겨울방학 세미나

EMNLP 2022 recap

NLP&AI 강명훈

Theme of the seminar

- Question Generation
  - Question Generation 연구가 활발히 이뤄지고 있는데 어떻게 접근해야 할까?
  - 가용할 수 있는 resource는 무엇일까?

- Fact Verification
  - 검증의 대상이 되는 불완전한 context 속에서 어떻게 정확한 검증 point를 찾을 수 있는가?
  - Fact verification task를 접근할 수 있는 새로운 framework를 어떻게 구상할 수 있을까?
  - 새로운 framework속에 question generation을 활용할 수 있을까?

- Writing
  - 제안하는 모델의 필요성을 현실의 문제와 어떻게 적절히 엮을 수 있을까?
  - Problem setting을 어떻게 명확히 할 수 있을까?

Today's paper

- Question Generation
  - Question Generation 연구가 활발히 이뤄지고 있는데 어떻게 접근해야 할까?
  - 가용할 수 있는 resource는 무엇일까?

Generative Language Models for Paragraph-Level Question Generation

Asahi Ushio and **Fernando Alva-Manchego** and **Jose Camacho-Collados**
Cardiff NLP, School of Computer Science and Informatics, Cardiff University, UK
{UshioA,AlvaManchegoF,CamachoColladosJ}@cardiff.ac.uk

Today's paper

- Fact Verification
  - 검증의 대상이 되는 불완전한 context 속에서 어떻게 정확한 검증 point를 찾을 수 있는가?
  - Fact verification task를 접근할 수 있는 새로운 framework를 어떻게 구상할 수 있을까?
  - 새로운 framework속에 question generation을 활용할 수 있을까?

Varifocal Question Generation for Fact-checking

Nedjma Ousidhoum* Zhangdie Yuan* Andreas Vlachos
Department of Computer Science and Technology
University of Cambridge
ndo24,zy317,av308@cam.ac.uk

Generating Literal and Implied Subquestions to Fact-check Complex Claims

Jifan Chen **Aniruddh Sriram** **Eunsol Choi** **Greg Durrett**
Department of Computer Science
The University of Texas at Austin
jfchen@cs.utexas.edu

Why should we refer this paper?

Generative Language Models for Paragraph-Level Question Generation

Asahi Ushio and **Fernando Alva-Manchego** and **Jose Camacho-Collados**
Cardiff NLP, School of Computer Science and Informatics, Cardiff University, UK
{UshioA,AlvaManchegoF,CamachoColladosJ}@cardiff.ac.uk

- Valuable resource for question generation
  - Resource for training: Benchmark dataset, Fine-tuned model 모두 공개
  - Resource for research: Multilingual, Multidomain 에서의 실험 내용 공개
    → 향후 QG task에서의 domain, language별 고려사항 제공

Preliminary knowledge

Generative Language Models for Paragraph-Level Question Generation

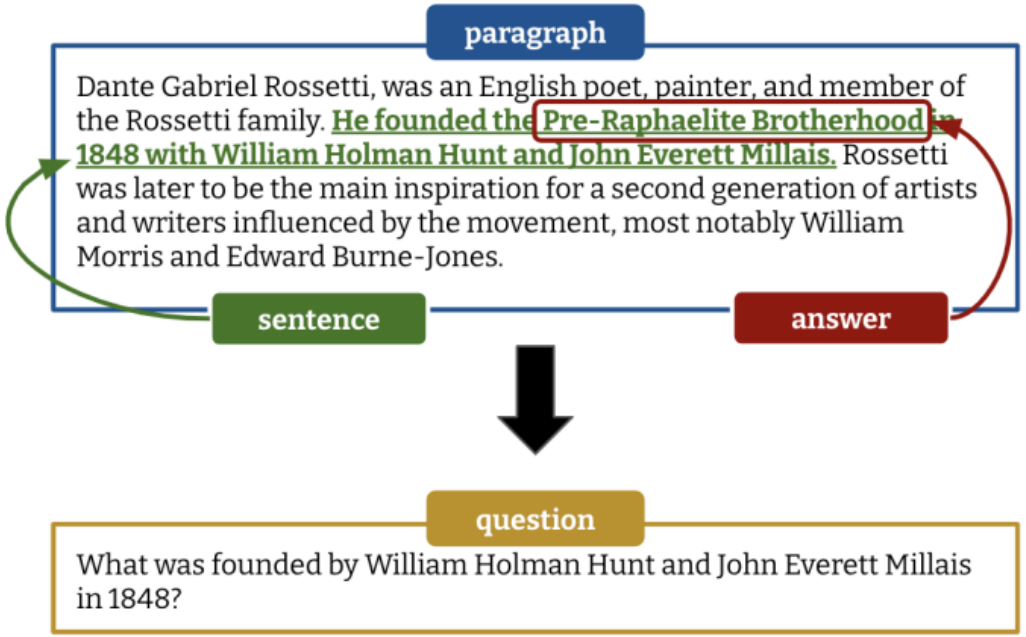


Figure 1: Overview of paragraph-level QG.

Task of generating a question given an **input context** consisting of a *document*, a *paragraph* or a *sentence*, and an *answer* where the question is anchored

즉, input으로 다양한 길이의 text가 사용될 수 있고 그 text를 활용하여 적절한 question을 만드는 것이 QG의 목표라 할 수 있음

Research Question: 이런 다양한 setting에 적용할 수 있는 QG는 어떻게 평가할 수 있을까?

→ 본 논문의 접근 point!

Introduction

- Problem setting: We need standardized approach for QG
  - QG task에서 backbone 모델로 어떤 PLM을 선정해야 하는지에 대한 기준 불명확
  - Automatic evaluation에 사용되는 BLEU, METEOR, ROUGE의 효용성에 대한 의문
  - 서로 다른 input을 사용하는 QG 모델들에 대한 비교 사례 없음

- Contribution
  - 표준화된 비교를 위한 Multilingual, Multidomain **QG-Bench 데이터셋** 제안
  - **T5, BART**를 이용한 input length, input domain, input language에 따른 실험결과 제공
  - **Automatic evaluation, Manual Evaluation** 결과를 모두 제공하고 유의미한 분석 제공
    - 두 지표 간의 상관관계 분석을 통해 기존 Automatic evaluation 지표 중 유의미한 지표, 한계점 등을 제시

QG-Bench

	Data size (train/valid/test)	Average character length (para./sent./ques./ans.)
SQuAD	75,722 / 10,570 / 11,877	757 / 179 / 59/ 20
SubjQA		
- *Book*	637 / 92 / 191	1,514 / 146 / 28 / 83
- *Elec.*	697 / 99 / 238	1,282 / 129 / 26 / 66
- *Grocery*	687 / 101 / 379	896 / 107 / 25 / 49
- *Movie*	724 / 101 / 154	1,746 / 146 / 27 / 72
- *Rest.*	823 / 129 / 136	1,006 / 104 / 26 / 51
- *Trip*	875 / 143 / 397	1,002 / 108 / 27 / 51
SQuADShifts		
- *Amazon*	3,295 / 1,648 / 4,942	773 / 111 /43 / 18
- *Wiki*	2,646 / 1,323 / 3,969	773 / 184 / 58 / 26
- *News*	3,355 / 1,678 / 5,032	781 / 169 / 51 / 20
- *Reddit*	3,268 / 1,634 / 4,901	774 / 116 / 45 / 19
Multilingual QG		
- *Ja*	27,809 / 3,939 / 3,939	424 / 72 / 32 / 6
- *Es*	77,025 / 10,570 / 10,570	781 / 122 / 64 / 21
- *De*	9,314 / 2,204 / 2,204	1,577 / 165 / 59 / 66
- *Ru*	40,291 / 5,036 / 5,036	754 174 / 64 / 26
- *Ko*	54,556 / 5,766 / 5,766	521 / 81 / 34 / 6
- *It*	46,550 / 7,609 / 7,609	807 / 124 / 66 / 16
- *Fr*	17,543 / 3,188 / 3,188	797 / 160 / 57 / 23

Table 1: Statistics of of all datasets integrating into our question generation benchmark after unification.

