Towards abstractive summarization in Hungarian

Márton Makrai[1,2], Ákos Máté Tündik[1], Balázs Indig[3,4], György Szaszák[1]

[1] BME Department of Telecommunications and Media Informatics
{tundik,szaszak}@tmit.bme.hu
[2] MTA Research Institute for Natural Sciences
makrai.marton@ttk.hu
[3] Eötvös Loránd University Department of Digital Humanities
indig.balazs@btk.elte.hu
[4] Digital Heritage National Laboratory

Abstract. We publish an abstractive summarizer for Hungarian, an encoder-decoder model initialized with huBERT, and fine-tuned on the ELTE.DH corpus of former Hungarian news portals. The model produces fluent output in the correct topic, but it hallucinates frequently. Our quantitative evaluation on automatic and human transcripts of news (with automatic and human-made punctuation) shows that the model is robust with respect to errors in either automatic speech recognition or automatic punctuation restoration.
Keywords: summarization, pre-trained model, automatic speech recognition, punctuation

1 Introduction

Automatic text summarization requires several complex language abilities: understanding the text, discriminating what is relevant, and writing a short synthesis (most commonly a couple of sentences). *Extractive* systems select sentences or words from the input document, while *abstractive* models are supposed to paraphrase the content. Extractive methods suffer from some limitations, including weak coherence between sentences, inability to simplify complex and long sentences, and unintended repetition (Hasan et al., 2021). In the past few years, pre-trained deep language models have achieved great advancements in both extractive and abstractive summarization (Edunov et al., 2019; Liu and Lapata, 2019; Rothe et al., 2020), but these models are heavily data-driven, and multilingual abstractive summarization datasets (Giannakopoulos et al., 2015; Scialom et al., 2020; Ladhak et al., 2020; Hasan et al., 2021) miss Hungarian.

We applied the pre-training-based method to Hungarian by fine-tuning an encoder-decoder model. The model has been initialized with huBERT (Nemeskey, 2020), the freely available Hungarian deep language model, as both the encoder and the decoder. The fine-tuning corpus consisted of news from former Hungarian portals. These articles have been crawled by the Hungarian National laboratory of Digital Heritage (Indig et al., 2020). Then the model is also evaluated on a Hungarian TV broadcast database Varga et al. (2015).

Qualitative analysis shows that our model produces fluent texts whose topic closely match that of the input, but the summaries contain much hallucination. The name of the model, `fosztogatnak2osztogatnak` refers to this phenomenon, more specifically an old Hungarian pun, which can be roughly translated as follows: People ask the Yerevan radio if it is true that Moskvitches are being handed out for free in Yerevan. Reply: "The news is true. However, not in Yerevan, but in Tbilisi. Not Moskvitches, but Volgas. And they are not handed out *(*Hungarian: *osztogatnak)* but plundered *(fosztogatnak)*. We share our model at `https://huggingface.co/BME-TMIT/foszt2oszt`.

2 Related work

2.1 Evaluation methods

One of the greatest problems of summarization research resides in evaluation. A common metric to automatically evaluate summaries is Recall-Oriented Understudy for Gisting Evaluation (ROUGE) (Lin, 2004), which counts the overlap of words or word units. Variants of this originally recall-oriented metric remain a strong baseline.

Fabbri et al. (2021) demonstrate the usefulness of some simple data statistics. The first three are *extractiveness* measures introduced by Grusky et al. (2018): *extractive fragment coverage* is the percentage of words in the summary that are from the source article, measuring the extent to which a summary is a derivative of a text; *density* is the average length of the extractive fragment to which each summary word belongs; and *compression ratio* is defined as the word ratio between the articles and the summary. Fabbri et al. also include the percentage of n-grams in the summary not found in the input document as a *novelty score,* and that of repeated n-grams in the summary as a *redundancy score*. Human evaluations show that summaries generated by recent models such as Pegasus (Zhang et al., 2019), BART (Lewis et al., 2020), and T5 (Raffel et al., 2020) score higher in some respect than reference summaries.

