# From News to Summaries: Building a Hungarian Corpus for Extractive and Abstractive Summarization

**Botond Barta**[1,2], **Dorina Lakatos**[1,2], **Attila Nagy**[2], **Milán Konor Nyist**[2], **Judit Ács**[1]
[1]HUN-REN Institute for Computer Science and Control
[2]Department of Automation and Applied Informatics
Budapest University of Technology and Economics
botondbarta@sztaki.hu, dorinapetra@gmail.com, attila.nagy234@gmail.com
nyist.milan78@gmail.com, acsjudit@sztaki.hu

**Abstract**

Training summarization models requires substantial amounts of training data. However for less resourceful languages like Hungarian, openly available models and datasets are notably scarce. To address this gap our paper introduces HunSum-2 an open-source Hungarian corpus suitable for training abstractive and extractive summarization models. The dataset is assembled from segments of the Common Crawl corpus undergoing thorough cleaning, preprocessing and deduplication. In addition to abstractive summarization we generate sentence-level labels for extractive summarization using sentence similarity. We train baseline models for both extractive and abstractive summarization using the collected dataset. To demonstrate the effectiveness of the trained models, we perform both quantitative and qualitative evaluation. Our dataset, models and code are publicly available, encouraging replication, further research, and real-world applications across various domains.

**Keywords:** abstractive summarization, extractive summarization, Hungarian

## 1. Introduction

The goal of Automatic Text Summarization is to produce a short, concise text, which retains key information from a longer article (Mani and Maybury, 1999). The advent of pre-trained language models has significantly advanced the field with a large body of research now concentrated on leveraging these models for more effective and coherent summaries (Liu and Lapata, 2019a). The two main approaches to summarization are extractive and abstractive.

Extractive summarization methods identify and extract salient sentences or tokens directly from the source document to construct the summary (Cao et al., 2016; Cheng and Lapata, 2016). These models are generally less coherent, but faster and less prone to faithfulness related problems compared to their abstractive counterpart (Li et al., 2021; Dreyer et al., 2023). In recent years, pre-trained language models such as GPT (Brown et al., 2020), PEGASUS (Zhang et al., 2020) and T5 (Raffel et al., 2020) have shown promising results in generating abstractive summaries. Although these models produce very fluent summaries, they tend to hallucinate inconsistent or contradictory content compared to the source document (Maynez et al., 2020).

In this paper, we build a dataset for Hungarian summarization and release it as open-source[1] alongside models trained on the data. We construct an abstractive summarization corpus[2] by performing a thorough cleaning and preprocessing of Hungarian segments from the Common Crawl dataset. Using the crawled news articles we also generate an extractive summarization corpus[3] by selecting the most similar article sentence for each lead sentence based on their sentence embeddings. We train both abstractive and extractive models on this corpus and evaluate them both quantitatively and qualitatively.

## 2. Related work

The CNN-DM corpus (Nallapati et al., 2016) was the first large-scale English abstractive summarization dataset which was constructed by scraping news outlets. Their summaries used human-generated summary bullets on the page. Another English-language summarization dataset is XSum (Narayan et al., 2018) which uses specific HTML classes on the page to collect the summary. Several different monolingual datasets have been inspired by XSum such as the French OrangeSum (Kamal Eddine et al., 2021) or the Russian Gazeta (Gusev, 2020). We follow a similar methodology later on in our paper. For Hungarian summarization Yang et al. (2021) build a corpus from two major Hungarian news sites (overlapping with our dataset) and train BERT-like models (Devlin et al., 2019). Agócs and Yang (2022) train multilingual and Hungarian models based on PreSumm (Liu and Lapata, 2019b). Makrai et al. (2022) train an encoder-decoder model based on huBERT (Nemeskey,

[1]https://github.com/botondbarta/HunSum
[2]SZTAKI-HLT/HunSum-2-abstractive
[3]SZTAKI-HLT/HunSum-2-extractive



*LREC-COLING 2024*, pages 7503–7509

2020) using the ELTE.DH corpus (Indig et al., 2020). Yang (2022) train BART-based models (Lewis et al., 2020) for abstractive summarization. Yang (2023) fine-tune PEGASUS and multilingual models mT5 and mBART for Hungarian abstracive summarization. We do our best effort to compare models trained on our dataset to prior works. Most works in Hungarian only released models and not the datasets, so any comparative analysis has to be taken with a grain of salt. A prior version of this dataset was released as HunSum-1 (Barta et al., 2023) with less preprocessing, fewer data sources and no extractive summaries.

