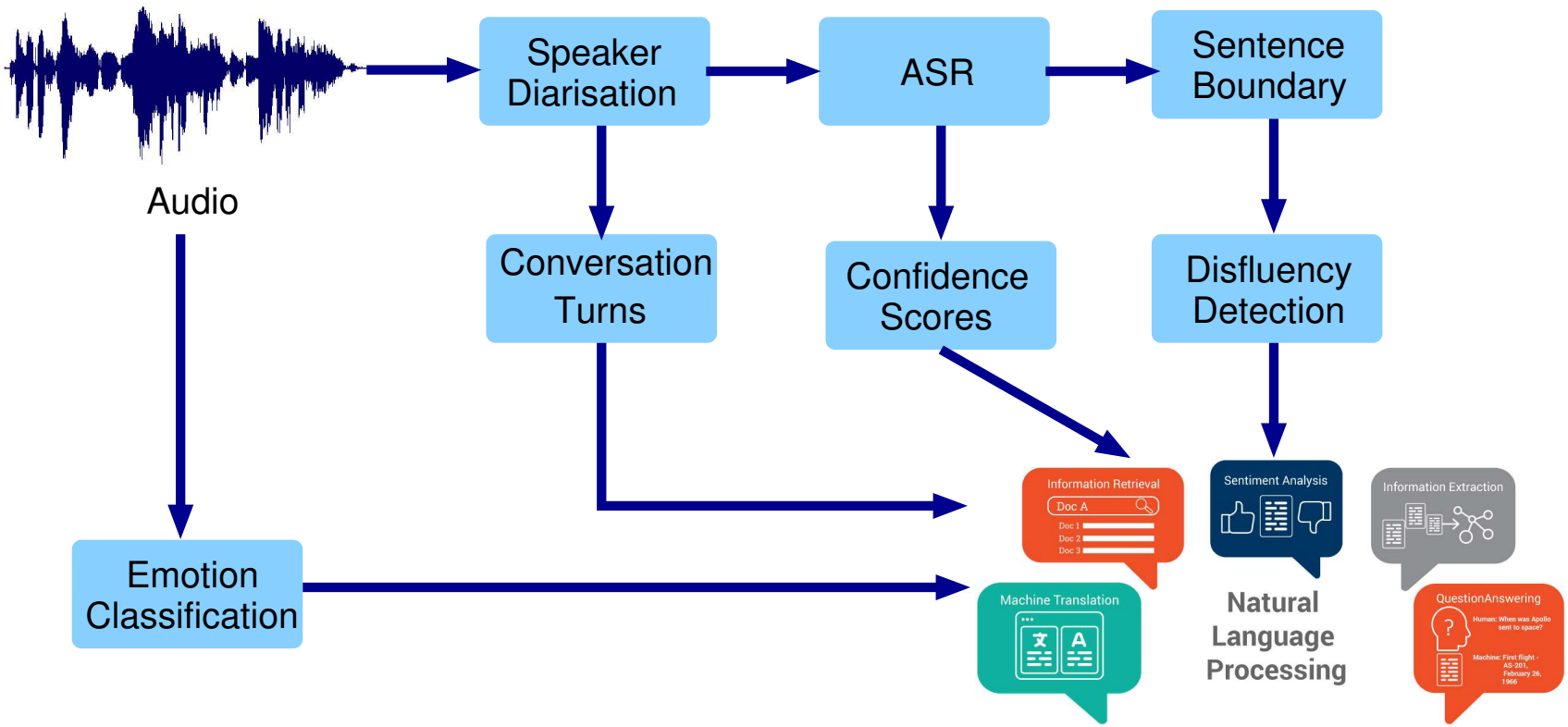
Abstractive Summarization:

Generate new summary content
eg. Neural generative summarization

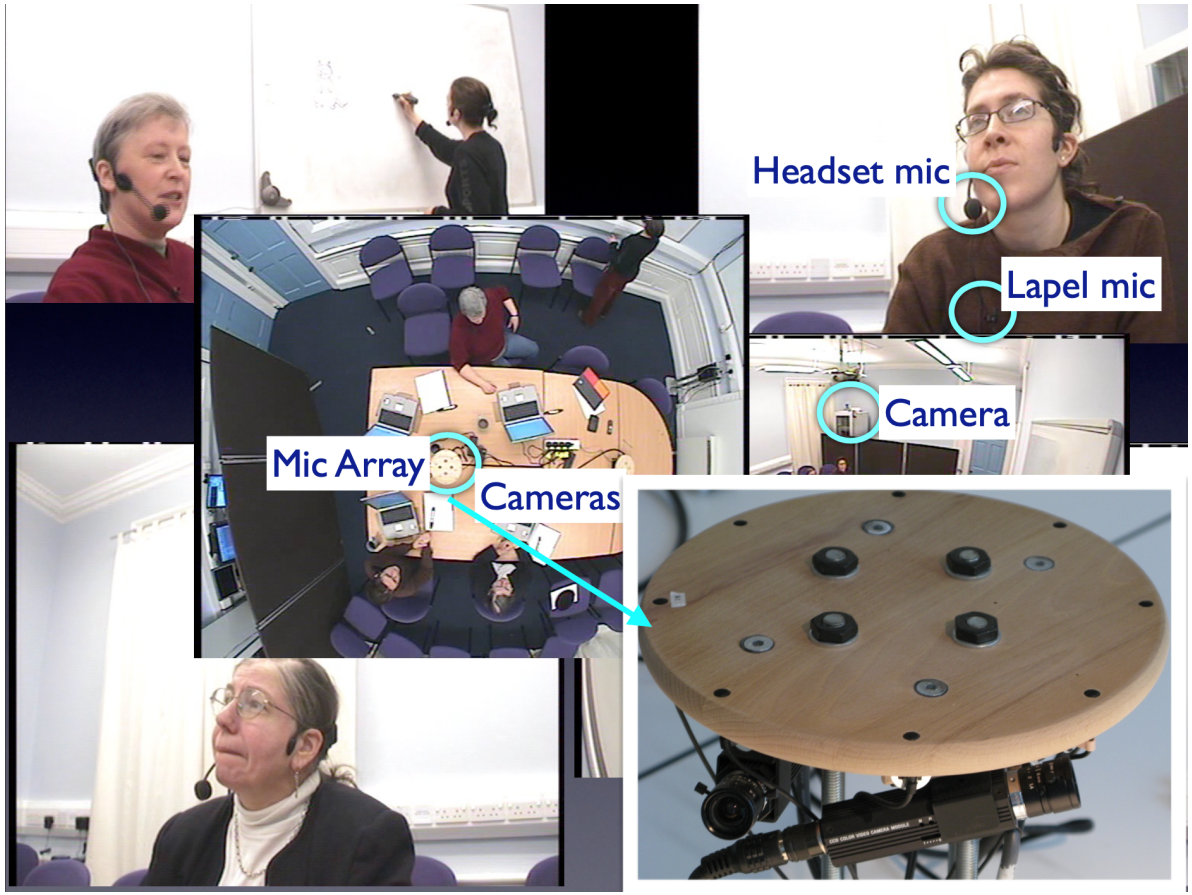


- This talk will focus on abstractive summarisation
  - sequence-to-sequence models state-of-the-art approach
  - extractive summarisation a sequence labelling task

