Towards abstractive summarization in Hungarian

Márton Makrai, Ákos Máté Tündik, Balázs Indig, György Szaszák
BME TMIT

MSZNY 2022. január 28.

TMiT

- summarization
  - news, science, stories, instructions, emails, patents, & legislative bills
  - extractive and abstractive
- fine-tuning paradigm
  - sequence-to-sequence initialized with a transformer
  - Rothe, Narayan, and Severyn (2020)
- multilingual abstractive summarization datasets miss Hungarian
- ELTE.DH corpus of former Hungarian news portals
- qualitative eval
  - fluent output in the correct topic, but much hallucination
- ROUGE, tokenized unigram/bigram F-score
- generation alternatives
- Spoken document summarization
  - Hungarian TV Broadcast database (manual and ASR transcripts)
  - apply stemming and punctuation (manual and automatic)
  - Best ROUGE-scores with ASR transcripts
- text length needs further study both on encoder and decoder side
- https://huggingface.co/BME-TMIT/foszt2oszt

TMiT

Agenda

1. Multilin summ, the ELTE.DH corpus, generation alternatives
2. Evaluation faithfulness with question answering
3. Spoken document summarization
4. Examples
5. Conclusions, Future Plans

TMiT

Multilingual summarization datasets

- We follow MLSum (Scialom et al., 2020), a
  - multilingual summarization dataset obtained from online newspapers
  - 1.5M+ article-summary pairs in 5 languages
    - French, German, Spanish, Russian, and Turkish besides English
  - cross-lingual comparative analyses based on state-of-the-art systems.
  - strong baselines from multilingual abstractive text generation models,
  - two theoretically independent factors of cross-lingual summarization
    - data: e.g. structure of the article, the abstractiveness of the summaries, quantity – domain adaptation
    - language: e.g. metric biases due to different morphological type – multilingual datasets are the only means to study these
- Giannakopoulos et al. (2015) and Ladhak et al. (2020)
  - multilingual summarization benchmarks
- Hasan et al. (2021): a dataset, a crawling curation tool, and summarization model checkpoints for multilingual summarization
- All these sources miss Hungarian
- abstractive summarization for Hungarian
  - Yang Zijian et al. (2021)
  - Yang Zijian Győző (2022)

TMiT

ELTE.DH corpus of former Hungarian news portals

- hír, lead (Yang Zijian et al., 2021)
- ELTE DH: megszűnt magyar hírportálok anyaga

site	cikk (>50)	van lead (>20)
Magyar idők	163 609	82 %
válasz	84 714	86 %
vs	51 302	93 %
abcúg	2 798	94 %
mosthallottam	389	80 %

- kiszűrjük, ha a cikk vagy a lead túl rövid (Scialom et al., 2020)
  - vagyis a főleg audiovizuális tartalmú cikkeket
- tanító/validáló/teszt vágás időrendi alapon
  - új témák jelennek meg az idők során
  - megakadályozza, hogy egy a tanulóadatban szereplő eseményről másik hírportálon írt cikket kelljen kivonatolni

TMiT

Generation alternatives

				stemmed		
	rouge-1	rouge-2	rouge-l	rouge-1	rouge-2	rouge-l
fosztogatnak2osztogatnak	**19.85**	**06.71**	**17.15**	**26.95**	**10.80**	**22.57**
num_beams=4	19.46	06.32	16.81	26.76	10.44	22.35
num_beams=6	19.51	06.18	16.79	26.61	10.32	22.26
top_k=50 (Fan, Grangier, and Auli, 2018)	19.08	05.97	16.83	25.75	09.67	21.94
diversity_penalty=0.5	18.60	05.95	16.41	25.70	09.43	21.76
von Platen	18.47	05.53	16.02	25.83	09.24	21.69
top_p=0.9 (Holtzman et al., 2020)	18.44	05.65	16.11	25.27	09.24	21.33
temperature=0.7	17.80	05.63	15.55	23.64	09.20	19.87

