Two Reproductions of *Language Model as an Annotator: Exploring DialoGPT for Dialogue Summarization*

•••

HumEval Workshop 21. 5. 2024

The Original Experiment (Feng et al., 2021)

- Dialogue Summarization
- Can DialoGPT be used as an annotator to improve the quality of summaries?
- Baselines:
    - Pointer-Generator Networks (PGN)(See et al., 2017)
    - Hierarchical Meeting summarization Network (HMNet)(Zhu et al., 2020)
- Proposed Systems:
    - PGN + DialoGPT for keyword extraction (D_{KE})
    - PGN + DialoGPT for redundancy detection (D_{RD})
    - PGN + DialoGPT for topic segmentation (D_{TS})
    - PGN + all of the above annotations (D_{ALL})
- + a human-written reference

The Human Evaluation

Original

- SAMSum and AMI datasets
- Evaluated Qualities
  - Informativeness (1-5)
  - Consciseness (1-5)
  - Coverage (1-5)
  - Overall Impression (1/0)
- Chinese NLP PhD students as evaluators
- Terminal annotation interface
- $10 per annotator

ReproHum

- AMI dataset
- Evaluated Qualities
  - Informativeness (1-5)
- PhD students with various L1s and backgrounds
- Google Form
- Approx. $115 per annotator