2.2 Factual faithfulness and hallucination

Huang et al. (2021) review experiments for solving the problem with neural encoder-decoder models pioneered by the Seq2Seq framework that while they produce summaries that are more abstractive, more readable, closer to human-edited summaries, they distort the article or generate fabrication of factual information in it. Previous evaluation methods of text summarization are not suitable for detecting this. The current research in response, unfortunately yet limited to English, is predominantly divided into fact-aware evaluation metrics to select outputs without factual inconsistency and new summarization systems optimized towards factual consistency.

Maynez et al. (2020) conduct a large scale human evaluation of neural extreme abstractive summarization models (extremity means that the summary is

a single sentence) and the types of hallucinations they produce, and find substantial amounts of hallucinated content in all model generated summaries.

They are interested in how frequently abstractive summarizers hallucinate content; whether models hallucinate by manipulating the information present in the input document (intrinsic hallucination) or by adding information not directly inferable from the input document (extrinsic); how much hallucinated content is factual, even when unfaithful; and whether there are automatic means of measuring these hallucinations (at least for English).

Their experiments suggest that intrinsic and extrinsic hallucinations happen in more than 70% of 1-sentence summaries; the majority of hallucinations are extrinsic, which potentially could be valid abstractions that use background knowledge, but over 90% of them are erroneous; models initialized with pre-trained parameters perform best both on automatic metrics and human judgments of faithfulness/factuality: they have the 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 for English. For Hungarian, these high level models are unfortunately yet unavailable.

The same limitation makes the findings of Gabriel et al. (2021) less relevant for Hungarian. They introduce a meta-evaluation framework for evaluating factuality evaluation metrics. They define conditions to evaluate factuality metrics on diagnostic factuality data across three summarization tasks. They find that 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) find that under similar settings, extractive summarizers are in general better than their abstractive counterparts thanks to strength in faithfulness and factual-consistency; milestone techniques (See et al., 2017) such as copy, coverage, and hybrid extractive/abstractive methods bring specific improvements but also show limitations; pre-training techniques, and in particular sequence-to-sequence pre-training, are highly effective for improving text summarization, with BART giving the best results.

2.3 Multilingual and medium-resource summarization

We follow the footsteps of MLSum (Scialom et al., 2020), a multilingual summarization dataset obtained from online newspapers. MLSum contains 1.5M+ article-summary pairs in 5 languages (French, German, Spanish, Russian, and Turkish) besides English. Scialom et al. conduct cross-lingual comparative analyses based on state-of-the-art systems. They offer strong baselines from multilingual abstractive text generation models, and distinguish between two theoretically independent factors to explain differences of results in cross-lingual summarization with different approaches: data (e.g. structure of the article, the abstractiveness of the summaries, quantity) and language (e.g. metric biases due to a different morphological type). The former has more to do with domain

adaptation, while the latter motivates further the development of multilingual datasets, since they are the only means to study such phenomena.

Giannakopoulos et al. (2015) and Ladhak et al. (2020) publish multilingual summarization benchmarks. Hasan et al. (2021) publish a dataset, a crawling curation tool, and summarization model checkpoints for multilingual summarization. All these sources miss Hungarian.

Most similar to us, Yang Zijian et al. (2021) report extractive and abstractive summarization systems for Hungarian. Their best results are obtained with huBERT, the same model we utilize. Another similarity is that their model is fine-tuned on news (HVG and index.hu), besides the MARCELL legislative corpus (Váradi et al., 2020). They mention the problem of using lead as summarization: Often the function of the lead is to attract attention and not to summarize, e.g. one of their articles is about damages caused by storms and the payments by the insurers. The lead is only about the insurers, it does not even mention the storm and the damages. Unfortunately, their model is not freely available.

2.4 Neural text generation

In the encoder-decoder setup, models are trained (or fine-tuned) to maximize the probability of the document-summary pairs provided in the training data, and at inference time, the output of the encoder is used for generating the summary. While fine-tuning is relatively computation intensive, and we experimented only with a couple of promising settings, we tried some methods to reduce hallucination – with limited success. Zarrieß and Schlangen (2018) give a survey of decoding methods in neural language generation grouped by their objectives: sequence-likelihood, diversity, and task-specific linguistic constraints or goals. The most well-known and de-facto standard decoding procedure remains beam search (Lowerre, 1976). Our experiments with some of the recent methods can be found in Section 3.2.

