Televíziós feliratok írásjeleinek visszaállítása rekurrens neurális hálózatokkal

Tündik Máté Ákos, Tarján Balázs, Szaszák György

Budapesti Műszaki és Gazdaságtudományi Egyetem,
Távközlési és Médiainformatikai Tanszék
e-mail:{tundik,tarjanb,szaszak}@tmit.bme.hu

Kivonat Az automatikus beszédfelismerő rendszerek (ASR) kimenete általában nem tartalmaz írásjeleket, pedig ezek döntően befolyásolják egy szöveg értelmét és értelmezhetőségét. Az írásjel-visszaállítás problémájára a közelmúltban eredményesen alkalmaztak rekurrens neurális hálókat (RNN). A valós idejű ASR kimenetekben (pl. televíziós feliratok) történő írásjel-visszaállítás azonban nagy kihívást jelent, mivel az alacsony késleltetés érdekében időben előre csak kis mértékben tekinthet a rendszer, illetve ASR-hibákkal is számolni kell, különösen informális vagy spontán beszéd esetén. Cikkünkben Maximum Entrópia (MaxEnt) és RNN írásjelező modelleket hasonlítunk össze magyar nyelvű, televíziós feliratok valós idejű és off-line feldolgozása során. A várakozásoknak megfelelően az RNN-alapú modellek teljesítménye felülmúlta a MaxEnt baseline rendszerét, illetve az ASR-hibák is jelentősen befolyásolták az írásjelek minőségét. Ezzel szemben az előretekintés mértékének korlátozása csak kisebb pontosságcsökkenést eredményezett. A hibák okainak feltárása érdekében műsortípusonként is kiértékeltük az írásjelező modellünket. Végezetül a modellünket egy angol nyelvű adatbázis segítségével más nemzetközi megoldásokkal is összehasonlítottuk.
Kulcsszavak: írásjel-visszaállítás, rekurrens neurális hálózatok, LSTM, maximum entrópia, ASR, feliratozás

1. Bevezetés

Az írásjelek fontossága nem csak abban rejlik, hogy az emberek számára olvashatóbbá teszik a szöveget, hanem a szövegfeldolgozó rendszerek is építenek rájuk a szintaktikai elemzés során [1,2]. Azaz hiányuk nem csak az emberi, de a gépi feldolgozást is megnehezíti. Az ASR-alapú diktáló rendszerek ezt a problémát általában úgy kerülik meg, hogy elvárják az írásjelek bemondását is, azonban könnyű belátni, hogy a beszélő ilyen fokú együttműködése nem várható el egy televíziós feliratozó-rendszer esetén.

Az írásjelek visszaállítása az ASR rendszerek kimenetén ismert probléma a beszédtechnológiában. Két alapvető megközelítést lehet megkülönböztetni: a prozódia alapút és a szövegalapút, bár gyakran kombinálva is használják őket. Általánosságban a prozódia alapú megközelítések számításigénye alacsonyabb és robusztusabbak az ASR hibákra, viszont a tanításukhoz szükséges címkézett,

akusztikai adatbázisok nehezebben hozzáférhetőek. Ezzel szemben a szövegalapú megközelítések többnyire pontosabb írásjelezést tesznek lehetővé, hála a nagy mennyiségű, aránylag könnyen hozzáférhető szövegkorpuszoknak, ugyanakkor érzékenyebbek az ASR-hibákra és több számítást igényelnek.

A korai írásjel-visszaállító megoldások az ASR-rendszerek n-gram nyelvi modelljébe épített rejtett eseményként tartalmazták az írásjeleket [3]. Ezek a modellek azonban hatalmas korpuszok bevonását igénylik, hogy csökkentsék az adatelégtelenségből fakadó problémát. Később megjelentek bonyolultabb szekvenciamodellezési megközelítések is: a transzdúcer-elvű keretrendszerek esetén egy írásjelezetlen szöveg kerül a bemenetre, melybe conditional random field (CRF), rejtett Markov-modellek (HMM), vagy maximum entrópia (MaxEnt) modellek segítségével helyezik el az írásjeleket [4,5]. A MaxEnt-alapú írásjelezési modellek lehetővé teszik a szöveges és prozódiai jellemzők hatékony kombinációját [6]. Egy átfogó tanulmányban [7] különböző jellemzőket hasonlítottak össze az írásjelezésre gyakorolt hatásuk tekintetében, MaxEnt-modellt használva. A legerősebb prediktív szöveges jellemzők közé a szóalakok és a szófaji (POS) címkék tartoztak, míg a legjobb prozódiai jellemző a szavak közötti szünetek időtartama volt. A Cho és társai által javasolt megoldás az írásjelezésre mint nyelven belüli fordítási feladatra tekint, írásjelezetlenből írásjeleket tartalmazó szószekvenciákat modellezve, mellyel jelentősen csökkentette az időkésleltetést [8].

A közelmúltban megjelentek a mély neurális hálózatokon alapuló megoldások: adott egy relatíve hosszú szó-kontextus, melyben a szavaknak egy beágyazási (Embedding) réteg révén átalakított alacsony dimenziójú vektor-reprezentációját vesszük, majd ezeket átvezetve kétirányú, rekurrens neurális hálózaton (RNN) [9] az írásjelek predikcióját megkapjuk mindegyik szóhoz. Az RNN-eket sokféle szekvenciacímkézési feladathoz alkalmazták már sikeresen, mivel képesek nagy kontextusok modellezésére, valamint a szavakból olyan jellemzőket tudnak kinyerni, amivel az adatelégtelenség problémája áthidalható. Tilk és Alumäe kétirányú RNN modelljében GRU [10] cellákat alkalmazva, ún. attention (figyelmi) mechanizmussal kiegészítve - mely segít még jobban fókuszálni az írásjelek szókörnyezetére közvetlen összeköttetések révén -, felülmúlta a korábbi észt és angol IWSLT eredményeket [11]. Egy nemrégiben készült tanulmányban [12] a kapitalizáció és az írásjelek helyreállítását egymással korreláló, többszörös szekvenciacímkézési feladatként kezelték, kétirányú RNN modellekkel. A [13]-ban a szerzők egy olyan prozódiai jellemzőket használó RNN-alapú központozási megoldást javasoltak, melynek hátterében a megnyilatkozások fonológiai frázisokra történő bontása és az elhelyezendő írásjelek közötti kapcsolat megállapítása állt.

