Automatikus összefoglaló generálás magyar nyelvre BERT modellel

Yang Zijian Győző[1,2,3], Perlaki Attila[1], Laki László János[2,3]

[1]Eszterházy Károly Egyetem, Informatikai Kar
3300 Eger, Leányka út 4.
{yang.zijian.gyozo, perlaki.attila,}@uni-eszterhazy.hu
[2]MTA-PPKE Magyar Nyelvtechnológiai Kutatócsoport
[3]Pázmány Péter Katolikus Egyetem, Információs Technológiai és Bionikai Kar
1083 Budapest, Práter u. 50/a.
{yang.zijian.gyozo, laki.laszlo}@itk.ppke.hu

Kivonat Cikkünkben különböző automatikus magyar nyelvű összefoglalást generáló neurális modelleket mutatunk be. Kétféle összefoglaló módszert különböztetünk meg. Az első módszer az absztraktív, amely a meglévő szövegből kinyeri a hasznos információt, majd erre támaszkodva próbál értelmes összefoglaló szöveget generálni. A másik módszer az extraktív, melybnek lényege, hogy a meglévő szövegből azokat a mondatokat vagy kifejezéseket nyeri ki, amelyek leginkább leírják a szöveg tartalmi lényegét. A rendszer a kinyert szövegekrészeket használja fel összefoglalóként. Kutatásunkban a „state-of-the-art" nyelvi reprezentációs modellnek számító BERT modellt használtuk. A rendszer tanításához különböző neurális modelleket alkalmaztunk. Extraktív összefoglaláshoz kipróbáltunk egy lineáris osztályozó, egy RNN és egy Transformer modellt. Az absztraktív modell tanításához Transformer modellt használtunk.
Kulcsszavak: extraktív összefoglaló, absztraktív összefoglaló, BERT

1. Bevezetés

A nagy mennyiségű írott szövegek rendszerezéséhez és átláthatóságához elengedhetetlen azok kivonatolása. Erre napjainkban automatikus rendszerek léteznek, melyek feladata a hosszabb szövegek, szövegrészek összefoglalása – text summarization – oly módon, hogy önálló folyékony szövegként leírja az egész dokumentum lényegét.

Kétféle automatikus összefoglaló megközelítést különböztetünk meg: absztraktív és extraktív. Absztraktív szöveg-összefoglalásnak hívjuk azt, amikor egy szöveg alapján olyan szöveget generálunk, amely kivonata az eredeti szövegnek. Tartalmazza a lényeget, tömörebben, rövidebben fogalmazza meg azt. Ez a módszer hasonlít leginkább az emberi összefoglaláshoz. A módszer legnagyobb nehézsége, hogy olyan szöveget kell generálni, ami nemcsak nyelvileg helyes, hanem tartalmaznia kell az eredeti szöveg mondanivalóját is. Ez két igen nehéz feladat, amelyekből külön-külön is több kutatás folyik. A másik megközelítés az

extraktív összefoglaló generálás, amely az eredeti szövegből nyeri ki a lényegre vonatkozó szövegrészleteket. Ez a feladat annyiban könnyebb, hogy nem kell nyelvileg helyes mondatot generálni, elég csak a meglévő szövegből megkeresni a lényegre vonatkozó részeket.

Kutatásunk célja, hogy megvizsgálja a jelenlegi legjobb eredményt elért összefoglaló módszert magyar nyelvre. Továbbá szeretnénk egy működő magyar nyelvű absztraktív és extraktív összefoglaló rendszert létrehozni.

