Abstractive summarization of fact check reports with pre-trained transformer tuning on extractive summaries

Author: Ing. Peter Vajdečka Supervisor: prof. Ing. Vojtěch Svátek, Dr.

Motivation and goals

Fact-checking reports are frequently too long for a casual reader, and contain auxiliary parts not directly relevant for judging the claim veracity. Automated creation of fact check report summaries is thus a topical task.

Goals:

- ☐ Review modern approaches to automated text summarization and identify methods suitable for generating a summary of a fact check report as a particular kind of document.
- ☐ Gather data from fact-checking sites, in particular, demagog.cz for Czech and politifact.com for English (scraped).
- ☐ Propose hybrid summarization model to be comparable with state-of-the-art summarizing models in the fact-checking domain.

Related works

Our proposed method was compared to the following works in terms of the ROUGE metric (or ROUGE RAW – a language independent metric):

On the Politifact dataset:

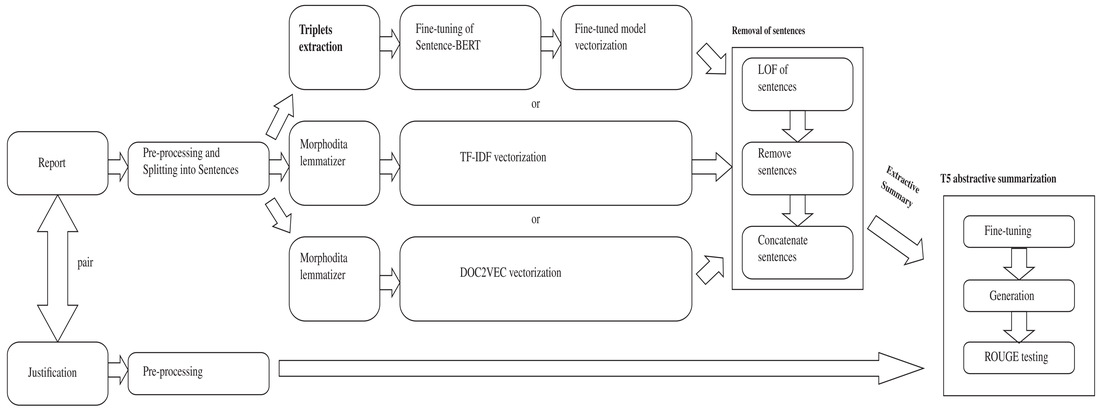
- ☐ Atanasova et al. [1] explored a supervised BERT-based technique for jointly predicting the truth of a claim and extracting supporting explanations from fact-checked claims.
- ☐ Kazemi et al. [2] utilized *GPT-2* for abstractive summarization and Biased TextRank for extractive summarization as alternative approaches.

On the SumeCzech dataset (not fact-checking focused, but the only summarization dataset available for Czech):

- ☐ Straka et al. [3] applied different extractive techniques and proposed the ROUGE RAW metric for the Czech language.
- ☐ Marek et al. [4] came up with named entities summarization, even improving results of [3].

Proposed hybrid method

The core contribution of our hybrid approach is extractive summarization based on Local outlier factor (LOF) and sentence representation by fine-tuned Sentence-BERT. The input pair consists in a fact check report and a justification (as its manually created summary). Sentence-splitting is applied to the report, and embeddings of separated sentences are created using alternative methods such as TF-IDF, DOC2VEC or Sentence-BERT; it was fine-tuned Sentence-BERT that returned the best vector representation. We then compute the normalized LOF. Sentences above a certain LOF threshold (optimized during the experiments) are removed, and the remaining ones become an extractive summary of the report, which together with its corresponding justification enters the T5 transformer to generate abstractive summaries.



Evaluation results

Source	System	ROUGE 1	ROUGE 2	ROUGE L
Atanasova 2020 (University of Copenhagen)	Explain-Extractive	35.7	13.51	31.58
	Explain-MT	35.13	12.9	30.93
Kazemi 2021 (University of Michigan)	TextRank	27.74	7.42	23.24
	GPT-2	24.01	5.78	21.15
	Biased TextRank	30.90	10.39	26.22
Present work	T5 Baseline	38.12	18.90	35.71
	SBERT+ LOF+T5 (13 % of sentences removed)	38.35	18.88	35.88
	Claim + T5 Baseline	39.19	20.56	36.92
	CLAIM + SBERT+LOF+T5 (13 % of sentences removed)	39.45	21.08	37.27
	CLAIM + SBERT+LOF+T5 (11 % of sentences removed)	39.76	21.37	37.54
	CLAIM + SBERT fine-tuned +LOF+T5 (13 % of sentences removed)	**40.76**	**22.00**	**38.36**
	CLAIM + SBERT fine-tuned +LOF+T5 (11 % of sentences removed)	39.55	20.69	37.11
	CLAIM + Morphodita + TF-IDF+LOF+T5 (13 % of sentences removed)	39.91	20.62	37.40
	CLAIM +Morphodita + TF-IDF+LOF+T5 (11 % of sentences removed)	39.86	20.59	37.30
	CLAIM + Morphodita + DOC2VEC+LOF+T5 (13 % of sentences removed)	38.58	19.62	36.20
	CLAIM + Morphodita + DOC2VEC+LOF+T5 (11 % of sentences removed)	39.04	20.65	36.70

