

# "Az invazív medvék nem tolerálják a szukis agressziót"

## Magyar GPT-2 kísérleti modell

Yang Zijian Győző

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

MTA-PPKE Magyar Nyelvtechnológiai Kutatócsoport  
1083 Budapest, Práter u. 50/a.  
yang.zijian.gyozo@itk.ppke.hu

**Kivonat** A GPT-2 egy kizárólag dekóderrel rendelkező autoregresszív transzformer modell, amely elsősorban szöveggeneráló feladatokra alkalmas. A kutatásomban betanítottam egy kísérleti GPT-2 modellt magyar nyelvre, majd azt tovább finomhangoltam különböző nyelvtechnológiai feladatokra. A kísérleteimben az előtanított GPT-2 modellel mondat-szintű és tokenszintű szövegosztályozás, absztraktív szövegösszefoglaló generálás, hírgenerálás és versgenerálás feladatokra finomhangoltam. Az eredmények alapján a GPT-2 kevésbé teljesít jól szövegosztályozás feladatokra. Absztraktív szövegösszegzés feladatában versenyképes eredményeket ért el, azonban a szöveggenerálás eredményei még érdekesebbek. A GPT-2 modellemmel létrehoztam egy hír- és egy versgenerátort.

**Kulcsszavak:** GPT-2, mondat szintű osztályozás, tokenszintű osztályozás, szentiment analízis, NER, NP, absztraktív összefoglaló generálás, szöveggenerálás, hírgenerálás, versgenerálás

## 1. Bevezetés

A BERT alapú modellek kizárólag enkóderrel rendelkeznek, amelyek kiválóan alkalmasak különböző osztályozási feladatokra, mint például mondat szintű vagy tokenszintű osztályozásra. Azonban dekóder hiányában kevésbé alkalmasak szöveggenerálásra, amelyhez általában egy dekóder hozzácsatlakoztatásával oldják meg a problémát. Ezzel ellentétben az autoregresszív modelleket elsősorban szöveggenerálásra tervezték, a közös tulajdonságuk, hogy dekóderrel rendelkeznek, amely a tanítás során mindig csak a szöveg bal oldalát látja, így megtanulja a következő szó/szavak predikciójának képességét. Ilyenek a Facebook OpenAI<sup>1</sup> GPT modelljei (Radford és Narasimhan, 2018), amelyek csak dekóderrel rendelkeznek.

<sup>1</sup> <https://openai.com>

A kutatásom során betanítottam egy kísérleti GPT-2 modellt, majd azt tovább finomhangoltam különböző feladatokra. A GPT-2 modellel mondat- és tokenszintű szövegosztályozás, absztraktív összefoglalás, hír- és versgenerálás feladatain tanítottam. A címben szereplő szokatlan mondatot az egyik finomhangolt GPT-2 modellem (SZUM modell) generálta. Külön érdekesség, hogy az idézőjeleket is a modell generálta.

Bár a GPT modellek egyik alapja a nagy mennyiségű adat használata, azonban erőforrások hiányában kisebb adatmennyiségen tanítottam a GPT-2 modellel, ezért is neveztem el kísérleti modellnek. Azonban kevesebb adattal is rendkívül látványos eredményeket értem el.

Modelljeim és szkriptjeim megtalálhatóak a Github<sup>2</sup> és Hugging Face<sup>3</sup> oldalakon.

## 2. GPT modellek és Kapcsolódó irodalom

A GPT (generative pre-training) (Radford és Narasimhan, 2018) egy kizárólag dekóder (decoder-only) tartalmazó transformer modell. A kizárólag dekóder architektúra abban különbözik az enkóder-dekóder architektúra dekóderétől, hogy hiányzik belőle az enkóder-dekóder figyelmi (encoder-decoder attention) réteg, mivel nincsen enkóder.

Az OpenAI<sup>4</sup> kutatócsoport projektje az úgynevezett félig felügyelt tanítás (semi-supervised learning) kategóriájába tartozik. A paradigma már eddig is komoly érdeklődésre tett szert, főleg az olyan alkalmazások terén, mint a szekvenciacímkezés (sequence labeling) vagy a szövegosztályozás (text classification). Nemrégiben számos kutatási eredmény rávilágított arra, hogy a szövegbeágyazásos (word embedding) (Collobert és mtsai, 2011) eljárások is komoly mértékben javíthatják a modellek teljesítményét. Mindezek azonban a szavak szintjén értelmezhető megközelítések, az OpenAI csapata azonban ennél magasabb szintű szemantikai vizsgáldást tűzött ki célul. Ezeknek a megoldásoknak egyre nagyobb szerep jut, ugyanis a szavaknál magasabb szintű beágyazások a nagyobb szövegegységek vektorreprezentációját teszik lehetővé. A felügyelet nélküli előtanítás (unsupervised pre-training) lingvisztikai információk megragadását teszi lehetővé, megfelelő transzformer háló kiválasztásával pedig hosszabb távú információkinyerés is megvalósítható. A tanítási segédobjektívák (auxiliary training objectives) használata a teljesítmény növekedésével járhat, mint például a Rei által bemutatott szekvenciacímkezés feladat esetén (Rei, 2017).

