# Absztraktív összefoglalás arab nyelvre

Kahla Mram[1], Yang Zijian Győző[1,2,3]

[1]Pázmány Péter Katolikus Egyetem, Információs Technológiai és Bionikai Kar
[2]MTA-PPKE Magyar Nyelvtechnológiai Kutatócsoport
1083 Budapest, Práter u. 50/a.
{kahla.mram, yang.zijian.gyozo}@itk.ppke.hu
[3]Nyelvtudományi Kutatóközpont
1068 Budapest, Benczúr u. 33.
yang.zijian.gyozo@nytud.hu

**Kivonat** Kutatásunkban arab nyelvre tanítunk különböző absztraktív összefoglaló modelleket. A jelen tanulmány a kutatásunk jelenlegi fázisát mutatja be. Arab nyelvre az absztraktív összefoglalás területén kevés kutatás történt, ezért korábbi kutatásunk során első feladatként saját adatot kellett gyűjteni. Adatgyűjtés után sikeresen finomhangoltunk különböző enkóder-dekóder architektúrájú transzformer modelleket. Kísérleteinkben kipróbáltuk a PreSumm és a többnyelvű mBART módszereket. A PreSumm módszerrel ezen a területen „state of the art" eredményt értünk el. Jelen tanulmány ezt a kutatási sorozatot folytatja. Kutatásunk során saját egynyelvű és többnyelvű BART modell tanításával kísérleteztünk, valamint az mT5 modellt próbáltuk arab összefoglaló generálásra finomhangolni. Kísérletünk során korlátozott mennyiségű adattal kísérleteztünk, célunk az volt, hogy megvizsgáljuk ezen módszerek alkalmazhatóságát. Kutatásunkkal ezért várakozásunknak megfelelően nem tudtuk felülmúlni a korábban elért legjobb eredményünket. Azonban így is versenyképes eredményeket tudtunk elérni, amelyek további kutatásoknak adnak teret, ez azonban nagyobb mennyiségű adat és infrastruktúra előfeltételt is megkövetel.
**Kulcsszavak:** arab absztraktív összefoglalás, BART, mBART, PreSumm, mT5

## 1. Bevezetés

Az összefoglaló generálás a nyelvtechnológia egyik kiemelt feladata lett. Kétféle összefoglaló módszert különböztetünk meg. Az első az extraktív, amikor a meglévő szövegből kiválasztjuk azokat a szövegrészeket, amelyek összefoglalóként funkcionálhatnak, ez gyakorlatilag egy osztályozási feladat. A másik módszer az absztraktív, amikor az emberhez hasonlóan a modell egy adott szövegből önállóan megfogalmaz egy összefoglalót. Az utóbbi módszerrel a modell olyan kifejezéseket is használhat, amelyek nem szerepeltek az eredeti szövegben. Kutatásunk elsősorban az arab nyelvre koncentrál. Az arab beszélt nyelvnek számos dialektusa van, de írás szempontjából erősen szabványosított. Ez nagyban segít az arab

írásos szövegek feldolgozásában, azonban számos nem anyanyelvű felhasználó is létrehoz elektronikus tartalmakat, ami nehezíti a szövegfeldolgozást. Az arab szövegek feldolgozását tovább nehezíti az a jelenség, hogy a rövid magánhangzók nincsenek jelölve a szövegben, ezért az olvasónak mélyebb nyelvi tudással kell rendelkeznie, ha meg szeretné érteni a szöveget. Az arab nyelv a nagy nyelvek között szerepel, nagy mennyiségű szövegadatbázissal. Szövegösszegzéssel kevesen foglalkoztak eddig, ilyen jellegű korpusz nem volt elérhető. Korábbi kutatásunkban összegyűjtöttük az első összefoglaló generálásra alkalmas arab nyelvű szövegkorpuszt, majd különböző transzformer modelleket finomhangoltunk. A jelen kutatásban saját kísérleti egynyelvű és többnyelvű BART modellekkel kísérleteztünk, illetve egy mT5 modellt finomhangoltunk absztraktív összefoglalásra.

