Automatikus írásjel-visszaállítás és Nagybetűsítés statikus korpuszon transzformer modellen alapuló neurális gépi fordítással

Yang Zijian Győző

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

Kivonat Cikkemben egy írásjelvisszaállító és nagybetűsítő programot mutatok be, amelyet a jelenkori „state-of-the-art" transzformer modellen alapuló neurális gépi fordító rendszerrel tanítottam be. A mobil eszközökön történő üzenetírás elterjedésével és a minél gyorsabb szövegbevitelre való törekvéssel tömeges jelenséggé vált a hibás szövegek írása. Ennek egyik következménye, hogy a interneten elérhető – főleg a szociális médiából származó – korpuszok egy része hibás. Többek között írásjelek hiányoznak, vagy végig kisbetűvel írnak. Az így létrejött korpuszok nem alkalmasak különböző kutatásokhoz, csak tisztítás után. A tisztítás folyamata időigényes, ezért igény van különböző korpusztisztító módszerekre. Az általam létrehozott rendszer, annak ellenére, hogy semmilyen morfológiai és szintaktikai elemzőt nem használ, közel 81%-os f-mértékkel tudja helyesen visszaállítani az alapírásjeleket és elvégezni a nagybetűsítést magyar nyelv esetében.
Kulcsszavak: írásjel-visszaállítás, nagybetűsítés, neruálisháló-alapú gépi fordítás, NMT, transzformer modell

1. Bevezetés

Napjainkban a számítógépes nyelvészek számára nagy lehetőséget nyújtanak az interneten elérhető nagy mennyiségű szövegek. Számos részterületen használjuk a weboldalakról összegyűjtött korpuszokat, mint például a gépi fordítás, a szövegkivonatolás vagy az érzelemdetektálás. Ezekhez a feladatokhoz viszont nélkülözhetetlen, hogy a vizsgált szöveg a lehető legjobb minőségű legyen.

A mobil eszközökön írt szövegek és üzenetek esetében tömegjelenséggé vált az ékezetes betűk és írásjelek elhagyása. Ennek következményeként léteznek olyan korpuszok is, amelyek egy része ékezet- és írásjelmentes, vagy kisbetűvel írták. Így nem működnek rajtuk a természetes szövegen betanított szövegfeldolgozó modellek.

Magyar nyelvre Dömötör és Yang (2018) végeztek kutatást különböző korpuszokban a nem sztenderd hibák előfordulására. Első sorban beszélt nyelvi és személyes alkorpuszokból indultak ki, hiszen ott lehet nagyobb mennyiségben

hibás szöveget találni. Kimutatták, hogy a nem sztenderd hibák közel 30%-át kiugróan az írásjelek és nagybetűk elhagyása teszi ki.

Az elmúlt években a neurálishálózat-alapú módszerek eredményei túlszárnyalták az addigi legjobb rendszereket. Ez a nyelvtechnológia területén is megmutatkozik, ezért célom az volt, hogy megvizsgáljam az írásjel- és nagybetűvisszaállítás problémáját a jelenlegi „state-of-the-art" NMT-alapú rendszerrel.

2. Kapcsolódó irodalom

Beszédtechnológia terén az írásjel-visszaállítás egy fontos feladat. Nehézsége abban rejlik, hogy a szöveg dinamikusan változik és mindig az adott környezethez igazodva kell visszaállítani az írásjelet.

Öktem és mtsai (2017) az automatikus beszédfelismerés feladatában az RNN neurális hálózat segítségével állítják vissza az írásjeleket. Transformers modell alapú írásjel- és nagybetű-visszaállítást Vāravs és Salimbajevs (2018) végezték lett és angol nyelvre. Kutatásukban a neurális gépi fordítást használták rendszerük betanítására. Nguyen és mtsai (2019) kutatásukban transformer modellel az írásjel- és nagybetű-visszaállítás mellett főnévi csoportok felismerés feladatát is belevették a tanításba. Alam és mtsai (2020) az angol mellett a kevés tanítóanyaggal rendelkező bengáli nyelvre tettek kísérletet az írásjelek visszaállítására. Mivel kevés a nyersanyag, ezért a különböző BERT (Devlin és mtsai, 2019) alapú modellek finomhangolásával tanítottak be modelleket, amelyekkel a problémát megoldják.