A nyelvi modellek képesek megtanulni számos nyelvelemzési feladat felügyelet nélküli megoldását, amennyiben kellően nagy méretű adathalmazon vannak betanítva, mint például a több millió honlap tartalmát felölelő WebText. A GPT-2 (Radford és mtsai, 2019) modell egy 1,5 milliárd paraméterrel rendelkező transzformer modell, amely nyolcból hét feladaton csúcsteljesítményt ért el az úgynevezett zero-shot üzemmódban. A GPT-2 választ próbál adni arra a

<sup>2</sup> <https://github.com/nytud/neural-models>

<sup>3</sup> <https://huggingface.co/NYTK>

<sup>4</sup> <https://openai.com>

kérdésre, hogy milyen módon hat a nagyobb adathalmazon való tanítás a nagy méretű nyelvi modellek teljesítményére. Számos hasonló munka próbálja ezt a megközelítést kísérleti úton átültetni a gyakorlatba. Például Jozefowicz és mtsai (2016) RNN-alapú (Recurrent Neural Network) nyelvi modelleket teszteltek az 1 milliárd szavas referencia feladaton (1 Billion Word Benchmark). A csoport kiemelkedő eredményei között szerepelnek, hogy az egyik modelljük 51,3-ról 30,0-ra javította a perplexitást, míg egy másik modellegyüttesük új csúcstól állított fel perplexitás terén 41,0-ról 23,7-re csökkentve azt (Jozefowicz és mtsai, 2016). Bajgar és mtsai (2017) egy Attention-Sum Reader modellt használtak, amelyet a Book Test adathalmazon tanítottak be, amely hatvanszor akkora méretű, mint a korábban megjelent, hasonló felépítésű Children’s Book Test<sup>5</sup> (CBT) adathalmaz. Kiemelendő azonban, hogy a Book Test adathalmazon betanított modell pontosabb eredményeket ért el a betanítást követően a hatvanszor kisebb CBT-n (Bajgar és mtsai, 2017). Hestness és mtsai (2017) több gépi tanulós modell teljesítményét vizsgálták a modell kapacitásának és a tanító adathalmaz méretének függvényében. A csoport eredményei összhangban állnak a GPT-2 esetében megfigyelt tendenciákkal (Hestness és mtsai, 2017). Karpathy és mtsai (2016) az RNN-alapú modellek, azon belül is a Long-Short Term Memory (LSTM) variáns kiemelkedő teljesítményének hátterét kutatták, továbbá az utóbbi modellvariánszt idézetek, zárójelek valamint sorhosszúság vizsgálatára alkalmazták (Karpathy és mtsai, 2016). Fontos kiemelni azt is, hogy a Liu és mtsai (2018) által jegyzett publikációban a Wikipédia cikkek generálására betanított kizárólag de-kóder architektúrájú modell megtanulta lefordítani a neveket nyelvek között (Liu és mtsai, 2018).

A nyelvi modellek generatív vagy egyéb feladatok megoldása során mutatózó teljesítményének javítására többféle stratégia is lehetséges. Az egyik út a paraméterszám és a számítási kapacitás együttes növelése, a másik kizárólag a paraméterszám emelésére törekszik, a harmadik stratégia pedig csak a számítási kapacitást variálja. Számos példa létezik transzformer-alapú modellek paraméterszámának és azzal együtt közel arányosan a FLOPS per token arány szisztematikus növelésére, és ezáltal jobb teljesítmény elérésére. Kísérleteinkben azt vizsgálták, hogy a 300 milliós (Devlin és mtsai, 2019), 8,3 milliárdos (Shoeybi és mtsai, 2020) és 11 milliárdos (Raffel és mtsai, 2020) paraméterszámú modellek milyen hatást gyakorolnak a teljesítményre nézve. Számos munka törekedett a teljesítmény javítására kizárólag a paraméterszám növelésével, azaz a modell információtároló kapacitását megnövelték a számítási kapacitás emelése nélkül. Erre példa a TED előadások szöveges átiratait tartalmazó korpuszon betanított több nyelvről több nyelvre fordító modell, amely 50 illetve 100 milliárd paraméterrel üzemel (Aharoni és mtsai, 2019). A harmadik stratégiát képviselő, kizárólag számítási kapacitás növelésével operáló modellfejlesztésre példa az adaptív számítási idő (adaptive computation time, ACT) algoritmus alkalmazása RNN-alapú modelleknél oly módon, hogy optimalizálja a számítási lépések számát a bemenet beérkezése és a kimenet kiküldése között (Graves, 2017). Hasonló módon számítási kapacitást növelnek a Transzformerek és RNN-alapú modellek

<sup>5</sup> <https://paperswithcode.com/dataset/cbt>

előnyös tulajdonságait ötvöző Univerzális Transzformerek (Dehghani és mtsai, 2019).

