Kombinált központozási megoldások magyar nyelvre pehelysúlyú neurális hálózatokkal

Tündik Máté Ákos, Szaszák György

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

Kivonat Napjainkban a rekurrens neurális hálókon alapuló szekvenciamodellezés hatékony eszköznek bizonyult több, a természetesnyelv-feldolgozás (NLP) témaköréhez tartozó probléma megoldásában. Ide sorolhatjuk az írásjelek gépi úton történő visszaállítását, vagyis az automatikus központozást is, melynek során a szó- és/vagy akusztikai eseménysorozathoz írásjeleket rendelünk. Ezt a technikát pl. a beszédfelismerő központozatlan kimenetére alkalmazva a szöveg sokkal olvashatóbbá, érthetőbbé válik. Cikkünkben pehelysúlyú kombinált központozási megoldásokat mutatunk be, melyhez karakter- és szószintű beágyazás (embedding) vektorokat, valamint egy 39 dimenziós akusztikai jellemzővektort is felhasználunk. Kísérleteinket két magyar nyelvű, hírműsorokat, illetve felolvasást tartalmazó korpuszon végeztük el. Eredményeinkkel igazoljuk, hogy a kombinált módszerekkel hatékonyabb tud lenni az írásjelek visszaállítása, mintha csak egy-egy szöveges vagy akusztikus komponensre támaszkodnánk.

Kulcsszavak: írásjel-visszaállítás, CNN, RNN, LSTM, karakter, szó, akusztika, prozódia, ASR

1. Bevezetés

Napjainkban nagy népszerűségnek örvend a kutatók között a gépi beszédfelismerés(ASR) kimenetének minél sokoldalúbb feldolgozása, melynek során a nyers szövegből egy ún. információgazdag átirat (rich transcription) keletkezik. Ehhez segítséget nyújtanak a rekurrens neurális hálózatokon alapuló írásjelező modellek is [1,2,3,4], melyek a korábbi, erre a problémára alkalmazott módszerek teljesítményét is felülmúlják [4]. Tipikusan szöveges [4] vagy prozódiai jellemzők [3] használatosak, de pl. angol nyelvre előfordulnak kombinált módszerek is [5].

Mivel az írásjelek az írott nyelv szerves részét képezik, a szöveges jellemzőkkel történő modellezésük különösebb magyarázatot nem igényel. Az [1,2,4] cikkek szerzői olyan, szóbeágyazásokat használó egy- és kétirányú rekurrens neurális hálózatokat (RNN) hoztak létre, mely széles szövegkontextusokból képes sokféle jellemzőt tanulni, ezáltal az írásjeleket hatékonyan beilleszteni a központozatlan szövegbe. A [4] szerzői az RNN-módszerek magasabb hatékonyságát egy tradicionális, Maximum Entrópia-alapú megoldással szemben is demonstrálták.

A [3] szerzői egy hasonló BiRNN architektúrát hoztak létre, de itt az írásjelek elhelyezése prozódiai jellemzők alapján detektált fonológiai frázisok segítségével történik meg, mivel ezek magas korrelációt mutatnak az írásjelekkel. Ezen módszer esetén szükséges a frázisok előzetes modellezése, mely pusztán akusztikai jellemzőkre támaszkodik (a frázisok detektálása az alapfrekvencia, az energia és ezek deriváltjainak segítségével lehetséges), ezért a központozásra alkalmazott modell nem függ pl. a felvételen elhangzott szavaktól/szósorozattól (ez ASR-hibák esetén előnyös).

A két különböző módszer összehasonlításakor mindegyikben találhatunk előnyös tulajdonságot. A szövegalapú jobb összteljesítményre képes (fedés, pontosság és F-pontszám tekintetében), és hatékonyabb a vesszők visszaállításában. Ezzel szemben a prozódián alapuló modell jóval robusztusabb az ASR-hibákkal szemben (a szóhibák továbbterjedése teljesen blokkolt), a mondatvégi írásjelek (mely legtöbb esetben a pont) predikciója precízebb. Tekintve, hogy a frázisok több szóból is állhatnak, így a modell nem minden szóhatárra jósol írásjelet, kombinálása a szöveges jellemzőket felvonultató rendszerrel ezért nem bizonyult sikeresnek. Így cikkünkben a "nyers" akusztikai jellemzőket használjuk fel, melyek kinyerése minden szó/szóhatár esetén megtörténik. Ennek hátránya, hogy bár a szavaktól maguktól továbbra sem függ a prozódiai modell, a hipotetikus szóhatároktól viszont már igen. Mint látni fogjuk, szerencsére ez nem okoz érdemi pontosságveszteséget, cserébe viszont lehetőségünk nyílik közel végponttól végpontig egyetlen modell tanítására.