A televíziós társaságok az ASR technológiát széles körben használják a feliratok készítéséhez, különösen az élő programokhoz, amelyeknek közel valós idejű feldolgozása kis késleltetést igényel [14]. Cikkünkben ehhez a feladathoz illeszkedő, alacsony késleltetésű, szövegalapú automatikus írásjelező modellek teljesítményét vizsgáljuk meg. Bemutatunk egy egyszerű felépítésű, RNN-alapú írásjel-visszaállító modellt, kétirányú LSTM cellákkal és a szóbeágyazást elvégző Embedding réteggel, majd összehasonlítjuk teljesítményét egy MaxEnt-alapú baseline megoldással. Különös figyelmet fordítunk az alacsony késleltetésű megoldá-

sokra. Mindkét rendszer teljesítményét automatikus és manuális átiratokon is kiértékeljük, megkülönböztetve az on-line és off-line működtetéshez alkalmas beállításokat. Magyar nyelvű televíziós műsorszórásból származó anyagokon kívül, az IWSLT angol nyelvű adathalmazán [15] elvégzett kísérleteink eredményeit is bemutatjuk, így összehasonlítva modellünk teljesítményét a legmodernebb rendszerekkel. A [13] által ismertetett tisztán prozódia alapú megközelítésen kívül nem ismerünk semmilyen előzetes munkát a magyar nyelvű beszédátiratokon történő írásjelek visszaállítására.

Cikkünk az alábbi struktúra szerint épül fel: először bemutatjuk az általunk használt adatbázisokat, illetve a kísérletekhez használt modelleket. Ezt követően áttérünk a magyar és angol nyelvű írásjelezési kísérletek ismertetésére és diszkussziójára. Végezetül a tanulságok levonása után néhány jövőbeni lehetséges kutatási irányt vázolunk fel.

2. Adatbázisok

2.1. Magyar nyelvű adatbázis

Az írásjel-visszaállítási kísérleteinkhez használt magyar nyelvű adatbázist a Médiaszolgáltatás-támogató és Vagyonkezelői Alap (MTVA) bocsátotta rendelkezésünkre. Az adatbázis különböző TV-műfajokhoz tartozó, kézzel készített feliratokat tartalmaz, amelyek lehetővé teszik számunkra, hogy különböző társalgási formákat tartalmazó műsorokon, például időjárás-előrejelzéseken, hírműsorokon, hírháttér-beszélgetéseken, magazinokon, sporthíreken és sportmagazinokon értékeljük ki az írásjelező megoldásunkat. A leggyakoribb és egyben a szöveg érthetősége szempontjából legfontosabb írásjeleket állítjuk vissza: a vesszőt, a mondatvégi pontot, a kérdőjelet és a felkiáltójelet. A kettőspontokat és a pontosvesszőket vesszővel helyettesítettük, minden más írásjeltől eltekintettünk. A tanítóanyag 20% -át validációs célból leválasztottuk, valamint egy külön, reprezentatív tesztkészletet használunk, amelynek nincs átfedése sem a tanító, sem a validációs halmazzal. Az adatokhoz kapcsolódóan további statisztikák a 1. táblázatban olvashatók.

1. táblázat. A magyar nyelvű adatbázis statisztikái

Műfajok	Tanító és validációs halmaz					Teszthalmaz					
	#Szavak	#Vessző	#Pont	#Kérdő	#Felki	#Szavak	#Vessző	#Pont	#Kérdő	#Felki	WER
Időjárás	478K	40K	31,5K	30	730	2,4K	250	200	0	20	6,8
Híradó	3493K	279K	223K	3,5K	4,6K	17K	1,5K	1K	20	50	10,1
Sporthírek	671K	55K	39,5K	280	2K	6K	500	400	2	30	21,4
Hírháttér	4161K	533K	225K	26,5K	4K	46,8K	6,3K	2,6K	250	130	24,7
Sportmagazin	-	-	-	-	-	22,7K	2K	1,4K	100	50	30,3
Magazin	4909K	732K	376K	72K	36K	10,4K	1,5K	700	150	70	38,7
Vegyes	1526K	187K	102K	11K	11,4K	30,7	4K	1,7K	280	150	-
Összesen	**15238K**	**1826K**	**997K**	**113K**	**58,8K**	**136K**	**16K**	**8K**	**800**	**500**	**24,2**

A kézzel készített feliratok mellett automatikus leiratokat is használtunk kísérleteinkben. Ezek egy, a televíziós műsorok élő feliratozásához optimalizált

ASR rendszerrel készültek [16]. Az ASR nyelvi modelljének tanítóhalmaza megegyezett az írásjelezési modellével, emellett hozzávetőlegesen 500 órányi beszédet használtunk fel az akusztikai modell tanításához. Az automatikus átiratok átlagos szóhibaaránya (WER) körülbelül 24% volt, azonban a műfajtól függően nagy változatosságot mutatott (lásd 1. táblázat). A "Vegyes" kategóriában nem állt rendelkezésre hanganyag az adatbázisban.

2.2. Angol nyelvű adatbázis

Az IWSLT adathalmaz angol nyelvű TED előadások átirataiból áll, és a közelmúltban az angol írásjel-visszaállítási kísérletek egyik népszerű feladatává vált [11,12,15,17]. Kísérleteink során ugyanazt a tanító, validációs és teszthalmazt használtuk, mint a fenti tanulmányok. Ezek rendre 2,1M, 296K és 13K szót tartalmaznak. Az adatbázis csak vesszőt, pontot és kérdőjelet tartalmaz, így csak ezek visszaállítására nyújt lehetőséget.

