文書要約タスクの評価指標の現状と課題

名古屋地区NLPセミナー
2023年7月3日

鴨井 遼

自己紹介

鴨井 遼

研究テーマ：言語生成モデルの評価、信頼性の向上など

経歴

- 慶應義塾大学理工学部数理科学科（統計学士）
  - カーネギーメロン大学　派遣交換留学
- テキサス大学オースティン校（CS修士）
  - Amazon - Applied Scientist Intern（2021年7月～12月）
- ペンシルベニア州立大学（博士課程　2023年8月～）

概要

- 文書要約タスクの評価手法（Factualityの評価）
  - Entailment-based metrics, QA-based metrics, LLM-based metrics
  - どの評価指標を研究するべきか？(Kamoi et al., EACL2023)
- 長い入力への対応
  - Entailment-based metrics (Kamoi et al., arXiv2023 など)
  - LLM-based metrics

Ryo Kamoi, Tanya Goyal, Greg Durrett. 2023. Shortcomings of Question Answering Based Factuality Frameworks for Error Localization. *EACL*.
Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, Greg Durrett. 2023. WiCE: Real-World Entailment for Claims in Wikipedia. *arXiv preprint arXiv:2303.01432*.

- アメリカの大学院への進学について

文書要約の評価基準

- Fluency
  - 要約に含まれる各文が文法的に正しい
- Coherence
  - 要約に含まれる文同士の関係に問題がない（文章全体として一貫性がある）
- Relevance
  - 元の文書の重要な部分が要約に含まれている
- Consistency = Factuality
  - 元の文書と矛盾する内容や、元の文書に含まれない情報が、要約に含まれない

最近のモデルは文法的に正しく一貫性のある文章は生成できる

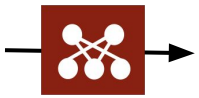
「重要な部分」の定義が難しい

→ **Factualityの評価が近年の研究の主な課題とされている**

文書要約におけるFactual Error

Original Document

[...] BBC Weather said 50mm of rain fell in Cambridgeshire in an hour, damaging the banks of the River Nene in March. [...] "The exact number of properties affected cannot be confirmed, but we understand that we are assisting currently at least 60 properties." [...]



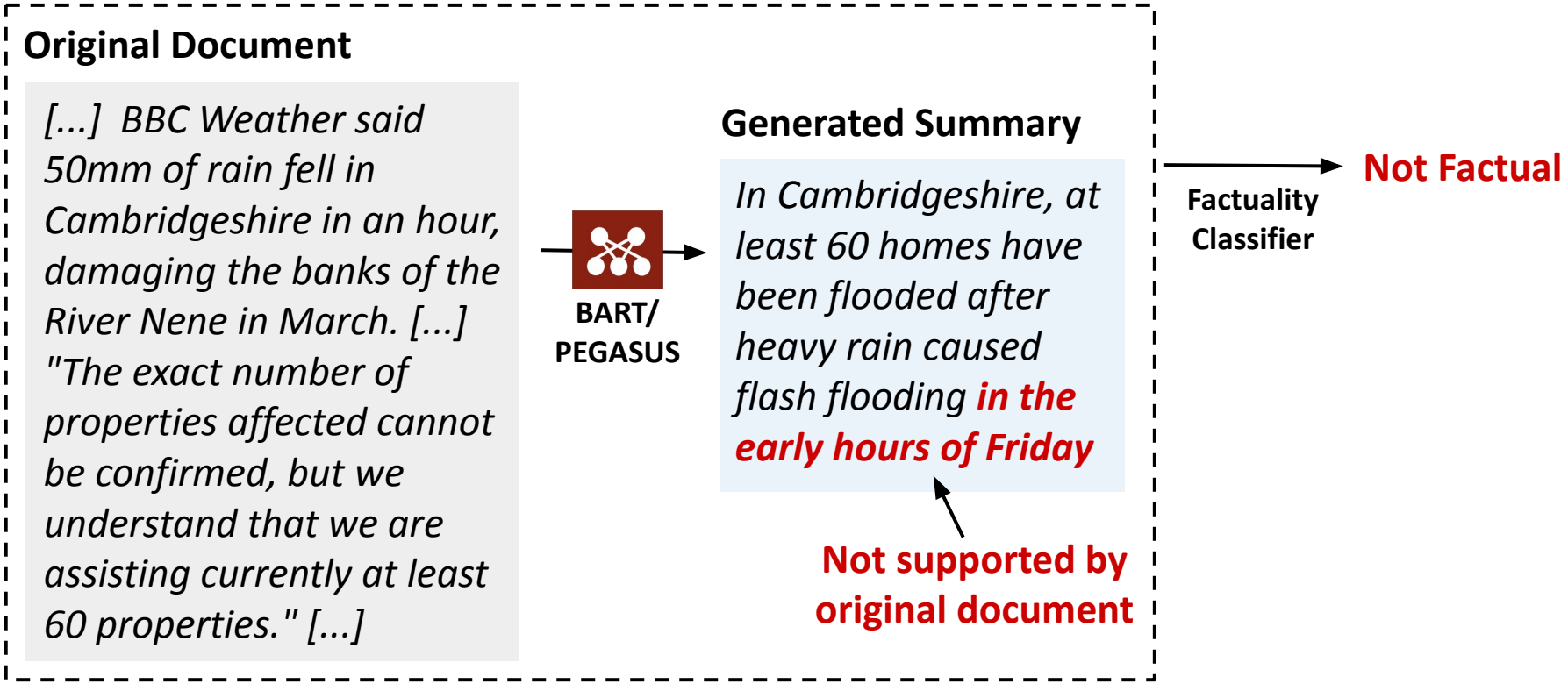
BART/ PEGASUS

Generated Summary

In Cambridgeshire, at least 60 homes have been flooded after heavy rain caused flash flooding ***in the early hours of Friday***

Not supported by original document

文書要約の評価（Factuality Classification）



文書要約の評価方法 (Factuality)

- Classic Scores: ROUGE, BERTScore
  - 性能が悪い
- Neural Metrics
  - Entailment-based metrics
  - QA-based metrics
  - LLM-based metrics
- Human Evaluation
  - 再現性が低い
  - コストが高い
  - 品質を上げることは容易ではない

文書要約の評価方法 (Factuality)

- Classic Scores: ROUGE, BERTSco
- Neural Metrics
  - Entailment-based metrics
  - QA-based metrics
  - LLM-based metrics
- Human Evaluation

	Model (↓)	Non-LLM Models				
✓	DAE	67.2				
✓	SummaC	96.8				
✓	QAFactEval	93.6				
	Prompt Group →	ZS	FS	Pers	CoT	GwE
✓	Dav001	61.2	56.8	52.9	61.6	58.1
✓	Dav002	74.5	81.3	57.9	78.5	73.2
✓	Dav003	82.3	78.4	62.4	85.5	76.8
✓	GPT3.5-turbo	84.3	82.9	75.1	84.0	86.3
✓	GPT4	91.3	90.1	66.3	85.7	78.0

Balanced Accuracy
on Synthetic FactCC (Laban et al., 2023)