Továbbhaladva, fontosnak tartjuk annak az esetnek a megvizsgálását is, amikor az írásjelező modellünk bemeneteként az egyes szavakból származó, karaktersorozatokból adódó információt használjuk fel. A karakteralapú modellek népszerűek a természetesnyelv-feldolgozás (NLP) területén is; segítségükkel lehetséges a szövegek szófaji címkézése [6], a nyelvi modellezés [7] és a névelem-felismerés [8]. Számos példát találunk az irodalomban olyan modellekre is, amelyek a karakterekből és a szavakból származó információt hatékonyan kombinálják, pl. névelem-felismerésre [9], gépi fordításra [10], vagy akár szentimentelemzésre [11].

A [12] cikk szerzői karakteralapú írásjelező modellt hoztak létre, melynek teljesítménye alig maradt el egy szóalapú, feltételes véletlen mező (Conditional Random Field, CRF) technológiát használó megoldással szemben. A karakteralapú modellek segítenek az adatelégtelenségi (data sparsity) probléma áthidalásában. Ez különösen fontos az agglutináló magyar nyelv esetén, ahol rengeteg szóalak használatos, ugyanakkor a valós-idejű gépi megoldásoknál kényszerként csak egy kötött szótárméret engedélyezett. Karaktersorozatokban gondolkodva ilyen kötöttséggel nem kell számolni, így a ritka szavak, mint karaktersorozat-inputok, tovább javíthatják az írásjelező modellek predikciós képességét.

Az alacsony szintű (pl. a karakterekből adódó) jellemzők hatékony kinyerése leginkább a konvolúciós neurális hálózatokhoz (CNN) köthető. A gépi látás mellett a beszédtechnológiai kutatásokban is sikerrel alkalmazták ezt a modellt [13,14], utalva arra, hogy nemcsak az emberek, hanem a mesterséges intelligencia is képes az alacsony szintű információk 'intuitív' észlelésére, mely hozzásegíthet a szöveg vagy beszéd értelmezéséhez [15]. A [12] cikk nyomán azt

gondoljuk, érdemes a karakterszintű automatikus írásjelezést magyar nyelvre is megvizsgálni.

Ezen túlmenően, a szöveges (karakter- és szóbeágyazások) és akusztikai jellemzőket együttesen felhasználva, három darab kétkomponensű, és egy darab, három komponensből álló kombinált központozó rendszert is bemutatunk.

Cikkünk az alábbi struktúra szerint épül fel: a 2. fejezetben bemutatjuk a kísérleteinkhez használt adatbázisokat. Ezt követően a 3. fejezetben ismertetjük a pehelysúlyú szó- és a karakteralapú, valamint az akusztikus jellemzőkön alapuló különálló modelleket, valamint ezek kombinált változatait. Továbbhaladva, a 4. fejezetben ismertetjük kísérleti eredményeinket. Végül az eredményekre vonatkozó tanulságokat levonva, felvázoljuk jövőbeni terveinket.

2. Adatbázisok

Az írásjel-visszaállítási kísérleteinkhez két magyar nyelvű adatbázist használtunk fel: a BABEL-t [16] és a Magyar Híranyag-adatbázist [17]. A tanítást és kiértékelést külön-külön végeztük el a két adatbázisra, azok jelentős különbségei miatt. Mindkét adatbázis nagyságrendileg 3-3 óra beszédet tartalmaz. 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.

Megjegyezzük, hogy mind a szó-, mind a karakteralapú beágyazások tanításához kiegészítő, csak szövegesen elérhető adatbázisokat használtunk fel a [18] irodalomban bemutatottaknak megfelelően.

Az anyagokat 60%-20%-20% arányban osztottuk fel tanítás, validálás és tesztelés céljából; a prozódiai írásjelező modell teljes egészében a BABEL-en illetve a Magyar Híranyag-adatbázison tanult, a [18]-ban ismertetett korpuszon előtanított szó- és karakteralapú módszerek esetén pedig adaptációt hajtottunk végre, tudástranszfert alkalmazva. A BABEL esetében a korpusz szövegrészei részben ismétlődnek; erre gondosan odafigyeltünk a tanító, validáló, és tesztelő halmazok összeállításakor.

A hírkorpuszon 35%-os, a BABEL-en mintegy 50%-os szóhibaarányt mértünk. (Az agglutináló, illetve egybeírandó szóösszetételekben gazdag nyelvekre a WER mindig jóval magasabb az angol nyelven mérthez képest a hasonló felismerési feladatok esetében [19].)