2. Kapcsolódó irodalom

Az extraktív módszer a legfontosabbnak ítélt mondatok kiemelésével (és szükség szerinti egyesítésével) dolgozik, a neurális modell szempontjából ez osztályozási problémaként jelenik meg: mely mondatok választhatók ki arra, hogy az összefoglalóban is szerepeljenek. Az egyik legkorábbi neurális hálózaton alapuló extraktív rendszer a SummaRuNNer (Nallapati és mtsai, 2017), amely egy RNN enkóder segítségével oldja meg a problémát. A Refresh (Narayan és mtsai, 2018) Rouge metrikán alapul, melynek segítségével megerősítéses tanulásos módszerrel rangsorolják a mondatokat a szövegben. A Latent (Zhang és mtsai, 2018) célja a kulcsszavak legpontosabb követése helyett az emberi munkával készült absztraktokhoz való minél közelebbi hasonlóság elérése volt. A Sumo (Liu és mtsai, 2019) olyan módszert alkalmaz, amely a dokumentumból kinyerhető több-gyökerű függőségi fa-struktúrákra épül, és az összefoglaló lehetséges formájának előbecslésén alapszik. A NeuSum (Zhou és mtsai, 2018) a mondatok pontozásával és szelektálásával közelíti meg a problémát.

Az absztraktív módszer neurális megközelítésben olyan problémaként mutatkozik meg, ahol egy adott szekvenciát egy másik szekvenciába kell transzformálni. Az enkóder a forrás-dokumentumból tokeneket azonosít be, azokat feltérképezi, majd a dekóder tokenről tokenre állít elő ebből egy új szöveget. A PTgen (See és mtsai, 2017) egy mutatókat (pointer) generáló eszköz, amely a forrásszövegben szavakat azonosít be, ezután egy közvetítő (coverage) mechanizmus az összefoglalóba kerülő szavakat tartja meg. A Deep Communicating Agent (Celikyilmaz és mtsai, 2018) olyan ágens-alapú megközelítés, ahol az ágensek együtt reprezentálják a feldolgozandó dokumentumot és ennek dekódolásákoz kapcsolódik egy hierarchia-figyelő ágens. A Deep Reinforced Modell (Paulus és mtsai, 2018) közvetítés-alapú (coverage), ahol a dekóder a már generálásra került szöveget is figyeli. A BottomUp (Gehrmann és mtsai, 2018) tartalomszűrő eljárása előbb meghatározza, mely szövegrészek kerülhetnek az összefoglalóba, majd a dekóder már csak ezeken dolgozik.

Magyar nyelven az OpinHu rendszer (Miháltz, 2010) rendelkezik összefoglaló funkcióval. A rendszer kulcsszavakat és szövegkontextust használ az információkinyerésre.

3. Az összefoglaló rendszer

Ebben a fejezetben mutatjuk be az összefoglaló rendszer részeit és a mögötte lévő korpuszt. Továbbá megismertetjük a BERT modell architektúráját, valamint az ezen alapuló absztraktív és extraktív modelleket.

3.1. A BERT modell

Yang Liu és Mirella Lapata szöveg-összefoglalással kapcsolatos munkája (Liu és Lapata, 2019) az előtanított nyelvi modellek (ELMo, GPT, BERT) közül a BERT modellt (lásd 1. ábra bal oldala) emeli ki. Ez a modell rendelkezik szó-, mondat- és pozícióreprezentációval is, amely nagyméretű szövegkorpuszon alapszik. A legtöbb esetben az előtanított modellek olyan természetes nyelvi feldolgozási problémák esetén alkalmazhatók, ahol mondat- illetve bekezdés-szintű értelmezés, osztályozás szükséges. Cikkükben bemutatják, hogy a szöveg-összefoglalás feladata túlmutat az egyszerű szó- vagy mondatfordításon.

A BERT (Devlin és mtsai, 2019) modell egy előre tanított nyelvi reprezentáció, a „Bidirectional Encoder Representations from Transformers" rövidítése, a Google terméke. A BERT modell tanítása két lépésből áll: „pre-training" és „fine-tuning". A „pre-training" fázisban egy általános nyelvi reprezentációt tanítanak, majd ezen modell kimeneti paramétereinek segítségével a „fine-tuning" fázisban egy feladatspecifikus modellt tanítanak be. A BERT modell tanítása úgy történik, hogy a szövegből először WordPiece (Wu és mtsai, 2016) tokenizálóval egy általános nyelvfüggetlen szótárat hoznak létre, majd a tokenizált szöveg véletlenszerűen kiválasztott 15%-át elmaszkolják, végül a modell ezeket az elmaszkolt szövegrészeket próbálja kitalálni. Ezután végeznek egy becslést a következő mondatra, melyből 50% valódi és 50% véletlenszerű mondat. A tanításhoz kétirányú Transformer modellt (Vaswani és mtsai, 2017) használnak.

