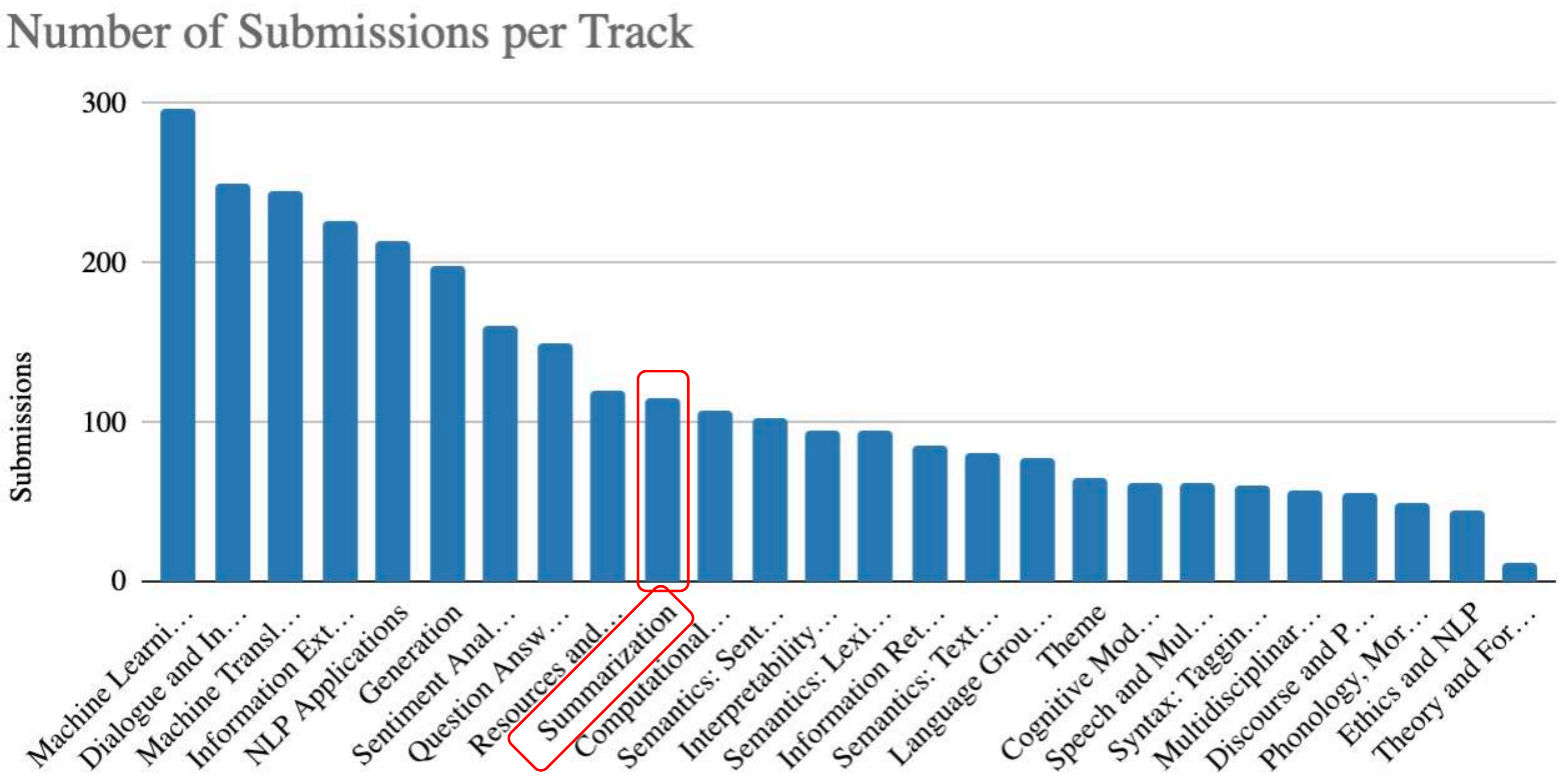
ACL2020 Summarization

Xiachong Feng

2020-6-16

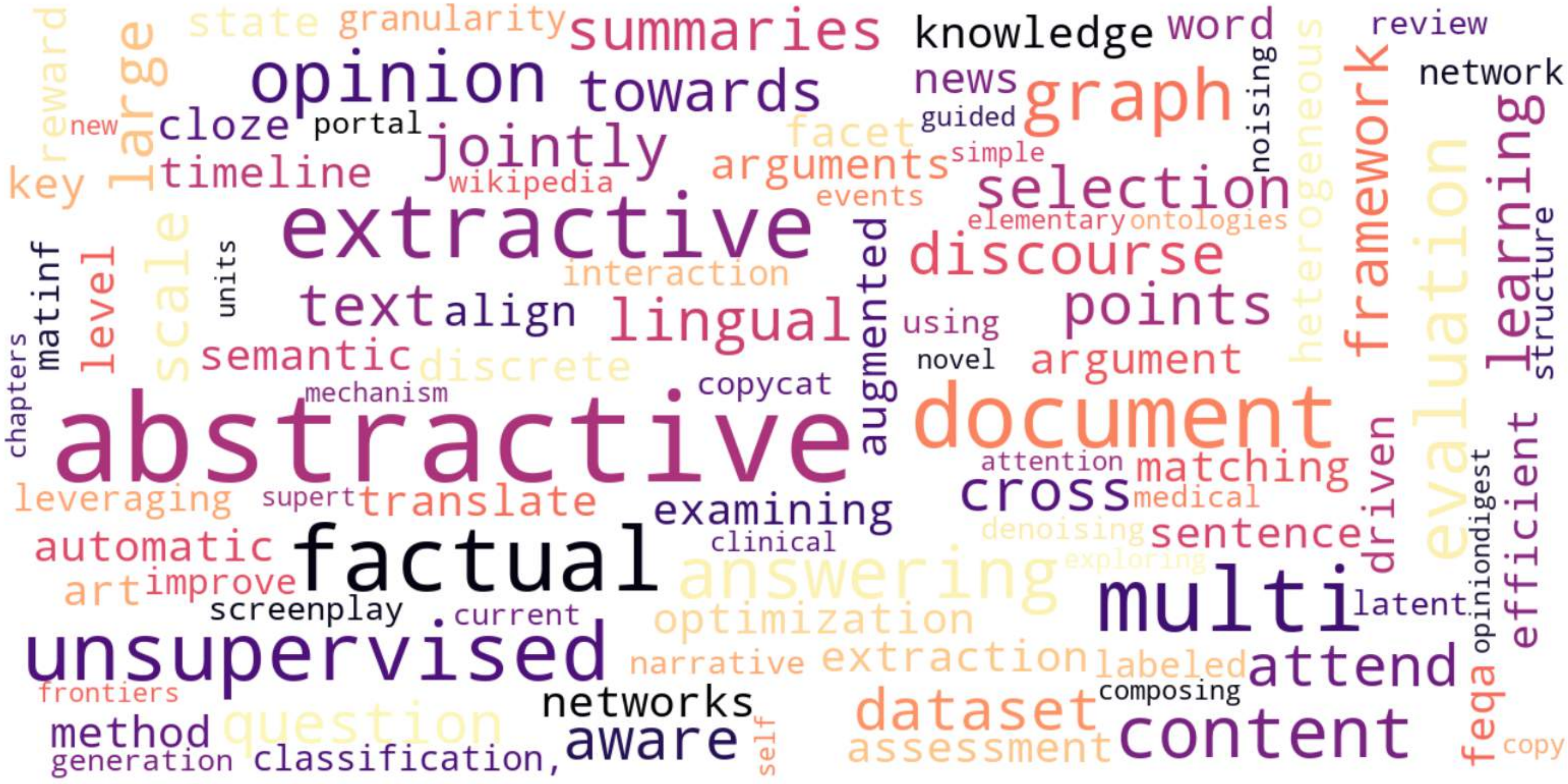
Overview



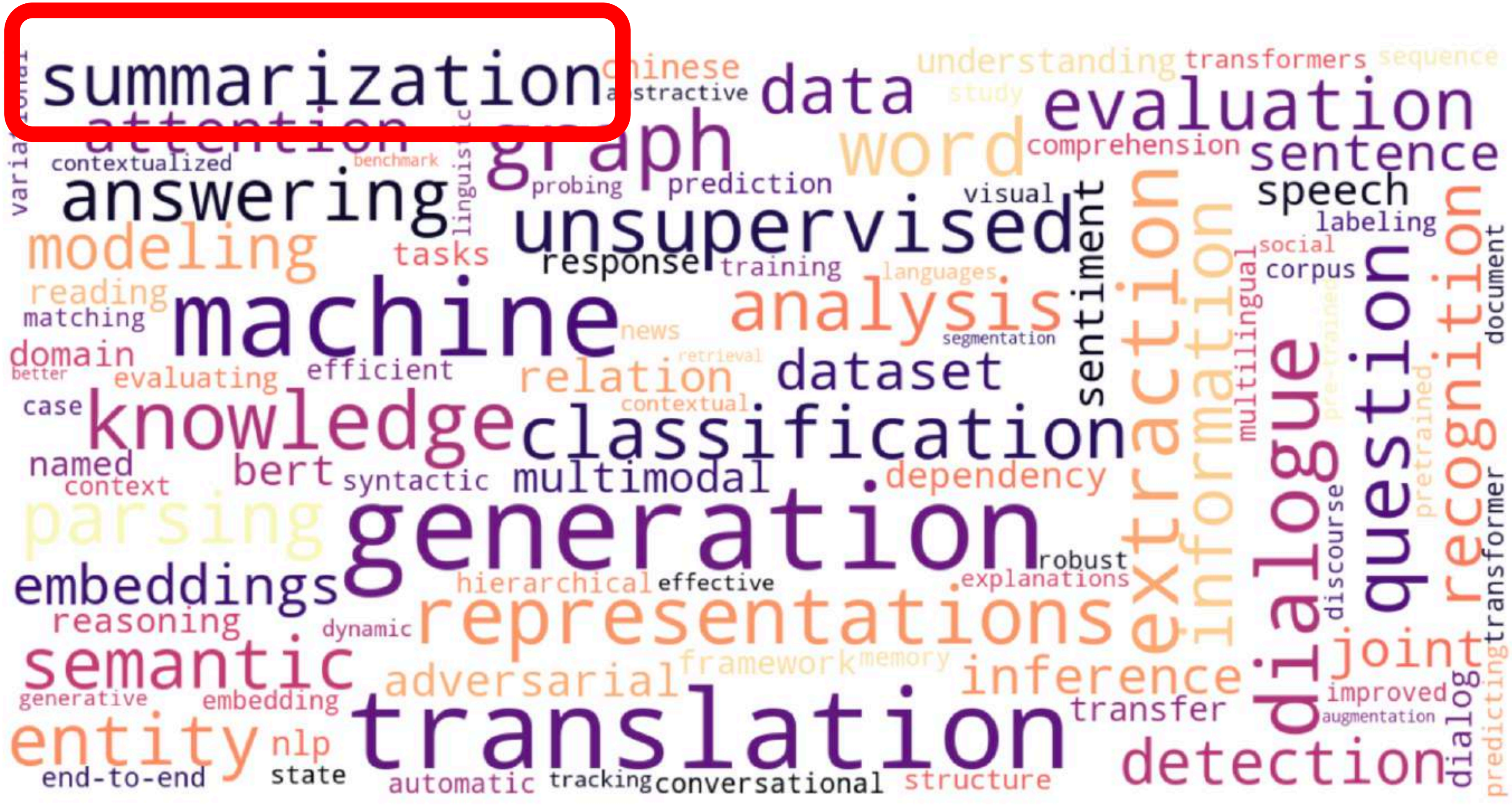
Track	Submissions	Accepted	% Accepted
Summarization	115	30	26.1

https://acl2020.org/blog/general-conference-statistics/

Overview-ACL20



Overview-All



https://github.com/roomylee/ACL-2020-Papers

Topics

- Factuality (6)
- Graph-based Methods (2)
- Opinion Summarization (2)
- Dataset (2)
- Others