Three Additional Reproductions

Repro #1

- Computer Science
- BSc degree
- Chinese as L1
- Evaluated only informativeness

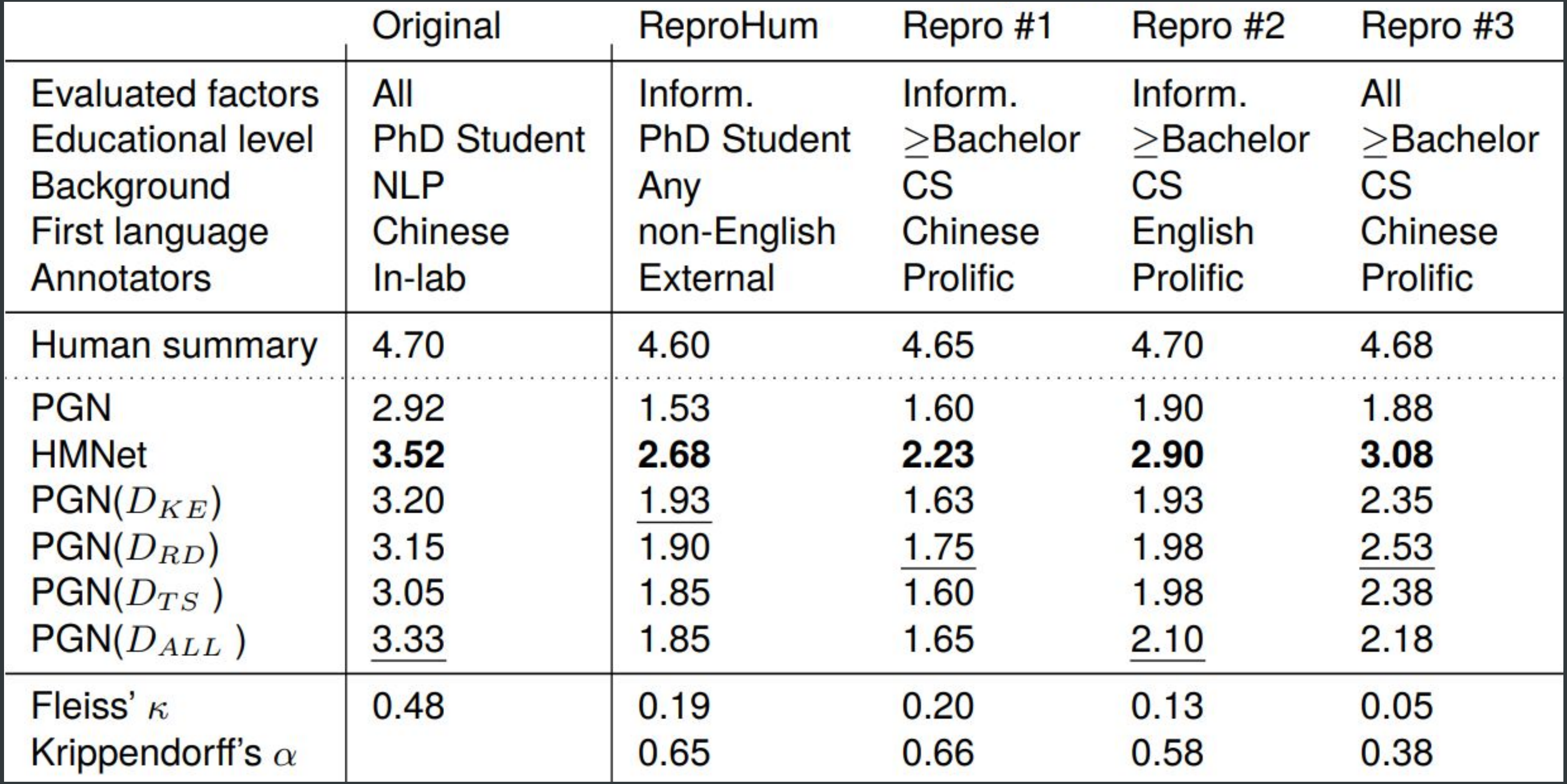
Repro #2

- Computer Science
- BSc degree
- English as L1
- Evaluated only informativeness

Repro #3

- Computer Science
- BSc degree
- Chinese as L1
- Evaluated all four criteria

Results



	Original	ReproHum	Repro #1	Repro #2	Repro #3
Evaluated factors	All	Inform.	Inform.	Inform.	All
Educational level	PhD Student	PhD Student	$\geq$Bachelor	$\geq$Bachelor	$\geq$Bachelor
Background	NLP	Any	CS	CS	CS
First language	Chinese	non-English	Chinese	English	Chinese
Annotators	In-lab	External	Prolific	Prolific	Prolific
Human summary	4.70	4.60	4.65	4.70	4.68
PGN	2.92	1.53	1.60	1.90	1.88
HMNet	**3.52**	**2.68**	**2.23**	**2.90**	**3.08**
PGN(D_{KE})	3.20	1.93	1.63	1.93	2.35
PGN(D_{RD})	3.15	1.90	1.75	1.98	2.53
PGN(D_{TS})	3.05	1.85	1.60	1.98	2.38
PGN(D_{ALL})	3.33	1.85	1.65	2.10	2.18
Fleiss' κ	0.48	0.19	0.20	0.13	0.05
Krippendorff's α		0.65	0.66	0.58	0.38

Summary

1. *"HMNet gets the best score in informativeness and coverage"*, which was **confirmed by our reproductions**.
2. *"Our method can achieve higher scores in all three metrics"*, which is **in line with** the results of **our reproductions**.
3. *"We also find there is still a gap between the scores of generated summaries and the scores of golden summaries"* – the gap seems **substantially larger than in the original study**.
4. **L1, level of education or field of study do not seem to have a significant impact on the results** of human evaluation in the summarisation task.

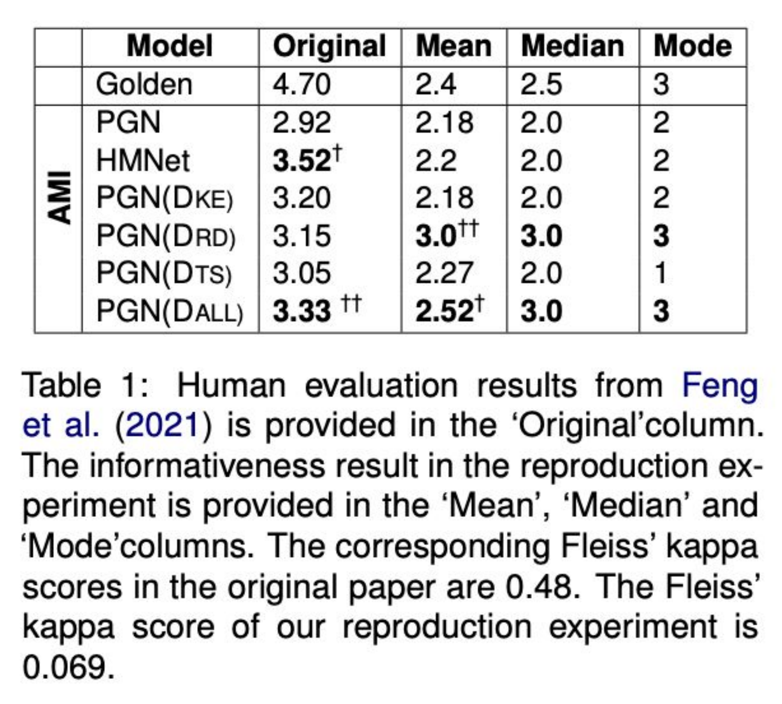
Reproduction by NLLG & University of Mannheim

One Additional Reproductions

Reproduction

- Phd degree
- Chinese as L1
- High English proficiency
- Evaluated only informativeness