A GPT-3 (Brown és mtsai, 2020) az első stratégiát követi, vagyis a paraméterszám és a számítási kapacitás együttes növelésével éri el a kívánt teljesítményt. 175 milliárd paraméterével, finomhangolás nélkül képes state-of-the-art eredmények elérésére számos feladat tekintetében. Nyelvi modellek hatékony önfelügyelt tanításához kiemelten fontos, hogy megfelelően nagy méretű, annotálatlan korpuszok álljanak rendelkezésre. Erre a célra számos lehetőség mutatkozik, például korábbi munkák különböző nyelvi modellek betanítására használták a Wikipédiát, a Gigaword-öt (Graff és mtsai, 2003), nem publikus Google News korpuszt, RealNews adatbázist (Zellers és mtsai, 2019), vagy a WikiText korpuszt (Merity és mtsai, 2017). A nagy méretű adathalmazok létrehozása egyre inkább a webes adatkivonatolás (internet scraping) technikával történik. Az egyik legtöbbit használt ilyen módon előállított szöveges adathalmaz a Common Crawl<sup>6</sup>. A Common Crawl-t használó projektek sokszor rávilágítanak az ezzel kapcsolatos olyan jellegű felmerülő problémákra, mint az adattisztítás vagy az adatok szűrése. Egyre inkább bevett gyakorlat, hogy több adatforrást kombinálnak össze nyelvi modellek betanításához.

A GPT modellek általános jellemzője, hogy szótárként BPE kódolást használnak. A jelenlegi 3+1 féle GPT modell összehasonlítva:

- GPT: 12 réteg, 12 figyelmi fej; 768 szóbeágyazás méret; 512 szöveghossz; 117 millió paraméter;
- GPT-2: 48 réteg, 12 figyelmi fej; 1600 szóbeágyazás méret; 1024 szöveghossz; 1,5 milliárd paraméter;
- GPT-3: 96 réteg, 96 figyelmi fej; 12888 szóbeágyazási méret; 2048 szöveghossz; 175 milliárd paraméter;
- GPT Neo (Black és mtsai, 2021): mesh-tensorflow könyvtár implementáció, hogy GPT-3 típusú modelleket tudjunk tanítani.

Kutatásom során a GPT-2 modellel kísérleteztem.

### 3. Kísérletek

A kutatásom során egy kísérleti GPT-2 modellt tanítottam elő, majd azt tovább finomhangoltam négy különböző feladatra.

#### 3.1. Előtanítás

A GPT előtanítását egy TDS cikk<sup>7</sup> alapján végeztem el, a cikkben megadott beállításokkal. Egy GPT-2 kísérleti modell került betanításra, amelyhez bekezdésekre bontott magyar Wikipédiát (Nemeskey, 2020) (méretei az 1. táblázatban láthatóak) használtam. A tanításhoz használt hiperparaméterek: blokk méret:

<sup>6</sup> <https://commoncrawl.org>

<sup>7</sup> <https://towardsdatascience.com/train-gpt-2-in-your-own-language-fc6ad4d60171>

100; batch méret: 12; buffer méret: 1000; tanulási ráta: 3e-5; epoch: 10; szótár méret: 33.000. A tanításhoz 4 darab GeForce GTX 1080Ti GPU kártyát használtam. A betanítás időtartama körülbelül 3 nap volt.

### 3.2. Finomhangolás

Az előtanított GPT-2 modellem teszteléséhez az alábbi finomhangolási feladatokat végeztem el:

- Mondatszintű szövegosztályozás: szentiment analízis 2 osztályos (2-CLASS) és 5 osztályos (5-CLASS)
- Tokenszintű szövegosztályozás: névelemfelismerés (NER) és főnévi csoport felismerés (NP)
- Absztraktív összefoglaló generálás (SZUM)
- Szöveggenerálás: hírgenerálás (HÍR) és versgenerálás (VERS)

A modellek tanításához az alábbi korpuszokat használtam fel, a korpuszok méretei az 1. táblázatban láthatóak:

- **SZENT**: A mondatszintű szentiment analízis feladatra a Precognox Kft.<sup>8</sup> által készített Hungarian Twitter Sentiment Corpus<sup>9</sup>-t használtam fel, amelyet két részkorpuszra bontottam fel:
  - 2-CLASS: bináris osztályozású részkorpusz. Az értékeket a következőképpen konvertáltam: 1-es és 2-es értékeket 0-ás értékre, mint negatív értékek, 4-es és 5-ös értékeket 1-es értékre, mint pozitív értékek. A 3-as értéket figyelmen kívül hagytam, hogy elkerüljem a félreérthető eseteket.
  - 5-CLASS: Az eredeti 5 osztályos korpusz. 1: nagyon negatív, 2: negatív, 3: semleges, 4: pozitív, 5: nagyon pozitív.
- **NER**: A névelemfelismerés feladathoz a NYTK-NerKor (NerKor) (Simon és Vadász, 2021) és a Szeged NER (SzNer) korpuszt (Szarvas és mtsai, 2006) használtam.
- **NP**: A főnévi névcsoport felismeréshez a Szeged Treebank korpuszt (Csendes és mtsai, 2005) használtam.
- **SZUM**: Az absztraktív összefoglaló feladathoz a H+I korpuszt használtam, amelyet Yang és mtsai (2021) korábbi kutatásaik során is alkalmaztak (Yang és mtsai, 2021). A választásom azért erre esett, hogy össze tudjam hasonlítani az eredményemet a kutatásukkal.
- **HÍR**: A hírgenerálás feladatához a HVG (online és nyomtatott) és NOL (Népszabadság Online - NOL.hu) korpusz (H+N) cikkeit használtam fel. A választás azért esett ezekre a korpuszokra, mert bekezdésekre vannak bontva, amelyeket fel tudtam használni a tanításhoz.
- **VERS**: A versgenerálás feladatához a Petőfi Sándor összes költeményei című művet használtam, amelyet a Magyar Elektronikus Könyvtár oldaláról<sup>10</sup> töltöttem le.