3. Írásjel-visszaállító módszerek

3.1. Maximum Entrópia (MaxEnt) Modell

A maximum entrópia (MaxEnt) modellt eredetileg Ratnaparkhi javasolta a szófaji címkézésre [18]. Minden mondatot tokenek (szavak) szekvenciájaként írunk le. Minden egyes tokenhez egy sor egyedi jellemző rendelhető, melyek segítenek a kimeneti címkék (jelen esetben az írásjelek) meghatározásában. Lévén, hogy ez egy felügyelt tanítási technika, a kimeneti címkék a token-sorozathoz hozzá vannak rendelve. A jellemzők meghatározásához a MaxEnt modell meghatároz a rendelkezésre álló címkék és az aktuális kontextus között egy együttes eloszlást, amely egy rádiusz paraméterrel szabályozható.

Maxent írásjelezési modellünkben csak a kisbetűs szóalakokat használtuk fel jellemzőként mind magyar, mind angol nyelven. Ehhez a Huntag nyílt forráskódú, nyelvfüggetlen, Markov modell-alapú, MaxEnt szekvenciacímkéző programot használtuk fel [19].

Mint említettük, a MaxEnt modell rádiusz paramétere határozza meg a figyelembe vett kontextus méretét. Alapértelmezés szerint a múltbeli és a jövőbeli kontextust egyenlő mértékben veszi figyelembe. Erre a beállításra *off-line módként* fogunk hivatkozni. A jövőbeli kontextus figyelembevétele azonban a késleltetést növeli; ezt limitálva, az *on-line módú* modellek teljesítményét is kiértékeltük. Cikkünkben az alábbi, kerek zárójeles jelöléssel utalunk a múltbeli és jövőbeli kontextus méretére; (5,1) azt jelenti, hogy valójában az adott modellben maximum 5 múltbeli és 1 jövőbeli tokent veszünk figyelembe az adott szóhoz tartozó írásjel predikciójához.

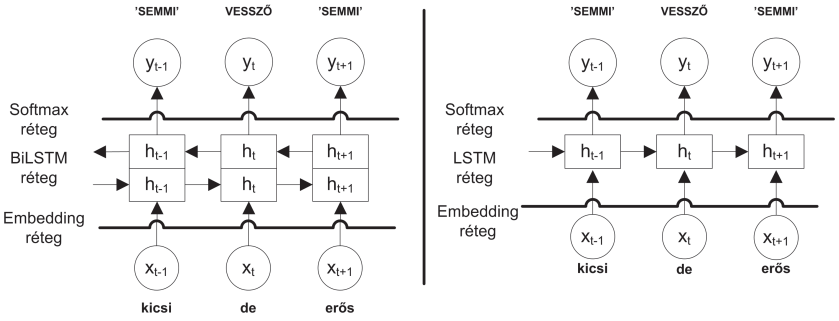
3.2. Rekurrens neurális háló

A tanító, a validációs és a teszthalmazt rövid, fix hosszúságú szekvenciákra osztjuk fel, köztes átfedések nélkül. A különböző szóalakok számát limitáljuk, a tanítóhalmaz k-leggyakoribb szavából szótárt képezve, a kieső szavakat pedig egy

közös *Ismeretlen* címkével látjuk el. A modellhez saját szóbeágyazási mátrixot képzünk az előre tanított beágyazási modell és a szótárban szereplő szavak segítségével.

Kísérleteinkben egy egyirányú és egy kétirányú RNN modell teljesítményét vizsgáljuk meg. A modellben az aktuális szót megelőző időpillanatra jósoljuk az írásjelet. A kísérletekhez használt RNN-architektúrákat az 1. ábrán mutatjuk be.

Az RNN-modellek (WE-LSTM és WE-BiLSTM, a Word Embedding (szóbeágyazás) rövidítéséből) a következőképpen épülnek fel: a szóbeágyazási mátrix alapján a modellnek átadott szószekvenciák a szóbeágyazási térbe (x_t reprezentálja az x szóhoz tartozó n-dimenziós szóbeágyazási vektort t időpillanatban). Ezek a reprezentációk a következő, rejtett rétegbe kerülnek, amely LSTM vagy BiLSTM rejtett cellákból áll, melyek a x_t kontextus rögzítéséért, az információ kinyeréséért felelősek. A kimenetet egy $softmax$ aktivációs függvény használata után kapjuk meg, mely az y_t kimeneti címkék eloszlását a jelenlegi szó x_t előtti időpillanatra (slot-ra) adja meg. Láthatjuk, hogy ez egy egyszerű felépítésű modell, ezáltal alacsony késleltetést, valós idejű működést tesz lehetővé.



1. ábra: A WE-BiLSTM (bal oldalon) és a WE-LSTM (jobb oldalon) RNN modell szerkezete

A magyar írásjelező modelleket a tanítókorpusz 100 000 leggyakoribb szavával tanítottuk, a kimaradt szavakhoz egy közös "*Ismeretlen*" szimbólumot rendeltünk. Az RNN-alapú írásjel-helyreállítási modellekhez egy 600 dimenziós, előre tanított magyar nyelvű szóbeágyazási modellt használtunk [20]. Angol nyelvű RNN-modelljeinkben egy 100 dimenziós, népszerű szóbeágyazási modellt használtunk, a "GloVe"-ot. A tanítás során az RNN modell súlyait a kategorikus keresztentrópia költségfüggvény alapján módosítjuk, valamint minden egyes epoch-ban frissítjük a szóbeágyazásokat is.

