



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

Botond Barta, Dorina Lakatos, Attila Nagy, Milán Nyist, Judit Ács

[botondbarta@sztaki.hu](mailto:botondbarta@sztaki.hu)

# Automatic text summarization

## Extractive

- Identify salient information in the text
- Typically selects a few sentences
- Computationally easier
- Coherence issues
- Redundancy issues
- No gold standard training sets

## Abstractive

- Generate new text
- Length can be adjusted within reason
- Computationally harder
- Hallucination
- Factual errors
- Grammaticality
- Training data can be harvested

# HunSum-2 dataset

Crawled and parsed 27 Hungarian news sites with leads from the CommonCrawl

- Title
- Lead
- Text
- Date, tags etc.

As an abstractive summarization task:

- Text →lead
- Text →title

2022. november. 22. 18:03 · TECH

## Válságálló az informatikusok fizetése, 30 százalékkal nőttek a bérek

szerző: hvg.hu

TUDOMÁNY +

**A senior IT-s szakemberek bruttó fizetése bőven a milliós kategóriában mozog, de a kezdő programozók is megkereshetik a bruttó 600–850 ezer forintot.**

A jelenlegi nehéz gazdasági helyzet sok szektorra hatással van, azonban az informatikai iparágat, és azon belül az IT-munkaerőpiacot látszólag kevésbé érinti. Hazai és külföldi szinten a tavalyi évhez képest sok munkakörben növekedtek a bérek, átlagosan közel 30 százalékkal.

A Bluebird összegzése szerint 2022-ben tovább nőtt a kereslet a Cloud megoldásokkal foglalkozó szakemberek iránt. Devops mérnökökre szintúgy nagy az igény, főleg Devops Cloud szakértők esetében, ugyanis megközelítőleg 20 százalékkal magasabb bért kapnak AWS, Azure, vagy GCP tapasztalatukért.

# Cleaning

- Constraints on the length:
  - Lead characters < article characters
  - Minimum article characters: 200
  - Maximum article characters: 15000
  - Minimum lead tokens (~words): 6
  - Maximum lead sentences: 5
  - Minimum article sentences: 6
- Remove non-Hungarian articles with fastText
- Article and lead deduplication with LSH - 0.45 threshold
- Minimum lead-article similarity using sentence-transformers - 0.17 threshold
- 1.82M articles in total