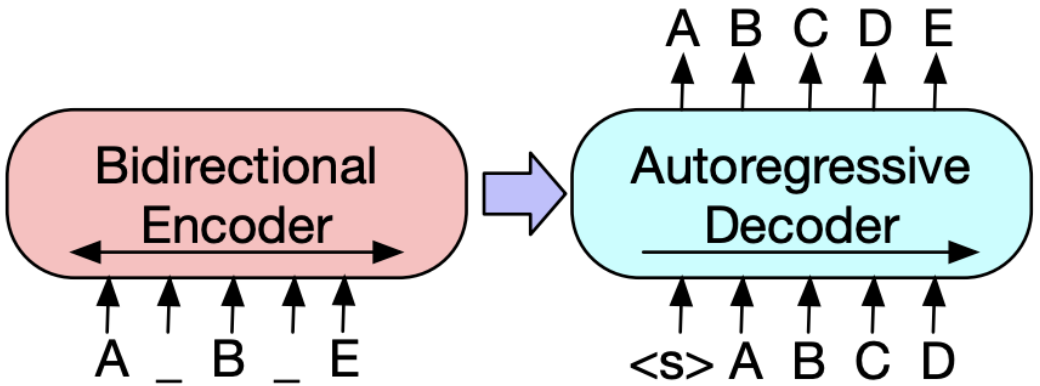
- Details

데이터 형식
- paragraph, sentence, question, answer
  - answer는 항상 sentence 안에 포함되어 있음

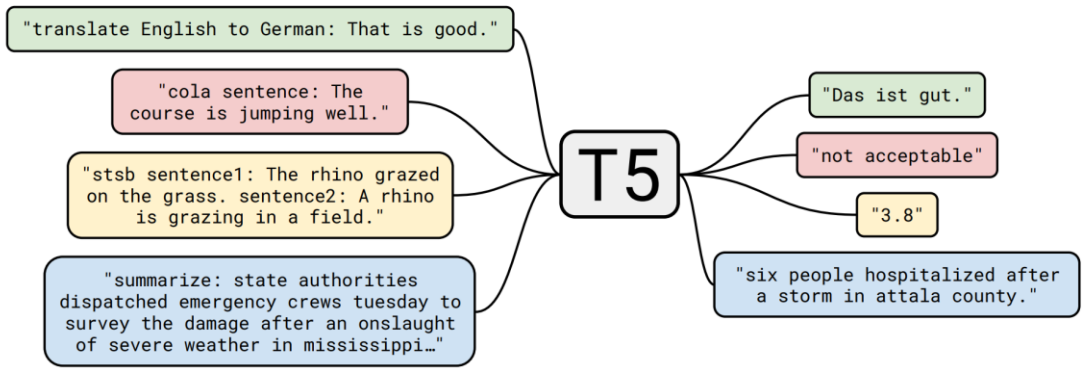
English dataset
- SQuAD (single domain)
- SQuADShifts (multi domain)
  - SQuAD와 text style은 동일하나 새로운 domain에 해당하는 데이터
- SubjQA (multi domain)
  - SQuAD와 text style이 다른 주관적인 질문과 그에 대한 답변이 포함된 데이터셋

Multilingual dataset
- Japanese: JAQuAD, Korean: KorQuAD
- German: GerQuAD, French: FQuAD
- Russian: SberQuAD, Spanish: Spanish SQuAD
- Italian: Italian SQuAD

Model



BART



T5

Training

- English: BART, T5 모델 사용
- Multilingual: mBART, mT5 사용

Input: $x = [c_1, \ldots, \texttt{<hl>}, a_1, \ldots, a_{|a|}, \texttt{<hl>}, \ldots, c_{|c|}]$

[CLS] The Super Bowl 50 was played at [HL] Santa Clara, California [HL] .

Output: question sentence

Experimental setup

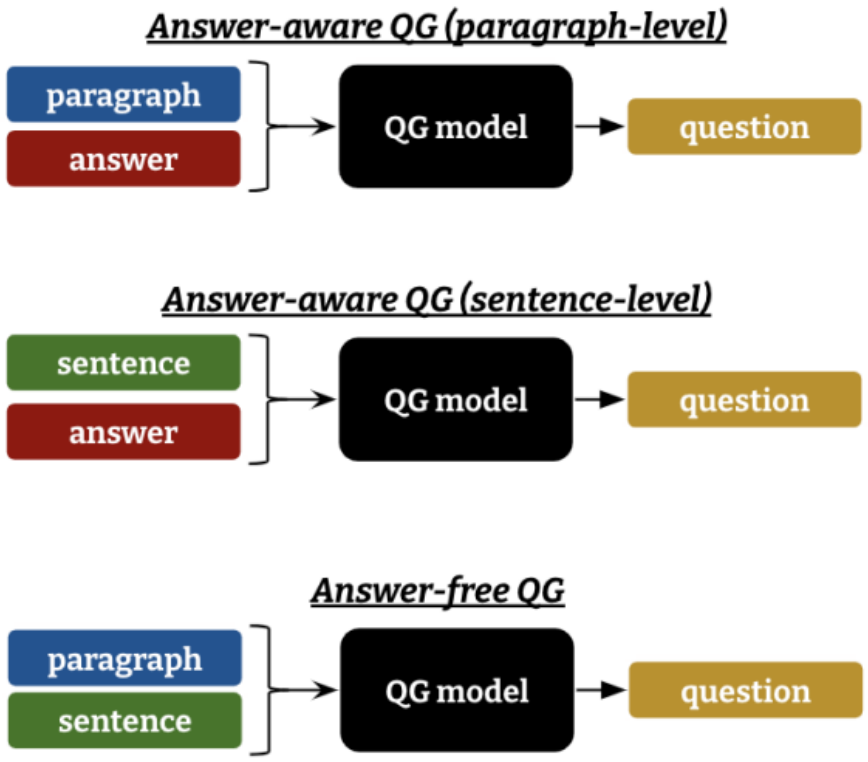


Figure 2: Input variations of QG models.

Evaluation Metric

- N-gram base metric: BLEU4, METEOR, ROUGE-L
- Model base metric: BERT score, MoverScore

Ablation Setting

- Model Input
  - Pargarph-level, Sentence-level, Answer-free
- Domain Adaptation
  - In-domain fine-tuning,
    zero-shot transfer from SQuAD,
    in-domain + SQuAD fine-tuning

Manual Evalautaion (human evaluataion)

500개의 case에 대해 5명의 annotator들이 평가 (리커트 3점 척도)

- Answerability : whether question can be answered by given input
- Grammaticality: grammatical correctness
- Understandability: whether question is easy to be understood

Experimental result 1: main result

Model	Param	B4	R-L	MTR	BS	MS
NQG (Du et al.)	30M	12.28	39.75	16.62	-	-
UniLM (Dong et al.)	340M	22.78	51.57	25.49	-	-
UniLMv2 (Bao et al.)	110M	24.70	52.13	26.33	-	-
ProphetNet (Qi et al.)	340M	23.91	52.26	26.60	-	-
ERNIE-G (Xiao et al.)	340M	25.40	52.84	26.92	-	-
$BART_{BASE}$	140M	24.68	52.66	26.05	90.87	64.47
$BART_{LARGE}$	400M	26.17	53.85	27.07	**91.00**	64.99
$T5_{SMALL}$	60M	24.40	51.43	25.84	90.45	63.89
$T5_{BASE}$	220M	26.13	53.33	26.97	90.84	64.74
$T5_{LARGE}$	770M	**27.21**	**54.13**	**27.70**	**91.00**	**65.29**

Analysis

- T5 계열의 모델이 BART대비 우수한 성능을 보임
  - T5 small은 적은 파라미터 수에도 불구하고
    다른 모델들과 비교하여 견줄만한 성능을 보임
- Multilingual 실험 결과, English 대비 다른 언어에서는
  다국어 모델이 낮은 성능을 보임
  - 특히 German, French같이 데이터 수가 적은 언어에서 성능이 가장 안 좋음

