Többnyelvű modellek és PEGASUS finomhangolása magyar nyelvű absztraktív összefoglalás feladatára

Yang Zijian Győző

Nyelvtudományi Kutatóközpont
1068 Budapest, Benczúr u. 33.
yang.zijian.gyozo@nytud.hu

Kivonat Napjaink egyik legfontosabb és legkutatottabb nyelvtechnológiai területe az absztraktív szövegösszefoglaló készítése. Mind a kutatásban, mind az iparban egyre nagyobb igény keletkezik a feladat megoldására. Az elmúlt években magyar nyelven is elindultak a kutatások ezen a területen, voltak különböző kísérletek magyar és többnyelvű előtanított neurális nyelvmodellek finomhangolásával. Jelen kutatásomban elsősorban a többnyelvű modellek finomhangolására tettem a hangsúlyt. Arra kerestem a választ, hogy a más nyelvekre, akár feladatokra előtanított modellek hogyan teljesítenek magyar nyelvre, illetve azok a többnyelvű modellek, amelyek angol vagy más nyelven a legjobb eredményt érték el absztraktív összefoglalás területén, adaptálhatóak-e magyar nyelvre. A kísérletem kiterjedt a manapság rendkívül népszerű mT5-re, a magyar nyelvi előtudással nem rendelkező mBART modellre és az M2M100 gépi fordítás feladatára előtanított 100 nyelvű neurális modellre. Az utóbbi két modell esetén a kérdés, hogy egy modell, amely nem rendelkezik magyar tudással a finomhangolás során meg tud-e tanulni magyarul megoldani egy feladatot, illetve, bár rendelkezik magyar tudással, de gépi fordításra tanított modell módosítható-e absztraktív összefoglaló generálás feladatára. Végül, de nem utolsó sorban, az angol nyelvre egyik legjobban teljesítő PEGASUS modellt finomhangoltam magyar absztraktív összefoglaló feladatra. Ezzel a kutatással kísérletet tettem egy angol nyelvű modellt magyar nyelvre adaptálni és arra kerestem a választ, hogy vajon ez lehetséges-e és van-e értelme. Eredményeim azt mutatják, hogy mindegyik modell finomhangolható és adaptálható magyar nyelvre, sőt az mT5 és az mBART esetében sikerült felülmúlni az eddigi legjobban teljesítő magyar BART modellt.
Kulcsszavak: absztraktív összefoglalás, mT5, mBART, M2M100, PEGASUS

1. Bevezetés

Napjaink egyik legnépszerűbb nyelvtechnológiai feladata a hosszú szövegek összefoglalása. Mind a kutatásban, mint az iparban kiemelt terület. Külön cégek[1]

[1] https://www.getabstract.com

alakulnak a feladat megoldására. Kétféle szövegösszefoglaló módszer létezik: extraktív és absztraktív. Az extraktív összefoglalás esetében a feladat megjelölni, kiemelni a szövegben lévő fontosnak ítélt szövegrészeket. Az absztraktív összefoglalás esetében viszont a bemeneti szöveg alapján egy új szöveg megfogalmazása a cél, amely tartalmazza a bemeneti szöveg tartalmi lényegét. Mindkettő feladat nehéz, mivel emberek között sincsen minden esetben teljes egyetértés abban, hogy mi a fontos egy szövegben, vagy mi a legfontosabb mondanivalója egy szövegnek. A feladatot tovább nehezíti, hogy az is számít milyen célra készül az összefoglaló. Egy marketing célra használt, figyelemfelkeltő összefoglaló mást tart fontosnak, mint egy szakmai összefoglaló. Mindezen szempontok mellett az absztraktív összefoglaló feladata annyiban nehezedik, hogy a meglévő szöveg alapján egy összefüggő, logikusan felépített, nyelvtanilag helyes szöveget kell generálnia. Azonban egy jól működő automatikus absztraktív összefoglaló szoftver rengeteg időt és energiát megspórolhat egy cég számára.