A Google betanított két többnyelvű modellt is[1]: kisbetűsített és nem kisbetűsített. A modellek tanításához kiválasztották az első 104 nyelvet, amely a legnagyobb Wikipédiával rendelkezik. A egyes nyelvek Wikipédia mérete igen különbözik, az adat közel 20%-át teszi ki az angol Wikipédia, ezért normalizálással kontrollálták a mintavételezést, hogy kiküszöböljék ezt a problémát. Ezután minden nyelvet, hasonlóan az angolhoz, tokenizálásnak vetették alá, amelynek négy lépése volt: kisbetűsítés, ékezetek eltávolítása, írásjelek leválasztása, whitespacek kezelése. A nem kisbetűsített modell tanítása is ezeken a lépéseken esett át, a WordPiece szótár segítségével kezelik a nem kisbetűs és ékezetes szavakat.

Természetesen a magyar nyelv is része ennek a modellnek. Kutatásunkhoz a nem kisbetűsített többnyelvű modellt (BERT-Base, Multilingual Cased) használtuk.

[1] https://github.com/google-research/bert/blob/master/multilingual.md

3.2. A korpusz

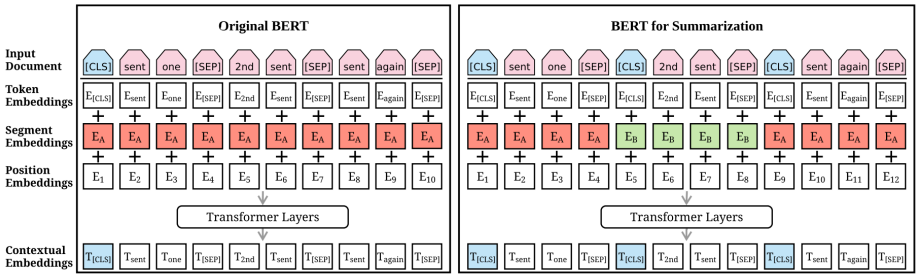
Tanító- és tesztkorpuszként a hvg.hu által rendelkezésünkre bocsátott nyomtatott és online hírlapból vett cikkeket, valamint a hozzájuk tartozó leadeket használtuk fel. A korpusz tulajdonságai:

- Nyomtatott cikkadatbázis (hetilap): 1994-2017
  - 35.513 cikk; 34.409.106 token; 2.045.255 type
- Online cikkadatbázis (napilap): 2012-2017
  - 374.064 cikk; 87.366.132 token; 3.544.622 type
- Összesen: 346.873 cikk; 121.772.523 token; 4.365.813 type;
- Cikkek témái: gazdaság, politika, tudomány, sport, kultúra, pszichológia, blog
- Kísérlethez:
  - Tanítóanyag: 343.000 cikk
  - Tesztanyag: 1790 cikk (eredtileg 1873 cikk volt, csak a rendszer kivette azokat a cikkeket, amelyek háromnál kevesebb mondattal rendelkeztek)
  - Validálás: 2000 cikk
  - Forrásszöveg (cikkek) átlagos bekezdéshossza: 317,37 szó; 15,36 mondat
  - Célszöveg (lead) átlagos bekezdéshossza: 26,21 szó; 1,56 mondat

Mivel nem mindegyik cikkhez tartozott lead, ezért miután a nyomtatott és az online cikkeket összetettük, kivettük azokat a cikkeket, amelyekhez nem tartozott lead (ezért látható eltérés a tulajdonságban az Összesen résznél).

