国际人工智能会议
AAAI 2021 论文北京预讲会

Natural Language Inference over Context — Investigating Contextual Reasoning over Long Texts

Hanmeng Liu
Westlake University, Zhejiang University
liuhanmeng@zju.edu.cn

Contextual Reasoning is Crucial for Human Inference and Hard for Machines

Example:

P: Ten new television shows appeared during the month of September. Five of the shows were sitcoms, three were hour-long dramas, and two were news-magazine shows. By January, only seven of these new shows were still on the air. Five of the shows that remained were sitcoms.

H1: *At least one of the shows that were cancelled was an hour-long drama.*

Entailment ✓ Contradiction Neutral

H2: *There is no hour-long drama remained on the air.*

Entailment **Contradiction** ✓ Neutral

H3: *Television viewers prefer sitcoms over hour-long dramas.*

Entailment Contradiction **Neutral** ✓

Reasoning over long texts introduces new scenarios
Logical reasoning, especially.

A Dataset for Natural Language Inference Addressing Contextual Reasoning

NLI

Two texts:
- Premise p
- Hypothesis h

Three labels:
- Entailment
- Contradiction
- Neutral

ConTRoL

	ConTRoL
Construction Method	Exams
Context Type	Passage
# of passages	1,970
# of premise-hypothesis pairs	8,325
# of multi-paragraph	4,171
Avg. length of multi-paragraph	757
# of single-paragraph	4,154
Avg. length of single-paragraph	148
Vocab size (premise)	54,265
Vocab size (hypothesis)	14,323
Avg. premise length	452
Avg. hypothesis length	12
Lexical overlap (Entailment)	4.87%
Lexical overlap (Neutral)	4.19%
Lexical overlap (Contradiction)	5.49%

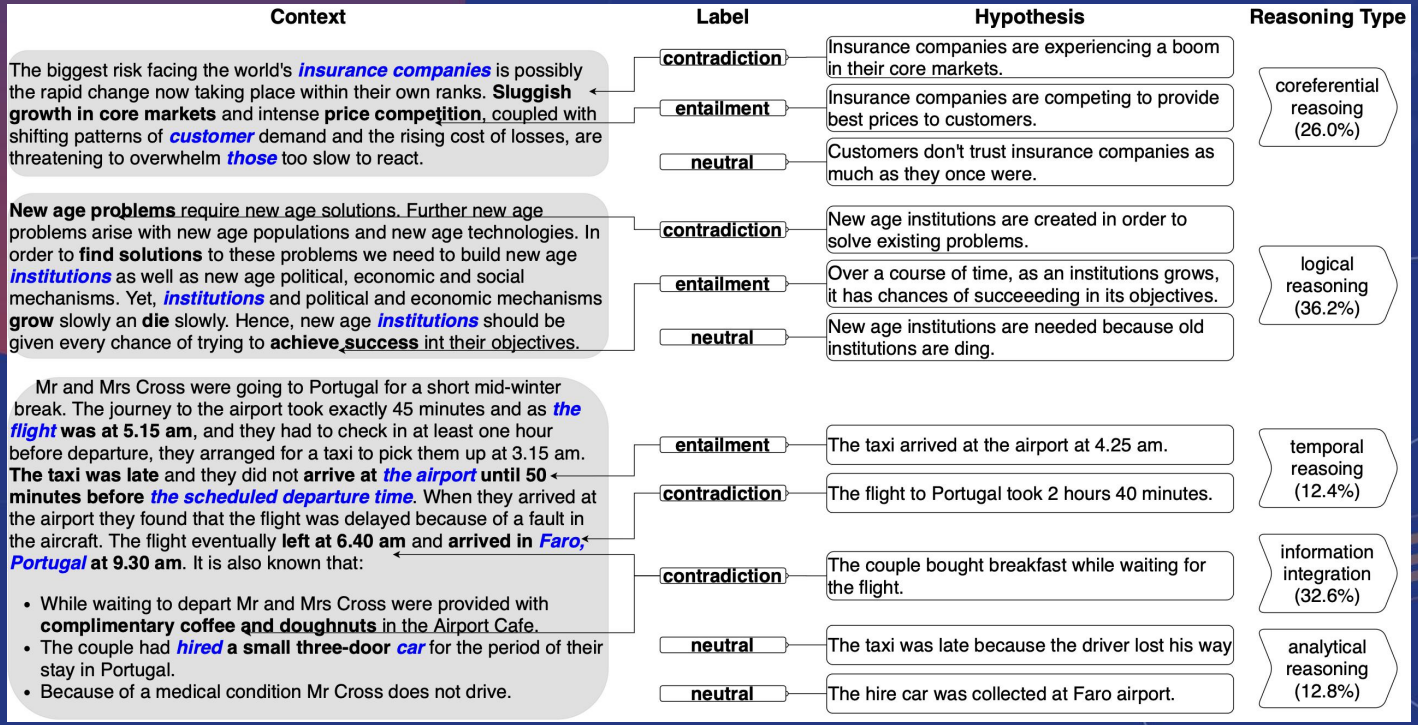
From sentence-level NLI to paragraph-level NLI (Reasoning Abilities)