Table: ROUGE F-scores obtained with alternative generation strategies

- diversity penalty: num_beam_groups=5, Vijayakumar et al. (2016)
- temperature: Ackley, Hinton, and Sejnowski (1985)

TMiT

Agenda

1. Multilin summ, the ELTE.DH corpus, generation alternatives
2. Evaluation faithfulness with question answering
3. Spoken document summarization
4. Examples
5. Conclusions, Future Plans

TMiT

Evaluation faithfulness with question answering I

in Hungarian?

- neural encoder-decoder models distort the article
  - or generate fabrication of factual information in it
- Previous evaluation methods cannot detect this
- Huang et al. (2021) review experiments for solving the problem
  - fact-aware evaluation metrics
  - new summarization systems optimized towards factual consistency

TMiT

Evaluation faithfulness with question answering II

in Hungarian?

- Maynez et al. (2020): a large scale human evaluation of neural extreme abstractive summarization models
  - extreme: the summary is a single sentence
  - types of hallucinations
  - substantial amounts of hallucinated content
    - in all model generated summaries
  - how frequently abstractive summarizers hallucinate content
  - intrinsic: manipulating the information present in the input
  - extrinsic: adding information not directly inferable
  - how much hallucinated content is factual, even when unfaithful
  - whether there are automatic means of measuring these hallucinations
  - intrinsic and extrinsic hallucinations in more than 70% of 1-sentence summaries
  - the majority of hallucinations are extrinsic
    - potentially could be valid abstractions that use background knowledge, but over 90% of them are erroneous
  - models initialized with pre-trained parameters perform best

TMiT

Evaluation faithfulness with question answering III

in Hungarian?

- both on automatic metrics and human judgments
- highest percentage of factual within extrinsic hallucinations
  * at least on in-domain summarization

- Textual entailment measures better correlate with faithfulness than standard ones
  - potentially leading the way to automatic evaluation metrics as well as training and decoding criteria

- Gabriel et al. (2021): a meta-evaluation framework for evaluating factuality evaluation metrics
  - They define conditions to evaluate factuality metrics on diagnostic factuality data across three summarization tasks
  - question-answering metrics improve
    - over standard metrics that measure the factuality of English summaries across domains
    - but their performance is highly dependent on the way in which questions are generated
- A survey by Huang et al. (2020)

TMiT

Evaluation faithfulness with question answering IV

in Hungarian?

- under similar settings, extractive summarizers are better than their abstractive counterparts thanks to strength in faithfulness and factual-consistency
- milestone techniques (See, Liu, and Manning, 2017)
  - copy, coverage, and hybrid extractive/abstractive methods
  - specific improvements but also limitations
- pre-training techniques are highly effective for improving text summarization
  - BART giving the best results
  - in particular sequence-to-sequence pre-training

TMiT

Agenda

1. Multilin summ, the ELTE.DH corpus, generation alternatives
2. Evaluation faithfulness with question answering
3. Spoken document summarization
4. Examples
5. Conclusions, Future Plans

TMiT

Spoken document summarization

- Challenges:
  - ASR error propagation: transcription errors
  - Missing punctuation marks and capitalization: segmentation errors after restoration
  - Spoken vs. written text: grammatically different structures (possible model training mismatch)
- Related work: extractive summarization (Tündik, Kaszás, and Szaszák, 2019)
  - punctuation errors slightly more critical than ASR errors
    - misplaced sentence boundaries –> changes of N-gram sequences –> changes of ROUGE scores
- Hypotheses for abstractive summarization:
  - punctuation errors become less impactful
  - impact of ASR errors and different structure more relevant
  - ASR errors yield more hallucination