	Model	B4	R-L	MTR	BS	MS
English	$mT5_{SMALL}$	21.65	48.95	23.83	90.01	62.75
	$mT5_{BASE}$	**23.03**	**50.67**	**25.18**	90.23	63.60
	mBART	**23.03**	50.58	25.10	**90.36**	**63.63**
Russian	$mT5_{SMALL}$	16.31	31.39	26.39	84.27	62.49
	$mT5_{BASE}$	17.63	33.02	28.48	85.82	64.56
	mBART	**18.80**	**34.18**	**29.30**	**87.18**	**65.88**
Japanese	$mT5_{SMALL}$	30.49	50.88	29.03	80.87	58.67
	$mT5_{BASE}$	**32.54**	52.67	**30.58**	81.77	59.68
	mBART	32.16	**52.95**	29.97	**82.26**	**59.88**
Italian	$mT5_{SMALL}$	7.37	21.93	17.57	80.80	56.79
	$mT5_{BASE}$	**7.70**	**22.51**	**18.00**	**81.16**	**57.11**
	mBART	7.13	21.69	17.97	80.63	56.84
Korean	$mT5_{SMALL}$	10.57	25.64	27.52	82.89	82.49
	$mT5_{BASE}$	**12.18**	**28.57**	29.62	**84.52**	**83.36**
	mBART	10.92	27.76	**30.23**	83.89	82.95
Spanish	$mT5_{SMALL}$	9.61	24.62	22.71	84.07	59.06
	$mT5_{BASE}$	**10.15**	**25.45**	**23.43**	**84.47**	**59.62**
	mBART	9.18	24.26	22.95	83.58	58.91
German	$mT5_{SMALL}$	0.43	10.08	11.47	79.90	54.64
	$mT5_{BASE}$	**0.87**	11.10	13.65	80.39	55.73
	mBART	0.75	**11.19**	**13.71**	**80.77**	**55.88**
French	$mT5_{SMALL}$	**8.55**	**28.56**	**17.51**	**80.71**	**56.50**
	$mT5_{BASE}$	6.14	25.88	15.55	77.81	54.58
	mBART	0.72	16.40	7.78	71.48	50.35

Experimental result 2: model input

	Model	B4	R-L	MTR	BS	MS
Answer-free	$BART_{BASE}$	21.97	49.70	23.72	90.38	63.07
	$BART_{LARGE}$	23.47	50.25	24.94	90.28	63.28
	$T5_{SMALL}$	21.12	47.47	23.38	89.64	62.07
	$T5_{BASE}$	22.86	49.51	24.52	90.03	62.99
	$T5_{LARGE}$	24.27	51.30	25.67	90.41	63.97
Sent-level	$BART_{BASE}$	23.86	51.43	25.18	90.70	63.85
	$BART_{LARGE}$	23.86	51.43	25.18	90.70	63.85
	$T5_{SMALL}$	23.23	50.18	24.80	90.36	63.18
	$T5_{BASE}$	24.33	51.81	25.81	90.73	64.00
	$T5_{LARGE}$	25.36	52.53	26.28	90.88	64.44
Para-level	$BART_{BASE}$	24.68	52.66	26.05	90.87	64.47
	$BART_{LARGE}$	26.17	53.85	27.07	**91.00**	64.99
	$T5_{SMALL}$	24.40	51.43	25.84	90.45	63.89
	$T5_{BASE}$	26.13	53.33	26.97	90.84	64.74
	$T5_{LARGE}$	**27.21**	**54.13**	**27.70**	**91.00**	**65.29**

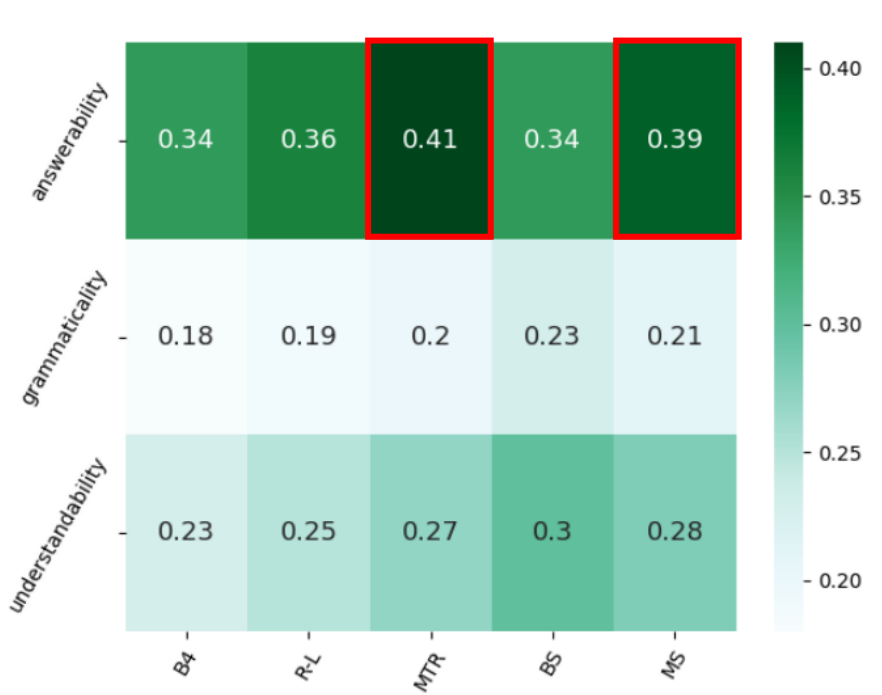
Analysis

- Paragraph-level 로 input을 줄 때 모든 경우에 비하여 성능이 가장 좋음
  - QG에서 global context를 주는 것이 좋은 방법일 수 있음
- Answer-free 경우에도 어느 정도의 성능을 보장함
  - 그럼에도 해당 case의 T5-large 모델은 para-level의 T5-small 모델보다 낮은 성능을 보임
- 다만 ans vs sent 격차는 sent vs para 대비 큼
  - 이는 QG task에서 모델의 input에 answer가 들어가는 것의 중요성을 알 수 있음

☞ global context를 주면서 answer-free한 경우의 성능을 높이는 바가 향후 중심 연구주제가 될 것

Experimental result 3: manual evaluation

Model	Manual Metric			Automatic Metric				
	Ans.	Gra.	Und.	B4	R-L	MTR	BS	MS
NQG	1.21	2.35	2.63	3.33	14.30	33.53	88.27	58.25
$BART_{LARGE}$	2.70	2.89	2.93	16.15	29.93	51.35	**90.95**	65.44
$T5_{SMALL}$	2.51	2.83	2.90	13.43	27.38	48.86	90.41	64.27
$T5_{LARGE}$	**2.80**	**2.93**	**2.95**	**17.56**	**30.42**	**52.00**	90.94	**66.09**
- sent-level	2.47	2.91	**2.95**	14.88	27.49	48.97	90.76	64.53
- answer-free	2.46	2.91	**2.95**	13.62	26.82	47.37	90.20	64.00



Analysis

- T5 Large 모델은 3가지 평가 척도 모두에서 3점에 가까운 성능을 보임

Additional Analysis

Manual metric과 Automatic Metric간의 spearman 상관관계를 측정

- Answerability를 제외하고는 Manual, Automatic metric간의 상관관계가 낮았음
- Answerability에 한하여 METEOR, MoverScore는 상대적으로 높은 상관관계를 보임
- 어떠한 automatic metric이 manual metric과의 높은 agreement를 달성하지 못함

☞ QG task에 적절한 Metric 제안 또한 후속 연구의 중심이 될 것

Conclusion & Insights

- Conclusion
  - 향후 QG task 수행에 유용한 QG-Bench 데이터셋과 BART, T5 기반 QG모델을 제안
  - 현재 QG task에서 부족한 점을 실험으로서 증명함 (Multilingual, Automatic Metric 등)

- Insights
  - 다른 task에서 사용되던 데이터셋을 최대한 활용하여 설정한 problem setting에 맞게 주조하는 좋은 reference 논문
  - 현재 QG에서 가능한 점과 부족한 점을 동시에 알려주는 논문으로써 이 task에 접근하고자 하는 연구자들에게 좋은 논문
    - 후속 연구에서 related works에 꼭 들어갈 논문

Why should we refer this paper?