2.5 Spoken document summarization

A challenging task is spoken document summarization, where additional factors can make the problem a bit more complex. In Tündik et al. (2019), the authors were analysing what additional distortion effects can arise during summarization in the case the document source is audio, and an automatic speech recognizer (ASR) is used to obtain the transcripts. These transcripts are then used as bases for the same summarization algorithm which is used for text documents. The three main differences and challenges are the following:

- ASR errors can propagate further into the processing pipeline;
- Due to missing punctuation marks and capitalization, tokenization (crucial in extractive summarization) lacks the necessary cues;
- Spoken documents may follow a different structure and can grammatically be different from written documents, hence there is a mismatch with models trained on written documents.

In Tündik and Szaszák (2019) the authors evaluated punctuation efforts on Hungarian and English ASR output. In Tündik et al. (2019) a similar punctuation model was applied to ASR output and assessed for distortions in the generated summaries. In the case of extractive summarization, punctuation errors were found to be slightly more critical than ASR errors, although this finding is limited by the summarization and its evaluation approaches, as in extractive summary, misplaced sentence boundaries will result in different N-gram sequences and hence most likely different ROUGE scores. With abstractive summarization, we hypothesize that punctuation errors become less impactful, but the difference between spoken and written documents becomes more relevant, beside the impact of ASR errors in both cases. Following Tündik et al. (2019) where extractive summarization was the target, here we focus on the effect of word and/or punctuation errors in case of abstractive summarization task. It is especially interesting whether the mismatch between spoken and written document styles and the ASR errors lead to more hallucinations than in the written based summarization.

3 Experimental Results

3.1 The ELTE.DH corpus: former Hungarian news portals

The used corpus contains a subset of the continuously growing *ELTE.DH corpus* (Indig et al., 2020). The ELTE.DH corpus is created by crawling the archives of Hungarian news portals into the ISO standard *WARC (Web ARChive)* format. The downloaded material was converted into standard TEI XML format (Schreibman et al., 2008) by carefully extracting the metadata and the text content and the result was deposited at Zenodo.org repository where it is available for research purposes upon request for the sake of reproducibility. Compared to the common crawling methods, these XML files not just have paragraphs, but have all the textual data (and no boilerplate) with the formatting kept. All available metadata for each article and the typological formatting is normalized to eliminate portal specific markup. The standard TEI XML format is widely used on the field of digital humanities as long-term archiving format for textual data which ensures its usability for various tasks.

For most of the portals, the special first paragraph of each article, which we will call the *lead*, is commonly used to summarize the content and attract the reader to read further for details. This intention, if it is marked by some kind of emphasis, is identified and noted in the XML files, and can be used for the summarisation task. For the purposes of `fosztogatnak2osztogatnak`, we recorded the site of origin, the date, the title, the lead, and the body text of each article.

Table 1 shows how many articles and leads we have after a rather generous length-filtering. (Unlike Straka et al. (2018), we did not drop articles with a low text-to-abstract ratio. Future work may investigate this direction.) The train/validation/test split has been designated on a chronological basis, similarly to Scialom et al. (2020). We ensure a ratio of 8:1:1 by assigning articles

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

Table 1. The number of articles (consisting of at least 50 characters) crawled from each former news site in our corpus, and the ratio of the articles that have a lead consisting of at least 20 characters

before 2017 October 9 to the training set, those after 2018 May 30 to the test set, and those in between to the development set. The rationale of a chronological split is that new topics appear over time, and this method prevents asking the model to extract an article about an event that is present in the training data from another portal. Our pilot results reported in the next section are based on one percent of the validation set. The test set is not used for anything, but it is delineated for future work.

3.2 Fine-tuning and results on ELTE.DH corpus

In fine-tuning and inference, we followed a jupyter notebook[1] by Patrick von Platen. Most hyper-parameters are the same as those by von Platen, but we found it advantageous to change the minimum length of the summary to 8 word-pieces (instead of 56), and the number of beams in beam search to 5 (instead of 4). We experimented with other parameter settings on one percent of the validation set, but they led to inferior ROUGE-scores (F1-score of stemmed bigrams), as shown in Table 2. The stemmed score is more informative, because when the unstemmed version returns 0, we can still learn something from the stemmed metric. Stemming in ROUGE has originally been motivated with the more accurate measurement of semantic compliance. One can argue that nowadays the models can achieve such semantic quality that a fluidity/fluency becomes also a key evaluation criterion. However, our model has difficulties with factual faithfulness, rather than fluency, so stemmed ROUGE seem more relevant.