## 2. Kapcsolódó irodalom

Arab nyelvre elsősorban extraktív összefoglalás területén végeztek kutatásokat. Az első extraktív rendszer a Lakhas (Douzidia és Lapalme, 2004), ami a szövegből 10 szót vonatolt ki összegzésnek, majd gépi fordító segítségével angol nyelvre fordította, hogy össze tudja hasonlítani a rendszert más rendszerekkel. A kiértékelést a ROUGE metrikával végezték el. Al Qassem és mtsai (2019) egy fuzzy logikán alapuló megközelítéssel, főnevek kivonatolásával végezték az összegzést. A SumSat (Lakhdar és Chéragui, 2019) három módszert ötvözve, hibrid módon összegez: egy szöveg környezetének szemantikai feltárása, indikátorként használható kifejezések kiválasztása, illetve az összefoglaló generálása a reprezentatív kifejezésekkel.

Absztraktív összefoglalás szempontjából Azmi és Altmami (2018) az extraktív összefoglaló rendszerből kiindulva egy négy lépéses absztrakt összefoglaló módszert javasol. Első lépés a téma szegmentálása, második a címsor generálása, harmadik az extraktív összefoglaló generálása, végül negyedik lépés a mondatcsökkentés. Al-Maleh és Desouki (2020) kutatásukban a cikkek első bekezdéseiből generáltak címsorokat, amelyek összefoglalóként funkcionálnak. Az összefoglaló modellhez egy enkóder-dekóder architektúrájú rekurrens hálózatot alkalmaztak. Elmadani és mtsai (2020) a PreSumm (Liu és Lapata, 2019) módszerrel és a többnyelvű BERT (Devlin és mtsai, 2019) modellel finomhangoltak extraktív és absztraktív modelleket egyaránt. Kahla és mtsai (2021) szintén kísérleteztek a PreSumm módszerrel, viszont emellett az mBART modellt is sikerült finomhangolniuk. Kutatásukban többnyelvű (cross-lingual) finomhangolásokkal növelték a rendszer teljesítményét.

Kutatásunk során Kahla és mtsai (2021) kutatását tovább gondolva egynyelvű és többnyelvű BART modelleket finomhangoltunk, illetve napjaink egyre népszerűbb mT5 modelljét próbáltuk ki.

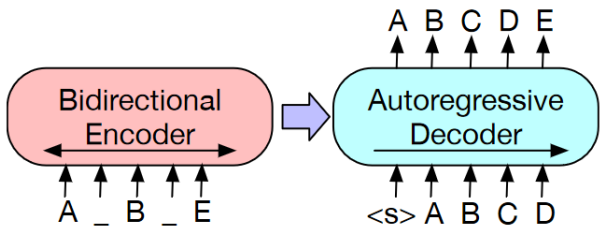
## 3. Felhasznált modellek

A **BART** (Lewis és mtsai, 2020) modell egy enkóder-dekóder architektúrán alapuló transzformer modell (lásd 1. ábra), amelyet a Fairseq (Facebook AI Rese-

arch Sequence-to-Sequence Toolkit) fejlesztett[1]. Az enkóder kétirányú (Bidirectional), a dekóder autoregresszív (Autoregressive). A korábbi kutatások alapján a csak enkóder típusú modellek (pl. BERT (Devlin és mtsai, 2019)) kiválóan alkalmasak magas minőségű szövegreprezentáció képezésére, azonban szöveggenerálás feladataira kevésbé. A csak dekóder típusú autoregresszív modellek (pl. GPT (Radford és Narasimhan, 2018)) a szöveggenerálás feladatain nyújtanak magas eredményt. A BART a két architektúra előnyeit ötvözi, ezért kiválóan alkalmas szövegösszefoglaló generálásra. Korábban kétféle BART modellt publikáltak:

