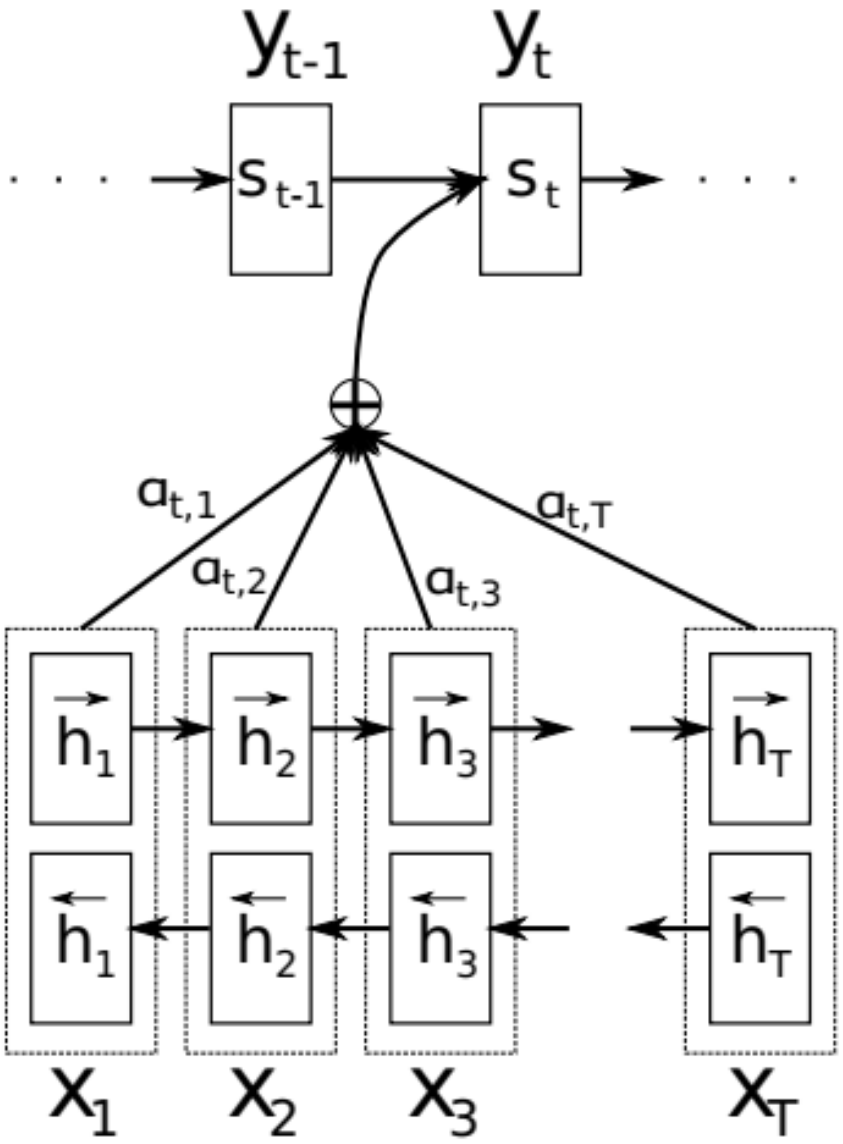
CS772: Deep Learning for Natural Language Processing (DL-NLP)

Attention; Summarization

Pushpak Bhattacharyya
Computer Science and Engineering Department
IIT Bombay

Week 11 of 18mar24

1-slide recap (Bahadanu et al, 2015)



$$h_t = f(x_t, h_{t-1})$$

$$c = g(\{h_1, h_2, h_3, \ldots h_{T_x}\})$$

$$P(\bar{y}) = \prod_{t=1}^{T} P(y_t \mid \{y_1, y_2, y_3, \ldots y_{t-1}\}, c)$$

$$P(y_t \mid \{y_1, y_2, y_3, \ldots y_{t-1}\}, c) = g(y_{t-1}, s_t, c)$$

$$p(y_i \mid y_1, y_2, \ldots y_{i-1}) = g(y_{i-1}, s_i, c_i)$$

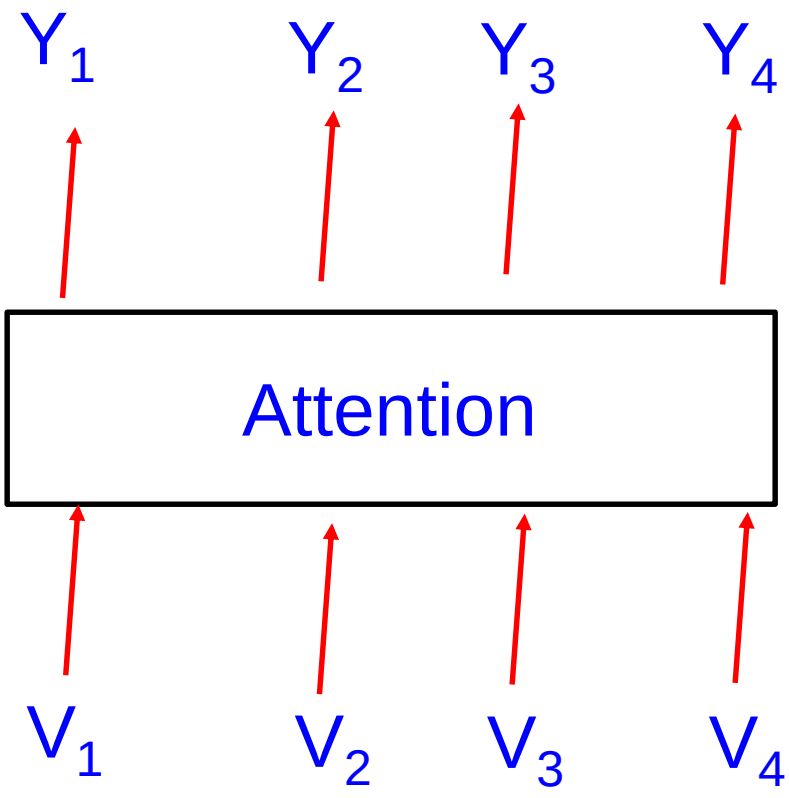
$$s_i = f(s_{i-1}, y_{i-1}, c_i)$$

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})}$$

$$e_{ij} = a(s_{i-1}, h_j)$$

Self Attention Block



Bank of the river

Word Embedding and Contextual Word Embedding

- Consider the phrase "*bank of the river*"
- Word embeddings of '*bank*', '*of*', '*the*', '*river*': V_1, V_2, V_3, V_4
- Now create a 'score' vector S_i for each word vector
- S_1: ($V_1.V_1$, $V_1.V_2$, $V_1.V_3$, $V_1.V_4$)
- Similarly, S_2, S_3, S_4

S-matrix

$$S = \begin{bmatrix} S_{11} & S_{12} & S_{13} & S_{14} \\ S_{21} & S_{22} & S_{23} & S_{24} \\ S_{31} & S_{32} & S_{33} & S_{34} \\ S_{41} & S_{42} & S_{43} & S_{44} \end{bmatrix}$$

S-scaled matrix

$$S - scaled = \frac{1}{\sqrt{d_k}} \times \begin{bmatrix} s_{11} & s_{12} & s_{13} & s_{14} \\ s_{21} & s_{22} & s_{23} & s_{24} \\ s_{31} & s_{32} & s_{33} & s_{34} \\ s_{41} & s_{42} & s_{43} & s_{44} \end{bmatrix}$$

W-matrix

$$W = \begin{bmatrix} w_{11} & w_{12} & w_{13} & w_{14} \\ w_{21} & w_{22} & w_{23} & w_{24} \\ w_{31} & w_{32} & w_{33} & w_{34} \\ w_{41} & w_{42} & w_{43} & w_{44} \end{bmatrix}$$

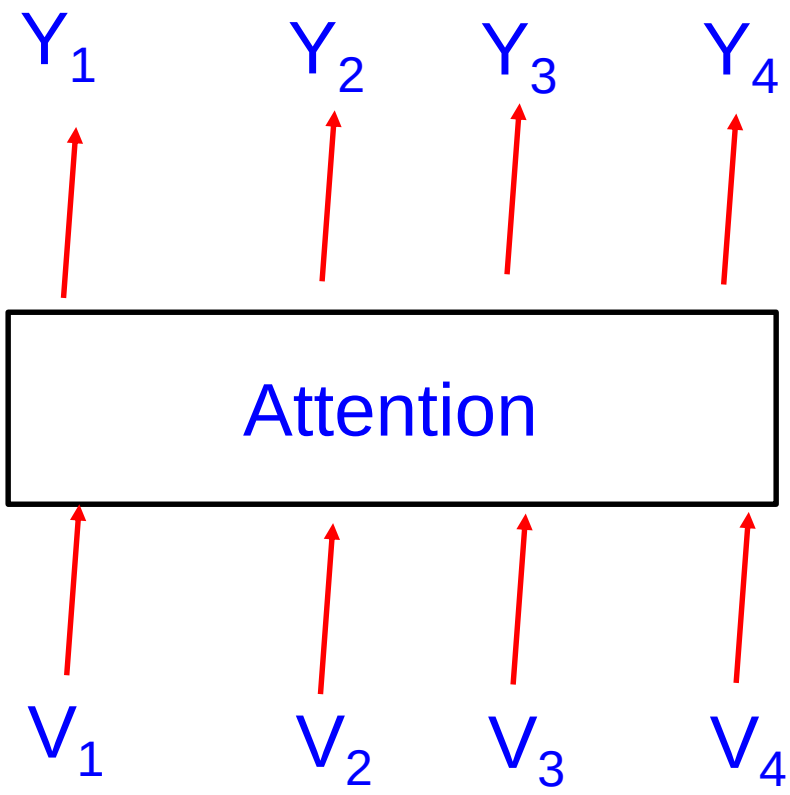
$$W_i - vector = soft\max\left(\frac{S_i - vector}{\sqrt{d_k}}\right)$$

Y-matrix

$$Y = \begin{bmatrix} y_{11}\, y_{12}\, y_{13}\, y_{14} \\ y_{21}\, y_{22}\, y_{23}\, y_{24} \\ y_{31}\, y_{32}\, y_{33}\, y_{34} \\ y_{41}\, y_{42}\, y_{43}\, y_{44} \end{bmatrix}$$

$$Y_i - vector = w_{11}.V_1 + w_{12}.V_2 + w_{13}.V_3 + w_{14}.V_4$$

Attention Block



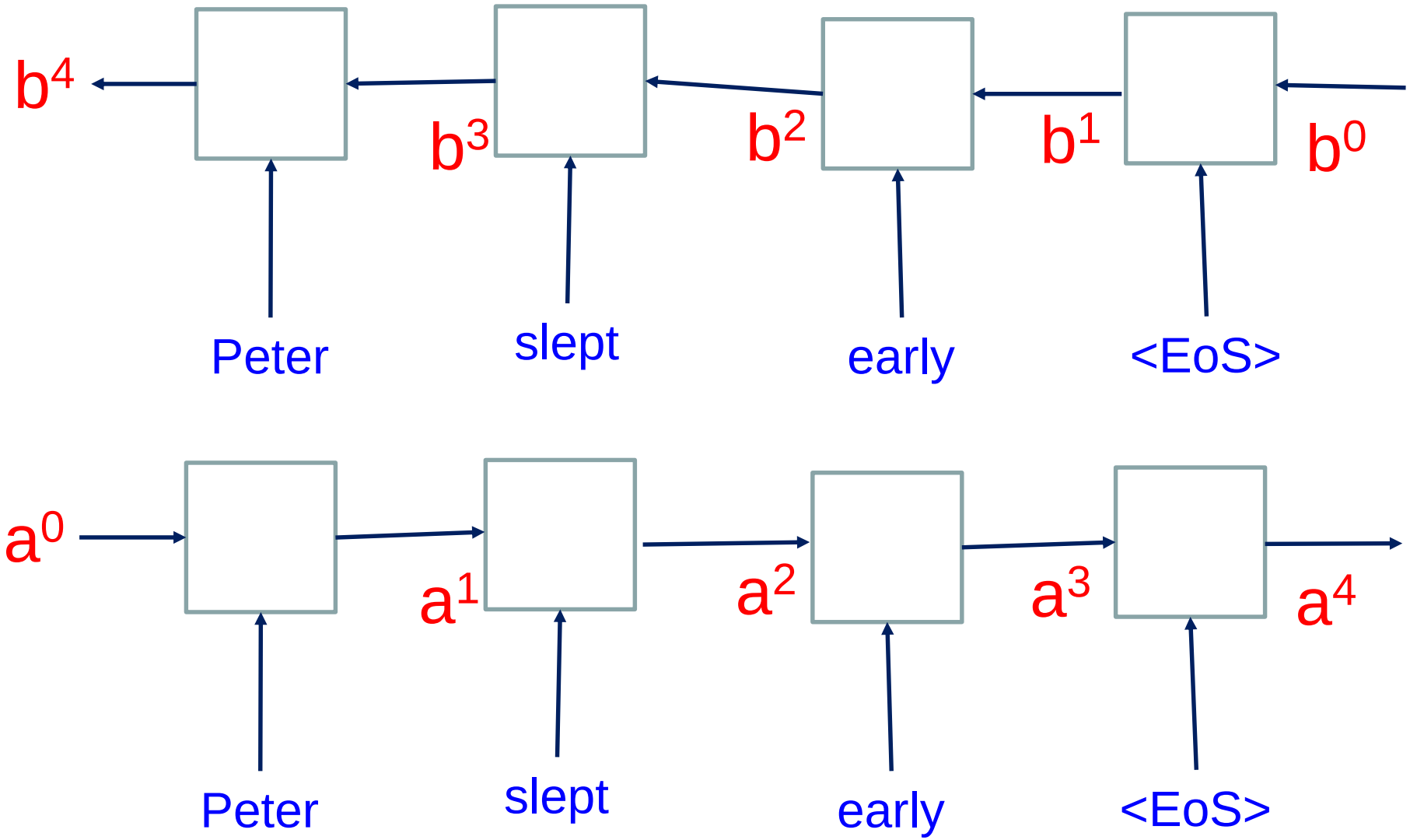
Bank of the river

Attention illustrated through *Peter slept early ←→ piitar jaldii soyaa*

Following the lecture on attention by

Andrew Ng

Illustration of attention computation- **Bi-GRU/LSTM** for input encoding



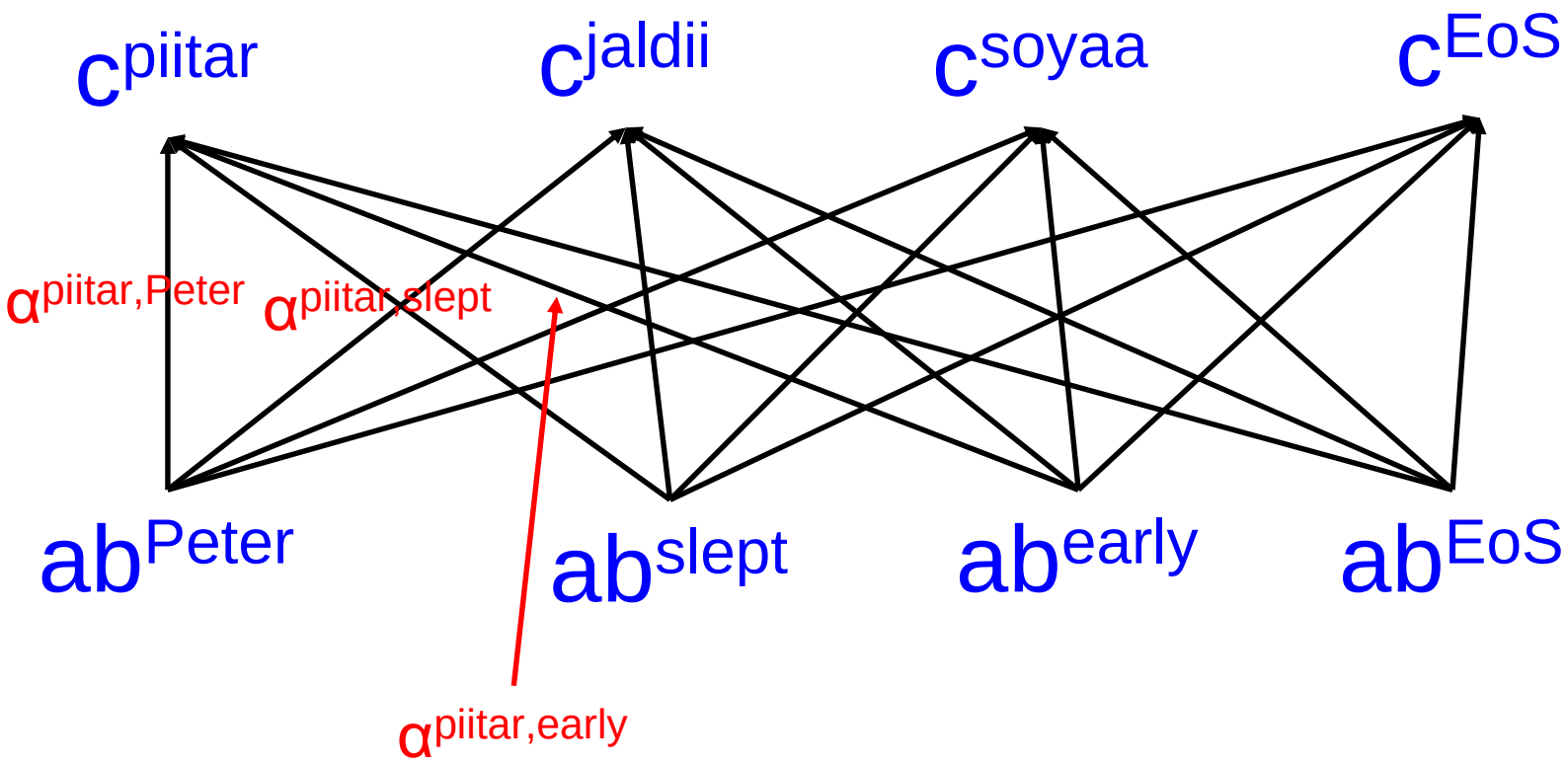
a^is and b^is are activations or states of the bi-directional units

Concatenate the ‘a’s and ‘b’s

- ab^0= concat(a^0, b^4), call this ab^{init}
- ab^1= concat(a^1, b^3), call this ab^{Peter}
- ab^2= concat(a^2, b^2), call this ab^{slept}
- ab^3= concat(a^3, b^1), call this ab^{early}
- ab^4= concat(a^4, b^1), call this ab^{EoS}

ab^{init}, ab^{Peter}, ab^{slept}, ab^{early}, ab^{EoS} are the feature vectors representing the tokens *init*, *Peter*, *slept*, *early* and <*EoS*> (EoS→ End of Sentence)

