Magyar hírek automatikus kivonatolása előtanított mély nyelvmodellel – terv

Makrai Márton, **Szaszák György**
BME TMIT

ELTE DH 2021. május 20.

Áttekintés

1. [3]Magyar [4]hírek [1]kivonatolása [2]előtanított mély nyelvmodellel
2. Az automatikus kivonatolás néhány problémája

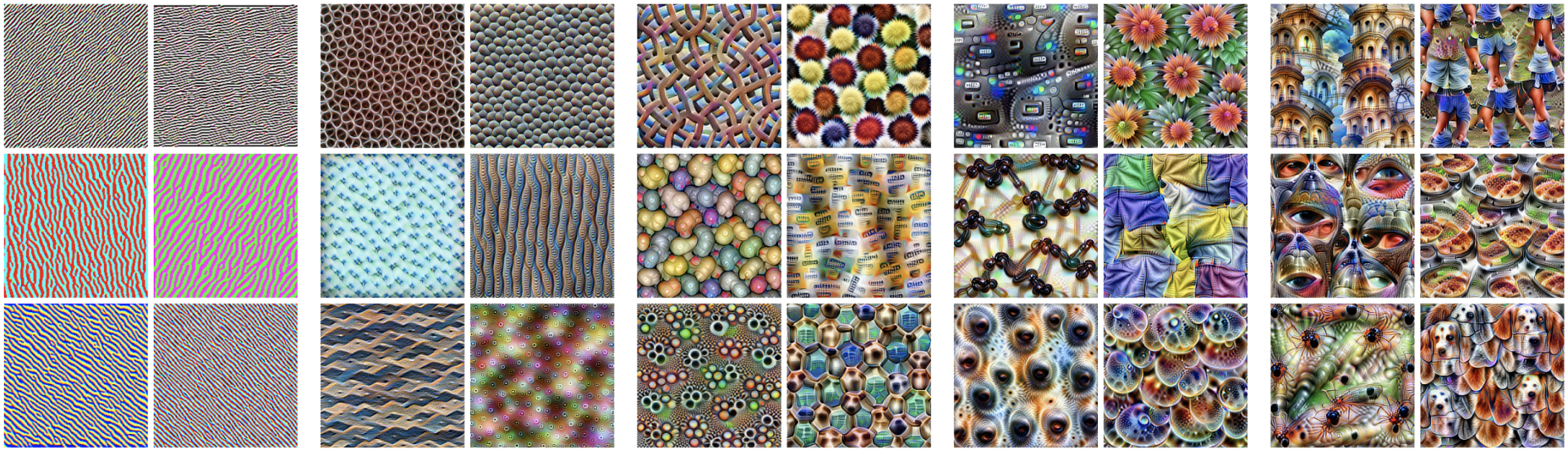
A projekt

- Szövegkivonatolás
  - hírek, tudomány, történetek, utasítások, e-mailek, szabadalmak és törvényjavaslatok
  - extraktív és absztraktív
- a projekten belüli előzmény: beszéd