Kutatásomban első sorban az absztraktív összefoglalás módszerére tettem a hangsúlyt. Kísérleteimben a témában népszerű mT5 és mBART többnyelvű modelleket finomhangoltam magyar nyelvre. A feladat érdekessége, hogy az mBART modell nem rendelkezik magyar tudással, azonban a szóelem tokenizálásnak köszönhetően a modell könnyen be tudja olvasni a magyar szöveget is. Emellett kísérletet tettem egy nem összefoglalásra szánt modell finomhangolására. Kutatásomban egy gépi fordító modellt, az M2M100 modellt próbáltam ki. Az M2M100 rendelkezik magyar tudással, azonban gépi fordítás feladatára tanították. Kutatásomban arra kerestem a választ, hogy kell-e magyarul tudnia egy modellnek, hogy magyar feladatot oldjon meg, illetve számít-e a feladat, vagy inkább a modell architektúrája a fontosabb. Végül, de nem utolsó sorban kísérletemben azt kutattam, hogy egy angol nyelvre előtanított modell, a PEGASUS, tovább finomhangolható-e magyar nyelvű feladatra, illetve milyen módszereket kell alkalmazni, hogy egyáltalán képes legyen valamennyire magyarul megtanulni.

Modelljeim elérhetőek a Hugging Face oldalunkon[2]: NYTK/summarization-hi-mbart-large-50-hungarian, NYTK/summarization-hi-mt5-base-hungarian és NYTK/summarization-hi-pegasus-hungarian.

2. Kapcsolódó irodalom

Az utóbbi években mind a nyelvtechnológiában, mind az absztraktív összefoglalás területén, a transzformer (Vaswani és mtsai, 2017) architektúra hozta a nagy áttörést. A transzformer architektúra megjelenésével egy időben a nyelvtechnológiai feladatok folyamata is megváltozott. Az új irány a kétlépcsős tanítás lett. Első lépésként egy általános nyelvmodellt tanítanak, amellyel a modellt felruházzák általános nyelvi tudással, ez az előtanítás folyamata, majd második lépésként az előtanított nyelvmodellt tovább tanítják, vagy másnéven finomhangolják egy adott konkrét feladatra. Bizonyos esetekben, hogy jobb eredményt érjenek el,

[2] https://huggingface.co/NYTK

az előtanítás folyamata során olyan feladatokkal tanítják a modellt, hogy azt könnyebb legyen egy adott feladatra, például összefoglaló generálásra, tanítani.

A szövegösszefoglaló generálás egy szöveggenerálási feladat, ami megkövetel a modelltől egy generálásra alkalmas architektúrát, amit dekódernek hívnak. A csak dekóderrel rendelkező modellek, mint például a GPT (Radford és Narasimhan, 2018; Radford és mtsai, 2019; Yang, 2022a), bár képesek szövegösszefoglalásra, nem teljesítenek olyan jól, mint az enkóder-dekóder architektúrájú modellek. Az első sikeres próbálkozások is enkóder-dekóder architektúrával történtek. 2019-ben a PreSumm (Liu és Lapata, 2019) módszerrel érték el a legjobb eredményt. A PreSumm módszer lényege, hogy egy előtanított BERT modellhez (Devlin és mtsai, 2019), csatlakoztatnak egy üres dekódert, amelyet finomhangolnak. Magyar nyelvre is ezzel a módszerrel érték el az első sikereket (Yang és mtsai, 2021), ahol enkódernek a huBERT (Nemeskey, 2021) modellel sikerült a legjobb eredményeket elérni.

Későbbiekben olyan autoregresszív modellekkel érték el a legjobb eredményeket, ahol már az előtanítás során is egy enkóder-dekóder architektúrát alkalmaztak. Az egyik legjobb eredményt elért modell 2020-ban a BART (Lewis és mtsai, 2020) modell volt, ahol az előtanítást úgy végezték el, hogy az az összefoglaló generálás finomhangolásának kedvezzen. Ehhez hasonló a PEGASUS is, ami csak az összefoglaló generálás feladatára alkalmas és több versenyben is a legjobb eredményt produkálta (Zhang és mtsai, 2020).

A BART angol nyelvű modell, ezért, hogy kiterjesszék a módszer előnyeit más nyelvekre is, elkészítették először az mBART (Liu és mtsai, 2020), majd az mBART-50 (Tang és mtsai, 2020) modelleket. A sima mBART 25, míg az mBART 50 nyelv tudását tartalmazza, azonban nincsen közöttük a magyar nyelv.

2021-ben Hasan és mtsai (2021) létrehozták az XL-Sum összefoglaló generálásra szánt korpuszt, ami 44 nyelvet tartalmaz. A korpusz közel egymillió BBC cikket és azok leadjeit tartalmazza. A korpusz segítségével a kevés forrással rendelkező arab nyelvekre sikerült jó minőségű szövegösszefoglaló modelleket tanítaniuk. A magyar nyelv azonban nincsen közöttük.

Pfeiffer és mtsai (2020) többnyelvű transzfer (cross-lingual transfer) tanuláson alapuló módszerekkel kísérleteztek többnyelvű modellek segítségével (Pfeiffer és mtsai, 2021, 2022).