3.3. Az extraktív és az absztraktív modellek

A BERT modell hangolható („fine-tune”) más feladatokra is, mint például összefoglaló generálásra. Az összefoglaló generáláshoz az összefoglalókból (a mi esetünkben leadekből) képzett szegmensvektorok (mondatvektorok) bemenetként szolgálnak az egyes Transformer rétegek számára. Kétféle összefoglaló modellt tudunk behangolni így: extraktív és absztraktív modellek.
Az extraktív modell: A BERT modell kimenetére rákötnek egy plusz réteget, ami specifikus feladatra alkalmas. A mi esetünkben egy olyan réteget kötnek rá, mely segítségével a cikkben lévő minden egyes mondatra mond egy valószínűséget, hogy az milyen mértékben írja le a cikk tartalmi lényegét. Az, hogy egy mondat mennyire foglalja össze a cikket, a leadeket használja fel. A plusz réteg lehet egyrészt egy sima lineáris osztályozó szigmoid függvénnyel, másrészt egy LSTM architektúrájú RNN, harmadrészt egy Transformer modell. A rendszer a betanított modellel kiválasztja és rangsorolja a cikkből azt a 3 mondatot, ami leginkább leírja annak tartalmi lényeget.
Az absztraktív modell: Az absztraktív modell megfeleltethető egy enkóder-dekóder alapú gépifordító rendszernek, ahol a forrásnyelv maga a dokumentum, a célnyelv pedig annak összefoglalója. A tanításhoz ebben az esetben a forrásnyelvi oldalon a BERT modellt használjuk, míg a célnyelvi oldalon a tanító anyagunk leadjeit használjuk.



1. ábra: A BERT és az összefoglaló BERT architektúrája (Liu és Lapata, 2019)

4. Kísérletek

Első lépésünk az volt, hogy előfeldolgozást végeztünk az eredeti szövegeken, mely az alábbi lépésekből állt. A cikkeket először mondatokra bontottuk, majd tokenizáltuk. Ezekhez az e-magyar tokenizálóját, a quntoken (Mittelholcz, 2017) eszközt használtuk. Ezt követően a szöveget az összefoglaló rendszer számára JSON formátumra alakítottuk. A rendszer ezután két speciális elemet illeszt be, az egyik a szöveg elejét jelzi, a másik a mondathatárokat. Ezután az előfeldolgozott fájlokkal különböző neurális modelleket tanítottunk be.

Kutatásunkban először megmértük az alapmódszer (baseline) teljesítményét, amely a cikk első három mondatát veszi összefoglalóként.

Következő lépésként betanítottunk három modellt az extraktív összefoglalóhoz:

- Lineáris osztályozó (BERT-Class), ahol a BERT modell kimenetére egy szigmoid függvénnyel ellátott lineáris réteg van kötve.
- Rekurrens neurális modell (BERT-RNN), melyben a BERT modell kimenetére egy bidirekciós LSTM réteg van kötve.
- Transformem modell (BERT-TransExt), ahol a BERT modell kimenetére egy Transformer modell van kötve.

Az absztraktív összefoglalóhoz kipróbáltunk két modellt:

- Baseline Transformer modell (BERT-TransAbs): egy alapértelmezett Transformer modell.
- Baseline absztraktív Transformer modell (BERT-TransAbs-baseline): Yang Liu és Mirella Lapata kutatásában (Liu és Lapata, 2019) az absztraktív modellre behangolt („fine-tuned”) baseline modell.

A modellek tanításhoz a Yang és társa (Liu és Lapata, 2019; Liu, 2019) által implementált eszközöket[2] használtuk fel.

A beállítási paraméterek extraktív modellek esetén:

[2] https://github.com/nlpyang

– Általános paraméterek:
  • dropout: 0,1; learning rate: 2e-3; batch size: 3000; tanítási lépésszám: 100000
– Transformer modell:
  • head: 8; belső réteg: 2; feedforward méret: 2048
– RNN modell:
  • rnn méret: 768

A beállítási paraméterek absztraktív modell esetén:

– dropout: 0,1; learning rate: 0,05; batch size: 3000; tanítási lépésszám: 200000; rejtett rétegek neuronszáma (enkóder, dekóder): 512; rétegek száma (enkóder, dekóder): 6; feedforward méret (enkóder, dekóder): 2048

5. Eredmények

A kiértékeléshez a ROUGE (Lin, 2004) módszert használtuk. A ROUGE (Recall-Oriented Understudy for Gisting Evaluation) egy fedés alapú módszer, ami a gépi fordítás során használt BLEU metrikán alapszik. Maga a ROUGE több almetódust is tartalmaz, melyek közül a méréseinkhez a ROUGE-1, ROUGE-2 és a ROUGE-L módszereket használtuk. A ROUGE-1 egy unigram, míg a ROUGE-2 egy bigram fedést számoló algoritmus. A ROUGE-L a leghosszabb közös szósorozatot vizsgálja bekezdés és mondat szinten.

A 1. táblázatban láthatók a különböző modellek teljesítményei. Az alapmódszer (baseline) eredménye teljesített a leggyengébben. Extraktív modell esetén BERT-RNN modell érte el a legjobb eredményt. Itt érdemes megjegyezni, hogy angol nyelv esetében ez a Transformer modell volt. Az eredmények csak azt mutatják, hogy az gép által kiválasztott mondat mennyire hasonlít a leadre. Lehetséges problémaforrás, hogy sok esetben a leadnek nem összefoglaló, hanem figyelemfelkeltő szerepe van.

Az absztraktív modell eredményeit tekintve igen alacsony a fedés, ami önmagában csak annyit jelent, hogy az összefoglaló nem hasonlít a leadre, de a kimenetet nézve sajnos egyelőre nem tudjuk értékelni még ezeket az eredményeket, mert a rendszer túltanult és mindenre ugyanazt a mondatot adta eredményül. A továbbiakban csak az extraktív modelleket fogjuk elemezni.

A 2. táblázatban látható az extraktív modellek viselkedése egymáshoz viszonyítva. Látható, hogy az esetek közel 7%-ában pontosan ugyanabban a sorrendben ajánlották ugyanazokat a mondatokat összefoglalásnak. Továbbá az látható, hogy a 3 modell közel 30%-ban ugyanazt a 2 mondatot választotta ki, és szintén közel 30%-ban pontosan egy közös mondatot választottak. Az arányokat nézve nagyon ritka eset az, amikor nem volt közös mondat. A páros összehasonlításokat nézve az szembetűnő, hogy az RNN és az osztályozó modell sokkal hasonlóbban választottak mondatokat, mint a Transformer és az osztályozó modell.

Az egyik legalapvetőbb extraktív összefoglaló módszer az, hogy kiválasztjuk a forrásszöveg első néhány mondatát (Liu és Lapata, 2019). A 3. táblázatban láthatjuk azokat az eredményeket, amelyek azt mutatják, hogy a különböző modellek milyen arányba választották a forrásszöveg első három mondatát. A rendszer

Model	ROUGE-1	ROUGE-2	ROUGE-L
Extraktív			
baseline	54,58	27,25	45,52
BERT-Class	55,26	28,21	46,23
BERT-RNN	**55,46**	**28,29**	**46,27**
BERT-TransExt	54,76	27,71	45,97
Absztraktív			
BERT-TransAbs	27,73	2,89	23,85
BERT-TransAbs-baseline	16,04	1,36	13,72

1. táblázat. ROUGE fedés eredmények

	Egyező mondatok száma		
	3 db	2 db	1 db
BERT-RNN - BERT-Class	33,18%	46,42%	12,35%
BERT-RNN - BERT-Trans	20,95%	42,35%	23,85%
BERT-Trans - BERT-Class	20,89%	32,35%	24,08%
BERT-RNN - BERT-Trans - BERT-Class	13,02%	35,92%	32,12%
BERT-RNN - BERT-Trans - BERT-Class (sorrend is egyezik)	6,93%		

2. táblázat. A különböző extraktív modellek viselkedése egymáshoz viszonyítva

a forrásszövegből rangsorolva 3 mondatot ajánl összefoglalónak. Az eredményből azt láthatjuk, hogy a Transformer modell első ajánlásnak közel 80%-ában választ a forrásszöveg első három mondatából, az esetek felében az első mondatot választja ki annak. Másik kiemelkedő eredmény az RNN modell viselkedése, amely közel 72%-ban választja a forrásszöveg első mondatát valamelyik ajánlásnak. Az esetek közel 40%-ában választja az első mondatot első ajánlásnak.