Factuality

Factuality-Good Analysis (1)

- On Faithfulness and Factuality in Abstractive Summarization

Input Document

勒布朗·詹姆斯的外号是“小皇帝”，湖人队的科比·布莱恩特的外号是“黑曼巴”。

Summary 1：勒布朗·詹姆斯的外号是“黑曼巴”

Summary 2：湖人队的勒布朗·詹姆斯

Intrinsic hallucinations

Summary 3：科比·布莱恩特获得五次NBA总冠军

Summary 4：勒布朗·詹姆斯原先先效力于魔术队

Extrinsic hallucinations

Factual Hallucinations

Summary 5：勒布朗·詹姆斯的外号是“小皇帝”

Faithfulness Summary 5 Factuality Summary 2,3,5

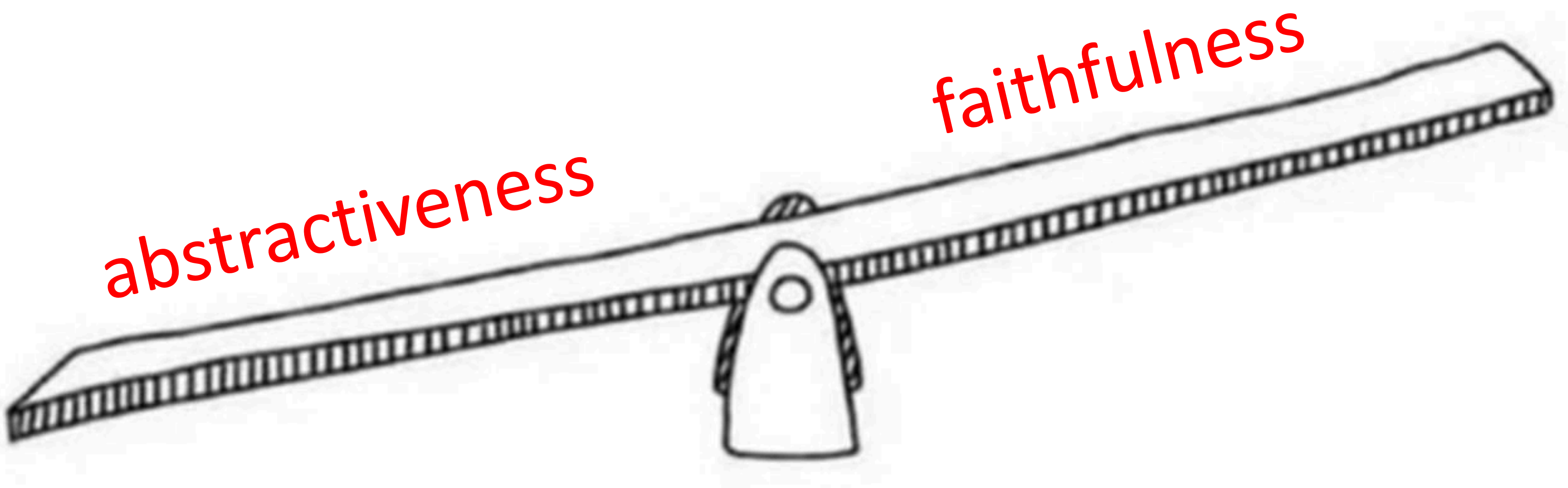
Factuality-Good Analysis (1)