ROUGE has three hyper-parameters: whether we apply morphological stemming to summaries (both the gold and the generated one), n (1, 2 or "longest"), and the direction of the comparison (precision, recall, or F-score). We used spaCy[2] for stemming, and the `rouge` Python package for computing ROUGE. Comparing these $2\times3\times3$ settings, scores obtained in all pairs of settings correlate well (Pearson ≥ 0.47 on the output of our model).

[1] `https://github.com/patrickvonplaten/notebooks/blob/master/BERT2BERT_for_CNN_Dailymail.ipynb`

[2] `https://github.com/spacy-hu/spacy-hungarian-models`

[3] num_beam_groups=5, (Vijayakumar et al., 2016)

				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 et al., 2018)	19.08	05.97	16.83	25.75	09.67	21.94
diversity_penalty=0.5[3]	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 (Ackley et al., 1985)	17.80	05.63	15.55	23.64	09.20	19.87

Table 2. ROUGE F-scores obtained with alternative generation strategies

Tables 3 to 5 illustrate the text generation on the ELTE.DH corpus with some examples.

Title	Lakóháznak ütközött egy kisrepülőgép San Diegóban
Body	A híradás szerint a baleset szombaton történt egy Beechcraft Bonanza típusú kisrepülőgéppel, amelynek a fedélzetén négyen voltak. Brian Fennessy, a San Diegó-i tűzoltóság parancsnoka elmondta, hogy a gép pilótája súlyos hajtóműhibáról értesítette az irányítótornyot röviddel azután, hogy felszállt a Mongtomery-Gibbs repülőtérről...
Summary	Lezuhant egy kisrepülőgép az egyesült államokbeli San Diegóban, a balesetben négyen meghaltak – közölte a helyi tűzoltóság.
Phenomena	Correctly recognized topic (flight accident), extra information about the country is also correct: San Diego is in the US.

Table 3. Text generation on the ELTE.DH corpus - Example 1

Title	Civil szakértők bírálják a MÁV-Startot
Body	... a vasúttársaság 2015 végén jelentősen ritkította a legforgalmasabb fővonalak gyorsvonatait... Az év végével megszüntetik a Budapest–Berlin éjszakai vonatot... MÁV-Start ... A vasúttársaság lapunk kérdésére azt közölte,
Summary	A Magyar Államvasutak Zrt. és a MÁV - START Zrt. megállapodott abban, hogy megszüntetik a Budapest – Bécs vasútvonalat – közölte a vasúttársaság szerdán az MTI - vel.
Phenomena	Correctly mentioned topic related to Hungarian railway companies, but wrong information about discontinuing Budapest-Wien railway line operation. Hungarian telegraph office and the time are referenced incorrectly.

Table 4. Text generation on the ELTE.DH corpus - Example 2

Title	Útravaló mindenkinek
Body	...fogalmazott Prőhle Gergely, a Petőfi Irodalmi Múzeum főigazgatója a sajtótájékoztatón a Márvány utcai iskolában, amelynek üres tűzfala Arany János hatalmas portréját kapta... a költő születésének 200. évfordulójára rendezett emlékév tavaly lezárult...
Summary	Emléktáblát avattak az Arany János - emlékév keretében a Petőfi Irodalmi Múzeumban (PIM) a Kossuth - és József Attila - díjas költő tiszteletére. A köztéri alkotás a költészet napja alkalmából készült.
Phenomena	The topic is correct: the (200th) anniversary of Poet János Arany's birth, but instead of a portrait painting of him on a firewall, a memorial plaque is mentioned with a wrong place and date. He did not receive Kossuth- and József Attila-prizes either.