Magyar nyelvre Tündik és mtsai (2018) az RNN hálózat segítségével állítják vissza az írásjeleket.

Kutatásom nem a dinamikusan változó szövegekre koncentrál, hanem a statikus korpuszokra. Módszerem első sorban korpusztisztításra alkalmas.

3. A korpusz és a fordító rendszer

A korpusz-alapú gépi fordító rendszer lényege, hogy transzformációt képez tetszőleges forrás- és célnyelvi mondatok között, ahol a rendszer betanításához nem kell más, mint egy kétnyelvű párhuzamos korpusz. Az írásjelek és nagybetűk helyreállítására a gépi fordítás módszereit választottam, mivel az írásjellel ellátott nagybetűkkel rendelkező mondatok grammatikailag, szókincsileg és szószerkezetileg nagyon hasonlóak az írásjel és nagybetű nélküli párjukhoz.

A neurális hálózat tanításához nagy mennyiségű tanítóanyagra van szükség, melynek előállítása a jelen feladathoz igen könnyű. A tanítóanyag létrehozásához annyit kellett tenni, hogy egy egynyelvű korpusz írásjeleit eltávolítom és a szöveget kisbetűsítem.

A korpusz létrehozásához az online elérhető Open Subtitles[1] nevű angol-magyar párhuzamos korpuszának magyar oldali szövegét használtam. A korpusz

[1] http://opus.nlpl.eu/OpenSubtitles-v2018.php

TV és mozi filmekre létrehozott feliratokból áll. Ennek megfelelően főleg rövidebb, informális mondatokat tartalmaz. A gépi fordító rendszer célnyelvi korpuszának előállításához a mondatokban az írásjeleket kitöröltem, majd a szöveget kisbetűsítettem.

A korpusz megközelítőleg 29 millió szegmensből áll, melyből 5000 mondatot validációs és 3000 mondatot tesztelési célra elkülönítettem. A korpusz az egyik legnagyobb szabadon hozzáférhető párhuzamos tanítóanyagnak számít, ellenben mérete elmarad az egynyelvű tanítóanyagokétól. Választásom azért esett erre az adathalmazra, mert több párhuzamos kutatásom során is használom, és néhány koprusztisztító lépést már előzetesen eszközöltünk rajta. Kivettem azokat a mondatokat, amelyek speciális karaktereket (pl. kínai, japán, cirill stb.) tartalmaztak, valamint a teszt halmaz mondatait kézzel kijavítottam. Mérete elégséges a neurális hálózatok helyes betanítására, valamint a tanítási idő is viszonylag kezelhető marad (1-2 nap). Végül utolsó szempont, hogy a feliratok gyakran hasonló mondatszerkezetűek, mint a beszéltnyelvi mondatok, amelyekben a legtöbb nem sztenderd hiba található.

A 2010-es évek első felére a statisztikai gépi fordítórendszerek elérték teljesítőképességük határát. Az alapjait képező módszert és a létrehozott keretrendszereket a kutatók nagyon sok befektetett munkája ellenére lényegében nem sikerült tovább javítani. Az áttörést (Bahdanau és mtsai, 2015) rendszere hozta el, ami egy figyelmi modellel támogatott enkóder-dekóder architektúrájú NMT rendszer volt. A modell lényege, hogy kettéválasztja a fordítás folyamatát két elkülöníthető részre. A kódolás során lényegében egy RNN-alapú seq2seq modellt hoz létre, tehát a szóbeágyazási modellhez hasonlóan a fordítandó modellekből egy n-dimenziós vektort készít. 1. ábrán ez a vektor felel meg az ábra közepén látható piros/sötét node-nak. A második fázis a dekódolás, ahol a mondatvektorból generálja ki a célnyelvi mondatot egy RNN réteg segítségével.