TMiT

Dataset

- Hungarian TV Broadcast database with ASR transcripts (Varga et al., 2015):
  - 10 snippets: sport news, weather forecasts and broadcast news
  - Human-made summaries prepared by 3 independent annotators based on gold (MT-MP) transcripts
  - Automatic punctuation to the transcripts: (Tündik et al., 2018)
- Transcript Types:
  - **MT-MP**: Manual Transcript with Manual Punctuation
  - **AT-MP**: ASR Transcript with Manual Punctuation
  - **MT-AP**: Manual Transcript with Automatic Punctuation
  - **AT-AP**: ASR Transcript with Automatic Punctuation

TMiT

Results

	No stemming			Stemming		
Transcript Type	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
MT-MP	10.02	**1.79**	8.55	15.32	4.45	12.30
MT-AP	9.90	1.75	8.88	14.40	4.03	12.86
AT-MP	**10.60**	1.49	**9.33**	15.56	4.44	**13.70**
AT-AP	10.15	1.58	8.46	**15.76**	**5.1**	12.41

- Overall:
  - Categories with erroneous texts has comparable or better performance than the gold transcripts
  - Significant increase with stemming
- Genre-level:
  - Weather Forecast: highest scores for ROUGE-1 and ROUGE-L
    - the length of the summary is the most appropriate for this category
  - Long sessions (news) not fit to the current length –> lower scores
    - topic segmentation ?

TMiT

Agenda

1. Multilin summ, the ELTE.DH corpus, generation alternatives
2. Evaluation faithfulness with question answering
3. Spoken document summarization
4. Examples
5. Conclusions, Future Plans

TMiT

Examples I.

Sport News	"...a Veszprém férfi kézilabdacsapata. Császár Gábor együttesével a Paris Saint-Germainnel találkozik a férfi kézilabda Bajnokok Ligájában..."
Summary	"...A magyar férfi kézilabda - válogatott a franciaországi Európa - bajnokságra készülődött, és a csoportkörbe jutásért játszik. A magyar válogatott a margitszigeti Eb - selejtezőn a franciákkal találkozik, és az Eb - n is ott lesz. A franciaországi Eb - re készülő magyar válogatott az utolsó pillanatban került be a nyolcaddöntőbe...'
Phenomena	The topic is hit by the algorithm (handball), but it mentions countries (Hungary and France) instead of specific clubs (Veszprém and Paris Saint-Germain). There is a mismatch in the competition (Champions League vs. European Championship), mentioning different stages (e.g. eight-final) from the article, and an unreferenced location ('margitszigeti')

TMiT

Examples II.

Sport News	"...ma esti Manchester United-Bayern München BL-negyeddöntő..." "...a Barcelona eddig százötször játszott az Atlético ellen..."
Summary	"A Manchester United és az Atlético Madrid összecsapása lesz a labdarúgó Bajnokok Ligája negyeddöntőjének első felvonása. A Barcelona és a Bayern München összecsapása az elődöntő első felvonásán, a Manchester City és a Barcelona pedig az elődöntőben találkozik a fináléban. A Manchester City a címvédő Real Madridot fogadja, a Barcelona a Barcelonát fogadja. A másik ágon a Barcelona és az FC Barcelona találkozik."
Phenomena	Unreferenced teams came up (Real Madrid, Manchester City), Repetitions for Barcelona, New/swapped pairings (MU - Atlético, Barcelona - Bayern), Quarter-final and semi-final in the same sentence

TMiT

Examples III.

Weather Forecast	"...a keleti országrészben pedig helyi záporok, egy-két helyen zivatar kialakulása...." "...abban a térségben nem lesz ritka harminc fok feletti hőmérséklet, meg sem. Ezzel. Európa legmelegebb pontja lesz ma az a térség..." "..a jövő hét közepétől kezdve helyenként a hőmérséklet még a harminc fokot is megközelítheti..."
Summary	"A jövő héten is folytatódik a változékony időjárás, a napsütés mellett egyre többfelé várható zápor, zivatar. A hét második felében akár 30 Celsius - fokig is emelkedhet a hőmérséklet, a hétvégén viszont akár 30 fok is lehet - derül ki az Országos Meteorológiai Szolgálat előrejelzéséből.'
Phenomena	The same temperature is mentioned but with digits and text form (30 vs. thirty), even the types of rainfall are correct, and the unreferenced Hungarian Meteorological Service (OMSZ) is nicely included in the sentence.