Spoken Language Processing Framework



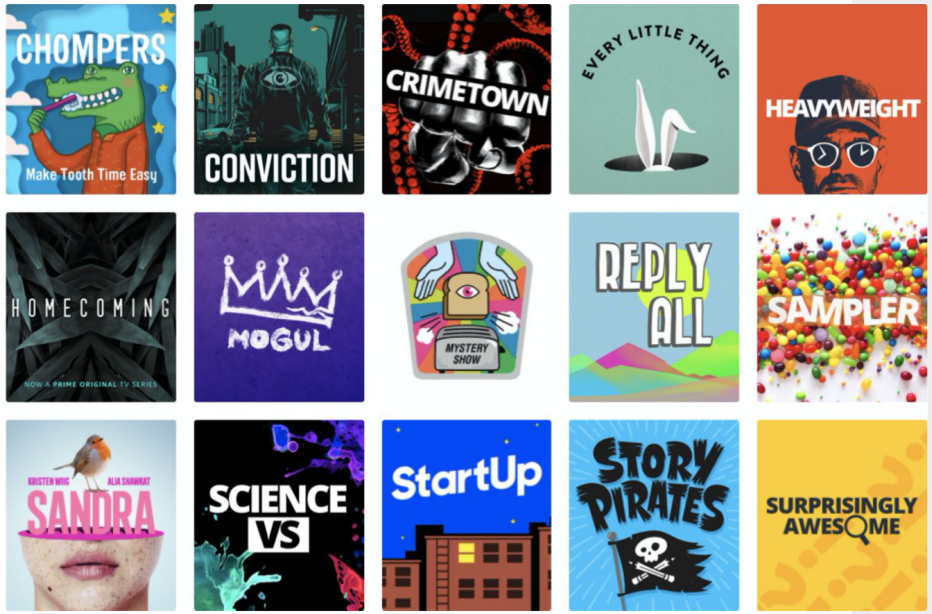
Meeting Summarisation: AMI Meeting Corpus [2]



- Standard task from the AMI corpus data
  - limited quantity of data, extensively annotated

Spotify Podcast Challenge [3]

A New English-Language Corpus

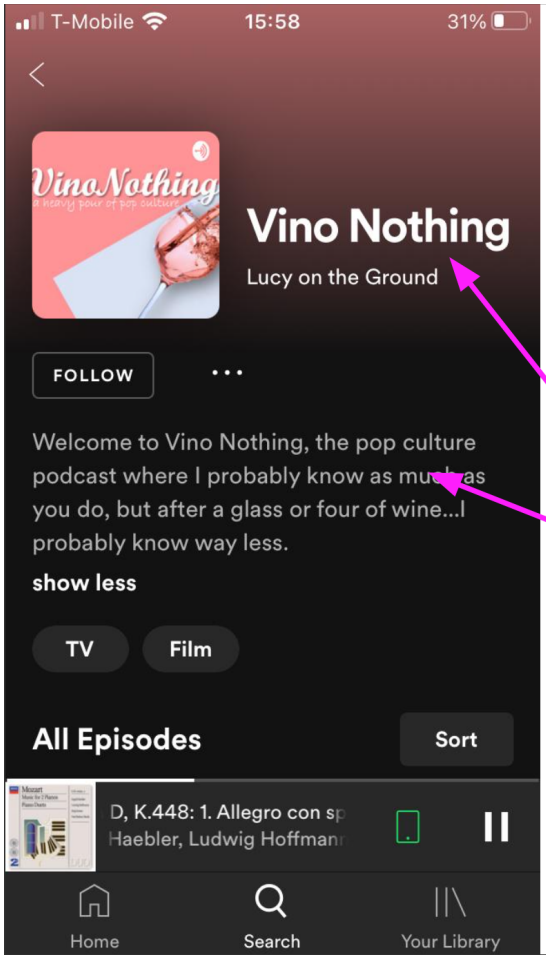


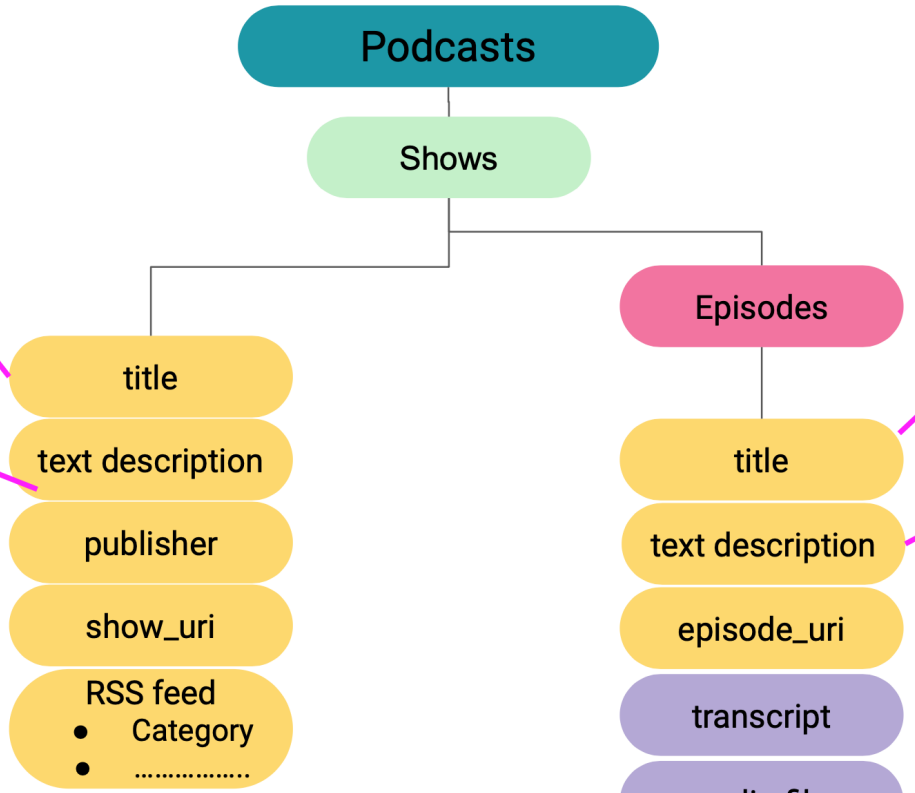
We released the first fully-transcribed, large-scale podcast dataset

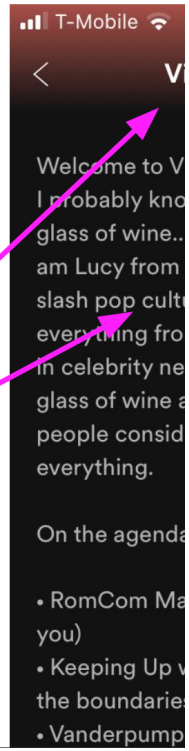
2TB of data - 100k episodes with audio

- Large-scale corpus, audio and "summaries" supplied
  - ASR generated transcriptions (18.1%) also available

