Automatikus hibajavítás statikus szövegeken

Gulás Máté[1,2], Yang Zijian Győző[1], Dömötör Andrea[1,2], Laki László János[1]

[1]MTA-PPKE Magyar Nyelvtechnológiai Kutatócsoport
1083 Budapest, Práter u. 50/a.
[2]Pázmány Péter Katolikus Egyetem Bölcsészet- és Társadalomtudományi Kar
2087 Piliscsaba, Egyetem u. 1.
gulas.mate@hallgato.ppke.hu,
{yang.zijian.gyozo, domotor.andrea, laki.laszlo}@itk.ppke.hu

Kivonat Kutatásunk célja egy olyan neurális hálózat alapú automatikus hibajavító eszköz létrehozása, amely képes a korpuszok sztenderdizálására. A különböző nyelvtechnológiai feladatok modelljeinek betanításához fontos, hogy a tanítókorpuszok minél kevesebb zajt illetve hibát tartalmazzanak, hiszen a gyenge minőségű tanítókorpuszok rendszerint rosszabb eredményekhez vezethetnek. Az interneten elérhető szövegek nagy része informális, nem ellenőrzött forrásból (pl. közösségi média, fórumok) származik. Tanulmányunkban a közösségi médiában gyakran előforduló gyakori hibákra fókuszálunk. Célunk feltárni és elemezni a hibatípusokat, majd az előfordulásuk alapján statisztikát készíteni. A kiszámolt hiba-előfordulások arányát felhasználjuk egy hibajavító modell tanítására. Kutatásunkban egy transzformer modellen alapuló neurális gépi fordító rendszert használtunk fel a hibajavító modell tanítására. Eredményeink azt mutatják, hogy a neurális gépi fordítás módszere alkalmas a feladatra, azonban több olyan hibatípus is létezik, amelyek további kutatást igényelnek.
Kulcsszavak: hibajavítás, hiba-előfordulás, korpusztisztítás, neurális gépi fordítás

1. Bevezetés

Napjaink számítógépes nyelvészeti megoldásainak nagy részéhez elengedhetetlenek a nagyméretű korpuszok. A különböző nyelvtechnológiai feladatok modelljeinek betanításához fontos, hogy a tanítókorpuszok minél kevesebb zajt illetve hibát tartalmazzanak, hiszen a gyenge minőségű tanítókorpuszok rendszerint rosszabb eredményekhez vezetnek. Eredendően hibátlan korpuszt találni szinte lehetetlen, hiszen az interneten elérhető szövegek nagy része informális, nem ellenőrzött forrásból (pl. közösségi média, fórumok) származik, de még az ellenőrzött szövegekben is lehet hibákat találni.

Ennek a problémának a megoldására egy neurális gépi fordítás (NMT) alapú hibajavítót javaslunk. Ennek segítségével egyszerűen tudunk szövegkorpuszokat javítani. Hibajavítónk magában foglalja az előforduló hibák széles skálájának javítását.

2. Kapcsolódó irodalom

Magyar nyelvre Dömötör és Yang (2018) végeztek kutatást a nem sztenderd hibák előfordulására korpuszokban. Első sorban beszélt nyelvi és a 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. 15% körül van az elütések, helyesírási és nyelvi hibák, 10% körül a nehezen értelmezhető beszéltnyelvi szöveg, kicsivel kevesebb, de ugyancsak 10% körül az ékezetek hiánya, 5% körül a szegmentálási hibák és végül kevesebb mint 2% körül az idegen szavak. Kutatásunkban ezeket a hibákat vizsgáltuk meg még részletesebben, kifejezetten közösségi médiákból származó kommentek és szövegek között.

A hibák visszaállításának gépi fordítással történő első megközelítése Brockett és mtsai (2006) nevéhez fűződik. Tanulmányukban statisztikai gépi fordítást (SMT) használtak azzal a céllal, hogy az angolt második nyelvként tanulók főnevekkel kapcsolatos hibáit javítsák. Az interneten nagy számban előforduló főnévi hibák 61,81%-át sikerült javítaniuk módszerükkel.

Az ő megközelítésükre alapozva többek között Felice és mtsai (2014) egy szabály és SMT alapú hibrid hibajavító rendszert alkottak, míg Susanto és mtsai (2014) egy klasszifikációs és SMT alapú rendszer kombinációját mutatták be.