Entailment-based Metrics

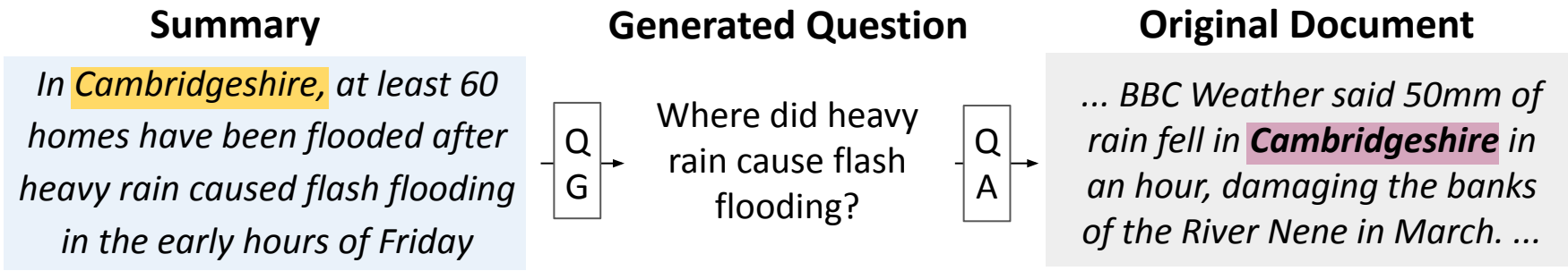
Natural Language Inference (自然言語推論): 前提と仮説の含意関係を推測するタスク

Premise	Label	Hypothesis
Letters		
Your gift is appreciated by each and every student who will benefit from your generosity.	*neutral*	Hundreds of students will benefit from your generosity.
Telephone Speech		
yes now you know if if everybody like in August when everybody's on vacation or something we can dress a little more casual or	*contradiction*	August is a black out month for vacations in the company.
9/11 Report		
At the other end of Pennsylvania Avenue, people began to line up for a White House tour.	*entailment*	People formed a line at the end of Pennsylvania Avenue.

(MNLI Dataset)

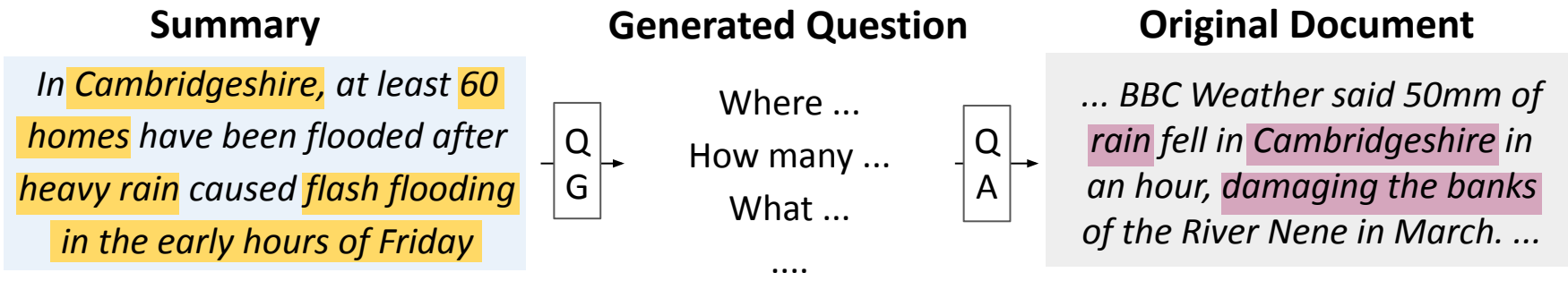
文書要約において、元の文書が要約文を含意していれば、要約文はFactualだと言えるのでは？ → Entailment-based Metrics

QA-based Metrics



score(span) = ***similarity***(*Cambridgeshire*, *Cambridgeshire*)

QA-based Metrics



$$\frac{1}{|S|} \sum_{\text{span} \in S} \text{score(span)} = \textbf{\textit{similarity}}(\quad , \quad)$$

スパンごとに質問を生成し、スパンごとのスコアを平均することでsummary-levelのFactuality scoreを得る

LLM-based metrics

- 2023年に入ってからChatGPTなどLLMを使った評価手法が数多く提案されている
- 現状では、それほど新しい技術を提案している研究はなく、
  Chain-of-Thought Promptなど既存の技術を評価に適用している

> Decide if the following summary is consistent with the corresponding article. Note that consistency means all information in the summary is supported by the article.
>
> Article: [Article]
>
> Summary: [Summary]
>
> Explain your reasoning step by step then answer (yes or no) the question:

(Luo et al., 2023)

結局、どれを使うべきなのか？

新しい文書要約手法を提案する場合など:
現状ではEntailment-based、QA-based、LLM-basedの指標のどれかについて明確に優位性が示されているわけではないので、複数の指標を示すのが無難

- Entailment-based: SummaC
- QA-based: QuestEval, QAFactEval
- LLM-based: 色々あるが、とりあえずZero-shot Chain-of-Thought?

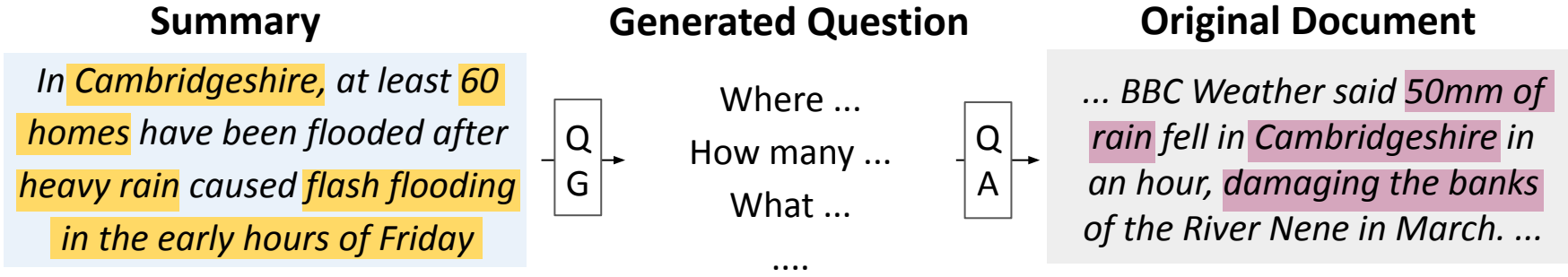
今後はどれを研究していくべきか？

QA-based metricsの研究はおすすめしない (Kamoi et al., EACL2023)

- 解決が難しい問題があり、性能改善のボトルネックとなっている
- Entailment-based metricsに対して優位性がないと考えられる

Ryo Kamoi, Tanya Goyal, Greg Durrett. 2023. Shortcomings of Question Answering Based Factuality Frameworks for Error Localization. *EACL*.