Spotify Podcast Data





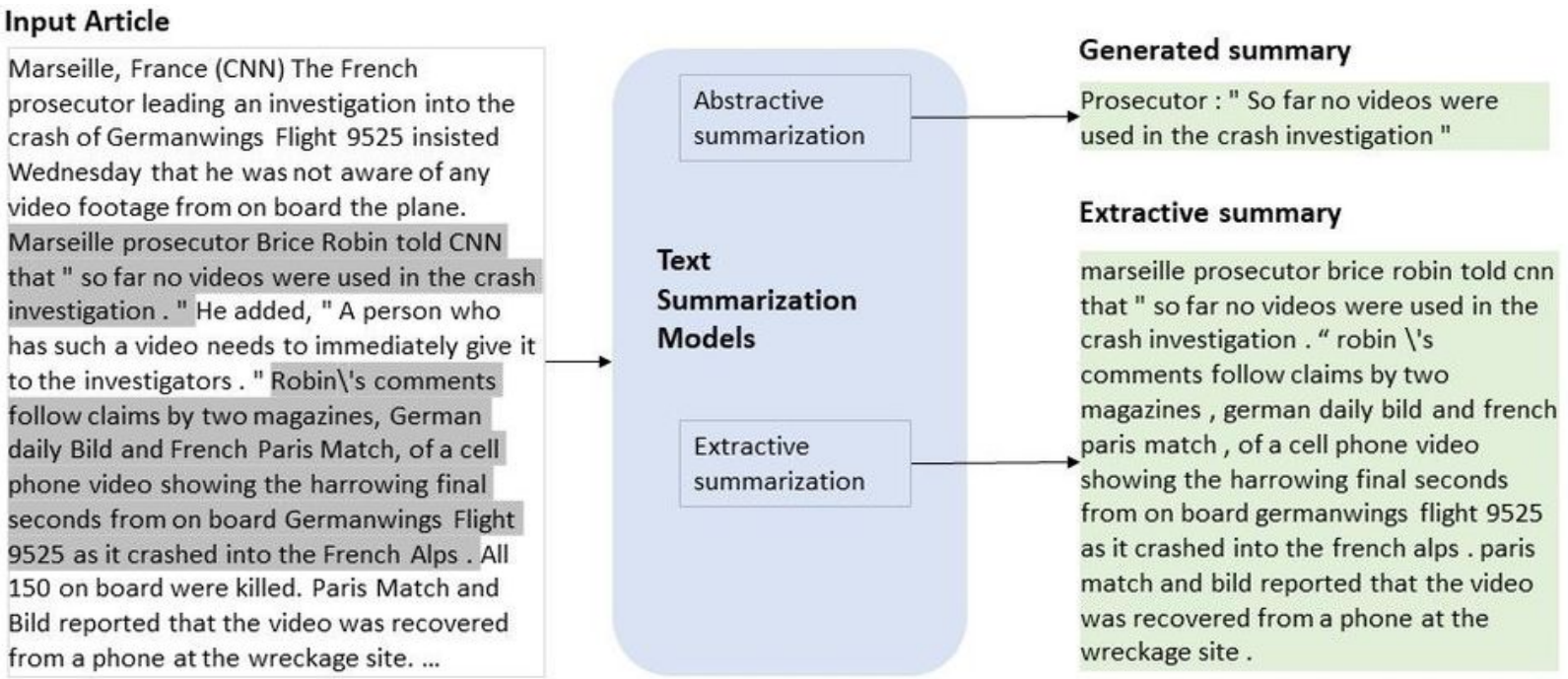


Summarisation Assessment [6]

- In common with many NLP tasks evaluation is challenging:
  - ideally human subjective evaluation - expensive
- Often used automatic metric ROUGE: N-Gram overlap
  Recall-Oriented Understudy for Gisting Evaluation
  - ROUGE-1: unigram overlap counts
  - ROUGE-2: bigram overlap counts
  - ROUGE-L: longest matching sequence
- ROUGE can yield N-Gram recall and precision metrics
  - can compute combined F-scores F_α

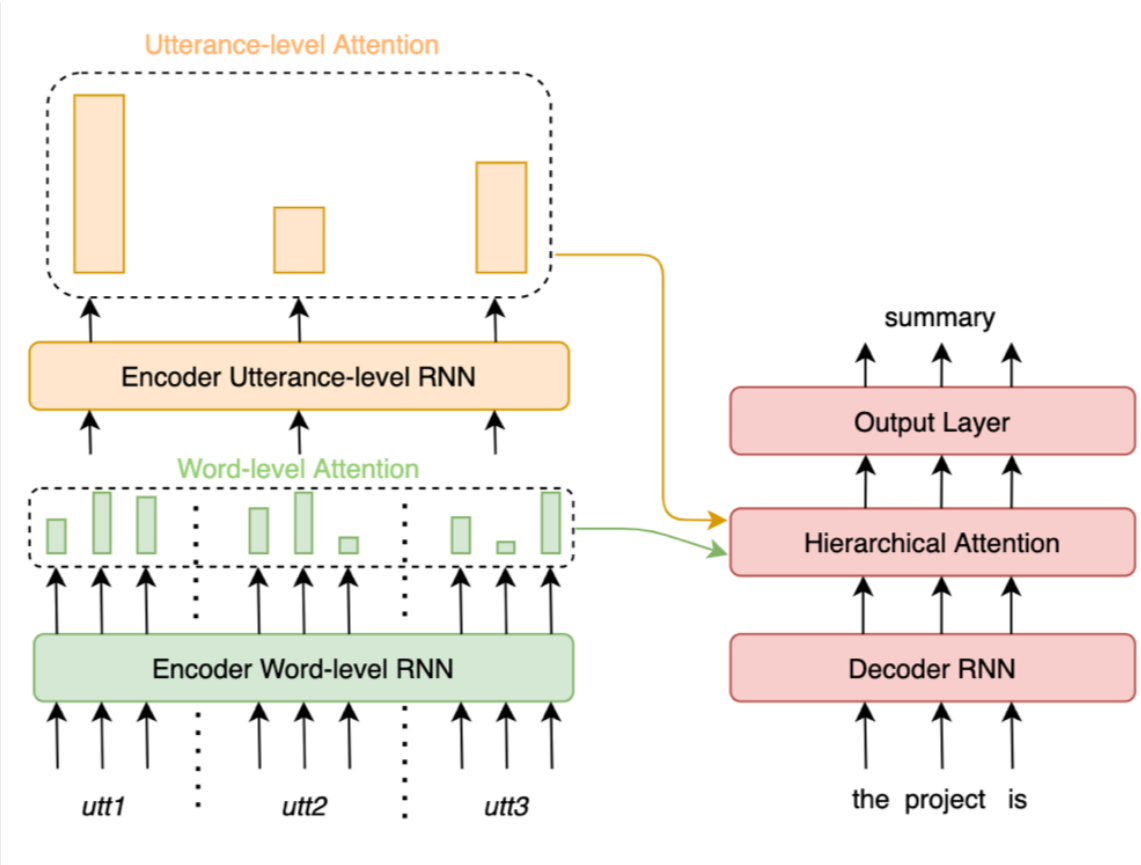
Abstractive Summarisation

Summarisation



- Two forms of summarisation normally considered:
  - extractive: most relevant phrases from document
  - abstractive: given the document generate a summary

Hierarchical Model [7]