Szisztematikus, kimerítő keresés (grid search) alapú optimizációt hajtottunk végre az RNN-ek hiperparaméterein, a validációs halmaz elemeit értékelve. A szekvenciák hosszát, a szótár méretét, a rejtett állapotok számát, a mini-batch méretét, és az optimizáló típusát változtattuk. Korai leállítást (early stopping, *Patience*) is használunk a túltanítás elkerülése érdekében. A 2. táblázat összefoglalja a magyar és az angol WE-BiLSTM és WE-LSTM modellekben használt

hiperparaméterek végső értékeit, beleértve azokat is, amelyeket a [11] cikkből vettünk át, a minél jobb összehasonlíthatóság érdekében.

2. táblázat. A WE-BiLSTM és WE-LSTM modellek hiperparaméterei

Nyelv	Model	Szekv. Hossza (#szavak)	Szótár Mérete (#szavak)	Szóbeágyazás Dim.	#Rejtett állapotok	Batch mérete	Optimalizáló	*Patience*
HUN	WE-BiLSTM	200	100 000	600	512	128	RMSProp	3
HUN	WE-LSTM				256			2
EN	WE-BiLSTM	200	27 244	100	256			2
EN	WE-LSTM	250	([11] révén)	([11] révén)				

A MaxEnt-hez hasonlóan az RNN modellnél is megkülönböztetjük az alacsony késleltetéshez adaptált on-line módot és a robusztus off-line módot, attól függően, hogy figyelembe vesszük-e a jövőbeli kontextust. Az RNN-alapú írásjelező rendszerek implementálásához a Keras keretrendszert [21] használtuk, a tanítást GPU-n végeztük el. Az RNN-modellek forráskódja nyilvánosan elérhető [1].

Vizsgálataink előkészítő fázisában a szóalakok mellett más szöveges jellemzők használatát is fontolóra vettük (lemmák, szófaji címkék (amit [22] is javasolt) és morfológiai elemzésből származó címkék), ezekkel azonban nem értünk el jobb eredményeket, így további kísérleteink során csak a szóalakokra támaszkodtunk.

4. Kísérleti eredmények

A következő fejezetben bemutatjuk a magyar és angol nyelvű írásjelezési kísérleteink eredményeit. A kiértékeléshez standard információ-visszakeresési mutatókat használtunk: Pontosság (Pr), Felidézés (Rc) és F1-érték (F1). Ezenkívül megadtuk a Slot Error Rate (SER) [23] értéket is, amely egy metrikában egyszerre tükrözi az írásjel-visszaállításhoz kapcsolódó hibák minden lehetséges típusát - beszúrásokat (Ins), helyettesítéseket (Sub) és törléseket (Del):

$$SER = \frac{C(Ins) + C(Subs) + C(Del)}{C(slotok\_szama)}, \tag{1}$$

ahol $C(.)$ a számláló operátor, a slot-ok pedig a szavak utáni helyek a szövegben, ahova írásjelet helyezhetnek a modellek.

4.1. Magyar eredmények

Először összehasonlítjuk a MaxEnt szekvenciacímkéző baseline rendszer teljesítményét (lásd 3.1 alfejezet) az RNN-alapú írásjelező rendszerével (lásd 3.2. alfejezet), a magyar adathalmazon, két különböző konfigurációban. Az *on-line*

[1] https://github.com/tundik/HuPP

módban az aktuális szó előtti időpillanat írásjel predikciója jelenik meg, ez még épp akkora késleltetést eredményez, amely még valós idejű alkalmazáshoz elfogadható. A legjobb írásjelezési eredményének elérése érdekében az ún. *off-line módban* a jövőbeli kontextust is felhasználjuk.

Az eredményeket két táblázatban mutatjuk be: a kézi úton előállított feliratokon végzett írásjel-visszaállítás eredményei a 3. táblázatban, az automatikus (ASR) feliratokon végzett visszaállítás eredményei a 4. táblában találhatók. A MaxEnt modellek (i, j) jelölésében i a múltbeli, míg j a jövőbeli kontextust kontrolláló rádiusz paraméter értékét jelöli. Amint láthatjuk, a vesszők predikciója minden módszer és konfiguráció esetében kiemelkedik a többi írásjelé közül. Erre az lehet a legvalószínűbb magyarázat, hogy magyar nyelven a vesszők nagy aránya megbízhatóan becsülhető a slot-ot követő szó alapján (pl. 'hogy'). Ezzel ellentétben a mondatvégi pontok megbízható visszaállításához elengedhetetlennek tűnik az akusztikus jellemzők figyelembevétele is [13].

3. táblázat. Írásjel-visszaállítási eredmények magyar nyelvű, manuális feliratokon

Referencia-átirat	Model	Vessző			Pont			Kérdőjel			Felkiáltójel			SER
		Pr	Rc	F1	Pr	Rc	F1	Pr	Rc	F1	Pr	Rc	F1	
Off-line mód	MaxEnt-(19,19)	72,5	59,6	65,5	52,1	40,0	45,2	**55,7**	21,8	31,3	31,1	31,5	31,3	63,5
	WE-BiLSTM	**72,9**	**71,2**	**72,0**	**59,1**	**56,1**	**57,6**	52,4	**38,7**	**44,5**	**51,3**	**36,1**	**42,4**	**50,1**
On-line mód	MaxEnt-(25,1)	71,8	58,1	64,2	47,5	35,7	40,8	50,4	16,2	24,5	29,3	**33,3**	31,2	66,9
	WE-LSTM	**72,7**	**69,5**	**71,1**	**56,2**	**48,3**	**52,0**	**60,4**	**31,1**	**41,1**	**61,1**	29,4	**39,7**	**53,6**

Ahogy a 3. táblázatban látható, a magyar nyelvű, manuálisan készült feliratok halmazán közel 20% relatív hibacsökkenés (SER) érhető el az RNN-alapú írásjelező modellel a baseline MaxEnt rendszerhez képest. A WE-BiLSTM és a WE-LSTM modellek különösen jól teljesítenek a pontok, kérdőjelek és felkiáltójelek helyreállításában, mivel a nagy kontextusokban rejlő információt sokkal hatékonyabban képesek kihasználni, mint a MaxEnt címkézők. A jövőbeni kontextus korlátozása az on-line konfigurációban sokkal kisebb mértékben befolyásolta az írásjelezés eredményességét, mint amire számítottunk. A jövőbeni szószekvencia-elemekből kinyert információk főként akkor hasznosak, ha a feladat megköveteli a felidézés maximalizálását, egyébként a pontosság tekintetében a WE-LSTM is alkalmas az írásjelek helyreállítására.