– BART-base: 6 réteg enkóder és 6 réteg dekóder; 12 figyelmi fej; 768 szóbeágyazás dimenzió; bementi hossz: 512; 140 millió paraméter
– BART-large: 12 réteg enkóder és 12 réteg dekóder; 16 figyelmi fej; 1024 szóbeágyazás dimenzió; 1024 bemeneti hossz; 400 millió paraméter

Az **mBART** (Liu és mtsai, 2020) egy több nyelven előtanított BART modell. Az előtanításhoz a Common Crawl adatbázisból kivonatolt 25 nyelvet tartalmazó CC25 (Wenzek és mtsai, 2020) korpuszt használták. Az mBART modellel végzett kísérletek rávilágítottak arra, hogy abban az esetben hasznos igazán a célnyelvtől eltérő nyelveken történő előtanítás, amennyiben a célnyelven rendelkezésre álló egynyelvű adathalmaz redukált méretű. Az mBART modellel végzett munka rámutat a többnyelvű előtanításban rejlő lehetőségek transzfer tanulásos (transfer learning) irányba való felhasználhatóságára.




1. ábra: BART modell architektúrája (Lewis és mtsai, 2020).

A **T5** (Text-To-Text Transfer transzformer) (Raffel és mtsai, 2020) a Google által készített enkóder-dekóder típusú modell. Az utóbbi időben a nyelvtechnológia területén kiemelt jelentőséggel bír a transzfer tanulás, amelynek során a nyelvi modellt egy adatokban gazdag feladaton tanították be, majd ezt követően került finomhangolásra egy soron következő célfeladatra. Ideális esetben a modell az előtanítás során olyan általános tudásra tesz szert, amely átvihető, és sikeresen alkalmazható a célfeladatok megoldásában. A T5 projekt alapötlete, hogy minden szövegelemzési feladatot (fordítás, kérdések megválaszolása, osztályozás stb.) szövegből szöveg (text-to-text) problémaként közelít meg, azaz

[1] https://github.com/pytorch/fairseq/tree/master/examples/bart

szöveg a bemenet és a modell ez alapján szöveget generál kimenetként (lásd 2). Itt fontos kiemelni a BERT-alapú modellekkel szemben mutatkozó alapvető különbséget a felépítésben: a T5 esetében mind a bemenet, mind pedig a kimenet szöveg formátumú, míg a BERT-alapú modellek esetében a bemenet szöveges, a kimenet azonban vagy egy osztályozó címke vagy pedig csak valamilyen bemenetből származó töredék.

A T5 projekt elsődleges célja nem az, hogy új módszerek kerüljenek kifejlesztésre, a munka mögött álló csapat elsődleges motivációja az, hogy bemutassák a terület jelenlegi állását, és összehasonlítsák az elérhető technikákat. Emellett a jelenlegi megközelítések határait is próbálják megállapítani azáltal, hogy szisztematikus módon és nagy mértékben megnövelt paraméterszámmal (modellek betanítása 11 milliárd paraméterig) kísérleteznek. A modell tanításához felhasznált korpusz a Colossal Clean Crawled Corpus (rövidítve C4), amely egy több száz gigabájtnyi világhálóról összegyűjtött és tisztított angol nyelvű szöveget tartalmaz. Az T5 esetében a paraméterek száma alapján 5 különböző méretű modell került betanításra:

– Small (300 millió paraméter), Base (580 millió paraméter), Large (1,2 milliárd paraméter), XL (3,7 milliárd paraméter), XXL (13 milliárd)

Az **mT5** (Xue és mtsai, 2021) a T5 több nyelvre kiterjesztett verziója. Az mT5 létrehozása során a szerzők törekedtek arra, hogy minél inkább megőrizzék a T5 strukturális jegyeit, ezért az mT5 örökölte a szövegből szöveg (text-to-text) tulajdonságot és az általános előtanítás menetét is, amelyhez szintén rendkívül nagy méretű korpuszt használtak.