Varifocal Question Generation for Fact-checking

Nedjma Ousidhoum* Zhangdie Yuan* Andreas Vlachos
Department of Computer Science and Technology
University of Cambridge
ndo24,zy317,av308@cam.ac.uk

- QG in fact verification domain
  - QG가 QA외의 도메인에서 어떻게 적용될 수 있는지에 관한 insight를 주는 논문
  - 한 개의 context에 대해서 다양한 방면으로 질문을 생성할 수 있는 방법을 참조할 수 있는 논문
  - 방법의 novelty보다 제안 필요성을 잘 서술한 논문

Introduction

- Problem setting: Solving paradox of QG in fact verification
  - Fan et al.(2020)에 따르면 팩트체킹을 할 때 claim과 관련된 question을 줄 경우 사람의 fact checking에 할애되는 시간이 감소되었다
  - 모델이 자동으로 question을 생성하는 QG를 이에 적용해볼 수 있음
    - 그런데 일반적인 Answer-aware QG setting을 사용하기에는 문제가 존재
  - Claim안에 answer는 명시적으로 존재하지 않고 fact checker들이 찾아야할 대상인데 어떻게 answer 없이 question을 생성할 수 있을까?
- Contribution
  - Claim 속의 focal point를 추출하여 claim 속 검증이 필요한 question을 생성하는 Varifocal 모델 제안
  - 제안한 Varifocal 모델의 우수성을 비교실험을 통해 증명
  - 생성한 question의 품질을 human evaluation을 통해서 증명

Model

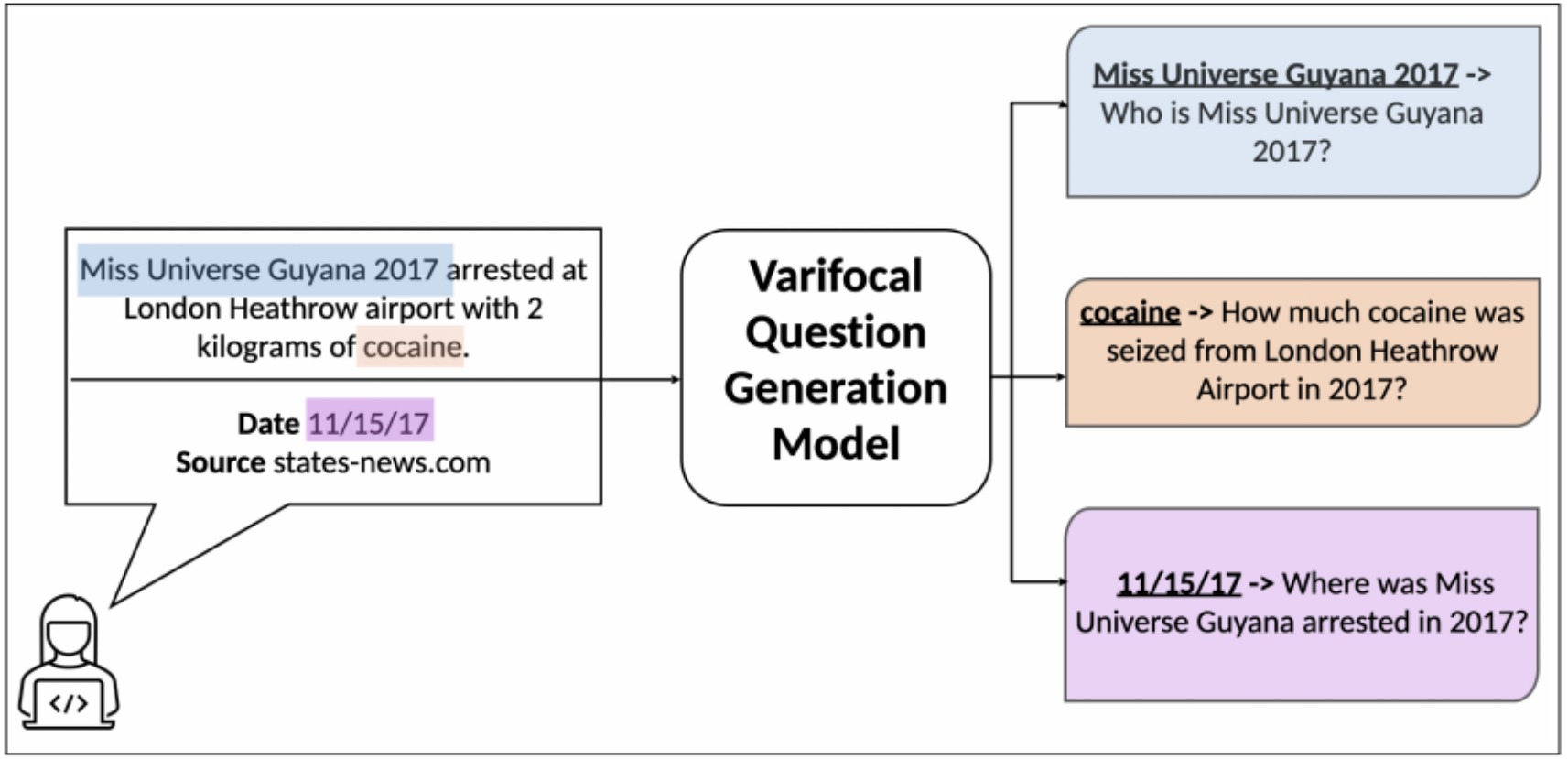
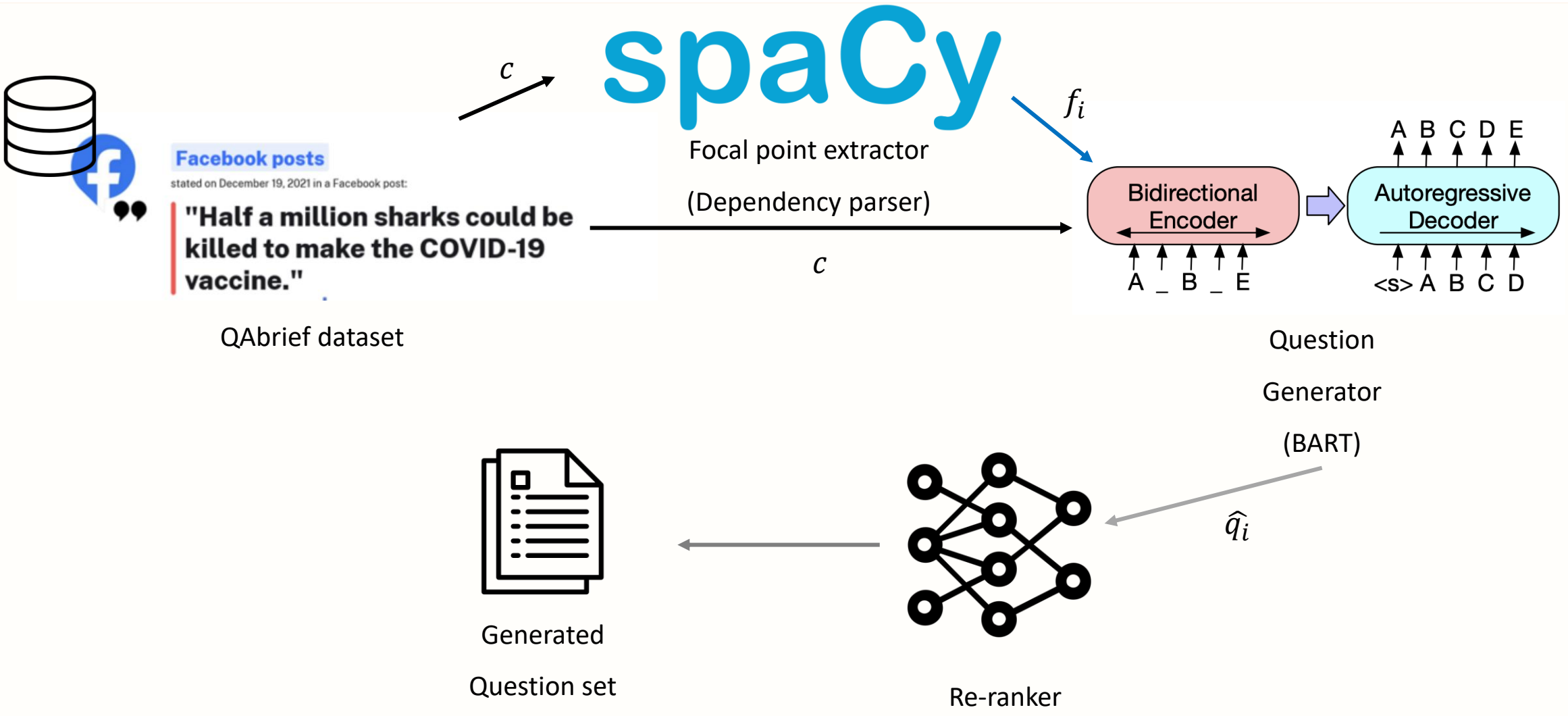


Figure 1: The architecture of Varifocal. We use a dependency parser to extract the different focal points, i.e. spans, then generate questions based on them. We rank the generated questions using a re-ranker and return the top n questions. The example in the figure was generated by our system. We show three highlighted focal points along with the (output) questions they led to.