QA-based metricsの優位性とは？

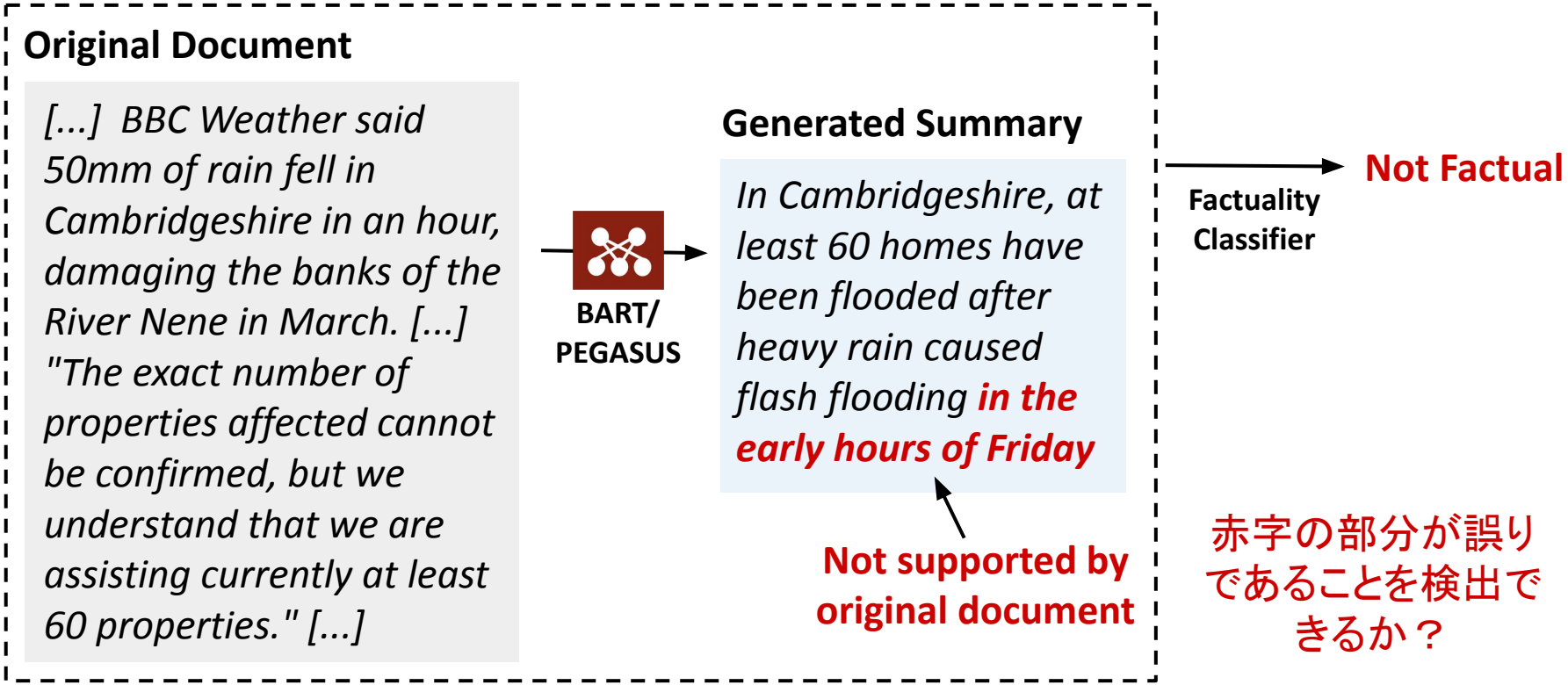


$$\frac{1}{|S|}\sum_{\text{span}\in S}\text{score(span)} = \textbf{\textit{similarity}}(\quad,\quad)$$

- 生成された問題と回答を見ることで、モデルの出力を解釈することができる
- スパンごとにスコアが出るので、fine-grainedな評価を得られる

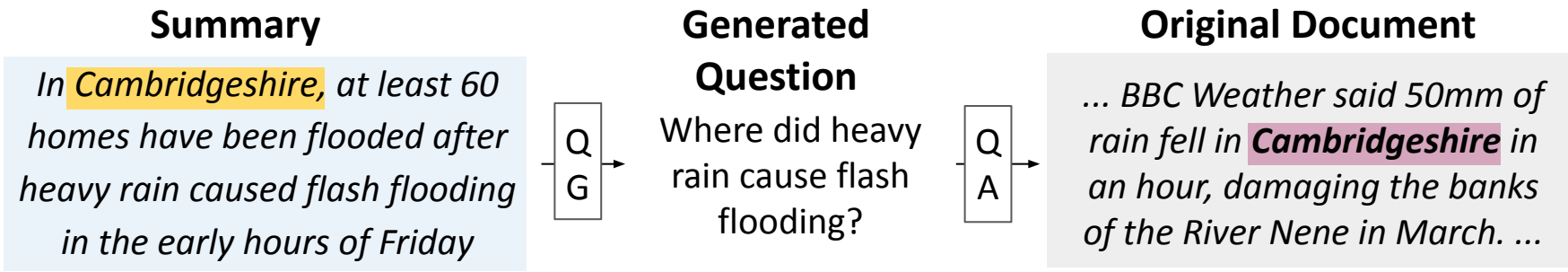
本当に正しい？

スパンレベルの評価



赤字の部分が誤りであることを検出できるか？

実験設定



score(span) = ***similarity***(*Cambridgeshire*, *Cambridgeshire*)

このスコアを用いてスパンごとの誤り
を検出できるかを評価する

実験設定

Task: Span-level factuality classification (binary)

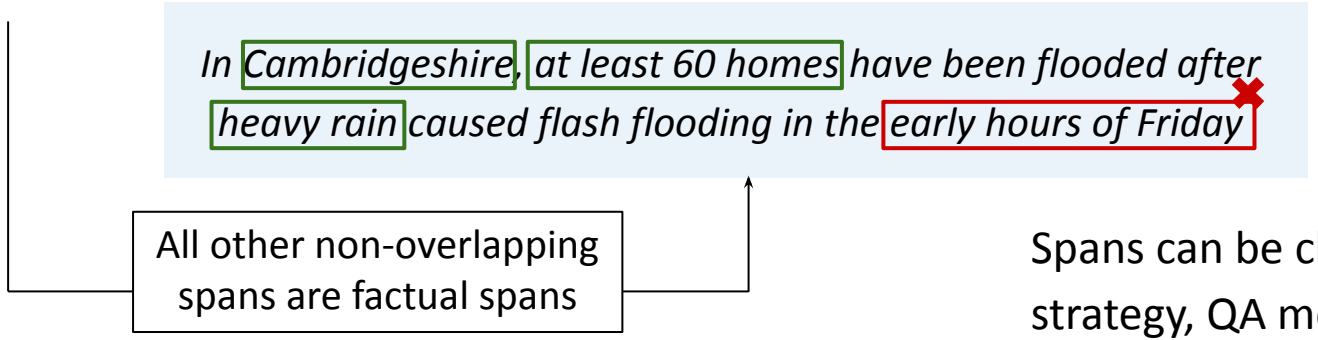
Available ground truth data: Human annotated non-factual spans.

In Cambridgeshire, at least 60 homes have been flooded after heavy rain caused flash flooding in the early hours of Friday

実験設定

Task: Span-level factuality classification (binary)

Available ground truth data: Human annotated non-factual spans.

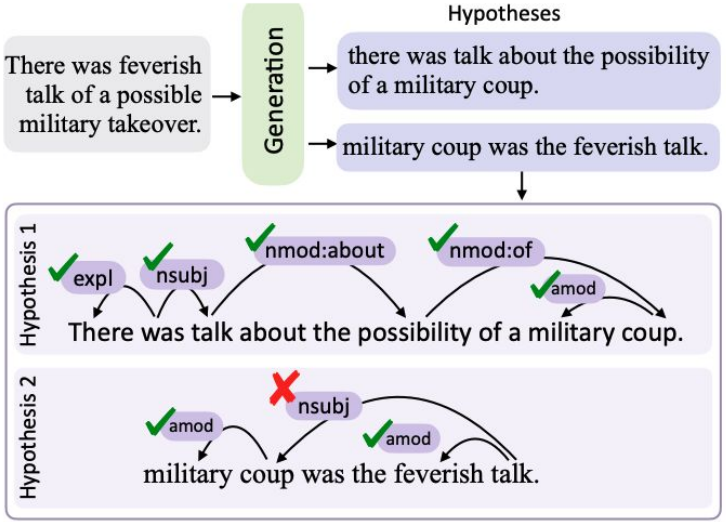


Spans can be chosen using any strategy, QA metrics typically focus on NPs + NEs.