előtanított mély nyelvmodellek

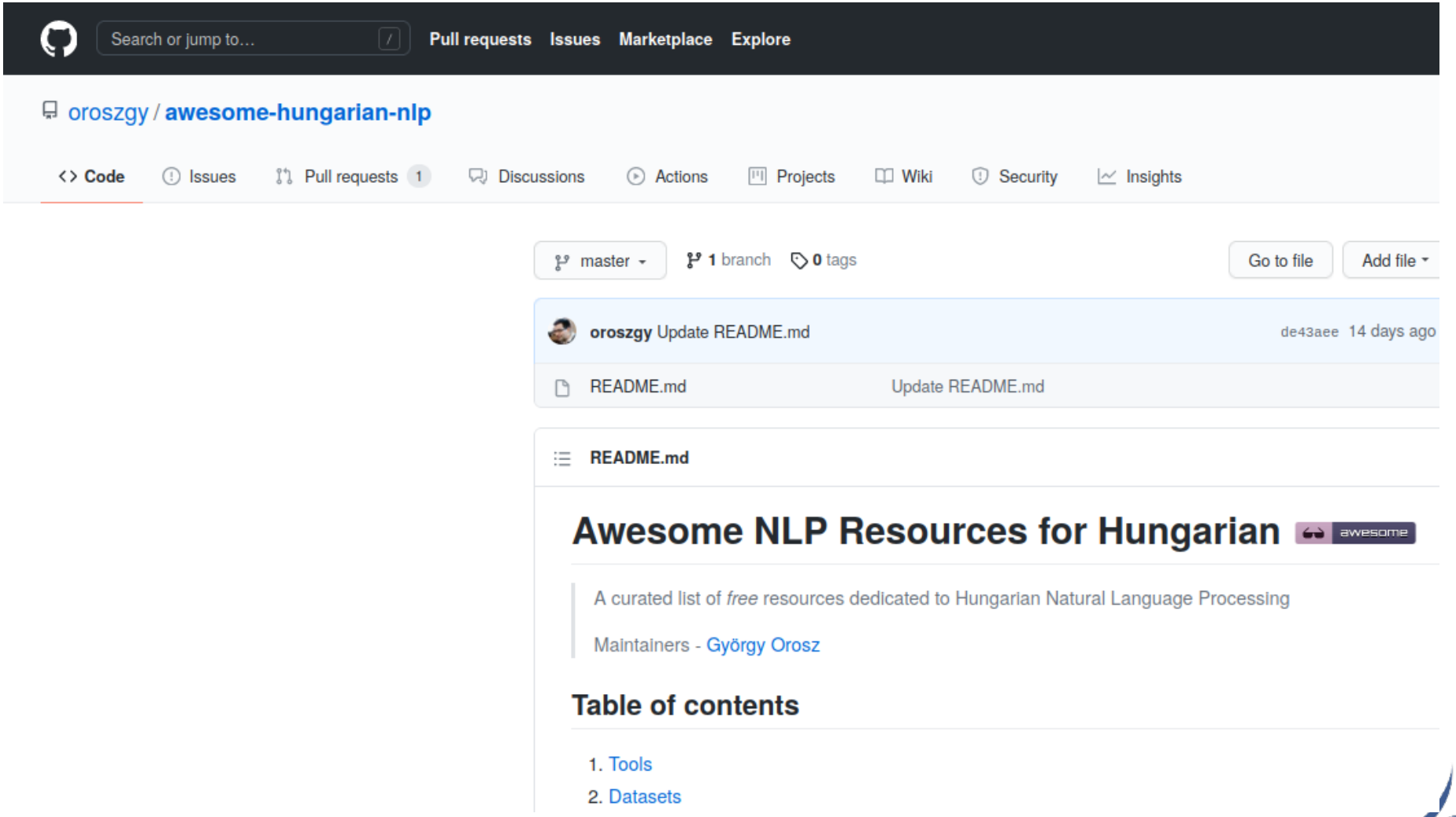
aka. kontextualizált szóbeágyazások

(Peters et al., 2018; Rogers, Kovaleva, and Rumshisky, 2020)