Model


c
spaCy
Focal point extractor
(Dependency parser)
c
f_i
Facebook posts
stated on December 19, 2021 in a Facebook post:
"Half a million sharks could be killed to make the COVID-19 vaccine."
QAbrief dataset
Bidirectional Encoder
Autoregressive Decoder
A B C D E
A _ B _ E
<s> A B C D
Question Generator (BART)
$\hat{q}_i$
Re-ranker
Generated Question set

Dataset : QABrief Dataset

2020 EMNLP main accepted

Fact Verification domain에 적용하기 위한 QA 데이터셋

Data collection

- DATACOMMONS, MULTIFC에서 claim을 추출
- Annotator을 이용하여 claim과 fact-checking article이 주어질 때 question generation
- 생성된 question과 상응하는 answer를 검색엔진을 이용하여 찾기

Dataset : QABrief Dataset

Train	Number of Claims	5,897
	Number of QA Pairs	18,281
Valid	Number of Claims	500
	Number of QA Pairs	1,431
Test	Number of Claims	500
	Number of QA Pairs	1,456
Avg Number Questions/Claim		3.16
Avg Number Words in Questions		10.54
Avg Number Words in Answers		43.56

Table 1: **Statistics of QABRIEFDATASET**

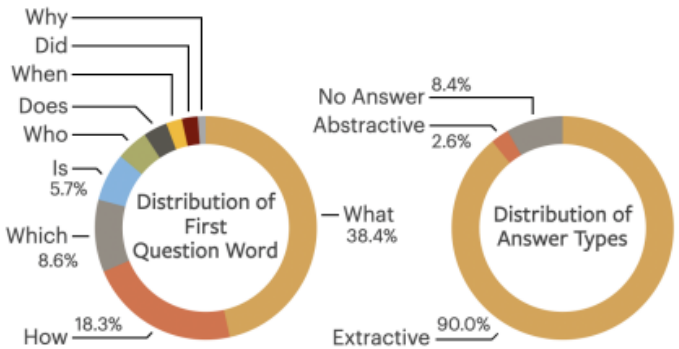


Figure 4: **Question and Answer Types**

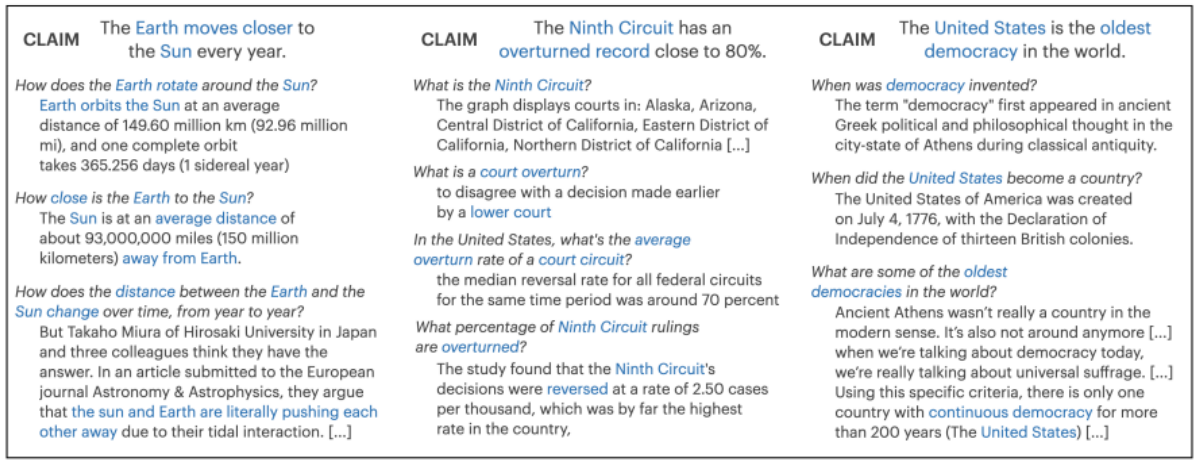


Figure 3: **Examples of QABriefs in QABRIEFDATASET**

2020 EMNLP main accepted

Fact Verification domain에 적용하기 위한 QA 데이터셋

Data collection

- DATACOMMONS, MULTIFC에서 claim을 추출
- Annotator을 이용하여 claim과 fact-checking article이 주어질 때 question generation
- 생성된 question과 상응하는 answer를 검색엔진을 이용하여 찾기

Experimental setup

Train	Number of Claims	5,897
	Number of QA Pairs	18,281
Valid	Number of Claims	500
	Number of QA Pairs	1,431
Test	Number of Claims	500
	Number of QA Pairs	1,456
Avg Number Questions/Claim		3.16
Avg Number Words in Questions		10.54
Avg Number Words in Answers		43.56

Table 1: **Statistics of QABRIEFDATASET**

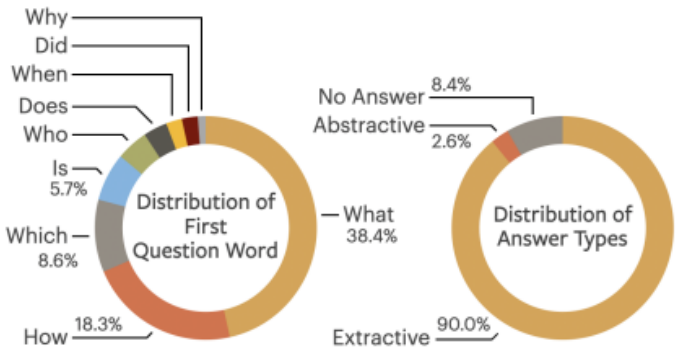


Figure 4: **Question and Answer Types**

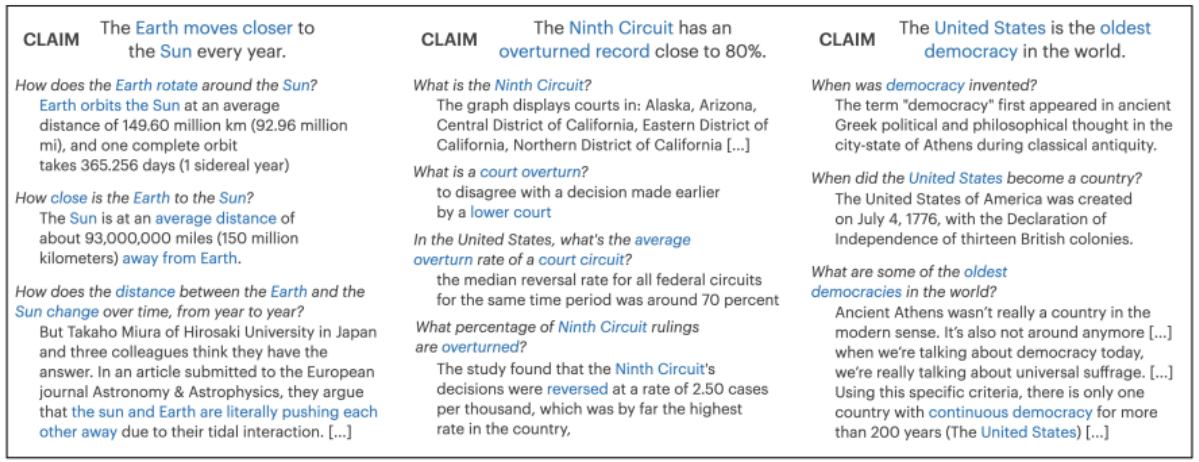


Figure 3: **Examples of QABriefs in QABRIEFDATASET**

Evaluation Metric

- BLEU2, BLEU4, chrF, METEOR, ROUGE-1, ROUGE-2, ROUGE-L, TER

Manual Evalautaion (human evaluataion)

500개의 case에 대해 5명의 annotator들이 평가 (리커트 3점 척도)

- Intelligibility : whether question is fluent
- Clarity: clear enough to be answered confidently using a search engine
- Relevance: only related to the mentioned entities
- Informativenes: whether question helps to fact-check the claim

Experimental result: automatic evaluation

System	BLEU-2	BLEU-4	chrF	METEOR	ROUGE-1	ROUGE-2	ROUGE-L	TER
BART	25.63	11.83	37.25	31.57	33.66	14.02	33.34	0.804
wh-BART	21.88	10.54	40.1	33.62	29.97	13.19	29.46	0.8697
SQuAD	25.85	11.95	38.60	30.67	32.70	13.63	31.84	0.809
Varifocal	29.98	15.17	41.12	34.84	37.27	17.64	36.77	**0.755**
Varifocal+Meta	**30.18**	**15.54**	**43.17**	**37.02**	**38.19**	**18.37**	**37.59**	0.764

Table 1: Automatic evaluation results on the QABriefs test set. For all scores higher is better except for TER.

