23rd Nordic Conference on Computational Linguistics (NoDaLiDa 2021)

Error Analysis of using BART for Multi-Document Summarization: A Study for English and German Language

Timo Johner, Abhik Jana, and Chris Biemann

Language Technology Group, Dept. of Informatics, Universität Hamburg, Germany

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Motivation

Recent research using pre-trained language models for multi-document summarization tasks have shown great potential for summarization.

...But lacks a deep investigation of potential erroneous cases and their possible application in languages beyond English

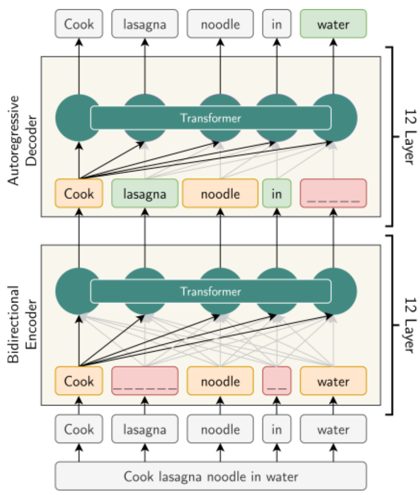
Approach

1. Reproduce — 2. Fine-Tune — 3. Analyse

- Reproduce recent pre-trained and fine-tuned results for multi-document summarization with the BART model, introduced by Lewis et al. (2020), on two English datasets.
- Adapt the model for German language by fine-tuning on a German MDS dataset, achieving state-of-the-art results with a margin of 3.48 – 8.67%.
- Analyse erroneous cases and cross-lingual error similarities for both languages regarding factfulness and topic delimitation.
- Investigate extractiveness of generated summaries.

Model

- BART model (Lewis et al., 2020)



- Generalizes concepts of bidirectional encoders from BERT (Devlin et al., 2019) and autoregressive decoders from GPT-2 (Radford et al., 2019).
- Combine text generation and contextual embeddings
- Apply the SDS approach to MDS
- Transformation through token masking, token deletion, text infilling, sentence permutation and document rotation.

- Make use of the pre-trained BART model and fine-tune the model on the three different datasets
- For MDS: merging multiple source documents to one single source document
- Remove duplicates through n-gram blocking

Datasets

1. CNN/DailyMail (Hermann et al., 2015):
   - single-document summarization news dataset
   - 311,971 news articles (~800 words on avg.)
   - Abstractive summaries
2. Multi-News (Fabbri et al., 2019):
   - Multi-document summarization news dataset
   - 250,000 news articles (~2,100 words on avg.)
   - 56,216 summaries with 2-10 source documents
3. auto-hMDS (Zopf, 2018):
   - Largest german dataset for multi-document summarization
   - 10,454 articles with different topics
   - 2,210 summaries with (4,73 sources on avg.)

Experimental Results

- Results on the CNN/DM, Multi-News and auto-hMDS dataset (top to bottom).
- The fine-tuned BART model achieves results comparable to the baselines for the CNN/DM dataset
- The fine-tuned BART model on Multi-News produces comparable results but takes more source documents into account

Method	R-1	R-2	R-L
LEAD-3 (Liu and Lapata)	40.42	17.62	36.67
BERTSUMABS (Liu and Lapata)	41.72	19.39	38.76
BERTSUMEXTABS (Liu and Lapata)	42.13	**19.60**	**39.18**
BART pre-trained	25.98	11.26	17.50
BART fine-tuned	**42.21**	19.10	35.38

Method	R-1	R-2	R-L
HI-MAP (Fabbri et al.)	40.08	14.90	19.70
BART DYNE-1 (Hokamp et al.)	**43.90**	**15.80**	**22.20**
BART DYNE-5 (Hokamp et al.)	43.20	13.60	20.40
BART pre-trained	30.67	10.05	16.99
BART fine-tuned	40.58	15.50	21.73

	100 words		200 words	
Method	R-1	R-2	R-1	R-2
RANDOM (Zopf)	18.57	1.85	25.53	3.25
LEAD (Zopf)	12.29	2.61	10.56	2.28
TOP-5 SENTENCES	21.71	4.28	19.61	3.87
LEXRANK	29.76	6.58	23.81	5.61
BART pre-trained	28.48	8.79	20.84	6.02
BART fine-tuned	**38.43**	**12.93**	**30.24**	**9.09**

➔ The fine-tuned BART model on auto-hMDS produces a state-of-the-art performance for German MDS with 38.43 (R-1), 12.93 (R-2) for 100 words and 30.24 (R-1), 9.09 (R-2) for 200 words

Example: Erroneous Case

- The table shows erroneous summaries based on the Multi-News dataset (Top) and the auto-hMDS dataset (Bottom).
- The model produces coherent summaries that tend to produce made-up and inaccurate facts.
- "who died in 2013"?

Summary (model generated) R-1 = 67.59, R-2 = 29.91, R-L = 31.41
[...] The former James Bond star, 65, who was trained as a commercial artist and worked as an illustrator, just auctioned off one of his paintings for $1.4 million, depicting the singer, who died in 2013. Other auction highlights included a Pierce Brosnam original painting, which sold for

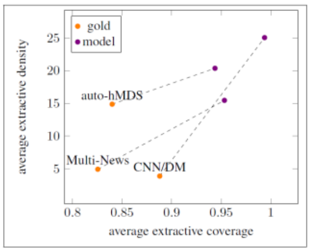
Summary (model generated) R-1 = 55.88, R-2 = 11.94, R-L = 30.88
Andrew Johnson (* 29. Dezember 1808 in Raleigh (North Carolina, USA; † 15. April 1865 in Greeneville, Tennessee) war der dritte Vizepraesident der Vereinigten Staaten, der durch den Tod seines Vorgaengers ins Amt kam und der erste nach einem Attentat. Als Hauptaufgabe seiner Praesidentschaft galt die sogenannte Reconstruction, der Wiederaufbau [...]

Extractiveness

- Measure extractiveness based on extractive coverage and extractive density (Grusky et al., 2018)
- Summaries are mainly built from extractive fragments or even whole paragraphs

$$\textsc{Coverage}(A, S) = \frac{1}{|S|} \sum_{f \in \mathcal{F}(A,S)} |f|$$

$$\textsc{Density}(A, S) = \frac{1}{|S|} \sum_{f \in \mathcal{F}(A,S)} |f|^2$$



Conclusion

- First attempt to use BART for German MDS
- Achieve SOTA for German MDS
- Analyse erroneous cases and extractiveness of BART cross-lingual on the different datasets
- Give impulse on further improvement regarding MDS

References

- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, USA.
- Alexander Richard Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-news: A largescale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084, Florence, Italy.
- Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana, USA.
- Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustaf Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *In Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 1*, pages 1693–1701, Montreal, Quebec, Canada.
- Chris Hokamp, Demian Gholipour Ghalandari, Nghia The Pham, and John Glover. 2020. Dyne: Dynamic ensemble decoding for multi-document summarization. *arXiv preprint arXiv:2006.08748*.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pretraining for natural language generation, translation,and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online.
- Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirica Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3721–3731, Hong Kong, China.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- Markus Zopf. 2018. Auto-hMDS: Automatic construction of a large heterogeneous multilingual multidocument summarization corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC)*, pages 3228–3233, Miyazaki, Japan.