Table 1: Politifact results

System	Test set								
	ROUGE RAW 1			ROUGE RAW 2			ROUGE RAW L		
	P	R	F	P	R	F	P	R	F
T5 Baseline	31.10	17.84	21.53	11.38	6.54	7.83	24.78	14.42	17.29
Claim + T5 Baseline	31.16	18.35	22.08	11.80	6.79	8.23	24.80	14.86	17.73
Claim + SBERT + LOF + T5 (24 % of sentences removed)	31.95	17.33	21.43	12.01	6.31	7.82	25.30	13.85	17.04
Claim + SBERT fine-tuned + LOF + T5 (24 % of sentences removed)	**32.73**	18.75	22.66	**12.97**	7.23	**8.82**	**26.29**	15.11	**18.25**
Claim + TF-IDF + LOF + T5 (24 % of sentences removed)	30.58	**19.92**	**23.08**	11.70	**7.51**	8.74	24.03	**15.82**	18.24
Claim + DOC2VEC + LOF + T5 (24 % of sentences removed)	31.41	16.89	20.82	11.50	6.06	7.49	25.29	13.78	16.89

Table 2: Demagog results (all for variants of present work)

Text → Headline		Test set									Out-of-domain test set								
Source	System	ROUGE RAW 1			ROUGE RAW 2			ROUGE RAW L			ROUGE RAW 1			ROUGE RAW 2			ROUGE RAW L		
		P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
SumeCzech [3]	first	7.4	13.5	8.9	1.1	2.2	1.3	6.5	11.7	7.7	6.7	13.6	8.3	1.3	2.8	1.6	5.9	12.0	7.4
	random	5.9	10.3	6.9	0.5	1.0	0.6	5.2	8.9	6.0	5.2	10.0	6.3	0.6	1.4	0.8	4.6	8.9	5.6
	textrank	6.0	**16.5**	8.3	0.8	2.3	1.1	5.0	**13.8**	6.9	5.8	**16.9**	8.1	1.1	3.4	1.5	5.0	**14.5**	6.9
	tensor2tensor	8.8	7.0	7.5	0.8	0.6	0.7	8.1	6.5	7.0	6.3	5.1	5.5	0.5	0.4	0.4	5.9	4.8	5.1
Named entities [4]	Seq2Seq	16.1	14.1	14.6	2.5	2.1	2.2	14.6	12.8	13.2	13.1	11.8	12	2	1.7	1.8	12.1	11	11.2
	Seq2Seq–NER	**16.2**	14.1	**14.7**	2.5	2.1	2.2	**14.7**	12.8	**13.3**	13.7	11.9	12.4	2	1.7	1.8	12.6	11.1	11.4
Present work (only 10 % of training data)	T5	15.4	11.0	12.5	3.2	2.3	2.6	14.2	10.1	11.5	15.9	11.9	13.2	4.4	3.2	3.6	14.9	11.2	12.4
	T5-SBERT-LOF (16 % of sentences removed)	15.8	11.4	12.9	**3.5**	**2.5**	**2.8**	14.6	10.6	11.9	**16.5**	12.4	**13.7**	**4.8**	**3.5**	**3.9**	**15.4**	11.6	**12.9**

Table 3: SumeCzech results

Conclusion

- ☐ On the ROUGE metrics, results for the proposed hybrid summarization approach outperform previous studies [1, 2, 3, 4] on all three data sets (see Tables 1, 2 and 3).
- ☐ We plan to extend the experiments to other domains and datasets (different from fact-checking) and aim to improve the ranking of input sentences by applying other features beyond the LOF score.

References

[1] Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. Generating fact checking explanations. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7352–7364. Association for Computational Linguistics, 2020.

[2] Ashkan Kazemi, Zehua Li, Verónica Pérez-Rosas, and Rada Mihalcea. Extractive and abstractive explanations for fact-checking and evaluation of news. In *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 45–50, Online, June 2021. Association for Computational Linguistics.

[3] Milan Straka, Nikita Mediankin, Tom Kocmi, Zdeněk Žabokrtskỳ, Vojtěch Hudeček, and Jan Hajic. Sumeczech: Large czech news-based summarization dataset. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[4] Petr Marek, Štěpán Müller, Jakub Konrád, Petr Lorenc, Jan Pichl, and Jan Šedivỳ. Text summarization of czech news articles using named entities. *arXiv preprint arXiv:2104.10454*, 2021.