Magyar nyelvre transzformer alapú neurális összefoglaló modelleket először Yang és mtsai (2021) készítettek a PreSumm módszerrel. Később Yang (2022b) végzett kísérleteket magyar nyelvű BART és GPT-2 (Yang, 2022a) modellek előállítására, amelyeket absztraktív összefoglalás feladatára is finomhangolta. Emellett Makrai és mtsai (2022) végzett sikeres kísérleteket huBERT alapú összefoglaló modellek tanítására.

Kutatásomban többnyelvű modellekkel és az angol nyelvű PEGASUS modellel kísérleteztem. Az mT5, az mBART, az M2M100 és a PEGASUS modellt finomhangoltam magyar nyelvű absztraktív szövegösszefoglaló generálás feladatára.

3. Korpusz és modellek bemutatása

A modellek tanításához a HI (Yang és mtsai, 2021; Yang, 2022b) korpuszt használtam fel. Az összehasonlíthatóság végett a korpuszom megegyezik a Yang (2022b) által használt HI korpusszal és annak tanító és teszt halmazával:

– HI (HVG + index.hu): 559 162 szegmens; cikk: 147 099 485 token; lead: 16 699 600 token; átlagos tokenszám a cikkekben: 263,07; átlagos tokenszám a leadekben: 29,87; tesztanyag: 3000 szegmens.

A kutatásomban négy modellel végeztem kísérletet: mT5, mBART, M2M100, PEGASUS.

Az **mT5** (Xue és mtsai, 2021) a T5 (Text-To-Text Transfer Transformer) (Raffel és mtsai, 2020) többnyelvű változata. A T5 a Google modellje, amellyel a transzfer tanulás tulajdonságát kutatták. A kutatásban több különböző nyelvtechnológiai feladatot tanítottak be egyszerre egy modellnek. Minden feladatot szövegből szöveg (sequence-to-sequence) feladatként kezeltek. Mivel egy modellnek tanították be őket, a feladatok segíthetik egymást. A T5 modell többnyelvű változata az mT5, azonban az mT5 esetében az előtanítás során nem tanították be a modellt különböző nyelvtechnológiai feladatokra, hanem BERT modellhez hasonlóan tanították elő. A modellt nagy mennyiségű adaton, a 101 nyelvből álló, általuk erre a célra létrehozott mC4 korpuszon tanították elő. A modell rendelkezik magyar tudással. A modellnek több különböző mérete van, a kutatásomban az mT5-base (580 millió paraméter) modellt használtam.

Az **mBART** (Liu és mtsai, 2020) a BART modell (Lewis és mtsai, 2020) többnyelvű változata. A BART modell előtanításához olyan feladatokat határoztak meg, amellyel később könnyebben tudják összefoglaló generálás feladatára finomhangolni a modellt. A feladatok a következőek: token maszkolás, token törlés, szövegrész maszkolás, mondatok összekeverése, dokumentum rotáció (egy véletlenszerűen kiválasztott tokennél elválasztják a szöveget, majd úgy forgatják a szöveget, hogy a kiválasztott token lesz az első tokenje a dokumentumnak). Az mBART esetében többnyelvű szövegen végezték el a leírt előtanítási módszert. Az mBART első változatához 25 nyelvet alkalmaztak a CC25 (Wenzek és mtsai, 2020; Conneau és mtsai, 2020) korpusz segítségével. Később ezt a modellt kiegészítették további 25 nyelvvel, így jött létre az mBART-50 (Tang és mtsai, 2020), amellyel egy 50 nyelvből 50 nyelvre fordító gépi fordító modellt tanítottak. A magyar nyelv nem része az 50 nyelvnek. Azonban a SentencePiece tokenizálásnak (Kudo és Richardson, 2018) köszönhetően fel lehet vele dolgozni a magyar nyelvű szöveget is. A kutatásomban az mBART-large-50 modellt (610 millió paraméter) alkalmaztam.

Az **M2M100** (Aharoni és mtsai, 2019) a Fairseq többnyelvű gépi fordító modellje. A modell tanításához 100 nyelvet használtak fel. Létrehoztak a feladatra egy angol centrikus 100 nyelvű párhuzamos korpuszt. Architektúrában enkóder-dekóder transzformert használtak, a többnyelvűség eléréséhez nyelvi kóddal látták el a szöveg elejét, mind az enkóder, mind a dekóder részben. A modell rendelkezik magyar tudással. Kutatásomban az M2M100_1.2B modellt (1,2 milliárd paraméter) alkalmaztam.