Generate context vector ‘c’ for each decoded word in the output



Attention weights

$\alpha^{piitar,Peter}$, $\alpha^{piitar,slept}$, $\alpha^{piitar,early}$

$\alpha^{piitar,Peter}+\alpha^{piitar,slept}+\alpha^{piitar,early}=1$

- $\alpha^{soyaa,Peter}$
- $\alpha^{soyaa,slept}$
- $\alpha^{soyaa,early}$

‘α’s sum to 1

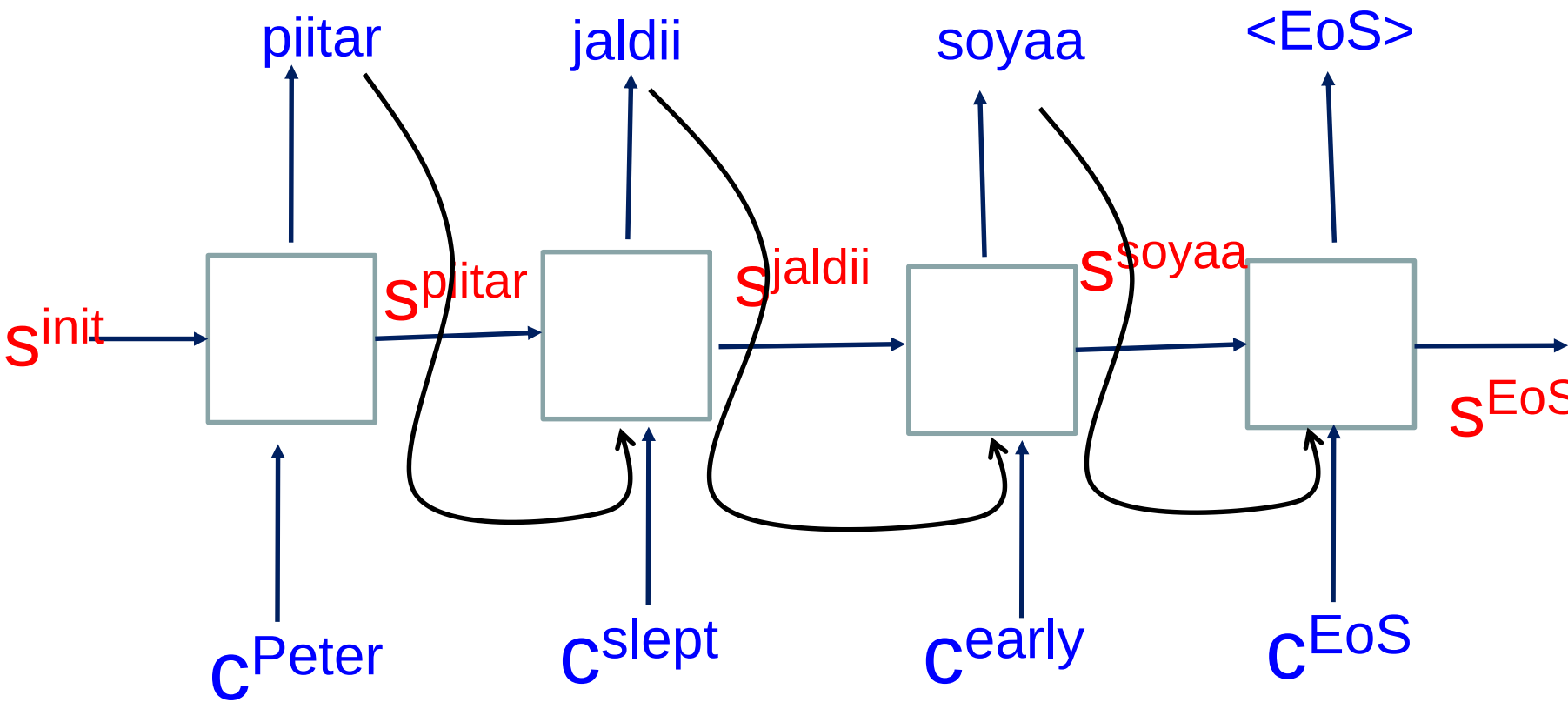
- $\alpha^{jaldii,Peter}$
- $\alpha^{jaldii,slept}$
- $\alpha^{jaldii,early}$
- ‘α’s sum to 1

Attention weight matrix- *Peter slept early ←→ piitar jaldii soyaa*

Hindi (col) --> **English (row)** **|** **\/**	**PIITAR (पीटर)**		**JALDII (जल्दी)**		**SOYAA (सोया)**
PETER	$\alpha^{piitar,Peter}$		$\alpha^{jaldii,Peter}$		$\alpha^{soyaa,Peter}$
SLEPT	$\alpha^{piitar,slept}$		$\alpha^{jaldii,slept}$		$\alpha^{soyaa,slept}$
EARLY	$\alpha^{piitar,early}$		$\alpha^{jaldii,early}$		$\alpha^{soyaa,early}$

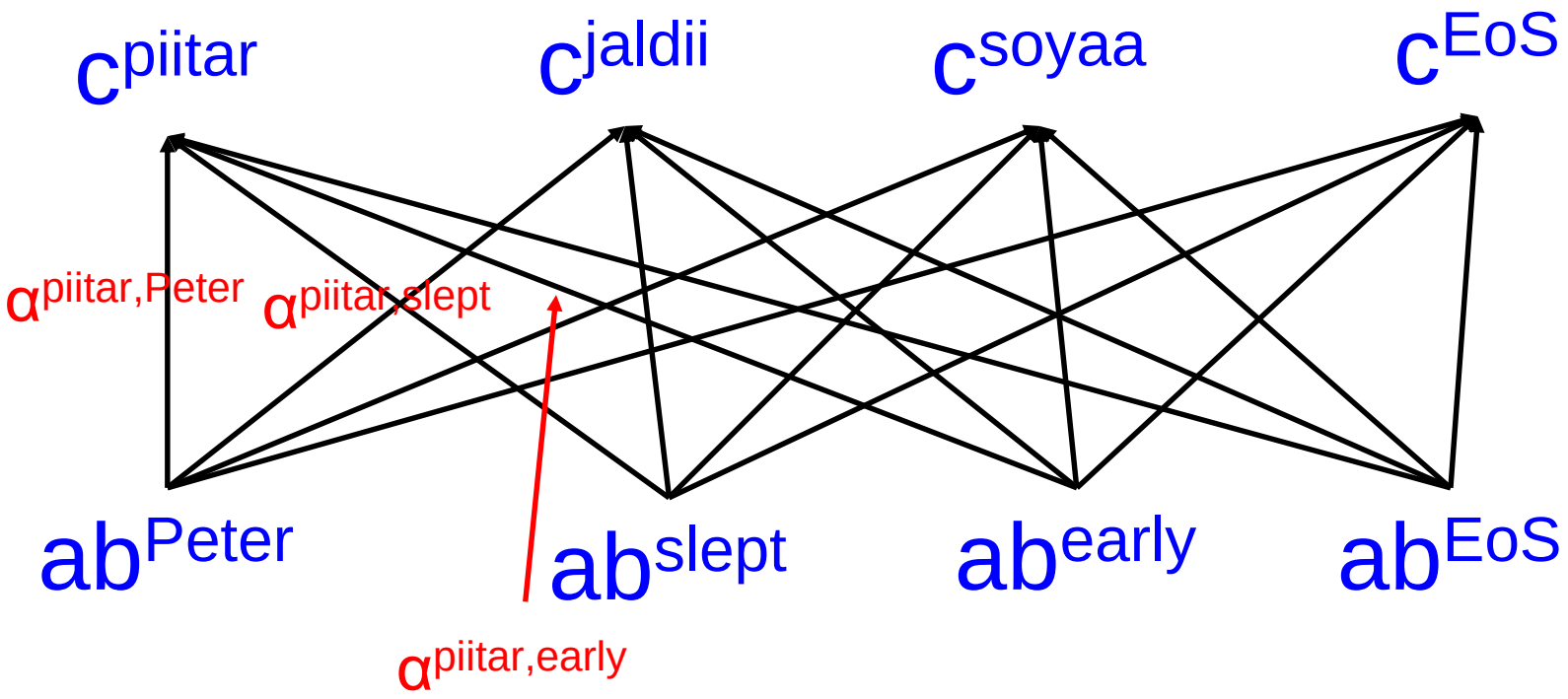
Columns sum to 1

Consider Decoder



s^is are activations or states of the decoder units

Context vectors



$$c^{piitar} = \alpha^{piitar,Peter} \cdot ab^{Peter} + \alpha^{piitar,slept} \cdot ab^{slept} + \alpha^{piitar,early} \cdot ab^{early}$$

Expressions for ‘α’s

$$\alpha^{piitar,Peter} = \frac{e^{f(piitar,Peter)}}{e^{f(piitar,Peter)} + e^{f(piitar,slept)} + e^{f(piitar,early)} + e^{f(piitar,EoS)}}$$

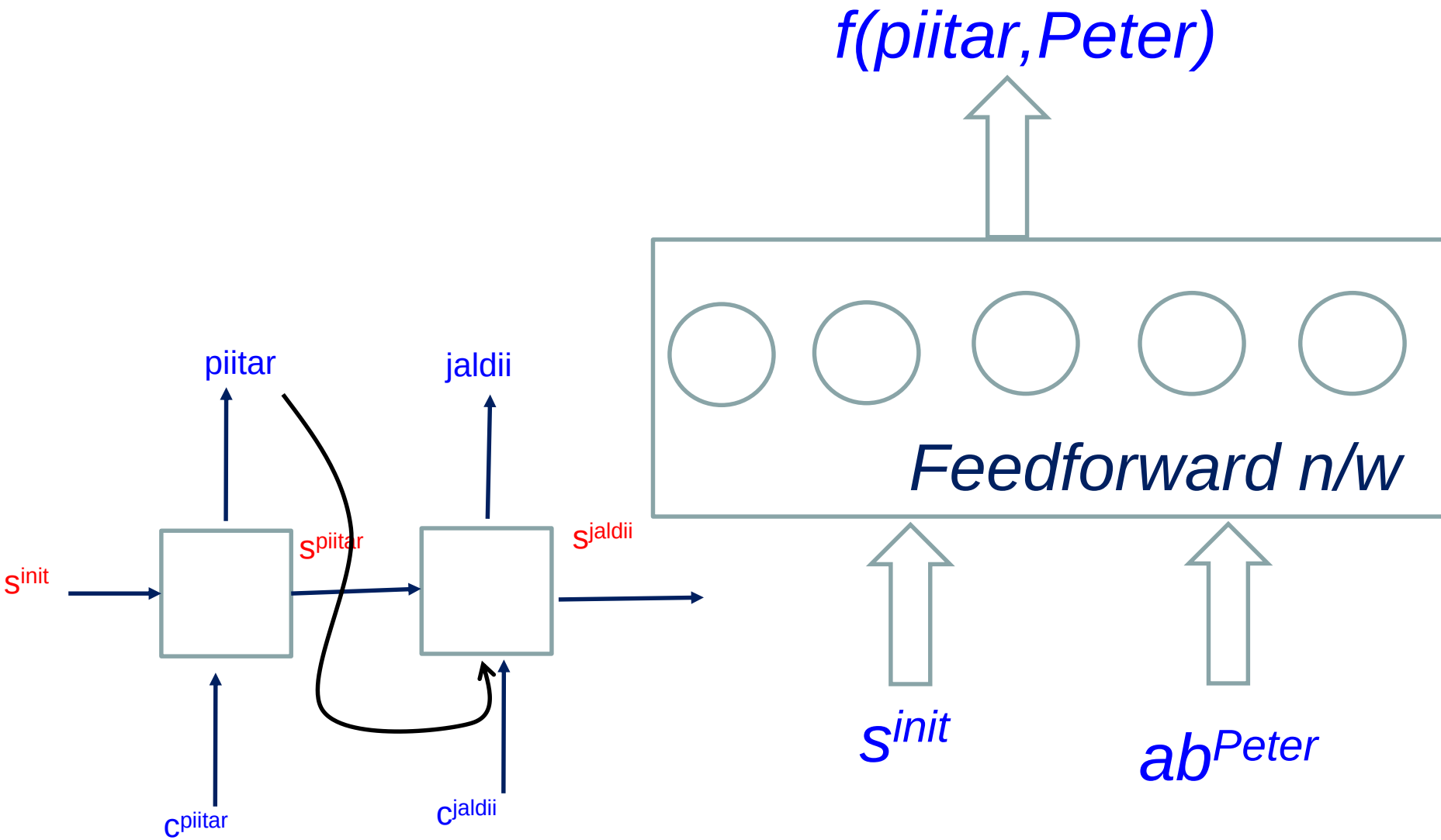
$$\alpha^{piitar,slept} = \frac{e^{f(piitar,slept)}}{e^{f(piitar,Peter)} + e^{f(piitar,slept)} + e^{f(piitar,early)} + e^{f(piitar,EoS)}}$$

$$\alpha^{piitar,early} = \frac{e^{f(piitar,early)}}{e^{f(piitar,Peter)} + e^{f(piitar,slept)} + e^{f(piitar,early)} + e^{f(piitar,EoS)}}$$

$$\alpha^{piitar,EoS} = \frac{e^{f(piitar,EoS)}}{e^{f(piitar,Peter)} + e^{f(piitar,slept)} + e^{f(piitar,early)} + e^{f(piitar,EoS)}}$$

Similarly for other ‘α’s

Function *f(.,.)*


f(piitar,Peter)
Feedforward n/w
piitar
jaldii
s^{init}
s^{piitar}
s^{jaldii}
c^{piitar}
c^{jaldii}
s^{init}
ab^{Peter}

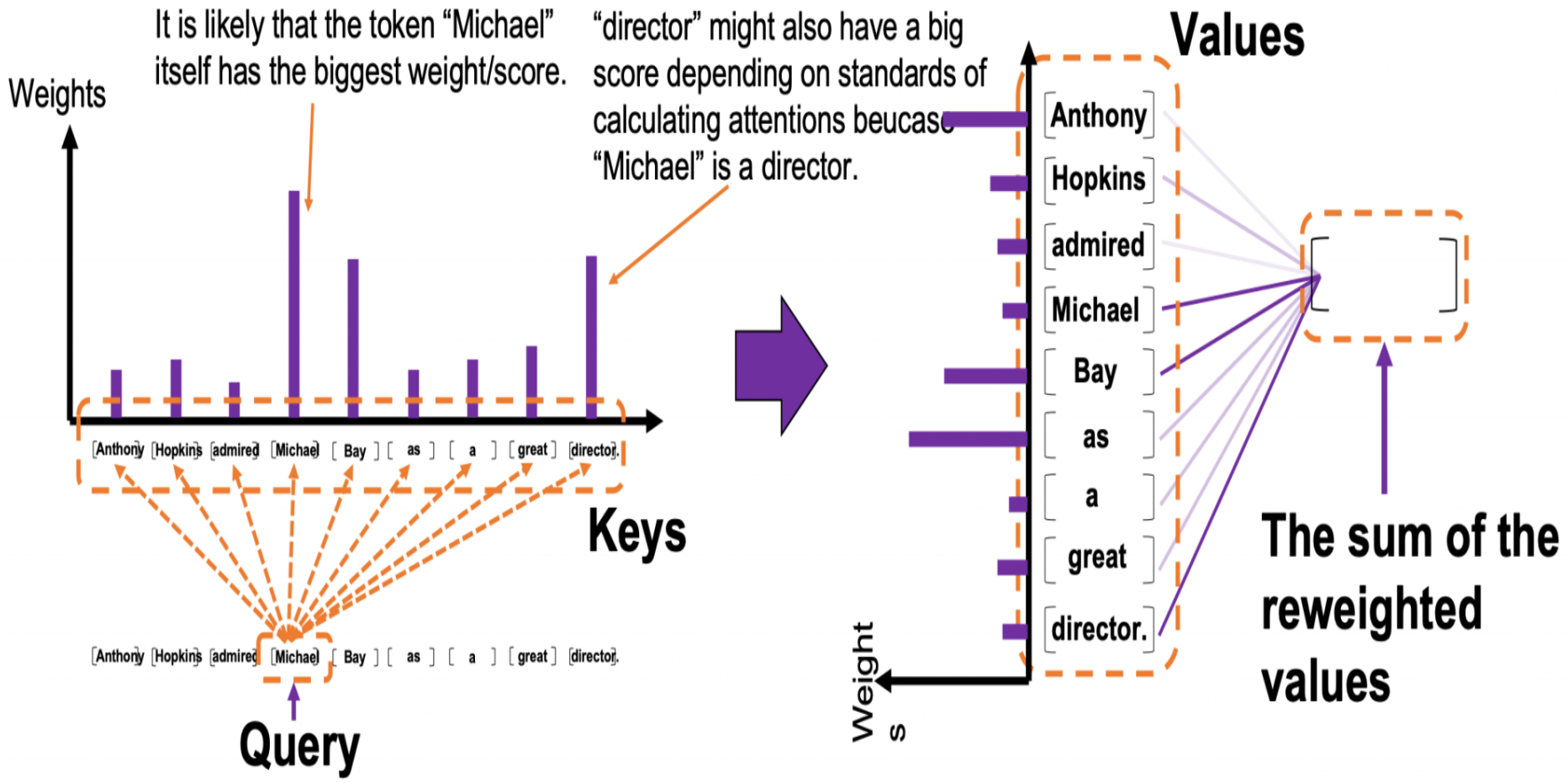
Important observations on self attention

- In the input sequence, pairs of words differ in their strength of association
- For example for an adjective-noun combination, adjective's attention should be stronger for the noun than for other words in the sentence
- So the key questions are:
  - What to attend to
  - With how much attention to attend to

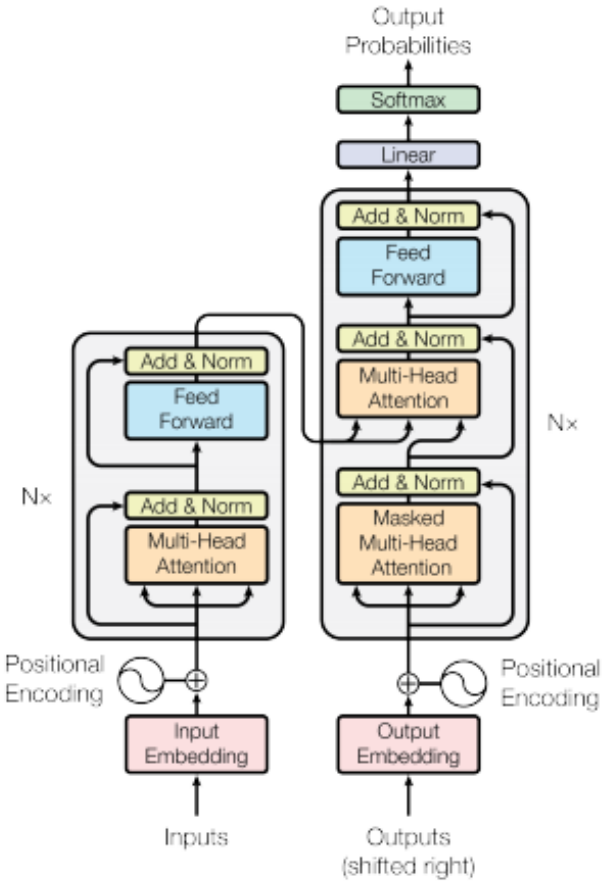
Attention that is non-self

- When the decoder generates the output sequence, attention is a 2-part attention
- Each output token should attend to whatever token has been output before
- Additionally, it should attend to the tokens in the input sequence

Fundamental concepts- “Attention”, “query”, “key”, “value”



Putting it all together



Decoder layer also has a cross-attention layer

Decoder ➔ masking for future time-steps while computing self-attention

There are residual connections & layer-normalization between layers

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. "Attention is all you need." NeurIPS (2017).

http://nlp.seas.harvard.edu/2018/04/03/attention.html
http://jalammar.github.io/illustrated-transformer/

Transformer has led to tremendous advances in MT

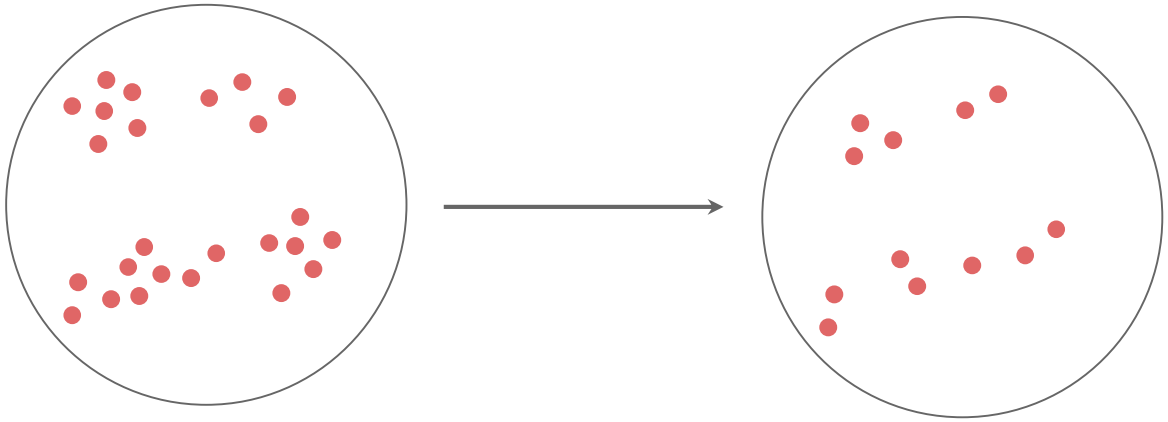
Encoder architectures like BERT based on Transformer have yielded large improvements in NLU tasks

Transformer models are the de-facto standard models for many NLP tasks

Cut to the Chase, Please: Summarizing Textual Data

Summarization Team @ CFILT Lab
Swaroop Nath
Tejpalsingh Siledar
Sri Raghava
Rupasai Rangaraju

What is Summarization?



Summarization → The act of distilling out a representative from the data

Data can be numeric, textual, etc.

Summary stats, such as mean, variance

What is Summarization?