Comparison

Dataset	Task	Reasoning	Context	Source
SQuAD (Rajpurkar et al. 2016)	Reading Comprehension	✓	Passage	Wikipedia
WIKIHOP (Welbl, Stenetorp, and Riedel 2017)	Reading Comprehension	✓	Document	Wikipedia
HOTPOTQA (Yang et al. 2018)	Reading Comprehension	✓	Document	Wikipedia
Cosmos QA (Huang et al. 2019)	Reading Comprehension	✓	Passage	Webblog
Social IQA (Sap et al. 2019)	Reading Comprehension	✗	Sentence	Social
WINOGRANDE (Sakaguchi et al. 2019)	Coreference Resolution	✗	Sentence	Diverse
CommonsenseQA (Talmor et al. 2019)	Reading Comprehension	✗	Sentence	Diverse
MuTual (Cui et al. 2020b)	Next Utterance Prediction	✓	Dialogue	Exam
ReClor (Yu et al. 2020)	Reading Comprehension	✓	Passage	Exam
LogiQA (Liu et al. 2020)	Reading Comprehension	✓	Passage	Exam
RTE (Dagan, Glickman, and Magnini 2005)	Natural Language Inference	✗	Sentence	Diverse
SNLI (Bowman et al. 2015)	Natural Language Inference	✗	Sentence	Captioning
WNLI (Wang et al. 2018)	Natural Language Inference	✗	Sentence	Fiction
QNLI (Wang et al. 2018)	Natural Language Inference	✗	Sentence	Wikipedia
MultiNLI (Williams, Nangia, and Bowman 2018)	Natural Language Inference	✗	Sentence	Diverse
Dialogue NLI (Welleck et al. 2018)	Natural Language Inference	✗	Sentence	Persona-chat
SciTaiL (Khot, Sabharwal, and Clark 2018a)	Natural Language Inference	✗	Sentence	Science
Adversarial NLI (Nie et al. 2019)	Natural Language Inference	✗	Paragraph	Diverse
AlphaNLI (Bhagavatula et al. 2019)	Natural Language Inference	✗	Sentence	Diverse
ConTRoL	Natural Language Inference	✓	Passage	Exam

Contextual Reasoning Explores Reasoning Types that are Difficult

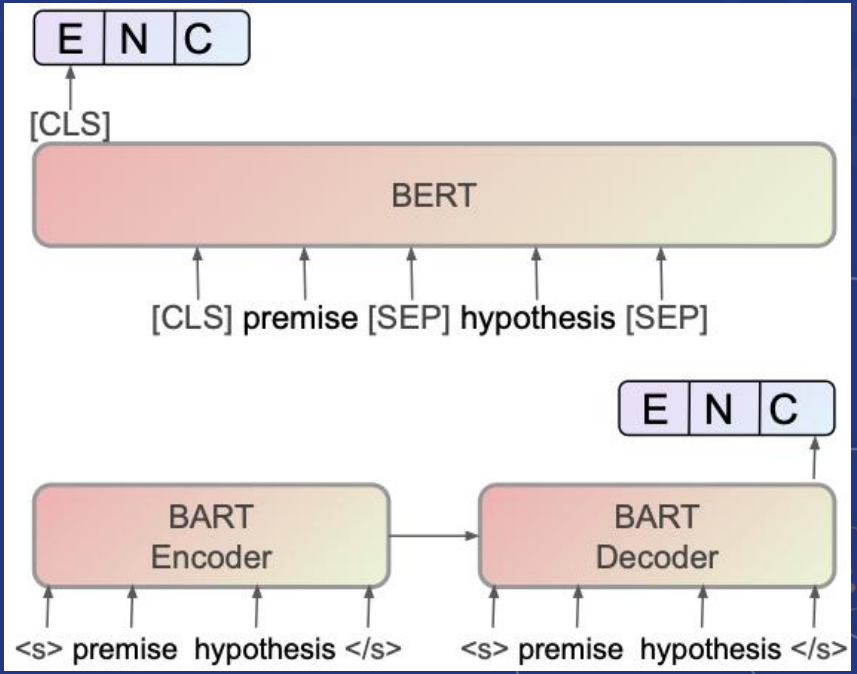
Reasoning types:



Pre-trained Language Models and the NLI Framework

Language models

- BERT
- RoBERTa
- XLNet
- Longformer
- BART



Models Performance on the ConTRoL dataset

Result

	Overall		Entailment			Neutral			Contradiction		
	Acc	Avg.F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Human	87.06	93.15	94.83	95.65	95.24	93.33	91.21	92.26	93.02	90.91	91.95
Ceiling	94.40	97.26	99.16	99.16	99.16	97.72	93.75	95.69	96.09	97.79	96.93
BERT-base	47.39	46.22	43.84	54.40	42.45	39.67	51.07	50.21	41.65	52.68	46.00
BERT-large	50.62	49.49	45.15	59.32	45.96	44.21	53.52	53.19	44.68	56.27	49.31
RoBERTa	45.90	45.67	40.99	51.24	45.38	**47.93**	44.34	45.96	44.19	47.54	45.67
Longformer	49.88	46.22	43.24	58.88	45.64	46.28	54.74	46.81	44.71	56.74	46.22
XLNet	54.85	**54.93**	46.15	62.22	**54.13**	47.11	59.94	55.74	46.63	61.06	**54.93**
BART	**56.34**	54.18	**50.23**	**67.32**	49.12	44.21	**62.99**	**59.57**	**47.03**	**65.09**	53.85
BART-NLI	45.02	42.33	39.85	53.49	40.87	43.80	46.79	43.83	41.73	49.92	42.30
BART-NLI-FT	60.95	57.41	62.58	61.54	58.67	42.15	78.29	56.17	50.37	68.91	57.39

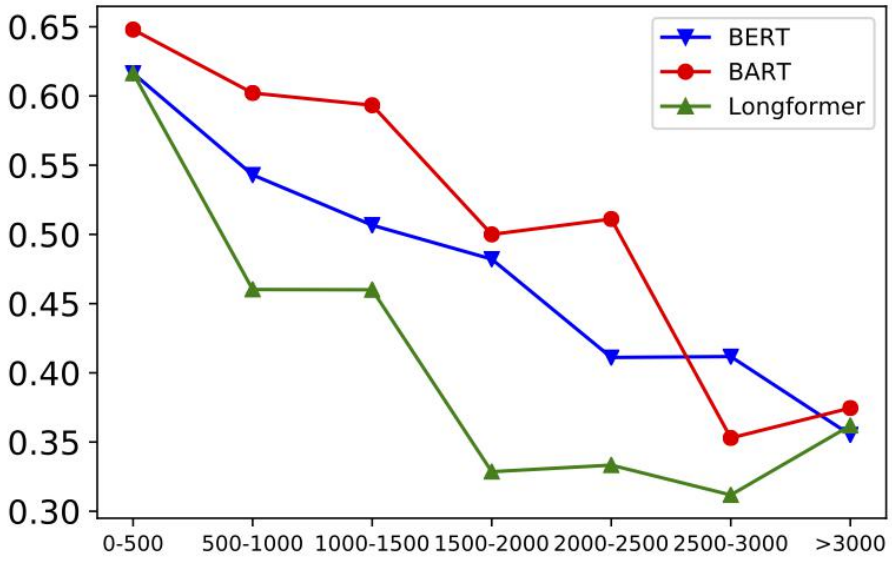
SOTA models' Huge Performance Drop Compared with Existing NLI datasets

Comparison

Benchmark	# Train	# Test	BERT	SOTA Model	SOTA Performance	Human
MultiNLI	393k	20k	85.9	T5-11B (Raffel et al. 2019)	92.0	92.8
QNLI	105k	5.4k	92.7	ALBERT (Lan et al. 2019)	99.2	91.2
RTE	2.5k	3k	70.1	T5-11B (Raffel et al. 2019)	92.5	93.6
WNLI	634	146	65.1	T5-11B (Raffel et al. 2019)	93.2	95.9
ConTRoL	8.3k	804	50.6	BART-NLI-FT	61.0	94.4