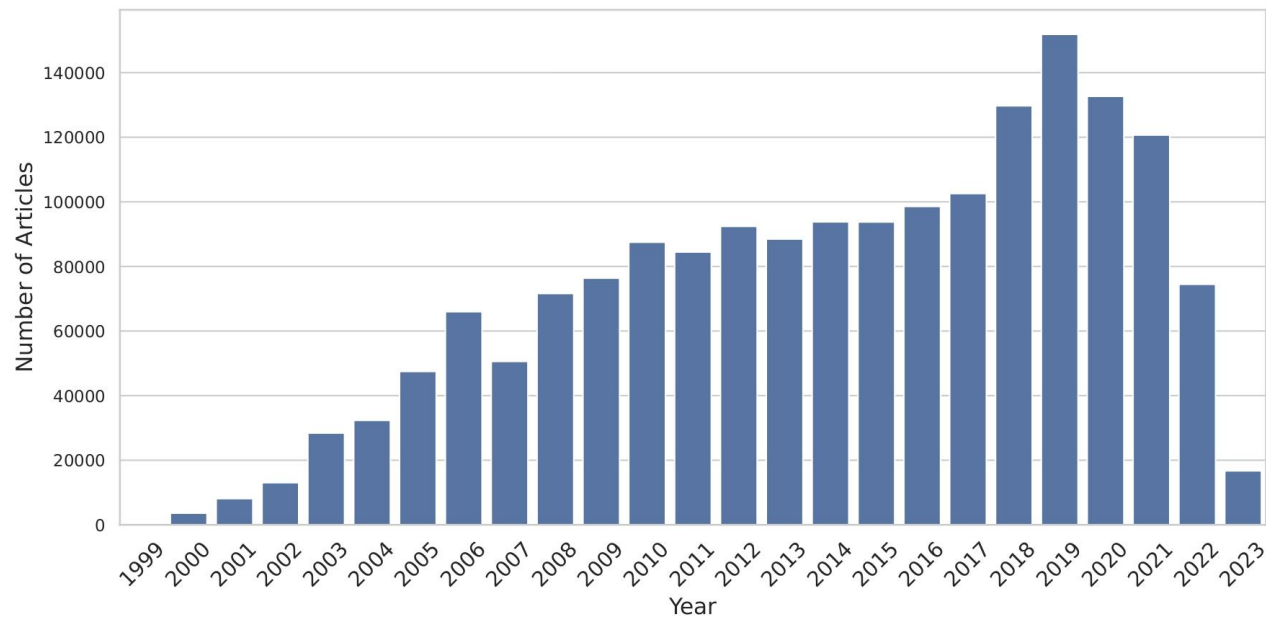
# Dataset statistics

- Novel N-gram ratio (NNG-n)
- Compression (CMP)
- Redundancy (RED-n)

Dataset	NNG-1	NNG-2	NNG-3	CMP	RED-1	RED-2
CNN/DM	13.20	52.77	72.22	90.90	13.73	1.10
XSum	35.76	83.45	95.50	90.40	5.83	0.16
XL-Sum (English)	32.22	80.99	94.57	92.97	6.56	0.20
HunSum-2	41.12	77.31	88.74	89.1	11.78	0.51

# Dataset statistics

1.82M articles after cleaning and deduplication



<i>regional</i>	346k
24.hu	307k
origo.hu	293k
hvg.hu	206k
kisalfold.hu	161k
index.hu	159k
delmagyar.hu	153k
nlc.hu	100k
nepszava.hu	28k
portfolio.hu	23k
m4sport.hu	20k
metropol.hu	12k
telex.hu	6k

# Extractive Dataset

Requires binary labeling at sentence level

- Calculate sentence embeddings for lead and article sentences
  - Using paraphrase-multilingual-MiniLM-L12-v2 from the sentence-transformers package
- Assessing similarity by cosine distance.
- Select the most similar sentence from the article for each sentence in the lead
- Maximize the total similarity across all summary sentences
  - Assignment problem

# Baseline Models

- Abstractive
  - Bert2Bert
    - Encoder-Decoder based model
    - Both side initialized with the pretrained huBERT
  - mT5
    - Pretrained multilingual transformer
- Extractive
  - BERTSum
    - BERT based encoding
    - Extra CLS tokens before each sentence
    - CLS token classification