A **PEGASUS** (Zhang és mtsai, 2020) modell tanításával azt célozták meg, hogy már az előtanítás során olyan feladatokat adtak a modellnek (hasonlóan az BART modellhez), amelyekkel az összefoglaló készítésének képességét erősítik. Az előtanítás folyamatában egyszerre kettő feladatot adtak a modellnek: szómaszkolás és célmondat maszkolás. A szómaszkolás hasonló a BERT modellhez, véletlenszerűen lemaszkoltak kettő szót. Az új megközelítés a célmondat maszkolás volt. Extraktív összegzés módszerével kiválasztották azt a mondatot, amely leginkább jellemez egy bekezdést, majd ezt a mondatot lemaszkolták. Ezzel a módszerrel a modell megtanulta azt, hogy melyek a fontos részek egy szövegben. A PEGASUS finomhangolással 12 különböző összefoglaló feladaton ért el piacvezető teljesítményt. Kutatásomban a PEGASUS-large modellt (568 millió paraméter) alkalmaztam.

4. Kísérletek

Első kísérletként megvizsgáltam, hogy a különböző modellek tokenizálói hogyan tudták feldolgozni a magyar szövegeket. Az összehasonlíthatóság végett azokra a szavakra néztem meg, amelyeket Nemeskey (2020) vizsgált a cikkében. Az átláthatóság kedvéért kivettem a speciális karaktereket, mint a nyelvi kódokat az M2M100 (__hu__) és az mBART (en_XX) esetében, valamint a szegmens végét jelölő címkét (</s>), amelyek mind a négy esetben voltak. A 1. táblázatban láthatóak a többnyelvű modellek által tokenizált szavak. Ami kiemelendő, az az mBART modell tokenizálása, ami nagyban hasonlít a huBERT tokenizálójához. Ez azért meglepő, mert direkt módon nem tettek bele magyar tanítóadatot az előtanítás során, azonban a tokenizálásból arra következtettem, hogy mégis sok magyar szöveg kerülhetett bele. A három modellből az M2M100 a legtöredezettebb, ami várható volt, hiszen a szótárban 100 nyelv szavaival kell osztozkodniuk a magyar szavaknak. Akkor mutatkozik meg a többnyelvű modellek hátránya, amikor több magyar specifikus betű, mint 'é', 'ú', 'ó' vagy 'ő', szerepel a szóban, ilyenkor a tokenizáló jobban széttördeli a szavakat.

A 2. táblázat mutatja a PEGASUS tokenizálójának kimenetét. A 'PEGASUS eredeti' oszlopban található az eredeti modell tokenizálása. Látható, hogy a modell nem ismeri a magyar ékezetes magánhangzókat. Ez nem meglepő, hiszen a modellt egy válogatott korpuszon tanították elő, amiben csak angol nyelvű szöveg volt. Ezért ahhoz, hogy kezelni tudja a magyar szavakat, hozzá kellett adni a hiányzó magyar ékezetes magánhangzókat. Az eredeti szótárban egyedül a kis 'é' betű szerepelt, így a szótárhoz első körben a következő magánhangzókat adtam:

– á, í, ó, ö, ő, ú, ü, ű, Á, É, Í, Ó, Ö, Ő, Ú, Ü, Ű

Azonban a tokenizáló tesztelése során vettem észre, hogy a SentencePiece esetében (mivel ez nem szó alapú, hanem szövegalapú) ha egy token önmagában szerepelt a szótárban, az azt jelentette, hogy ez a token egy szó része volt és nem az adott szó kezdőtokenje. Vagyis, ha egy szó ékezetes magánhangzóval kezdődött, akkor hozzá kellett tenni egy speciális karatert az ékezetes magánhangzó

	huBERT	mT5	mBART	M2M100
Nemzeti	Nemzeti	Nem zeti	Nemzeti	Nemzeti
Andersen	Andersen	Andersen	Andersen	Andersen
labdarúgó	labdarúgó	labda rúg ó	labdarúgó	lab dar ú gó
zambiai	z amb iai	z ambia i	z ambia i	z amb iai
megmaradt	megmaradt	meg marad t	megmarad t	meg mar adt
hétfő	hétfő	hét fő	hétfő	hét fő
keddtől	kedd től	ked d től	kedd től	k edd től
edényben	edény ben	ed ény ben	e dé ny ben	ed ény ben
Hétfőn	Hétfőn	H ét fő n	H ét fő n	H ét fő n
tájékoztatták	tájékoztat ták	tájékoztat ták	tájékoztat ták	tájé kozt att ák
leggazdagabb	leggazdagabb	leg gazdag abb	leg gazda ga bb	leg gaz dag abb
elpártolt	el párt olt	el párt olt	el pár to lt	el p árt olt

1. táblázat. Többnyelvű modellek tokenizálása

elé, ami azt jelölte, hogy előtte egy szóköz szerepelt. Így a következő tokeneket is hozzá kellett adni a szótárhoz:

– _á,_í,_ó,_ö,_ő,_ú,_ü,_ű,_Á,_É,_Í,_Ó,_Ö,_Ő,_Ú,_Ü,_Ű

A magyar ékezetes magánhangzók hozzáadásával egy újabb problémát generáltam. Mivel az eredeti szövegben nem voltak ékezetes magánhangzók, ezért olyan szóösszetétel sem, amiben ékezetes betűk lettek volna. Így a tokenizáló leválasztotta az ékezetes magánhangzókat a körülötte lévő betűkről. Ennek az lett a következménye, hogy ahol két ékezetes magánhangzó között volt egy betű, az magára maradt, mint önálló token. Vagyis, egy szóközt illesztett be a magára maradt betű elé. A szótárban nem szerepelt egy-egy szóköz nélküli betű, ezért hozzá kellett adni őket, hogy helyesen tudja tokenizálni a magyar ékezetes betűvel rendelkező szavakat. Így a teljes magyar ábécé betűinek kicsi és nagy változatait, valamint a speciális karakterrel ellátott (space jelölő) ékezetes magánhangzókat hozzáadtam a szótárhoz.

A magyar ábécé betűinek hozzáadására több módszer is adott volt, azonban a meglévő függvényekkel való hozzáadása túlságosan széttöredezte a szöveget, gyakorlatilag karakteralapúvá alakította a tokenizálót. Ez azért történt, mert mind a token hozzáadása (add_token), mind a speciális token hozzáadása (add_special_token) függvény valószínűség nélkül a szótár elejére teszi be a hozzáadott tokeneket. Így a tokenizálás ezekkel az újonnan hozzáadott betűkkel kezdi a tokenizálást, vagyis karakter alapon kezdi a tördelést. Ezért bele kellett nyúlnom a szótárba. A fent említett tokeneket valószínűségekkel a szótár végére adtam hozzá. Az eredeti szótár utolsó token valószínűségnél egyenletesen kisebb valószínűségeket rendeltem az új tokenekehez. A transzformer finomhangoló implementációja[3] alapján a modell az új tokenekhez automatikusan véletlenszerűen rendel súlyokat, majd a modellt átméretezi az új szótárnak megfelelően. Ilyen

[3] https://github.com/huggingface/transformers/tree/main/examples/pytorch/summarization

módon az eredetileg 96.103 tokenből álló szótárból készítettem egy 96.200 méretűt. A módosítást a 'tokenizer.json' nevű fájlban tudtam elvégezni, amivel csak a PegasusTokenizerFast függyvény tud dolgozni, ezért a finomhangoló szkriptet át kellett írni, hogy a PegasusTokenizerFast függvényt használja.

A 2. táblázat 'PEGASUS kiegészített' oszlopában látható a magyar ékezetes magánhangzók hozzáadása utáni tokenizálás eredménye. Bár nagyon töredezett, de tudja kezelni a magyar szöveget. Az 'edényben' szóban látható, hogy az 'y' betűt leválasztotta a 'ny' betűről. Mint fent említettem, az 'é' betű szerepelt az eredeti szótárban is és megvizsgálva az 'én' szóelem benne van a szótárban. Hiába tettem bele a 'ny' betűt, az 'én' szóelemnek nagyobb a valószínűsége, ezért egybetartotta a tokenizáló, és leválasztotta az 'y'-t.

	PEGASUS eredeti	PEGASUS kiegészített
Nemzeti	Nem ze ti	Nem ze ti
Andersen	Andersen	Andersen
labdarúgó	lab dar <unk> g <unk>	lab dar ú g ó
zambiai	zambia i	zambia i
megmaradt	me g mara d t	me g mara d t
hétfő	hé tf <unk>	hé tf ő
keddtől	ked d t <unk> l	ke dd t ő l
edényben	ed én y ben	ed én y ben
Hétfőn	Hé tf <unk> n	Hé tf ő n
tájékoztatták	t <unk> j é ko z t att <unk> k	t á j é ko z t att á k
leggazdagabb	leg gaz dag abb	leg gaz dag abb
elpártolt	el p <unk> r tol t	e lp á r to lt

2. táblázat. PEGASUS modell tokenizálása

Kísérleteim során a kiválasztott modelleket absztraktív összefoglaló feladatra finomhangoltam a HI korpuszon. A finomhangoláshoz a Hugging Face szkriptjét[4] alkalmaztam. A mérésekhez 4 darab NVIDIA A100 (80GB) GPU-t használtam az alábbi paraméterekkel:

- mT5: batch méret: 4; bemeneti szöveghossz: 1024; kimeneti szöveghossz: 256; epoch: 40; tanulási ráta: 0,0005; prefix: 'summarize: ';
- mBART: batch méret: 4; bemeneti szöveghossz: 1024; kimeneti szöveghossz: 256; epoch: 40; tanulási ráta: 0,00002;
- M2M100: batch méret: 4; bemeneti szöveghossz: 1024; kimeneti szöveghossz: 256; epoch: 40; tanulási ráta: 0,00002; src_lang: hu; tgt_lang: hu; warmup: 15 000;
- PEGASUS: batch méret: 4; bemeneti szöveghossz: 1024; kimeneti szöveghossz: 256; epoch: 40; tanulási ráta: 0,00002; warmup: 15 000;

[4] https://github.com/huggingface/transformers/tree/main/examples/pytorch/summarization

Az összehasonlíthatóság végett, az epoch számot meghagytam ugyanannyinak (40) mint Yang (2022b) kutatásában. A modellek méretei miatt 4-es batch mérettel tudtam dolgozni. Az előnye ezeknek a nagyobb modelleknek, hogy képesek nagyobb 1024-es bemeneti szöveghosszal dolgozni. A HI cikkek átlagos tokenszáma 263, így nem szükséges longformer modellek alkalmazása.

Az mT5 finomhangolása során a modell tulajdonsága miatt, hozzá kellett adni egy prefixet. Ezt leszámítva gyakorlatilag az alapbeállításokkal lehetett finomhangolni.

Az mBART modell esetében már módosítani kellett a paramétereken. A tanulási rátát kisebbre kellett venni, hogy jobban konvergáljon a modell, azonban mást nem kellett módosítani rajta.

Az M2M100 modellel még több módosítást kellett végezni. Először a modell tulajdonságának megfelelően be kellett állítani, hogy mind a forrásnyelv, mind a célnyelv magyar legyen. Az első kísérlet után mindössze 20,13 ROUGE1 értéket sikerült elérni a HI korpuszon, ezért 0,0005 tanulási rátát lejjebb vettem 0,00002-re (több tanulási rátát is kipróbáltam, ez bizonyult a legjobbnak). Ezzel sikerült 28,91 ROUGE1 értéket elérni. Ezután a warmup funkciót is bekapcsoltam, amivel végül a legjobb eredményt sikerült elérnem.

Végül a PEGASUS modellel kellett a legtöbb módosítást végezni. Először is a szótárat kellett kiegészíteni, hogy fel tudja dolgozni a magyar nyelvű szöveget. Ezen kívül az M2M100 tapasztalataiból kiindulva kisebbre állítottam a tanulási rátát és hozzáadtam a warmup lépést is.

5. Eredmények

A modellek kiértékeléséhez a ROUGE (Lin, 2004) metrikákat használtam. A gép által generált összefoglalókat összehasonlítottam az eredeti leadekkel. A 3. táblázat mutatja a modellek eredményeit. Az eredmények a következő formátumban láthatóak: ROUGE-1/ROUGE-2/ROUGE-L. Az összehasonlíthatóság végett betettem Yang (2022b) legjobban teljesítő modelljét a BART-base-1024 modellt. Olyan szempontból is releváns ez a modell, hogy ez is 1024 bemeneti hosszal dolgozott.

Az eredményekben megfigyelhető, hogy az mT5 és az mBART magasan felülmúlta a magyar modellt. Viszont az M2M100 és a PEGASUS modelleknek nem sikerült. Ez igazából várt eredmény, hiszen az M2M100 gépi fordításra előtanított modell, míg a PEGASUS finomhangolás nélkül nem tud egyáltalán magyarul. A meglepő eredmény mégis az, hogy versenyképes teljesítményt nyújtottak. A PEGASUS esetében a legváratlanabb az eredmény, hiszen egy modell ami nem tudott egyáltalán magyarul a finomhangolás során megtanult magyarul is és közben nem felejtette el a szövegösszefoglaló generálás tudását sem. Valamint az mBART modell eredménye is meglepő, hiszen közvetlenül nem tettek bele magyar tudást, azonban a tokenizálásból azt lehetett leszűrni, hogy látott már magyar szöveget. Ez segíthette a modellt a magyar nyelvre való finomhangolásra.