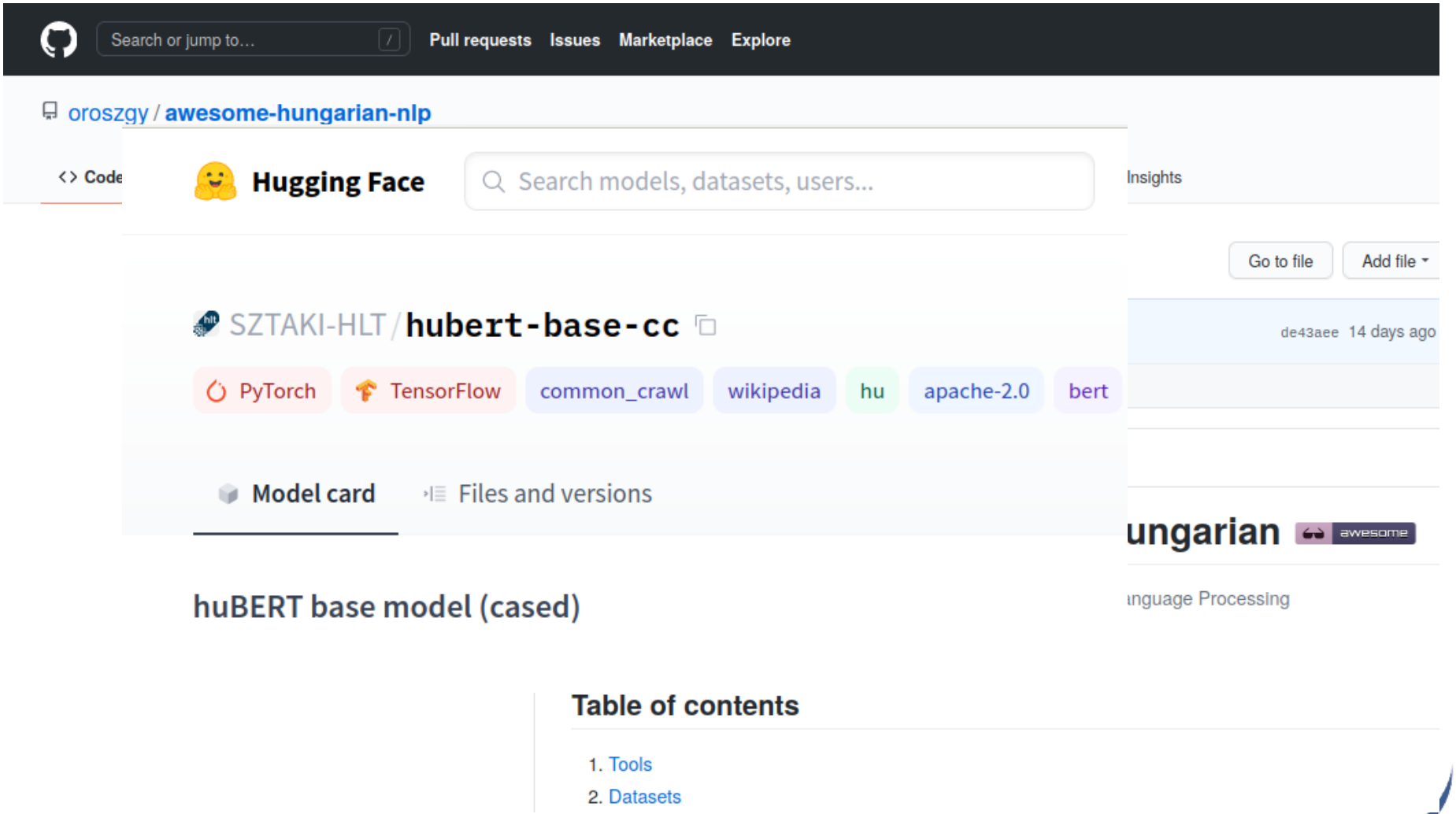


látás	élek → textúrák → minták → részek → objektumok
nyelv	morfológia → szintaxis → szemantika

magyar nyelvtechnológia, huBERT



magyar nyelvtechnológia, huBERT



Magyar hírek kivonatolása

- hír, lead
- 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

Áttekintés

1. [3]Magyar [4]hírek [1]kivonatolása [2]előtanított mély nyelvmodellel

2. Az automatikus kivonatolás néhány problémája

a kivonatolás kiértékelése (Fabbri et al., 2021)

- ROUGE: közös *n*-gramok – jól korrelál a relevanciával
  - relevancia: csak a fontos, nem hosszú, nem redundáns
- új *n*-gramok vs fedés – konzisztencia (∼precision), gördülékenység
- ismétlődő *n*-gramok – (in)koherencia
- sűrűség: a kivágatok hossza
  – koherencia, konzisztencia, gördülékenység
- tavaly óta jobbak a modellek, mint a referencia (T5, BART, Pegasus)
  - és mint az a tartomány, amire a ROUGE-t kitalálták

mennyire absztraktív, *n*-gram

- különböző szövegkonvenciók (Liu et al., 2019)
  - az elején összpontosítja a fontos információt
    ↔ egyenletesebben oszlik el az egészben
- különböző összefoglaló stílusok
  - bőbeszédű vs. távirati; extraktív vs. absztraktív
- az absztraktív összefoglalókban a másolás mennyiségét szabályozni
- tipikus felhasználási helyzetek (Song et al., 2020):
  - a gépi kivonatok nem tartalmazhatnak túl sok másolt tartalmat engedély nélkül
  - $\geqslant$ 11 egymást követő szó az EU szabványai szerint szellemi alkotás, szerzői jog védi
  - viszonylag extraktív kívánatos:
    kevésbé valószínű a tartalmi hallucináció és
    jobban megőrzi a jelentést

többnyelvű kivonatolás, MLSum (Scialom et al., 2020)