3. Írásjel-visszaállító módszerek

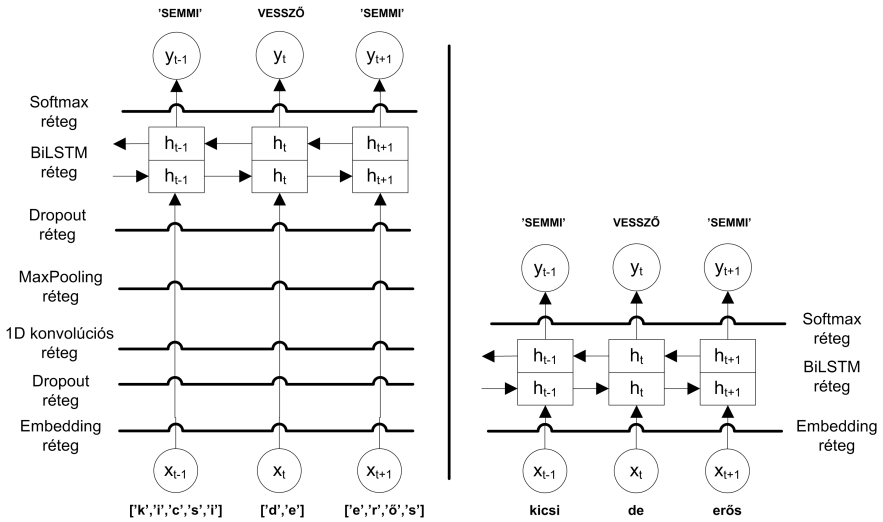
3.1. Szóalapú modell

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 kétirányú RNN modell teljesítményét vizsgáljuk meg. A modellben az aktuális szót megelőző időpillanatra jósoljuk az írásjelet. A kísérletekhez használt szóalapú RNN-architektúrát az 1. ábrán mutatjuk be.

A szóalapú RNN-modell ("W") a következőképpen épül 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) kerülnek. Ezek a reprezentációk a következő, rejtett rétegbe továbbítódnak, amely BiLSTM rejtett cellákból áll, ezek a 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 x_t szó előtti időpillanatra (slot-ra) adja meg.



1. ábra: A karakteralapú "C" RNN modell (bal oldalon) és a szóalapú "W" RNN modell (jobb oldalon) szerkezete

A "W" modellt a tanítókorpusz leggyakoribb 100 ezer szavával tanítottuk, valamint különböző dimenziószámú, előre tanított magyar nyelvű szóbeágyazási modelleket is kipróbáltunk [20]. A tanítás során az RNN modell súlyait a kategorikus keresztentrópia költségfüggvény alapján módosítjuk, valamint minden egyes epoch-ban frissítjük a szóbeágyazásokat is.

3.2. Karakteralapú modell

A "C" karakteralapú modellünk az 1. ábrán látható módon épül fel. A modell – hasonlóan a szóalapú megoldáshoz – fix hosszúságú szekvenciákat fogad a bemenetén, melyben a szavak karaktersorozatokként reprezentáltak. Minden egyes

karakter a karakterek által alkotott beágyazási térbe kerül. Fontos eltérés a szóalapú modellhez képest, hogy a karakteralapú modell tanításának kezdetén a beágyazási tér vektorai (előtanítás nélkül) véletlenszerűen inicializáltak. Másfelől, a karakteralapú modell előnye, hogy nem szükséges az OOV-szavakat kezelni, az összes karaktert tartalmazó szótár limitált számossága miatt; így az ezen szavakból képzett karaktersorozatok is befolyásolják/segítik a jellemzőtanulást.

A beágyazási transzformációt követően, az 1D-konvolúció művelete (különböző súlyozású konvolúciós szűrőivel) számos reprezentációt készít a transzformált bemenetből. Ezekből a reprezentációkból a dimenziócsökkentést elvégző MaxPooling réteg segítségével egy új, jóval tömörebb jellemzővektor keletkezik. Végül a BiLSTM réteg ismét a t időpillanat kontextusának rögzítéséért, az információ kinyeréséért felelős. A kimeneti írásjelcímkék posterior valószínűségeit ismét egy $softmax$ aktivációs függvény használata után kapjuk meg. A köztes Dropout rétegek célja, hogy elkerüljük a modell túltanulását.