Table 5. Text generation on the ELTE.DH corpus - Example 3

3.3 Robustness against speech recognition and punctuation errors

Scenarios Now we turn to the assessment of semantic bias introduced by the presence of ASR and/or punctuation errors in case of abstractive summaries. We create therefore 4 kinds of transcripts to be compared:

- **MT-MP**: Manual Transcript with Manual Punctuation: this is a simple human made gold transcript, which includes punctuation for {. , ? !};
- **AT-MP**: ASR Transcript with Manual Punctuation: we use an ASR transcript, but restore punctuation from the gold transcripts (based on their timestamps, followed by human check);
- **MT-AP**: Manual Transcript with Automatic Punctuation: we remove punctuation from the gold transcript, and predict punctuation automatically;
- **AT-AP**: ASR Transcript with Automatic Punctuation: ASR transcripts are punctuated with the model described in Tündik et al. (2018).

Dataset We use 10 snippets (blocks) from a Hungarian TV broadcast database covering various genres, such as sport news, weather forecasts and news Varga et al. (2015) with its corresponding ASR solution gently provided by SpeechTex Ltd for the experiments. We have overall 500 sentences and 8k word tokens in total. We use the Kaldi version of the ASR in Varga et al. (2015) (with Kaldi decoder) by 6.8%, 10.1%, and 21.4% Word Error Rates (WER) on weather forecasts, news and sport news, respectively. For AP (automatic punctuation) we use the model from Tündik et al. (2018) and obtain F1-measures in the range of 60-70% on MT (manual transcript) and 45-50% on AT (ASR transcript).

We use human made summaries prepared by 3 independent annotators based on the MT-MP scenario transcripts.

Table 6 and Fig 1 show the results (F1-scores of different ROUGE-metrics).

The tendency is different compared to what the authors of Tündik et al. (2019) experienced in case of extractive summarization; categories with erroneous texts (mistakes are coming either from ASR or punctuation) are often