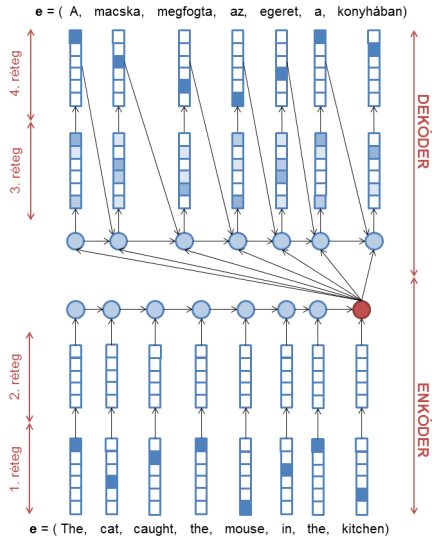
Innentől számítva az NMT rendszerek átvették a vezető szerepet az SMT-től. 2017-ben a Google cég munkatársai (Vaswani és mtsai, 2017) publikálták és szabadon hozzáférhetővé tették az úgynevezett multi-attention réteggel támogatott NMT rendszerüket. Ezt a szakirodalomban transzformer-alapú architektúrának nevezik. A módszer lényege, hogy az eddigi egy helyett több figyelmi réteget helyeztek el a rendszerben, ami segítségével nagymértékben nőtt a többértelmű szavak fordításának minősége.

Munkám során a Marian NMT(Junczys-Dowmunt és mtsai, 2018) nevű keretrendszert használtam, ami egy c++ nyelven íródott szabadon hozzáférhető programcsomag. Könnyen telepíthető, jól dokumentált, memória- és erőforrás-optimális implementációjának köszönhetően[2] az akadémiai felhasználók és fejlesztők által leggyakrabban használt eszköz (Barrault és mtsai, 2019).

Manapság a neurális hálózat alapú modellek tanításához részszó (subword) tokenizálót (Sennrich és mtsai, 2015) használnak, hogy csökkentsék a szótárok méretét, és közben kezeljék az ismeretlen szavak problémáját.

A BPE (Byte Pair Encoding) egy adattömörítő eljárás, ahol a leggyakoribb bájtpárokat egy olyan bájttal helyettesítjük, amely nem szerepel magában az

[2] https://marian-nmt.github.io/



1. ábra: Enkóder-dekóder architektúra vázlatos rajza

adatban. Az eljárás a korpuszon először egy karakteralapú szótárat hoz létre, ahol minden szót karakterek sorozataként ábrázol. Ezután gyakoriság alapján a gyakori karaktersorozatokat önálló tokenekként kezeli. Ezzel az adat tömörítése mellett az ismeretlen szavak kezelését is megoldja, hiszen a részszavakból előállítható egy olyan összetétel, amely nem szerepelt eredetileg a korpuszban.

Ezt a módszert fejlesztették tovább (Kudo és Richardson, 2018). Az általuk létrehozott Sentence Piece nevű eszköz egy felügyelet nélküli szöveg tokenizáló és detokenizáló, melyet elsősorban a neurálishálózat-alapú gépi tanulásos feladatokhoz fejlesztettek ki. Implementálva van benne a BPE metrika, ami egy unigram nyelvmodellel (Kudo, 2018) van súlyozva. Használatával elhagyhatók a költséges nyelvspecifikus előfeldolgozási lépések, mint például a tokenizálás vagy a kisbetűsítés. A módszer lényege, hogy a természetes szöveget úgy alakítja át, hogy abban a különböző „szavak” száma korlátos legyen, valamint az így létrejött tanítóanyagban nem lesznek ismeretlen szavak. Ennek köszönhetően a neurális hálózatok paraméterszáma nagymértékben csökkenthető.

4. Kísérletek

Először megszámoltam (lásd 1. táblázat), hogy hányszor szerepelnek a számomra releváns esetek, vagyis az írásjelek és a nagybetűs szavak.