3.3. Akusztikai-prozódiai modell

A [3] szerzői az automatikus központozáshoz fonológiai frázisszegmentálásból származó prozódiai jellemzőket használtak fel. Mivel ezt a szegmentálást egy külön Rejtett Markov-modell végzi el, ehelyett vizsgálatainkhoz csak a frázisszegmentálást segítő akusztikai-prozódiai jellemzőket tartottuk meg a neurális háló alapú írásjelezés esetén. Az alapfrekvencia és átlagos energia kinyerése egy 150 ms-os ablakban történik (mel skálára bontás nélkül), 10 ms-onként mintavételezve, 5-pontos medián szűrővel simítva. Az x_t szóhoz tartozó jellemzővektorba az alapfrekvencia- és energiaértékek első- és a másodrendű deriváltjai is bekerültek (d_t), melyeket az alábbi regressziós képlet segítségével számítottunk ki $W = 30$ keret hosszú kontextust figyelembe véve:

$$d_t = \frac{\sum_{i=1}^{W/2} i(x_{t+i} - x_{t-i})}{2\sum_{i=1}^{W/2} i^2} \tag{1}$$

Ahol a beszédfelismerő szóhatárt feltételez, ott újabb, két 6-dimenziós jellemzővektor kinyerése történik meg; egy a szóhatárt megelőző 15 keretben, egy pedig az utána következő 15 keretet befoglalva. Ehhez alap statisztikai értékeket számítunk; a minimum-, maximum- és átlagértékek kerülnek a 6x6 dimenziós vektorokba. A bemeneti vektort végül az aktuális szót megelőző szó időtartamával, és a két szó között eltelt szünetértékkel egészítjük ki. Az akusztikai alapú "P" modellünk is pehelysúlyú; a jellemzővektort egy kétirányú LSTM rétegbe irányítjuk, ezt követően pedig a softmax réteg felel a kimeneti írásjelért. Sajnos kevés hanganyag állt rendelkezésünkre, azonban az akusztikus modellünk tanításához ez is elegendőnek bizonyult; a modellre vonatkozó "legjobb" hiperparamétereket az 1. táblázat mutatja be.

A [3] szerzői által ismertetett módszerrel szemben ugyan szükségünk van az ASR szolgáltatta szóhatárokra, de mint látni fogjuk, ez a modell szóhiba-tűrését nem csökkentette érdemben. Úgy véljük, ez a technika kellően robusztus és mégis

egyszerű, mivel a dinamikus (első- és másodrendű derivált) jellemzők segítségével a legfontosabb prozódiai sajátosságokat, azok kontextusát tudjuk kinyerni a szóhatárokon (lokális hangsúlymintázatok, intonáció és szünet tükrében).

3.4. Hiperparaméterek

Szisztematikus, kimerítő keresés (grid search) alapú optimalizációt hajtottunk végre a "C", W" és a "P" modellek hiperparaméterein, a validációs halmaz elemeit értékelve. A szekvenciák hosszát, a rejtett állapotok számát, a minibatch méretét, és az optimizáló típusát mindhárom modell esetén változtattuk, valamint korai leállítást (early stopping, *Patience*) is használtunk, a túltanítás elkerülése érdekében. A szöveges modellek esetén a szótár méretét, valamint a szó- illetve karakterbeágyazási dimenziót is konfiguráltuk. Emellett a "C" modell esetén a konvolúciós szűrők számát és azok hosszát, a MaxPooling-ablak méretét és a bemenet átlapolásánál alkalmazott lépésközt változtattuk. Az 1. táblázat összefoglalja a modelljeinkben használt hiperparaméterek végső értékeit.

1. táblázat. A szöveges és a prozódiai alapú modellek hiperparaméterei

Bemenet	**Modell**	**Szekv. Hossza**	**Szótár Mérete**	**Beágyazási dimenzió**	**Rejtett állapotok**	**Batch mérete**	**Optimalizáló**	**Szűrők hossza**	**#Szűrők**	**Lépésköz**	**MaxPooling ablakméret**	*Patience*
Szavak	"W"	200	100,000	300	512	128	RMSProp	N/A	N/A	N/A	N/A	3
Karakterek	"C"	200	100	80	512	128	RMSProp	6	70	2	25	3
Prozódia	"P"	200	N/A	N/A	512	16	RMSProp	N/A	N/A	N/A	N/A	3

A központozó rendszerek implementálásához a Keras keretrendszert [21] használtuk, a tanítást GPU-n végeztük el.

3.5. Hibrid modellek

A különböző inputokat páronként kombinálva ("karakter és szó" ("C+W"), "karakter és prozódia" ("C+P"), "szó és prozódia" ("W+P")) három különböző hibrid modellt vizsgáltunk meg, valamint egy negyediket is, mely mindhárom bemenetet egyszerre dolgozza fel ("C+W+P"). A hibrid modellekhez a különálló előtanított karakter- ("C") és szóalapú ("W") modellek súlyait is felhasználtuk, kombinálva a prozódia ("P") alapú modell bemenetével. Az összekapcsolás a "C" és/vagy "W" modellek softmax kimeneti aktivációs rétegeit megelőző BiLSTM rétegeken történt meg, illetve a softmax rétegekkel történő összeillesztést is kipróbáltuk. Az összeillesztett alsóbb rétegekhez hozzáadtunk még egy új, közös BiLSTM réteget és egy új softmax réteget; így állt össze a teljes hibrid hálózat.