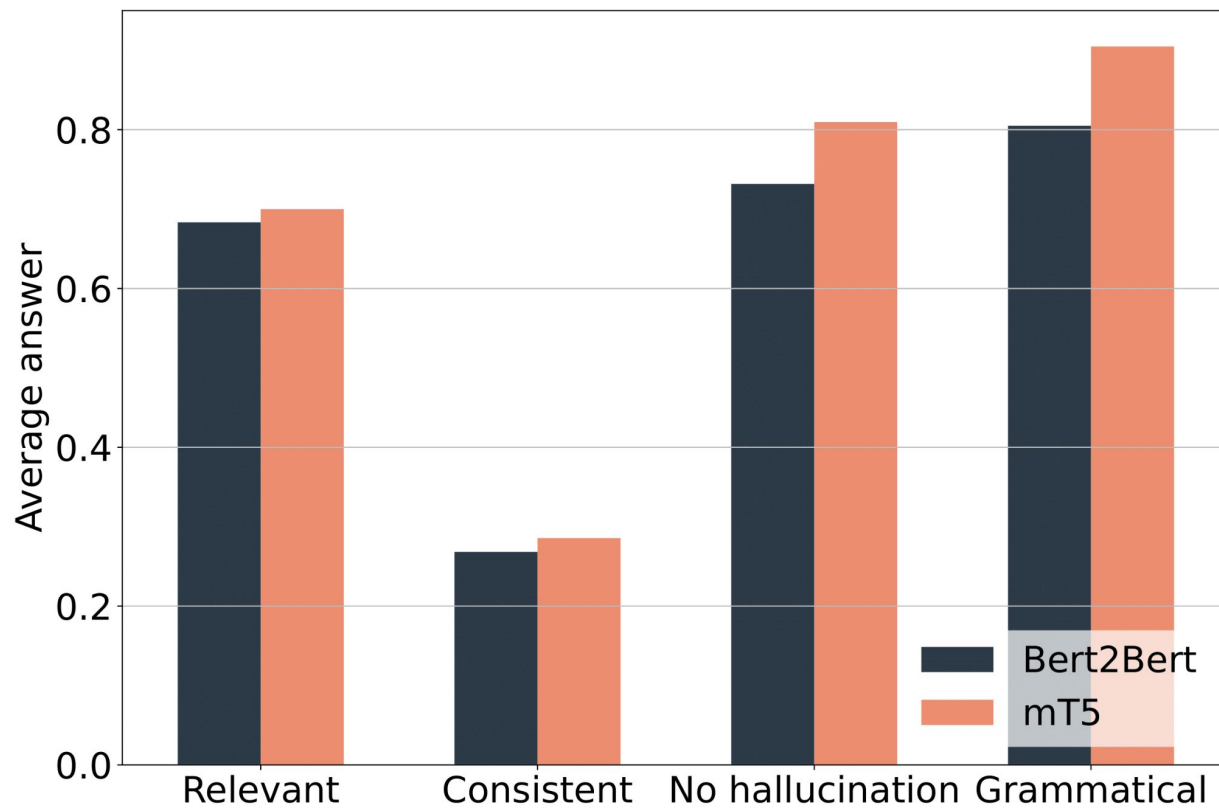
# Results

<b>Model</b>	<b>R-1</b>	<b>R-2</b>	<b>R-L</b>	<b>BertScore</b>
Bert2Bert	40.95	14.18	27.42	78.81
mT5-base	40.06	12.67	25.93	78.64
extractive	<b>49.85</b>	<b>20.12</b>	<b>33.46</b>	<b>79.18</b>
hi-mbart-large-50	31.63	13.26	22.82	77.77
hi-mt5-base	29.53	11.34	21.35	76.99
foszt2oszt	26.87	8.03	20.19	75.84

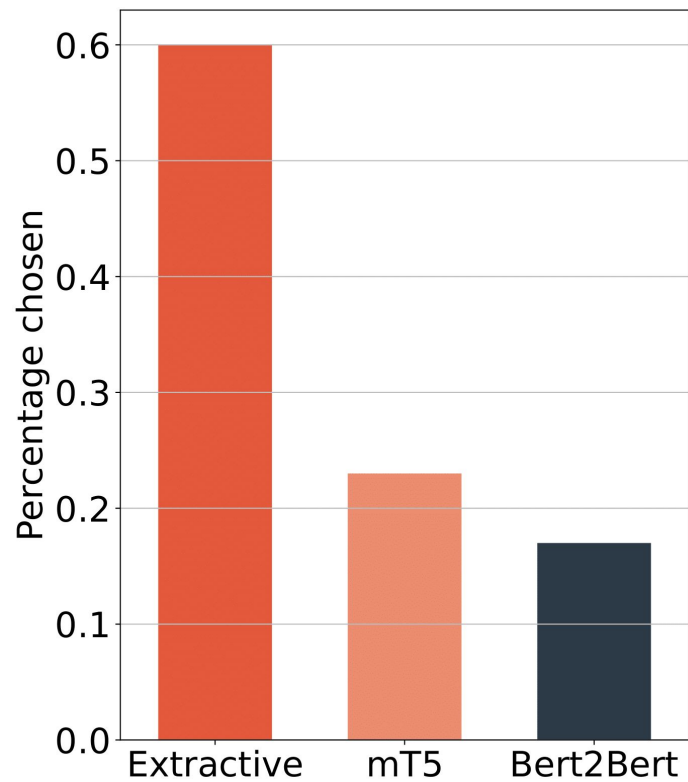
# Human evaluation

- 60 article
- 4 annotators
- Every article is evaluated by 3 annotators
- Questions:
  - **Relevant:** Does the summary convey what the article is about?
  - **Consistent:** If the answer to **Topic OK** is Yes, does the summary only contain information that is consistent with the article?
  - **No hallucination:** If the answer to **Topic OK** is Yes, does the summary only contain information that can be inferred from the article?
  - **Grammatical:** Is the summary grammatically correct?
  - **Best model:** Which model created the best summary?
- Average pairwise Cohen kappa: 60%

# Human evaluation



# Human evaluation



# Key Takeaways

- **HunSum-2**: Largest open-source Hungarian corpus for abstractive and extractive summarization
- Thorough cleaning, preprocessing and deduplication
- Baseline models
- Human evaluation on 60 articles with 4 annotators and 5 questions
- Our code: <https://github.com/botondbarta/HunSum>
- <https://huggingface.co/datasets/SZTAKI-HLT/HunSum-2-extractive>
- <https://huggingface.co/datasets/SZTAKI-HLT/HunSum-2-abstractive>