Metric: F1

手法

- **Exact Match**
  Span is factual if 1-gram overlap with input is 1.
- **Dependency Arc Entailment (DAE)**
  (Goyal et al. 2020, 2021)
  Trained factuality model that outputs non-factual spans
  (projected from dependency arcs).



QA Metrics

- **QuestEval (Scialom et al., EMNLP 2021)**
- **QAFactEval (Fabbri et al., NAACL 2022)**

Differ in which QA, QG models are chosen, similarity function, etc.

Span-Levelの結果

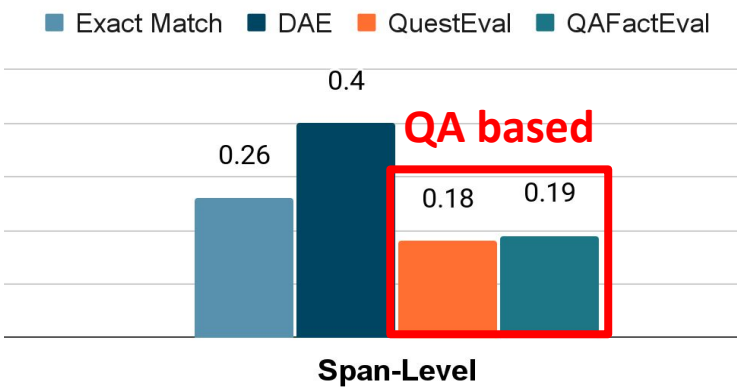
Baselines

- **Exact Match**
  Span is factual if 1-gram overlap with input is 1.
- **Dependency Arc Entailment (DAE)**
  Trained factuality model that outputs non-factual spans
  (projected from dependency arcs).

QA Metrics

- **QuestEval**
- **QAFactEval**

F1 score on GD21 (XSum) dataset

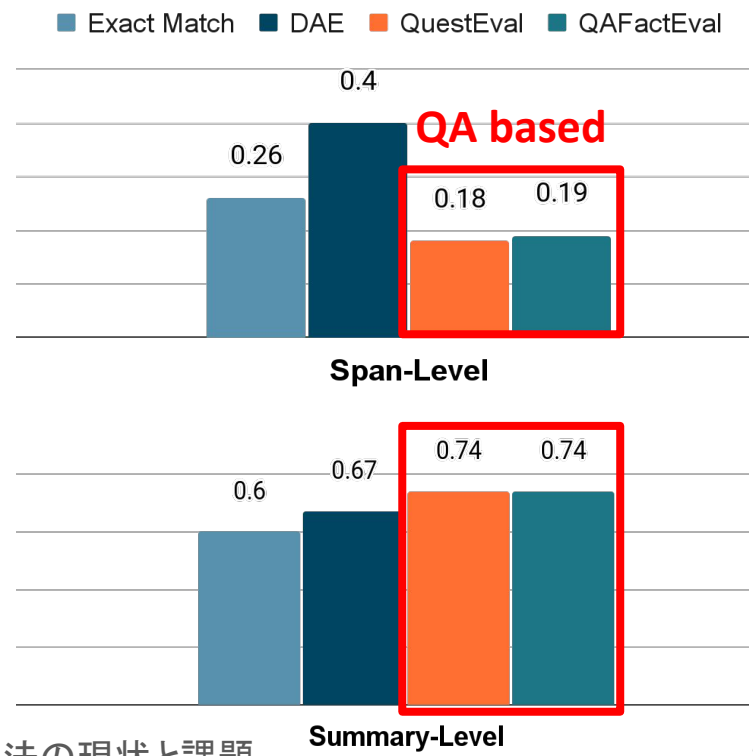


Span-levelでは単純なベースライン手法がQA-basedを上回る

Summary-levelは？

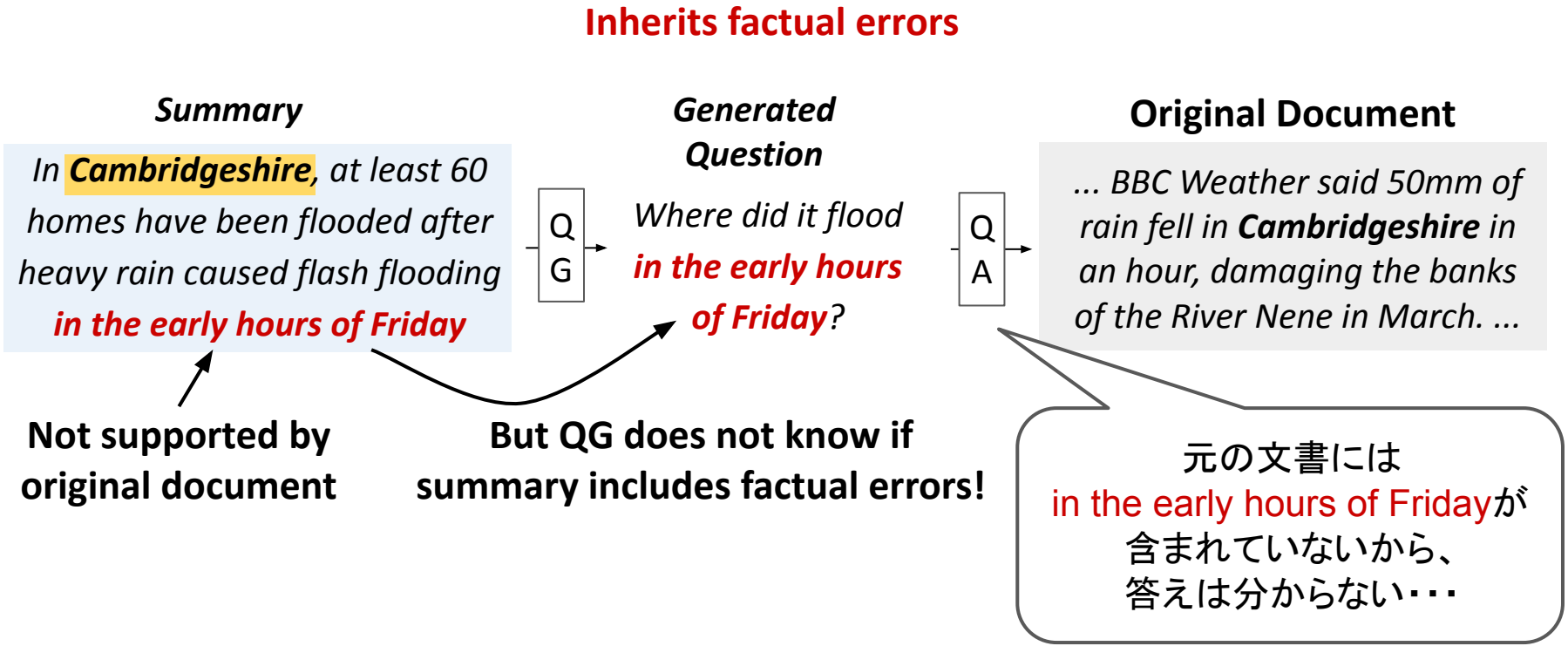
Span-LevelとSummary-Levelの比較

F1 score on GD21 (XSum) dataset



Summary-levelではQA-basedが優れている

なぜSpan-Levelのエラー検出ができないのか



質問・回答を読んでも何も解釈できない

Summary

*In **Cambridgeshire**, at least 60 homes have been flooded after heavy rain caused flash flooding **in the early hours of Friday***

Not supported by original document

Q: Where has heavy rain caused flash flooding **in the early hours of Friday**?
A: ❌ Unanswerable

Q: How many homes have been flooded **in the early hours of Friday**?
A: ❌ Unanswerable

Q: What caused flash flooding **in the early hours of Friday**?
A: ❌ Unanswerable

Inherited Errorは回避できるか？

Original Document

[...] The weather also hit Norfolk and Lincolnshire, […] BBC Weather said 50mm of rain fell in Cambridgeshire in an hour, damaging the banks of the River Nene in March. […]

Generated Summary

*In **Cambridgeshire**, at least 60 homes have been flooded after heavy rain caused flash flooding **in the early hours of Friday***

Long Question (Contains Extrinsic Errors)

Q: Where has heavy rain caused flash flooding **in the early hours of Friday**?
A: ❌ Unanswerable

If generated questions **inherit errors** in summary, the answer will be erroneous

生成する質問の長さを短くすればInherited Errrosは減らせるか？

Inherited Errorは回避できるか？

Original Document

*[...] The weather also hit **Norfolk and Lincolnshire**, […] BBC Weather said 50mm of rain fell in Cambridgeshire in an hour, damaging the banks of the River Nene in March. […]*

Generated Summary

*In **Cambridgeshire**, at least 60 homes have been flooded after heavy rain caused flash flooding **in the early hours of Friday***

Long Question (Contains Extrinsic Errors)

Q: Where has heavy rain caused flash flooding **in the early hours of Friday**?

A: ❌ Unanswerable

Short Question (Under-specified)

Q: Where was there heavy rain?

A: ❌ Norfolk and Lincolnshire

Short questions can be **underspecified for source document** and cannot be answered correctly

Inherited Errorは回避できるか？

Original Document

*[...] The weather also hit Norfolk and Lincolnshire, […] BBC Weather said 50mm of rain fell in Cambridgeshire in an hour, **damaging the banks** of the River Nene in March. […]*

Generated Summary

*In **Cambridgeshire**, at least 60 homes have been flooded after heavy rain caused flash flooding **in the early hours of Friday***

Long Question (Contains Extrinsic Errors)

Q: Where has heavy rain caused flash flooding **in the early hours of Friday**?

A: ❌ Unanswerable

Intermediate Question (Just-Right Length)

Q: Where has heavy rain **caused flash flooding**?

A: ✔ Cambridgeshire

Short Question (Under-specified)

Q: Where was there heavy rain?

A: ❌ Norfolk and Lincolnshire

Questions with "Just right" length is required for good span-level evaluation

But it is **not possible** to infer correct lengths when QG generates questions

なぜSummary-levelでは性能が高いのか？

- 長い質問を生成すれば、inherited errrosが発生するのでlocalizationはできなくなるが、誤りは検出することができる
  - Inherited errorsは誤りを含む要約文でしか発生しないため、inherited errorsによって「誤りがある」と判定されることはsummary-levelの評価には悪影響がない
- ちなみに、Fabbri et al. (2021)は実験的にQA-based metricsでは長い質問を生成することでsummary-levelの性能が高くなることを報告している

Long Question (Contains Extrinsic Errors)

Q: Where has heavy rain caused flash flooding **in the early hours of Friday**?
A: ❌ Unanswerable

If generated questions **inherit errors** in summary, the answer will be erroneous

QA-basedはEntailment-basedに対して優位性はあるか？

- どちらの手法も解釈可能ではない
- 長い質問を用いて情報を比較することは、sentence-levelのentailmentを評価していることとほとんど同じ
  - 長い質問がUnanswerableかを判定する ≒ Entailmentの評価
- 長い質問を生成しても、要約文には元の文書の全ての情報が入っているわけではないため、Under-specified questionを完全に防ぐことはできない
- 生成された答えが正しいかどうか（一致しているか）の判定は容易ではない

Long Question (Contains Extrinsic Errors)

Q: Where has heavy rain caused flash flooding **in the early hours of Friday**?
A: ❌ Unanswerable

If generated questions **inherit errors** in summary, the answer will be erroneous

概要

- 文書要約タスクの評価手法（Factualityの評価）
  - Entailment-based metrics, QA-based metrics, LLM-based metrics
  - どの評価指標を研究するべきか？(Kamoi et al., EACL2023)
- **長い入力への対応**
  - **Entailment-based metrics (Kamoi et al., arXiv2023 など)**
  - LLM-based metrics

Ryo Kamoi, Tanya Goyal, Greg Durrett. 2023. Shortcomings of Question Answering Based Factuality Frameworks for Error Localization. *EACL*.
Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, Greg Durrett. 2023. WiCE: Real-World Entailment for Claims in Wikipedia. *arXiv preprint arXiv:2303.01432*.

- アメリカの大学院への進学について

Entailment-based Metrics

Natural language inference (自然言語推論): 前提と仮説の含意関係を推測するタスク

Premise	Label	Hypothesis
Letters		
Your gift is appreciated by each and every student who will benefit from your generosity.	*neutral*	Hundreds of students will benefit from your generosity.
Telephone Speech		
yes now you know if if everybody like in August when everybody's on vacation or something we can dress a little more casual or	*contradiction*	August is a black out month for vacations in the company.
9/11 Report		
At the other end of Pennsylvania Avenue, people began to line up for a White House tour.	*entailment*	People formed a line at the end of Pennsylvania Avenue.

(MNLI Dataset)

文書要約において、元の文書が要約文を含意していれば、要約文はFactualだと言えるのでは？ → Entailment-based Metrics

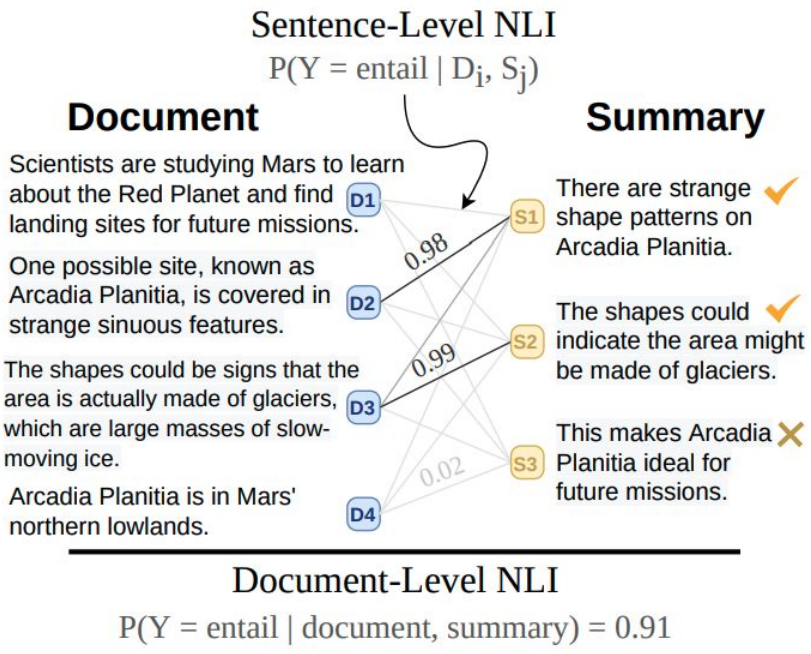
Entailment-based Metricsの課題1：モデルの入力サイズ

小規模の言語モデルは512～1024トークンくらいまでしか入力できない

↓

文書要約への適用には工夫が必要

解決策1：長い文書を分割する



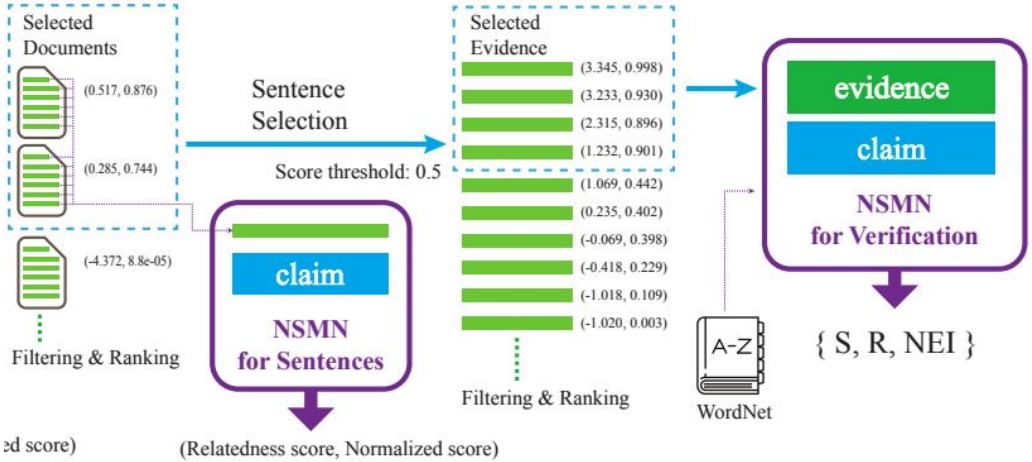
SummaC (Laban et al., 2021)

長い文書を文（もしくは段落など）に分割し、文同士の含意関係を用いて文書同士の含意関係を近似する

欠点

- 文脈が考慮されていない
- 元の文書の全体を読まないと含意関係が判断できない場合もある

解決策2：長い文書から必要な部分だけを取り出す



(Nie et al., 2018)

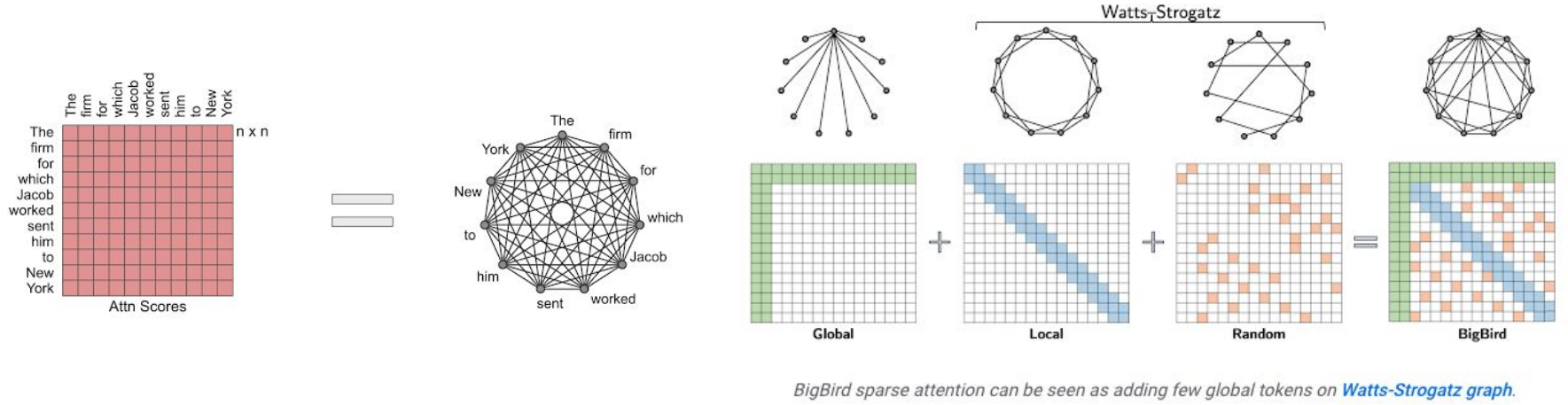
含意関係の判断に必要な部分を検索し、検索された文を用いて判定する

欠点

- そもそも検索が難しい
- やはり文脈は失われてしまうことが多い

解決策3：モデルの入力サイズを大きくする

Big Bird (Zaheer et al., 2020)



Longformer (Beltagy et al, 2020)
など入力長の長いモデルを使う

欠点

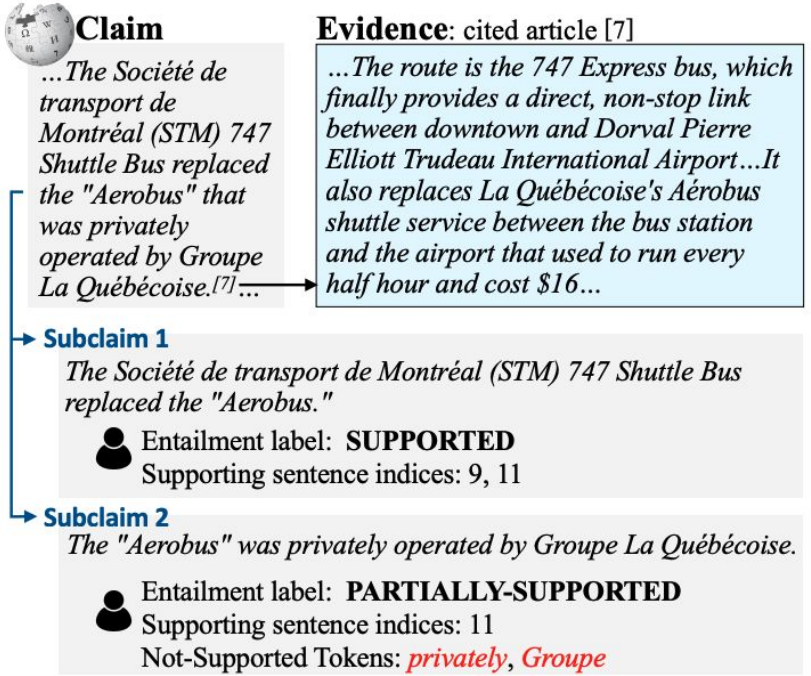
- モデルの入力サイズの拡大には限界がある
- 学習が難しい (Yin et al., 2021)

課題2：評価・学習のためのデータセットが不足している

既存のNLIデータセットの問題点

- 短い前提・仮説のデータセットが多い
- 人工的に作られた仮説を含むデータセットが多い
    - 特に負例（含意されていない例）を作成するときに、単語を入れ替えるなどの単純な加工を行う場合がある（例：DocNLI (Yin et al., 2021)）
- Fine-grainedなアノテーションを含むデータセットは少ない

WiCE: Real-World Entailment for Claims in Wikipedia

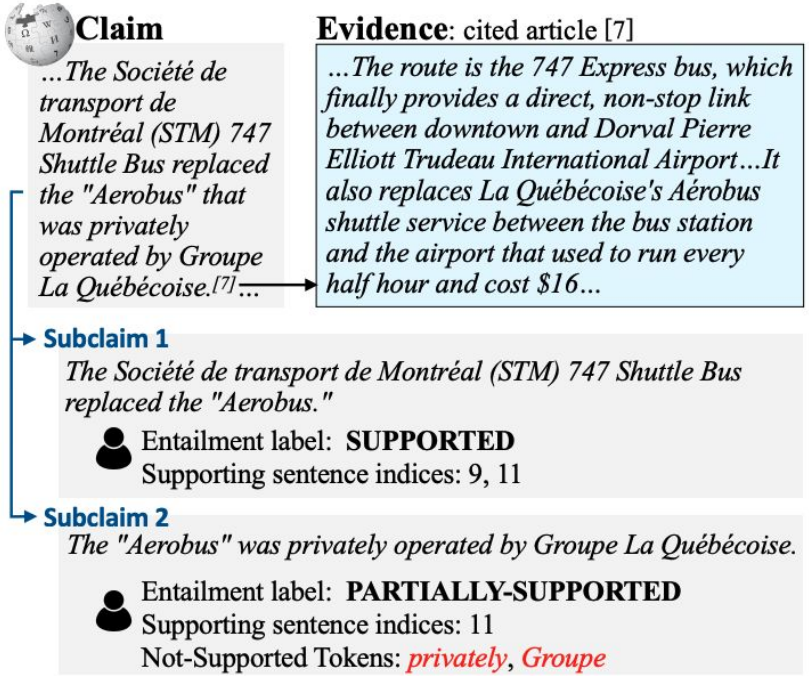


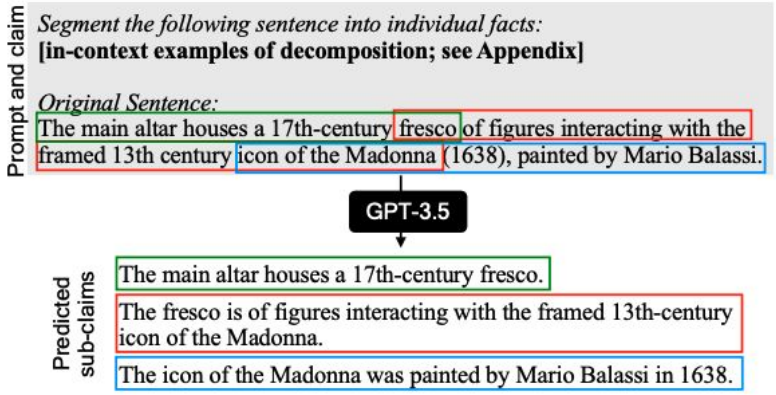
WiCE (Kamoi et al., 2023)

- Wikipediaの文を仮説、引用されているウェブサイトを前提とする
- Wikipediaでは引用されているウェブサイトに完全に文章が含意されていない場合も多くあるため、
  自然な負例が含まれている

Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, Greg Durrett. 2023. WiCE: Real-World Entailment for Claims in Wikipedia. *arXiv preprint arXiv:2303.01432*.

WiCE: Claim-Split

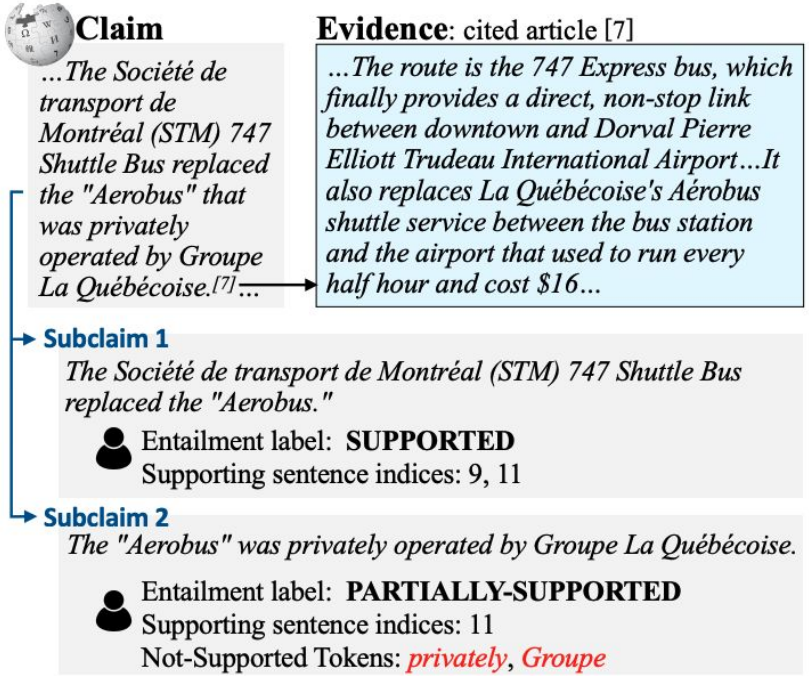




Claim-Split

- GPT-3などの大規模言語モデルを用いて、few-shotで長い仮説を分割
- 数%は誤りが含まれるので、人手で修正

WiCE: Real-World Entailment for Claims in Wikipedia



Fine-grainedなアノテーション

- Claim Split：長い仮説を自動的に分割し、細かいSub-claimレベルでのアノテーションを提供
- ウェブサイトの中で、どの文が仮説を含意しているのかを明示
- 仮説の一部しか含意されていない場合、含意されていない部分を明示

WiCE: 既存手法のパフォーマンス

Evidence Sentences Retrieval

- Human Performanceはかなり高い
- Fine-tuningを行ったモデルであっても、Human Performanceと比較すると大きく劣る

Train Data	Claim			Subclaim		
	F1	P	R	F1	P	R
BM25	14.3†	8.2†	90.3	9.8†	5.7†	72.2
ANLI	22.6†	24.7†	35.7†	30.5†	28.8†	40.5†
WiCE	44.6†	44.1†	61.0†	35.8†	33.6†	49.4†
ANLI+WiCE	56.5†	55.5†	70.0†	43.1	40.4	58.6
w/ evidence context						
WiCE	**61.4**	**59.2**	**74.4**	**43.5**	**40.9**	56.1
ANLI+WiCE	59.0†	56.9	70.8†	41.9	36.4	**63.9**
Human	90.9¶	92.2¶	92.6¶	90.7	92.3	91.6

Table 6: Performance of T5-3B on the **evidence retrieval task** of WiCE. †: Worse than T5-3B finetuned on WiCE (w/ context) with p-value < 0.05 in a paired bootstrap test. ¶: For claim level human performance, we take the union set of retrieved sentences at subclaim level. Human performance is on 50 random examples.

WiCE: 既存手法のパフォーマンス

Entailment Classification

- Chunk（256トークン以下）に分割して、ChunkごとのEntailment Scoreの最大値をとるだけでも一定の性能が得られる
- Retrievalの改善によってEntailment Classificationの性能を向上できる
- Human Performanceと比較すると低い

Unit	Train Data	Claim F1	Claim Acc	Subclaim F1	Subclaim Acc
sent	ANLI+WiCE	58.0†	62.8†	81.2†	77.3†
chunk	WiCE	65.3†	77.1	85.1†	82.7†
chunk	ANLI+WiCE	**72.1**	**79.1**	**87.3**	**85.0**
–	Human	83.3	92.0	94.4	94.4

EvidenceをSentenceもしくはChunkに分割し、Local Entailment Scoreの最大値を用いる

Setting	Claim F1	Claim Acc	Subclaim F1	Subclaim Acc
MAX best (Table 5)	72.1	79.1	87.3	85.0
top-k	71.1	79.3	86.6‡	84.2‡
top-k (w/ context)	**72.9**	**79.9**	**88.4**	**87.0**
oracle	78.0	84.4	88.7	87.7
Human	83.3	92.0	94.4	94.4

Retrieval + NLI

概要

- 文書要約タスクの評価手法（Factualityの評価）
  - Entailment-based metrics, QA-based metrics, LLM-based metrics
  - どの評価指標を研究するべきか？(Kamoi et al., EACL2023)
- **長い入力への対応**
  - Entailment-based metrics (Kamoi et al., arXiv2023 など)
  - **LLM-based metrics**

Ryo Kamoi, Tanya Goyal, Greg Durrett. 2023. Shortcomings of Question Answering Based Factuality Frameworks for Error Localization. *EACL*.
Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, Greg Durrett. 2023. WiCE: Real-World Entailment for Claims in Wikipedia. *arXiv preprint arXiv:2303.01432*.