Far below human ceiling performance

Qualitative and Quantitative Detailed Analysis



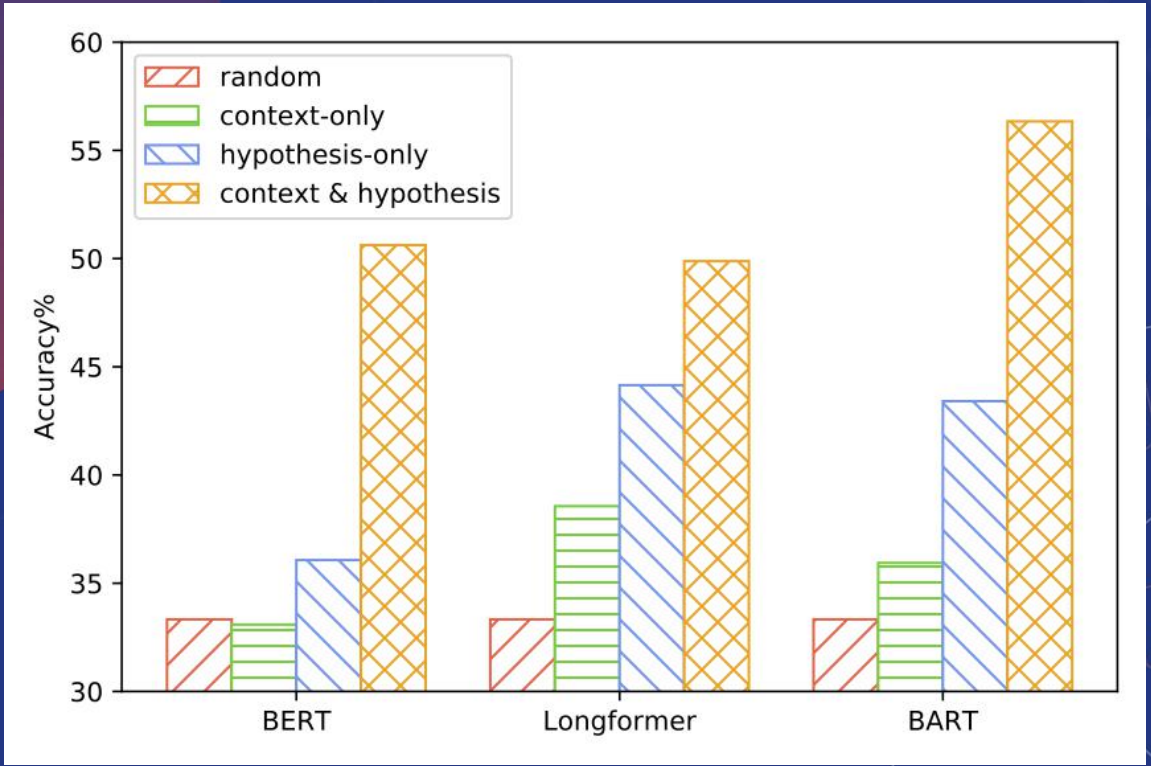
Across different context lengths

Reasoning Type	BERT	BART
Coreferential Reasoning	74.64	74.92
Analytical Reasoning	67.96	69.65
Temporal Reasoning	56.44	57.34
Information Integration	40.07	43.39
Logical Reasoning	40.76	43.20

Across different reasoning types

Corpus Bias

Ablation study



Case Study

The challenge in ConTRoL

In both cases, the correct answer is not explicitly mentioned in the premise, but need contextual reasoning to infer.

P1: Three athletes each receive a first, second and third prize for a different sporting event. Either *Anne* or ***Josie*** got the second prize for Tennis. *Anne* got the same prize for throwing the javelin as ***Josie*** got for swimming. *Tanya* got the first prize for swimming, and her prize for the javelin was the same as ***Josie***'s for tennis and *Anne*'s for swimming.

H1: *Josie was best with the javelin.*

Entailment **Contradiction** ✓ **Neutral** ✗

P2: Two masked gunmen held up ***the only bank in Tuisdale*** at 10.30 on Wednesday 23 May. ***They made a successful getaway with over 500,000***. The police say that three men are helping them with their enquiries. It is also known that: Four people work at the bank. Six customers were in the bank at 10.30. No shots were fired. Ms Grainger left the bank at 10.28 on Wednesday 23 May. All the people in the bank were made to lie on the floor face down on their stomachs. The police chased the getaway car for 16 km, and then lost it. An alarm alerted the police to the hold-up. A red Ford Mondeo drove away from the bank at high speed at 10.30 on Wednesday 23 May.

H1: *As a goodwill gesture, Tuisdales other bank provided emergency access to cash for customers after their ordeal.*

Entailment **Contradiction** ✓ **Neutral** ✗

国际人工智能会议
AAAI 2021 论文北京预讲会

THANKS

2020.12.19