How important is the background of your annotators?

Evaluating Summarization Models: Investigating the Impact of Education and Language Proficiency on Reproducibility

Mateusz Lango Patrícia Schmidtová Simone Balloccu Ondřej Dušek
{lango, schmidtova, balloccu, odusek}@ufal.mff.cuni.cz

CHARLES UNIVERSITY

ÚFAL

Original Experiment

- Feng et al. “Language Model as an Annotator: Exploring DialoGPT for Dialogue Summarization”, ACL 2021
- Dialogue Summarization task
- Can annotations by DialoGPT 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
- Dataset: AMI and SAMSum (not included in the reproduction)

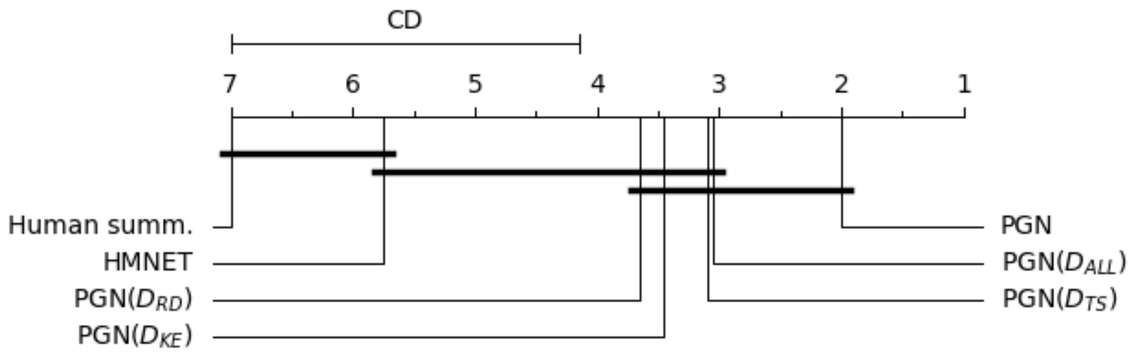
	Original	ReproHum	Repro #1	Repro #2	Repro #3
Evaluated factors	All	Inform.	Inform.	Inform.	All
First language	Chinese	non-English	Chinese	English	Chinese
Educational level	PhD Student	PhD Student	at least Bachelor degree		
Background	NLP	Any	Computer Science		
Annotators	In-lab	External	Prolific crowdsourcing platform		
Interface	Text file/bash	Google Forms			
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	<u>1.93</u>	1.63	1.93	2.35
PGN(D_{RD})	3.15	1.90	<u>1.75</u>	1.98	<u>2.53</u>
PGN(D_{TS})	3.05	1.85	1.60	1.98	2.38
PGN(D_{ALL})	<u>3.33</u>	1.85	1.65	<u>2.10</u>	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

Main Results

- Human reference got very consistent scores, **automatic summaries got lower scores** in all reproductions
- The author's extensions are ranked between the PGN baseline and HMNet, but their **relative rankings vary greatly** (small std usually 0.04-0.07)
- Differences between baseline PGN and all extensions are **not statistically significant**
- Different selection of annotators does not lead to significant differences, **evaluation of all quality factors gives most similar results**
- The inter-annotator **agreement is much lower in the reproduced studies** compared to the original experiment

Assessing reproducibility

	Pearson	Spearman	RMSE
ReproHum	**0.99**	0.85	1.16
Repro #1	0.98	**0.88**	1.35
Repro #2	0.98	**0.88**	1.00
Repro #3	0.97	0.68	**0.77**



Summary

- The main claims of the original study were confirmed by our reproductions:
  - *"HMNet gets the best score in informativeness and coverage"*, which was **confirmed by our reproductions**.
  - *"Our method can achieve higher scores in all three metrics”*, which is **in line with** the results of **our reproductions**.
  - *"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**.
- **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.
- **Evaluating all quality factors from the original study seems influential** while designing reproduction studies

Presented at HumEval co-located with LREC-COLING 2024, Torino, Italy

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

European Research Council
Established by the European Commission