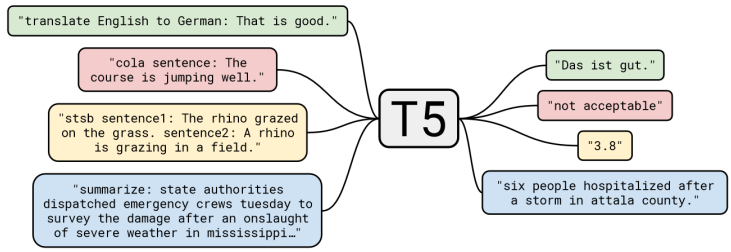
Az mT5 betanításához az mC4 korpuszt használták, amely a T5 tanítására alkalmazott C4 többnyelvű változata. Az mC4 101 különböző nyelvű szövegeket tartalmaz. A T5 modellel összehasonlítva az mT5 nagyobb paraméterszámokkal rendelkezik, ez a nagyobb szótárméret következménye. Fontos megjegyezni, hogy T5-alapú modellek az enkóder-dekóder struktúrát követik, ezért paraméterszámuk általában kétszer akkora, mint egy hasonló méretű csak enkóder struktúrájú modell.

Az mT5 kutatás megmutatta, hogy a T5 modell kiválóan alkalmazható többnyelvű kontextusban is, továbbá rendkívül magas eredményeket tud elérni különböző referenciafeladatokban.

## 4. Felhasznált korpuszok

Az egynyelvű BART modellünk tanításához erőforrás hiányában jelen kutatáshoz egy csökkentett méretű arab Wikipédia szöveget használtunk. A korpusz előállításához első körben 150.000 szegmensnyi szöveget töltöttünk le. A BART tanításához olyan bekezdések kellenek, amelyek legalább két pont írásjellel rendelkező mondatot tartalmaznak. Ezért első lépésként ez alapján szűrtük a szöveget, a szűrés után 9.773 bekezdésünk maradt.

A többnyelvű BART modellünk tanításához angol, magyar és arab Wikipédiából vett bekezdéseket vettünk, amelyek minimum kettő pont írásjellel rendelkeztek (ez az elvárás a BART esetében). A kiegyensúlyozottság végett 10.000




2. ábra: T5 modell (Raffel és mtsai, 2020).

szegmenst vettünk mind az angol mind a magyar korpuszból. Az 1. táblázatban láthatóak az előtanításhoz használt korpuszokra jellemző kvantitatív tulajdonságok.

| | Arab | Angol | Magyar |
|---|---|---|---|
| Szegmens | 9.773 | 10.000 | 10.000 |
| Token | 761.371 | 1.357.875 | 818.420 |
| Type | 108.982 | 60.248 | 139.996 |
| Átlagos mondatszám | 1,73 | 5,06 | 4,36 |
| Átlagos tokenszám | 77,897 | 135,78 | 81,84 |

1. táblázat. Előtanításhoz felhasznált korpuszok tulajdonságai.

A finomhangoláshoz ugyanazt a korpuszt (Arab-Szum) használtuk, mint az előző kutatásunkban (Kahla és mtsai, 2021). Továbbá végeztünk transzfer tanulásos kísérletet is, amihez vegyesen válogattunk angol és magyar szegmenseket (Multi-Szum). A kiegyensúlyozottság végett 20.000 angol és 20.000 magyar szegmens került kiválasztásra a finomhangolási korpuszba. Az angol szegmenseket a CNN/Daily Mail korpuszból (Nallapati és mtsai, 2016) vettük, míg a magyar szegmenseket a HVG korpuszból (Yang és mtsai, 2021). A jelen kutatáshoz kiválasztott finomhangolási korpusz részkorpusza az előző kutatáshoz használt angol és magyar finomhangolási korpuszoknak. A 2. táblázatban láthatóak a korpuszokra jellemző kvantitatív tulajdonságok.