4. Kísérleti eredmények

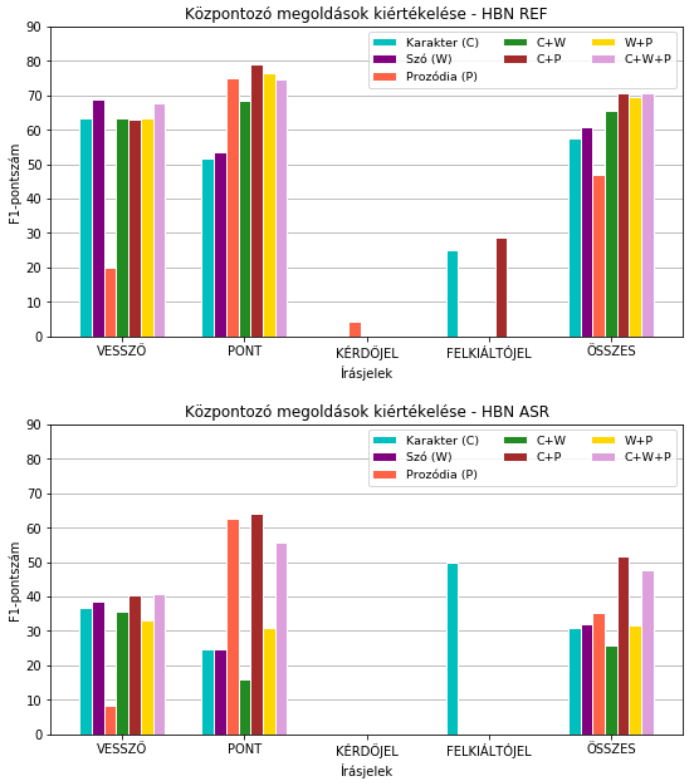
A következő fejezetben bemutatjuk a magyar nyelvű írásjelezési kísérleteink eredményeit. A részletes kiértékelést egy standard információ-visszakeresési mutató, az F1-érték mentén mutatjuk be, melyet az írásjelekre vonatkozó Pontosság (Pr)

és Fedés (Rc) értékekből származtattunk. Ezenkívül a legjobban teljesítő modellekhez megadjuk a Slot Error Rate (SER) [22] értéket is, amely egy metrikában egyszerre tükrözi az írásjel-visszaállításhoz kapcsolódó hibák minden lehetséges típusát - beszúrásokat (Ins), helyettesítéseket (Sub) és törléseket (Del), N helyes találat mellett:

$$SER = \frac{C(Ins) + C(Subs) + C(Del)}{C(slotok\_szama = N + Subs + Del)}, \quad (2)$$

ahol $C(.)$ a számláló operátor, a slot-ok pedig azon szavakat követő helyek a szövegben, amelyekben helyesen szerepel írásjel.

A Magyar Híranyag-adatbázis (HBN) kézi illetve az ASR átiratairaira vonatkozó eredmények a 2. ábrán láthatók.

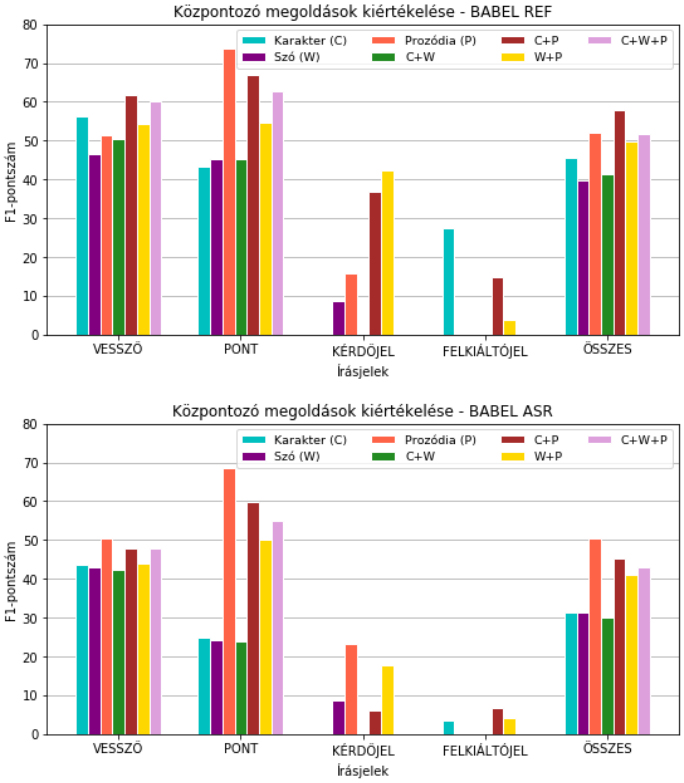


2. ábra: A Magyar Híranyag-adatbázisból származó kézi és ASR feliratok központozása