## 3. Methods

### 3.1. Dataset collection

We use the freely available Common Crawl dataset[4] as a basis for constructing the corpus. It contains petabytes of crawled web pages from the past 25 years and it is available on Amazon S3 in WARC format. Retrieval and deduplication of the raw dataset by domains was done using the downloader created by Nemeskey (2020). We pick 27 Hungarian news sites including most major Hungarian-language news sites to build our corpus. The selected sites all have a dedicated lead article field to make extracting the summary easier. The final raw dataset was 290 GB of data in HTML format. We then extracted the relevant parts from each article: the lead, the article, the title, the creation date and optionally some tags. We apply the following preprocessing steps and constraints:

- Remove links, image captions and embedded social media from articles.
- Remove galleries.
- Discard articles that are a part of a live blog.
- Discard articles where the article text is shorter than the lead.
- Discard articles shorter than 200 characters or longer than 15,000 characters or have fewer than 6 sentences.
- Discard articles with leads shorter than 6 tokens or longer than 5 sentences.
- Remove low-quality or incorrectly scraped data points. We assess quality by calculating the similarity between the leads and articles using the `paraphrase-multilingual-MiniLM-L12-v2` from the `sentence-transformer` package and remove those with a similarity score below 0.17.

[4]https://commoncrawl.org/

Through exploratory data analysis we also removed problematic patterns in the data, such as lottery and sports results, where the data was not applicable to summarization.

For tokenization and sentence splitting, we used the quntoken[5] package, for language detection we used FastText (Joulin et al., 2017). We also remove near-duplicate documents with Locality Sensitive Hashing (LSH) with a similarity threshold of 0.45. If two articles were classified as similar, we kept the more recent one. The preprocessed and deduplicated dataset contains 1.82 million documents. Distribution by year and source with the average sentence and token numbers can be seen in Figure 1 and Table 2. We also compute a number of commonly used descriptive statistical measures about the dataset such as Novel N-gram ratio (NNG-n) (Narayan et al., 2018), compression (CMP) (Bommasani and Cardie, 2020) and redundancy (RED-n) (Hasan et al., 2021) listed in Table 1.

We split the final dataset with stratified sampling using the news sources to train-dev-test with the dev and tests being 1998 documents. This split is released alongside the entire dataset on Huggingface. We carry out all of our experiments on this split and encourage further works to do so for comparable results.

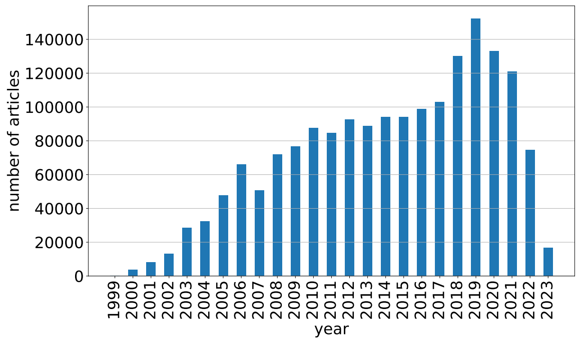


Figure 1: Number of articles by year.

### 3.2. Abstractive Summarization

We trained baseline models using our dataset. As there is no publicly available Hungarian generative model, we experimented with mT5 (Xue et al., 2021), the multilingual version of the T5 model. Another model we experimented with is the Hungarian version of the BERT model, huBERT (Nemeskey, 2020), which we fine-tuned as an encoder-decoder architecture (Bert2Bert).

We fine-tuned these models on our dataset using the parameters in Table 3. The BERT models have a maximum input length of 512 tokens, and for comparison purposes we also truncated the input

[5]https://github.com/nytud/quntoken

| NNG-1 | NNG-2 | NNG-3 | CMP | RED-1 | RED-2 |
|---|---|---|---|---|---|
| 41.12 | 77.31 | 88.74 | 89.1 | 11.78 | 0.51 |

Table 1: Intrinsic evaluation of the dataset.