<sup>8</sup> <https://www.precognox.hu>

<sup>9</sup> <http://opendata.hu/dataset/hungarian-twitter-sentiment-corpus>

<sup>10</sup> <https://mek.oszk.hu/01000/01006/>

	Tanító	Validáció	Teszt
Magyar Wikipédia (bekezdés #)	2.020.956	-	3.000
2-CLASS (szegmens #)	2.468	-	269
5-CLASS (szegmens #)	3.600	-	400
NerKor (mondat #)	1.633.952	211.392	210.884
SzNER (mondat #)	8.484	514	932
NP (mondat #)	65.679	8.209	8.209
SZUM (cikk+lead pár #)	559.162	3.000	3.000
HÍR (cikk #)	589.914	-	65.545
VERS (vers #)	735	-	80

1. táblázat. Felhasznált korpuszok méretei.

**A mondatszintű szentiment analízis** esetében, a Huggingface transformers által rendelkezésre bocsátott „text classification” könyvtárát<sup>11</sup> használtam, a következő módosított paraméterekkel: tanulási ráta: 2e-5; batch méret: 32; maximális szekvenciahossz: 128; epoch: 10. Az eredmény táblázatba (lásd 2. táblázat) a 10 epochból a legjobbat írtam ki. A kiértékeléshez a pontosság (accuracy) metrikát alkalmaztam.

**A NER és NP** feladatokra történő finomhangolásnál a Huggingface transformers „token classification” könyvtárát<sup>12</sup> használtam, a következő módosított paraméterekkel: tanulási ráta: 5e-5; batch méret: 4; maximális szekvenciahossz: 128; epoch: 10. Az eredmény táblázatba (lásd 2) a 10 epochból a legjobbat írtam ki. A kiértékeléshez az F-mérték metrikát alkalmaztam.

**Az absztraktív összefoglalás** finomhangolásához a GPT-2 cikkben (Radford és mtsai, 2019) alkalmazott módszert használtam. A hírek szövegeit és a hozzájuk tartozó leadeket az alábbi formátumba alakítottam át:

– 1 sor: [hír szövege] TL;DR: [lead szövege]

Majd ezt a korpuszt használtam fel egy nyelvi modell finomhangolásához az előtanított GPT-2 modell segítségével. A nyelvi modell finomhangolásához a Huggingface transformers „language modeling” könyvtárát<sup>13</sup> használtam, a következő módosított paraméterekkel: tanulási ráta: 5e-5; batch méret: 4; blokk méret: 512; epoch: 10.

**A hírgenerálás** feladatához H+N korpusz híreinek bekezdéseit használtam, egy sor egy hír, a bekezdések <n> szeparátorral vannak elválasztva. A zajok kiszűréséhez kivettem a tíz és annál több bekezdésből álló cikkeket, valamint a kettőnél kevesebb bekezdésből álló cikkeket. Továbbá végeztem tisztításokat: kivettem azokat a cikkeket, amelyek hivatkozásokat és szkripteket tartalmaztak.

<sup>11</sup> <https://github.com/huggingface/transformers/tree/master/examples/pytorch/text-classification>

<sup>12</sup> <https://github.com/huggingface/transformers/tree/master/examples/pytorch/token-classification>

<sup>13</sup> <https://github.com/huggingface/transformers/tree/master/examples/pytorch/language-modeling>

A tanításhoz ugyanazt a nyelvi modell finomhangoló kódot használtam mint, amelyet az absztraktív összefoglaló feladatnál használtam. A finomhangolást a következő módosított paraméterekkel végeztem el: tanulási ráta: 5e-5; batch méret: 5; blokk méret: 512; epoch: 10.

A **versgenerálás** feladatához ugyanazt a könyvtárat használtam, mint a hírgenerálás feladatában, a következő módosított paraméterekkel: tanulási ráta: 5e-5; batch méret: 5; blokk méret: 512; epoch: 10.

### 3.3. Eredmények

A 2-es táblázatban láthatóak az osztályozási (mondat- és tokenszintű) feladatok eredményei. Az összefoglalt eredmények alapján elmondható, hogy a GPT-2 modellek nem tudják felülmúlni a jelenlegi state-of-the-art huBERT modellt. Ez az eredmény várható volt, mivel az általam használt modellek számos jellemzőjük tekintetében gyengébbek (kevesebb tanítóadat, kisebb batch méret stb.). Ennek ellenére, mégis magas eredményeket érnek el. A szentiment analízis kísérletekben például csak 5-7 százalékkal maradnak alul a huBERT-hez képest.

	NerKor (F1)	SzNER (F1)	NP (F1)	2-class (Pontosság)	5-class (Pontosság)
huBERT	90,18	97,51	96,97	85,92	68,50
GPT-2	69,43	88,06	85,02	80,37	61,00

2. táblázat. NER, NP és Szentiment analízis eredmények.

A GPT-2 modellek erősebb oldala a szövegenerálás. A szövegeneráló GPT-2 modellek kvantitatív kiértékeléséhez a perplexitást választottam, mivel az eredeti GPT-2 kutatásban is ezt alkalmazták. A 3. táblázatban láthatóak a modellek perplexitás értékei. Az összefoglaló generáló modell 19,5 perplexitást ért el, ami kiemelendő, hiszen az eredeti GPT-2 kutatásban 18,4-t értek el.