A szövegalapú ("C" és "W") modellekkel leginkább a vesszők visszaállítása lehetséges, mind a kézi (REF), mind az ASR átiratokon. Ezek a modellek a pont predikciójának tekintetében gyengébb mutatóval rendelkeznek, szemben a

prozódiai "P" modellel, mely ebben a tekintetben jól teljesít, viszont gyenge a vesszők jóslásában. Ígéretes tehát a két jellemzőkészlet kombinálása: a szövegalapú komponenseket a prozódiával kombinálva (akár párban ("C+P", "W+P"), akár hármasban ("C+W+P") további javulás tapasztalható az automatikus írásjelező modellek teljesítményében. A legjobb eredményt a "C+P" inputkombinációval értük el, mind a kézi átiratokon ($F1 = 70,7\%; SER = 45,1\%$), mind az ASR-kimeneten ($F1 = 51,8\%; SER = 78,2\%$).

A BABEL adatbázis kézi illetve az ASR átirataira vonatkozó eredmények a 3. ábrán láthatók.



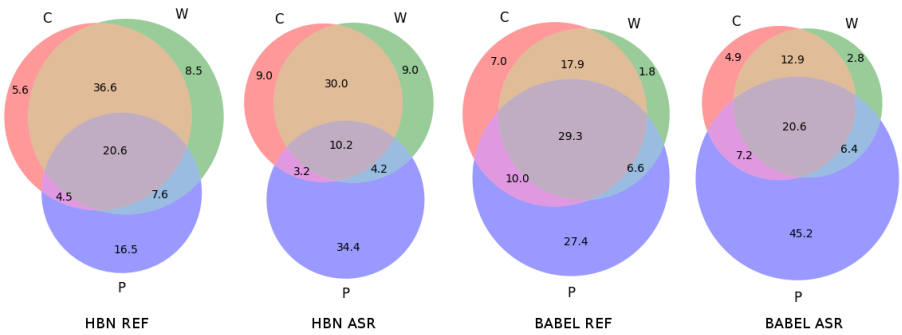
3. ábra: A BABEL adatbázis kézi és ASR átiratainak központozása

A HBN adatbázissal összevetve kiugró a "P" modell szerepe, amit a BABEL kontrollált és gondos artikulációjú felvételeivel magyarázunk. A legjobb eredményt a BABEL-es kézi átiratokon szintén a "C+P" inputkombinációval értük el ($F1 = 58,0\%; SER = 55,6\%$), viszont az ASR-kimeneten a "P" modell önmagában a leghatékonyabb, a magas szóhibaarány következtében a szöveges jellemzőkkel kiegészített hibridek nem tudtak magasabb teljesítményt elérni ($F1 = 50,3\%; SER = 71,7\%$).

A kérdések azonosításában meglepő módon a "P" modell a BABEL-en is gyengén teljesít, ennek okát az adatok kiegyensúlyozatlanságában látjuk, míg a HBN adatbázis esetén arra vezetjük vissza, hogy a kérdések és felkiáltások nem a megfelelő intonációval realizálódnak, hanem a deklaratív irányba tolódnak el jelentősen (de kevés is a minta ezekre a mondatokra az adatbázisban). Ezeket a feltételezéseinket lehallgatással is ellenőriztük, de megfelelnek a [5] irodalom megfigyeléseinek is.

Megjegyezzük, hogy a kísérleteket angol nyelvre is elvégeztük, noha hely hiányában az arra vonatkozó eredményeket nem mutatjuk be részletesen; ott a "P" modell kézi átiraton lényegesen gyengébb, de ASR átiraton szintén felerősödő szerepét tapasztaltuk. Fontos különbség, hogy angol nyelven a mondatvégi írásjelek (pontok) a "W" és a "C" modellekkel is pontosabban detektálhatók voltak kézi átiraton, mint a "P" modellel, illetve az ASR kimeneten a "W+P" hibrid bizonyult a leghatékonyabbnak, igaz „csupán” $p < 0,05$ szignifikanciaszint mellett.