A 4. táblázatban látható néhány példa a különböző modellek kimeneteire. Láthatunk először példát arra, amikor teljesen megegyezik mind az ajánlott mondatok, mind a sorrend (a 2. táblázat alapján az esetek 6,93%-a). Majd mutatunk példát arra, amikor az ajánlott mondatok megegyeznek, de más sorrendben ajánlanak (a 2. táblázat alapján az esetek 13,02%-a). Ezután láthatunk néhány példát arra, amikor közel hasonló eredményeket adtak a különböző modellek. A példában a BERT-Class és a BERT-RNN modellek ugyanazokat a mondatokat ajánlották, csak más sorrendben (a 2. táblázat alapján az esetek 33,18%-a). A Transformer modell harmadik ajánlása különbözik a másik kettő modellétől. Végül egy olyan példát lehet látni, ahol eléggé különböző ajánlásokat adtak a modellek, a példában egy közös mondat van csak.

6. Összegzés

Létrehoztunk egy magyar nyelvű szöveg-összefoglaló rendszert, amellyel jelenleg extraktív összefoglalást tudunk készíteni hírlap cikkekből. A rendszer tanításhoz a jelenleg „state-of-the-art” nyelvi reprezentáció modellt, a Google által kuta-

	1. ajánlás	2. ajánlás	3. ajánlás	Összesen
	1. mondata a forrásszövegnek			
BERT-Class	38,60%	18,83%	11,56%	68,99%
BERT-RNN	40,89%	18,27%	12,79%	71,96%
BERT-Trans	52,46%	7,04%	6,48%	65,98%
	2. mondata a forrásszövegnek			
BERT-Class	17,21%	28,38%	15,53%	61,12%
BERT-RNN	19,05%	27,21%	15,92%	62,18%
BERT-Trans	16,65%	41,62%	6,76%	65,03%
	3. mondata a forrásszövegnek			
BERT-Class	11,90%	14,08%	24,36%	50,34%
BERT-RNN	11,73%	15,59%	22,07%	49,39%
BERT-Trans	11,28%	17,65%	42,23%	71,17%
	Összesen			
BERT-Class	67,71%	61,28%	51,45%	
BERT-RNN	71,68%	61,06%	50,78%	
BERT-Trans	80,39%	66,31%	55,47%	

3. táblázat. A forrásszöveg első három mondatának kiválasztásának arányai

tott többnyelvű BERT modellt használtuk. Az extraktív összefoglalóhoz többféle modellt is kipróbáltunk, melyek közül az RNN érte el az legjobb eredményt. Az absztraktív összefoglaláshoz Transformer alapú neurális hálót használtunk. Sajnos az absztraktív modellünk még nem ért el értékelhető eredményt, de az extraktív modellek már működnek és eredményeinkben kielemeztük működéseit. Továbblépésként az absztraktív modellekkel szeretnénk értékelhető eredményt elérni.

Köszönetnyilvánítás

Jelen kutatás a Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal Mesterséges Intelligencia Nemzeti Kiválósági Programja támogatásával a 2018-1.2.1NKP-2018-00008 azonosítójú projekt keretében valósult meg.

A kutatást az EFOP-3.6.1-16-2016-00001 „Kutatási kapacitások és szolgáltatások komplex fejlesztése az Eszterházy Károly Egyetemen” című projekt támogatta.

Hivatkozások

Celikyilmaz, A., Bosselut, A., He, X., Choi, Y.: Deep communicating agents for abstractive summarization. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). pp. 1662–1675. Association for Computational Linguistics, New Orleans, Louisiana (Jun 2018)