- Utterances expected to influence summaries differently
  - combine word and utterance level attention-mechanism
  - integrates extractive-elements into abstractive process

Hierarchical Attention Mechanism

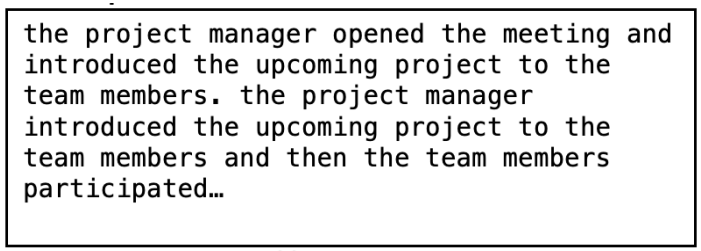
- Encoder uses word and utterance encoders
  - $\mathbf{h}_i^{\mathtt{u}}$ encoding of i^{th} document utterance
  - $\mathbf{h}_{ij}^{\mathtt{w}}$ encoding of j^{th} word of $i-th$ document utterance
- Decoder at step t attends (decoder state $\mathbf{d}_t$)
  - Utterance-level Attention: for the i^{th} utterance

$$\alpha_{ti}^{\mathtt{u}} = \frac{\exp(\mathbf{d}_t^{\mathtt{T}} \mathbf{W}^{\mathtt{u}} \mathbf{h}_i^{\mathtt{u}})}{\sum_{i'} \exp(\mathbf{d}_t^{\mathtt{T}} \mathbf{W}^{\mathtt{u}} \mathbf{h}_{i'}^{\mathtt{u}})}$$

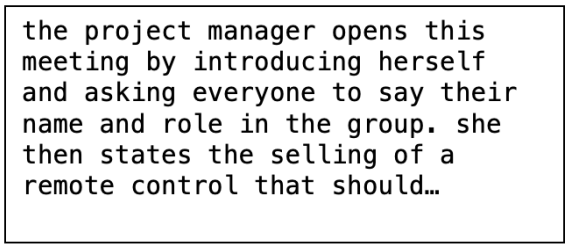
  - Word-level Attention: j^{th} word of i^{th} utterance

$$\alpha_{tij}^{\mathtt{w}} = \alpha_{ti}^{\mathtt{u}} \left(\frac{\exp(\mathbf{d}_t^{\mathtt{T}} \mathbf{W}^{\mathtt{w}} \mathbf{h}_{ij}^{\mathtt{w}})}{\sum_{j'} \exp(\mathbf{d}_t^{\mathtt{T}} \mathbf{W}^{\mathtt{w}} \mathbf{h}_{ij'}^{\mathtt{w}})} \right)$$

Impact of Limited Data



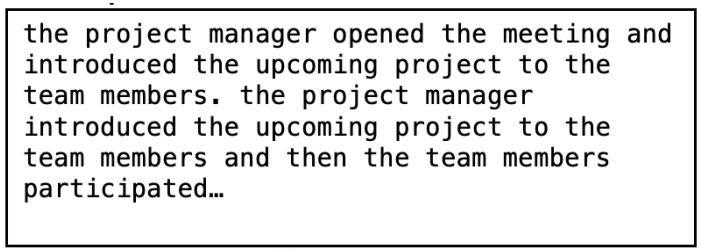
repetitive summary



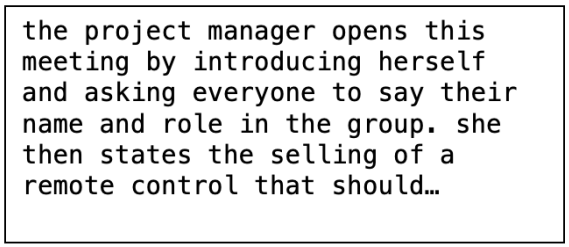
more diverse summary

- Tasks such as AMI have very limited training data
  - resulting summaries tend to lack diversity between sentences

Impact of Limited Data



repetitive summary



more diverse summary

- Tasks such as AMI have very limited training data
  - resulting summaries tend to lack diversity between sentences
- Use document utterance attention (α_t^{u}) to measure diversity
  - intra-sentence: for summary sentence k, length T_k

$$D_{\text{intra},k} = \frac{1}{\binom{T_k}{2}} \sum_{t_1=1}^{T_k-1} \sum_{t_2=t_1+1}^{T_k} \|\alpha_{t_1}^{\mathrm{u}} - \alpha_{t_2}^{\mathrm{u}}\|_2; \quad D_{\text{intra}} = \frac{1}{K} \sum_{k=1}^{K} D_{\text{intra},k}$$

  - inter-sentence: over K summary sentences

$$\bar{\alpha}_k^{\mathrm{u}} = \frac{1}{T_k} \sum_{t=1}^{T_k} \alpha_t^{\mathrm{u}}, \quad D_{\text{inter}} = \frac{1}{\binom{K}{2}} \sum_{k_1=1}^{K-1} \sum_{k_2=k_1+1}^{K} \|\bar{\alpha}_{k_1}^{\mathrm{u}} - \bar{\alpha}_{k_2}^{\mathrm{u}}\|_2$$

Multi-Task Learning

- Modify criterion, diversity and multi-task learning: minimise

$$\mathcal{L}(\boldsymbol{\theta}) = -\mathcal{L}_{\texttt{ml}}(\boldsymbol{\theta}) - \underbrace{(\lambda_1 \mathcal{L}_{\texttt{da}}(\boldsymbol{\theta}) + \lambda_2 \mathcal{L}_{\texttt{ex}}(\boldsymbol{\theta}))}_{\textit{Multi-Task}} + \underbrace{\lambda_1 \mathcal{L}_{\texttt{dv}}(\boldsymbol{\theta})}_{\textit{Diversity}}$$

  - $\mathcal{L}_{\texttt{ml}}(\boldsymbol{\theta})$: maximise likelihood of summary
  - $\mathcal{L}_{\texttt{da}}(\boldsymbol{\theta})$: maximise likelihood of dialogue act
  - $\mathcal{L}_{\texttt{ex}}(\boldsymbol{\theta})$: maximise likelihood of extractive summary
  - $\mathcal{L}_{\texttt{dv}}(\boldsymbol{\theta})$: maximise summary diversity

$$\mathcal{L}_{\texttt{dv}}(\boldsymbol{\theta}) = \frac{D_{\text{inter}}}{D_{\text{intra}}}$$

Diversity at Inference Time

- During inference encourage diversity
  - use simple unigram measure of diversity

$$\hat{y}_t = \arg\max_{y \in \mathcal{V}} \left\{ \log P_t(y) - \beta \left(\frac{\sum_{\tau=1}^{t-1} \delta(\hat{y}_\tau, y)}{t-1} \right) \right\}$$