A 4. ábrán Venn-diagramokkal mutatjuk be, hogy a "C","W" és "P" modellek milyen mértékben járulnak hozzá az írásjelek helyes visszaállításához (%-ban megadva). Habár a "P" modell összességében kevesebb írásjelet volt képes megfelelően beszúrni a szövegbe a HBN átiratokon (a gondosabban intonált BABEL-nél nem), de a szerepe látványos; a helyesen beszúrt írásjelek 15-25%-át egyedüliként fedte kézi átiratokon (REF), míg ASR kimeneten az írásjelek harmadát-felét egyedüliként képes detektálni. A kézi átiratokon a szövegalapú modellek (BABEL esetében enyhe) dominanciája figyelhető meg. A "P" modell az ASR átiratok esetén szépen javítja a központozó modell beszédfelismerési hibákkal szembeni robusztusságát.



4. ábra: A szövegalapú és prozódiai modellek kontribúciója a helyesen visszaállított írásjelek halmazát tekintve, a "HBN" és a BABEL korpuszon

Az eredmények alapján az alábbi következtetések rajzolódnak ki: a szakirodalomban is ismert a "P" modell jó szóhibatűrése, amelyet mi is demonstráltunk.

Magyar nyelvre a mondatvégi írásjeleket a szövegalapú modellek pontatlanabbul jelezték előre a vesszőkhöz képest, sőt, angol nyelvre is pontosabb ezen írásjelek predikciója. Ezt a magyar kevésbé kötött szórendjére és a szóalakok relatíve magas számára vezetjük vissza: az agglutináló nyelvek esetén - mint a magyar - a szóbeágyazások alkalmazása kevésbé hatékony; ennek egyfelől az az oka, hogy a több különböző előforduló szóalak miatt az OOV-arány is általában magasabb, mint például az angol nyelvben (esetünkben HBN-re az OOV-arány 8,6%, BABEL-re 11,8% volt), míg nagyobbrészt azt feltételezzük, hogy a kevésbé kötött szórend miatt nagy szókontextusra (akár a beágyazás alapjául szolgáló skip-gram kontextusablakán is kívülre) kiterjedő nagyobb változékonyság miatt a beágyazások szemantikus kapcsolatokat jósló képessége kevésbé robusztus. Karakter N-gramokkal és az ASR-szótár elemeire illesztett szóbeágyazásokkal lehetne javítani az OOV miatti problémán, ezzel a szemantikai pontosságot is növelve, ahogy azt a [23] cikkben több nyelvre be is mutatták. Ezek bevonásától jelen cikkben eltekintettünk, és a hibrid modellekben a karakteralapú modellünkből kinyert jellemzőket használtuk fel, amely kisebb mértékben, de szintén javította a robusztusságot.

Ezzel a hipotézissel összhangban van a szövegalapú modellek vesszőkre vonatkozó magas predikciós képessége is: a magyar nyelvben (is) két esetben gyakori a vesszők használata; egyrészt a kötőszavak előtt (melynek szerepe a különböző tagmondatok elválasztása), másrészt a felsorolásban. Az előbbi esetben a kötőszóhoz tartozó szóbeágyazás általában ismert, az utóbbi esetben pedig tipikusan a szemantikailag hasonló szavakat kapcsoljuk össze. Mindkét esetben megmutatkozik a szóbeágyazások segítő szerepe, melyet másutt a kevésbé kötött szórend miatti „csere-bere” lehetősége itt nem befolyásol.

Összegezve az írásjelezési eredményeket, az agglutináló és kötetlen szórendű magyar nyelven szignifikáns teljesítménynövekedés érhető el a karakterszintű és a prozódiai jellemzők bevonásával, a szóalapú baseline modellünkkel összehasonlítva. Kiemelve az ASR átiratok központozását, a karakter-prozódiai jellemzőpárost használó hibrid modell segítségével közel 40%-os relatív javulás érhető el F1-érték tekintetében, a valós ASR felhasználási körülményeket jól reprezentáló HBN korpuszon, mely $p < 0,01$ érték mellett is szignifikáns.

5. Összegzés

Cikkünkben különböző automatikus írásjelező modelleket mutattunk be, szöveges jellemzők (karakterek és szavak) és prozódiai jellemzők egyenkénti, valamint kombinált használatával. Fontos kiemelni a prozódiai modellek teljesítményét, mely a gépi beszédfelismerésből származó hibák ellenére is képes a hatékony írásjelezésre, szemben a szóalapú modellel, mely meglehetősen érzékeny azokra. Kismértékben a karakteralapú modell is emeli a szóhibatűrést. A szóalapú modell teljesítményét befolyásolja az is, hogy a kevésbé kötött szórend és a nagy szótárméret miatt a szóbeágyazások által biztosított szemantikai modellező képesség és koherencia is csak korlátozottabb mértékben tud érvényesülni. A karakter-prozódia jellemzőket együttesen használó hibrid modell bizonyult a leghatéko-