A hírgeneráló modell 22,06 perplexitást ért el, ami gyengébb, mint ami az eredeti publikációban szerepel. Ez a várakozásomnak megfelel, hiszen kevesebb adaton tanítottam, azonban ez nem is probléma, hiszen nem célozom visszaismételnéni a híreket, hanem a modellnek egy adott témában kreatívan kell magától előállítania tartalmakat. A hírgeneráló modell megtalálható a Hugging Face oldalon<sup>14</sup>.

A versgeneráló modell perplexitása már jóval gyengébb, 47,5 körüli érték. Ez azonban egy vers esetében nem feltétlenül negatívum, hiszen nem szóról szóra szeretném visszaidézni a szerző verseit, a célozom, hogy nagyobb kreativitást adjak a modellnek. A versgeneráló modell megtalálható a Hugging Face oldalon<sup>15</sup>.

A 4. táblázatban látható az absztraktív összefoglaló generálás eredménye. A generálás folyamán megadtam a tesztanyag cikk szövegét, majd utána hozzáillesztettem a „TL;DR:” kifejezést, jelezve, hogy összefoglalót várok generált

<sup>14</sup> <https://huggingface.co/NYTK/text-generation-news-gpt2-small-hungarian>

<sup>15</sup> <https://huggingface.co/NYTK/text-generation-poem-petofi-gpt2-small-hungarian>

	Perplexitás
SUM	19,52
HÍR	22,06
VERS	47,46

3. táblázat. A szöveggeneráló GPT-2 nyelvi modellek perplexitása.

szöveggént. A modellel ezután 3 szöveget generáltattam, majd mindegyik szöveg első mondatát összekonkatenáltam. Így eredményként 3 mondat alkotja a generált összefoglalást.

Az összehasonlíthatóság végett beillesztettem a Yang és mtsai (2021) kutatásából a huBERT alapú PreSumm modell legjobb eredményét. A nemzetközi kutatásokban az F-mértéket (F) szokás megjeleníteni, azonban mind a PreSumm, mind a GPT-2 megoldása több mondatot is visszaad eredményül. A GPT-2 által javasolt megoldásban konkrétan 3 mondatot. Ilyen szempontból releváns a ROUGE fedés értékek megjelenítése, hiszen ha sok szöveget generál, magasabb lesz a fedés (R) és alacsonyabb a pontosság (P). Az eredmények azt mutatják, hogy a GPT-2 által javasolt összefoglalás generáló módszer (ami nem egy „hagyományos” szövegből szöveg – seq2seq – enkóder-dekóder architektúrán alapuló) versenyképes eredményt képes produkálni. Ha az F-mértéket nézzük, ROUGE-1 esetében még magasabb eredményt is el tudott érni a GPT-2.

	PreSumm	GPT-2
ROUGE-1 R	57.07	40,90
ROUGE-1 P	14.63	17,53
ROUGE-1 F	22.42	<b>23,06</b>
ROUGE-2 R	26.97	11,89
ROUGE-2 P	6.64	4,94
ROUGE-2 F	10.24	6,56
ROUGE-L R	48.28	27,46
ROUGE-L P	12.19	11,35
ROUGE-L F	18.72	15,04

4. táblázat. Absztraktív összefoglaló eredménye.

Az 5. táblázatban látható egy példa összefoglaló generálásra. Közelebbről megvizsgálva a generált eredményt más képet kapunk a kvantitatív eredményhez képest. A példában az látható, hogy releváns mondatokat generált a modell, azonban inkább „hallucinációk”, mint összefoglalás. Ez egy nagy különbség egy seq2seq módszerhez képest. A seq2seq nehezen tud kilépni az adott bemeneti szöveg korlátaiból, míg ezzel a nyelvi modellezés módszerrel való összegzés sokkal „kreatívabb”, ami a jelen feladatban inkább hátrányt jelent, mint előnyt. Összehasonlításképpen a PreSumm által generált összefoglalót is betettem, ennek eredményeként az látható, hogy a PreSumm is „hallucinál”, de sokkal kevesebbet és relevánsabb kifejezéseket. A PreSumm generált egy „autóalkatrész-műhely” kife-



jezést, ami nem volt a cikkben, azonban az eredeti leadben látható, hogy valóban egy autóalkatrész-műhely gyulladt ki. Ez egy kifejezetten precíz „hallucináció”. A GPT-2 ehelyett „ruhaterelő cég” és „családi ház” kifejezéseket „hallucinált”, amelyek egyáltalán nem relevánsak.

---

**Cikk szövege:**

Az Erika utcai kétszáz négyzetméteres műhelyhez az érdi tűzoltókat és a megyei katasztrófavédelem műveleti szolgálatát riasztották . A tűzoltók egy oxigén- és egy disszagálpalackot hoztak ki az ingatlanból , az utóbbit folyamatosan hűtötték , így végül nem volt szükség a kilövésére . A műhelyben keletkezett tűz veszélyeztette a szomszédos száz négyzetméteres irodaépületet is , de három vízsugárral sikerült a lángokat megfékezni . Személyi sérülés nem történt .