Analysis

- BART: QABriefs의 claim, golden question을 [SEP]으로 구분해서 입력으로 넣음
- wh-BART: 위 모델에 What, Why, How 등의 의문문으로 question을 생성하도록 forcing
- SQuAD: SQuAD 데이터로 먼저 pretraining 한 뒤에 QABrief로 fine-tuning
  - 다만 SQuAD에서 answer를 사용하는 것이 아닌 제안 방법처럼 focal points를 대신 사용
- Varifocal: SQuAD pretrain + 제안 방법
- Varifocal + Meta: SQuAD pretrain + 제안 방법 + meta 데이터 input으로 추가

Experimental result: manual evaluation

Avg	I	C	R	Info
Gold	**0.97**	0.91	0.79	1.72
SQuAD	0.83	0.84	0.77	1.91
BART	0.85	0.76	0.67	1.49
Varifocal	**0.97**	**0.94**	**0.93**	**2.33**
Varifocal+Meta	0.93	0.91	0.89	2.10

Table 2: Average of intelligibility (I), clarity (C), relevance (R), and informativeness scores per system based on our human evaluation.

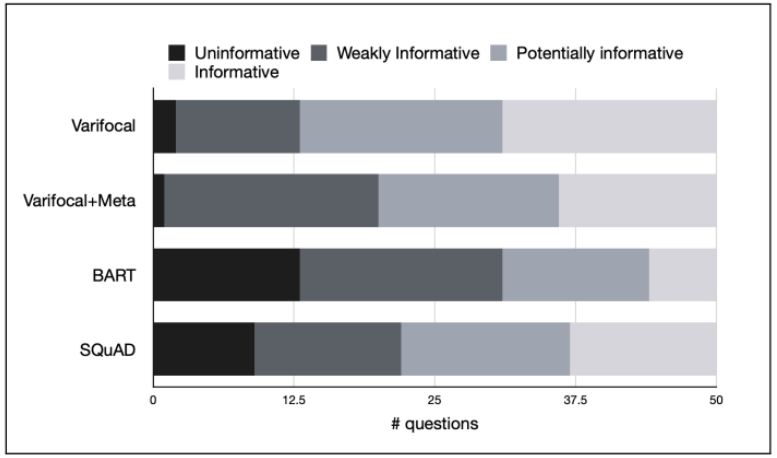


Figure 2: The distribution of informativeness scores across the different systems. Brighter means better.

- Analysis
  - 모든 부분에서 제안한 모델 Varifocal이 가장 높은 점수를 보임
  - 다만 informativeness에서 inter annotator agreement는 낮음 (0.26)

Conclusion & Insights

- Conclusion
  - QG를 fact verification에 적용하려고 했음
  - Fact verification domain에서 기인하는 answer-free 문제를 적용하고자 focal points를 사용함

- limitations
  - Focal points를 추출하는 데에 단순히 dependency parser를 사용하고 그 이유를 밝히지 않음
  - claim마다 focal points의 개수를 normalizing 하지 않음
    - focal points간의 중요성을 고려하지 않음
  - 실제 fact verification에서 제안한 QG가 최종 classification에 도움이 될 수 있는지를 검증하지 않음

☞ 개선할 점이 많은 연구이므로 후속 연구로 접근하기 좋은 주제

Why should we refer this paper?

Generating Literal and Implied Subquestions to Fact-check Complex Claims

Jifan Chen **Aniruddh Sriram** **Eunsol Choi** **Greg Durrett**
Department of Computer Science
The University of Texas at Austin
jfchen@cs.utexas.edu

- Sub-questions help with fact verification!
  - 현재 fact verification의 naïve한 접근법을 벗어나는 아이디어와 인사이트를 제시하는 논문
  - 복잡하고 암시적인 context를 담은 sentence단위 claim에서 literal, implied sub question을 담은 데이터셋을 제공하는 논문
  - 실제 subquestion을 input으로 줄 때 retriever의 성능 향상 증대를 preliminary experiment로 증명함

Preliminary knowledge

Generating Literal and Implied Subquestions to Fact-check Complex Claims

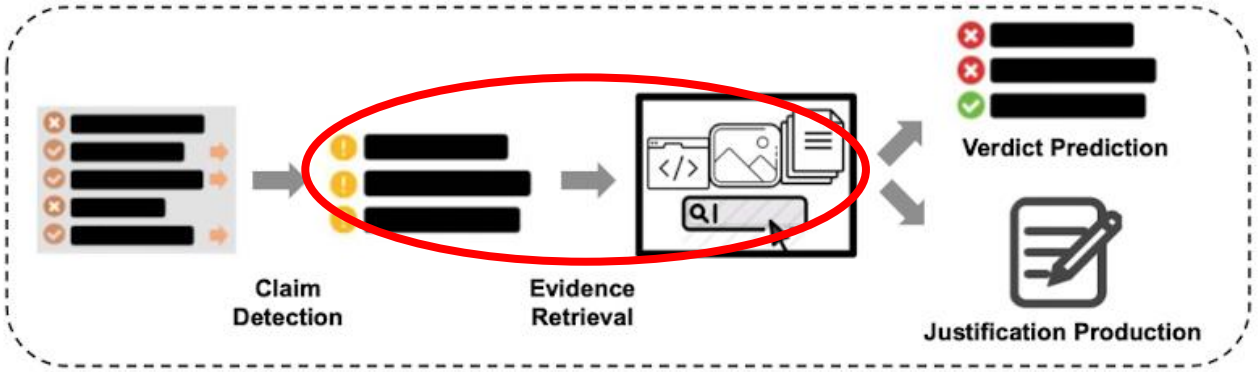


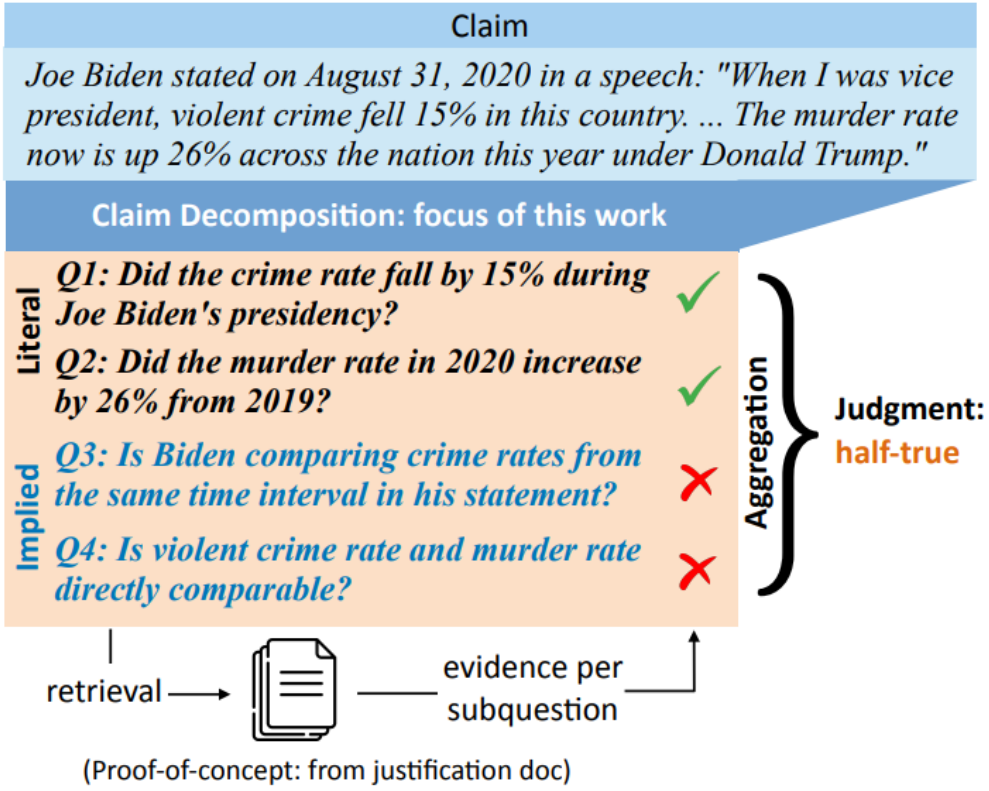
Figure 2: A natural language processing framework for automated fact-checking.