Summarization → The act of distilling out a representative from the data

Data can be numeric, textual, etc.

OUR FOCUS!!

Summarizing Textual Data

The **team's social handles** have been teasing a **titular change** for a while now, with the 'Bangalore' in the name expected to make way for the more **vernacular and official 'Bengaluru'**. They are now also a **champion franchise**. But while the **WPL coronation** is a big win for **brand and fandom alike**, it is likely to have **little bearing on IPL 2024**, save perhaps for the **vibes**.

Changes have been afoot on the **men's side** since the end of last season, when a **three-year run of playoffs appearances** was broken. **Andy Flower** arrived with his **sparkling CV as the head coach**. **Mo Bobat** has taken over as **Director of Cricket**. These changes . . . IPL so as to give the **new leadership group enough time before the purported mega auction to analyse first-hand and chart the roadmap for the team's future**. . . .

The passage discusses **anticipated changes in the Bangalore cricket team, including a potential name change to "Bengaluru"**. Despite recent WPL success, it's unlikely to impact IPL significantly. New leadership aims to analyze team dynamics before a mega auction. Challenges include adapting to a new bowling lineup and maximizing home advantage.

Summarizing Textual Data

The **team's social handles** have been teasing a **titular change** for a while now, with the 'Bangalore' in the name expected to make way for the more **vernacular and official 'Bengaluru'**. They are now also a **champion franchise**. But while the **WPL coronation** is a big win for **brand and fandom alike**, it is likely to have **littl** **24**, save perhaps for

238 words

Changes have been afoot on the **men's side** since the end of last season, when a **three-year run of playoffs appearances** was broken. **Andy Flower** arrived with his **sparkling CV as the head coach**. **Mo Bobat** has taken over as **Director of Cricket**. These changes . . . IPL so as to give the **new leadership group enough time before the purported mega auction to analyse first-hand and chart the roadmap for the team's future**. . . .

The passage discusses **anticipated changes in the Bangalore cricket team, including a potential name change to "Bengaluru"**. Despite s, it's unlikely to impact IPL significantly. New nalyze team dynamics before a mega auction adapting to a new bowling lineup and maximizing home advantage.

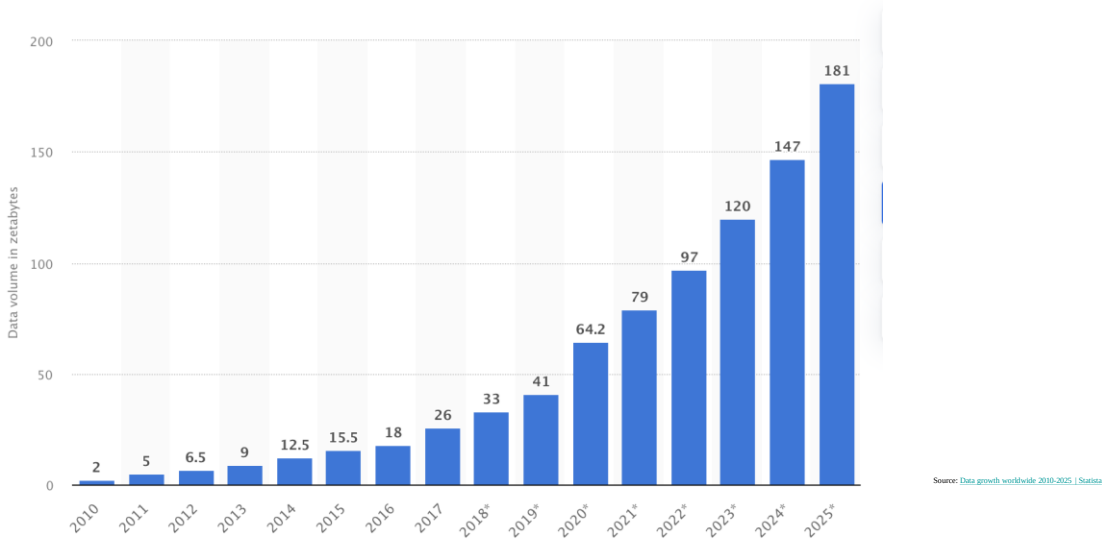
34 words

Why would you need Summarization?

In 2020, the amount of digital data was 64.2 zettabytes

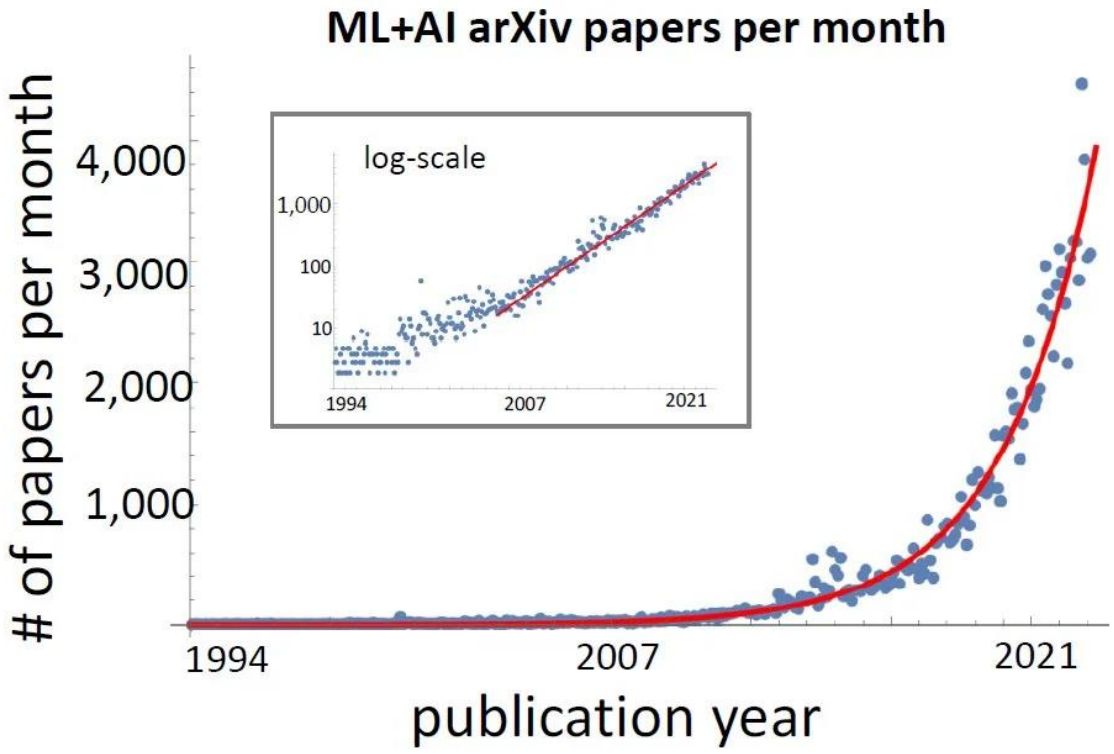
In 2025, the expected amount of digital data is 180 zettabytes

That's 64.2 **BILLION** Terabytes



Source: Data growth worldwide 2010-2025 | Statista

Why would you need Summarization?

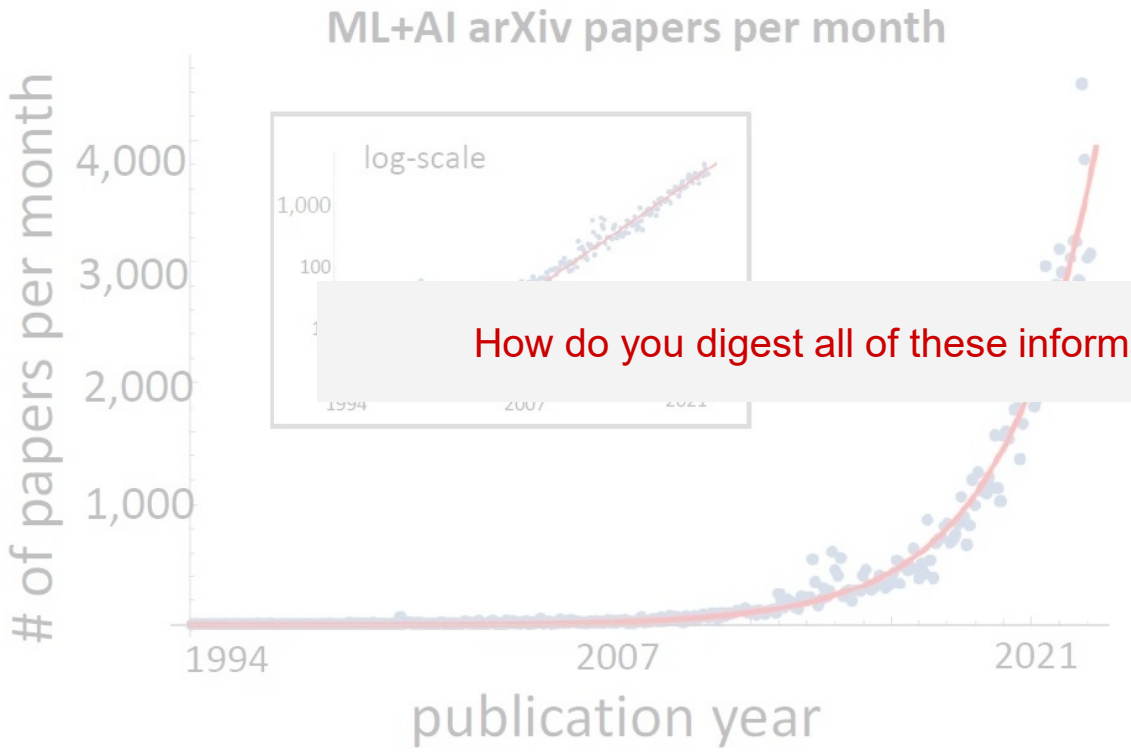


Source: "The number of AI papers on arXiv per month grows exponentially with doubling rate of 24 months." : r/singularity

Computational Linguistics (cs.CL) has **1021** submissions already in March, 2024

Source: Computation and Language authors/titles Mar 2024

Why would you need Summarization?



How do you digest all of these information?

Computational
stics (cs.CL) has
ubmissions
already in March, 2024

Source: "The number of AI papers on arXiv per month grows exponentially with doubling rate of 24 months." : r/singularity

Source: Computation and Language authors/titles Mar 2024

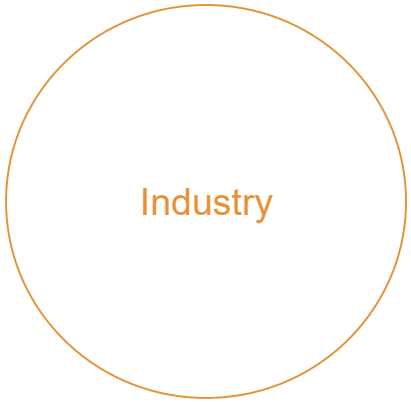
Benefits of Summarization

Saves your time!

Removes redundancy in data → Storage Efficiency!

Use Cases of Summarization

Hey ChatGPT, give me a gist of this PPT?



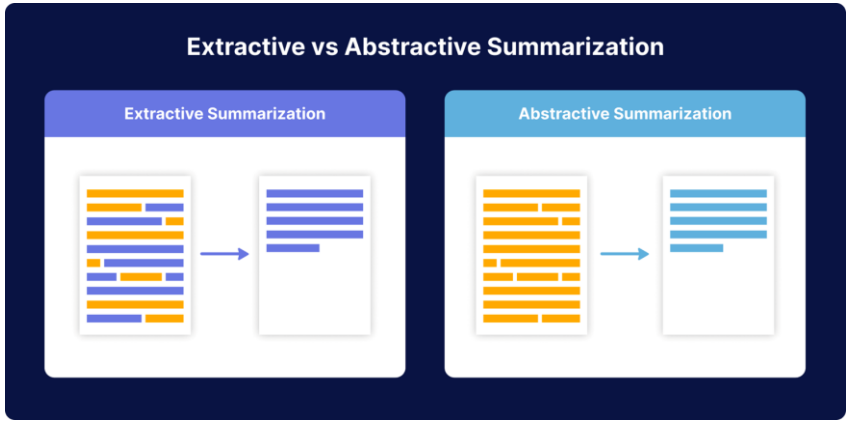
Onboarding Employees

Summarizing Design Docs

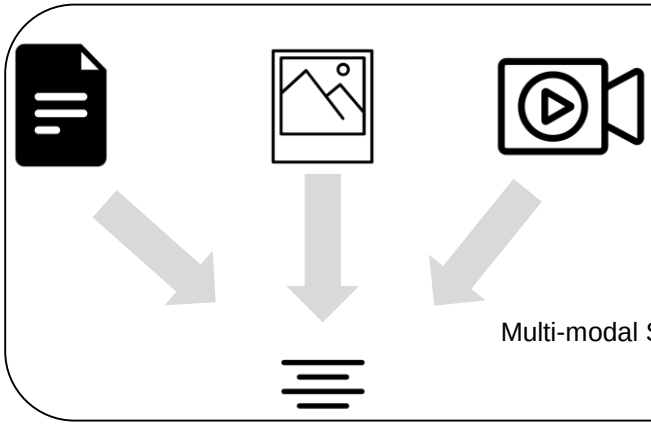
Catchy Marketing Posts

Types of Summarization

Single Document Summarization



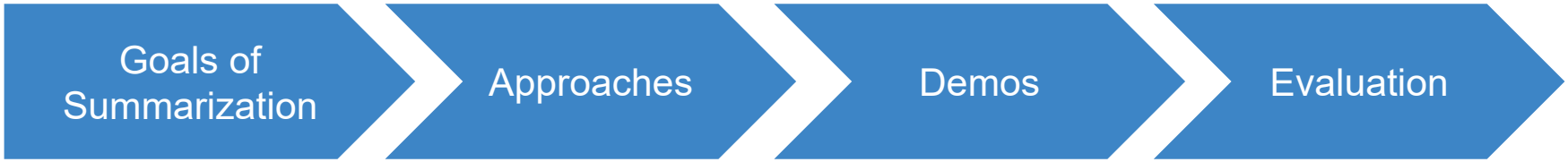
Multi Document Summarization



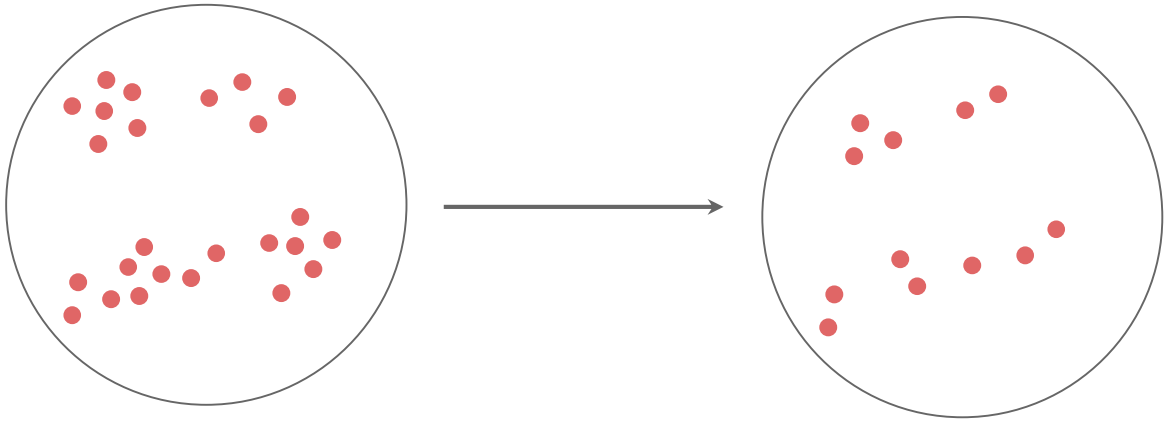
Roadmap | rough

- Details on Summarization
  - What qualities do you need in your summarizer? Coverage and Conciseness!
  - How do you plan to achieve them?
    - Statistical
    - Deep Learning
    - PLMs
  - Demos
  - Evaluation
    - Lexical metrics
    - Semantic metrics
    - Human Evaluation

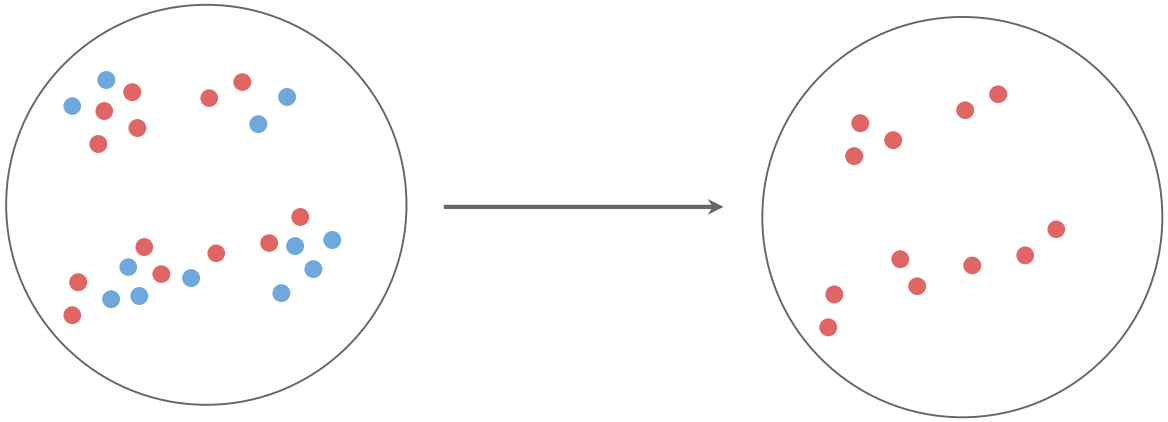
Roadmap



Goals of Summarization



Goals of Summarization



The way that these points are distributed is maintained

Coverage

Nearby points of chosen points were dropped

Concisenes
s

Coverage and Conciseness

Coverage

You have to include **all** important details from the document

Concisenes
s

You have to include **only** important details from the document

Extractive Summarization | Statistical Approaches

Key Idea →

1. Formulate a scoring scheme for sentences.
2. Rank them and select top-*k*.

Scoring Schemes

1. Term Frequency based Approaches

Adding frequency of all words in a sentence

$$\text{score} = \frac{\Sigma_w f(w)}{L}$$

Length of the sentence

Scoring Schemes

1. Term Frequency based Approaches
2. TF-IDF based Approaches
3. Feature based Approaches

```
columns = ['tf-score', 'length-score', 'position-score', 'paragraph-score', 'cue-words-score', 'paragraph_id']
```

Term-Frequency

Are there interesting words in the sentence?

Scoring Schemes

1. Term Frequency based Approaches
2. TF-IDF based Approaches
3. Feature based Approaches
4. Topic-Modelling based Approaches

How many topic words does this sentence have?

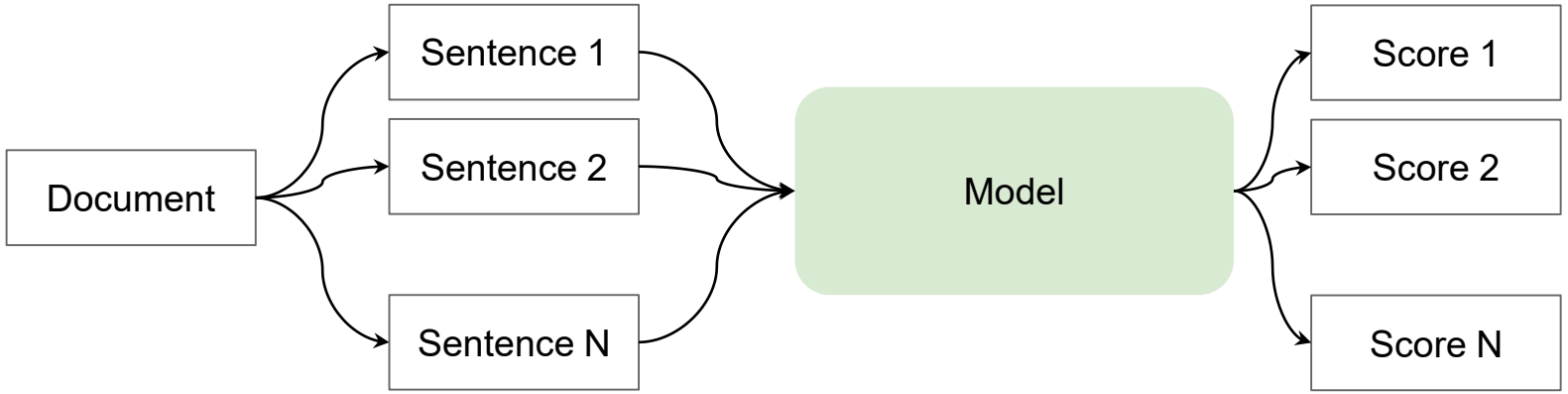
github.com/swaroop/nlp-ext-summ

Extractive Summarization | Era of Deep Learning

Key Idea is similar

Scoring is now done using DL models

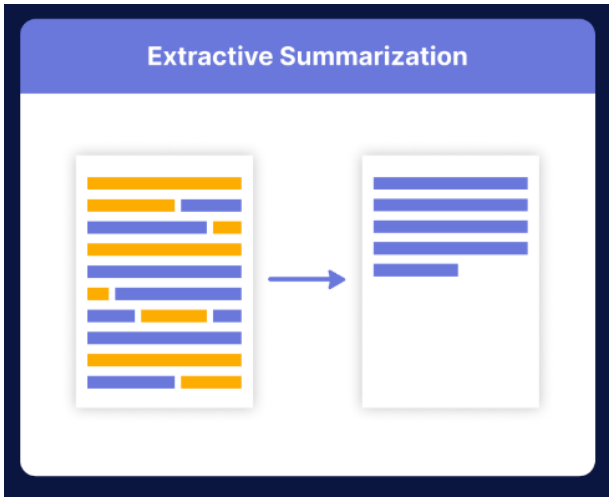
1. Formulate a scoring scheme for sentences.
2. Rank them and select top-*k*.