## 5. Kísérletek

BART kutatásunk során előtanítottunk egy egynyelvű és egy többnyelvű (angol, magyar, arab) BART base modellt. A Facebook nem tette közzé az előtanítás implementációját, ezért a Hugging Face transzformers[2] könyvtárai által biztosított előtanítási függvényeket használtuk. A BART előtanításához a BartFor-

[2] https://huggingface.co/transzformers/model_doc/bart.html

| | Arab | | Angol | | Magyar | |
|---|---|---|---|---|---|---|
| | Cikk | Lead | Cikk | Lead | Cikk | Lead |
| Szegmens | 21.508 | | 20.000 | | 20.000 | |
| Token | 6.929.974 | 2.867.754 | 15.795.098 | 1.050.273 | 5.387.638 | 602.136 |
| Type | 290.138 | 178.614 | 169.709 | 56.902 | 397.628 | 99.166 |
| Átlagos mondatszám | 14,42 | 1,47 | 28,69 | 1 | 11,19 | 1,56 |
| Átlagos tokenszám | 412,05 | 35,131 | 789,76 | 52,51 | 269,38 | 30,10 |

2. táblázat. Finomhangoláshoz felhasznált korpuszok tulajdonságai.

CausalLM[3] függvényt használtuk. A BartForCausalLM a BART modell dekóder önálló része, melynek a tetején egy nyelvmodell réteg található. Ez alkalmas a következő szó prediktálására (causal language modeling). A modell tovább finomhangolható BART finomhangolási feladatokra. A kutatásunk során egy kísérleti arab egynyelvű és egy kísérleti háromnyelvű BART base modellt tanítottunk elő:

- **Arab BART**: Egynyelvű arab BART base modell, 512 bemeneti szöveghossz, közel 19.808 bekezdésnyi arab Wikipédia szövegen tanítva. A szótárméret: 40.000.
- **Multi BART**: háromnyelvű BART base modell, 512 bemeneti szöveghossz, bekezdés alapú Wikipédia szövegeken tanítva: 19.808 arab, 20.000 angol és 20.000 magyar. A szótárméret: 50.000.

Az egynyelvű BART modell előtanításához az alábbi hiperparamétereket használtuk: 512 bemeneti szöveghossz; batch méret: 8/GPU (4 db GeForce GTX 1080 + 4 db GeForce RTX 2080); epoch szám: 50; tanulási ráta: 2e-6; fp16.

A többnyelvű BART modell előtanításához az alábbi hiperparamétereket használtuk: 512 bemeneti szöveghossz; batch méret: 6/GPU (4 db GeForce GTX 1080 + 4 db GeForce RTX 2080); epoch szám: 50; tanulási ráta: 8e-7; fp16.

Finomhangolásos kísérleteinkben kettő BART modellt tanítottunk:

- **BART arab szum**: Arab BART finomhangolva Arab-Szum korpuszon.
- **BART multi transz**: Multi BART finomhangolva a Multi-Szum korpuszon, majd azt tovább finomhangoltuk az Arab-Szum korpuszon.

Az egynyelvű BART finomhangolásához az alábbi hiperparamétereket használtuk: 512 maximum bemeneti és 256 maximum kimeneti szöveghossz, batch méret: 8/GPU (4 db GeForce GTX 1080 - 12GB) méret, epoch szám: 120, tanulási ráta: 2e-5, warmup lépés: 5000; fp16.

A többnyelvű kísérlet során először finomhangoltuk a Multi BART modellünket a háromnyelvű Multi-Szum korpuszon az alábbi hiperparaméterekkel: 512 maximum bemeneti és 256 maximum kimeneti szöveghossz; batch méret 5/GPU (4 db GeForce GTX 1080 + 4 db GeForce RTX 2080); epoch szám: 40; tanulási ráta: 5e-5, warmup lépés: 5000; fp16.

Majd a többnyelvű finomhangolás után tovább finomhangoltuk az Arab-Sum korpuszon, az alábbi hiperparaméterekkel: 512 maximum bemeneti és 256 maximum kimeneti szöveghossz; batch méret 5/GPU (4 db GeForce GTX 1080 +

[3] https://huggingface.co/transzformers/model_doc/bart.html#bartforcausallm

4 db GeForce RTX 2080); epoch szám: 80; tanulási ráta: 5e-5, warmup lépés: 5000; fp16.

A többnyelvű kísérletek összeségében szintén 120 (40+80) epoch szám mellett tanultak. Azt tapasztaltuk, hogy a magas epoch szám nem okoz túltanulást, inkább egyre finomabb dolgokat tanult meg.

Végül kísérleteztünk az mT5 modellel:

– **mT5 arab szum**: mT5 small modell finomhangolása az Arab-Szum korpuszon.

Az mT5 finomhangolása alábbi hiperparaméterekkel történt: 512 maximum bemeneti és 256 maximum kimeneti szöveghossz; batch méret 2/GPU (4 db GeForce GTX 1080); epoch szám: 40; tanulási ráta: 2e-5, warmup lépés: 5000; prefix: "summarize: ". Az fp16 paramétert nem használtuk, mivel a T5 típusú modellek esetében az fp16 használatával nem konvergál a tanítás. Az erőforrásaink korlátai miatt nem tudtuk az mT5 nagyobb modelljeit kipróbálni.

## 6. Eredmények

A 3. és a 4. táblázatban láthatóak a modellek tulajdonságai, illetve a méréseink eredményei. A dupla vonal alatti modelleket tanítottuk a jelen kutatásunkban. A „+” jel jelöli azokat a modelleket, amelyekhez a mostani kutatásunk során előtanítást is végeztünk. Korábbi kutatásunkban a PreSumm sajátosságai miatt a fedés mértékeket publikáltuk, azonban a nemzetközi sztenderd szerint az F-mérték a mérvadó, ezért a modelleket újra teszteltük az F-mérték alapján. Röviden a modellekről:

– *AraBERT*: AraBERT (Antoun és mtsai, 2020) finomhangolása arab korpuszon PreSumm eszközzel
– *mBERT*: többnyelvű BERT (Devlin és mtsai, 2019) finomhangolása arab korpuszon PreSumm eszközzel
– *mBERT + hun*: mBERT finomhangolása magyar HVG korpuszon (Yang és mtsai, 2021) PreSumm eszközzel, majd tovább finomhangolása arab korpuszon
– *mBERT + eng*: mBERT finomhangolása angol CNN/Daily Mail korpuszon PreSumm eszközzel, majd tovább finomhangolása arab korpuszon
– *mBART-50*: mBART-50 (Tang és mtsai, 2020) finomhangolása arab korpuszon
– *mBART-50-rus*: Gazeta korpuszon (Gusev, 2020) (52.400 szegmens) finomhangolt mBART-50, majd tovább finomhangolva arab korpuszon

Az összevethetőség végett a 3. táblázatban feltüntettük a paraméterszámokat, az általunk felhasznált előtanításhoz (Elő) és finomhangoláshoz (Finom) használt korpuszok méreteit és azt, hogy melyik modell milyen nyelvi tudással rendelkezik.

