Detecting Response Generation Not Requiring Factual Judgment

Ryohei Kamei[1]♣, Daiki Shiono[1], Reina Akama[1,2], Jun Suzuki[1,2]
[1] Tohoku University, [2] RIKEN ♣ ryohei.kamei.s4@dc.tohoku.ac.jp

Abstract

- Created a dataset annotated with 4 sentence types
- Exp.1: Classification task using some models, and the best model had an accuracy of 88%
- Exp.2: Classification models' accuracy improves as the number of training data is increased

Background

- Factuality of dialogue responses is an open issue
- Previous study to detect/reduce hallucinations, defined as responses not based on given knowledge[1]
- Our Goal: Achieve both attractiveness and factuality

Idea

- Expressing personal opinions and feelings is also crucial for dialogue systems
- Dialogue dataset annotated with a new label indicated whether a **factual correctness judgment was required**

Knowledge
On April 18, 2017, Facebook announced React Fiber, a new core algorithm of React framework library for building user interfaces.

Knowledge-based Dialogue Response
I did a little dabbing myself in web dev, it's really fun!
FaceBook also announced React Fiber, a new coree algorithm, you may want to check that out as well!
→ Existing Label: Hallucination

Sentence 1
I did a little dabbing myself in web dev,it's really fun!
→ Our Label: (iii)

Sentence 2
FaceBook also announced React Fiber, a new coree algorithm, you may want to check that out as well!
→ Our Label: (iv)

Dataset Construction

1. Wizard of Wikipedia[2] responses are split into sentences
2. Four labels were annotated by three Yes/No questions

	Label Explanation	# of sample
(i)	agreement, disagreement, interjections	141 (10.7%)
(ii)	suggestions, advice	110 (8.4%)
(iii)	subjective opinions, personal thoughts	540 (41.0%)
(iv)	objective information	526 (39.9%)

Experiments

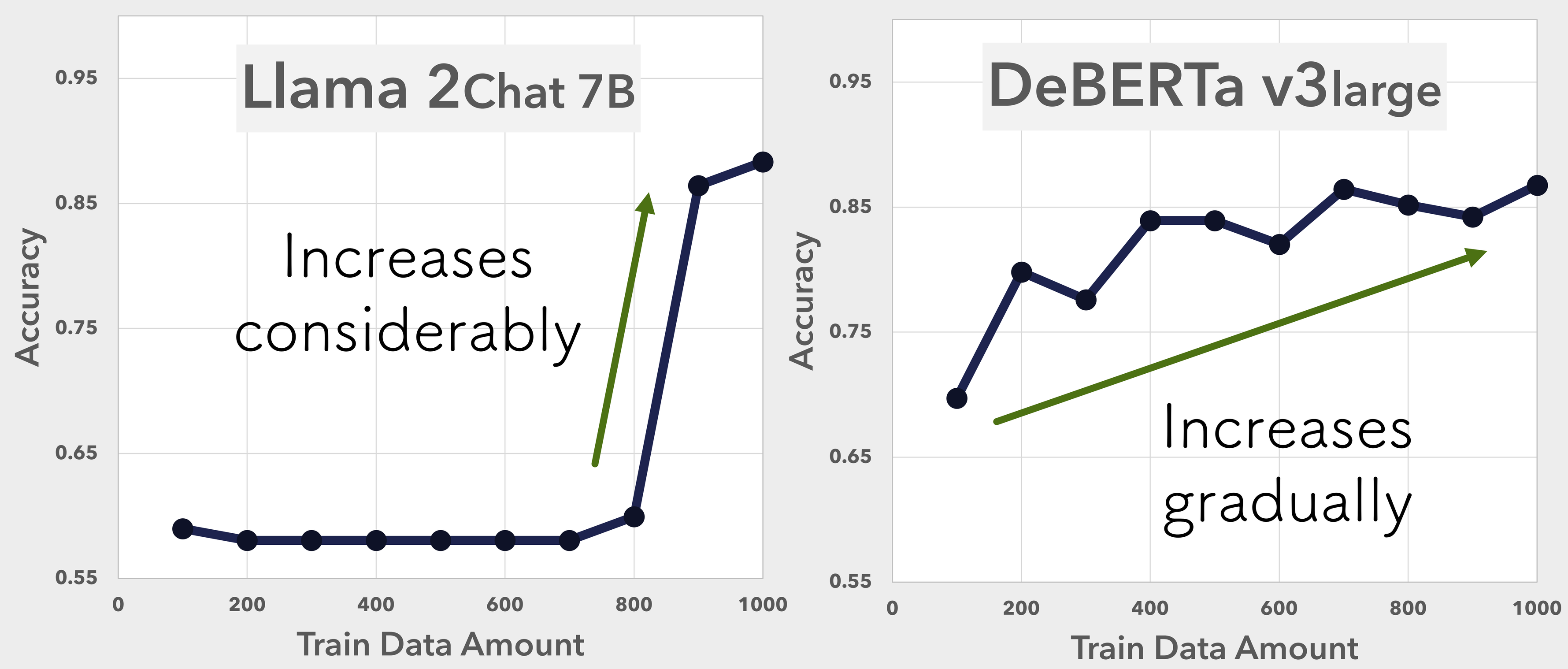
Exp.1 Detecting response Not requiring factual judgement

Models: GPT-4, Llama 2$_{\text{Chat 7B}}$, DeBERTa v3$_{\text{large}}$, etc.
Metrics: Accuracy, Precision, Recall, F1-Score

model	architecture	fine-tuning	Accuracy	Precision	Recall	F1-Score
GPT-3.5	decoder	✗	57.73	58.17	96.74	72.65
GPT-4	decoder	✗	57.73	58.99	89.13	71.00
Llama 2$_{\text{Chat 7B}}$	decoder	✗	58.99	58.60	**100.0**	73.9
Llama 2$_{\text{Chat 7B}}$	decoder	✓	**88.33**	**91.53**	88.04	**89.75**
DeBERTa v3$_{\text{large}}$	encoder	✓	86.75	85.83	81.95	83.85
RoBERTa$_{\text{large}}$	encoder	✓	84.23	87.39	72.93	79.51
BERT$_{\text{large}}$	encoder	✓	83.28	80.77	78.95	79.85

- ✅ Highest accuracy by Llama 2$_{\text{Chat 7B}}$ with fine-tuning
- ❌ Decoder models without fine-tuning almost predict "Not Requiring Factual Judgment"
- ✅ Encoder models with fine-tuning have higher Precision and slightly lower Recall

Exp.2 Relation between train data amount and accuracy



- ✅ Further accuracy is expected using more data

Examples: predicted wrongly when the amount of train data is small but predicted correctly when it is large

1) It was first documented all the way back to 1481.
2) Toews is great, he was the third overall pick in the 2006 NHL draft.

- ✅ By training, model can predict "factual judgment is required" when proper nouns or dates are present

Future Work

- Collect large-scale data and improve the performance of classification models
- Clarify the reason for the sudden increase in accuracy when the number of training data exceeds 800
- Apply classification models to dialogue systems

[1] Dziri+, FaithDial: A Faithful Benchmark for Information-Seeking Dialogue(TACL 2022) [2] Dinan+, Wizard of Wikipedia: Knowledge-Powered Conversational Agents(ICLR 2019)