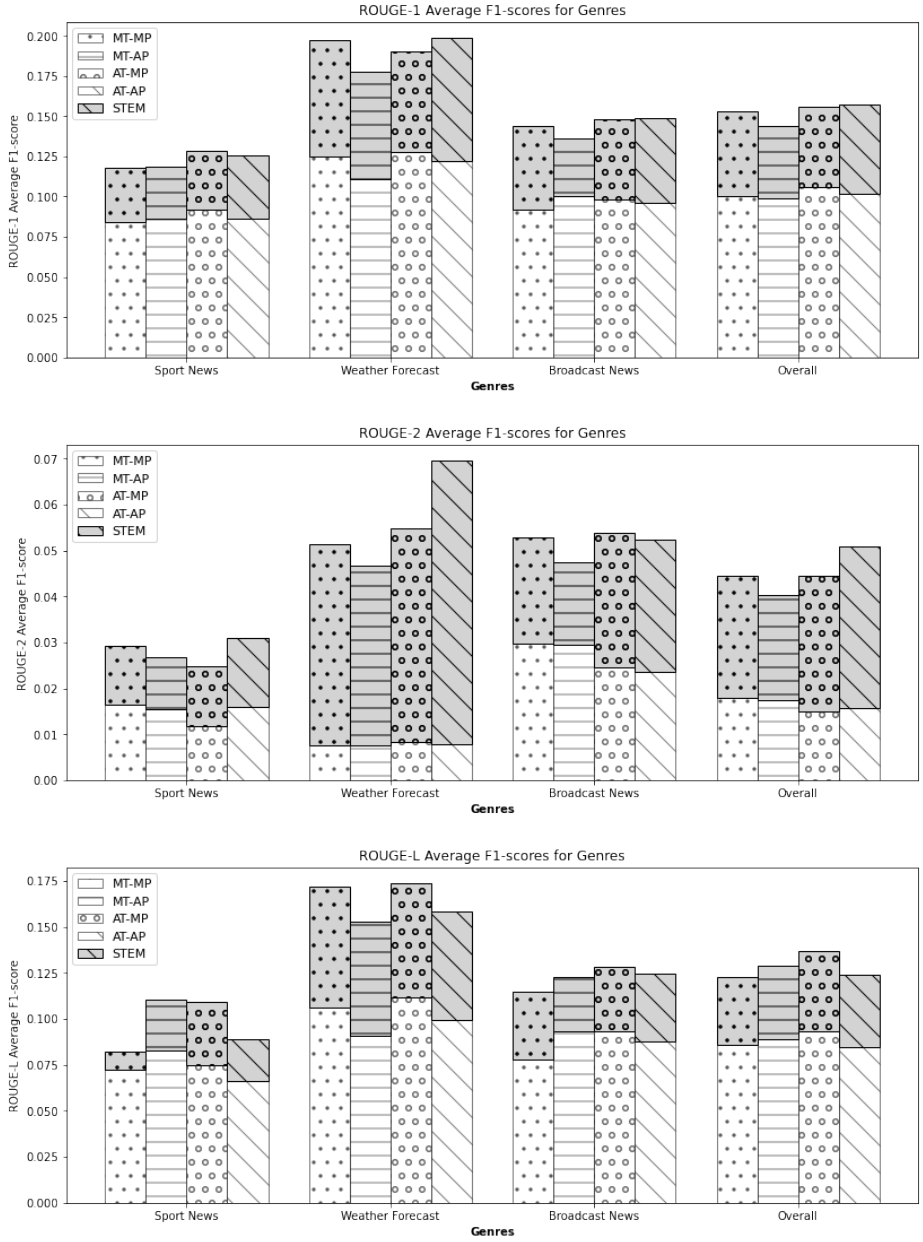

Fig. 1: Genre-based ROUGE-scores (with and without stemming)

	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

Table 6. ROUGE-scores obtained on TV Broadcast Dataset

close to the performance with gold transcripts in ROUGE, or even better. Moreover, significant increase is revealed using stemming algorithm. Although it is not a typical step to involve, this difference is subjected to the morphologically rich Hungarian language, e.g. a relevant noun (focus) in a singular vs. in a plural form. Surprisingly, in case of stemming, AT-AP transcript - based abstractive summaries provide the highest ROUGE-1 and ROUGE-2 F1-scores. The genre-level investigations uniformly confirm the positive role of stemming, the increases are shown with grey.

Summaries of Weather Forecast have the highest scores For ROUGE-1 and ROUGE-L, showing that the length of the summary is the most appropriate for this category. The length of the abstractive summaries shall be fine-tuned in the future, because there are various topics in the sport news and the broadcast news typically, which cannot be compressed; usually one main topic is emphasized by the generator, or smaller information blocks were combined like e.g. *nagyerdei stadion* (the name of a football stadium) + *gazdagréti uszoda* ('swimming pool in Gazdagrét') → *nagyerdei uszoda* ('Nagyerdei Swimming Pool'). Sometimes the polarity of the utterance changes, e.g. negation is dropped. Unfortunately, since ROUGE-scores are just simple numerical approximations, they hide these hallucination errors, so in case of abstractive summaries, subjective evaluation is also needed in the future.

Tables 7 to 10 illustrate the text generation on the Broadcast Dataset with some examples (possibly containing ASR and/or punctuation errors).

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')

Table 7. Text generation on the Broadcast Dataset - Example 1

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 finálében. 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

Table 8. Text generation on the Broadcast Dataset - Example 2

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.

Table 9. Text generation on the Broadcast Dataset - Example 3

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.

Table 10. Text generation on the Broadcast Dataset - Example 4

4 Conclusions

In this paper we were investigating encoder-decoder based abstractive summarization for Hungarian language. Beside fine-tuning the baseline huBERT based model, we were also looking into aspects of using the summarizer on spoken documents transcribed by ASR. We evaluated our model with objective scores, i.e. ROUGE. As the fine-tuning and evaluation were performed with corpora using news lead as target summary, this evaluation is to some extent biased, but with lacking human reference this is still the best evaluation scenario we could think of. In terms of spoken documents, the ASR can introduce two types of errors, word errors and punctuation errors. A bit surprisingly these errors did not significantly impact summarization performance, although the before mentioned mismatch resulting from using the leads as targets limits the applicability of these findings. In the future we would like to focus both on the improvement w.r.t. ROUGE-scores, e.g. deeper investigation of the min/max length parameter of the decoder part, and/or also changing the encoder part to fit better to long broadcast data, and on the subjective evaluation of our summarizer. Eventually a finer method designed for semi-supervised or unsupervized target extraction is also in our interest, so that we can use more accurate references than the leads, which are not always true summaries of the following document.