Models	Hallucinated I	Hallucinated E	Hallucinated I ∪ E	Faith.	+Fact.
PtGen	19.9	**63.3**	75.3	24.7	27.3
TConvS2S	17.7	71.5	78.5	21.5	26.9
TranS2S	19.1	68.1	79.3	20.7	25.3
BertS2S	16.9	64.1	**73.1**	**26.9**	**34.7**
Gold	**7.4**	73.1	76.9	23.1	—

1. intrinsic and extrinsic hallucinations happen frequently
2. the majority of hallucinations are extrinsic, 90% of extrinsic hallucinations were erroneous.
3. models initialized with pretrained parameters perform best both on automatic metrics and human judgments of faithfulness/factuality. they have the highest percentage of extrinsic hallucinations that are factual

Factuality-Good Analysis (2)

- FEQA: A Question Answering Evaluation Framework for Faithfulness Assessment in Abstractive Summarization
- We find that current models exhibit a trade-off between abstractiveness and faithfulness: outputs with less word overlap with the source document are more likely to be unfaithful.



Factuality-Method (1)

- Improving Truthfulness of Headline Generation

- Focus: headline generation
- Drawbacks of dataset:
  - They assumed the lead (first) sentence of an article as a source document and its corresponding headline as a target output.
- Reason:
  - untruthful supervision data used for training the model.

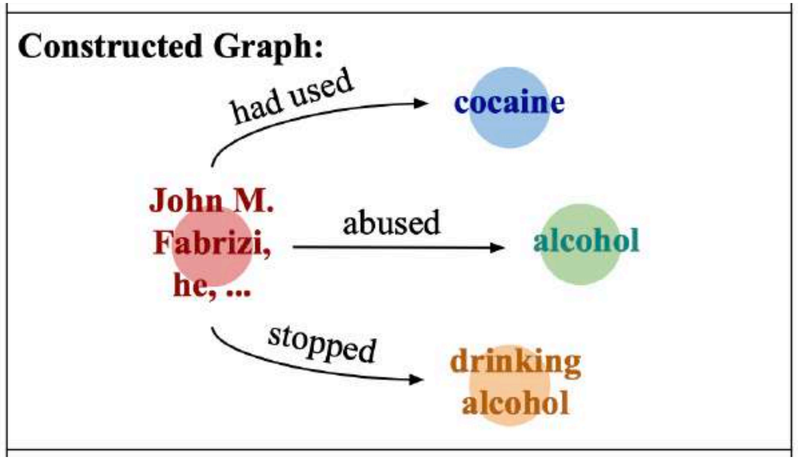
Factuality-Method (1)

- Mehod
  1. Human annotate each doc-summary pair a entailment label
  2. Use RoBERTa to train a entailment model
  3. For each instance in the dataset
  4. Filter out non-entail instances
  5. Use clean data with self-training to train the model
- Reslut
  - headline generation model trained on filtered supervision data shows no clear difference in ROUGE scores but remarkable improvements in automatic and manual evaluations of the generated headlines.

Factuality-Method (2)

- **Knowledge Graph-Augmented Abstractive Summarization with Semantic-Driven Cloze Reward**
- Question : unfaithful content
- Method :

Input Article of New York Times:
John M. Fabrizi, the mayor of Bridgeport, admitted on Tuesday that he had used cocaine and abused alcohol while in office.
Mr. Fabrizi, who was appointed mayor in 2003 after the former mayor, Joseph P. Ganim, went to prison on corruption charges, said he had sought help for his drug problem about 18 months ago and that he had not used drugs since.
About four months ago, he added, he stopped drinking alcohol.



Summary by Human:
The Week column. Mayor John Fabrizi of Brigeport, Conn, publicly admits he used cocaine and abused alcohol while in office; says he stopped drinking alcohol and sought help for his drug problem about 18 months ago.

coreference resolution && open information extraction

Factuality-Method (2)