Evidence-based Fact-verification NLP system의 구성은 4가지 모듈로 구성됨

1. Claim Detection: 주어진 문장이 검증이 필요한지 여부 확인 → 필요한 경우 claim이 됨
2. Evidence Retrieval: claim을 지지 (support) 혹은 반박 (refute)할 수 있는 evidence를 검색
3. Verdict Prediction: 검색된 evidence를 이용하여 claim에 대한 최종 판별 진행
4. Justification Production: claim 판별 이유에 대한 generation진행

Introduction

- Problem setting: prediction doesn't explain much for fact checking



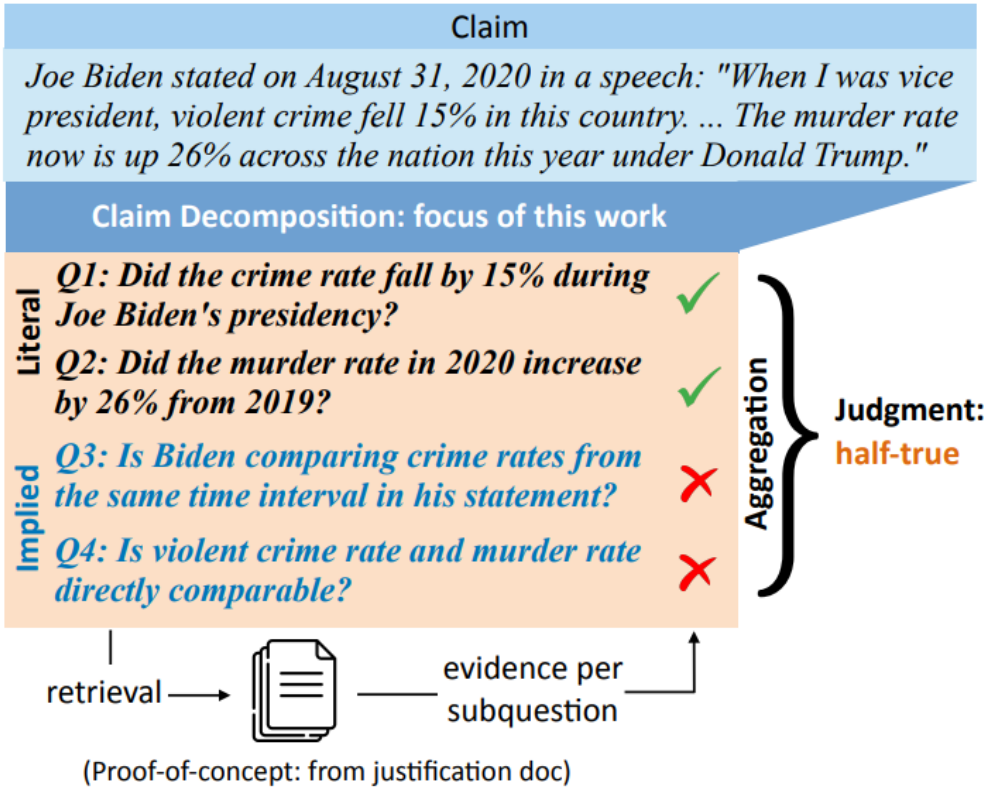
다음 claim의 fact checking 결과가 half-true라고 나왔을 때,
우리는 **정확히 어떤 점이 거짓에 해당하고 진실**인지를 알 수 없다

↓

1) claim의 검증 사항을 sub-question으로 분해하고
2) 각 sub-question별 fact checking을 진행한다면
3) 그 결과를 종합할 때 최종 검증 결과를 이해할 수 있다

Introduction

- Problem setting: prediction doesn't explain much for fact checking



- Contribution

  - claim속 복잡한 context를 sub-question으로 분해한 ClaimDecomp 데이터셋 제안
    - literal, implied sub-question모두 제시
    - 각 sub-question은 yes, no로 대답이 가능한 question
  - 제안한 sub-question을 retriever의 input으로 사용할 때의 모델 성능 향상 실험을 제공
    - 새로운 fact verification framework 개발 가능성 암시

Dataset Construction

Claim: A Facebook post stated on January 31, 2021: "Nancy Pelosi bought $1.25 million in Tesla stock the day before Joe Biden signed an order "for all federal vehicles" to be electric."

Justification: An image shared on Facebook claims that Nancy Pelosi bought $1.25 million in Tesla stock the day before Biden signed an order for all federal vehicles to be electric, implying that she sought to profit from inside information about new government policies. The House speaker did report transactions involving Tesla stock, but the post misrepresented the purchases and Biden's policies to create the false impression that the transactions represented improper insider trading in Tesla shares.

Annotation: Question	Answer	Question Source	
Were the stock purchases improper insider trading?	No	Claim ○	Justification ●
Does the executive order Biden signed require all federal vehicles to be electric?	Unknown	Claim ●	Justification ○
Did Nancy Pelosi buy 1.25 million Tesla stock the day before Joe Biden signed an order about electric vehicles?	Unknown	Claim ●	Justification ○

Figure 2: An example of our annotation process. The annotators are instructed to write a set of subquestions, give binary answers to them, and attribute them to a source. If the answer cannot be decided from the justification paragraph, "Unknown" is also an option. The question is either based on the claim or justification, and the annotators also select the relevant parts (color-coded in the figure) on which the question is based.

Politifact에서 수집한 claim과 justification (article)이 주어지면

1. yes-no 응답이 가능한 question 생성
2. 생성한 question에 대한 answer 기입 (yes, no, unknown)
3. 생성한 question의 출처 표기 (claim, justification)

Dataset Analysis 1

Question Type	# Questions	R1-P	R2-P	RL-P
Literal	2.15	0.56	0.30	0.47
Implied	1.02	0.28	0.09	0.22

Table 4: Number of questions of each type per claim and their lexical overlap with the claim measured by ROUGE-1, ROUGE-2, and ROUGE-L precision (how many n-grams in the question are also in the claim).

Domain knowledge (38.8%)	**Claim**: "When President Obama was elected, the market crashed … Trump was up 9%, President Obama was down 14.8% and President Bush was down almost 4%. There is an instant reaction on Wall Street." **Question**: Did Obama cause the stock market crash when he was elected? (**Domain knowledge of whether the stock market is correlated with the election.**)
Context (37.6%)	**Claim**: With voting by mail, "you get thousands and thousands of people … signing ballots all over the place." **Question**: Is there a greater risk of voting fraud with mail-in ballots? (**Need to know the background that the claim is about the potential risks of mail-in ballots.**)
Implicit meaning (16.5%)	**Claim**: Nancy Pelosi bought $1.25 million in Tesla stock the day before Joe Biden signed an order "for all federal vehicles" to be electric. **Question**: Were the stock purchases improper insider trading? (**The claim implies this purchase is insider trading.**)
Statistical rigor (7.1%)	**Claim:** "No other country witnesses the number of gun deaths that we do here in the U.S., and it's not even close." **Question:** Is the United States the country with the the highest percentage of gun deaths? (**Highest number of gun deaths does not entail highest percentage of gun deaths.**)

Figure 4: Four types of reasoning needed to address subquestions with their proportion (left column) and examples (right column). It shows that a high proportion of the questions need either domain knowledge or related context.

1. 285개의 sub-question에 대한 분석

- 1개의 claim당 평균적으로 2개의 Literal question과 1개의 Implied question 존재

2. Implied question type에 대한 분석
- Domain Knowledge: claim에 대한 domain-specific한 knowledge(politics, legal 등)가 담긴 question
- Context: claim의 background knowledge, context가 담긴 question
- Implicit meaning: claim에 암시된 내용을 가지고 생성된 question
- Statistical rigor: claim에 제시된 통계 수치에 대한 해석을 사용한 question

Dataset Analysis 2

Claim: The group With Honor stated on September 10, 2018 in a TV ad: Kentucky Rep. Andy Barr "would let shady payday lenders take advantage of our troops" and that he took "$36,550 from payday lenders."

CLAIMDECOMP	Fan et al. (2020)
1 Has Barr received $36,550 from payday lenders?	1 What are Payday lenders? **helpful background but not precisely about claim**
2 Did Barr vote for legislation that would weaken restrictions for payday lenders?	2 What's the maximum amount you can get from payday lenders? **useful context but not directly about claim**
3 Are there any protections for service members using payday lending services?	3 What percentage of US troops use a payday lender? **useful context but not directly about claim**
4 Has Barr's voting record directly affected protection for veterans against payday lenders?	