TMiT

Examples IV.

Broadcast News	"...a fogyatékossággal élők munkába állása, valamint a női foglalkoztatás eredményei is nagyban hozzájárultak az országos foglalkoztatási adatok javulásához - jelentette ki a szociális és családügyért felelős államtitkár. Soltész Miklós hangsúlyozta, a megváltozott munkaképességű nyilvántartott álláskeresők száma az elmúlt kormányzati ciklusban tizenhét és fél ezerre csökkent.
Summary	'A kormány elkötelezett a fogyatékkal élők foglalkoztatásának javítása mellett, a kormány a következő években is folytatja a munkaerőpiaci integrációt – mondta a szociális ügyekért és társadalmi felzárkózásért felelős államtitkár. Soltáry - Siklósi Anita kiemelte : a kormány célja, hogy minél több embert vonjanak be a munkaerőpiacra, és a jövőben is legyen lehetőség a pályakezdőkre."
Phenomena	A hard topic is recognized (integration of people with disabilities into the job market), but a wrong (non-existent!) person is mentioned.

TMiT

Agenda

1. Multilin summ, the ELTE.DH corpus, generation alternatives
2. Evaluation faithfulness with question answering
3. Spoken document summarization
4. Examples
5. Conclusions, Future Plans

TMiT

Conclusions, Future Plans

- Abstractive summarization for Hungarian language:
  - Fine-tuning of the baseline huBERT based model
  - Experiments on spoken documents transcribed by an ASR
  - Punctuation and ASR errors surprisingly does not impact the ROUGE-scores
- Future Plans:
  - Improving the objective ROUGE-scores of the model
    - choosing other references instead of the leads
    - setting the min/max length parameter of the decoder part
    - changing the encoder part to fit better to long broadcast data
  - Subjective evaluation of the summarizer

TMiT

Bibliography I

Ackley, David H, Geoffrey E Hinton, and Terrence J Sejnowski (1985). "A learning algorithm for Boltzmann machines". In: *Cognitive science* 9.1, pp. 147–169 (cit. on p. 6).

Fan, Angela, David Grangier, and Michael Auli (July 2018). "Controllable Abstractive Summarization". In: *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*. Melbourne, Australia: Association for Computational Linguistics, pp. 45–54. DOI: 10.18653/v1/W18-2706. URL: https://aclanthology.org/W18-2706 (cit. on p. 6).

Gabriel, Saadia et al. (Aug. 2021). "GO FIGURE: A Meta Evaluation of Factuality in Summarization". In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Online: Association for Computational Linguistics, pp. 478–487. DOI: 10.18653/v1/2021.findings-acl.42. URL: https://aclanthology.org/2021.findings-acl.42 (cit. on p. 10).

TMiT

Bibliography II

Giannakopoulos, George et al. (Sept. 2015). "MultiLing 2015: Multilingual Summarization of Single and Multi-Documents, On-line Fora, and Call-center Conversations". In: *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Prague, Czech Republic: Association for Computational Linguistics, pp. 270–274. DOI: 10.18653/v1/W15-4638. URL: https://aclanthology.org/W15-4638 (cit. on p. 4).

Hasan, Tahmid et al. (Aug. 2021). "XL-Sum: Large-Scale Multilingual Abstractive Summarization for 44 Languages". In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Online: Association for Computational Linguistics, pp. 4693–4703. DOI: 10.18653/v1/2021.findings-acl.413. URL: https://aclanthology.org/2021.findings-acl.413 (cit. on p. 4).

Holtzman, Ari et al. (2020). "The Curious Case of Neural Text Degeneration". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=rygGQyrFvH (cit. on p. 6).

Huang, Dandan et al. (Nov. 2020). "What Have We Achieved on Text Summarization?" In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 446–469. DOI: 10.18653/v1/2020.emnlp-main.33. URL: https://aclanthology.org/2020.emnlp-main.33 (cit. on p. 10).

TMiT

Bibliography III

Huang, Y et al. (2021). "The Factual Inconsistency Problem in Abstractive Text Summarization: A Survey". In: (cit. on p. 8).

Ladhak, Faisal et al. (Nov. 2020). "WikiLingua: A New Benchmark Dataset for Cross-Lingual Abstractive Summarization". In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, pp. 4034–4048. DOI: 10.18653/v1/2020.findings-emnlp.360. URL: https://aclanthology.org/2020.findings-emnlp.360 (cit. on p. 4).

Maynez, Joshua et al. (July 2020). "On Faithfulness and Factuality in Abstractive Summarization". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 1906–1919. DOI: 10.18653/v1/2020.acl-main.173. URL: https://aclanthology.org/2020.acl-main.173 (cit. on p. 9).

Rothe, Sascha, Shashi Narayan, and Aliaksei Severyn (2020). "Leveraging Pre-trained Checkpoints for Sequence Generation Tasks". In: *Transactions of the Association for Computational Linguistics* 8, pp. 264–280. DOI: 10.1162/tacl_a_00313. URL: https://aclanthology.org/2020.tacl-1.18 (cit. on p. 2).

TMiT

Bibliography IV

Scialom, Thomas et al. (Nov. 2020). "MLSUM: The Multilingual Summarization Corpus". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 8051–8067. DOI: 10.18653/v1/2020.emnlp-main.647. URL: https://aclanthology.org/2020.emnlp-main.647 (cit. on pp. 4, 5).

See, Abigail, Peter J. Liu, and Christopher D. Manning (July 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)*. Vancouver, Canada: Association for Computational Linguistics, pp. 1073–1083. DOI: 10.18653/v1/P17-1099. URL: https://aclanthology.org/P17-1099 (cit. on p. 11).

Tündik, Máté Akos, Valér Kaszás, and György Szaszák (2019). "Assessing the Semantic Space Bias Caused by ASR Error Propagation and its Effect on Spoken Document Summarization.". In: *Proc. Interspeech* (cit. on p. 13).

Tündik, Máté Akos et al. (2018). "User-centric Evaluation of Automatic Punctuation in ASR Closed Captioning". In: *Proc. Interspeech 2018*, pp. 2628–2632 (cit. on p. 14).

TMiT

Bibliography V

Varga, Ádám et al. (2015). "Automatic Close Captioning for Live Hungarian Television Broadcast Speech: A Fast and Resource-Efficient Approach". In: *Proceedings of SPECOM*. Springer, pp. 105–112 (cit. on p. 14).

Vijayakumar, Ashwin K. et al. (2016). "Diverse Beam Search: Decoding Diverse Solutions from Neural Sequence Models". In: *CoRR* abs/1610.02424. URL: http://arxiv.org/abs/1610.02424 (cit. on p. 6).

Yang Zijian, Győző et al. (May 2021). "Abstractive text summarization for Hungarian". In: *Annales Mathematicae et Informaticae* 53. Selected papers of the 2020 Conference on Information Technology, pp. 299–316 (cit. on pp. 4, 5).

Yang Zijian Győző (2022). "BARTerezzünk! - Messze, messze, messze a világtól, - BART kísérleti modellek magyar nyelvre". In: *XVIII. Magyar Számítógépes Nyelvészeti Konferencia*. Szeged, Magyarország: Szegedi Tudományegyetem, Informatikai Intézet (cit. on p. 4).

TMiT