A 4. táblázatban mutatok egy-egy példát a modellek összefoglalóiról. Egy olyan példát választottam, ahol a legtöbb modell hallucinált vagy rosszul kom-

	HI
BART-base-1024	31,86/14,59/23,79
mT5	33,30/15,97/24,65
mBART	**35,17/16,46/25,61**
M2M100	30,84/13,21/22,54
PEGASUS	30,36/13,11/21,57

3. táblázat. Absztraktív összefoglaló modellek eredményei

binált információkat, hogy lássuk a modellek gyengeségeit. A BART-base-1024 modell: 'Magyar Fotográfiai Központ', '160 fotóművész'; mT5: 'Magyar Nemzeti Galéria', 'Krisztina körúti épület'; mBART: 'Capa Központban nyílik a kiállítás', M2M100: 'Városi Múzeum'. A hallucinációk mellett a BART-base-1024 modell esetében a fogalmazás területén is vannak hibák. A példákból az látszik, hogy valóban a BART-base-1024 összefoglalója a legrosszabb. Az mT5 összefoglalója a legrészletesebb és leghosszabb, azonban a hallucinációi miatt fenntartásokkal kell kezelni, de a hosszúsága miatt nagyobb fedést tud produkálni, ez magyarázhatja a magas F1 mértékét is. Az mBART és az M2M100 a legszűkszavúbbak. A PEGASUS eredménye elég jó, de ha közelebbről megnézzük, akkor inkább extraktív módon készítette az összefoglalót. Az első mondat a cikk egyik hosszabb mondatának részmondata, amit levágott és egy ponttal lezárta, talán ez a lezárás az egyetlen absztraktív művelet benne. A második mondat egy az egyben szerepel az eredeti cikkben. Az eredmény nem meglepő, hiszen az előtanított modell részben extraktív feladatra volt tanítva. A beam érték feljebb állításával ez módosulna, de akkor az összehasonlíthatóság elveszne. Viszont az extraktív tulajdonsága miatt nincsen hallucináció a PEGASUS kimenetében. Ilyen szempontból ha a számok területén a leggyengébb is, mégis a legjobbnak mondható, mivel nincsen benne félrevezető információ. A jövőben szeretném továbbfejleszteni a kutatást abba az irányba, hogy egy angol nyelvű és egy magyar nyelvű szótár egyesítésével létrehozzak egy új, kétnyelvű szótárat.

6. Összegzés

A kutatásom során különböző többnyelvű és egy angol nyelvű neurális nyelvmodellt finomhangoltam magyar nyelvű absztraktív összefoglaló generálás feladatára. A kutatás érdekessége, hogy olyan modelleket is kipróbáltam, mint az mBART, ami ugyan összefoglaló generálásra való, azonban nincsen benne közvetlenül magyar nyelvi tudás. Illetve az M2M100 modellt, amelyben ugyan van magyar tudás, de gépi fordításra tanították elő. A kísérleteimben megmutattam, hogy mindegyik modellt lehet finomhangolni magyar összefoglaló generálás feladatára. Az eredményeimben nem várt módon, az mBART érte el a legjobb eredményt a HI korpuszon. Ez azt a kérdést veti fel, hogy mennyi magyar tudás szükséges az előtanításban, hogy utána magyar nyelvű feladatokra tudjuk adaptálni. A tokenizálás kísérletben láthattuk, hogy bár közvetlen módon nem tettek bele az mBART modellbe magyar anyagot, azonban mégis feltételezhetjük, hogy