- Model : ASGARD

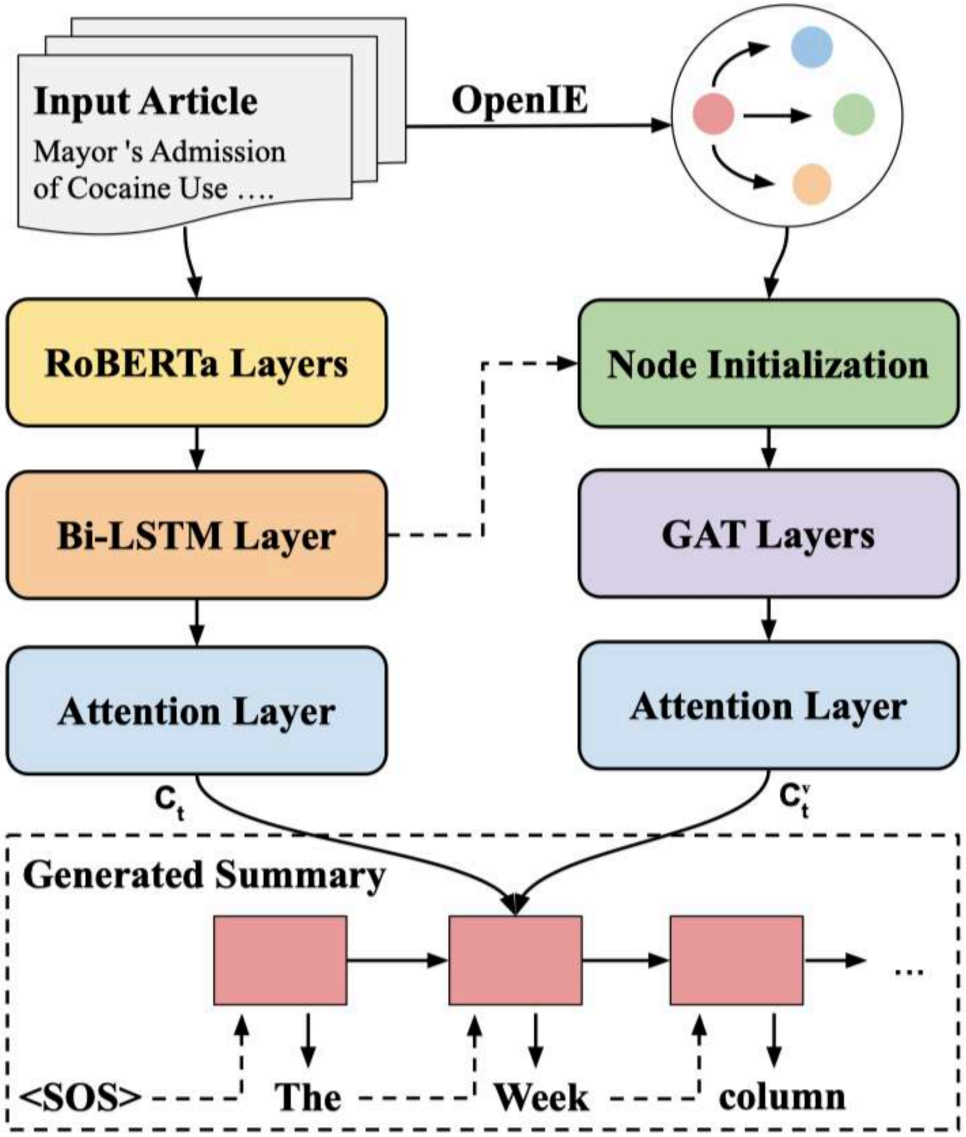
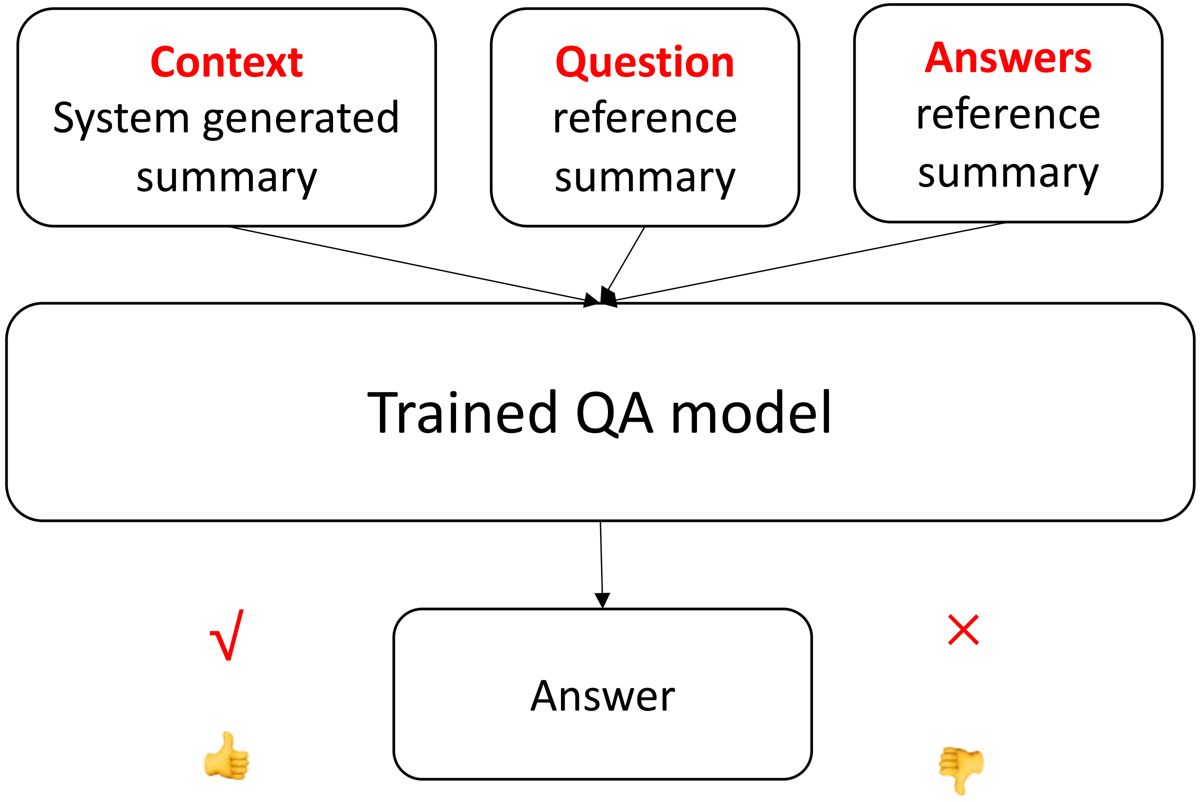


Figure 2: Our ASGARD framework with document-level graph encoding. Summary is generated by attending to both the graph and the input document.

Factuality-Method (2)

- Self-critical Policy Gradient
- Reward：
  - ROUGE
  - Multiple Choice Cloze Reward.



Factuality-Method (2)

- Salient Context:
  - greedy search to select the best combination of sentences that maximizes ROUGE2 F1 with reference to human summary.
  - further include a sentence in the salient context if it has a ROUGE-L recall greater than 0.6 when compared with any sentence in the reference.
- Question Construction.
  - argument pair questions
  - predicate questions
- QA model
  - RoBERTa
  - concatenate the <u>salient context</u>, the <u>question</u>, and each of the <u>four candidate answers</u>
  - Predict ：[CLS] representation

Reference Summary:
Federal Reserve *increases* **interest rates**.
IE Output:
⟨ **Federal Reserve**, *increases*, **interest rates** ⟩

Salient Context:
Federal Reserve *signals* positivity about the market. **Fed** increases benchmark interest rate again this May. **American economy** *keeps* the high growth rate. Jerome H. Powell *discussed* **potential risks**.
IE Outputs:
1. ⟨ **Federal Reserve**, *signals*, positivity ⟩
2. ⟨ **American economy**, *keeps*, the high growth rate ⟩
3. ⟨ Jerome H. Powell, *discussed*, **potential risks** ⟩

⇓

Multiple Choice Cloze Questions:
Argument Pair Question: ____ increases ____.
A. **Federal Reserve**, **interest rates** (✓)
B. **interest rates**, **Federal Reserve** (swapping args in A)
C. **American economy**, **interest rates** (replacing arg using triple 2)
D. **Federal Reserve**, **potential risks** (replacing arg using triple 3)

Predicate Question: Federal Reserve ____ interest rates.
A. increases (✓) B. signals C. keeps D. discussed

Factuality-Method (3)

Optimizing the Factual Correctness of a Summary: A Study of Summarizing Radiology Reports