Acknowledgement

Our model was fine-tuned on a server of the SZTAKI HLT group, which kindly provided access to it. We are grateful to SpeechTex Ltd. for letting us use their ASR engine on some spoken documents to simulate spoken document summarization with an ASR + Summarizer cascade. We thank Dávid Nemeskey for help in uploading the model to Hugging Face. We thank NKFIH for the financial support of the experiments and project (under contract FK-124413).

Bibliography

Ackley, D.H., Hinton, G.E., Sejnowski, T.J.: A learning algorithm for boltzmann machines. Cognitive science 9(1), 147–169 (1985)

Edunov, S., Baevski, A., Auli, M.: Pre-trained language model representations for language generation. 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). pp. 4052–4059. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019), https://aclanthology.org/N19-1409

Fabbri, A.R., Kryściński, W., McCann, B., Xiong, C., Socher, R., Radev, D.: Summeval: Re-evaluating summarization evaluation. Transactions of the Association for Computational Linguistics 9, 391–409 (2021)

Fan, A., Grangier, D., Auli, M.: Controllable abstractive summarization. In: Proceedings of the 2nd Workshop on Neural Machine Translation and Generation. pp. 45–54. Association for Computational Linguistics, Melbourne, Australia (Jul 2018), https://aclanthology.org/W18-2706

Gabriel, S., Celikyilmaz, A., Jha, R., Choi, Y., Gao, J.: GO FIGURE: A meta evaluation of factuality in summarization. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. pp. 478–487. Association for Computational Linguistics, Online (Aug 2021), https://aclanthology.org/2021.findings-acl.42

Giannakopoulos, G., Kubina, J., Conroy, J., Steinberger, J., Favre, B., Kabadjov, M., Kruschwitz, U., Poesio, M.: 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. pp. 270–274. Association for Computational Linguistics, Prague, Czech Republic (Sep 2015), https://aclanthology.org/W15-4638

Grusky, M., Naaman, M., Artzi, Y.: 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). pp. 708–719. Association for Computational Linguistics, New Orleans, Louisiana (Jun 2018), https://aclanthology.org/N18-1065

Hasan, T., Bhattacharjee, A., Islam, M.S., Mubasshir, K., Li, Y.F., Kang, Y.B., Rahman, M.S., Shahriyar, R.: XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. pp. 4693–4703. Association for Computational Linguistics, Online (Aug 2021), https://aclanthology.org/2021.findings-acl.413

Holtzman, A., Buys, J., Du, L., Forbes, M., Choi, Y.: The curious case of neural text degeneration. In: International Conference on Learning Representations (2020), https://openreview.net/forum?id=rygGQyrFvH

Huang, D., Cui, L., Yang, S., Bao, G., Wang, K., Xie, J., Zhang, Y.: What have we achieved on text summarization? In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 446–469. Association for Computational Linguistics, Online (Nov 2020), https://aclanthology.org/2020.emnlp-main.33

Huang, Y., Feng, X., Feng, X., Qin, B.: The factual inconsistency problem in abstractive text summarization: A survey (2021)

Indig, B., Knap, Á., Sárközi-Lindner, Z., Timári, M., Palkó, G.: The ELTE.DH pilot corpus – creating a handcrafted Gigaword web corpus with metadata. In: Proceedings of the 12th Web as Corpus Workshop. pp. 33–41. European Language Resources Association, Marseille, France (May 2020), https://aclanthology.org/2020.wac-1.5

Ladhak, F., Durmus, E., Cardie, C., McKeown, K.: WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization. In: Findings of the Association for Computational Linguistics: EMNLP 2020. pp. 4034–

4048. Association for Computational Linguistics, Online (Nov 2020), https://aclanthology.org/2020.findings-emnlp.360

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 7871–7880. Association for Computational Linguistics, Online (Jul 2020), https://aclanthology.org/2020.acl-main.703

Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. pp. 74–81. Association for Computational Linguistics, Barcelona, Spain (Jul 2004), https://aclanthology.org/W04-1013

Liu, Y., Lapata, M.: Text summarization with pretrained encoders. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 3730–3740. Association for Computational Linguistics, Hong Kong, China (Nov 2019), https://aclanthology.org/D19-1387

Lowerre, B.: The HARPY Speech Recognition System. Ph.D. thesis, Carnegie Mellon University, Pittsburgh, PN, USA (1976)