nyabbnak a kézi átiratokon, míg az ASR esetben a karakter-prozódia párossal működő hibrid modellel (a HBN korpuszon), illetve a BABEL-en a prozódiai hálóval értük el a legjobb eredményt, F1 és SER tekintetében. Az írásjelekre kitérve, a prozódiai modell erőssége a pontok, míg a szövegalapúaké a vesszők visszaállítása. Úgy véljük, hogy a karakteralapú és prozódiai modellekhez kapcsolódó eredményeink és megfigyeléseink a többi agglutináló nyelvre is érvényesek lehetnek; ezeket további vizsgálatokkal lenne érdemes alátámasztani.

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.

Hivatkozások

1. Tilk, O., Alumäe, T.: LSTM for punctuation restoration in speech transcripts. In: Proceedings of Interspeech. (2015) 683–687
2. Tilk, O., Alumäe, T.: Bidirectional Recurrent Neural Network with Attention Mechanism for Punctuation Restoration. In: Proceedings of Interspeech. (2016) 3047–3051
3. Moró, A., Szaszák, G.: A phonological phrase sequence modelling approach for resource efficient and robust real-time punctuation recovery. In: Proceedings of Interspeech. (2017)
4. Tündik, M.Á., Tarján, B., Szaszák, G.: Low Latency MaxEnt-and RNN-Based Word Sequence Models for Punctuation Restoration of Closed Caption Data. In: International Conference on Statistical Language and Speech Processing, Springer (2017) 155–166
5. Klejch, O., Bell, P., Renals, S.: Sequence-to-sequence models for punctuated transcription combining lexical and acoustic features. In: Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on, IEEE (2017) 5700–5704
6. Hardmeier, C.: A neural model for part-of-speech tagging in historical texts. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. (2016) 922–931
7. Kim, Y., Jernite, Y., Sontag, D., Rush, A.M.: Character-aware neural language models. In: AAAI. (2016) 2741–2749
8. Klein, D., Smarr, J., Nguyen, H., Manning, C.D.: Named entity recognition with character-level models. In: Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4, Association for Computational Linguistics (2003) 180–183
9. Chiu, J.P., Nichols, E.: Named entity recognition with bidirectional LSTM-CNNs. arXiv preprint arXiv:1511.08308 (2015)
10. Chung, J., Cho, K., Bengio, Y.: A character-level decoder without explicit segmentation for neural machine translation. arXiv preprint arXiv:1603.06147 (2016)
11. dos Santos, C., Gatti, M.: Deep convolutional neural networks for sentiment analysis of short texts. In: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. (2014) 69–78

12. Gale, W., Parthasarathy, S.: Experiments in character-level neural network models for punctuation. Proc. Interspeech 2017 (2017) 2794–2798
13. Abdel-Hamid, O., Deng, L., Yu, D.: Exploring convolutional neural network structures and optimization techniques for speech recognition. In: Interspeech. Volume 2013. (2013) 1173–5
14. Tóth, L.: Combining time-and frequency-domain convolution in convolutional neural network-based phone recognition. In: Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on, IEEE (2014) 190–194
15. McNamara, D.S., Kintsch, E., Songer, N.B., Kintsch, W.: Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. Cognition and instruction **14**(1) (1996) 1–43
16. Roach, P., Arnfield, S., Barry, W., Baltova, J., Boldea, M., Fourcin, A., Gonet, W., Gubrynowicz, R., Hallum, E., Lamel, L., et al.: BABEL: An Eastern European multi-language database. In: Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference On. Volume 3., IEEE (1996) 1892–1893
17. Teleki, C., Szabolcs, V., Levente, T.S., Klára, V.: Development and evaluation of a Hungarian Broadcast News Database. In: Forum Acusticum. (2005)
18. Tündik, M.A., Szaszák, G.: Joint Word-and Character-level Embedding CNN-RNN Models for Punctuation Restoration. In: Proceedings of the 9th IEEE International Conference on Cognitive Infocommunications (CogInfoCom 2018), IEEE (2018) 135–140
19. Kurimo, M., Puurula, A., Arisoy, E., Siivola, V., Hirsimäki, T., Pylkkönen, J., Alumäe, T., Saraclar, M.: Unlimited vocabulary speech recognition for agglutinative languages. In: Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics -, Morristown, NJ, USA, Association for Computational Linguistics (2006) 487–494
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. Makhoul, J., Kubala, F., Schwartz, R., Weischedel, R.: Performance measures for information extraction. In: Proceedings of DARPA broadcast news workshop. (1999) 249–252
23. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. arXiv preprint arXiv:1607.04606 (2016)