Ahogy a gépi fordítás területén a mondatalapú NMT egyre jobb eredményeket produkált a kifejezés-alapú SMT-vel szemben, a hibajavítás terén is egyre több NMT alapú megközelítés született, kezdetben azonban a kifejezés-alapú SMT rendszerek felülmúlták az NMT rendszereket. Az első jelentős NMT alapú hibajavító rendszert Junczys-Dowmunt és mtsai (2018b) mutatták be. Kutatásukban az alacsony-erőforrású NMT-k számos módszerét ültették át NMT alapú hibajavító rendszerükbe és további modell-független módszereket vezettek be. Hibajavító rendszerük több mint 10%-kal jobb eredmény produkált, mint az addigi state-of-the art NMT alapú hibajavító, és 2%-kal jobb eredményt ért el a CoNLL-2014 benchmark szerint, mint az addigi nem-neurális state-of-the-art rendszer.

Egy további úttörő NMT alapú hibajavító rendszert mutatott be Chollampatt és Tou Ng. (2018), akik egy többrétegű, konvolúciós encode-decoder neurális hálót használtak. A hálót olyan beágyazásokkal inicializálták, melyek felhasználják a karakter n-gram információt.

A közelmúltban nagyon népszerű lett az NMT alapú hibajavítás, több kutatás is született a témában. Ezek közül érdemes kiemelni Zhao és mtsai (2019) munkáját, akik egy másolási mechanizmust vezettek be: a változatlan és a szótáron kívüli szavakat közvetlenül átmásolták a forrásmondatból a célmondatba.

Jelenleg nincs tudomásunk arról, hogy magyar szerzők foglalkoztak volna NMT alapú hibajavítással, vagy külföldi szerzők magyar nyelvű NMT hibajavító rendszerrel.

3. Hibatípusok informális szövegekben

A hibajavító első lépéseként az volt a célunk, hogy felmérjük, milyen típusú hibák fordulnak elő informális szövegekben. Ehhez a TrendMiner projekt korpuszát[1] használtuk (Miháltz és mtsai, 2015). A korpusz 1,9 millió magyar nyelvű, politikai témájú Facebook hozzászólást tartalmaz morfológiai elemzéssel együtt. A korpusz kiválóan alkalmas a hibák keresésére, hiszen egyrészt a közösségi média kommentek nem szerkesztett, nem ellenőrzött mondatokból állnak, másrészt a morfológiai elemzés Out of Vocabulary (OOV) címkéje segíti a hibás szövegek keresését. A szövegek jellemző hibáinak feltérképezésére tehát azokat a kommenteket használtuk fel, amelyekben szerepel legalább egy szó OOV címkével. Az így automatikusan kiválasztott kommentekből 350-et vizsgáltunk meg részletesen. Ezekben számos olyan hibát is találtunk, amelyeket a morfológiai elemzés nem jelölt ismeretlen szónak.

A talált hibatípusokat három fő kategóriába osztottuk be aszerint, hogy milyen nyelvi elemzési szint szükséges a generálásukhoz. Az így kapott három kategória és a hibák típusai az 1–2. táblázatokban láthatók.

1. Felszíni alakból generálható
Ékezetek hiánya
Kisbetű/nagybetű tévesztés
Írásjelek hiánya, hibái
Magánhangzó hosszúság/rövidség
Mássalhangzó hosszúság/rövidség
Plusz vagy hiányzó karakter
Dátumok, számjegyes kifejezések hibái
j-ly tévesztés
Informális rövidítések (*h, vmi*, stb.)
Egyéb gyakori speciális hibák: pl. *-ban/-ba, sem/se, lesz/lessz*
Elgépelések

1. táblázat. A TrendMiner korpusz gyakori hibatípusai (felszíni alakból generálható hibák)

Jelen munkánkban nem foglalkozunk a morfológiai és a szintaktikai elemzés után generálható hibákkal. Ennek oka többek között az, hogy egyfelől az egyszerű/felszíni alakból generálható hibák alkotják a talált hibatípusok legnagyobb részét, másrészt ezeket a típusú hibákat viszonylag gyorsan és könnyen elő lehet állítani.

[1] http://corpus.nytud.hu/trendminer/

Elemzés szükséges	
2. Morfológia	**3. Szintaxis**
Összetett szavak különírása	Hibás egyeztetés
Tagadószó egybeírása a következő szóval	Ragozás hibái vagy hiánya
Kopula egybeírása az előző szóval	
Jelzős szerkezetek egybe-/különírása	
Módosítós szerkezetek egybe-/különírása	
Véletlenszerű egybeírás (szóközhiány)	
Magánhangzó-harmónia megsértése	
Suksükölés	
Igekötők egybe-/különírása	

2. táblázat. A TrendMiner korpusz gyakori hibatípusai (elemzéssel generálható hibák)

4. Módszer

4.1. A tanítókorpusz előállítása

Második lépésként előállítottunk egy olyan párhuzamos korpuszt, amely mindkét oldalán ugyanazokat a mondatokat tartalmazza, annyi különbséggel, hogy az egyik oldalán ugyanannak a mondatnak a helyes, míg a másikon a hibás változata található. A tanítókorpusz létrehozásához az online elérhető Open Subtitles[2] nevű angol-magyar párhuzamos korpuszának magyar oldali szövegét használtuk. A korpusz TV és mozi filmekre létrehozott feliratokból áll. Ennek megfelelően főleg rövidebb, informális mondatokat tartalmaz. A tanítókorpusz 1 millió szegmensből, a validációs korpusz 5000 mondatból és a tesztkorpusz 3000 mondatból áll. A korpusz az egyik legnagyobb szabadon hozzáférhető párhuzamos tanítóanyagnak számít, ellenben mérete elmarad az egynyelvű tanítóanyagokétól. Választásunk azért esett erre az adathalmazra, mert több párhuzamos kutatásunk során is használjuk, és néhány korpusztisztító lépést már előzetesen eszközöltünk rajta. Kivettük azokat a mondatokat, amelyek speciális karaktereket (pl. kínai, japán, cirill stb.) tartalmaztak, valamint a teszt halmaz mondatait kézzel kijavítottuk. Mérete elégséges a neurális hálózatok helyes betanítására, valamint a tanítási idő is viszonylag kezelhető marad (1-2 nap).

A tanítókorpuszunk forrásnyelvi oldalán található (feltételezett) helyes mondatokat automatikusan „rontottuk el” az 1. táblázatban található, felszíni alakból generálható hibákkal. Az így kapott mondatok alkotják a párhuzamos korpuszunk hibás oldalát. Kutatásunkban nem használtunk morfológiai és szintaktikai elemzőt, ezért az automatikus elrontó szkriptünk nem állít elő morfológiai és szintaktikai hibákat. Az automatikus hibageneráló szkriptünk a 3. táblázat arányai alapján ront el szavakat, az egyes hibatípusok ugyanis ebben az arányban fordultak elő a vizsgált korpuszban.

Miután lefuttattuk a szkriptünket az 1 millió mondat (6.181.241 token) szavaira, az alábbi hibaarányt kaptuk a mesterségesen előállított korpuszban:

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

– szavak 6,55%-a hibás (egy szó több hibát is tartalmazhat)
– mondatok 32,04%-a hibás

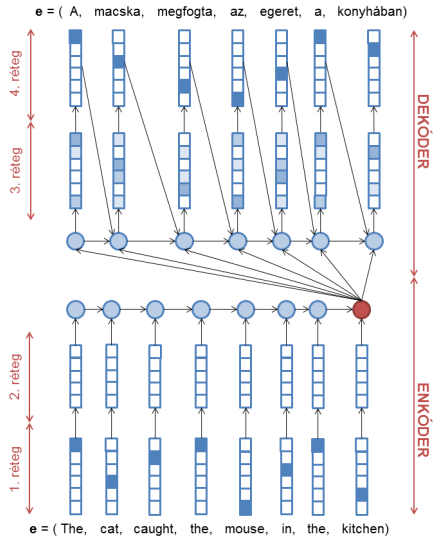
Hiba típusa	arány
Ékezetek hiánya	2,758%
Mondatkezdő kisbetű	9,43%
Mondatvégi írásjel hiánya	6,82%
Hosszú magánhangzó helyett rövid	6,84%
Rövid magánhangzó helyett hosszú	0,245%
Hosszú mássalhangzó helyett rövid	1,87%
Magánhangzó nyúlás l előtt (pl. *hátúl*)	1%
Vonatkozó névmások előtti vessző hiánya	22%
„hogy" előtti vessző hiánya	24%
„ly" betű helyett „j"	4,34%
„lesz" szó helyett „lessz"	9%
Extra pont	7,14%
Extra vessző	10,85%
Informális rövidítések	3%
Extra karakterek beszúrása	0,12%
Hiányzó karakter	10,85%
-ban, -ben helyett -ba, -be	1,42%

3. táblázat. Hibák aránya

4.2. A neurális gépi fordítórendszer

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 munka ellenére lényegében nem sikerült tovább javítani. Az áttörést (Bahdanau és mtsai, 2015) rendszere hozta el, ami egy figyelmi (attention) 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. Az 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.



1. ábra: Enkóder-dekóder architektúra vázlatos rajza

Munkánk során a Marian NMT (Junczys-Dowmunt és mtsai, 2018a) nevű keretrendszert használtuk, ami egy C++ nyelven íródott szabadon hozzáférhető programcsomag. Könnyen telepíthető, jól dokumentált, memória- és erőforrás-optimális implementációjának köszönhetően[3] az akadémiai felhasználók és fejlesztők által leggyakrabban használt eszköz (Barrault és mtsai, 2019).

4.3. A Sentence Piece tokenizáló

Az NMT rendszerek működése GPU processzorokon történik, melyek egyik szűk keresztmetszete a bennük található memória mérete. Ez határozza meg a létrehozható NMT rendszer szótárának a méretét. Egy szóalapú rendszer esetében az általánosságban 100K különálló szóban korlátozzák le a rendszert, így a további szavakat ismeretlenként kezeli.

(Sennrich és mtsai, 2015) ezt a problémát úgy oldották meg, hogy a szavak helyett úgynevezett subword (szótöredék) szintre csökkentették a legkisebb fordítási egységet. A BPE (Byte Pair Encoding) egy adattömörítő eljárás, ahol a leggyakoribb bájtpárokat egy olyan bájttal helyettesítjük, amely nem szerepel magában az 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

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

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.

(1) Sima szöveg: Petőfi Sándor egy nagyszerű költő.
SPM szöveg: Pető fi □ S ándor □egy □nagyszerű □költő .

A fenti példában látható az SPM (Sentence Piece modell) kimenete. A sima szöveg szavait gyakran előforduló karakter sorozatokra tördeli szét. Érdekes megfigyelni, hogy az eredeti mondat szóközeit is a szavakhoz csatolja és mint önálló karaktert (□) kezeli.

4.4. NMT tanítása

Az NMT tanításához a Marian neurális gépi fordítórendszert használtuk. A rendszer fontos jellemzője a SPM technológia. A rendszerünk tanításához az alábbi hyper-paramétereket használtuk:

- Sentence Piece: szótárméret: 16000; egy szótár a forrás- és egy a célnyelvi korpusznak; karakter lefedettség a teszt korpuszon: 100%
- Transzformer 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

A gépi fordítás forrásnyelve a hibageneráló szkriptünk által „elrontott" tanítókorpusz, míg a célnyelv az eredeti nem hibás tanítókorpusz.

A tanítás körülbelül 7 óra alatt végzett.

5. Eredmények

A 4. táblázat eredményein megfigyelhető, hogy a fedés szinte minden esetben alulmarad a pontossághoz képest, azaz a rendszer inkább "óvatos", mint alapos. Ezt az is okozhatja, hogy a tanítóanyagban a mondatpárok hibátlannak tekintett oldala valójában nem volt mindig hibátlan. A legjobb eredményt a tipikusan az internetes fórumokra jellemző hibákkal értük el (ly-j, lesz-lessz, rövidítések).

	Pontosság	Fedés	F-mérték	Esetek száma (res - gold)
Nagybetű	92,35%	90,41%	91,37%	327 - 334
Ékezetes szavak	88,44%	81,55%	84,85%	545 - 591
Mondatvégi írásjelek	82,42%	78,18%	80,25%	387 - 408
Mondatvégi pont	84,75%	86,06%	85,40%	328 - 323
Mondatvégi kérdőjel	75,67%	60,86%	67,46%	37 - 46
Mondatvégi felkiáltójel	61,11%	34,37%	44,00%	18 - 32
Vessző	93,47%	90,76%	92,10%	368 - 379
Rövid MGH	94,07%	79,37%	86,10%	135 - 160
Hosszú MSH	88%	68,75%	77,19%	50 - 64
Hosszú MGH	89,31%	70,05%	78,52%	131 - 167
Hosszú MGH l előtt	84,61%	70,96%	77,19%	26 - 31
ly-j	100%	92,85%	96,29%	13 - 14
lesz-lessz	100%	100%	100%	18 - 18
ban, ben	100%	60%	74,99%	3 - 5
rövidítések	100%	100%	100%	16 - 16