| | | Article | | Lead | |
|---|---|---|---|---|---|
| **Site** | **Count** | **tokens** | **sents** | **tokens** | **sents** |
| *regional* | 346 812 | 368.2 | 18.6 | 27.1 | 1.5 |
| 24.hu | 307 477 | 350.6 | 18.8 | 22.7 | 1.4 |
| origo.hu | 293 810 | 408.3 | 20.3 | 40.5 | 2.0 |
| hvg.hu | 206 719 | 382.4 | 17.1 | 30.0 | 1.5 |
| kisalfold.hu | 161 315 | 341.6 | 18.8 | 25.8 | 1.5 |
| index.hu | 159 545 | 526.2 | 26.1 | 42.5 | 2.2 |
| delmagyar.hu | 153 139 | 351.4 | 18.9 | 29.8 | 1.7 |
| nlc.hu | 99 674 | 385.2 | 22.1 | 26.1 | 1.7 |
| nepszava.hu | 28 493 | 468.2 | 21.4 | 33.2 | 1.6 |
| portfolio.hu | 22 766 | 470.2 | 21.5 | 54.3 | 2.1 |
| m4sport.hu | 19 673 | 397.7 | 24.8 | 28.7 | 1.3 |
| metropol.hu | 12 007 | 295.7 | 15.9 | 25.1 | 1.4 |
| telex.hu | 6 420 | 918.9 | 41.6 | 52.0 | 2.4 |

Table 2: Average length of the articles and leads. The *regional* category groups smaller, local news sites.

| **Parameter** | **Bert2Bert/mT5** |
|---|---|
| batch size | 13 |
| learning rate | 5e-5 |
| weight decay | 0.01 |
| warmup steps | 16000/3000 |
| patience | 6 |

Table 3: Hyperparameters for training abstractive summarization models.

in case of the mT5 model. The models were trained on a single NVIDIA A100 GPU with early stopping on the validation loss. The mT5 model stopped learning at 8.14 epoch, while the Bert2Bert model at 3.8.

### 3.3. Extractive Summarization

Extractive summarization models highlight sentences that summarize the article. Training such models requires binary labeling at the sentence level which is not available in our raw dataset. To transform our data into this form, we used sentence transformers to calculate the embedding of the lead and article sentences, and then for each lead sentence we selected the closest article sentence by cosine distance in such a way that the sum of similarities is maximised. The sentence embeddings were computed using the `paraphrase-multilingual-MiniLM-L12-v2` model.

We chose the BertSum (Liu, 2019) architecture using huBERT with a simple classifier layer at the end to train our baseline model for extractive summarization. To train our model we used the same train-dev-test split mentioned before. The model was trained for 21,000 steps using a batch size of 200 with a learning rate of 5e-5. We evaluated the model every 1000 steps on our validation set and stopped the training process when the evaluation loss had not decreased in 10 evaluation step. The model was trained on four NVIDIA A100 GPUs.

## 4. Results

### 4.1. Quantitative Evaluation

We evaluated our abstractive and extractive models using two automatic metrics: ROUGE (Lin, 2004) and BertScore (Zhang et al., 2019). The results can be seen in Table 4. The extractive model outperformed the abstractive models significantly in terms of ROUGE and slightly in terms of BertScore. This may be a biased comparison to some extent, since the extractivity of the dataset itself favors extractive models when making comparisons using metrics such as ROUGE. We also compared our models to other publicly available Hungarian abstractive summarization models. The ROUGE scores turned out considerably lower for these models with a multilingual BART model producing the highest ROUGE score. As these models' training and test data is not available, we only evaluated them on our test set, this likely explains the performance difference compared to our models. We also compared our best performing abstractive model

Bert2Bert with other models trained on monolingual summarization datasets in other languages. For most of them, only ROUGE scores have been published, therefore only these are shown in Table 5. Due to the varying sizes of the other publicly available datasets and their linguistic differences, it is not possible to draw any major conclusions except that the ROUGE scores of the models are roughly in the same range.

### 4.2. Qualitative Evaluation

Quantitative metrics cannot always reveal specific problems with abstractive summarization models, such as hallucinations or biases. For this reason, we conduct a qualitative analysis on a 60 document sample from the test set. We extend the questions used by Hasan et al. (2021) with an additional question about grammaticality. Each annotator has to answer the following questions for each model prediction:

- **Relevant:** Does the summary convey what the article is about?
- **Consistent:** Does the summary only contain information that is consistent with the article?
- **No Hallucination:** Does the summary only contain information that can be inferred from the article?
- **Grammatical:** Is the summary grammatically correct?