SummaRunner | Nallapati et al., 2017

Leveraging BERT for Extractive Summarization, Derek Miller, 2019

Extractive Summarization | Challenges



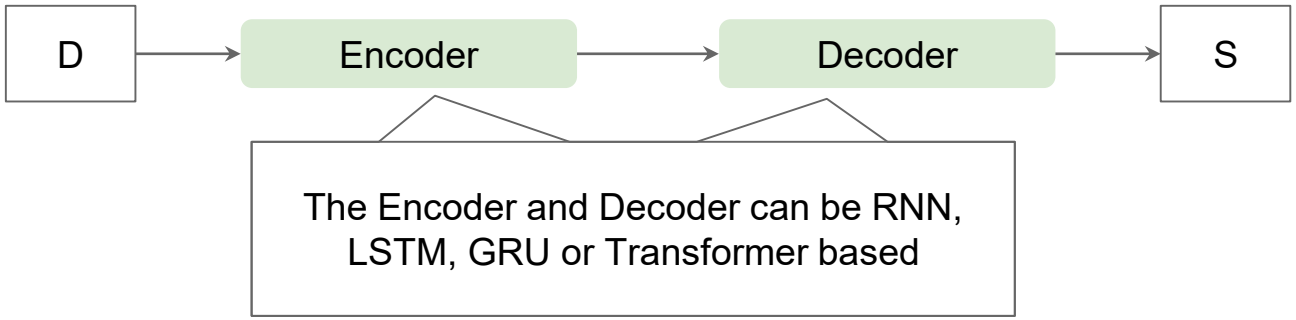
Coherence → It is not necessary for high ranked sentences to be contiguous in the document.

Sentences are stitched together → may not read **fluently**

Abstractive Summarization | Underlying Math

$$\mathcal{P}(S \mid D) = \mathcal{P}(s_n \mid s_1, s_2, \ldots, s_{n-1}, D)$$

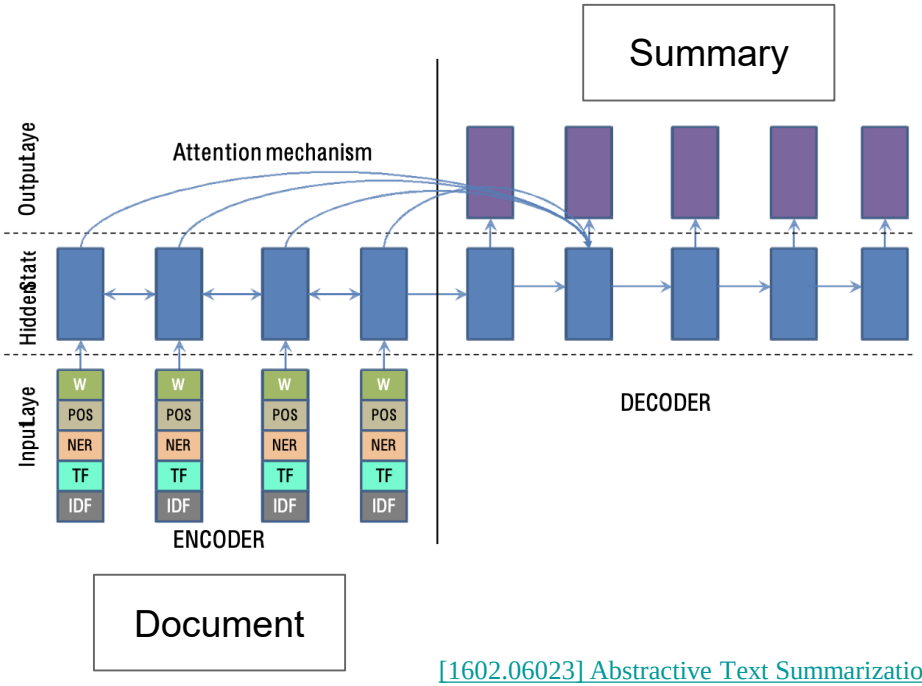
Formulation is the same Sequence-to-Sequence[1] formulation.



1. [1409.3215] Sequence to Sequence Learning with Neural Networks, Sutskever et al., 2014

Abstractive Summarization | RNNs

Encode-Attend-Decode Framework (*Quite popular during 2014-17*)



Word embedding, along with position, and other things are fed to the RNN-based encoder

An RNN-based decoder generates one token at a time by attending to all the encoded document tokens

[1602.06023] Abstractive Text Summarization Using Sequence-to-Sequence RNNs and Beyond, Nallapati et al., 2016

A Neural Attention Model for Abstractive Sentence Summarization - ACL Anthology, Rush et al., 2015

Abstractive Summarization | Era of Fine-Tuning

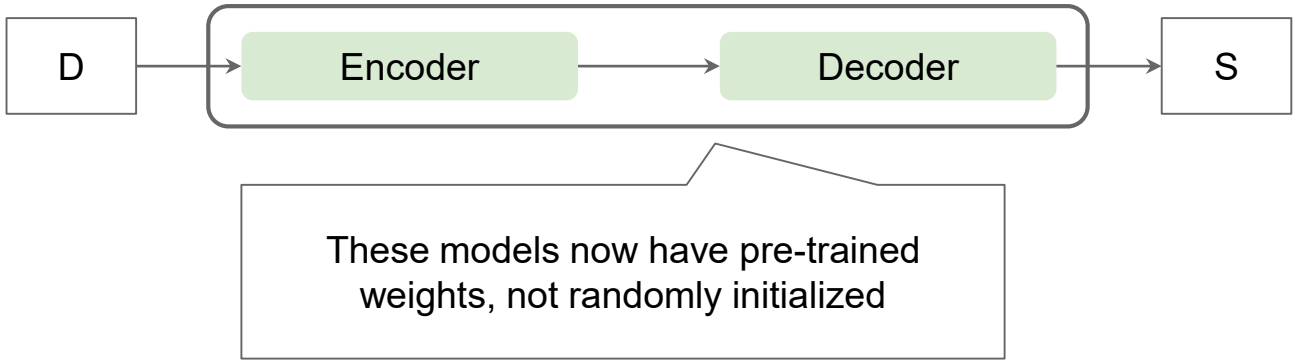
Transformers allowed for large-scale pre-training, with Language Modelling objective(s).

Pre-training allowed for the infusion of language knowledge prior to task knowledge.

Now the paradigm was → **pre-train** (to learn the language), then **fine-tune** (to learn the task)

Abstractive Summarization | Era of Fine-Tuning

Now the paradigm was → **pre-train** (to learn the language), then **fine-tune** (to learn the task)



Examples include BART, GPT-*, based abstractive summarization

[2006.01997] Automatic Text Summarization of COVID-19 Medical Research Articles using BERT and GPT-2, Kieuvongngam et al., 2020

[2006.09595] CO-Search: COVID-19 Information Retrieval with Semantic Search, Question Answering, and Abstractive Summarization, Esteva et al., 2020

Abstractive Summarization | Challenges

1. Pre-trained weights might have too **strong priors**, which might not be suitable to Summarization
2. Document might have **domain-specific rare words**, which might be difficult for the model to generate.
3. The training objective does not reflect the **goals of Summarization** (conciseness and coverage)

Abstractive Summarization | Summarization influenced Pre-Training

Pre-training is mostly task-agnostic. However, there are instances when downstream task influenced pre-training designs.

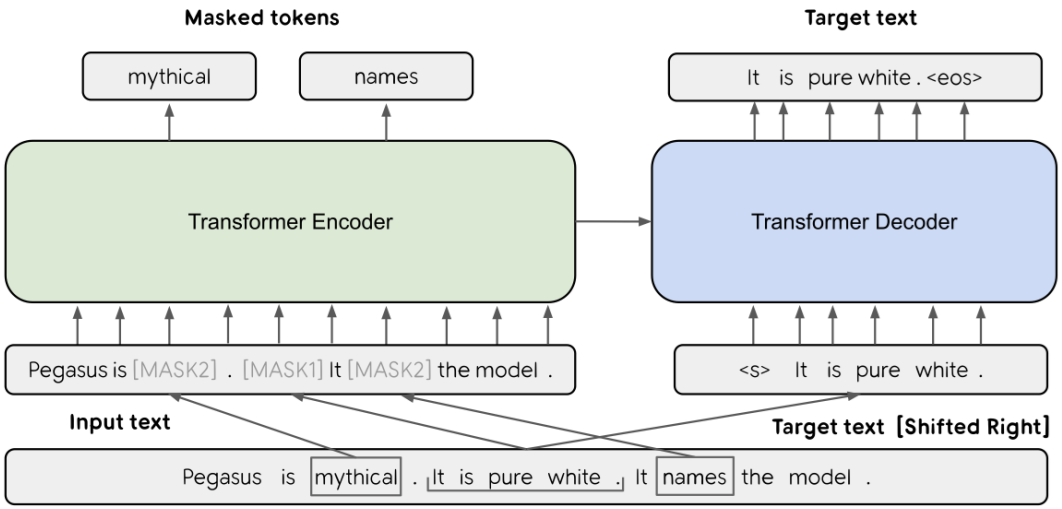
PEGASUS[1] is an instance → pre-training design influenced by abstractive summarization.

PEGASUS proposes Gap Sentence Generation (GSG) as an additional pre-training objective.

Motivation → predicting gap sentences helps the model better understand the document, it is a generalization over Masked Word Modelling (MLM)

1. [1912.08777] PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization, Zhang et al., 2019

Abstractive Summarization | PEGASUS Pre-Training



Encoder is trained to predict the masked tokens

Decoder is trained to predict the masked sentence

Abstractive Summarization | Pointer Generator Net

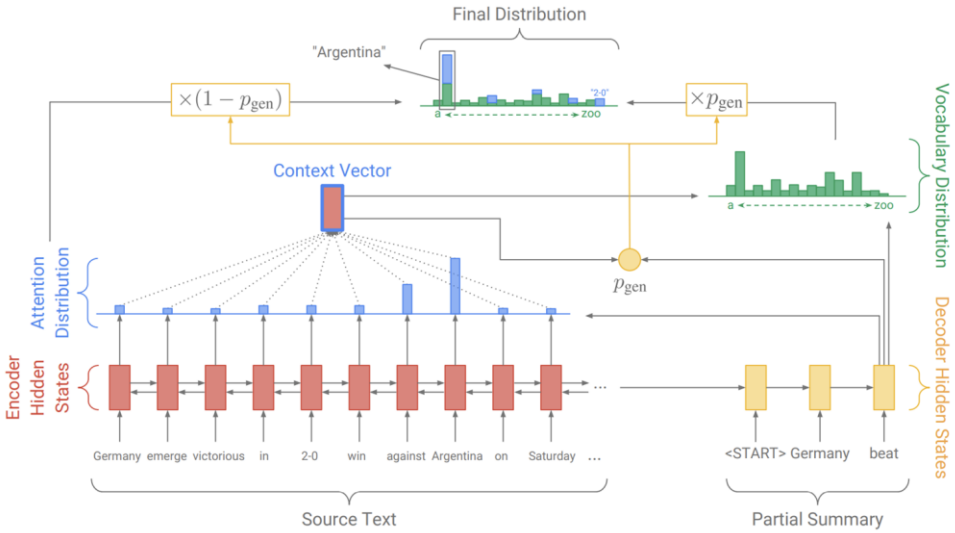
Pointer Generator Networks[1] were proposed to deal with UNK generations

It was difficult for models to generate domain-specific important words, these are rare → so statistics is biased to not favour them

Pointer Generator Networks use a gating mechanism to decide on whether to generate a token or copy one from input text.

1. [1704.04368] Get To The Point: Summarization with Pointer-Generator Networks, See et al., 2017

Abstractive Summarization | Pointer Generator Net

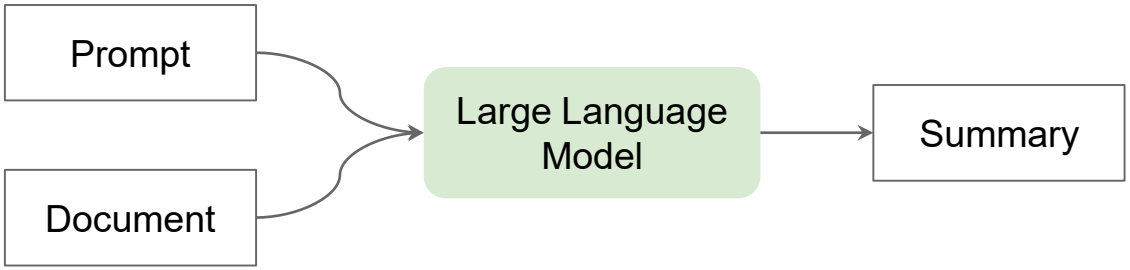


The final distribution is a superposition of attention distribution and vocabulary distribution

Summarization | Recent Trends

Prompt-based Inference is the latest trend in most NLP tasks.

Refining prompts to get better outputs is a core problem to solve.



[2309.04269] From Sparse to Dense: GPT-4 Summarization with Chain of Density Prompting, Adams et al., 2023

Prompt-based Summarization DEMOs

Evaluating Summaries

ROUGE: Measures lexical overlap between generated and reference summary

It measures how much of the reference summary is *recalled* by the generated summary

ROUGE-WE: Measures semantic overlap between generated and reference summary

It extends ROUGE by using word2vec to compute *cosine similarity*, instead of token-by-token equality.

BERTScore: Measures semantic overlap between generated and reference summary

It uses *BERT token embeddings to align tokens* in reference and generated summaries, and then compute the overall similarity.

[2007.12626] SummEval: Re-evaluating Summarization Evaluation, Fabbri et al., 2020

Reinforcement Learning & Query-focused Summarization

Query focused Summarization | What is it?

What is string pool in Java?

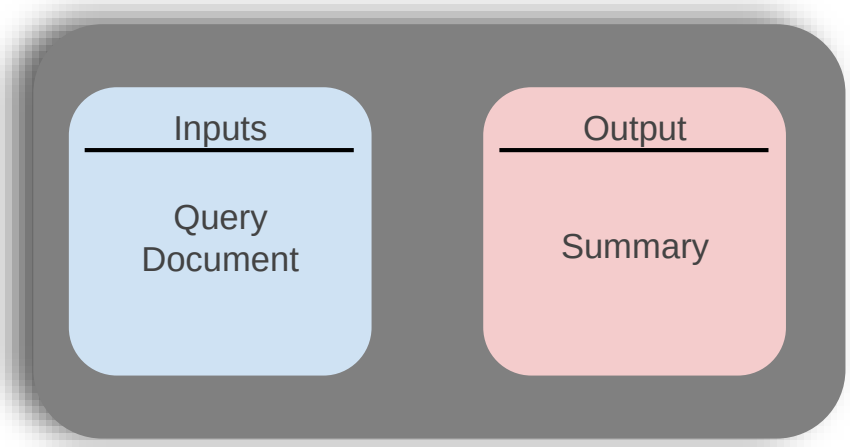
Query

Document

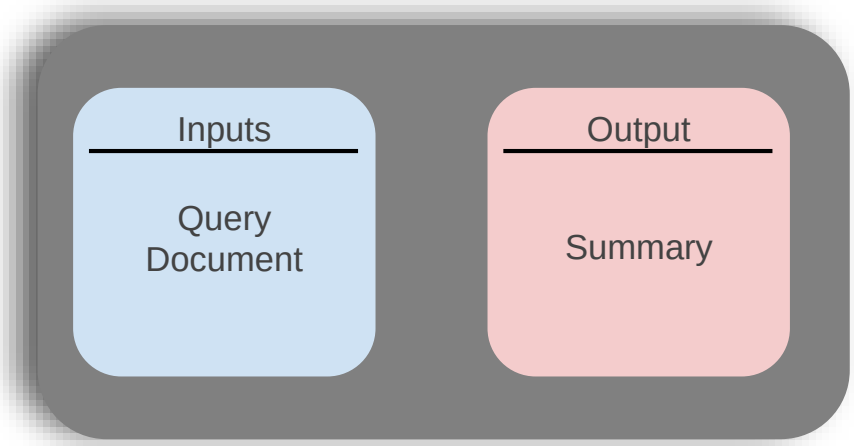
String pool is nothing but a storage area in Java heap where string literals are stored. It is also known as String Intern Pool or String Constant Pool. It is just like object allocation. By default, it is empty and privately maintained by the Java · · · The JVM performs some steps during the initialization of string literals that increase the performance and decrease the memory load . . .

Summary

String Pool, used to reduce the memory footprint, is a specific area in the memory allocated to the process used to store String literal declared within Java program.



Query focused Summarization | Approach?