Korábbi kutatásunkból az látható, hogy az elő-finomhangolt (más nyelven transzfer tanulással) modellekkel tudtunk növelni a rendszer minőségén.

| Modell | Paraméter # | Elő (token) | Finom (szegmens) | Nyelv |
|---|---|---|---|---|
| AraBERT | 136 millió | - | 19.808 | arab |
| mBERT | 110 millió | - | 19.808 | 104 nyelv |
| mBERT+hun | 110 millió | - | 442.739+19.808 | 104 nyelv |
| mBERT+eng | 110 millió | - | 286,817+19.808 | 104 nyelv |
| mBART-50 | 610 millió | - | 19.808 | 50 nyelv |
| mBART-50-rus | 610 millió | - | 19.808 | 50 nyelv |
| + BART arab szum | 140 millió | 761.371 | 19.808 | arab |
| + BART multi transz | 140 millió | 2.937.666 | 59.808+19.808 | arab, angol, magyar |
| mt5 arab szum | 300 millió | - | 19.808 | 101 nyelv |

3. táblázat. Modellek tulajdonságai.

A mostani eredményekből az látható, hogy az általunk tanított kísérleti BART modellek nem tudják felülmúlni a korábbi kutatásunkban elért eredményeket. Ez várakozásunknak megfelel, hiszen azokat a modelleket sokkal nagyobb adathalmazon tanították elő. Mind a többnyelvű BERT, mind a huBERT, vagy az mBART óriási mennyiségű adaton tanult szemben a körülbelül 30.000 szegmensű Wikipédia korpuszunkkal, de csak kevés értékkel marad le. Azonban AraBERT modellt így is szignifikánsan felülmúlja. Továbbá a korábbi kutatásunk tapasztalata, miszerint többnyelvű transzfer tanulással tovább növelhető a rendszer teljesítménye, újra bebizonyosodott. Az angol-magyar adatokkal hozzáadott korpuszon való elő-finomhangolás javított a rendszer minőségén.

Manuálisan vizsgálva az összefoglalókat, azt figyeltük meg, hogy az összefoglalók nyelvtanilag helyesek, a BART modellek kevés hibát vétenek és a témakör szintjén relevánsak. Az egyetlen típushiba, hogy gyakran belekever olyan elemeket, amelyek szemantikailag nem helytállóak.

Ezzel a kutatással bebizonyosodott, hogy képesek vagyunk saját BART modellt tanítani, valamint nagyobb erőforrás és tanítóanyag mellett tovább növelhető a modellek minősége.

| Modell | ROUGE-1 | ROUGE-2 | ROUGE-L |
|---|---|---|---|
| AraBERT | 0,772 | 0,008 | 0,772 |
| mBERT | 4,264 | 0,164 | 4,264 |
| mBERT+hun | 4,909 | 0,178 | 4,903 |
| mBERT+eng | 12,610 | 2,107 | 12,610 |
| mBART-50 | 5,952 | 0,312 | 5,921 |
| mBART-50-rus | 7,145 | 0,766 | 7,101 |
| + BART arab szum | 3,066 | 0,023 | 3,007 |
| + BART multi transz | 3,895 | 0,114 | 3,877 |
| mt5 arab szum | 6,851 | 0,294 | 6,840 |

4. táblázat. Arab összefoglaló generálás F-mérték eredmények.

Végül, de nem utolsósorban azt láthatjuk az mT5 small modell finomhangolásának eredményében (lásd 4. táblázat), hogy minőségében felülmúlja az mBERT,

mBERT + hun és az mBART-50 modelleket is. Elő-finomhangolás nélkül közel olyan magas eredményt ér el, mint az mBART-50-rus. Fontos megjegyezni, hogy ez egy small modell, ami paramétereit tekintve sokkal kisebb mint az mBART. Ezzel a méréssel azt láthatjuk, hogy nagyobb erőforrás mellett és esetleg nagyobb epoch szám mellett további eredményjavulást tudunk elérni.

További példák és modellek a projekt oldalunkon[4] érhetőek el.