Annotators are also asked, which summary they consider best, in that case the extractive model summary is also an option to select.

All annotators are native Hungarian speakers. Every data point was annotated by three annotators. The average majority answers are presented in Figure 2 where 1 means `Yes` and 0 means `No`. The average pairwise Cohen kappa between the annotators is 0.60 indicating moderate agreement. The results show that the mT5 model performs slightly better on all 4 questions. In general, close to 70% of the articles were classified as correctly capturing the gist of the document for both models. Factuality seems to be the biggest pain point as close to two thirds of the generations contained at least one inconsistency with the original article. Interestingly outputs that cannot be verified from the source sentence (extrinsic hallucinations) were produced less frequently, only in about 20% of cases for the mT5 model. For the question about the best model, the extractive model was chosen 60% of the time, while the mT5 model only reached 23%. Annotators felt that although extractive summaries were often less coherent, the factual mistakes and inconsistencies made abstractive summaries less desirable.

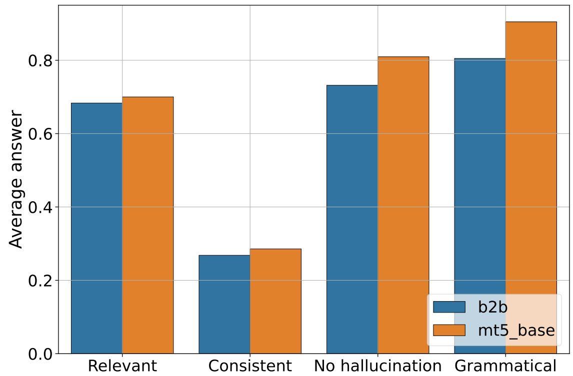


Figure 2: The average answers for the properties by models.

## 5. Conclusion

This paper presents a novel open-source Hungarian corpus designed for training both extractive and abstractive summarization models. The baseline models trained on the dataset have shown promising results both quantitatively and qualitatively with the extractive model performing best. Although the abstractive models produced fluent and grammatically correct sentences, the qualitative evaluation highlighted concerns particularly around factuality. Improving this is an exciting future direction both via making improvements to the dataset or experimenting with architectures that optimize for factual correctness. We encourage future works to use this dataset for benchmarking new methods for Hungarian summarization and hope that this will improve reproducibility in the field.

## 6. Bibliographical References

Ádám Agócs and Zijian Győző Yang. 2022. Absztraktív összefoglaló presumm módszerrel. In *XVIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2022)*.

Botond Barta, Dorina Lakatos, Attila Nagy, Milán Konor Nyist, and Judit Ács. 2023. HunSum-1: an Abstractive Summarization Dataset for Hungarian. In *XIX. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2023)*, pages 231–243, Szeged, Magyarország. Szegedi Tudományegyetem, Informatikai Intézet.

Rishi Bommasani and Claire Cardie. 2020. Intrinsic evaluation of summarization datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8075–8096, Online. Association for Computational Linguistics.

| Model | R-1 | R-2 | R-L | BertScore |
|---|---|---|---|---|
| Bert2Bert | 40.95 | 14.18 | 27.42 | 78.81 |
| mT5-base | 40.06 | 12.67 | 25.93 | 78.64 |
| extractive | **49.85** | **20.12** | **33.46** | **79.18** |
| hi-mbart-large-50 (Yang, 2023) | 31.63 | 13.26 | 22.82 | 77.77 |
| hi-mt5-base (Yang, 2023) | 29.53 | 11.34 | 21.35 | 76.99 |
| foszt2oszt (Makrai et al., 2022) | 26.87 | 8.03 | 20.19 | 75.84 |

Table 4: ROUGE and BertScore recall scores on the test set. ROUGE-1, ROUGE-2 and ROUGE-L scores are abbreviated as R-1, R-2 and R-L respectively.

| Dataset | Language | Size | R-1 | R-2 | R-L |
|---|---|---|---|---|---|
| HunSum-2 (ours) | Hungarian | 1.82M | 40.95 | 14.18 | 27.42 |
| CNN/DM (Nallapati et al., 2016) | English | 312K | 35.46 | 13.30 | 32.65 |
| OrangeSum (Kamal Eddine et al., 2021) | French | 30K | 32.67 | 13.73 | 23.18 |
| pn-summary (Farahani et al., 2021) | Persian | 93K | 44.01 | 25.07 | 37.76 |
| Gazeta (Gusev, 2020) | Russian | 60K | 32.10 | 14.20 | 27.90 |
| IlPost (Landro et al., 2022) | Italian | 44K | 38.91 | 21.38 | 32.05 |

Table 5: ROUGE scores on different monolingual abstractive summarization models.


Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Ziqiang Cao, Wenjie Li, Sujian Li, Furu Wei, and Yanran Li. 2016. AttSum: Joint learning of focusing and summarization with neural attention. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 547–556, Osaka, Japan. The COLING 2016 Organizing Committee.

Jianpeng Cheng and Mirella Lapata. 2016. Neural summarization by extracting sentences and words. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 484–494, Berlin, Germany. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Markus Dreyer, Mengwen Liu, Feng Nan, Sandeep Atluri, and Sujith Ravi. 2023. Evaluating the tradeoff between abstractiveness and factuality in abstractive summarization. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2089–2105, Dubrovnik, Croatia. Association for Computational Linguistics.

Mehrdad Farahani, Mohammad Gharachorloo, and Mohammad Manthouri. 2021. Leveraging parsbert and pretrained mt5 for persian abstractive text summarization. In *2021 26th International computer conference, computer society of Iran (CSICC)*, pages 1–6. IEEE.

Ilya Gusev. 2020. Dataset for automatic summarization of russian news. In *Artificial Intelligence and Natural Language: 9th Conference, AINL 2020, Helsinki, Finland, October 7–9, 2020, Proceedings 9*, pages 122–134. Springer.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

Balázs Indig, Árpád Knap, Zsófia Sárközi-Lindner, Mária Timári, and Gábor Palkó. 2020. The ELTE.DH pilot corpus – creating a handcrafted Gigaword web corpus with metadata. In *Proceedings of the 12th Web as Corpus Workshop*, pages 33–41, Marseille, France. European Language Resources Association.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Moussa Kamal Eddine, Antoine Tixier, and Michalis Vazirgiannis. 2021. BARThez: a skilled pretrained French sequence-to-sequence model. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9369–9390, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nicola Landro, Ignazio Gallo, Riccardo La Grassa, and Edoardo Federici. 2022. Two new datasets for italian-language abstractive text summarization. *Information*, 13(5):228.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. 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*, pages 7871–7880, Online. Association for Computational Linguistics.

Haoran Li, Arash Einolghozati, Srinivasan Iyer, Bhargavi Paranjape, Yashar Mehdad, Sonal Gupta, and Marjan Ghazvininejad. 2021. EASE: Extractive-abstractive summarization end-to-end using the information bottleneck principle. In *Proceedings of the Third Workshop on New Frontiers in Summarization*, pages 85–95, Online and in Dominican Republic. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yang Liu. 2019. Fine-tune BERT for extractive summarization. *CoRR*, abs/1903.10318.

Yang Liu and Mirella Lapata. 2019a. 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)*, pages 3730–3740, Hong Kong, China. Association for Computational Linguistics.

Yang Liu and Mirella Lapata. 2019b. 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)*, pages 3730–3740, Hong Kong, China. Association for Computational Linguistics.

Márton Makrai, Ákos Máté Tündik, Balázs Indig, and György Szaszák. 2022. Towards abstractive summarization in hungarian. In *XVIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2022)*.

Inderjeet Mani and Mark T Maybury. 1999. *Advances in automatic text summarization*. MIT press.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Dávid Márk Nemeskey. 2020. *Natural Language Processing Methods for Language Modeling*. Ph.D. thesis, Eötvös Loránd University.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Zijian Győző Yang, Ádám Agócs, Gábor Kusper, and Tamás Váradi. 2021. Abstractive text summarization for hungarian. volume 53, pages 299–316.

Zijian Győző Yang. 2022. Barterezzünk!-messze messze messze a világtól-bart kísérleti modellek magyar nyelvre. In *XVIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2022)*.

Zijian Győző Yang. 2023. Többnyelvű modellek és PEGASUS finomhangolása magyar nyelvű absztraktív összefoglalás feladatára. In *XIX. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2023)*, pages 381–393, Szeged, Magyarország. Szegedi Tudományegyetem, Informatikai Intézet.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with BERT. *CoRR*, abs/1904.09675.