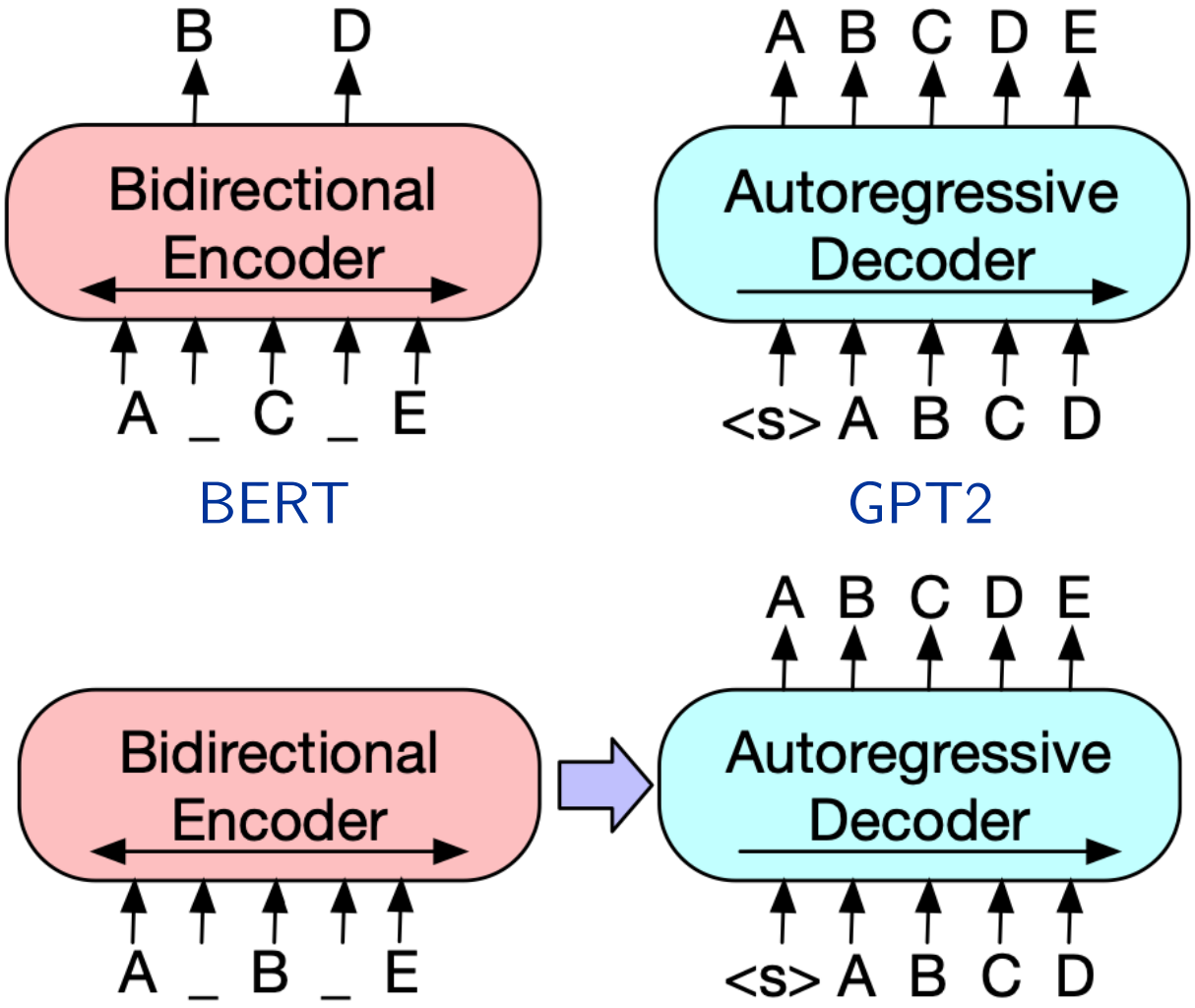
  where
  - β is the unigram bias constant
  - $\mathcal{V}$ is the decoding vocabulary
  - $\delta(\hat{y}_\tau, y)$ is 1 when $y = \hat{y}_\tau$.
- Word blocking is also popular for diversity
  - if predicted word y would yield a seen N-gram: set $P_t(y) = 0$

Performance on AMI Corpus

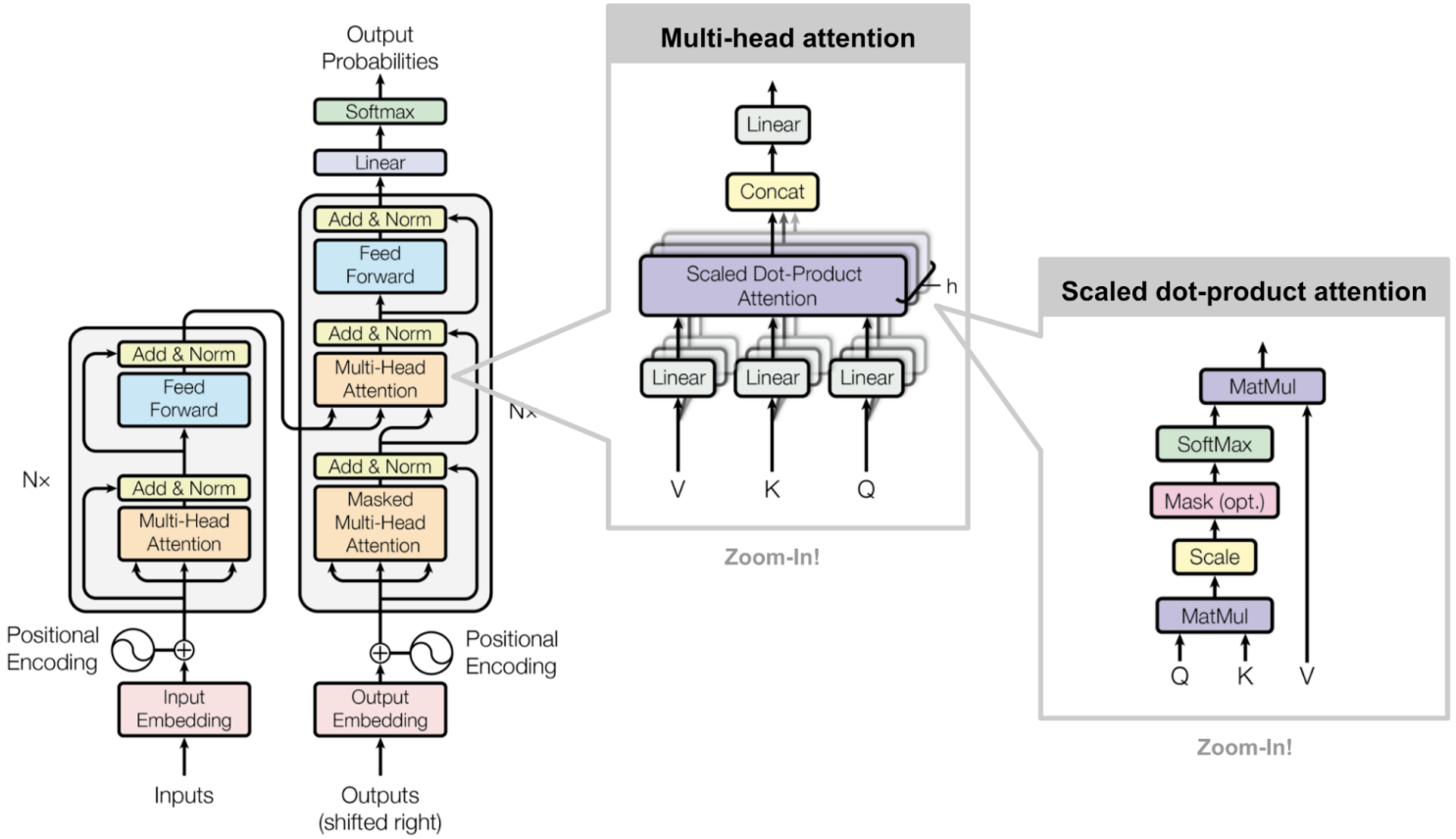
Setting	β	ROUGE-1	ROUGE-2	ROUGE-L
HIER	—	39.12±2.45	13.03±1.58	36.77±2.28
	20.0	42.39±1.91	11.60±1.60	39.27±1.96
HIER+MT	—	40.13±0.93	13.59±0.66	38.00±0.74
	5.0	42.69±1.03	13.94±0.86	39.94±0.94
HIER+DIV	—	41.94±0.25	12.87±0.28	39.30±0.56
	1.00	42.84±1.10	12.94±0.50	39.79±1.83
HIER+MT+DIV	—	44.46±0.11	14.51±0.12	41.12±0.13
	0.25	**44.36±0.58**	**14.62±0.21**	**41.10±0.25**