## 7. Összegzés

Kutatásunk során arab nyelvre tanítottunk különböző transzformer modelleket absztraktív összefoglaló generálás feladatában. A jelen tanulmány egy pillanatképet mutat a kutatásunk jelenlegi fázisáról, amelyben saját egynyelvű és többnyelvű BART előtanításával és finomhangolásával kísérletezünk. Továbbá a napjaink egyik népszerű T5 többnyelvű modelljét is finomhangoltuk. Kutatásunkat ezen fázisában kevés erőforrással és tanítóanyaggal végeztük el, ezért várakozásunknak megfelelően eredményeinkkel nem tudtuk felülmúlni a korábbi „state of the art" eredményét. Azonban bemutattuk, hogy így is versenyképes teljesítményt tudtak nyújtani, ami kiváló alapot képez a nagyobb erőforrással való kísérletek számára.

A jövőben a jelen tanulmányban bemutatott kísérleteket fogjuk elvégezni nagy teljesítményű szuperszámítógépeken és nagy mennyiségű adatokon.

## Hivatkozások

Al-Maleh, M., Desouki, S.: Arabic text summarization using deep learning approach. Journal of Big Data 7, 1–17 (2020)

Al Qassem, L., Wang, D., Barada, H., Al-Rubaie, A., Almoosa, N.: Automatic Arabic text summarization based on fuzzy logic. In: Proceedings of the 3rd International Conference on Natural Language and Speech Processing. pp. 42–48 (2019)

Antoun, W., Baly, F., Hajj, H.: AraBERT: Transformer-based model for Arabic language understanding. In: Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection. pp. 9–15. European Language Resource Association, Marseille, France (May 2020)

Azmi, A.M., Altmami, N.I.: An abstractive arabic text summarizer with user controlled granularity. Information Processing and Management 54(6), 903–921 (2018)

Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: 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). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019)

[4] http://nlpg.itk.ppke.hu/projects/summarize

Douzidia, F.S., Lapalme, G.: Lakhas, an Arabic summarization system. Proceedings of DUC2004 (2004)

Elmadani, K.N., Elgezouli, M., Showk, A.: BERT fine-tuning for Arabic text summarization. ArXiv abs/2004.14135 (2020)

Gusev, I.: Dataset for automatic summarization of russian news. In: Artificial Intelligence and Natural Language. pp. 122–134. Springer International Publishing, Cham (2020)

Kahla, M., Yang, Z.G., Novák, A.: Cross-lingual fine-tuning for abstractive Arabic text summarization. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021). pp. 655–663. INCOMA Ltd., Held Online (Sep 2021)

Lakhdar, S.M., Chéragui, M.A.: Building an extractive Arabic text summarization using a hybrid approach. In: International Conference on Arabic Language Processing. pp. 135–148. Springer (2019)

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)

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. pp. 3730–3740. Association for Computational Linguistics, Hong Kong, China (2019)

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., Zettlemoyer, L.: Multilingual Denoising Pre-training for Neural Machine Translation. Transactions of the Association for Computational Linguistics 8, 726–742 (11 2020)

Nallapati, R., Zhou, B., dos Santos, C., Caglar, G., Xiang, B.: Abstractive text summarization using sequence-to-sequence RNNs and beyond. In: Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning. pp. 280–290. Association for Computational Linguistics, Berlin, Germany (Aug 2016)

Radford, A., Narasimhan, K.: Improving language understanding by generative pre-training (2018)

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)

Tang, Y., Tran, C., Li, X., Chen, P.J., Goyal, N., Chaudhary, V., Gu, J., Fan, A.: Multilingual translation with extensible multilingual pretraining and finetuning (2020)

Wenzek, G., Lachaux, M.A., Conneau, A., Chaudhary, V., Guzmán, F., Joulin, A., Grave, E.: CCNet: Extracting high quality monolingual datasets from web

crawl data. In: Proceedings of the 12th Language Resources and Evaluation Conference. European Language Resources Association, Marseille, France (May 2020)

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., Raffel, C.: 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. pp. 483–498. Association for Computational Linguistics, Online (Jun 2021)

Yang, Z.G., Agócs, Á., Kusper, G., Váradi, T.: Abstractive text summarization for hungarian. Annales Mathematicae et Informaticae 53, 299–316 (2021)