	Példa
cikk	A több mint 160 fotót felvonultató tárlat azt szeretné megmutatni , ki volt Robert Capa - mondta el a hétfői sajtóbemutatón Fisli Éva , a tárlat kurátora . A látogatók így általa és róla készített felvételek , valamint korhű installációk segítségével a magyar származású Capát egyaránt megismerhetik mint emigránst , fotóriportert és haditudósítót , mint játékost és hőst . A 2014. január 12-ig látogatható tárlathoz Doku-tér is tartozik , ahol a látogatók elmélyülhetnek a Capa-irodalomban is . A fotográfus születésének 100. évfordulójához kapcsolódóan a 3-as metró kocsijainak hirdetőfelületein is látható jelenleg egy válogatás Capa-fotográfiákból , október 21-én pedig megnyílik az Ernst Múzeum épületében a Capa Központ .
lead	A világhírű magyar származású fotográfus " szerepeit " kutatja a Magyar Nemzeti Múzeum (MNM) Robert Capa - A játékos című , kedden nyíló kiállítása . Capa képei láthatók a 3-as metrón is , októberben pedig nyílik az Ernst Múzeum épületében a Capa Központ .
BART-base-1024	A Magyar Fotográfiai Központban több mint 160 fotóművész várja újra Robert Capa fotóit , köztük a korhű magyar származású fotográfiát .
mT5	A világhírű magyar származású fotográfus , Robert Capa születésének 100. évfordulója alkalmából nyílt kiállítás a Magyar Nemzeti Galéria budapesti , Krisztina körúti épületében .
mBART	Robert Capa születésének 100. évfordulója alkalmából fotókiállítás nyílik hétfőn Budapesten , a Capa Központban .
M2M100	Robert Capa világhírű fotográfusról nyílik kiállítás kedden Budapesten , a Városi Múzeumban .
PEGASUS	Több mint 160 fotót felvonultató tárlat azt szeretné megmutatni, ki volt Robert Capa. A látogatók így általa és róla készített felvételek, valamint korhű installációk segítségével a magyar származású Capát egyaránt megismerhetik mint emigránst, fotóriportert és haditudósítót, mint játékost és hőst.

4. táblázat. Különböző modellek összefoglalói

került bele magyar szöveg. Végül, de nem utolsó sorban, a PEGASUS angol nyelvű modellt adaptáltam magyar nyelvre. Megmutattam a kísérletemben, hogy a szótár bővítésével a modell megtanítható, finomhangolható magyar nyelvre.

Az mT5 és az mBART modellekkel felülmúltam és 'state-of-the-art' eredményeket tudtam elérni a magyar nyelvű absztraktív összefoglaló területén.

Hivatkozások

Aharoni, R., Johnson, M., Firat, O.: Massively multilingual neural machine translation. 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). Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019)

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 8440–8451. Association for Computational Linguistics, Online (Jul 2020)

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)

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)

Kudo, T., Richardson, J.: SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 66–71. Association for Computational Linguistics, Brussels, Belgium (Nov 2018)

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)

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)

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)

Makrai, M., Tündik, Á.M., Indig, B., Szaszák, G.: Towards abstractive summarization in hungarian. In: XVIII. Magyar Számítógépes Nyelvészeti Konferencia. pp. 505–219. Szegedi Tudományegyetem, Informatikai Intézet, Szeged, Magyarország (2022)

Nemeskey, D.M.: Egy emBERT próbáló feladat. In: XVI. Magyar Számítógépes Nyelvészeti Konferencia. pp. 409–418. Szegedi Tudományegyetem, Szeged (2020)

Nemeskey, D.M.: Introducing huBERT. In: XVII. Magyar Számítógépes Nyelvészeti Konferencia. pp. 3–14. Szegedi Tudományegyetem, Informatikai Intézet, Szeged, Magyarország (2021)

Pfeiffer, J., Goyal, N., Lin, X., Li, X., Cross, J., Riedel, S., Artetxe, M.: Lifting the curse of multilinguality by pre-training modular transformers. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 3479–3495. Association for Computational Linguistics, Seattle, United States (Jul 2022)

Pfeiffer, J., Vulić, I., Gurevych, I., Ruder, S.: MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 7654–7673. Association for Computational Linguistics, Online (Nov 2020)

Pfeiffer, J., Vulić, I., Gurevych, I., Ruder, S.: UNKs everywhere: Adapting multilingual language models to new scripts. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 10186–10203. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic (Nov 2021)

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

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners (2019)

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)

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (szerk.) Advances in Neural Information Processing Systems 30, pp. 5998–6008. Curran Associates, Inc. (2017)

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)
Yang, Z.G.: "Az invazív medvék nem tolerálják a suzukis agressziót" - Magyar GPT-2 kísérleti modell. In: XVIII. Magyar Számítógépes Nyelvészeti Konferencia. pp. 463–476. Szegedi Tudományegyetem, Informatikai Intézet, Szeged, Magyarország (2022a)
Yang, Z.G.: BARTerezzünk! Messze, messze, messze a világtól, BART kísérleti modellek magyar nyelvre. In: XVIII. Magyar Számítógépes Nyelvészeti Konferencia. pp. 15–28. Szegedi Tudományegyetem, Informatikai Intézet, Szeged, Magyarország (2022b)
Zhang, J., Zhao, Y., Saleh, M., Liu, P.: Pegasus: Pre-training with extracted gap-sentences forabstractive summarization. In: Thirty-seventh International Conference on Machine Learning (2020)