---

**Eredeti lead:**

Lángra kapott egy érdi autószerelő-műhelyben álló kisteherautó , a tűz átterjedt a műhely berendezési tárgyaira és tetőszerkezetére is csütörtök hajnalban – közölte a Pest megyei katasztrófavédelem .

---

**GPT-2-SZUM által generált összefoglaló:**

Kétezer liter mérgező gázt próbáltak ki Budapesten az érdi tűzoltók . Kigyulladt egy százezer forintos ruhaterelő cég telephelyének melléképülete péntek délután a Pest megyei Visegrádon . Meghalt egy ercsi családi ház tetőszerkezetében két munkás , amikor kigyulladt egy 1000 négyzetméteres műhely .

---

**PreSumm által generált összefoglaló:**

Kigyulladt egy autóalkatrész-műhely Érden , a tűzben senki sem sérült meg - közölte a Pest Megyei Katasztrófavédelmi Igazgatóság szóvivője . A tűzben senki nem sérült meg , de a tüzet sikerült megfékezni . A tüzet sikerült eloltani . Az épületből a szomszédos irodaépületet is ki kellett löni , de az ott keletkezett tüzet sikerült megfékeznie a tűzoltóknak – közölték a katasztrófavédelemtől . Az épületet kiürítették , a tüzet eloltották , senki nem sebesült meg – közölték szerdán . A tüzet a tűzoltók fékezték meg , a tűz a szomszédos irodaépületre is átterjedt .

---

5. táblázat. Példa absztraktív összefoglalás generálásra.

A GPT módszerek legnagyobb erőssége a szöveggenerálás. Az eredeti GPT-2 publikációkban és demókban is elsősorban ezen a területen mutatnak látványos példákat. Ezért kutatásomban ennek próbáltam utánajárni. A 6. és a 7. táblázatban láthatóak a szöveggenerálási példák. A SUM modell esetében a szöveggenerálás hosszúságát 256-ra állítottam. A modell három különböző szöveget generált. Kezdesnek megadtam néhány szót, majd arra generált a modell folytatást. Egyik érdekes dolog, hogy nem írtam vesszőt a kezdőszöveg végére, azt már a gép generálta, illetve a 3. példában tovább toldalékolta az általam megadott szöveget. Manuálisan megvizsgáltam a folytatásokat, ezek a szövegek ilyen formában nem szerepelnek a korpuszban, vagyis a folytatás szövegét „kreatív” módon állítja össze a modell. Továbbá érdekes még, hogy kellő hosszúságú szöveg generálása esetén az összefoglaló TL;DR: részt is legenerálja, ez azonban nem váratlan, hiszen ilyen szövegen tanult.

---

**Kezdő szöveg:**

Szeptember végén zárul a balatoni szezon

---

**1. generált szöveg:**

Szeptember végén zárul a balatoni szezon, a Magyar Fürdőszövetség pedig ősszel még két nagy meghívásos fürdőhotel építését tervezi. A rendezvényszervezők és a Balaton Szépe Bizottság üzleti tervében csaknem 100 millió forintot különítettek el a fürdőfelújításra és korszerűsítésre.

TL;DR:

A balatoni szezonkezdet közeleddével a kormányzati holdingok telítődnek.

---

**2. generált szöveg:**

Szeptember végén zárul a balatoni szezon, a rendezvény honlapja szerint a Balaton északi partján két kemping, a Balatonfüred-Balatonkenese és a balatonfüredi Tóni kemping működik, a kisebbeknek is három kempingje van. A szállást-köröket 650-550-en választják a Balaton partján, nyáron 500-1000 látogató is lehet.

A júniusban átadandó kenesei Siotour 2012 a Balaton partján a négy magyar nemzet üdülőhelyeinek, egy-egy kempingnek ad otthont.

TL;DR:

Balatonakarattya egykori üdülőfalujának termálvizes strandját bővítette a Magyar Turisztikai Ügynökség.

---

**3. generált szöveg:**

Szeptember végén zárul a balatoni szezonja. Csonthegynek ki is gyulladt a Velencei-tó csabai pihenőhelye, ami miatt a helyiek nem tudnak felkészülni a csütörtök délutáni kéthetes kempingezésre.

---

6. táblázat. Hírgenerálás példa a SUM modellel.

A HÍR modell esetében 512 hosszúságúra állítottam a generált szöveg hosszát és megállási feltételként a „\n” tokent adtam meg, hiszen a modell ezen token után új témában kezd el hírt generálni. A 7. táblázat példájában az látható, hogy az első szöveg rövidebb, míg a második és a harmadik szövegek kicsit hosszabbak. A bekezdéseket elválasztó tokent (<n>) is a gép magától generálta. A harmadik példa témájában már nagyon eltér az előző kettő és az eredeti mondattól. Sőt megfogalmazásában rendkívül szokatlan mondatokat generált a modell.

A generált szövegek általánosságban nézve nyelvtanilag helyesek és könnyen érthetőek. A mondatok viszonylag logikusan vannak felépítve (idő hiányában manuálisan történt a vizsgálat, kettő ember által). Azonban szemantikai szempontból szokatlan szövegeket kapunk, kevésbé összetartozó kifejezéseket láthatunk egy mondaton belül.

Végezetül, a 8. táblázatban látható a versgenerálás eredménye. A példában megadtam egy kezdő sort írásjelekkel, majd a modell erre generált folytatást. Ebben az esetben szintén 256-ra állítottam a kimenet méretét, azonban néhány sor után ismételni kezdte azokat, ezért levágtam belőle, az első 10 sor látható a példában. Ellenőriztem, ezek a sorok nem szerepelnek az eredeti művekben, a modell maga rakta össze ilyen formában őket. Külön érdekesség, hogy a sortöréseket is a modell generálta.

A modellek kipróbálhatóak a demó oldalunkon, ahol további példák is találhatóak: <https://juniper.nytud.hu/demo>.

---

**Kezdő szöveg:**

Szeptember végén zárul a balatoni szezon

---

**1. generált szöveg:**

Szeptember végén zárul a balatoni szezon. Csütörtökön a hagyományos kempingstrandot, a Festival Balett Singers Travel rendezvényét és a Mátyás-templom barokk orgonaeépítő üzemet alakítanak ki. A lehetőségekre a [www.volstadt.hu](http://www.volstadt.hu) honlapon várják az érdeklődőket.

---

**2. generált szöveg:**

Szeptember végén zárul a balatoni szezon. <n>

Az Egy Balaton Nap célja, hogy a fesztiválozók az idén is felcsendülhessenek kedvenc dalaikból. <n>

Az önkormányzat is részt vesz a fesztivál "témacsúcsán", és a látogatókat szeretettel várják a belvárosi strandokra. <n>

A település lakói idén júniusban kilenc helyszínen koccinhattak a

Balaton Soundon, idén a négynapos fesztiválon összesen több mint 250 ezer ember volt kíváncsi a legnépszerűbb műsorszámokra. <n>

A Strand Fesztivál szakmai partnerei a Balaton Soundon fellépett előadók, akik a Magyarország legnagyobb újdonságai - sportfesztiválok, ismert előadók és zenekarok - mellett rendhagyó, különleges kísérőprogramok, színházi események is várták a résztvevőket. <n>

Mint elmondta, a főváros strandjai ezen a nyáron is bővelkednek, hiszen augusztus 20-tól kezdve október 31-ig többször, több helyen lesz lehetőségük élő bejelentkezéseket tenni a fesztiválra. <n>

"A három nap minden más sportághoz hasonlóan a hazai vízilabda-válogatotté, az összeállításunk egyértelműen a magyar csapaté, ráadásul - akár négyéves lemaradással is - alapvetően van jelen a magyar vízilabda-válogatott: a tavaszi hónapokban eddig még csak egyszer szerepelt magyar csapat, amely Európa-bajnokságon szerepelt és azóta mindössze egyszer marad a döntőben.

A hazai úszósportban azonban valószínűleg nem lesz olyan, mint egy átlagos nyári napon, hiszen a hazai válogatott kiesése esetén sokáig nem áll szándékában a felnőtt nemzeti együttessel is kikerülni" - mondta a polgármester.

---

**3. generált szöveg:**

Szeptember végén zárul a balatoni szezon, az első fészek hamar ki is nyílik majd. <n>

A gyermek egyéves koráig anyatejet tartalmazó tápszert ad a borjúnak. A vádlott két éven át az erszényt, egy gyermekbántalmazási programot is megnyert, amelyből táplálkozni is tudott, de nem tudott uralkodni magzatokon. <n>

Másnap a másodrendű vádlott fiatal, büntetett előéletű testvérével és a nő egyik barátjával vizslatta a reggeli szekrényt, és a teljes adagot elfogyasztotta. A gyerekeket éjszaka, munkaidőben végzett tevékenységgel foglalkoztatta, éjszaka a mosott borral többször megivott, majd megvendégtelt és a magára hagyott pénzt kivette a biztonsági őrök kezéből. <n>

A gyerekeket szórakozás céljából kivette a tejesebdozból, hogy elrejtse az ablaküvegre szerelt, hideg falú ajtó rácsait, majd távoztak. Az ablakon keresztül odalett az a hét alsó tagozatos diák, akik éjjel - a szülők halálával egy időben - megpróbálták bezárni az ajtót. <n>

A gyerekek ki tudtak szabadulni az ajtókról, illetve elhajtottak a garázsban, a házba bejutni azonban már nem tudtak. A kislány a rendőrségen végzett beismerő vallomásaiban tagadta, hogy ő robbantotta volna fel a falat, és a körözés alatt is tagadta a bűncselekményt. Az ügyészség a vádiratban valamennyi gyilkosságot felfüggesztett börtönbüntetéssel, míg társait életfogytig tartó szabadságvesztéssel, illetve közügyektől eltiltással sújtotta. <n>

Az elsőrendű vádlott mindvégig tagadta bűnösségét, vallomását részben valótlanak vallotta.

---

7. táblázat. Hírgenerálás példa a HÍR modellel.

---

**Kezdő szöveg:**

Szegeden, január végén,

---

Szegeden, január végén, őszapófa.

Hogy mulatságos ünnepe vagy multság, az már rosz kéz,

Menyasszonynak végletes reménye?

Mégis el vannak választva.

Elbeszélve... szégyenszem.

Ne hagyj szemünknek lobogója,

S ím a szivárvány,

S, mi van belőlem... harsogó csonka szem.

„És ha e sugarat áldozánk,

Kínálok én is bokrétával!”

---

Szólj, hogy lesz az, hogy itt néhány perc.

8. táblázat. Versgenerálás eredménye.

---

## 4. Összegzés

Kutatásomban egy GPT-2 modellt tanítottam magyar nyelvre, amelyet különböző nyelvtechnológiai feladatokra finomhangoltam. A kísérleteim során a betanított GPT-2 modellt szövegosztályozás, absztraktív összefoglalás és szöveggenerálás feladatain tettem próbára. Az eredmények azt mutatták, hogy a GPT-2 modellem, mint autoregresszív modell, elsősorban szöveggenerálás feladataira teljesít jól, azon belül is a hír- és versgenerálás feladatában produkálnak figyelemre méltó eredményeket. Érdeemes őket kísérleti modelleknek tekinteni, hiszen az eredeti publikációkban a szerzők óriási mennyiségű adaton tanították modelleiket, ezzel ellentétben a modelljeimet csak a magyar Wikipédián és pár százezer hírrel tanítottam be.

## 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)
- Bajgar, O., Kadlec, R., Kleindienst, J.: Embracing data abundance: Booktest dataset for reading comprehension. In: Proceedings of the 5th International Conference on Learning Representations (ICLR 2017). Palais des Congrès Neptune, Toulon, France (2017)
- Black, S., Leo, G., Wang, P., Leahy, C., Biderman, S.: GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow (Mar 2021), <https://doi.org/10.5281/zenodo.5297715>, If you use this software, please cite it using these metadata.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A.,

- Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (szerk.) *Advances in Neural Information Processing Systems*. vol. 33, pp. 1877–1901. Curran Associates, Inc. (2020)
- Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.: Natural language processing (almost) from scratch. *J. Mach. Learn. Res.* 12(null), 2493–2537 (Nov 2011)
- Csendes, D., Csirik, J., Gyimóthy, T., Kocsor, A.: The szeged treebank. In: Matoušek, V., Mautner, P., Pavelka, T. (szerk.) *Text, Speech and Dialogue*. pp. 123–131. Springer Berlin Heidelberg, Berlin, Heidelberg (2005)
- Dehghani, M., Gouws, S., Vinyals, O., Uszkoreit, J., Łukasz Kaiser: Universal transformers. In: *Proceedings of the 7th International Conference on Learning Representations (ICLR 2019)*. New Orleans, LA, USA (2019)
- 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)
- Graff, D., Kong, J., Chen, K., Maeda, K.: English gigaword. *Linguistic Data Consortium, Philadelphia* 4(1), 34 (2003)
- Graves, A.: Adaptive computation time for recurrent neural networks (2017)
- Hestness, J., Narang, S., Ardalani, N., Diamos, G., Jun, H., Kianinejad, H., Patwary, M.M.A., Yang, Y., Zhou, Y.: Deep learning scaling is predictable, empirically (2017)
- Jozefowicz, R., Vinyals, O., Schuster, M., Shazeer, N., Wu, Y.: Exploring the limits of language modeling (2016)
- Karpathy, A., Johnson, J., Fei-Fei, L.: Visualizing and understanding recurrent networks. In: *Proceedings of the 4th International Conference on Learning Representations (ICLR 2016)*. Caribe Hilton, San Juan, Puerto Rico (2016)
- Liu, P.J., Saleh, M., Pot, E., Goodrich, B., Sepassi, R., Kaiser, L., Shazeer, N.: Generating wikipedia by summarizing long sequences. In: *Proceedings of the 6th International Conference on Learning Representations (ICLR 2018)*. Vancouver Convention Center, Vancouver, BC, Canada (2018)
- Merity, S., Xiong, C., Bradbury, J., Socher, R.: Pointer sentinel mixture models (2017)
- Nemeskey, D.M.: *Natural Language Processing Methods for Language Modeling*. Ph.D.-értekezés, Eötvös Loránd University (2020)
- 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), <http://jmlr.org/papers/v21/20-074.html>
- Rei, M.: Semi-supervised multitask learning for sequence labeling. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. pp. 2121–2130. Association for Computational Linguistics, Vancouver, Canada (Jul 2017)
- Shoeybi, M., Patwary, M., Puri, R., LeGresley, P., Casper, J., Catanzaro, B.: *Megatron-lm: Training multi-billion parameter language models using model parallelism* (2020)
- Simon, E., Vadász, N.: Introducing nytk-nerkor, A gold standard hungarian named entity annotated corpus. In: Ekstein, K., Pártl, F., Konopík, M. (szerk.) *Text, Speech, and Dialogue - 24th International Conference, TSD 2021, Olomouc, Czech Republic, September 6-9, 2021, Proceedings. Lecture Notes in Computer Science*, vol. 12848, pp. 222–234. Springer (2021)
- Szarvas, G., Farkas, R., Kocsor, A.: A multilingual named entity recognition system using boosting and c4.5 decision tree learning algorithms. In: Todorovski, L., Lavrač, N., Jantke, K.P. (szerk.) *Discovery Science*. pp. 267–278. Springer Berlin Heidelberg, Berlin, Heidelberg (2006)
- Yang, Z.G., Agócs, Á., Kusper, G., Váradi, T.: Abstractive text summarization for hungarian. *Annales Mathematicae et Informaticae* 53, 299–316 (2021)
- Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., Choi, Y.: Defending against neural fake news. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (szerk.) *Advances in Neural Information Processing Systems*. vol. 32. Curran Associates, Inc. (2019)