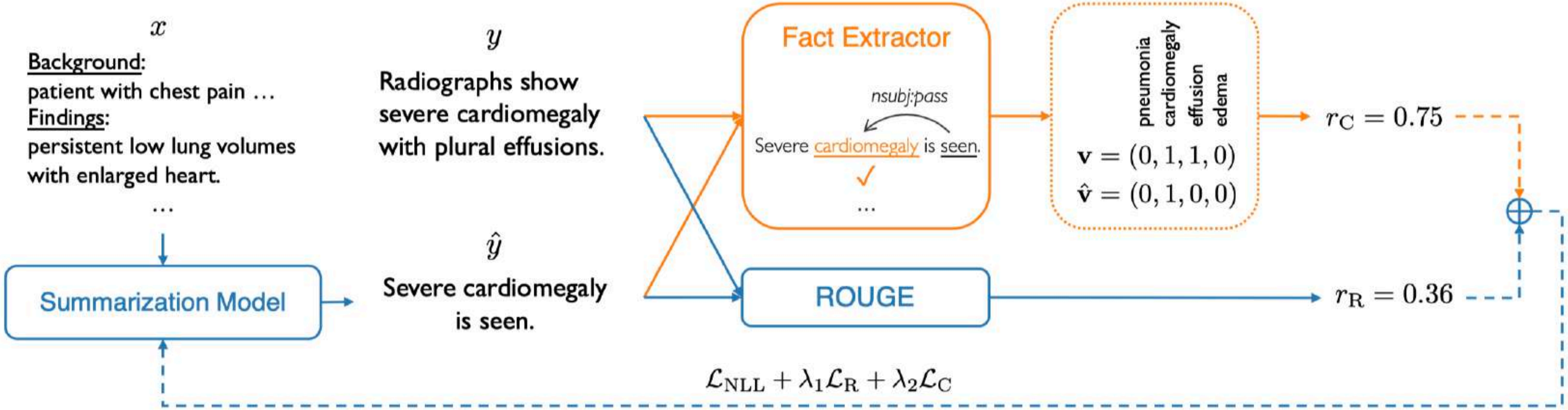
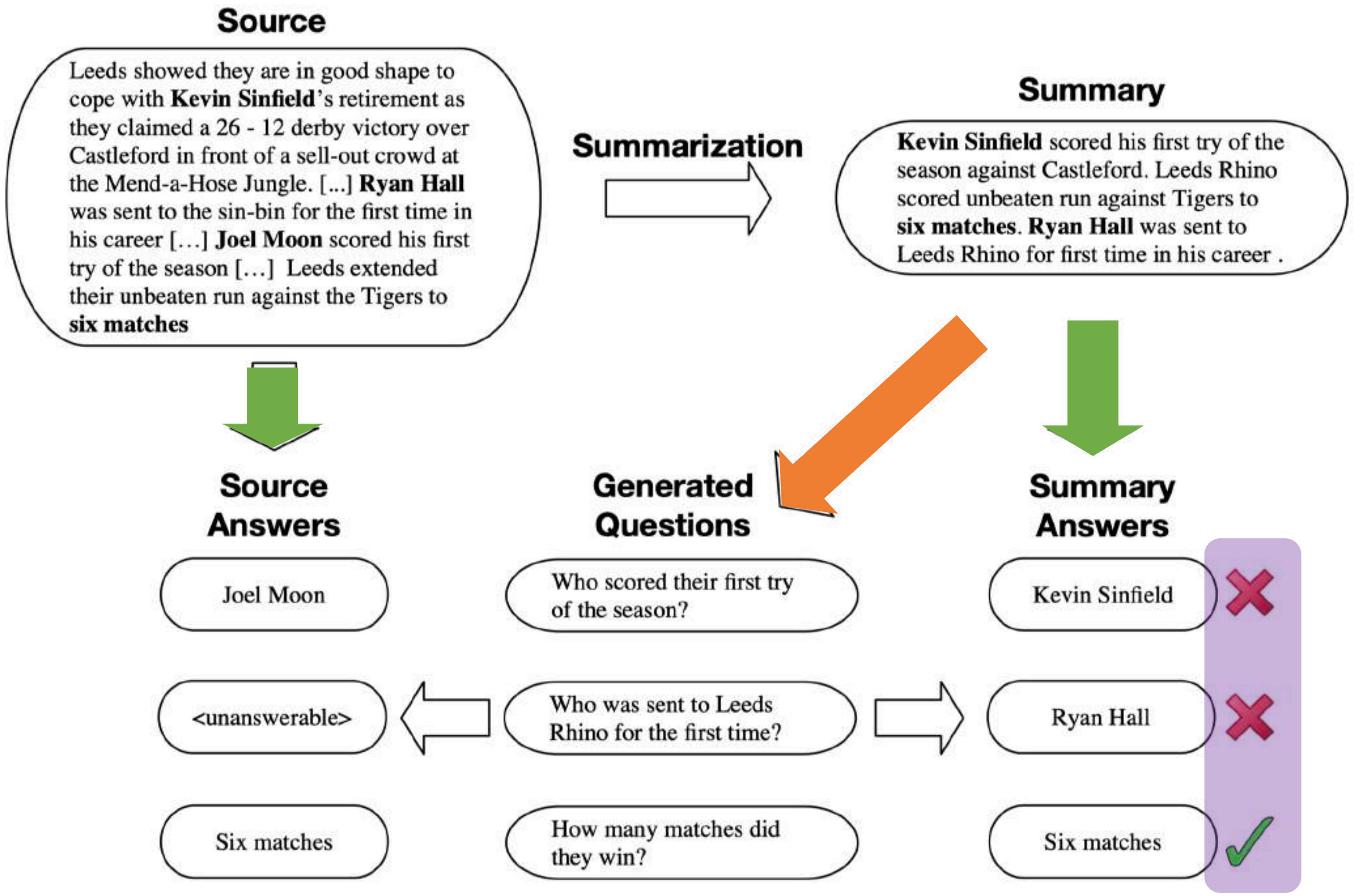


Figure 2: Our proposed training strategy. Compared to existing work which relies only on a ROUGE reward r_R, we add a factual correctness reward r_C which is enabled by a fact extractor. The summarization model is updated via RL, using a combination of the NLL loss, a ROUGE-based loss and a factual correctness-based loss. For simplicity we only show a subset of the clinical variables in the fact vectors $\mathbf{v}$ and $\hat{\mathbf{v}}$.

Factuality-Evaluation

Asking and Answering Questions to Evaluate the Factual Consistency of Summaries



- Answer conditional QG models, use named entities and noun phrases as answers candidates (BART, NewsQA)
- Extractive QA models (BERT, SQuAD2.0)
- Answer Similarity : token-level F1

Factuality Papers

Fact-based Content Weighting for Evaluating Abstractive Summarisation	ACL20
On Faithfulness and Factuality in Abstractive Summarization	ACL20
Improving Truthfulness of Headline Generation	ACL20
Knowledge Graph-Augmented Abstractive Summarization with Semantic-Driven Cloze Reward	ACL20
FEQA : A Question Answering Evaluation Framework for Faithfulness Assessment in Abstractive Summarization	ACL20
Optimizing the Factual Correctness of a Summary: A Study of Summarizing Radiology Reports	ACL20
Asking and Answering Questions to Evaluate the Factual Consistency of Summaries	ACL20

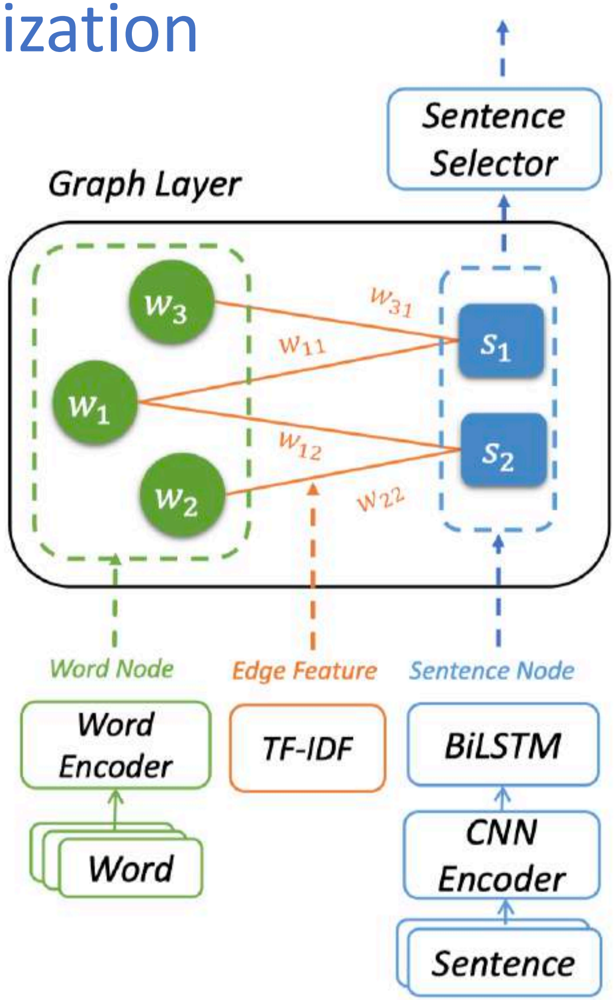
Boosting Factual Correctness of Abstractive Summarization with Knowledge Graph	
Ranking Generated Summaries by Correctness : An Interesting but Challenging Application for Natural Language Inference	ACL19
Evaluating the Factual Consistency of Abstractive Text Summarization	
Assessing The Factual Accuracy of Generated Text	KDD19
Faithful to the Original: Fact Aware Neural Abstractive Summarization	AAAI18
Ensure the Correctness of the Summary : Incorporate Entailment Knowledge into Abstractive Sentence Summarization	COLING18
Mind The Facts: Knowledge-Boosted Coherent Abstractive Text Summarization	

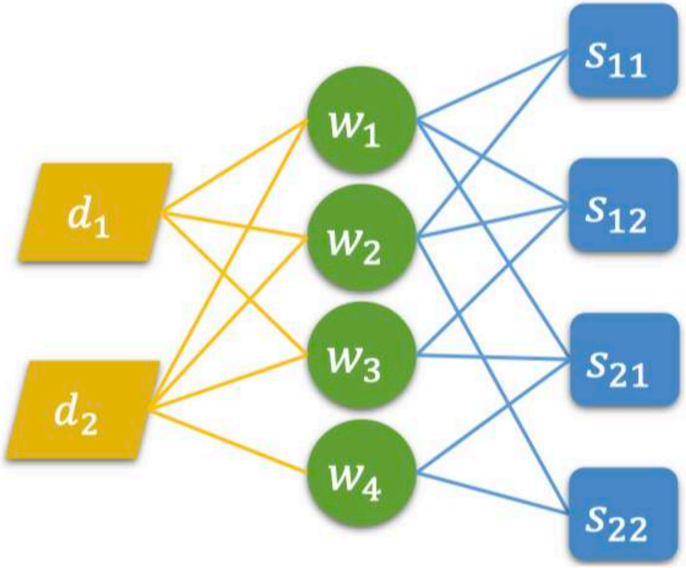
事实感知的生成式文本摘要 https://mp.weixin.qq.com/s/Aye9FBwG-v2JO2MLoEjo0g

Graph-Based

Heterogeneous Graph Neural Networks

- Heterogeneous Graph Neural Networks for Extractive Document Summarization





Multi-Document Summarization

- **Leveraging Graph to Improve Abstractive Multi-Document Summarization**

- Graph Construction
  - **Similarity graph :** tf-idf cosine similarities between paragraphs
  - **Topic graph :** topic relations between paragraphs. The edge weights are cosine similarities between the topic distributions of the paragraphs.
  - **Discourse graph :** discourse markers (e.g. however, moreover), co-reference and entity links

Multi-Document Summarization

Graph-informed Self-attention

$$\alpha_{ij} = softmax(e_{ij} + \boxed{\Re_{ij}})$$

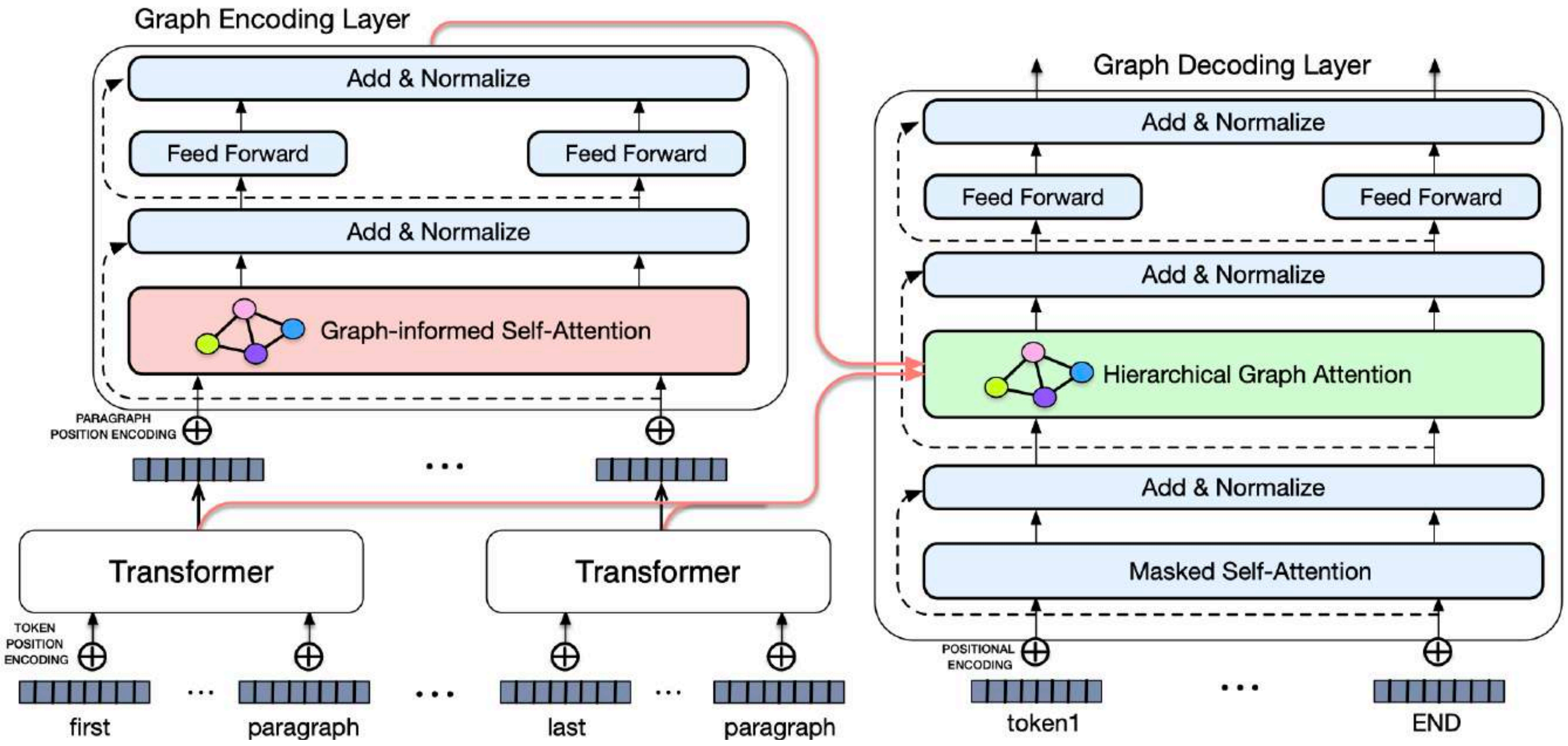
$$e_{ij} = \frac{(x_i^{l-1}W_Q)(x_j^{l-1}W_K)^T}{\sqrt{d_{head}}}$$

$$u_i = \sum_{j=1}^{L} \alpha_{ij}(x_j^{l-1}W_V)$$

$$\Re_{ij} = -\frac{(1-\mathbb{G}[i][j])^2}{2\sigma^2}$$

gaussian bias

$G[i][j]$ indicates the relation weights between paragraph P_i and P_j .



Opinion Summarization

Opinion Summarization

- Given
  - A set of reviews about a product (e.g., a movie or business).
- Output
  - Summary
- Challenge
  - Training data is not available and cannot be easily sourced

Summary	This restaurant is a hidden gem in Toronto. The food is delicious, and the service is impeccable. Highly recommend for anyone who likes French bistro.
Reviews	We got the steak frites and the chicken frites both of which were very good ... Great service ... || I really love this place ... Côte de Boeuf ... A Jewel in the big city ... || French jewel of Spadina and Adelaide , Jules ... They are super accommodating ... moules and frites are delicious ... || Food came with tons of greens and fries along with my main course , thumbs uppp ... || Chef has a very cool and fun attitude ... || Great little French Bistro spot ... Go if you want French bistro food classics ... || Great place ... the steak frites and it was amazing ... Best Steak Frites ... in Downtown Toronto ... || Favourite french spot in the city ... crème brule for dessert

Opinion Summarization (1)

- **OPINIONDIGEST: A Simple Framework for Opinion Summarization *ACL Short***

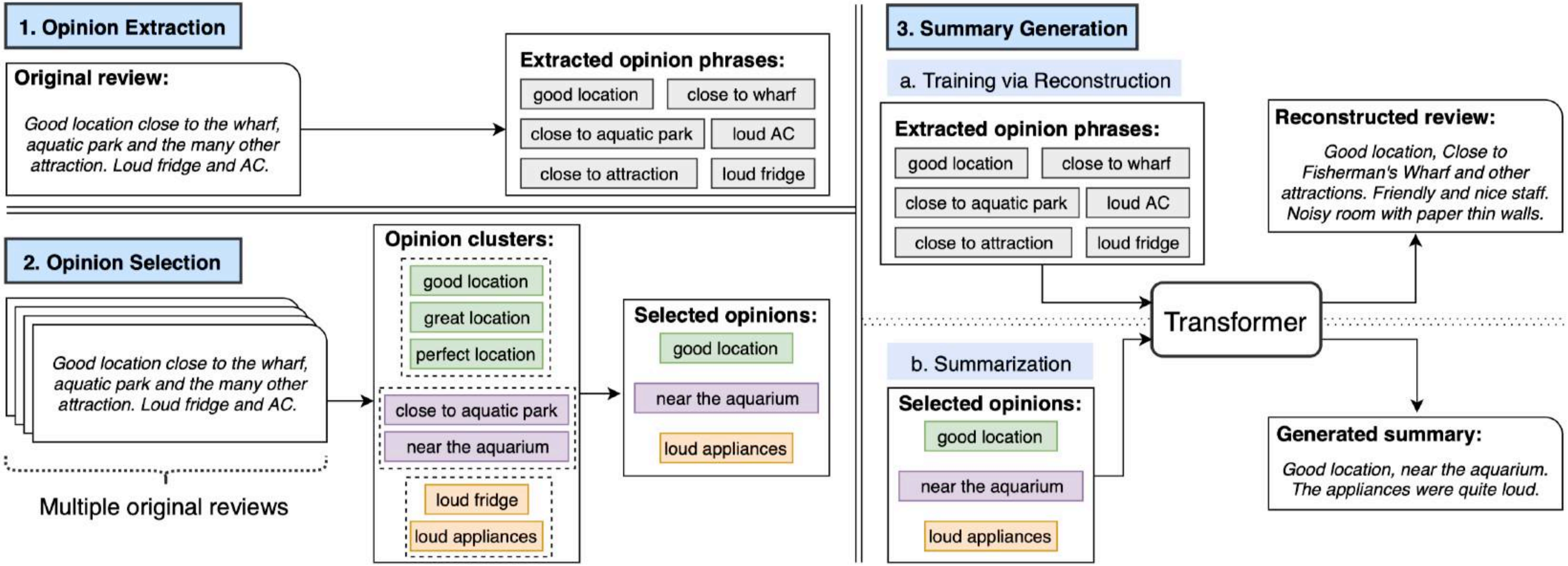


Figure 1: Overview of the OPINIONDIGEST framework.

Opinion Summarization (2)

- **Unsupervised Opinion Summarization with Noising and Denoising**

- Motivation: denoising can be seen as removing diverging information.

- Method:
  - Sample a review
  - Noising
  - Denoising

Opinion Summarization (2)