4. táblázat. Hibatípusok relatív eredményei

Ezek valószínűleg egyébként nem fordulnak elő az OpenSubtitles korpuszban, így valóban csak a hibás szöveg-generáló szkripttel kerülhettek a tanítóanyagba. A mondatvégi írásjelek pótlásában a felkiáltójel eltalálása bizonyult a legnehezebb feladatnak, ami nem meglepő, hiszen sok esetben az ember számára is nehéz eldönteni, hogy egy mondatot felkiáltónak szánt-e a szerzője. Ugyanez elmondható a kérdőjellel kapcsolatban is: az eldöntendő kérdések és a kijelentő mondatok között az írásjelen kívül nincs különbség. A vesszőhibákkal kapcsolatban meg kell említeni, hogy a rendszer csak a *hogy* és a vonatkozó névmások előtti vessző pótlására lett tanítva, ezeket viszonylag jó eredménnyel tudta teljesíteni. A vesszőhibák egyébként, amint a 3. táblázatban látható, az összes hiba csaknem felét teszik ki.

6. Összegzés

Kutatásunk célja egy olyan NMT alapú automatikus hibajavító eszköz létrehozása volt, amely képes a korpuszok sztenderdizálására. Ebben a tanulmányban a közösségi médiában gyakran előforduló gyakori hibákra fókuszáltunk. A korpuszméréseink szerint a legjellemzőbb hibatípus a vesszőhiány, rendszerünk ezt 90% körüli pontossággal és fedéssel tudja javítani. Hibajavítónk ezen kívül eredményes volt néhány tipikus, informális szövegekre jellemző hiba javításában, a kevésbé specifikus hibák esetén azonban nem volt annyira sikeres. Ennek oka lehet a tanítókorpusz "helyes" oldalának esetlegesen rossz minősége, illetve az is, hogy a generált hibák nem minden esetben eléggé „élethűek”.

A további feladataink között szerepel ezért a tanítókorpusz minőségének javítása, illetve a hibajavító kiterjesztése további, már elemzést igénylő hibatípusokra is. Továbbá a neurális gépi fordítás mellett szeretnénk kipróbálni az új

kontextuális szóbeágyazás-alapú modellek, mint a BERT jellegű modellek alkalmazását 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

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

Brockett, C., Dolan, W.B., Gamon, M.: Correcting esl errors using phrasal smt techniques. In: Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics. pp. 249–256. Association for Computational Linguistics, Stroudsburg, USA (2006)

Chollampatt, S., Tou Ng., H.: A multilayer convolutional encoder-decoder neural network for grammatical error correction. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (2018)

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)

Felice, M., Yuan, Z., Andersen, O., Yannakoudakis, H., Kochmar, E.: Grammatical error correction using hybrid systems and type filtering. In: Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task. pp. 15–24. Association for Computational Linguistics, Baltimore, Maryland (2014), http://www.aclweb.org/anthology/ W14-1702

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 2018a)

Junczys-Dowmunt, M., Grundkiewicz, R., Guha, S., Heafiled, K.: Approaching neural grammatical error correction as a low-resource machine translation task. 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. 595–606. Association for Computational Linguistics, New Orleans, Louisiana (2018b)

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

Miháltz, M., Váradi, T., Csertő, I., Fülöp, É., Pólya, T., Kővágó, P.: Beyond sentiment: Social psychological analysis of political facebook comments in hungary. In: Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA 2015). pp. 127–133. Association for Computational Linguistics, Lisboa, Portugal (2015)

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

Susanto, R.H., Phandi, P., Tou Ng, H.: System combination for grammatical error correction. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 951–962. Association for Computational Linguistics (2014), https:// doi.org/10.3115/v1/D14-1102

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

Zhao, W., Wang, L., Shen, K., Jia, R., Liu, J.: Improving grammatical error correction via pre-training a copy-augmented architecture with unlabeled data. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (2019)