Figure 5: Comparison between our decomposed questions with QABriefs (Fan et al., 2020). In general, our decomposed questions are more comprehensive and relevant to the original claim.

Fact-checking claim에 대한 QG를 하는 QAbriefs 데이터셋과의 비교

- QAbriefs의 질문은 claim과 직접적인 상관이 없는 Question이 많음
- QAbreifs의 질문은 사실 그 자체에 관한 단순한 질문이 많음 (What 류 질문, 개념적 정의에 관한 질문 등)
- ClaimDecomp의 경우 질문들이 claim과의 연관성이 높고 추론을 요하는 implied question이 존재

☞ QAbriefs 보다 고도화된 ClaimDecomp의 우수성

Experimental setup

Split	# unique claims	# tokens per claim	avg. # subquestions in single annotation	Answer % Yes	Answer % No	Answer % Unknown	Source % Justification	Source % Claim
Train	800	33.4	2.7	48.9	45.3	5.8	83.6	16.4
Validation	200	33.8	2.7	48.3	44.8	6.9	79.0	21.0
Validation-sub	50	33.7	2.9	45.2	47.8	7.0	90.4	9.6
Test	200	33.2	2.7	45.8	43.1	11.1	92.1	7.9

Table 1: Statistics of the CLAIMDECOMP dataset. Each claim is annotated by two annotators, yielding a total of 6,555 subquestions. The second column blocks (Answer % and Source %) report the statistics at the subquestion level; Source % denotes the percentage of subquestions based on the text from the justification or the claim.

1. Automatic QG
2. Classification
3. Evidence Retrieval

ClaimDecomp 데이터셋을 활용하여 **3가지 Research Question(RQ)**에 대한 실험 진행

RQ 1: Can we **automatically generate** literal & implied sub-question using PLM?
→ Generation task

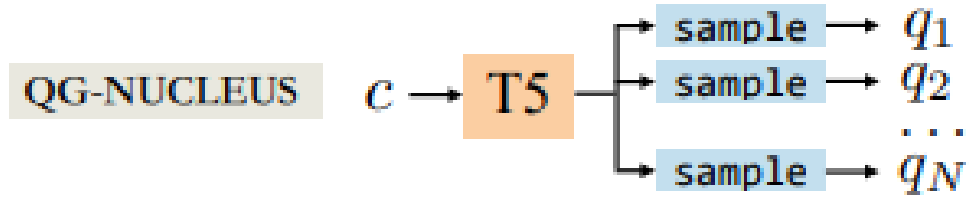
RQ 2: Whether answers to subquestions can be used to **determine the veracity of the claim**
→ classification task

RQ 3: Is claim decomposition (subquestions) **useful for evidence retrieving**?
→ retrieval task

Experimental result1: Automatic QG

QG-MULTIPLE $c, N \rightarrow$ T5 $\rightarrow q_1 [\mathrm{S}] q_2 [\mathrm{S}] \ldots q_N$

Model: T5-3B
Input: claim + all GT sub-questions
Output: predicted sub-question set



Model: T5-3B
Input: (claim, GT sub-question) pair
Output: (claim, predicted sub-question) pair

Model	R-all	R-literal	R-implied
QG-MULTIPLE	0.58	0.74	0.18
QG-NUCLEUS	0.43	0.59	0.11
QG-MULTIPLE-JUSTIFY	0.81	0.95	0.50
QG-NUCLEUS-JUSTIFY	0.52	0.72	0.18

Table 3: Human evaluation results on the Validation-sub set (N=146). R-all denotes the recall for all questions; R-literal and R-implied denotes the recall for the literal questions and the implied questions respectively.

Analysis

1. Literal vs Implied
- literal question에 generation에서는 높은 Recall이 도출되나
- Implied question generation에서는 낮은 Recall

2. Input Context
- Justification article을 input으로 추가했을 때 모두 성능 향상이 존재

3. Model configuration
- 모든 subquestion을 한번에 생성하는 MULTIPLE 방법이 NUCLEUS 방법보다 성능이 좋음 (특히 Implied QG 부분에서)

Experimental result2: Classification

Proposed methods	Macro-F1	Micro-F1	MAE
Question aggregation	0.30	0.29	1.05
Question aggregation*	0.46	0.45	0.73
Random (label dist)	0.16	0.18	1.68
Most frequent	0.06	0.23	1.31

Baseline

Table 6: Claim classification performance of our question aggregation baseline vs. several baselines on the development set. MAE denotes mean absolute error.

Sub-question 개별 yes-no 합이 최종 label prediction으로 이어질 수 있는지 실험

- 모델을 사용하는 것이 아니라 주어진 validation-sub 의 데이터셋을 그대로 사용할 때의 결과

최종 판별

$$\hat{v} = \frac{1}{N} \sum_{i=1}^{N} 1\,[a_i = 1]$$

Question aggregation *: 저자들이 임의로 판별에 관련 없는 sub-question 제외

Analysis

- 판별에 중요하지 않은 question을 제거하는 것은 예측에 도움
- 그럼에도 불구하고 단순 subquestions의 yes-no 합은 최종 판별과 동떨어져 있음

☞ Sub-question set가운데 중요도를 판단하는 것과 판별에 적절히 aggregation할 수 있는 후속 연구 필요

Experimental result3: Evidence Retrieval

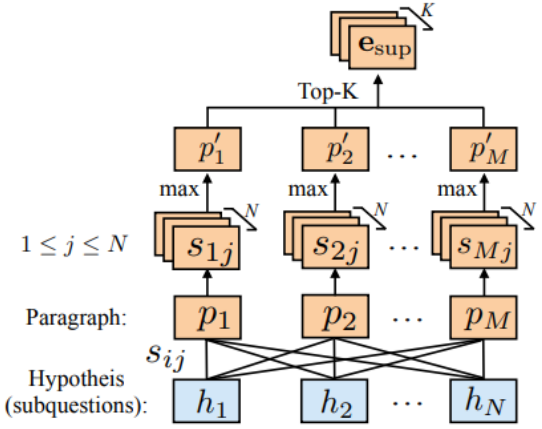


Figure 6: Illustration of evidence paragraph retrieval process. The notations corresponds to our descriptions in Section 6. K is a hyperparameter controlling the number of passages to retrieve.

1. Validation-sup에 대한 evidence human annotation 진행
2. NLI 모델을 활용하여 sub question을 support하는 evidence set retrieval
   1) preprocess: $q_N \xrightarrow{\text{GPT}-3} h_N$, 의문문을 평서문으로 변환
   2) NLI 모델을 이용하여 모든 h_N, p_M pair의 entailment probability 계산
   3) top-K entailment probability를 가지는 evidence 집합 e_{sup} 산출
3. NLI 모델을 활용하여 sub question을 refute하는 evidence set retrieval
4. support, refute별로 예측된 e_{sup} 의 합집합을 최종 예측 evidecne로 사용

Model	Decomposed claim predicted	Decomposed claim gold	Original claim
MNLI	41.0	48.8	35.2
NQ-NLI	38.8	34.5	40.9
DocNLI	**44.7**	**59.6**	36.9
BM25	36.2	47.5	39.2

Table 8: Evidence retrieval performance (F1 score) with the decomposed claims (from predicted and annotated (gold) subquestions) and the original claim on the Validation-sub set. A random baseline achieves 24.9 F1 and human annotators achieve 69.0 F1.

Analysis

- Sub-question을 사용하여 retrieval을 하는 것이 claim 단독 사용하여 retrieval하는 것보다 성능이 좋음 (DocNLI, BM25)

☞ QG가 현재의 fact-checking framework에 도입될 수 있는 가능성 암시

Conclusion & Insights

- Conclusion
  - claim을 여러 개의 sub-question으로 분해하는 claim decomposition task와 ClaimDecomp 데이터셋을 제안
  - ClaimDecomp 데이터셋을 이용하여 claim decomposition task의 중요성과 fact-checking system에서의 편입 가능성을 증명

- Insights
  - 짧은 길이의 문장에 여러 context가 포함되어 있는 claim을 n개의 sub-question으로 분해하는 새로운 task 제시
  - Introduction에서의 문제제기와 이를 설명하기 위한 figure가 일목요연하게 제시되어 introduction 구조 참고가 용이한 논문
  - Fact-checking에서의 QG 적용시 성능 개선의 가능성을 보여주는 논문
    - Evidence Retrieval에서의 성능 향상 실험

☞ QG가 결합된 새로운 Fact-checking framework 개발의 기반이 되는 연구

Thank you