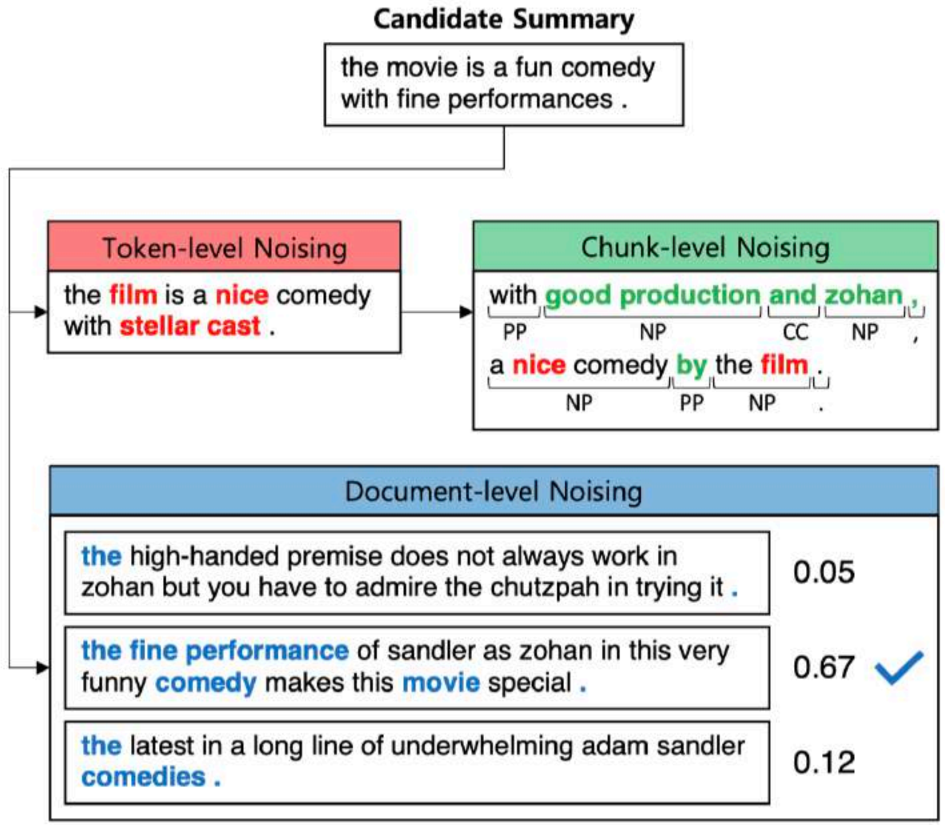


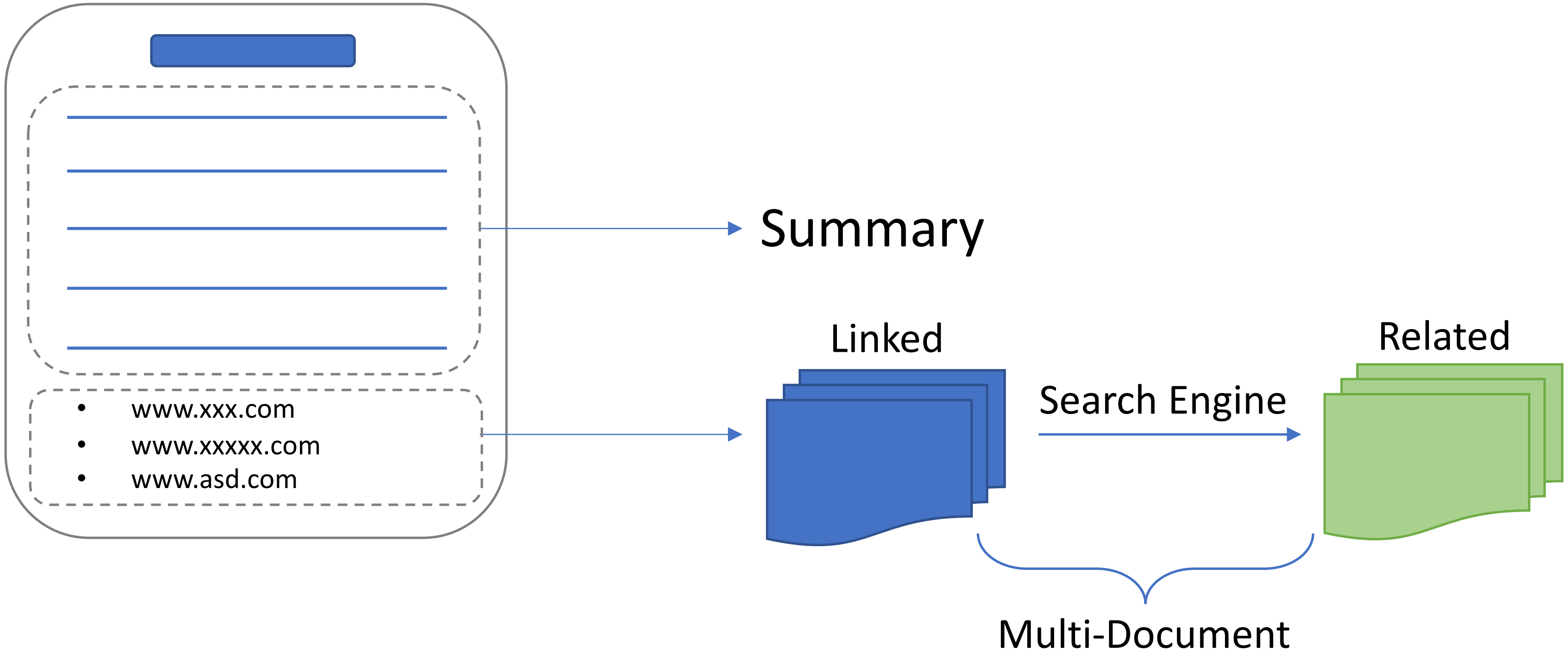
Figure 1: Synthetic dataset creation. Given a sampled candidate summary, we add noise using two methods: (a) segment noising performs token- and chunk-level alterations, and (b) document noising replaces the text with a semantically similar review.

- **Segment Noising**
  - Token-level
    - replace words
  - Chunk level
    - parse chunks for current review a
    - choose another review, parse chunks b
    - use b as template
    - generate a noise version of a
- **Document Noising**
  - choose N similar reviews

Dataset

Dataset

ID	Paper	Desp	Highlight
1	A Large-Scale Multi-Document Summarization Dataset from the Wikipedia Current Events Portal	多文档 新闻领域	10,200 clusters with one human-written summary and 235 articles per cluster on average.



Dataset

ID	Paper	Desp	Highlight
2	MATINF: A Jointly Labeled Large-Scale Dataset for Classification, Question Answering and Summarization	文摘、问答、分类	Multi-task

婴幼保健 Infant health care	*Class*
宝宝为什么总是吐舌头啊? Why does my baby always stick his tongue out ?	*Question*
我家宝宝出生快满四个月了，这几天我忽然发现宝宝总是吐舌头，而且口水也很多，那么这到底是咋回事啊? My baby is almost four months old. In these few days, I suddenly found that my baby always stick his tongue out and has a lot of saliva. So what is this?	*Description*
正常，不要担心的，小孩子都这个样子。宝宝吐舌头也是很正常的现象，你也不用过于担心，宝宝流口水可能是要长牙齿了。 Don't worry, it's normal. Kids are like this. It is also normal for your baby to stick his tongue out. You don't have to worry too much. Your baby's drooling may be a sign of teeth growing.	*Answer*

Figure 1: An example entry from MATINF.

Question + Description → Class

Text Classification

Question → Answer

Question Answering

Description → Question

Summarization

Others

Others

1. Extractive Summarization as Text Matching
2. Discourse-Aware Neural Extractive Text Summarization
3. Exploring Content Selection in Summarization of Novel Chapters
4. Examining the State-of-the-Art in News Timeline Summarization
5. From Arguments to Key Points: Towards Automatic Argument Summarization
6. Storytelling with Dialogue: A Critical Role Dungeons and Dragons Dataset
7. Facet-Aware Evaluation for Extractive Summarization
8. SUPERT- Towards New Frontiers in Unsupervised Evaluation Metrics for Multi-Document Summarization

1. Jointly Learning to Align and Summarize for Neural Cross-Lingual Summarization
2. Attend, Translate and Summarize: An Efficient Method for Neural Cross-Lingual Summarization
3. Multi-Granularity Interaction Network for Extractive and Abstractive Multi-Document Summarization
4. The Summary Loop: Learning to Write Abstractive Summaries Without Examples
5. Fact-based Content Weighting for Evaluating Abstractive Summarisation
6. Self-Attention Guided Copy Mechanism for Abstractive Summarization
7. Understanding Points of Correspondence between Sentences for Abstractive Summarization

Conclusion

1. 🔥🔥🔥 HOT : Factuality
2. Abstractive Papers > Extractive Papers
3. Cross-Lingual Summarization
4. Graph Neural Networks
5. Dataset Papers (Compared with ACL19)
6. Unsupervised Methods (Opinion Summarization)

Thanks!