4. táblázat. Írásjel-visszaállítási eredmények magyar nyelvű, gépi feliratokon

ASR átirat	Model	Vessző			Pont			Kérdőjel			Felkiáltójel			SER
		Pr	Rc	F1	Pr	Rc	F1	Pr	Rc	F1	Pr	Rc	F1	
Off-line mód	MaxEnt-(19,19)	**64,5**	55,8	59,9	41,1	31,2	35,6	**41,2**	8,8	14,4	48,8	17,1	25,4	79,2
	WE-BiLSTM	63,9	**67,7**	**65,7**	**50,5**	**49,0**	**49,8**	37,7	**24,1**	**29,4**	**60,9**	**24,0**	**34,4**	**70,1**
On-line mód	MaxEnt-(25,1)	**64,3**	54,9	59,2	38,9	29,4	33,5	36,0	7,1	11,9	47,1	20,6	28,6	81,3
	WE-LSTM	63,8	**65,1**	**64,4**	**47,8**	**42,0**	**44,7**	**48,5**	**20,5**	**28,9**	**61,8**	**21,7**	**32,1**	**73,1**

Amint azt a bevezetőben már vázoltuk, mind az ASR-hibák, mind a jövőbeli kontextus korlátozása azon tényezők közé tartoznak, amelyek nehezítik az írásjelek hatékony elhelyezését az élő TV-műsorok feliratainban. Az eredményeink azt mutatják, hogy a két faktor közül a jövőbeni kontextus kevésbé fontos az írásjelek robusztus helyreállításához, ami ellentmond a várakozásainknak. Ezzel szemben az ASR-hibák jobban összefüggnek az írásjelek hibáival: a manuálisról a gépi úton készült feliratokra való áttérés során a SER értéke 15-20%-kal növekedett (lásd a 4. táblázatot). Habár ezzel párhuzamosan a MaxEnt és RNN modellek közötti különbség is csökkent az ASR feliratok feldolgozásakor, de az RNN még mindig jelentősen felülmúlta a baseline rendszert.

4.2. Műfaji analízis

A magyar nyelvű adatbázis teszthalmazának feliratait 6 műfaji kategóriába lehet osztani (lásd 1. táblázat). Ebben az alfejezetben az egyes műfajokon mérhető írásjel-helyreállítási pontosságokat hasonlítjuk össze, feltételezve, hogy az informálisabb, spontánabb műfajok írásjelezése nehezebb feladat, hiszen ezek esetében magasabb az ASR-hibák aránya (WER). Az egyes műfajokra vonatkozó írásjelek egy részét nem értékeltük (lásd az "N/A" feliratokat a 1. táblázatban), ha a felidézést vagy pontosságot nem lehetett meghatározni a tévesztési mátrix alapján.

Mivel az RNN-alapú rendszer minden műfajt tekintve meghaladta a MaxEnt rendszer teljesítményét, a könnyebb olvashatóság érdekében úgy döntöttünk, hogy csak a WE-BiLSTM és WE-LSTM rendszerek eredményeit adjuk meg a 5. és 6. táblázatban.

5. táblázat. Műfajonkénti írásjelezési eredmények magyar nyelvű referencia átiratokon

Referencia átirat	Műfaj	Vessző			Pont			Kérdőjel			Felkiáltójel			SER
		Pr	Rc	F1	Pr	Rc	F1	Pr	Rc	F1	Pr	Rc	F1	
RNN Off-line mód	Időjárás	61,2	54,3	57,5	46,7	46,7	46,7	N/A	N/A	N/A	**90,0**	45,0	60,0	69,3
	Híradó	**89,9**	**84,4**	**87,1**	**84,3**	**90,7**	**87,3**	**91,7**	**50,0**	**64,7**	83,9	**56,5**	**67,5**	**20,0**
	Sporthír	68,3	60,6	64,2	49,4	51,4	50,4	N/A	N/A	N/A	75,0	30,0	42,9	67,0
	Hírháttér	80,4	74,5	77,3	63,9	64,9	64,4	63,0	46,4	53,5	88,9	18,5	30,6	38,7
	Sportmagazin	61,2	61,1	61,1	43,9	49,3	46,5	55,2	37,5	44,7	38,5	9,4	15,2	73,1
	Magazin	67,6	67,6	67,6	45,1	46,3	45,7	50,5	29,7	37,5	50,0	5,6	10,1	58,6
RNN On-line mód	Időjárás	60,2	57,5	58,8	45,7	37,9	41,4	N/A	N/A	N/A	87,5	35,0	50,0	70,6
	Híradó	**88,4**	**83,1**	**85,7**	**86,6**	**81,3**	**83,9**	**75,0**	40,9	**52,9**	**100,0**	**67,4**	**80,5**	**24,1**
	Sporthírek	68,7	57,2	62,4	42,4	37,5	39,8	N/A	N/A	N/A	90,0	60,0	72,0	74,2
	Hírháttér	80,1	74,0	76,9	66,7	54,8	60,1	63,0	**45,6**	**52,9**	77,6	29,2	42,5	40,8
	Sportmagazin	60,8	59,7	60,3	42,3	34,8	38,2	53,3	38,3	44,5	20,0	7,5	11,0	77,3
	Magazin	67,6	65,1	66,3	43,5	32,8	37,4	57,3	27,2	36,9	36,4	11,3	17,2	61,5

Ha összevetjük az eredményeket a 1. táblázat statisztikáival, láthatjuk, hogy az írásjelek helyreállítása a híradó, hírháttér és magazin műfajok esetén sikerült a legjobban. Ez nem meglepő, hiszen ezekre a műfajokra rendelkezünk a

legtöbb tanítóanyaggal. Azonban a viszonylag nagy különbség a három jól modellezett műfaj között azt sugallja, hogy ezen kívül más tényezők is befolyásolják az eredményeket. Véleményünk szerint ezek közül kiemelkedik az adott szöveg tervezettsége. A nyelvi modellezéshez hasonlóan, minél formálisabb a beszédstílus a felvételeken, annál hatékonyabb az írásjelek elhelyezése is, pl. a híradók esetében. Nyilvánvaló, hogy a társalgási (pl. hírháttér) és az informális (magazin) beszédstílusok sajátosságai (melyekben a megakadásjelenségek és az alapvető grammatikai szabályokkal nonkonform kifejezésmód aránya magasabb) még inkább megnehezítik a predikciót, és a formálisabb stílusokhoz képest több írásjelezési hibát vezetnek be a rendszerbe.

6. táblázat. Műfajonkénti írásjelezési eredmények magyar ASR átiratokon

ASR átirat	**Műfaj**	**Vessző**			**Pont**			**Kérdőjel**			**Felkiáltójel**			**SER**	**WER**
		Pr	Rc	F1	Pr	Rc	F1	Pr	Rc	F1	Pr	Rc	F1		
RNN Off-line mód	Időjárás	57,8	54,3	55,9	48,1	43,8	45,8	N/A	N/A	N/A	**100,0**	**60,0**	**75,0**	74,3	6,8
	Híradó	65,7	**82,2**	**73,0**	**63,5**	**82,7**	**71,8**	**52,9**	**40,9**	**46,2**	75,0	45,7	56,8	**57,4**	10,1
	Sporthírek	53,3	55,9	54,6	37,9	41,1	39,5	N/A	N/A	N/A	88,2	53,6	66,7	93,1	21,4
	Hírháttér	**70,4**	68,8	69,6	55,6	49,4	52,3	44,1	28,4	34,5	70,8	13,1	22,1	59,6	24,7
	Sportmagazin	52,8	58,0	55,3	37,5	41,1	39,2	42,9	18,8	26,1	N/A	N/A	N/A	93,2	30,3
	Magazin	59,6	59,9	59,7	34,6	29,1	31,9	13,9	19,4	16,2	16,7	1,4	2,6	82,3	38,7
RNN On-line mód	Időjárás	61,4	57,9	59,6	43,1	39,1	41,0	N/A	N/A	N/A	**88,9**	40,0	55,2	73,2	6,8
	Híradó	64,8	**80,6**	**71,8**	**62,4**	**73,5**	**67,5**	40,0	9,1	14,8	75,0	45,7	56,8	**61,8**	10,1
	Sporthírek	52,8	54,0	53,4	35,4	34,6	35,0	N/A	N/A	N/A	83,3	**53,6**	**65,2**	96,7	21,4
	Hírháttér	**70,2**	67,1	68,6	53,5	40,9	46,3	**46,4**	**25,6**	**33,0**	69,2	13,8	23,1	62,4	24,7
	Sportmagazin	53,1	55,4	54,2	35,6	30,9	33,1	41,4	22,7	29,3	16,7	5,7	8,5	94,0	30,3
	Magazin	58,5	59,9	59,2	36,0	22,3	27,6	46,3	12,0	19,1	42,9	4,2	7,7	83,2	38,7

Az időjárás-előrejelzés és a sportprogramok esetén a viszonylag magas SER érték jelzi annak fontosságát, hogy a tanítóhalmaznak elegendő mennyiségű témaspecifikus adatot kell tartalmaznia. Az ilyen alultanított témakörök esetén további tanítóanyagok bevonásával vagy adaptációs technikák alkalmazásával lehetne jobb eredményeket elérni.

Összehasonlítva a manuális és gépi feliratok írásjelezési hibáit, néhány érdekes következtetést vonhatunk le. A jól modellezett műfajok esetében a SER növekedése korrelál az ASR átirat szóhibaarányával (WER). Azonban a többi műfaj (időjárás-jelentés, sporthírek, sportmagazin) a SER és a WER között ilyen összefüggés nem fedezhető fel. A sporthírek viszonylag gyenge eredményeire külön kitérnénk. Míg az ASR átirata mérsékelt szóhibaarányt (24,7%) mutat, az írásjelezéshez kapcsolódó SER érték majdnem másfélszeresére nőtt (67% -ról 93-97%-ra). Feltételezzük, hogy ez a jelenség a sportprogramokban szereplő névelemek (Named Entity) megnövekedett mennyiségével függ össze, tekintve, hogy a legmagasabb OOV-arány (10%) itt figyelhető meg a teszthalmazban szereplő 6 műfaj közül.

4.3. Angol nyelvű eredmények

Ebben az alfejezetben néhány, a közelmúltban publikált írásjel-visszaállító megoldással hasonlítjuk össze a modellünket. Ebből a célból az IWSLT2011 adatbázist használjuk, amely TED előadások feliratait tartalmazza, és az angol nyelvű írásjelezési kísérleteknél egyfajta benchmark-nak tekinthető. A precízebb összehasonlíthatóság érdekében az alapértelmezett tanító, validációs és teszthalmazt használtuk, azonban a hiperparamétereket erre a feladatra újra optimizáltuk a validációs halmazon. Kiemeljük, hogy az IWSLT adatbázis nem tartalmaz mintákat felkiáltójelekhez.

Az angol írásjelezési eredményeket a 7. és 8. táblázatokban mutatjuk be. Amint látható, on-line módban a bemutatott egyirányú RNN modellünk (WE-LSTM) jelentősen felülmúlta a [11]-ben megjelenő ún. T-LSTM konfigurációt, amely az eddigi legjobb on-line eredményeket szolgáltatta ezen az adathalmazon legjobb tudomásunk szerint. Ha nem használunk előre tanított szóbeágyazási modellt (noWE-LSTM), eredményünk nagyon közel áll a T-LSTM konfigurációhoz.

7. táblázat. Írásjelezési eredmények angol nyelvű referencia átiratokon

Referencia átirat	Model	Vessző			Pont			Kérdőjel			SER
		Pr	Rc	F1	Pr	Rc	F1	Pr	Rc	F1	
Off-line mód	MaxEnt-(6,6)	45,6	26,7	33,7	59,4	57,0	58,2	52,4	23,9	32,8	77,2
	WE-BiLSTM	55,5	45,1	49,8	65,9	**75,1**	70,2	57,1	52,2	54,5	59,8
	T-BRNN-pre [11]	**65,5**	47,1	54,8	73,3	72,5	72,9	**70,7**	**63,0**	**66,7**	**49,7**
	Corr-BiRNN [12]	60,9	**52,4**	**56,4**	**75,3**	70,8	**73,0**	70,7	56,9	63,0	50,8
On-line mód	MaxEnt-(10,1)	44,9	23,7	31,0	53,4	50,1	51,7	50,0	21,7	30,8	83,2
	noWE-LSTM	47,3	**42,7**	44,9	60,9	50,4	55,2	**68,2**	32,6	44,1	76,4
	WE-LSTM	**56,3**	40,3	**47,0**	**61,2**	**60,5**	**60,8**	55,5	**43,5**	48,8	**68,1**
	T-LSTM [17]	49,6	41,1	45,1	60,2	53,4	56,6	57,1	43,5	**49,4**	74,0

8. táblázat. Írásjelezési eredmények angol nyelvű ASR átiratokon

ASR Átirat	Model	Vessző			Pont			Kérdőjel			SER
		Pr	Rc	F1	Pr	Rc	F1	Pr	Rc	F1	
Off-line mód	MaxEnt-(6,6)	40,6	23,9	30,1	56,2	53,5	54,8	31,6	17,1	22,2	84,0
	WE-BiLSTM	46,8	39,6	42,9	60,7	70,3	65,1	44,4	45,7	45,0	72,5
	T-BRNN-pre [11]	**59,6**	42,9	49,9	**70,7**	**72,0**	**71,4**	60,7	48,6	54,0	**57,0**
	Corr-BiRNN [12]	53,5	**52,5**	**53,0**	63,7	68,7	66,2	**66,7**	**50,0**	**57,1**	65,4
On-line mód	MaxEnt-(10,1)	42,6	23,9	30,7	53,2	48,9	51,0	33,3	17,1	23,0	87,0
	noWE-LSTM	40,2	**39,3**	39,7	56,2	46,6	51,0	**76,5**	38,2	**51,0**	86,5
	WE-LSTM	**48,8**	37,1	**42,2**	**57,6**	**57,3**	**57,4**	41,2	41,2	41,2	**78,3**
	T-LSTM [17]	41,8	37,8	39,7	56,4	49,3	52,6	55,6	**42,9**	48,4	83,7

Bár ebben a cikkben elsősorban az egyszerű, alacsony késleltetésű írásjelező modellek bemutatására fókuszáltunk, a WE-BiLSTM rendszert a jelenlegi state-of-the-art off-line megoldásokkal is összevetettük. A 7. és 8. táblázatokban látható, hogy mind a [11]-féle T-BRNN-pre modell, mind a [12]-féle Corr-BiRNN felülmúlta a WE-BiLSTM teljesítményét, főként a vesszők és a kérdőjelek beszúrását illetően. Az idézett írásjel-helyreállító rendszerek azonban sokkal összetettebb struktúrával rendelkeznek, és átlagosnak mondható kiépítettségű infrastruktúrán nem képesek valós időben működni, illetve bizonyos feladatokban elfogadhatatlan késleltetést visznek a rendszerbe. A WE-BiLSTM modellek esetén a pontok magas felidézési értékét a manuális és ASR feliratokban kedvező eredménynek tekintjük.

5. Összegzés

Cikkünkben bemutattunk egy alacsony késleltetésű, RNN-alapú, írásjel-visszaállító rendszert, amelynek a teljesítményét magyar és angol nyelvű adatbázisokon is kiértékeltük, valamint összehasonlítottuk egy Maximum Entrópia-alapú szekvenciacímkéző rendszerrel. Mindkét modellezési módszert kiértékeltük on-line módban, ahol a valós idejű működés lehetővé tétele érdekében csak a múltbéli szöveges jellemzők alapján hoztunk döntést; valamint off-line módban is, ahol mind a múltbeli, mind a jövőbeli jellemzőket figyelembe vettük. Az RNN-alapú megközelítés mindegyik tesztkonfigurációban jelentősen felülmúlta a MaxEnt baseline rendszer teljesítményét. Meglepő módon, azonban az on-line mód csak kis mértékben csökkentette az írásjelek helyreállításának pontosságát.

A magyar nyelvű felirat-adatbázisban a különböző televíziós műfajokon mérhető eredményeket összehasonlítva azt találtuk, hogy a szövegalapú modellekben az írásjelek helyreállításának pontossága (a nyelvi modellezéshez hasonlóan) a rendelkezésre álló tanítóadatok mennyiségétől és az adott feladat tervezettségétől függ. Megjegyezzük, hogy a magyar nyelvű beszédátiratok szövegalapú írásjel-helyreállításának témakörében nem ismerünk korábbi munkát.

Annak érdekében, hogy összehasonlítsuk modelljeinket a state-of-the-art RNN írásjelező rendszerekkel, angol nyelvű kísérleteket végeztünk az IWSLT adatbázison, mind on-line, mind az off-line üzemmódban. On-line írásjel-visszaállításban a mi WE-LSTM rendszerünk érte el a legjobb eredményt. Off-line módban a komplexebb megoldások természetesen felülmúlták a mi, elsősorban valós idejű működésre tervezett, egyszerű felépítésű megoldásunkat.