Table: ROUGE F_1 on the AMI test set - Hierarchical Settings.

Pre-Trained Model: BART [5]



Transformer: Reminder [10]



- standard sequence-to-sequence transformer model:
  - includes positional encoding as only self-attention is used

BART for Summarisation

- BART is suitable for abstractive summarisation
  - maximum positional encoding of 1024

Dataset	#Doc	Input	90^{th}%	Target
Podcast	106k	5,727	11,677	61.1

  - how to handle input podcasts/documents of longer length?
  - simplest approach is to truncate the length of the input

BART for Summarisation

- BART is suitable for abstractive summarisation
  - maximum positional encoding of 1024

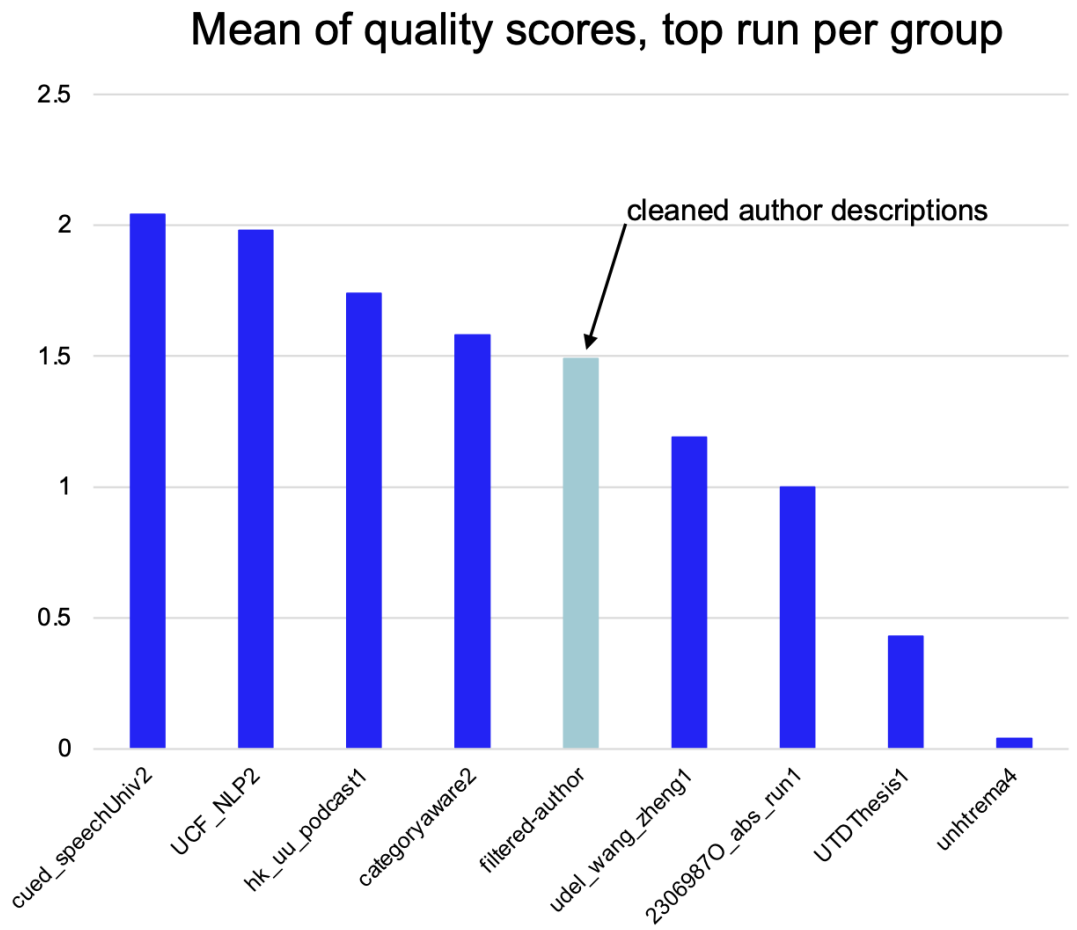
Dataset	#Doc	Input	90^{th}%	Target
Podcast	106k	5,727	11,677	61.1

  - how to handle input podcasts/documents of longer length?
  - simplest approach is to truncate the length of the input
- Exploit the Hierarchical Model utterance-level attention $\boldsymbol{\alpha}_t^{\mathrm{u}}$
  - compute average attention over generated T-length summary

$$\overline{\alpha}_i^{\mathrm{u}} = \frac{1}{T} \sum_{t=1}^{T} \alpha_{ti}^{\mathrm{u}}$$

  - simply select highest average attention input utterances
- Fine-tune BART model on this selected input

Spotify Results



- Human assessment of quality of summary
  - Excellent (4), Good (2), Fair (1), Bad (0)

Long-Span Dependencies

Document Length [4]

Dataset	#Doc	Input	90^{th}%	Target
Podcast	106k	5,727	11,677	61.1
arXiv	216k	8,584	16,108	367
PubMed	133k	3,865	7,234	260

- Statistics for three summarisation tasks given above
  - input length (document) can be very long
  - significantly longer than machine-translation scenario

Computational and Memory Challenge for Transformers

Document Length [4]

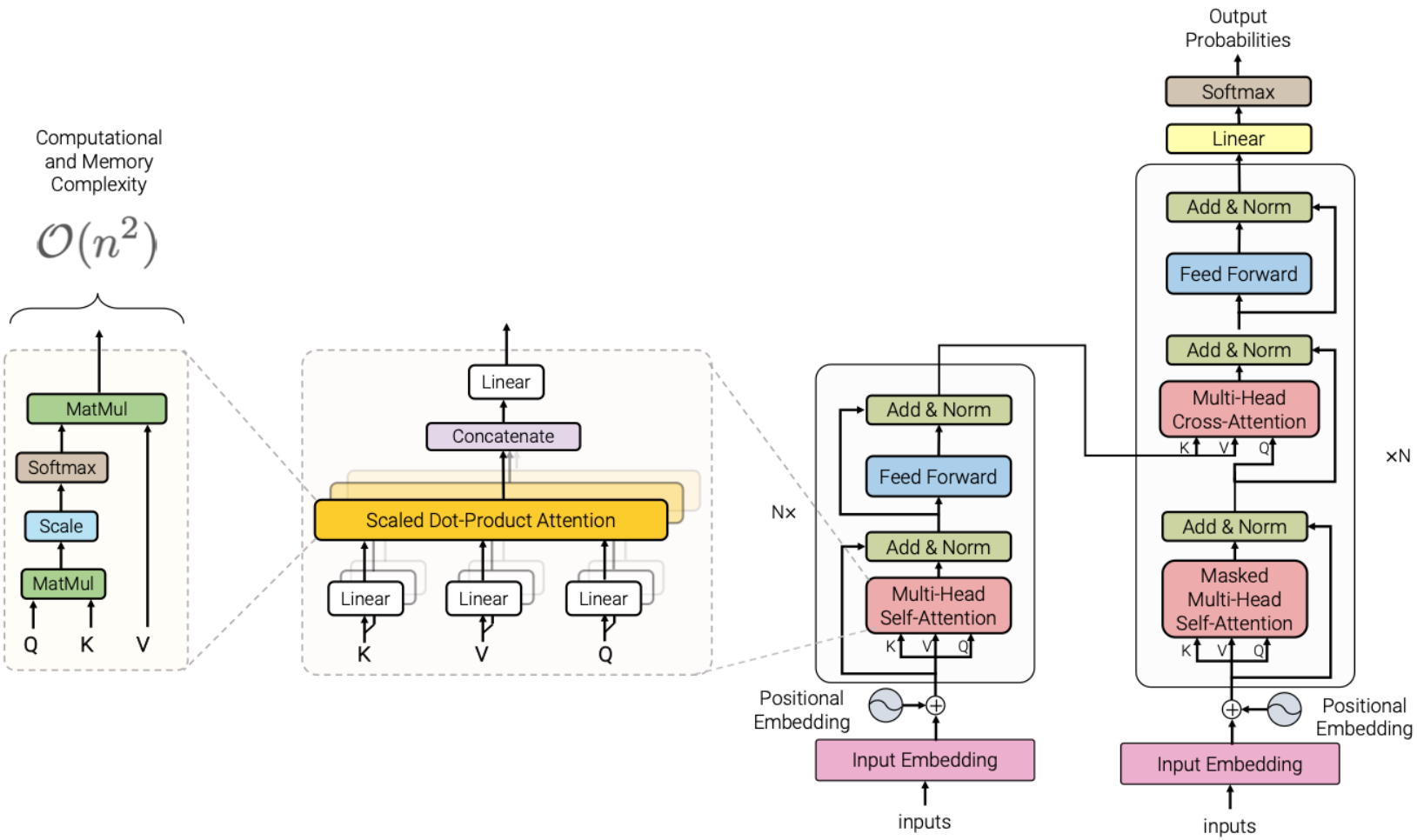
Dataset	#Doc	Input	90^{th}%	Target
Podcast	106k	5,727	11,677	61.1
arXiv	216k	8,584	16,108	367
PubMed	133k	3,865	7,234	260

- Statistics for three summarisation tasks given above
  - input length (document) can be very long
  - significantly longer than machine-translation scenario

  Computational and Memory Challenge for Transformers

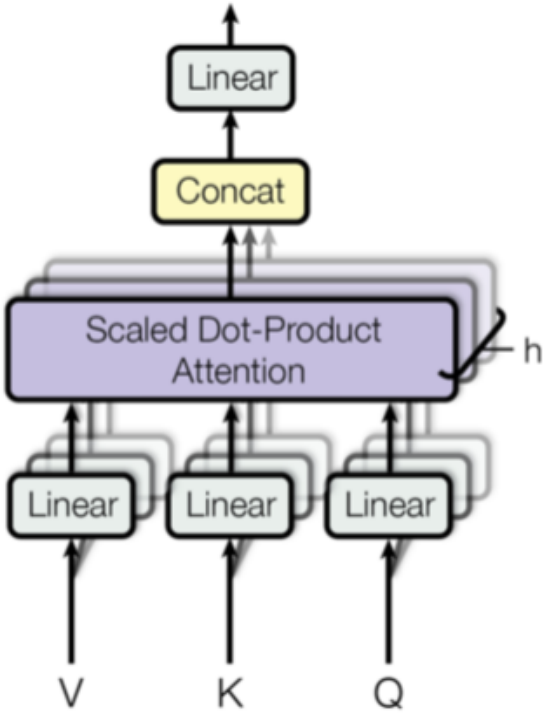
- Standard set of long-span tasks also available [8]

Transformer [10]



- Extending size of the positional encoding not sufficient
  - compute and memory increase at $\mathcal{O}(n^2)$: n input length

Encoder Multi-Head Self-Attention



h

- Notation
  - Q: query for attention
  - K: key for attention - length d_k
  - V: input values $(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n)$
- Scaled dot-product multi-head attention

$$\alpha_{\tau i}^{(j)} = \frac{1}{Z} \exp\left(\mathbf{k}_\tau^{(j)\mathsf{T}} \boldsymbol{q}_i^{(j)} / \sqrt{d_k^{(j)}}\right)$$

$$\boldsymbol{c}_i^{(j)} = \sum_{\tau=1}^{n} \alpha_{\tau i}^{(j)} \mathbf{W}_{\mathrm{v}}^{(j)} \boldsymbol{v}_\tau$$

$$\boldsymbol{c}_i = = \left[\begin{array}{ccc} \boldsymbol{c}_i^{(1)\mathsf{T}} & \ldots & \boldsymbol{c}_i^{(J)\mathsf{T}} \end{array} \right]^{\mathsf{T}}$$

- Self-attention $Q = K = V$

$$\boldsymbol{k}_i^{(j)} = \mathbf{W}_{\mathrm{k}}^{(j)} \boldsymbol{v}_i; \quad \boldsymbol{q}_i^{(j)} = \mathbf{W}_{\mathrm{q}}^{(j)} \boldsymbol{v}_i;$$

Computational Issue: Self-Attention

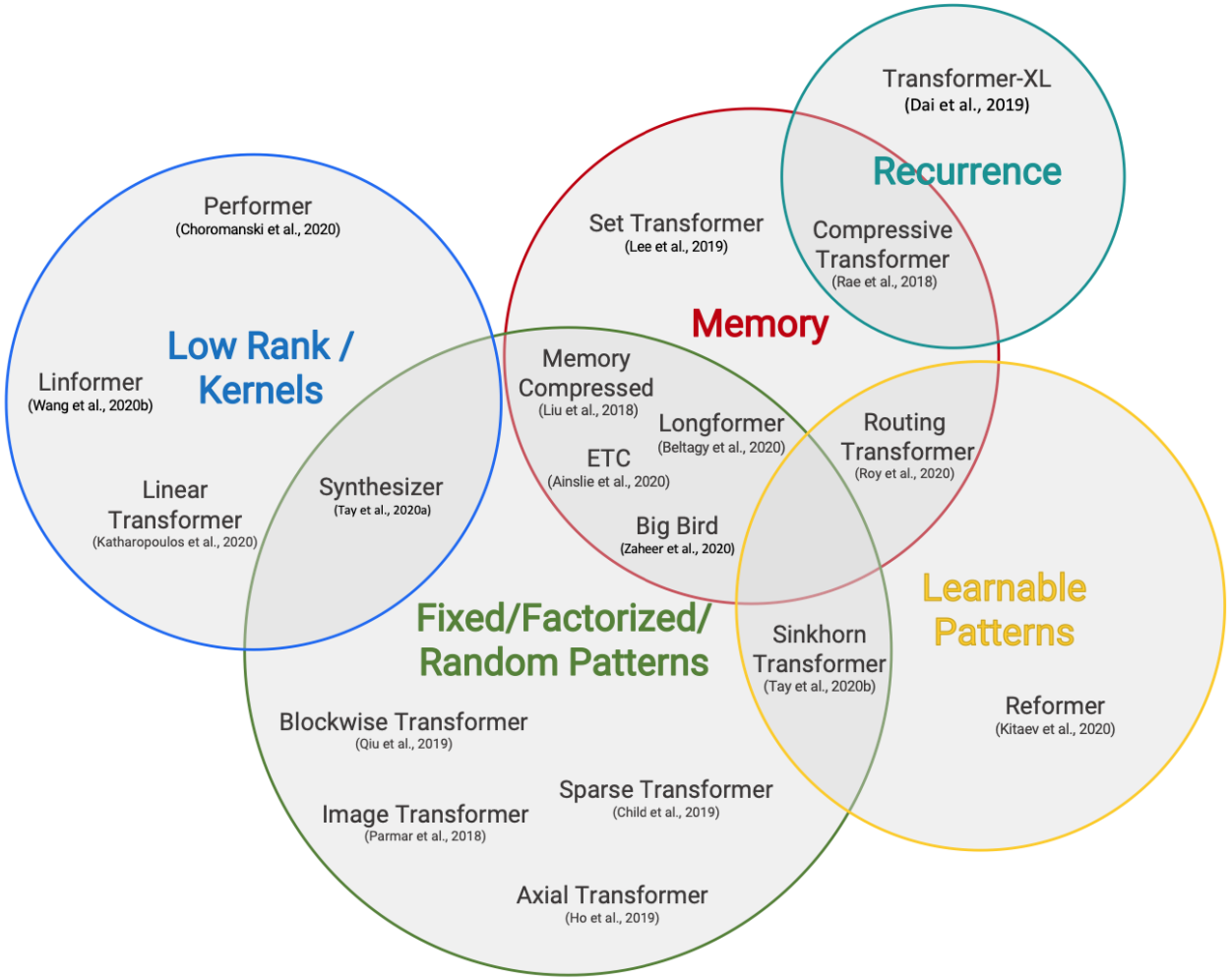
- Consider computation cost of self-attention:
  - For each location i compute $(n\times)$
    - for each attention mechanism j compute $\boldsymbol{c}_i^{(j)}$:

$$\boldsymbol{c}_i^{(j)} = \sum_{\tau=1}^{n} \alpha_{\tau i}^{(j)} \mathbf{W}_{\mathrm{v}}^{(j)} \boldsymbol{v}_\tau$$

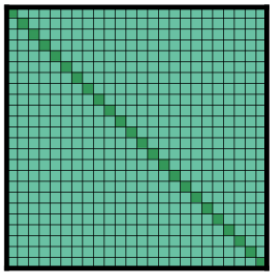
  - Yields computational cost cost $\mathcal{O}(n^2)$ for length n
- Also memory issues $\mathcal{O}(n^2)$ using PyTorch `autograd`

- For summarisation can use selection approaches
  - what about more general approaches?

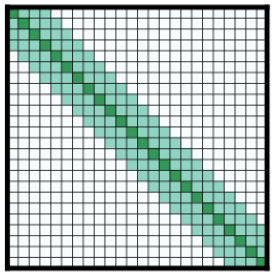
Efficiency Approaches [9]



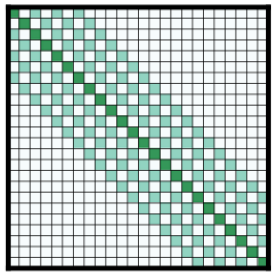
Longformer Attention [1]



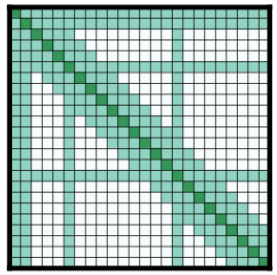
(a) Full n^2 attention



(b) Sliding window attention



(c) Dilated sliding window



(d) Global+sliding window

- Simplest approach is to restrict span of attention mechanism
  - sliding window uses span $2\beta + 1$

$$\boldsymbol{c}_i^{(j)} = \sum_{\tau=i-\beta}^{i+\beta} \alpha_{\tau i}^{(j)} \mathbf{W}_{\mathrm{v}}^{(j)} \boldsymbol{v}_\tau$$

  - dilation modifies attention "receptive field"
- Can combine with (limited) global attention mechanisms
  - only specific locations have complete attention span

[1] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," 2020.

[2] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner, "The AMI meeting corpus: A pre-announcement," in *Machine Learning for Multimodal Interaction, Second International Workshop, MLMI 2005, Edinburgh, UK, July 11-13, 2005, Revised Selected Papers*, ser. Lecture Notes in Computer Science, S. Renals and S. Bengio, Eds., vol. 3869. Springer, 2005, pp. 28–39. [Online]. Available: https://doi.org/10.1007/11677482_3

[3] A. Clifton, S. Reddy, Y. Yu, A. Pappu, R. Rezapour, H. Bonab, M. Eskevich, G. Jones, J. Karlgren, B. Carterette, and R. Jones, "100,000 podcasts: A spoken English document corpus," in *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 5903–5917. [Online]. Available: https://www.aclweb.org/anthology/2020.coling-main.519

[4] A. Cohan, F. Dernoncourt, D. S. Kim, T. Bui, S. Kim, W. Chang, and N. Goharian, "A discourse-aware attention model for abstractive summarization of long documents," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 615–621. [Online]. Available: https://www.aclweb.org/anthology/N18-2097

[5] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 7871–7880. [Online]. Available: https://www.aclweb.org/anthology/2020.acl-main.703

[6] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, July 2004, pp. 74–81. [Online]. Available: https://www.aclweb.org/anthology/W04-1013

[7] P. Manakul, M. J. F. Gales, and L. Wang, "Abstractive spoken document summarization using hierarchical model with multi-stage attention diversity optimization," in *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, H. Meng, B. Xu, and T. F. Zheng, Eds. ISCA, 2020, pp. 4248–4252. [Online]. Available: https://doi.org/10.21437/Interspeech.2020-1683

[8] Y. Tay, M. Dehghani, S. Abnar, Y. Shen, D. Bahri, P. Pham, J. Rao, L. Yang, S. Ruder, and D. Metzler, "Long range arena: A benchmark for efficient transformers," 2020.

[9] Y. Tay, M. Dehghani, D. Bahri, and D. Metzler, "Efficient transformers: A survey," 2020.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *CoRR*, vol. abs/1706.03762, 2017. [Online]. Available: http://arxiv.org/abs/1706.03762