- アメリカの大学院への進学について

LLMの入力サイズ

- 最近のLLMは入力サイズが長くなっているので、大抵の場合はLLM-based scoreを使うことができる
- しかし、入力サイズが数万トークン以上になるタスクもある
  - Academic paper summarization
  - Book summarization

Recurrent Transformer

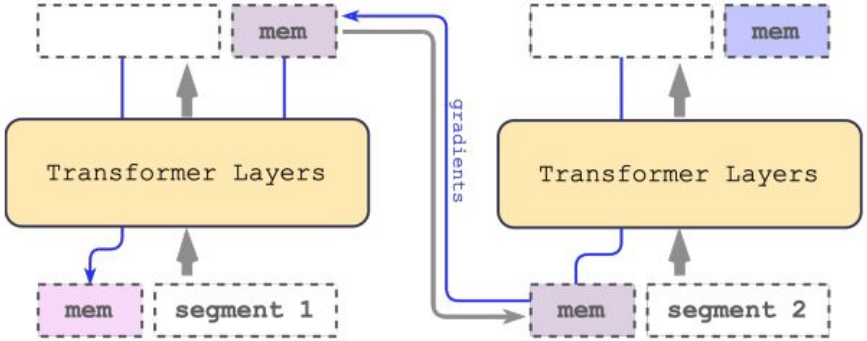


Figure 1: **Recurrent Memory Transformer.** Memory is added as tokens to the input sequence and memory output is passed to the next segment. During training gradients flow from the current segment through memory to the previous segment.

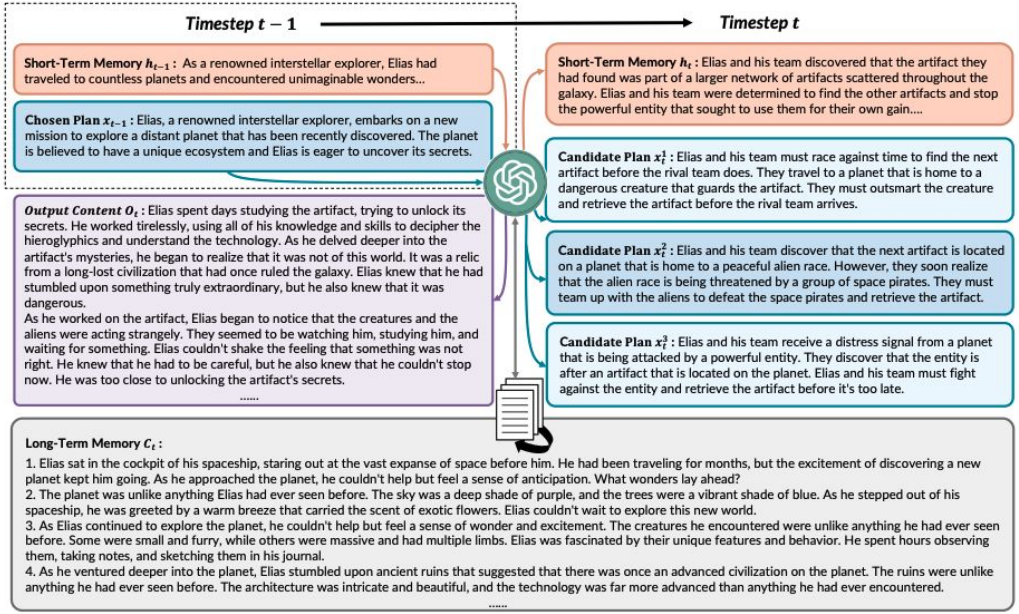
(Bulatov et al., 2022)

RNNの各ユニットをTransformerに変更したようなモデル

課題

- 長い文書を用いて学習する必要があり、どれくらいの規模に対応できるかは未知数

RecurrentGPT



(Zhou et al., 2023)

プロンプトの設計により、zero-shot LLMでRNNを再現する

課題

- プロンプトの長さによってメモリのサイズも制限される

RecurrentGPT

```
I need you to help me write a novel. Now I give you a memory (a brief summary) of 400 words, you should use it to store the key content of what has
been written so that you can keep track of very long context. For each time, I will give you your current memory (a brief summary of previous
stories. You should use it to store the key content of what has been written so that you can keep track of very long context), the previously
written paragraph, and instructions on what to write in the next paragraph. I need you to write:
1. Output Paragraph: the next paragraph of the novel. The output paragraph should contain around 20 sentences and should follow the input
instructions.
2. Output Memory: The updated memory. You should first explain which sentences in the input memory are no longer necessary and why, and then
explain what needs to be added into the memory and why. After that you should write the updated memory. The updated memory should be similar to the
input memory except the parts you previously thought that should be deleted or added. The updated memory should only store key information. The
updated memory should never exceed 20 sentences!
3. Output Instruction:  instructions of what to write next (after what you have written). You should output 3 different instructions, each is a
possible interesting continuation of the story. Each output instruction should contain around 5 sentences
    Here are the inputs:

    Input Memory:
    {short_memory}

    Input Paragraph:
    {input_paragraph}

    Input Instruction:
    {input_instruction}

    Input Related Paragraphs:
    {input_long_term_memory}

    Now start writing, organize your output by strictly following the output format as below:
    Output Paragraph:
    <string of output paragraph>, around 20 sentences.

    Output Memory:
    Rational: <string that explain how to update the memory>;
    Updated Memory: <string of updated memory>, around 10 to 20 sentences

    Output Instruction:
    Instruction 1: <content for instruction 1>, around 5 sentences
    Instruction 2: <content for instruction 2>, around 5 sentences
    Instruction 3: <content for instruction 3>, around 5 sentences

Very important: The updated memory should only store key information. The updated memory should never contain over 500 words! Finally, remember
that you are writing a novel. Write like a novelist and do not move too fast when writing the output instructions for the next paragraph. Remember
that the chapter will contain over 10 paragraphs and the novel will contain over 100 chapters. And this is just the begining. Just write some
interesting staffs that will happen next. Also, think about what plot can be attractive for common readers when writing output instructions. You
should first explain which sentences in the input memory are no longer necessary and why, and then explain what needs to be added into the memory
and why. After that, you start rewrite the input memory to get the updated memory.
```

まとめ

- 今後はEntailment-basedとLLM-basedの指標の研究がおすすめ
- さまざまな研究課題が残されている
  - NLIモデルおよびLLMを長い入力に対応させる
  - 解釈可能な評価指標、fine-grainedな出力ができる評価指標がほとんど存在しない
  - WiCEのように、長い入出力や多様なアノテーションを含むようなデータセットがさらに求められている