Rendszerünk jövőbeni továbbfejlesztésének egyik legfontosabb iránya, hogy egyesítjük a szöveges jellemzőket használó jelenlegi megoldásunkat a [13]-ban bemutatott prozódia alapú modellel. Az angol nyelvű modell további szöveges vagy akusztikai jellemzőkkel való kiterjesztése szintén ígéretes irány, miközben mindkét nyelven az alacsony késleltetést is szem előtt tartjuk.

Mindent összevetve munkánk egyik legnagyobb eredményének azt tekintjük, hogy sikerült létrehoznunk egy könnyű és gyors RNN modellt alacsony késleltetést megkövetelő írásjel-visszaállítási alkalmazások (pl. TV-műsorok feliratozása) támogatására. Ezen kívül fontos megemlíteni, hogy mindezt az agglutináló ma-

gyar nyelv esetére tettük meg, amely esetén sokkal kevésbé korlátozott a szórend, mint az angol nyelvben. A grammatikai funkciók sokkal kevésbé függnek a szórendtől, mint az utótagoktól (pl. esetvégződésektől), így az adatokban fellépő magasabb fokú változatosság is nehezebbé teszi a szekvencia modellezést.

Köszönetnyilvánítás

A szerzők köszönetüket fejezik ki a Nemzeti Kutatási, Fejlesztési és Innovációs Hivatalnak, amely az FK-124413 projekt keretében a cikkben ismertetésre került kutatást támogatta. Köszönjük továbbá a Pro Progressio Alapítvány (Tarján Balázs), valamint az NVIDIA támogatását (GPU biztosítása az RNN tanításokhoz).

Hivatkozások

1. Batista, F., Moniz, H., Trancoso, I., Mamede, N.: Bilingual experiments on automatic recovery of capitalization and punctuation of automatic speech transcripts. IEEE Trans. on Audio, Speech, and Language Processing **20**(2) (2012) 474–485
2. Paulik, M., Rao, S., Lane, I., Vogel, S., Schultz, T.: Sentence segmentation and punctuation recovery for spoken language translation. In: Proceedings of ICASSP, IEEE (2008) 5105–5108
3. Gravano, A., Jansche, M., Bacchiani, M.: Restoring punctuation and capitalization in transcribed speech. In: Proceedings of ICASSP, IEEE (2009) 4741–4744
4. Beeferman, D., Berger, A., Lafferty, J.: Cyberpunc: A lightweight punctuation annotation system for speech. In: Proceedings of ICASSP, IEEE (1998) 689–692
5. Lu, W., Ng, H.T.: Better punctuation prediction with dynamic conditional random fields. In: Proceedings of EMNLP, ACL (2010) 177–186
6. Huang, J., Zweig, G.: Maximum entropy model for punctuation annotation from speech. In: Proceedings of Interspeech. (2002) 917–920
7. Batista, F.: Recovering Capitalization and Punctuation Marks on Speech Transcriptions. PhD thesis, Instituto Superior Técnico (2011)
8. Cho, E., Niehues, J., Kilgour, K., Waibel, A.: Punctuation insertion for real-time spoken language translation. In: Proceedings of the Eleventh International Workshop on Spoken Language Translation. (2015)
9. Schuster, M., Paliwal, K.K.: Bidirectional recurrent neural networks. IEEE Trans. on Signal Processing **45**(11) (1997) 2673–2681
10. Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555 (2014)
11. Tilk, O., Alumäe, T.: Bidirectional recurrent neural network with attention mechanism for punctuation restoration. In: Proceedings of Interspeech. (2016) 3047–3051
12. Pahuja, V., Laha, A., Mirkin, S., Raykar, V., Kotlerman, L., Lev, G.: Joint learning of correlated sequence labelling tasks using bidirectional recurrent neural networks. arXiv preprint arXiv:1703.04650 (2017)
13. Moró, A., Szaszák, Gy.: A phonological phrase sequence modelling approach for resource efficient and robust real-time punctuation recovery. In: Proceedings of Interspeech. (2017)

14. Renals, S., Simpson, M., Bell, P., Barrett, J.: Just-in-time prepared captioning for live transmissions. In: Proceedings of IBC 2016. (2016)
15. Che, X., Wang, C., Yang, H., Meinel, C.: Punctuation prediction for unsegmented transcript based on word vector. In: Proceedings of LREC. (2016) 654–658
16. Tarján, B., Varga, Á., Tobler, Z., Szaszák, Gy., Fegyó, T., Bordás, Cs., Mihajlik, P.: Magyar nyelvű, élő közéleti- és hírműsorok gépi feliratozása. In: XII. Magyar Számítógépes Nyelvészeti Konferencia: MSZNY 2016, Szeged (2016) 89–99
17. Tilk, O., Alumäe, T.: LSTM for punctuation restoration in speech transcripts. In: Proceedings of Interspeech. (2015) 683–687
18. Ratnaparkhi, A., et al.: A maximum entropy model for part-of-speech tagging. In: Proceedings of EMNLP. (1996) 133–142
19. Recski, G., Varga, D.: A Hungarian NP chunker. The Odd Yearbook **8** (2009) 87–93
20. Makrai, M.: Filtering wiktionary triangles by linear mapping between distributed models. In: Proceedings of LREC. (2016) 2776–2770
21. Chollet, F.: Keras: Theano-based deep learning library. Code: https://github.com/fchollet. Documentation: http://keras.io (2015)
22. Ueffing, N., Bisani, M., Vozila, P.: Improved models for automatic punctuation prediction for spoken and written text. In: Proceedings of Interspeech. (2013) 3097–3101
23. Makhoul, J., Kubala, F., Schwartz, R., Weischedel, R.: Performance measures for information extraction. In: Proceedings of DARPA broadcast news workshop. (1999) 249–252