Maynez, J., Narayan, S., Bohnet, B., McDonald, R.: On faithfulness and factuality in abstractive summarization. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 1906–1919. Association for Computational Linguistics, Online (Jul 2020), https://aclanthology.org/2020.acl-main.173

Nemeskey, D.M.: Natural Language Processing Methods for Language Modeling. Ph.D. thesis, Eötvös Loránd University (2020)

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research 21(140), 1–67 (2020), http://jmlr.org/papers/v21/20-074.html

Rothe, S., Narayan, S., Severyn, A.: Leveraging pre-trained checkpoints for sequence generation tasks. Transactions of the Association for Computational Linguistics 8, 264–280 (2020), https://aclanthology.org/2020.tacl-1.18

Schreibman, S., Siemens, R., Unsworth, J.: A companion to digital humanities. John Wiley & Sons (2008)

Scialom, T., Dray, P.A., Lamprier, S., Piwowarski, B., Staiano, J.: MLSUM: The multilingual summarization corpus. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 8051–8067. Association for Computational Linguistics, Online (Nov 2020), https://aclanthology.org/2020.emnlp-main.647

See, A., Liu, P.J., Manning, C.D.: 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). pp. 1073–1083. Association for Computational Linguistics, Vancouver, Canada (Jul 2017), https://aclanthology.org/P17-1099

Straka, M., Mediankin, N., Kocmi, T., Žabokrtský, Z., Hudeček, V., Hajič, J.: SumeCzech: Large Czech news-based summarization dataset. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), Miyazaki, Japan (May 2018), https://aclanthology.org/L18-1551

Tündik, M.A., Kaszás, V., Szaszák, G.: Assessing the semantic space bias caused by asr error propagation and its effect on spoken document summarization. In: Proc. Interspeech (2019)

Tündik, M.A., Szaszák, G.: Leveraging a character, word and prosody triplet for an asr error robust and agglutination friendly punctuation approach. In: Proc. Interspeech (2019)

Tündik, M.A., Szaszák, G., Gosztolya, G., Beke, A.: User-centric Evaluation of Automatic Punctuation in ASR Closed Captioning. In: Proc. Interspeech 2018. pp. 2628–2632 (2018)

Váradi, T., Koeva, S., Yamalov, M., Tadić, M., Sass, B., Nitoń, B., Ogrodniczuk, M., Pęzik, P., Barbu Mititelu, V., Ion, R., Irimia, E., Mitrofan, M., Păiș, V., Tufiș, D., Garabík, R., Krek, S., Repar, A., Rihtar, M., Brank, J.: The MARCELL legislative corpus. In: Proceedings of the 12th Language Resources and Evaluation Conference. pp. 3761–3768. European Language Resources Association, Marseille, France (May 2020), https://aclanthology.org/2020.lrec-1.464

Varga, Á., Tarján, B., Tobler, Z., Szaszák, G., Fegyó, T., Bordás, C., Mihajlik, P.: Automatic Close Captioning for Live Hungarian Television Broadcast Speech: A Fast and Resource-Efficient Approach. In: Proceedings of SPECOM. pp. 105–112. Springer (2015)

Vijayakumar, A.K., Cogswell, M., Selvaraju, R.R., Sun, Q., Lee, S., Crandall, D.J., Batra, D.: Diverse beam search: Decoding diverse solutions from neural sequence models. CoRR abs/1610.02424 (2016), http://arxiv.org/abs/1610.02424

Yang Zijian, G., Agócs, Á., Kusper, G., Váradi, T.: Abstractive text summarization for hungarian. Annales Mathematicae et Informaticae 53, 299–316 (5 2021), selected papers of the 2020 Conference on Information Technology

Zarrieß, S., Schlangen, D.: Decoding strategies for neural referring expression generation. In: Proceedings of the 11th International Conference on Natural Language Generation. pp. 503–512. Association for Computational Linguistics, Tilburg University, The Netherlands (Nov 2018), https://aclanthology.org/W18-6563

Zhang, J., Zhao, Y., Saleh, M., Liu, P.J.: PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. CoRR abs/1912.08777 (2019), http://arxiv.org/abs/1912.08777