Az NMT tanításához a Marian neurális gépi fordítórendszert használtam. A rendszer fontos jellemzője a SPM technológia. A rendszerem tanításához az alábbi paramétereket használtam:

	Tanító	Valid	Teszt
nagybetűs szavak	19 586 281	53 237	31 558
"."	8 025 537	21 782	12 909
","	13 043 393	35 241	20 886
"?"	257 005	712	409
"!"	115 393	306	181
";"	53 674	149	90
":"	753 380	2046	1 270
„" (magyar kezdő (alsó) idézőjel)	834 363	2 183	1 286
"”" (magyar záró (felső) idézőjel)	820 681	2 142	1 317
"-" (kötőjel)	372 8116	10 155	5 939
"–" (nagykötőjel)	1 026 991	2 752	1 812

1. táblázat. Írásjelek és nagybetűs szavak előfordulásai a korpuszokban

– Sentence Piece: szótárméret: 16000; egy szótár a forrás- és egy a célnyelvi korpusznak; karakter lefedettség a teszt korpuszon: 100%
– Transformer modell: enkóder és dekóder rétegeinek száma: 6; transformer-dropout: 0,1;
– learning rate: 0,0003; lr-warmup: 16000; lr-decay-inv-sqrt: 16000;
– optimizer-params: 0,9 0,98 1e-09; beam-size: 6; normalize: 0,6
– label-smoothing: 0,1; exponential-smoothing

Kétféle modellt készítettem:

– alap: kisbetűsítés és alapírásjelek elhagyása:
  - ".": pont
  - ",": vessző
  - "?": kérdőjel
  - "!": felkiáltójel
– bővített: kisbetűsítés és bővített írásjelek elhagyása:
  - az alapírásjelek
  - ";": pontosvessző
  - ":": kettőspont
  - „": (magyar kezdő (alsó) idézőjel)
  - "”": (magyar záró (felső) idézőjel)
  - "-": (kötőjel)
  - "–": (nagykötőjel)

5. Eredmények

Kutatásom során megmértem a gép által adott szóalapú eredmény pontosságát (precision), fedését (recall) és az F-mértékét. Mivel a gépi fordítás során az eredetileg helyes szavak is megváltozhatnak, az összes szóra szükséges megvizsgálni a fordítás pontosságát (ALL). Végeztem külön kiértékelést csak azokra a szavakra, amelyek a kutatásom számára relevánsak (REL), beleértve az írásjelekkel rendelkező és nagybetűs szavakat egyaránt.

Emellett külön megmértem azt, hogy a modelljeim az írásjeleket (alap és bővített), valamint a nagybetűs szavakat milyen mértékben tudták visszaállítani.

A 2. táblázat eredményei alapján láthatjuk, hogy az általam létrehozott rendszer teljesítménye, amely transzformer modellt és Sentence Piece tokenizálót használ, az összes szóra nézve meghaladja a 92%-os F-mértéket. Az alapírásjelek és nagybetűk visszaállítását pedig 81%-os pontossággal tudja a rendszerem elvégezni. Az eredmények azt mutatják, hogy a nagybetűsítés feladatát pontosabban végzi, mint az írásjel-visszaállítást.

	Pontosság	Fedés	F-mérték
alap + nagybetű (ALL)	93,28%	91,43%	92,34%
alap + nagybetű (REL)	82,88%	79,65%	81,23%
bővített + nagybetű (ALL)	91,09%	90,31%	90,70%
bővített + nagybetű (REL)	78,72%	76,16%	77,42%
nagybetű	85,88%	83,33%	84,59%
alapírásjelek	79,60%	75,38%	77,43%
bővített írásjelek	73,89%	70,17%	71,98%

2. táblázat. Írásjel-visszaállítás és nagybetűsítés eredményei

	Pontosság	Fedés	F-mérték
alap "."	74,43%	73,09%	73,75%
alap ","	82,89%	77,95%	80,35%
alap "?"	67,15%	38,49%	48,93%
alap "!"	52,17%	9,75%	16,43%
bővített "."	70,88%	73,57%	72,19%
bővített ","	78,08%	78,77%	78,42%
bővített "?"	66,66%	36,67%	47,31%
bővített "!"	58,69%	14,91%	23,78%

3. táblázat. Alap írásjelek visszaállításának részletes eredményei

Végül kiértékeltem a két modell alapírásjeleinek visszaállításának teljesítményét külön-külön (lásd 3. táblázat). A 3. táblázat eredményei alapján a kérdőjelek és a felkiáltójelek értékei alacsonyak. Ez annak tulajdonítható, hogy a tesztanyagban ezek az írásjelek elég kevésszer szerepelnek (lásd 1. táblázat). Legjobban a vesszőket tudja a rendszer visszaállítani, közel 80%-os F-mértékkel.