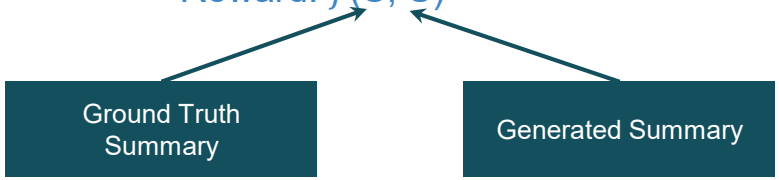
Employ Reinforcement Learning for Training

Model each token generation as an action taken by the agent

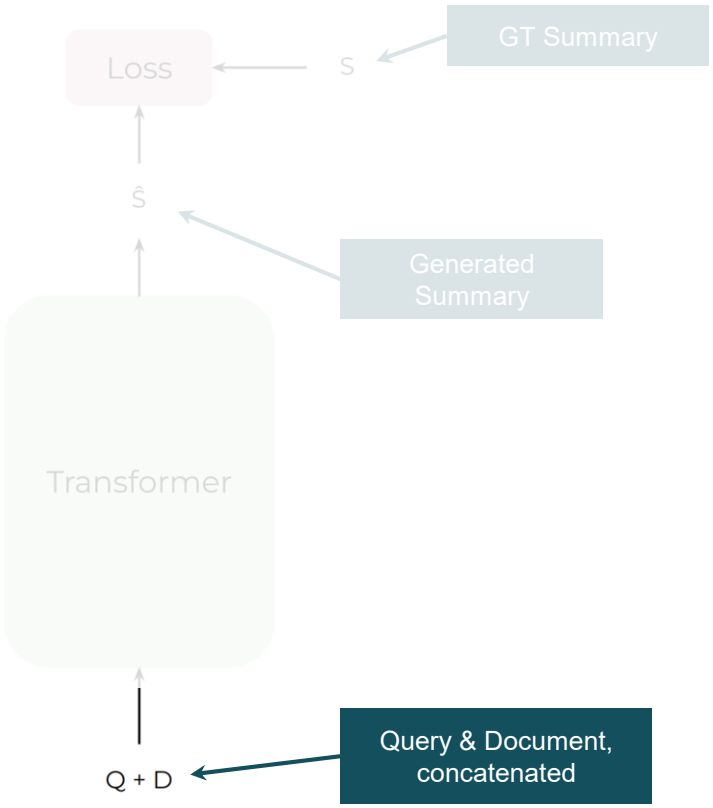
State: Query, Document, Partial Summary

Action: Token

Reward: $f(S, \hat{S})$



Why Reinforcement Learning? | Key Insight



Token-by-Token match

$$\mathcal{L}_{\mathcal{MLE}} = -\sum_{t=1}^{n} \Big\{ \mathbb{1}[y_t = y_t^*] \log \mathcal{P}(y_t^* | y_1^*, y_2^*, \cdots, y_{t-1}^*, \mathbf{q}, \mathbf{d}) \Big\}$$

$$\mathcal{L}_{\mathcal{PG}} = -\sum_{t=1}^{n} \Big\{ (R(\tau) - b) \log \mathcal{P}(y_t^* | y_1^*, y_2^*, \cdots, y_{t-1}^*, \mathbf{q}, \mathbf{d}) \Big\}$$

Any Generic Reward

Rewards for RL Training

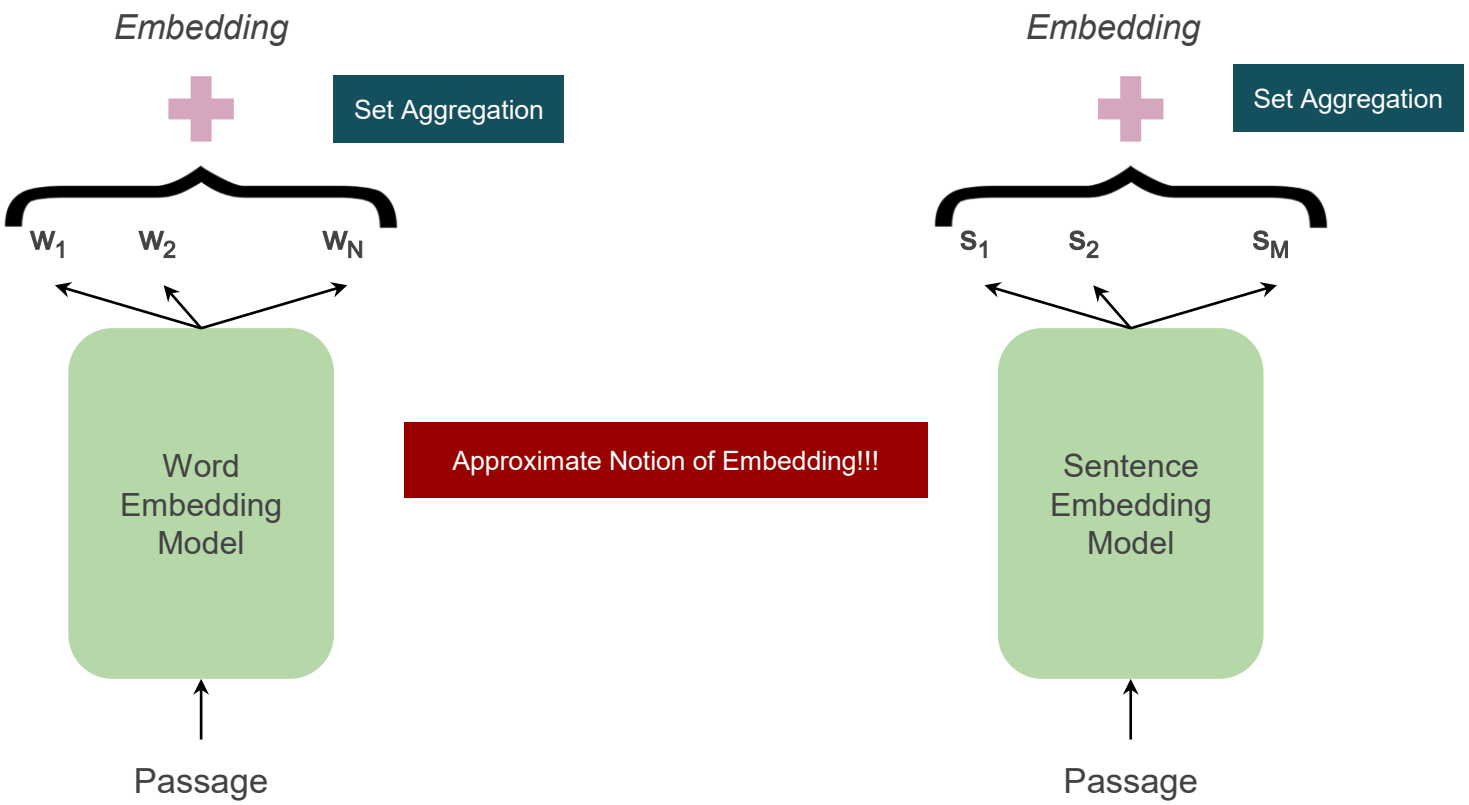
Reward	Description	Formula
ROUGE-L	Recall oriented reward to improve coverage	$\mathbb{ROUGE}_{\mathbb{L}}(GT, GN)$
BLEU	Precision oriented reward to generate concise summaries	$mean_{i=1}^{4}(\mathbb{BLEU}_i(GT, GN))$
SimCSE (Gao et al., 2021)	Semantic match obtained by averaging sentence embeddings	$cos(mean_m(\mathbb{E}_{\mathbb{S}}), mean_n(\mathbb{E}_{\mathbb{T}}))$
SBERT (Reimers and Gurevych, 2019)	Semantic match obtained by averaging sentence embeddings	$cos(mean_m(\mathbb{E}_{\mathbb{S}}), mean_n(\mathbb{E}_{\mathbb{T}}))$
SFPEG	Semantic match obtained using Passage Embedding	$cos(\mathbb{E}_{GT}, \mathbb{E}_{GN})$

Lexical Rewards

Semantic Rewards

Based on a novel Passage Embedding scheme we developed

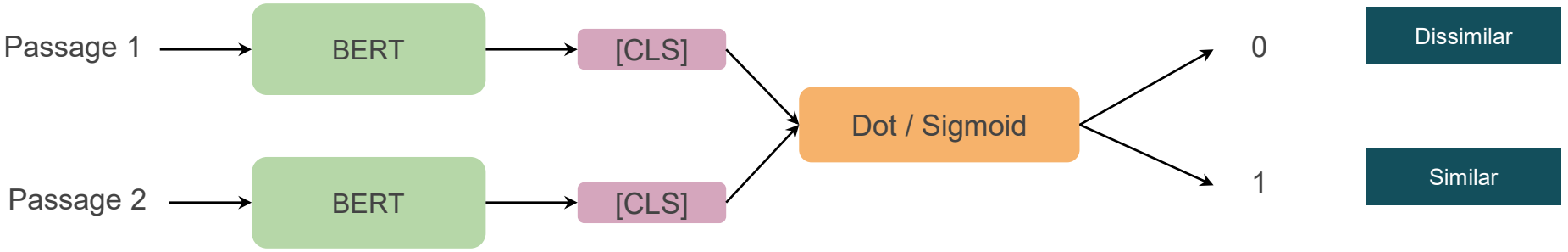
Passage Embedding | Contemporary Works



Passage Embedding | Cluster Hypothesis

Passages that answer to similar information needs tend to be clustered together

Cluster Hypothesis[1]



1. N. Jardine and C.J. van Rijsbergen. 1971. The use of hierarchic clustering in information retrieval. *Information Storage and Retrieval*, 7(5):217–240.

Contributions

Novel RL algorithm for QfS

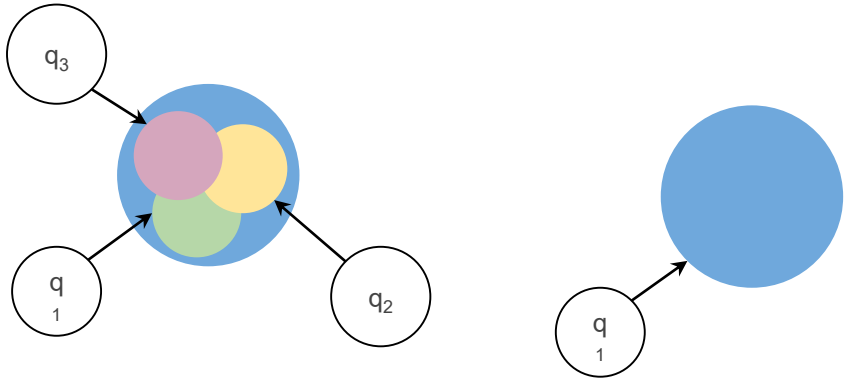
Tackles conflict of Teacher Forcing with application of RL (*see paper for discussion*)

Beats SOTA both in automatic and human evaluation

Human Curated test set

Tackles topic centralization

250 instances



Contributions

Novel RL algorithm for QfS

Tackles conflict of Teacher Forcing with application of RL (*see paper for discussion*)

Beats SOTA both in automatic and human evaluation

Human Curated test set

Tackles topic centralization

250 instances

Cluster hypothesis based novel Passage Embedding

2.21 ROUGE points improvement over competing rewards

Dataset to train Passage Embedding model

~8 million instances scraped from Reddit

Experiments | Automatic Evaluation

EXPT I

QfS on ELI5[1] test set

EXPT II

QfS on our test set

EXPT III

QfS on our test set, with random documents

Tests if the model truly uses the document, given the query

1. Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. Eli5: Long form question answering. In *Proceedings of ACL 2019*.

Experiments | Automatic Evaluation

EXPT I

QfS on ELI5[1] test set

EXPT II

QfS on our test set

EXPT III

QfS on our test set, with random documents

Replication of Lewis et al., 2020[2]

Our models

Model ID	Reward
BART SL	-
BART R	ROUGE-L
BART R-SEM	ROUGE-L + SimCSE
BART R-B	ROUGE-L + BLEU
BART R-SBERT	ROUGE-L + SBERT
BART R-SFPEG	ROUGE-L + SFPEG
BART SFPEG	SFPEG

1. Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. Eli5: Long form question answering. In *Proceedings of ACL 2019*.

2. Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. 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*, pages 7871–7880, Online. Association for Computational Linguistics.

Experiments | Human Evaluation

EXPT I

QfS on ELI5[1] test set

EXPT II

QfS on our test set

EXPT III

QfS on our test set, with random documents

Model	ELI5 dataset			Our dataset (EXPT-II)			Our dataset (EXPT-III)		
	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
Fan et al. (2019)	28.9	5.4	23.10	28.82	6.97	25.41	23.13	4.39	20.91
Lewis et al. (2020)	30.6	6.2	24.3	-	-	-	-	-	-
BART SL (b)	29.68	5.89	25.44	29.67	7.88	26.40	23.32	4.51	20.96
BART R	38.93	8.05	34.54	43.08	15.30	39.08	24.38	4.40	22.17
BART R-SEM	38.02	6.56	33.13	44.12	15.15	40.39	25.39	4.51	23.06
BART R-B	**39.52**	**8.25**	**34.92**	43.46	15.67	39.29	26.01	4.88	23.45
BART R-SBERT	36.9	6.36	32.8	42.93	15.10	39.56	25.41	4.47	23.19
BART R-SFPEG	39.40	6.92	34.10	**45.52**	**16.83**	**41.29**	25.89	4.99	23.50
BART SFPEG	34.76	5.96	29.66	44.30	15.79	40.39	25.82	4.63	23.61

Less difference → less focus on document

More difference → actually focuses on document

1. Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. Eli5: Long form question answering. In *Proceedings of ACL 2019*.

Experiments | Human Evaluation

Fluency

Likert Scale
0 (Very Bad) to 4 (Very Good)

Correctness

YES (1) or NO (0)

Model	Fluency	Correctness
BART SL	3.08 / 3.12	0.26 / 0.28
BART R	3.34 / 3.30	0.38 / 0.32
BART R-SEM	3.26 / 3.23	0.24 / 0.18
BART R-B	3.24 / 3.26	0.28 / 0.26
BART R-SEM	3.28 / 3.27	0.22 / 0.26
BART R-SFPEG	**3.46 / 3.48**	**0.48 / 0.44**
BART SFPEG	3.06 / 3.04	0.28 / 0.26
Human	4.0 / 3.98	1.0 / 1.0

Conclusion

RL leads to better Query-focused Summaries, both in automatic and human evaluation

RL helps learn QfS better, supported by huge difference in Expt `II` and Expt `III` scores (*see paper for discussion*)

Cluster Hypothesis → better Passage Embeddings, supported by better downstream (QfS) results

Qualitatively too, the summaries are much better (*see paper for extensive qualitative analyses*)

Conclusion

RL leads to better Query-focused Summaries, both in automatic and human evaluation

Qualitatively too, the summaries are much better (*see paper for extensive qualitative analyses*)

RL helps learn QfS better, supported by huge difference in Expt II and Expt III scores (*see paper for discussion*)

Cluster Hypothesis → better Passage Embeddings, supported by better downstream (QfS) results

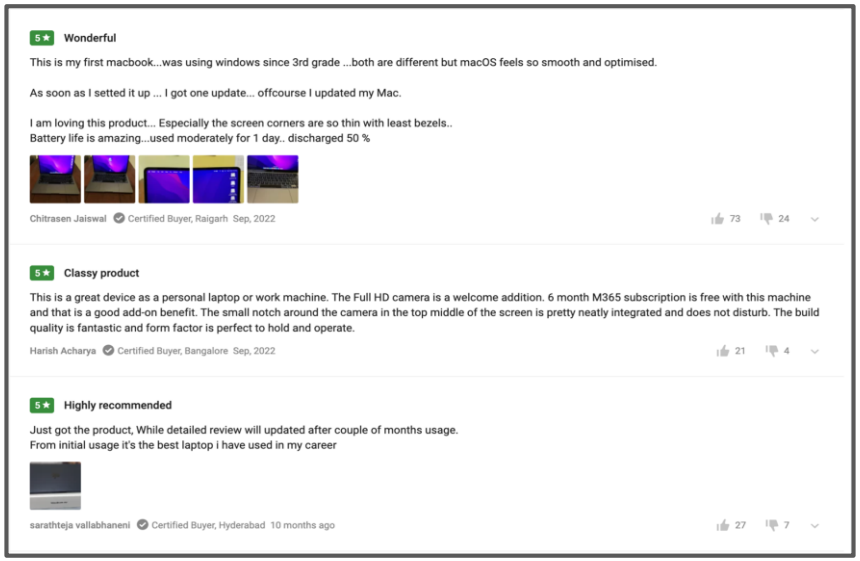
Reinforcement Learning >> Supervised Learning | Query focused Summarization

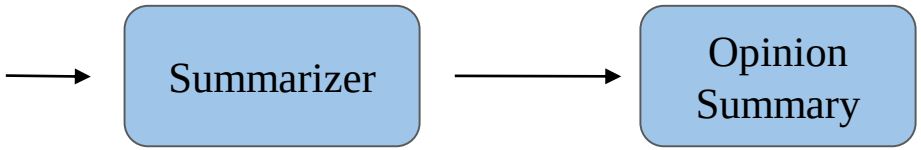
Opinion Summarization

What is Opinion Summarization?

Opinion Summarization is the process of condensing opinions present across opinionated sentences.

Problem of Opinion Summarization:

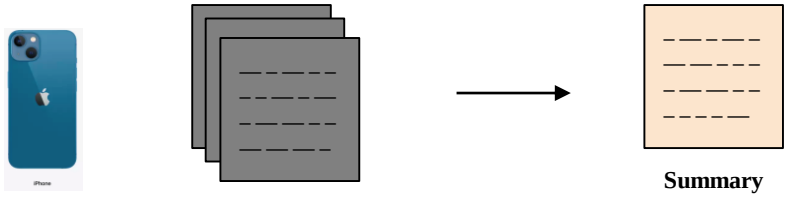




Cite: https://bit.ly/flipkart-macbook

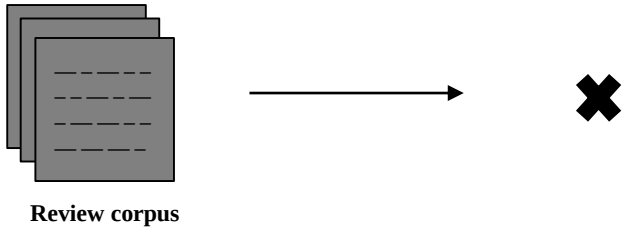
Opinion Summarization Training Paradigms

Supervised Opinion Summarization



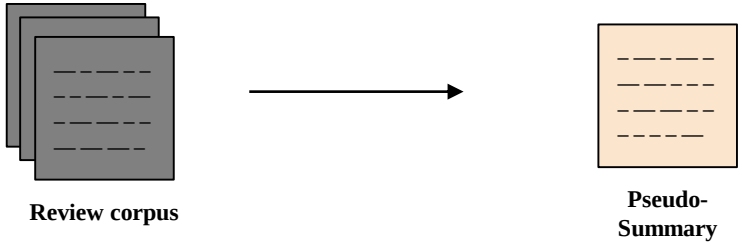
In supervised opinion summarization, we have a review corpus and a corresponding summary.

Unsupervised Opinion Summarization



In unsupervised opinion summarization, we only have a review corpus.

Self-supervised Opinion Summarization



In self-supervised opinion summarization, we use one of the reviews as a pseudo-summary.

Opinion Summarization is done primarily in an unsupervised/self-supervised way

Aspect-sentiment-based Opinion Summarization using Multiple Information Sources

Siledar, T., Makwana, J., and Bhattacharyya, P. (2023). Aspect-sentiment-based opinion summarization using multiple information sources. Proceedings of the 6th Joint International Conference on Data Science & Management of Data (10th ACM IKDD CODS and 28th COMAD)

Problem Statement

Build an opinion summarization system that takes

input data from the following 4 sources:

1. Product Description
2. Product Specification
3. Customer Reviews
4. Question-Answers

and gives summarized version of the product covering all of its salient aspects.

Input: Four sources of information

Output: Summary

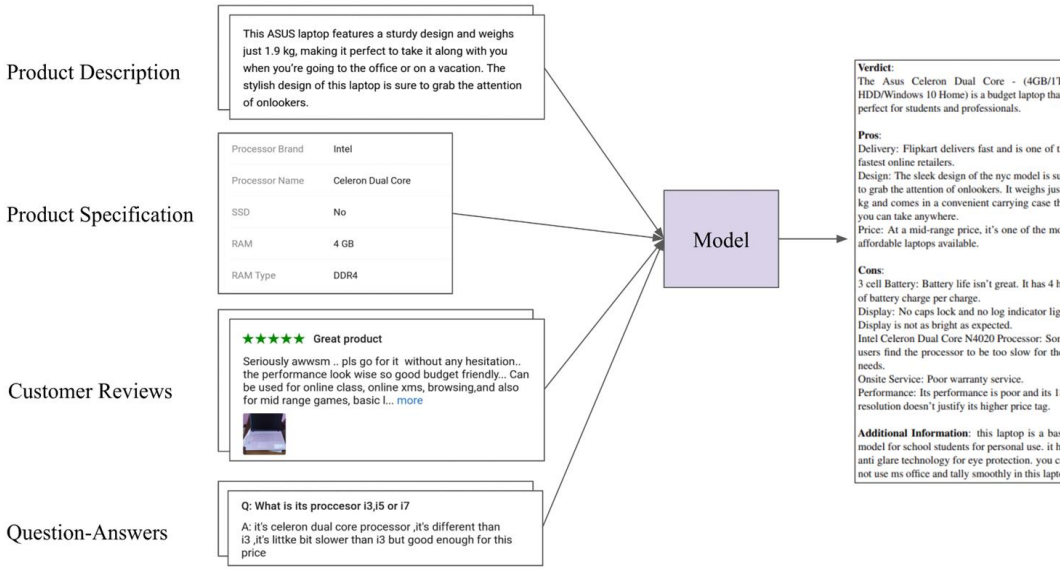
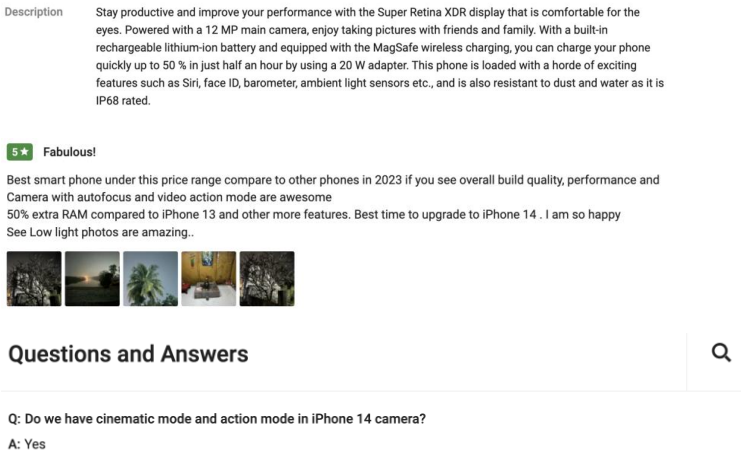


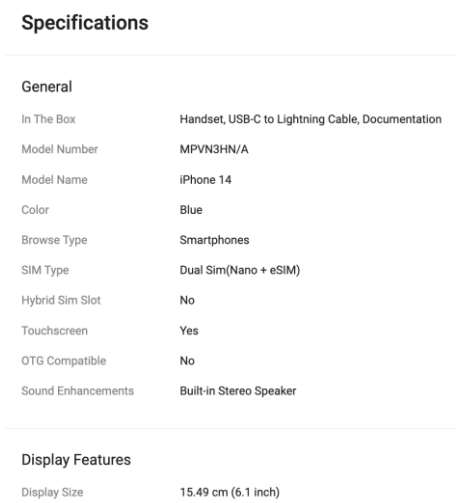
Figure 1: General overview of the problem statement

Motivation

Problem:

Traditionally, most approaches use only reviews to extract opinions and generate summaries. Information sources such as product description, specification and question-answers which contain vital information has not been utilized for generating summaries





All such additional sources have never been considered for opinion summarization

Dataset and Annotation

Training Data:

- [Bražinskas et al., 2021] presented the AMASUM dataset consisting of 33,324 abstractive summaries along with their respective customer reviews for more than 31,000 Amazon products with 326 reviews on average per product
- The summaries written by professionals follow the format of verdict, pros and cons, extracted from different review websites
- They used this dataset to train summarizer models

Dataset	Products	Reviews/Product	Summaries
AMASUM	31,483	326	33,324

Table 1: Statistics of AMASUM dataset which is used to train our summarizer models.

Test Data:

- Annotated products for three categories: Laptops, Mobiles and Tablets by employing 2 male annotators aged 23-27 for the task
- This is Flipkart test set which is used for evaluation of different models.

Category	No of Products
Laptops	25
Mobiles	99
Tablets	23
Total	147

Table 2: Statistics of annotated Flipkart test set.

Approach

Verdict Summary

We filter out top-5 rated reviews and summarize them using our fine-tuned summarizer

Pros/Cons Summary

We perform sentence-level filtering using an off-the-shelf aspect extractor model PyABSA.

For example :{aspect- battery, sentiment- positive}

We use our fine-tuned summarizer to generate the summary

Additional Information

We use a fine-tuned T5 model to generate sentences using question-answer pairs. We pass these sentences through a fine-tuned summarizer to generate the question-answer summary

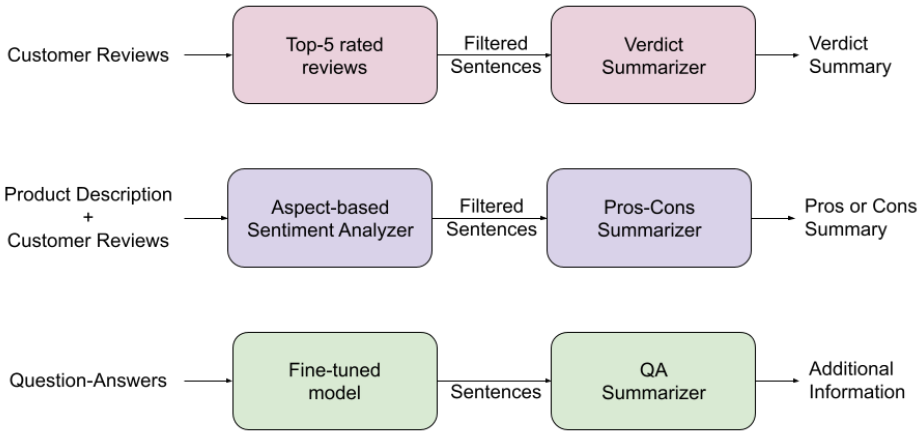


Figure 1. Block diagram depicting the process of our aspect-sentiment-based opinion summarization model.

Results and Analysis

	Model	Verdict			Pros			Cons			Add.		
		R1	R2	RL	R1	R2	RL	R1	R2	RL	R1	R2	RL
Other	RANDOM	10.72	0.60	9.80	13.11	0.61	12.37	10.23	0.45	9.56	-	-	-
	COPYCAT	17.05	1.78	14.50	18.48	0.79	16.35	12.20	0.88	10.34	-	-	-
	SELSUM	31.69	12.50	29.04	11.63	0.84	10.62	7.35	0.45	7.10	-	-	-
Our Work	ASBOS-C	37.27	19.22	35.32	19.63	1.92	18.35	16.74	1.47	15.87	-	-	-
	ASBOS-CD	37.27	19.22	35.32	20.15	2.17	18.82	16.83	1.54	15.99	-	-	-
	ASBOS-CS	37.27	19.22	35.32	24.01	6.86	23.08	21.11	6.98	20.01	-	-	-
	ASBOS-ALL	**37.27**	**19.22**	**35.32**	**24.36**	**7.91**	**23.11**	**21.17**	**7.07**	**20.08**	**8.89**	**1.38**	**8.55**

C denotes customer reviews, CD denotes customer reviews and product description, CS denotes customer reviews and product specification, ALL denotes all four information sources.

Our aspect-sentiment-based opinion summarizer using customer reviews (ASBOS-C) performs much better than the other models indicating the effect of using our approach of aspect-sentiment based selector.

Verdict:
The Asus Celeron Dual Core - (4GB/1TB HDD/Windows 10 Home) is a budget laptop that's perfect for students and professionals.

Pros:
Delivery: Flipkart delivers fast and is one of the fastest online retailers.
Design: The sleek design of the nyc model is sure to grab the attention of onlookers. It weighs just 1 kg and comes in a convenient carrying case that you can take anywhere.
Price: At a mid-range price, it's one of the most affordable laptops available.

Cons:
3 cell Battery: Battery life isn't great. It has 4 hrs. of battery charge per charge.
Display: No caps lock and no log indicator light. Display is not as bright as expected.
Intel Celeron Dual Core N4020 Processor: Some users find the processor to be too slow for their needs.
Onsite Service: Poor warranty service.
Performance: Its performance is poor and its 18k resolution doesn't justify its higher price tag.

Additional Information: this laptop is a basic model for school students for personal use. it has anti glare technology for eye protection. you can not use ms office and tally smoothly in this laptop.

Here the red part comes from the product description, the orange part comes from the product specification, the blue part is through question-answers whereas the rest is from customer reviews.

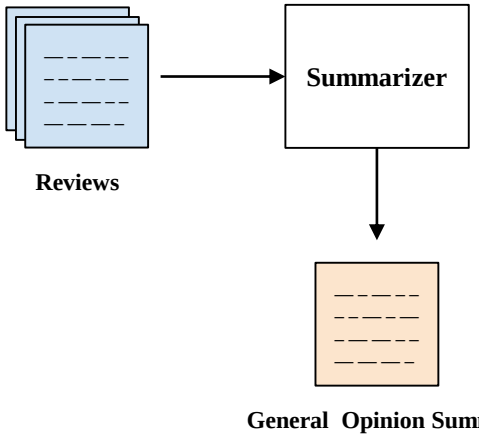
Synthesize, if you do not have: Effective Synthetic Dataset Creation Strategies for Self-Supervised Opinion Summarization in Ecommerce

Tejpalsingh Siledar, Suman Banerjee, Amey Patil, Sudhanshu Singh, Muthusamy Chelliah, Nikesh Garera, and Pushpak Bhattacharyya. 2023. Synthesize, if you do not have: Effective Synthetic Dataset Creation Strategies for Self-Supervised Opinion Summarization in E-commerce. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 13480–13491, Singapore. Association for Computational Linguistics.

Problem Statement

General Opinion Summarization

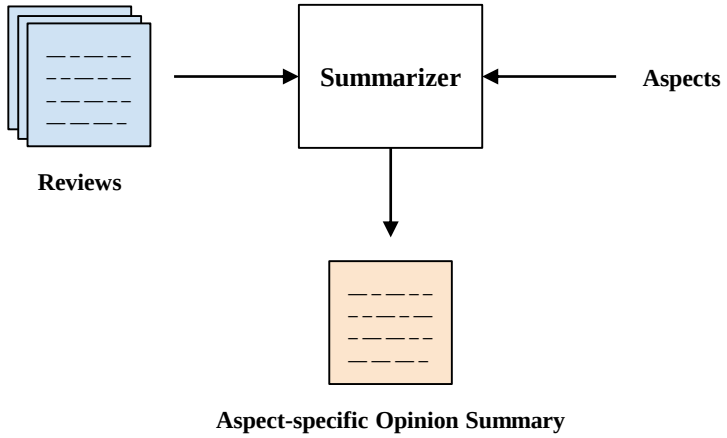
Input: Reviews
Output: General Opinion Summary



Example: A summary about the opinions of an Iphone product

Aspect-specific Opinion Summarization

Input: Reviews, Aspects
Output: Aspect-specific Opinion Summary



Example: A summary about the opinions of the battery life of an Iphone product

Motivation

General Opinion Summarization

Problem: Faithfulness issue in the generated summaries

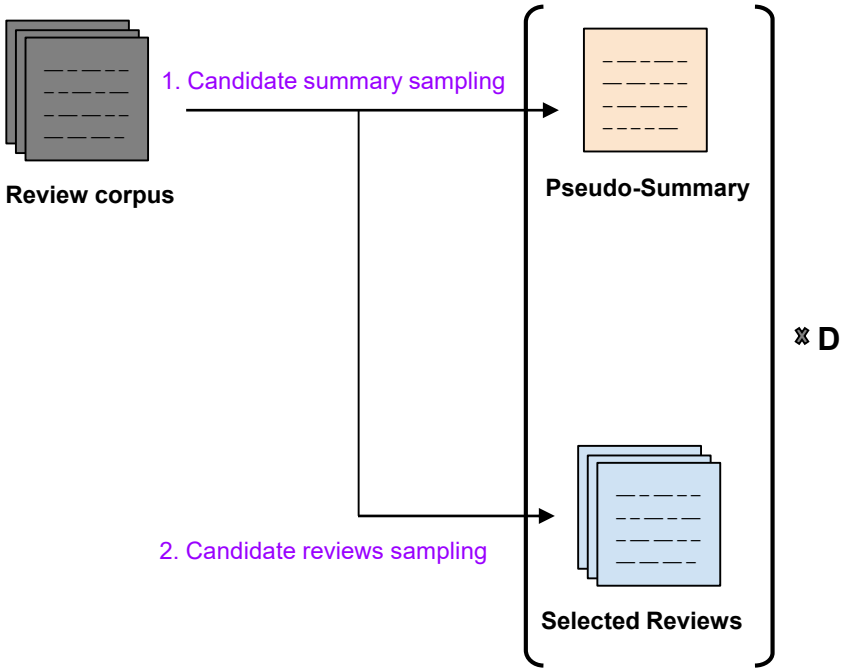
This is a nice teapot, but the color is not as bright as the picture. It is more of a dark turquoise than a light blue. I was hoping it would be more of an aqua blue, but it is more like a dark aqua. It still looks nice, but I would have preferred the color to be more like the photo.

Aspect-specific Opinion Summarization

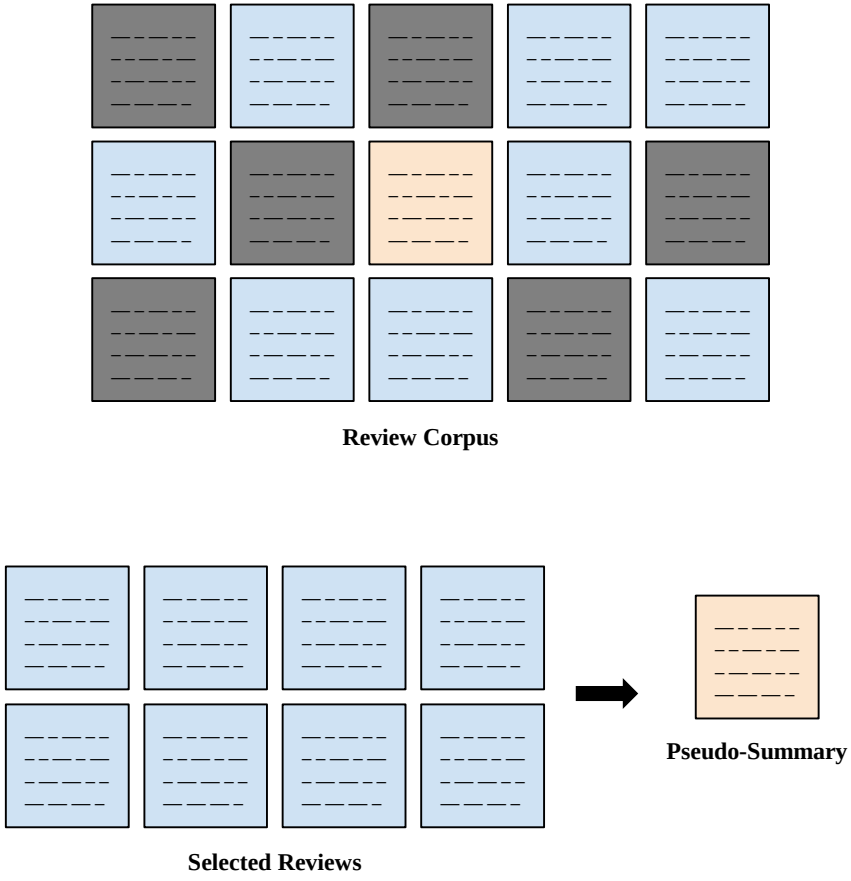
Problem: Models depend on pre-defined aspects and seed words for training, making them to not be generalizable.

Ports - the hdmi ports stopped working and the tv would go black for about three to five seconds every few minutes. it has all of the inputs for us, and more.
Picture - the picture quality was better than my 5 year old lcd samsung. the picture is decent, but not as good as the htpc.
Sound - i bought the tv at fry's after salesperson assured me headphones would work with this. picture is acceptable and sound is above average for the price.

General Opinion Summarization



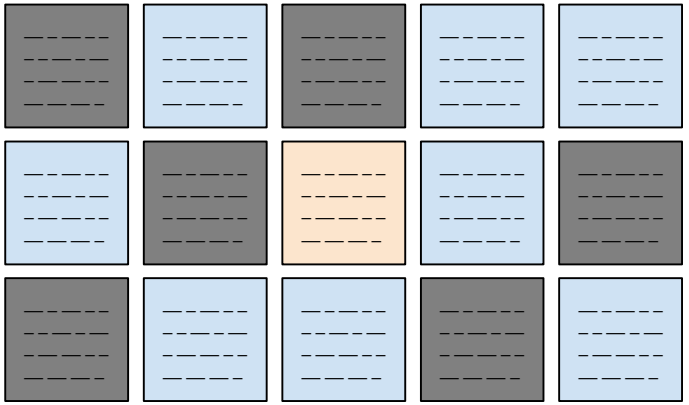
Self-supervised modeling for Opinion Summarization



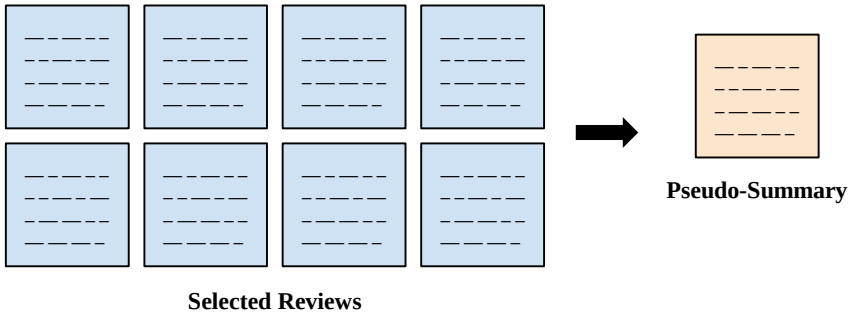
Our Approach

- Hypothesis: Proper curation of synthetic datasets can solve the issue of faithfulness
- Use lexical (R1) and semantic (cosine) similarity for pseudo-summary selection
- We construct a NxN matrix using the N reviews from the review corpus where each cell m_{ij} represents a score between review r_i and r_j
- We use (selected reviews, pseudo-summary) pairs with scores above a certain threshold as our training points

$$
\begin{array}{c} \\ r_1 \\ r_2 \\ . \\ . \\ r_N \end{array}
\begin{array}{c}
\begin{array}{cccccc} r_1 & r_2 & . & . & r_N \end{array} \\
\left[\begin{array}{cccccc}
0 & & & & & \\
 & 0 & & & & \\
 & & 0 & & & \\
 & & m_{ij} & 0 & & \\
 & & & & & 0
\end{array}\right]
\end{array}
$$



Review Corpus



Selected Reviews

Pseudo-Summary

Our Approach

	r_1	r_2	r_3	r_4	r_5	r_6	r_7	r_8	r_9	r_{10}	Top-8 mean
r_1	0.00	0.01	0.60	0.47	0.56	0.54	0.55	0.61	0.56	0.55	0.55
r_2	0.01	0.00	0.03	0.05	0.04	0.03	0.06	0.00	0.03	0.07	0.03
r_3	0.60	0.03	0.00	0.72	0.60	0.62	0.56	0.66	0.66	0.76	0.64
r_4	0.47	0.05	0.72	0.00	0.57	0.55	0.49	0.62	0.68	0.61	0.58
r_5	0.56	0.04	0.60	0.57	0.00	0.57	0.57	0.59	0.49	0.58	0.56
r_6	0.54	0.03	0.62	0.55	0.57	0.00	0.47	0.70	0.76	0.55	0.59
r_7	0.55	0.06	0.56	0.49	0.57	0.47	0.00	0.70	0.76	0.55	0.58
r_8	0.61	0.00	0.66	0.62	0.59	0.70	0.43	0.00	0.71	0.61	0.61
r_9	0.56	0.03	0.66	0.68	0.49	0.76	0.41	0.71	0.00	0.52	0.59
r_{10}	0.55	0.07	0.76	0.61	0.58	0.55	0.59	0.61	0.52	0.00	0.59

For a threshold say 0.6

Synthetic pairs →

$[r_1, r_4, r_5, r_6, r_7, r_8, r_9, r_{10}], r_3$

$[r_1, r_3, r_4, r_5, r_6, r_7, r_9, r_{10}], r_8$

Results

	Model	asp?	Oposum+ R1↑	Oposum+ R2↑	Oposum+ RL↑	Amazon R1↑	Amazon R2↑	Amazon RL↑
Extractive	Clustroid	✗	33.44	11.00	20.54	29.27	4.41	17.78
	LexRank	✗	35.42	10.22	20.92	29.46	5.53	17.74
	QT	✗	37.72	14.65	21.69	34.04	7.03	18.08
	$\text{Acesum}_{\text{ext}}$	✓	38.48	15.17	22.82	x	x	x
	$\text{SW-LOO}_{\text{ext}}$	✓	40.45	19.13	23.20	x	x	x
	$\text{NLI-LOO}_{\text{ext}}$	✓	39.79	18.33	23.49	x	x	x
Abstractive	MeanSum	✗	26.25	4.62	16.49	29.20	4.70	18.15
	CopyCat	✗	27.98	5.79	17.07	31.97	5.81	20.16
	Acesum	✓	32.98	10.72	20.27	x	x	x
	SW-LOO	✓	<u>36.19</u>	**12.17**	<u>21.11</u>	x	x	x
	NLI-LOO	✓	31.22	9.93	19.08	x	x	x
	PlanSum	✗	30.26	5.29	17.48	32.87	6.12	19.05
	ConsistSum	✗	x	x	x	33.32	5.94	<u>21.41</u>
	MultimodalSum	✗	33.08	7.46	19.75	34.19	7.05	20.81
	TransSum	✗	x	x	x	34.23	<u>7.24</u>	20.49
	COOP	✗	x	x	x	**36.57**	7.23	21.24
	Our Model	✗	**36.57***	8.79	**21.35**	<u>35.46</u>*	**7.30**	**21.50**

Table 1: Evaluation for general summaries on Oposum+ and Amazon test sets. asp? indicates systems that use human-specified aspects. Bold and underline indicate best and second-best scores using abstractive systems.* indicates pvalue < 0.05 on paired t-test against MultimodalSum. Our model outperforms existing models on the task of general opinion summarization.

Model	SummaC ↑	CTC ↑	FactCC ↑	FactGraph ↑
PlanSum	0.33	0.81	0.16	0.21
Multimodalsum	0.38	**0.85**	0.46	0.53
Our Model	**0.40**	**0.85**	**0.68**	**0.66**

Table 2: Faithfulness Evaluation. Our model outperforms alternatives on three faithfulness measuring metrics: SummaC, FactCC, and FactGraph.

Amazon	Faithfulness ↑	Coherent ↑	Concise ↑	Fluency ↑
PlanSum	-0.92	-0.91	-0.84	-0.72
Multimodalsum	0.41	**0.45**	0.40	0.31
Our Model	**0.51**	**0.45**	**0.41**	**0.41**

Table 3: Best-Worst Scaling. Our model receives better scores on three criteria. Best values are in bold.

Analysis (shorts product)

Gold Summaries

- These look very nice on. The mesh in them run a little tight though so you may want to go up in size. They are roomy in the pockets and can hold onto a lot of items. They are a great price for the quality and I would highly recommend them to anyone looking for a new pair of swim trunks.
- These swim trunks look great and the waist fits as expected for the size. They are also a good deal for the price. However, the inner mesh liner is much smaller than the size of the outer trunks. You will probably have to cut out the mesh liner to wear them comfortably.
- These are nice fitting shorts but the mesh liner felt tight and I cut it out. You might want to size up if you are unsure. The deep pockets come in handy. These shorts are stylish and they look good. They are more comfortable if the inner mesh liner is removed.

PlanSum

- i bought this for my husband and she loves it. it is a very good product and the fit is good. the only reason i didn't give it 5 stars is because it does not fit in the middle of the back of my shirt. i would recommend this to anyone who is looking for a good quality product.

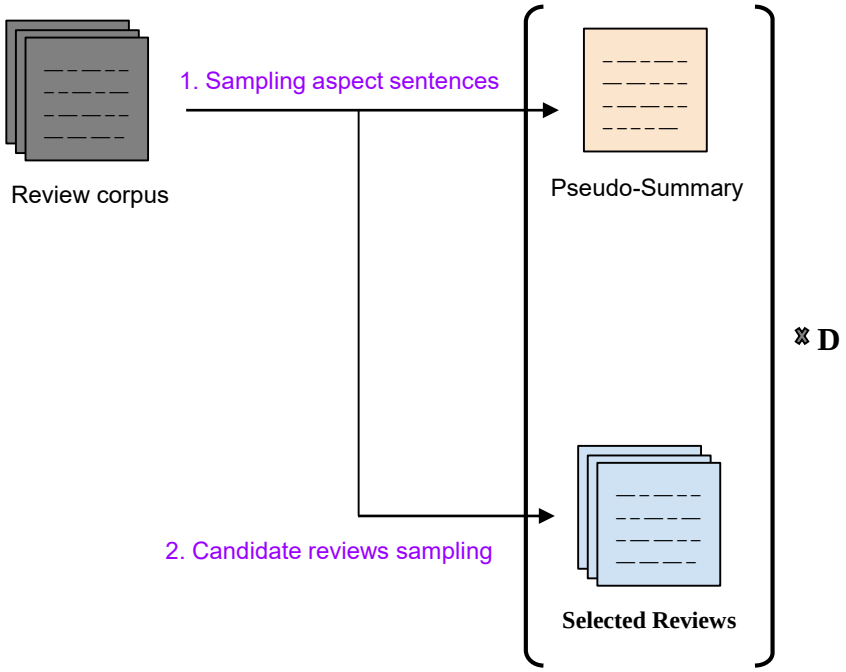
MultimodalSum

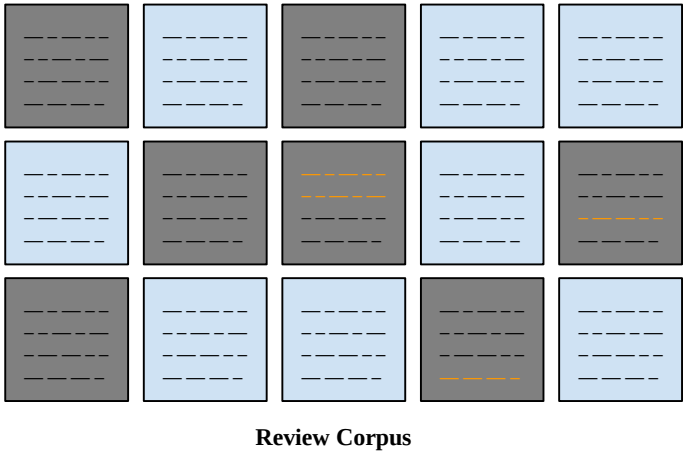
- I bought these for my husband to wear to the beach. They fit well and look good. The only thing I don't like is that the mesh liner is a little tight. He has to pull it up every time he goes in and out of the water.

Our Model

- These shorts are great. They fit well and look great. The only thing I would change is the liner. It is a little too small and I had to cut it out. Other than that, I would recommend these shorts.

Aspect-specific Opinion Summarization





Approach

- Hypothesis: An automatic aspect mapping and proper curation of synthetic datasets can remove the dependency on pre-defined aspects and seed words

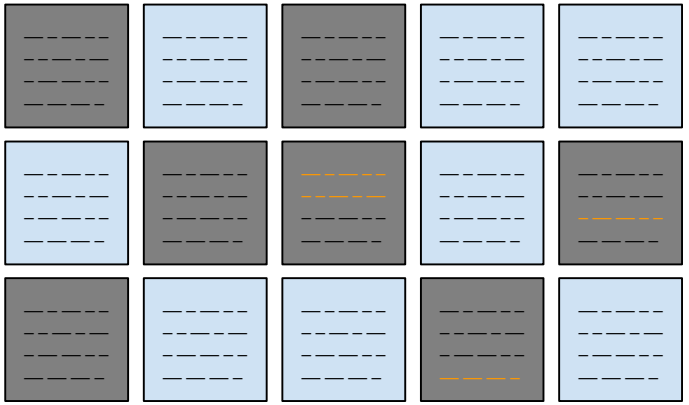
- Aspect mapping: Helps in identifying which sentence belongs to which aspect

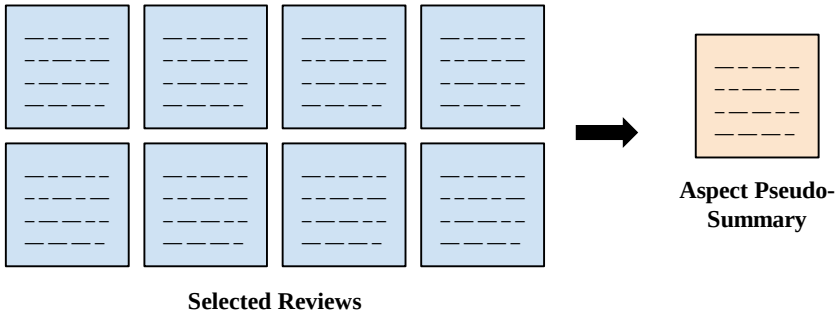
  display: screen, display, screen quality, ...
  camera: selfie camera, camera, camera quality, ...
  battery: battery life, battery backup, ...
  processor: processor, intel processor, ...

  We use an off-the-shelf aspect extractor to identify aspects and map them as shown above

- Synthetic Dataset: Create (selected reviews, aspects, pseudo-summary) triplets and use them as training points

  Example: We select a few sentences for the aspect say “battery” and call it the pseudo-summary. We then use the lexical and semantic scores to select reviews that can act as input.





Example of Aspect Mapping

Pre-defined aspect-seed words mapping	
Pre-defined mapping	looks: looks color stylish looked pretty quality: quality material poor broke durable size: fit fits size big space
Automatic aspect mapping	
Fine-grained clusters	selfie camera, camera, camera quality, back camera, ... screen, display, display, screen quality, ... battery life, battery backup, battery capacity, battery, ... processor, intel processor, ...
Coarse-grained clusters	display, display camera, camera, camera, ... battery, battery processor
Aspect mapping	display: screen, display, screen quality, ... camera: selfie camera, camera, camera quality, ... battery: battery life, battery backup, ... processor: processor, intel processor, ...

Results and Analysis

	Model	*asp?*	Oposum+ R1↑	Oposum+ R2↑	Oposum+ RL↑	Flipkart R1↑	Flipkart R2↑	Flipkart RL↑
Extractive	LexRank	✗	22.51	3.35	17.27	10.41	0.93	8.72
	QT	✗	23.99	4.36	16.61	14.11	1.71	9.56
	$\text{Acesum}_{\text{ext}}$	✓	26.16	5.75	18.55	x	x	x
	$\text{SW-LOO}_{\text{ext}}$	✓	28.14	6.10	19.51	x	x	x
	$\text{NLI-LOO}_{\text{ext}}$	✓	26.78	6.48	18.07	x	x	x
Abstractive	MeanSum	✗	24.63	3.47	17.53	10.64	1.33	9.78
	CopyCat	✗	26.17	4.30	18.20	13.48	1.92	10.35
	Acesum	✓	29.53	6.79	<u>21.06</u>	x	x	x
	SW-LOO	✓	<u>30.00</u>	**6.92**	20.76	x	x	x
	NLI-LOO	✓	28.90	6.60	20.11	x	x	x
	ASBOS	✓	23.45	4.37	16.85	<u>14.62</u>	<u>2.23</u>	<u>11.56</u>
	Our Model	✗	**30.95**	**6.92**	**21.73**	**20.15**	**2.86**	**15.99**

Table 1: Evaluation for aspect summaries on Oposum+ and Flipkart test sets. asp? indicates systems that use human-specified aspects. Bold and underline indicate best and second-best scores using abstractive systems. Our model without relying on any human-specified aspects or seed words learns the task of aspect-specific opinion summarization and outperforms alternatives.

AceSum

Ports - the hdmi ports stopped working and the tv would go black for about three to five seconds every few minutes. it has all of the inputs for us, and more.

Picture - the picture quality was better than my 5 year old lcd samsung. the picture is decent, but not as good as the htpc.

Sound - i bought the tv at fry's after salesperson assured me headphones would work with this. picture is acceptable and sound is above average for the price.

Our Model

Ports - i have had this tv for a few months now and i have had no issues with it, it has all of the hdmi ports i need and the picture quality is great.

Picture - the picture is great, the tv is easy to set up, and it has all of the inputs i need, but it is not a smart tv.

Sound - the sound is decent, but i have a surround sound system so i don't use the tv's speakers much, but it is a great tv for the price.

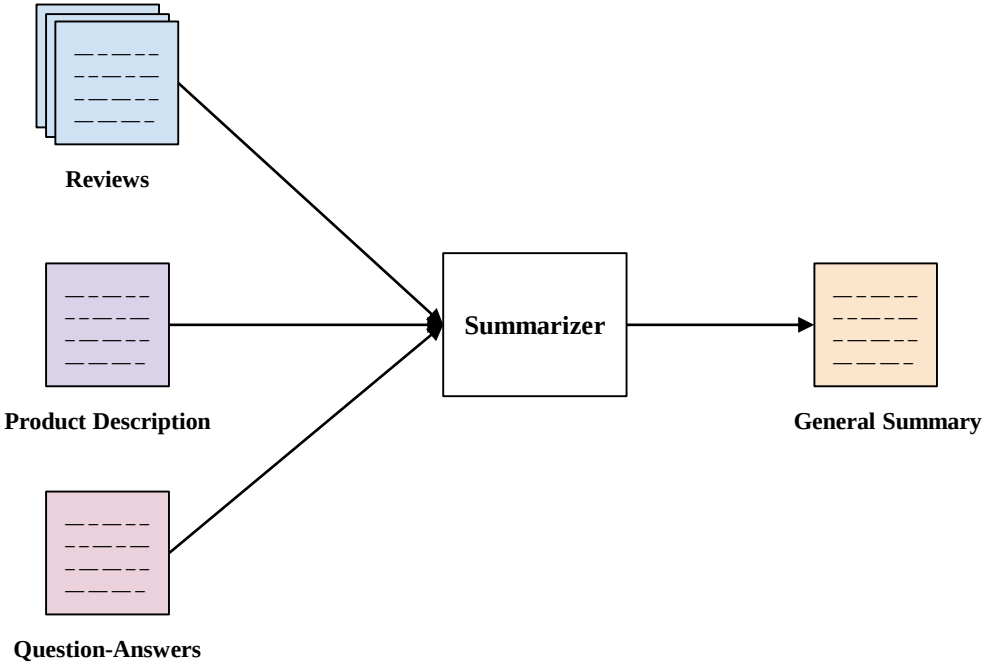
Product Description and QA Assisted Self-Supervised Opinion Summarization

Siledar, Tejpalsingh, Rupasai Rangaraju, Sankara Sri Raghava Ravindra Muddu, Swaprava Nath, Pushpak Bhattacharyya, Suman Banerjee et al. "Product Description and QA Assisted Self-Supervised Opinion Summarization."In Findings of the Association for Computational Linguistics: NAACL 2024

Problem Statement

Input: Reviews, Product Description, Question-Answers

Output: General Opinion Summary



MultimodalSum
I bought this product to scan my negatives. It does not work with Windows XP. I have tried to contact the company several times and have not received a response. I am very disappointed in the product. I would not recommend it to anyone.
Our Model (MEDOS)
I purchased the **VuPoint FS-C1-VP Film and Slide Digital Converter** to scan my **35mm** film and slide negatives. It is not compatible with Windows XP. <u>The software does not work with Windows 7 or 8</u>. I have tried to contact the company and they do not respond to my emails. I would not recommend this product to anyone.

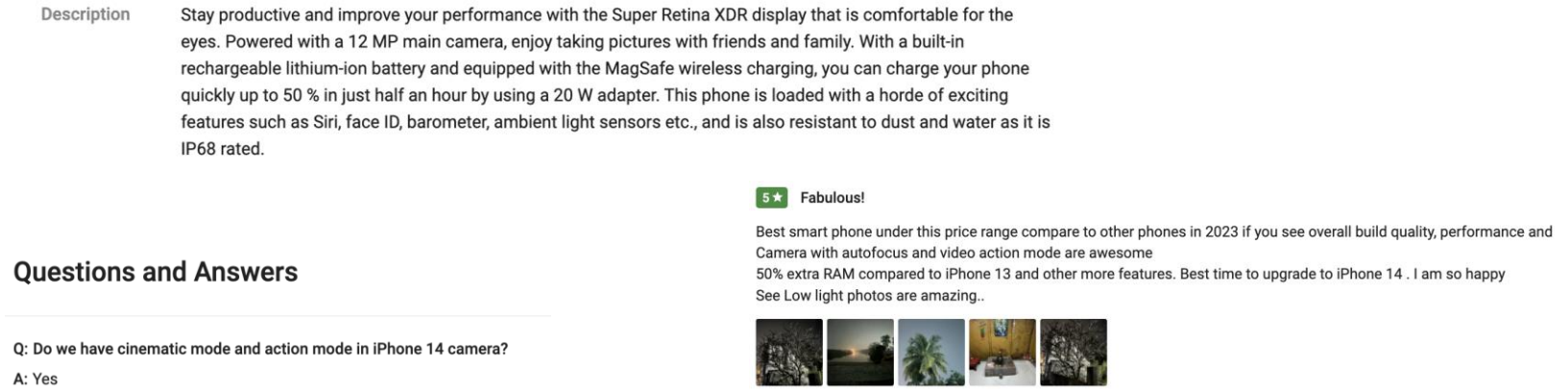
Table 1: Multimodal Sum vs. MEDOS generated summary for a product from the Amazon test set. Information assisted from product description and question-answers are in bold and underline respectively. Our model is able to capture essential information from the product description and question-answers, not found in reviews. This makes our model-generated summaries more informative while still retaining the consensus opinions from reviews as evident in the above example.

Motivation

Problem:

1. Our previous multi-source model uses simple rules to combine information from multiple sources
2. It depended on supervised datasets for training which are not available in large amounts

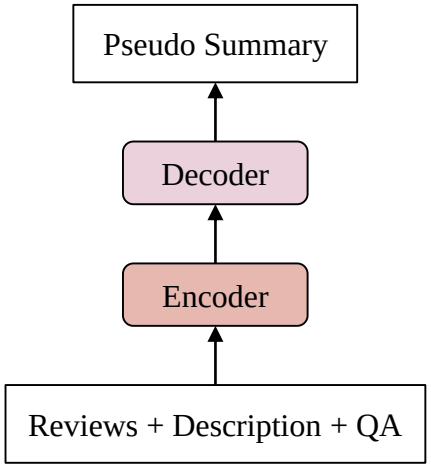
Can the model learn to automatically pick information from different sources for summarization?



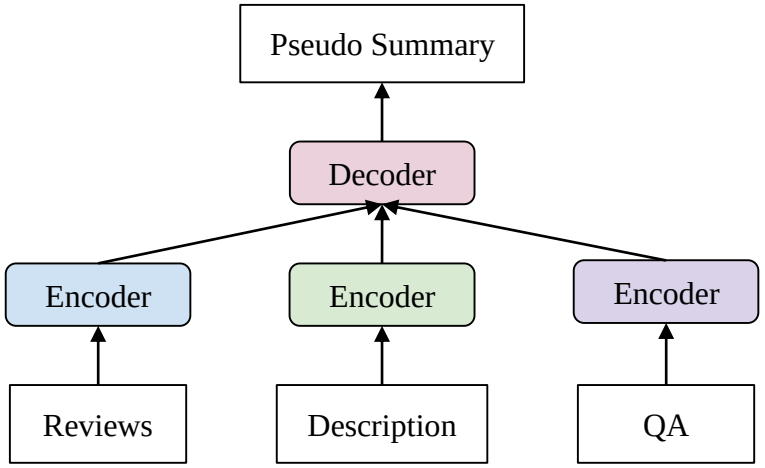
No end to end model that can pick essential details from multiple sources automatically to generate summaries

Approach

We create synthetic datasets using lexical and semantic similarity. Pseudo summary selection is done by selecting the review that is closest to description and question-answers and a set of reviews.



a) Single-Encoder Decoder Framework



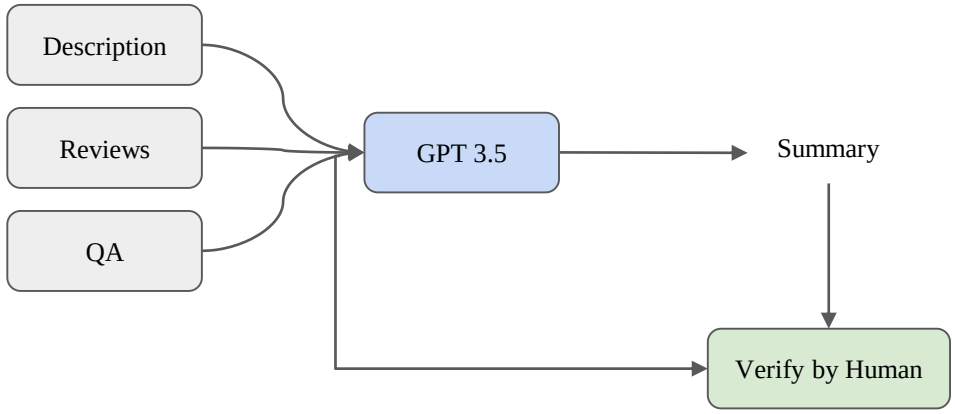
b) Multi-Encoder Decoder Framework (MEDOS)

We investigated two approaches

a) Single-Encoder Decoder Framework
b) Multi-Encoder Decoder Framework

Test set creation

- In the absence of any test sets that contain additional sources, we extended Amazon, Oposum+, and Flipkart to contain such sources and leveraged ChatGPT to annotate summaries using reviews and additional sources as input
- We prompt the GPT 3.5 and get the summaries verified by human raters



	Original			Extended (Ours)					
	Amazon	Oposum+	Flipkart	Amazon GPT-R	Oposum+ GPT-R	Flipkart GPT-R	Amazon GPT-RDQ	Oposum+ GPT-RDQ	Flipkart GPT-RDQ
#domains	4	6	3	4	6	3	4	6	3
#test set	32	30	145	32	30	145	32	30	145
#reviews/product	8	10	10	8	10	10	8	10	10
#summaries/product	3	3	1	**3**	**3**	**1**	**3**	**3**	**1**
#summaries	96	90	145	**96**	**90**	**145**	**96**	**90**	**145**
#descriptions	-	-	-	-	-	-	21	17	145
#question-answers	-	-	-	-	-	-	11	10	145

Table 1: Statistics for original and extended test sets. GPT-R indicates the use of reviews whereas GPT-RDQ indicates the use of reviews, description, and question-answers to generate summaries using ChatGPT. Bold represents our contributions. In the respective extended versions, reviews are the same as the original.