- Online újságokból nyerték, 1,5 millió + cikk/összefoglaló pár
  - 5 nyelv: francia, német, spanyol, orosz, török
  - + angol újságok a népszerű CNN / Daily mailből
- több nyelv SOTA rendszerének összehasonlító elemzése
- tudásátvivős *(transfer learning)* technikák
  - a *de facto* bevett paradigma az NLP-ben (Guzmán+ 2019)

1. nyelvmodell tanítása nagy több nyelvű korpuszon
2. modell finomhangolása egy vagy több nagy nyelv feladat-specifikus adatán
3. használat *(inference)*: a különböző előtanítási nyelvekre alkalmazzuk a modellt

- jelentős teljesítménybeli különbség az angol és a célnyelv között
- mennyi javulást hoz az önfigyelem (self-attention, ~transformer)
  - nyelvenként különböző: német > francia
  - hasonlóan, mint a gépi fordításban
  - talán a morfológia (szóalakok) miatt: szabad szórend (Döbrössy et al., 2019)

absztraktív kivonatolás előtanított enkóderrel

(Liu et al., 2019)

- az a nehéz, hogy csak az enkóder előtanított, a dekóder nem
  - ezért az enkódert kisebb tanulási rátával és simább decay-jel tanítják
- kétlépéses finomhangolás: extraktív majd absztraktív
- mennyire az elejéről veszi:

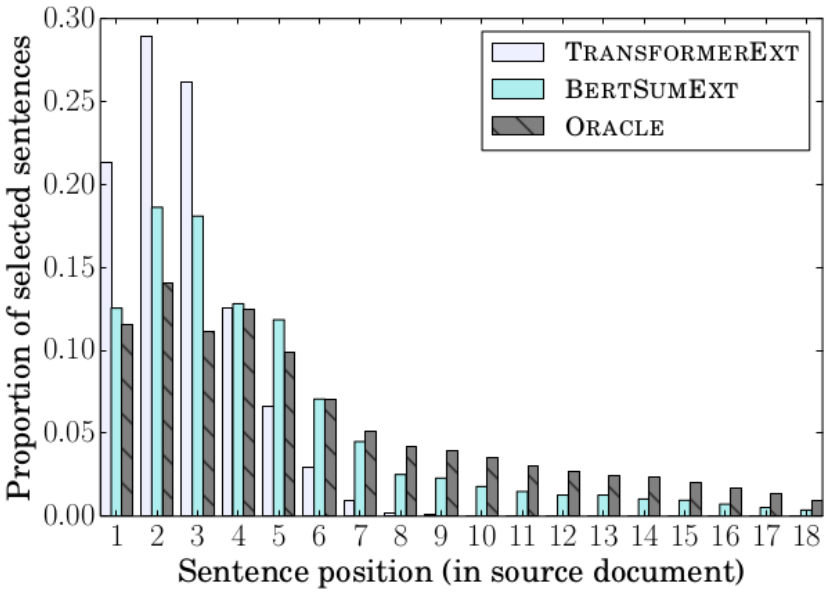
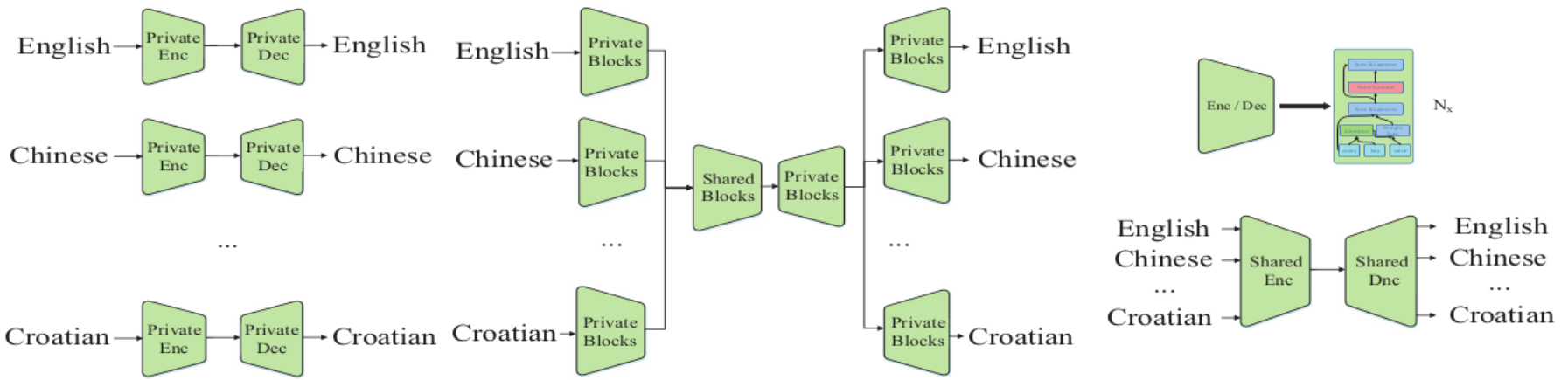