Modell	Példa
BERT-Class BERT-RNN BERT-Trans	1. Pótlóbuszok járnak az M3-as metró helyett Újpest-Központ... 2. Mintegy háromnegyed órával később közölték: helyreállt a rend... 3. Kérjük az arra közlekedők türelmét – írták a BKK....
BERT-Class	1. A napokban Tajvanról érkezett egy sisakkamerás felvétel... 2. Az incidenst egy motoros társaság közös csapatása során... 3. Jó kérdés, hogy mit szólt a barátnő a férfi magatartásához...
BERT-RNN	1. Az incidenst egy motoros társaság közös csapatása során... 2. Jó kérdés, hogy mit szólt a barátnő a férfi magatartásához... 3. A napokban Tajvanról érkezett egy sisakkamerás felvétel...
BERT-Trans	1. Jó kérdés, hogy mit szólt a barátnő a férfi magatartásához... 2. Az incidenst egy motoros társaság közös csapatása során... 3. A napokban Tajvanról érkezett egy sisakkamerás felvétel...
BERT-Class	1. A finn kormányfő ugyanakkor meg van győződve arról... 2. A finn kormány 2017 januárjától próbaképpen bevezetné... 3. A kormányfő az intézkedéstől a szociális juttatások rendszerének...
BERT-RNN	1. A finn kormány 2017 januárjától próbaképpen bevezetné... 2. A kormányfő az intézkedéstől a szociális juttatások rendszerének... 3. A finn kormányfő ugyanakkor meg van győződve arról...
BERT-Trans	1. A finn kormány 2017 januárjától próbaképpen bevezetné... 2. A kormányfő az intézkedéstől a szociális juttatások rendszerének... 3. A társadalmi kísérlet a 2015-ben hivatalba lépett...
BERT-Class	1. Az óceánparti San Sebastián baszk nagyvárosban az óvárosig... 2. Megrongálódott három híd, amely az Urumea folyón vezetett át. 3. A létesítmény igazgatója több mint kétmillióra euróra tette a kárt.
BERT-RNN	1. Zarauzban és az északspanyol part más fürdőhelyein épületek... 2. Megrongálódott három híd , amely az Urumea folyón vezetett át. 3. Az óceánparti San Sebastián baszk nagyvárosban az óvárosig...
BERT-Trans	1. Asztúria autonóm körzetben a hullámok lerombolták a luarcai... 2. Az óceánparti San Sebastián baszk nagyvárosban az óvárosig... 3. Mint az elpais.com, az El País című lap internetes portálja...

4. táblázat. Néhány példa az extraktív modellek kimeneteire

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)

Gehrmann, S., Deng, Y., Rush, A.: Bottom-up abstractive summarization. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 4098–4109. Association for Computational Linguistics, Brussels, Belgium (Oct-Nov 2018)

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.: Fine-tune bert for extractive summarization. In: IJCNLP. Hong Kong, China (2019)

Liu, Y., Lapata, M.: Text summarization with pretrained encoders. In: IJCNLP. Hong Kong, China (2019)

Liu, Y., Titov, I., Lapata, M.: Single document summarization as tree induction. 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. 1745–1755. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019)

Miháltz, M.: Opinhu: online szövegek többnyelv véleményelemzése. VII. Magyar Számítógépes Nyelvészeti Konferencia pp. 14–23 (2010)

Mittelholcz, I.: emtoken: Unicode-képes tokenizáló magyar nyelvre. XIII. Magyar Számítógépes Nyelvészeti Konferencia pp. 61–69 (2017)

Nallapati, R., Zhai, F., Zhou, B.: Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. pp. 3075–3081. AAAI'17, AAAI Press (2017)

Narayan, S., Cohen, S.B., Lapata, M.: Ranking sentences for extractive summarization with reinforcement learning. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). pp. 1747–1759. Association for Computational Linguistics, New Orleans, Louisiana (Jun 2018)

Paulus, R., Xiong, C., Socher, R.: A deep reinforced model for abstractive summarization. In: Proceedings of the 6th International Conference on Learning Representations. Vancouver, Canada (2018)

See, A., Liu, P.J., Manning, C.D.: Get to the point: Summarization with pointer-generator networks. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1073–1083. Association for Computational Linguistics, Vancouver, Canada (Jul 2017)

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)

Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, L., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G.S., Hughes, M., Dean, J.: Google's neural machine translation system: Bridging the gap between human and machine translation. Technical Report abs/1609.08144 (2016)

Zhang, X., Lapata, M., Wei, F., Zhou, M.: Neural latent extractive document summarization. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 779–784. Association for Computational Linguistics, Brussels, Belgium (Oct-Nov 2018)

Zhou, Q., Yang, N., Wei, F., Huang, S., Zhou, M., Zhao, T.: Neural document summarization by jointly learning to score and select sentences. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 654–663. Association for Computational Linguistics, Melbourne, Australia (Jul 2018)