	Info. ↑	Faith. ↑	Coh. ↑	Con. ↑	Flu. ↑
Human	3.88	3.91	3.68	3.83	3.62
GPT-R	4.02	4.13	4.02	4.09	3.98
GPT-RDQ	4.10	4.16	4.16	4.23	4.16

Table 2: Annotation quality. Both GPT-R and GPT-RDQ summaries score higher on all the metrics on average compared to human-annotated summaries. Scores range from 1-5. Info-informativeness, Faith-faithfulness, Coh-coherence, Con-conciseness, Flu-fluency

Results and Analysis

					Amazon			Amazon GPT-R			Amazon GPT-RDQ		
abs?	**Model**	**R**	**D**	**Q**	R1 ↑	R2 ↑	RL ↑	R1 ↑	R2 ↑	RL ↑	R1 ↑	R2 ↑	RL ↑
✗	Random	✓	✗	✗	27.86	3.87	16.68	20.69	1.56	12.55	18.83	1.45	12.03
✗	Oracle	✓	✗	✗	44.47	13.83	30.85	33.69	6.04	22.88	31.83	5.77	22.04
✗	Clustroid	✓	✓	✓	29.27	4.41	17.78	22.74	2.16	14.03	21.31	2.57	13.38
✗	LexRank	✓	✓	✓	29.46	5.53	17.74	22.82	3.08	13.77	19.30	4.31	12.90
✗	QT	✓	✓	✓	34.04	7.03	18.08	23.01	2.48	12.05	21.78	3.25	12.36
✓	CopyCat	✓	✗	✗	31.97	5.81	20.16	20.09	1.79	12.94	20.54	1.94	13.85
✓	PlanSum	✓	✗	✗	32.87	6.12	19.05	20.49	1.76	12.44	19.09	1.58	12.02
✓	ConsistSum	✓	✗	✗	33.32	5.94	**21.41**	-	-	-	-	-	-
✓	MultimodalSum	✓	✓	✗	34.19	7.05	20.81	21.43	1.58	13.20	20.39	2.08	12.83
✓	TransSum	✓	✗	✗	34.23	7.24	20.49	-	-	-	-	-	-
✓	COOP	✓	✗	✗	**36.57**	7.23	21.24	-	-	-	-	-	-
✓	T5-concat	✓	✓	✓	28.04	4.46	16.39	21.28	**2.57**	13.00	20.61	2.72	13.33
✓	BART-concat	✓	✓	✓	32.35	6.49	19.78	22.32	2.27	13.74	21.75	2.39	13.57
✓	**MEDOS**	✓	✓	✓	34.63	**7.48**	20.97	**23.92***	2.27*	**14.69***	**25.44***	**4.16***	**16.45***

Table 1: Results on Amazon test set and its extensions. R, D, Q indicate the presence of reviews, description, and question-answers respectively in the input. abs? indicate abstractive systems. Bold and underline indicate best and second-best scores using abstractive systems. * indicates pvalue < 0.05 on paired t-test against MultimodalSum. Overall our combination of SDC approach and MEDOS outperforms baselines across all three test sets.

Gold (Amazon GPT-RDQ)
The **VuPoint Film and Slide Scanner** presents a mixed picture. It claims ease & efficiency for converting film to digital format, but user experiences differ. While some commend its user-friendliness and recommend it, others report frustrating issues like bleached images, compatibility problems, and subpar results. The need for XP compatibility limits its usefulness for modern systems. It offers potential for simple scanning but may require persistence to achieve desired outcomes.
MultimodalSum
I bought this product to scan my negatives. It does not work with Windows XP. I have tried to contact the company several times and have not received a response. I am very disappointed in the product. I would not recommend it to anyone.
BART-concat
I bought this Scanner to scan my slides and film. I have Windows XP and it does not work with it. I tried to download the drivers from the web site but they are not available. I called the company and they said they would send me a new one but I have not received it yet. I am very disappointed.
Our Model (MEDOS)
I purchased the **VuPoint FS-C1-VP Film and Slide Digital Converter** to scan my **35mm film** and slide negatives. It is not compatible with Windows XP. The software does not work with Windows 7 or 8. I have tried to contact the company and they do not respond to my emails. I would not recommend this product to anyone.

Table 2: Comparative Analysis. ChatGPT-generated summary using reviews, description, and question-answers (GPT-RDQ) followed by different model generated summaries for an Amazon test set product. Information assisted from the description and question-answers are in bold and underline respectively. MEDOS is able to capture vital information from additional sources which won't be possible using only reviews.

Contributions

1. A novel synthetic dataset creation (SDC) approach that enables supervised training in the presence of additional sources without the need for any annotated training datasets. We propose a Multi-Encoder Decoder framework for Opinion Summarization (MEDOS) to effectively fuse information from reviews, product description, and question-answers (QA). To the best of our knowledge, we are the first to do multi-source self supervised opinion summarization.

1. Extensions to e-commerce test sets namely Amazon (Bražinskas et al., 2020) and Oposum+ (Amplayo et al., 2021) to include additional sources. For comparison, we extend: Amazon, Oposum+, and Flipkart by curating six new test sets: Amazon R, Amazon RDQ, Oposum+ R, Oposum+ RDQ, Flipkart R, and Flipkart RDQ leveraging ChatGPT to annotate summaries. We extend the test sets to contain 662 opinion summaries across six curated test sets

1. Experimental demonstrations of our SDC approach and MEDOS model in outperforming the SOTA model on nine test sets on average by 14.5% in ROUGE-1 F1

1. Comparative and qualitative analysis indicating the importance of sources such as product description and question-answers in generating more informative summaries compared to existing models

Evaluation using LLMs

Siledar, Tejpalsingh, Swaroop Nath, Sankara Sri Raghava Ravindra Muddu, Rupasai Rangaraju, Swaprava Nath, Pushpak Bhattacharyya, Suman Banerjee et al. "One Prompt To Rule Them All: LLMs for Opinion Summary Evaluation." *arXiv preprint arXiv:2402.11683* (2024).

Problem with ROUGE & BERT Score as Metrics

Opinion summaries are generally evaluated using Rouge or BertScore which are reference based metrics

Issue?

1. Requires test sets which are available only for a few domains.
2. Does not give a complete picture of how good the opinion summaries are.
3. ROUGE only focuses on n-gram overlap.

Solution

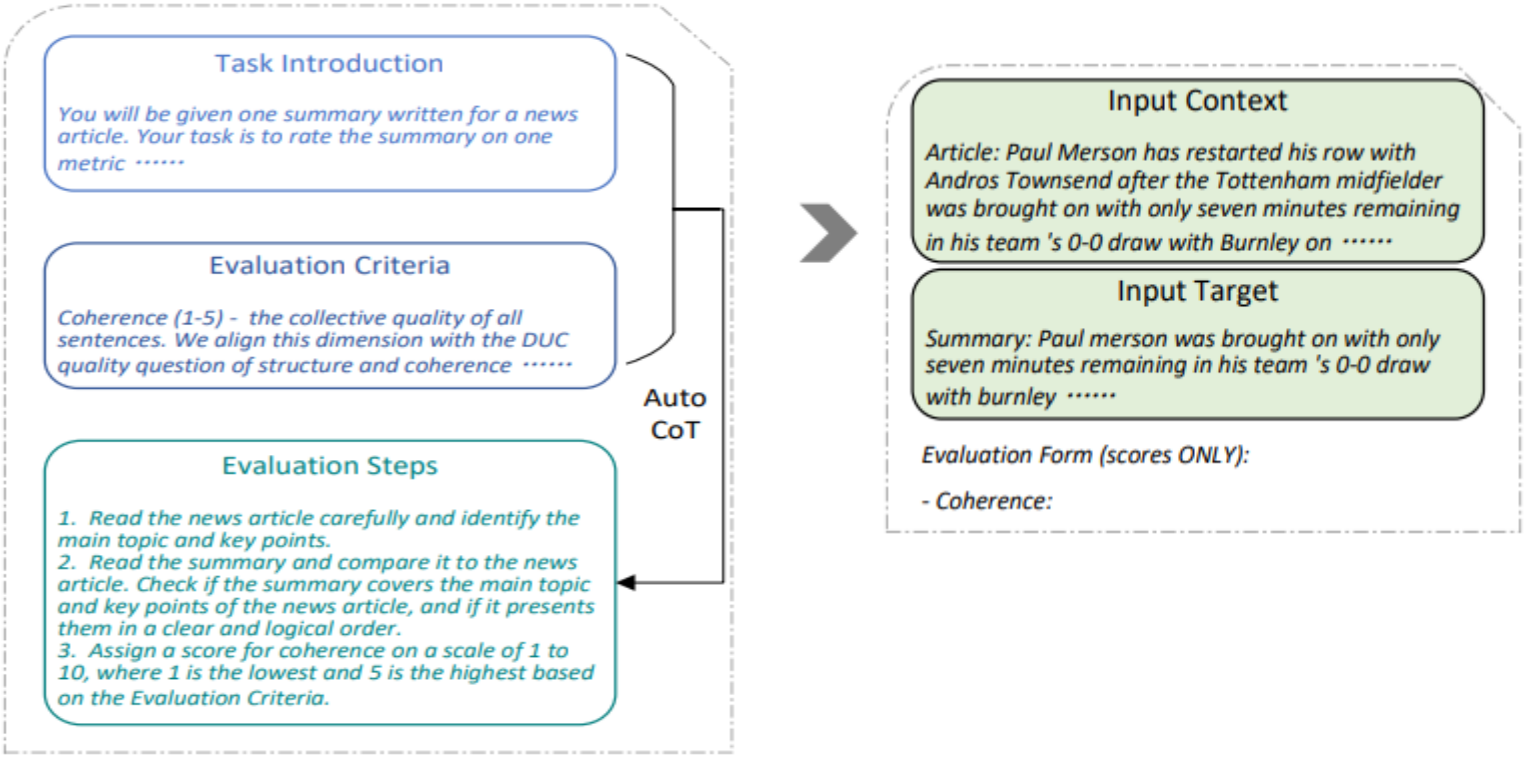
Automatically evaluate summaries using different reference summary-free metrics

Few Reference Summary-Free Metrics

- Fluency: The quality of summary in terms of grammar, spelling, punctuation, capitalization, word choice, and sentence structure and should contain no errors.
- Coherence: The collective quality of all sentences. The summary should be well structured and well-organized.
- Relevance: The summary should not contain opinions that are either not consensus or important
- Faithfulness: Every piece of information mentioned in the summary should be verifiable/supported/inferred from the reviews only.
- Aspect-Coverage: The summary should cover all the aspects that are majorly being discussed in the reviews.
- Sentiment Consistency: All the aspects being discussed in the summary should accurately reflect the consensus sentiment of the corresponding aspects from the reviews.
- Specificity: The summary should avoid containing generic opinions. All the opinions within the summary should contain detailed and specific information about the consensus opinions.

… and many more

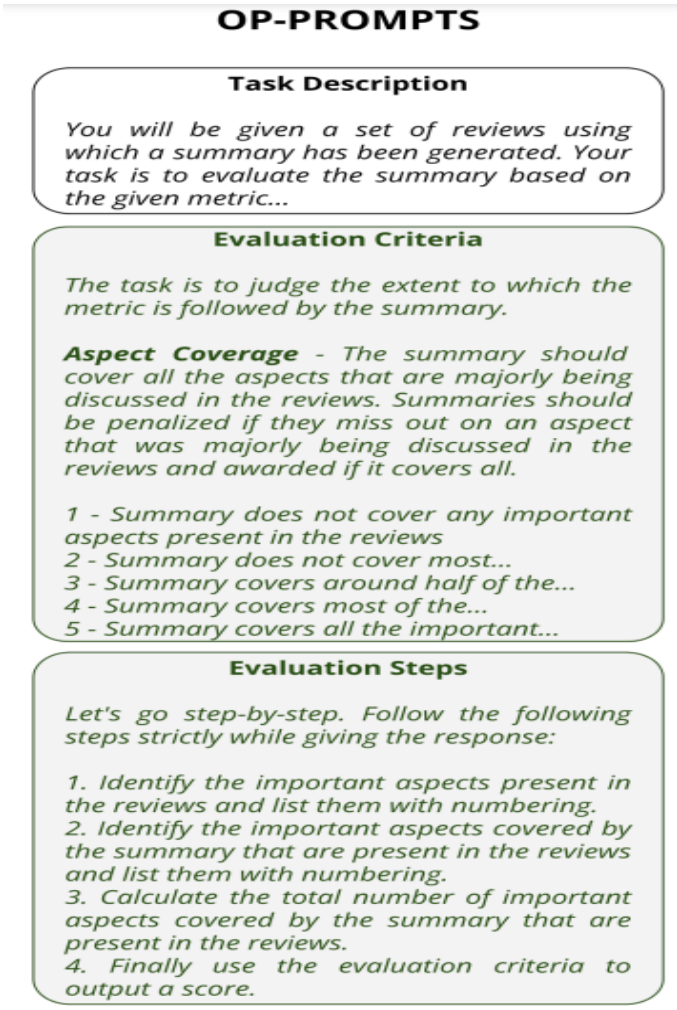
G-EVAL: NLG Evaluation using GPT-4 with Better Human Alignment



One Prompt To Rule Them All: LLMs for Opinion Summary Evaluation

Approach 1: OP-PROMPTS: Dimension dependent set of prompts

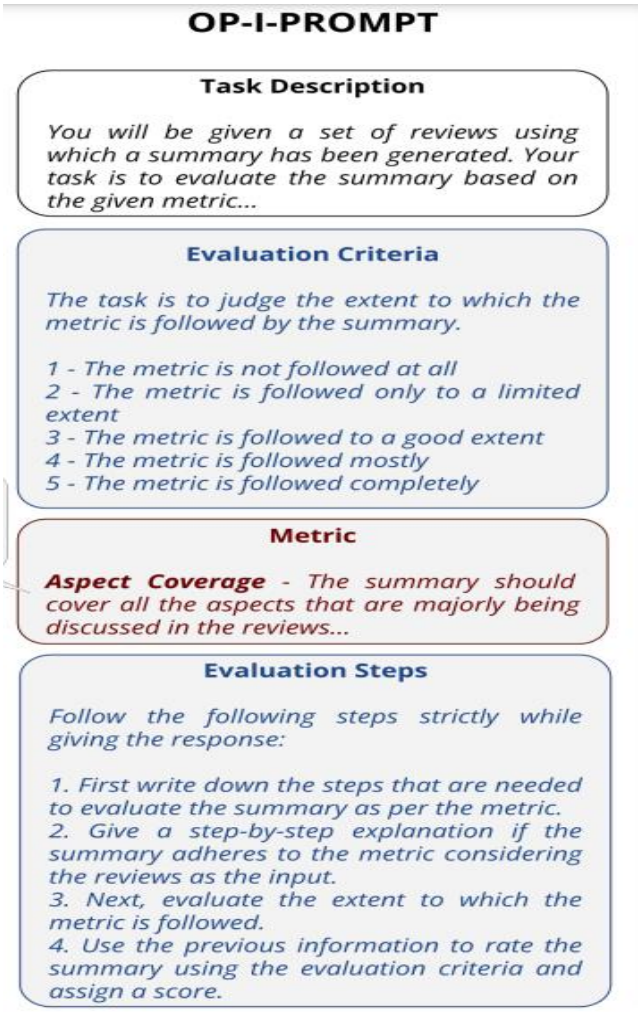
- No need of Chain of Thought (CoT) prompting.
- But for every metric we need to change the Evaluation Criteria prompt



One Prompt To Rule Them All: LLMs for Opinion Summary Evaluation

Approach 2: OP-I-PROMPTS: Dimension independent set of prompts

- Free from CoT prompting
- No need to change Evaluation criteria every time



SUMMEVAL-OP Benchmark Dataset

We created the SUMMEVAL-OP benchmark dataset for evaluating the opinion summaries on 7 dimensions.

#Products	32
#reviews/product	8
#summaries/ product	13 (from 13 different models including one human summary)
#total summaries in the dataset	13x32 = 416
#dimensions/summary	7 (Fluency, Coherence, Relevance, Aspect coverage, Faithfulness, sentiment consistency, Specificity)
#total dimensions in the dataset	7x416 = 2912

Human Annotations on SummEval-OP

3 Masters' students were hired to rate 7 dimensions of 13 summaries for each product in SummEval-OP dataset on a scale of 1-5. Evaluation is done in 2 rounds.

Round-1: Annotators asked to rate the summaries according to eval criteria.

Round-2: ratings where the scores of any rater differed from any other rater by 2 or more points were re-evaluated.

	Round-I ↑	Round-II ↑
fluency	0.55	0.84
coherence	0.43	0.73
relevance	0.50	0.79
faithfulness	0.63	0.86
aspect coverage	0.64	0.82
sentiment consistency	0.41	0.78
specificity	0.34	0.76
AVG	0.50	0.80

Table 1: Krippendorff's alpha coefficient (α) for Round-I and Round-II on 7 dimensions. As expected, we see an improvement in Round-II coefficient scores.

Inter-Rater Reliability

Results on Different Metrics

		FL ↑		CO ↑		RE ↑		FA ↑		AC ↑		SC ↑		SP ↑	
		ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ
SUMMEVAL-OP (Ours)	HUMANS	0.80	0.77	0.81	0.76	0.91	0.86	0.89	0.85	0.93	0.87	0.91	0.85	0.92	0.87
	ROUGE-1	-0.36	-0.28	-0.30	-0.24	-0.31	-0.23	-0.35	-0.26	-0.44	-0.32	-0.38	-0.29	-0.30	-0.23
	ROUGE-2	-0.23	-0.18	-0.14	-0.10	-0.17	-0.12	-0.21	-0.16	-0.26	-0.19	-0.24	-0.18	-0.14	-0.09
	ROUGE-L	-0.39	-0.32	-0.30	-0.23	-0.34	-0.25	-0.40	-0.30	-0.51	-0.37	-0.45	-0.33	-0.38	-0.27
	BERTSCORE	-0.32	-0.27	-0.28	-0.22	-0.29	-0.22	-0.34	-0.26	-0.51	-0.43	-0.41	-0.33	-0.37	-0.28
	BARTSCORE	-0.19	-0.15	-0.19	-0.14	-0.29	-0.22	-0.33	-0.25	-0.45	-0.35	-0.37	-0.28	-0.36	-0.27
	SUMMAC	0.23	0.20	0.18	0.14	0.30	0.25	0.25	0.21	0.24	0.19	0.25	0.20	0.26	0.21
	UNIEVAL	0.36	0.28	0.52	0.42	0.33	0.25	0.17	0.14	-	-	-	-	-	-
	G-EVAL-3.5	**0.63**	**0.55**	0.59	0.49	0.68	**0.56**	0.70	0.58	0.79	0.67	**0.73**	**0.61**	**0.75**	**0.63**
	OP-I-GPT-3.5	0.60	0.51	**0.61**	**0.51**	**0.69**	**0.56**	**0.71**	**0.59**	0.80	0.68	**0.73**	**0.61**	0.74	0.61
	G-EVAL-MISTRAL	0.50	0.43	0.54	0.45	0.52	0.42	0.54	0.44	0.61	0.49	0.55	0.46	0.62	0.50
	OP-MISTRAL	0.38	0.32	0.58	0.47	0.56	0.45	0.57	0.46	0.80	0.67	0.60	0.49	0.75	0.62
	OP-I-MISTRAL	0.54	0.45	0.58	0.47	0.59	0.47	0.63*	0.51*	**0.82***	**0.70***	**0.73***	**0.61***	0.71*	0.58*

Table 2: *Spearman* (ρ) and *Kendall Tau* (τ) correlations at summary-level on 7 dimensions for the SUMMEVAL-OP dataset. For closed-source, OP-I-PROMPT performs comparably to G-EVAL, whereas for open-source it outperforms alternatives. $*$ represents significant performance (p-value < 0.05) to G-EVAL-MISTRAL computed using Mann-Whitney U Test. HUMANS- averaged correlation of each annotator with the overall averaged ratings.

Scores of Various Models on Different