Figure 2: Proportion of extracted sentences according to their position in the original document.

bosnyák, horvát, kínai

Cao et al. (2020)

- absztraktív összefoglaló adat bosnyákra és horvátra
  177 406, illetve 204 748 példa
- kisebb (1/20), mint a gazdag erőforrású nyelvek kivonatolási adathalmazai
- maga a hír az eredeti szöveg, és a címet használja kivonatként
- többnyelvű modell kivonatolásra:



(a) Unshared encoder and decoder (b) Partially shared encoder and decoder (c) Fully shared encoder and decoder

emberi teljesítmény, kivonatolásra specifikus modellek

PEGASuS, T5, BigBird, mBART

- PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive (előtanítás mondatkitöltéssel absztraktív kivonatolásra)
  - új előtanítási feladat:
    fontos mondatok elhagyva/letakarva, a modellnek ki kell generálnia
  - bizonyos tesztekben emberi teljesítmény, és
    kis tanulóadaton (1000 példa) is jobb, mint a korábbiak
- T5
  - egységes keret:
    az összes szöveges nyelvi probléma szövegből-szöveget formátumban
  - prompt
  - SOTA eredmények
    - összefoglalás, kérdésmegválaszolás, szövegosztályozás
  - soknyelvű (101) változata is van, mT5

Irodalomjegyzék I

Cao, Yue et al. (2020). "MultiSumm: Towards a Unified Model for Multi-Lingual Abstractive Summarization". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 01, pp. 11–18 (cit. on p. 13).

Döbrössy, Bálint et al. (Aug. 2019). "Investigating Sub-Word Embedding Strategies for the Morphologically Rich and Free Phrase-Order Hungarian". In: *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*. Florence, Italy: Association for Computational Linguistics, pp. 187–193. DOI: 10.18653/v1/W19-4321. URL: https://www.aclweb.org/anthology/W19-4321 (cit. on p. 11).

Fabbri, Alexander R et al. (2021). "Summeval: Re-evaluating summarization evaluation". In: *Transactions of the Association for Computational Linguistics* 9, pp. 391–409 (cit. on p. 9).

Liu, Nelson F. et al. (2019). "Linguistic Knowledge and Transferability of Contextual Representations". In: pp. 1073–1094. DOI: 10.18653/v1/N19-1112. URL: https://www.aclweb.org/anthology/N19-1112 (cit. on pp. 10, 12).

Irodalomjegyzék II

Peters, Matthew et al. (2018). "Deep Contextualized Word Representations". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 2227–2237. DOI: 10.18653/v1/N18-1202. URL: http://aclweb.org/anthology/N18-1202 (cit. on p. 4).

Rogers, Anna, Olga Kovaleva, and Anna Rumshisky (2020). "A primer in bertology: What we know about how bert works". In: *arXiv preprint arXiv:2002.12327* (cit. on p. 4).

Scialom, Thomas et al. (2020). "Mlsum: The multilingual summarization corpus". In: *arXiv preprint arXiv:2004.14900* (cit. on pp. 7, 11).

Song, Kaiqiang et al. (2020). "Controlling the amount of verbatim copying in abstractive summarization". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 05, pp. 8902–8909 (cit. on p. 10).