Results



	Model	Original	Mean	Median	Mode
	Golden	4.70	2.4	2.5	3
AMI	PGN	2.92	2.18	2.0	2
	HMNet	**3.52**†	2.2	2.0	2
	PGN(DKE)	3.20	2.18	2.0	2
	PGN(DRD)	3.15	**3.0**††	**3.0**	**3**
	PGN(DTS)	3.05	2.27	2.0	1
	PGN(DALL)	**3.33** ††	**2.52**†	**3.0**	**3**

Table 1: Human evaluation results from Feng et al. (2021) is provided in the 'Original'column. The informativeness result in the reproduction experiment is provided in the 'Mean', 'Median' and 'Mode'columns. The corresponding Fleiss' kappa scores in the original paper are 0.48. The Fleiss' kappa score of our reproduction experiment is 0.069.

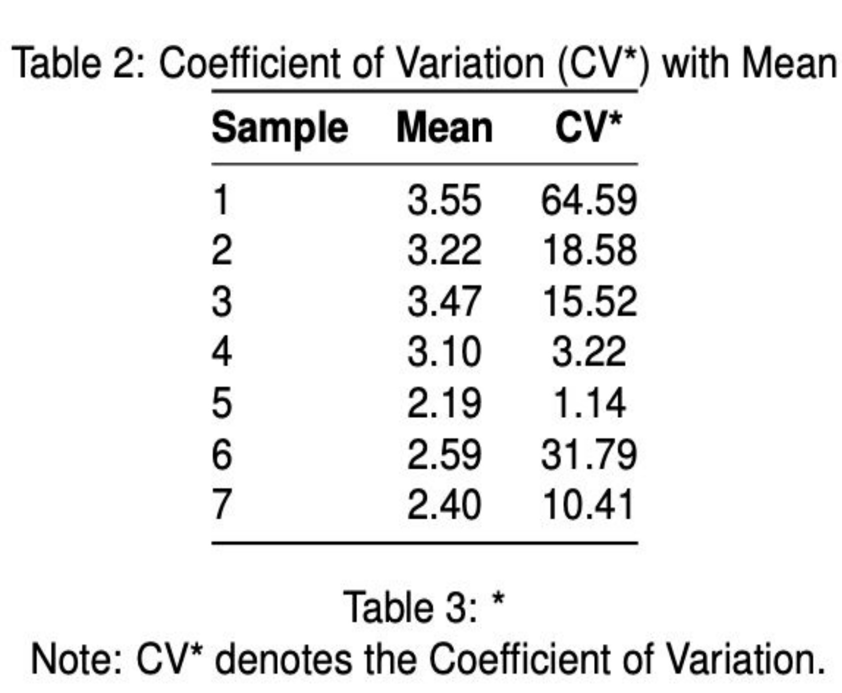


Table 2: Coefficient of Variation (CV*) with Mean

Sample	Mean	CV*
1	3.55	64.59
2	3.22	18.58
3	3.47	15.52
4	3.10	3.22
5	2.19	1.14
6	2.59	31.79
7	2.40	10.41

Table 3: *

Note: CV* denotes the Coefficient of Variation.

Summary

1. *"HMNet gets the best score in informativeness and coverage"*, which **couldn't be confirmed by our reproduction**.
2. *"Our method can achieve higher scores in all three metrics"*, which **is not in line with the results of our reproduction**.
3. *"We also find there is still a gap between the scores of generated summaries and the scores of golden summaries"* – **the gap seems substantially larger than in the original study.**
4. In our reproduction study, the inter-annotator agreement was notably lower
5. We were unable to confirm the effectiveness of the proposed approach in terms of informativeness.

Thank you and see you at the poster session!

Charles University

lango

schmidtova

balloccu

odusek

@ufal.mff.cuni.cz

Univ. of Mannheim

vivian.fresen
@adesso.de

ms.wuurbanek
@gmail.com

Steffen.eger @
uni-mannheim.de

erc
European Research Council
Established by the European Commission

This research was co-funded by the European Union (ERC, NG-NLG, 101039303) and by Charles University projects GAUK 40222 and SVV 260 698.

NLLG

adesso

This research was co-funded by the BMBF, via the grant "Metrics4NLG"

SPONSORED BY THE
Federal Ministry
of Education
and Research

References

- Xiachong Feng, Xiaocheng Feng, Libo Qin, Bing Qin, and Ting Liu. 2021. Language model as an annotator: Exploring DialoGPT for dialogue summarization. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 1479–1491, Online. Association for Computational Linguistics.
- Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.
- Chenguang Zhu, Ruochen Xu, Michael Zeng, and Xuedong Huang. 2020. A hierarchical network for abstractive meeting summarization with crossdomain pretraining. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 194–203, Online. Association for Computational Linguistics.