ACL2020 Summarization

Xiachong Feng
2020-6-16
Overview

Number of Submissions per Track

<table>
<thead>
<tr>
<th>Track</th>
<th>Submissions</th>
<th>Accepted</th>
<th>% Accepted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Summarization</td>
<td>115</td>
<td>30</td>
<td>26.1</td>
</tr>
</tbody>
</table>

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” Intrinsic
Summary 2: 湖人队的勒布朗·詹姆斯 hallucinations
Summary 3: 科比·布莱恩特获得五次NBA总冠军 Factual Hallucinations
Summary 4: 勒布朗·詹姆斯原先先效力于魔术队 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)**

<table>
<thead>
<tr>
<th>Models</th>
<th>Hallucinated</th>
<th>Faith.</th>
<th>+Fact.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>I</td>
<td>E</td>
<td>I ∪ E</td>
</tr>
<tr>
<td><strong>PtGEN</strong></td>
<td>19.9</td>
<td>63.3</td>
<td>75.3</td>
</tr>
<tr>
<td><strong>TConvS2S</strong></td>
<td>17.7</td>
<td>71.5</td>
<td>78.5</td>
</tr>
<tr>
<td><strong>TRANS2S</strong></td>
<td>19.1</td>
<td>68.1</td>
<td>79.3</td>
</tr>
<tr>
<td><strong>BERTS2S</strong></td>
<td>16.9</td>
<td>64.1</td>
<td>73.1</td>
</tr>
<tr>
<td><strong>GOLD</strong></td>
<td>7.4</td>
<td>73.1</td>
<td>76.9</td>
</tr>
</tbody>
</table>

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)

• Method
  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

• Result
  • 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.

**constructed Graph:**

```
 John M. Fabrizi, he, ...

 had used  cocaine

 abused  alcohol

 stopped  drinking alcohol
```

**Summary by Human:**
The Week column. Mayor John Fabrizi of Bridgeport, 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.

![Diagram showing the process of a trained QA model with context, question, and answers.]
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 salient context, the question, and each of the four candidate answers
  • Predict : [CLS] representation
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

<table>
<thead>
<tr>
<th>Title</th>
<th>Conference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fact-based Content Weighting for Evaluating Abstractive Summarisation</td>
<td>ACL20</td>
</tr>
<tr>
<td>On Faithfulness and Factuality in Abstractive Summarization</td>
<td>ACL20</td>
</tr>
<tr>
<td>Improving Truthfulness of Headline Generation</td>
<td>ACL20</td>
</tr>
<tr>
<td>Knowledge Graph-Augmented Abstractive Summarization with Semantic-Driven Cloze Reward</td>
<td>ACL20</td>
</tr>
<tr>
<td>FEQA: A Question Answering Evaluation Framework for Faithfulness Assessment in Abstractive Summarization</td>
<td>ACL20</td>
</tr>
<tr>
<td>Optimizing the Factual Correctness of a Summary: A Study of Summarizing Radiology Reports</td>
<td>ACL20</td>
</tr>
<tr>
<td>Asking and Answering Questions to Evaluate the Factual Consistency of Summaries</td>
<td>ACL20</td>
</tr>
<tr>
<td>Boosting Factual Correctness of Abstractive Summarization with Knowledge Graph</td>
<td>ACL19</td>
</tr>
<tr>
<td>Ranking Generated Summaries by Correctness: An Interesting but Challenging Application for Natural Language Inference</td>
<td>KDD19</td>
</tr>
<tr>
<td>Evaluating the Factual Consistency of Abstractive Text Summarization</td>
<td>AAAI18</td>
</tr>
<tr>
<td>Assessing The Factual Accuracy of Generated Text</td>
<td>COLING18</td>
</tr>
<tr>
<td>Faithful to the Original: Fact Aware Neural Abstractive Summarization</td>
<td>COLING18</td>
</tr>
<tr>
<td>Ensure the Correctness of the Summary: Incorporate Entailment Knowledge into Abstractive Sentence Summarization</td>
<td>COLING18</td>
</tr>
<tr>
<td>Mind The Facts: Knowledge-Boosted Coherent Abstractive Text Summarization</td>
<td>COLING18</td>
</tr>
</tbody>
</table>

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

\[ \alpha_{ij} = \text{softmax} (e_{ij} + \mathcal{R}_{ij}) \]
\[ e_{ij} = \frac{(x_i^{l-1}W_Q)(x_j^{l-1}W_K)^T}{\sqrt{d_{\text{head}}}} \]
\[ u_i = \sum_{j=1}^{L} \alpha_{ij} (x_j^{l-1}W_V) \]
\[ \mathcal{R}_{ij} = -\frac{(1 - G[i][j])^2}{2\sigma^2} \]

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

<table>
<thead>
<tr>
<th>Reviews</th>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>We got the steak frites and the chicken frites both of which were very good ...</td>
<td>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.</td>
</tr>
<tr>
<td>Great service ...</td>
<td></td>
</tr>
<tr>
<td>Côte de Boeuf ... A Jewel in the big city ...</td>
<td></td>
</tr>
<tr>
<td>moules and frites are delicious ...</td>
<td></td>
</tr>
</tbody>
</table>
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)

- **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

![Diagram showing synthetic dataset creation](image)

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.
Dataset
## Dataset

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

- www.xxx.com
- www.xxxxx.com
- www.asd.com

![Diagram](Diagram.png)
### Dataset

<table>
<thead>
<tr>
<th>ID</th>
<th>Paper</th>
<th>Desp</th>
<th>Highlight</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>MATINF: A Jointly Labeled Large-Scale Dataset for Classification, Question Answering and Summarization</td>
<td>文摘、问答、分类</td>
<td>Multi-task</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Infant health care</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Why does my baby always stick his tongue out?</td>
<td>Question</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Infant health care</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>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，那么这到底是怎么回事呢？</td>
<td>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?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Infant health care</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>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.</td>
<td></td>
</tr>
</tbody>
</table>

Figure 1: An example entry from MATINF.
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
7. Facet-Aware Evaluation for Extractive Summarization

1. Jointly Learning to Align and Summarize for Neural Cross-Lingual 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!