6. Összegzés

A kutatásommal létrehoztam egy írásjel- és nagybetű-visszaállító rendszert. A rendszer tanításához egy neuralishálózat-alapú gépi fordítórendszert használtam,

amely transzformer modellt és Sentence Piece tokenizálót használ. A rendszerem 81%-os F-mértékkel tudja helyesen visszaállítani az alapírásjeleket és a nagybetűket.

Továbblépési lehetőségként szeretném a modelleket a beszédtechnológia feladataira hangolni, valamint kipróbálni az új BERT alapú modellek teljesítményét is.

Köszönetnyilvánítás

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

Hivatkozások

Alam, T., Khan, A., Alam, F.: Punctuation restoration using transformer models for resource-rich and -poor languages. In: Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020). pp. 132–142. Association for Computational Linguistics, Online (Nov 2020)

Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: Bengio, Y., LeCun, Y. (szerk.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015), http://arxiv.org/abs/1409.0473

Barrault, L., Bojar, O., Costa-jussà, M.R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Koehn, P., Malmasi, S., Monz, C., MĂźller, M., Pal, S., Post, M., Zampieri, M.: Findings of the 2019 conference on machine translation (wmt19). In: Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1). pp. 1–61. Association for Computational Linguistics, Florence, Italy (August 2019), http://www.aclweb.org/anthology/W19-5301

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)

Dömötör, A., Yang, Z.G.: Így írtok ti: nem sztenderd szövegek hibatípusainak detektálása gépi tanulásos módszerrel. XIV. Magyar Számítógépes Nyelvészeti Konferencia pp. 305–316 (2018)

Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, T., Seide, F., Germann, U., Aji, A.F., Bogoychev, N., Martins, A.F.T., Birch, A.: Marian: Fast neural machine translation in C++. In: Proceedings of ACL 2018, System Demonstrations. pp. 116–121. Association for Computational Linguistics, Melbourne, Australia (Jul 2018)

Kudo, T.: Subword regularization: Improving neural network translation models with multiple subword candidates. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 66–75. Association for Computational Linguistics, Melbourne, Australia (Jul 2018)

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), https://www.aclweb.org/anthology/D18-2012

Nguyen, B., Nguyen, V., Nguyen, H., Pham, P., Nguyen, T.L., Do, T., Luong, C.: Fast and accurate capitalization and punctuation for automatic speech recognition using transformer and chunk merging. pp. 1–5 (10 2019)

Öktem, A., Farrús, M., Wanner, L.: Attentional parallel rnns for generating punctuation in transcribed speech. In: Camelin, N., Estève, Y., Martín-Vide, C. (szerk.) Statistical Language and Speech Processing. pp. 131–142. Springer International Publishing, Cham (2017)

Sennrich, R., Haddow, B., Birch, A.: Neural machine translation of rare words with subword units. CoRR abs/1508.07909 (2015), http://arxiv.org/abs/1508.07909

Tündik, M.Á., Tarján, B., Szaszák, Gy.: Televíziós feliratok írásjeleinek visszaállítása rekurrens neurális hálózatokkal. XIV. Magyar Számítógépes Nyelvészeti Konferencia pp. 183–195 (2018)

Vāravs, A., Salimbajevs, A.: Restoring punctuation and capitalization using transformer models. In: Dutoit, T., Martín-Vide, C., Pironkov, G. (szerk.) Statistical Language and Speech Processing. pp. 91–102. Springer International Publishing, Cham (2018)

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), http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf