DOI: 10.31915/NWS.2021.8

Automatikus kézírás-felismertetés Kiss József levelezésén

Szűcs Kata Ágnes
Petőfi Irodalmi Múzeum Digitális Bölcsészeti Központ (Budapest)
szucs.kata@dbk.pim.hu

Abstract
The digital edition of the József Kiss correspondence is a pilot project of the Centre for Digital Humanities, Petőfi Literary Museum. In addition to the processing of the personal and professional letters of the 19th-century writer, poet, and editor of the literary journal *A Hét* (The Week), the project is to explore the possibilities offered by the Transkribus software.

Handwritten Text Recognition is an emerging field of the digital humanities. The paper will discuss this artificial intelligence-based technology and our experiences in creating a Hungarian model. The best result has a 6,94% character error rate. Apart from the practical experience gained in testing, the paper discusses the possibilities of implementing HTR in public collections.

Keywords: HTR, handwritten text recognition, Transkribus, correspondence, 19th-century literature

1. Bevezető

Az alábbi esettanulmány a Petőfi Irodalmi Múzeum Digitális Bölcsészeti Központ Handwritten Text Recognition (HTR), azaz a kézírás-felismertetéssel kapcsolatos eddigi munkáját, és a Transkribus használatát hivatott bemutatni. Rövid elméleti áttekintés után, ismertetem a témában szerzett tapasztalatainkat, milyen eredményeket kaptunk a modellek tanításakor, és hogyan sikerült javítani rajtuk.

A modellépítés alapját a PIM Kézirattárában folyamatos feldolgozás alatt álló Kiss József-levelezés képezi. A 493 leltári tételből álló korpusz 1683 darab személyes vagy szakmai levelet jelent, melyeknek címzettje vagy írója Kiss József, a 19–20. század fordulóján élt költő és lapszerkesztő. Az ő nevéhez kötődik *A Hét* című hetilap alapítása, mely csomópontot jelentett a kialakulóban lévő polgári értelmiség képviselőinek, mintegy alapot teremtve a későbbi irodalmi folyóirat, a Nyugat számára is. A levelezés kiadása hiánypótló és irodalomtörténeti szempontból is jelentőséggel bír. Emellett lehetőséget ad arra, hogy a digitalizált kéziratok gyűjteményi kezelésére, digitális források kiadására alakítsunk ki alternatívákat.

2. Automatikus kézírás-felismertetés (HTR)

A kézírás felismerését célzó technika sokáig együtt fejlődött az OCR-rel (optikai karakterfelismerés), ahol a szkennelt dokumentumok nyomtatott szövege válik gép által olvashatóvá. Az OCR technológiában az egyes karakterek képezik a felismerés

alapját, melyeket előre megadott mintákkal hasonlít össze. A HTR különálló kutatási területté fejlődött a 2000-es évek óta, a kézírások különbözősége és a feladat számítási komplexitása miatt.[1]

Az egyik fő különbség a két technológia között, hogy a HTR egy szegmentált sor szövegében lévő összes karakter felismerésére fókuszál.[2] A gépi tanulással támogatott technológia képes a vizuális jegyek elsajátítására (így nem kell külön mintákat létrehozni), a neurális háló segítségével pedig több egymást átfedő szövegsor képéből képes karakter valószínűséget számítani.[3][4] Természetesen a HTR technológiát nem csak kézzel írt szövegeken lehet alkalmazni.

3. Transkribus
Szegmentálás, szöveg átírása

Az ingyenesen letölthető és használható szoftver megkönnyíti a kéziratokkal való munkát. Az átírás sorról sorra történik a digitális fakszimile folyamatos jelenlétében. Verziókövető rendszerrel van ellátva a program, tehát minden mentés bármikor visszaállítható. Felhő alapú, ezért a megfelelő elővigyázatosságok mellett egy gyűjteményen egyszerre többen is dolgozhatnak párhuzamosan. Sokrétű címkézési és metaadatolási rendszer van beépítve. A fájlokat többféle kimeneti formátumban lehet exportálni (.pdf, .tei, .docx, txt, .xlsx, .zip, stb.) A Transkribus emellett lehetőséget biztosít az automatikus kézírás-felismertetésre és a kézírást felismerő modell létrehozására is.

Magyar nyelven még nem készült nyilvánosan elérhető modell, így azt a program segítségével kezdtük el építeni.[5] Ehhez a szkennelt kéziratképeket fel kell tölteni a Transkribus szervereire, majd előállítani egy minimum 5–15000 szót tartalmazó, átírással rendelkező korpuszt. A nyomtatott szövegek esetén kevesebb is elengedő lehet. Az átírást a Transkribuson belül is létre lehet hozni, de meglévő átirat akár utólag is hozzárendelhető a képekhez (Text2Image funkció).[6]

Az átíráshoz a képeket szegmentálni kell. A digitális képfeldolgozásban a szegmentálás egy kép több szegmensre (pixelhalmazokra, más néven képobjektumokra) történő

1 Katuščak Dušan, „Automated Transcription of Handwritten Text: READ and TRANSKRIBUS (An Experiment with Transcribing Letters of Andrej Kmeť)", 2019. október 20.

2 Puigcerver Joan, „Are Multidimensional Recurrent Layers Really Necessary for Handwritten Text Recognition?", in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 01, 2017, 67–72, https://doi.org/10.1109/ICDAR.2017.20.

3 Dietrich Felix, „OCR vs. HTR or "What Is AI, Actually?"", *READ-COOP* (blog), elérés 2021. június 2., https://readcoop.eu/insights/ocr-vs-htr/.

4 Puigcerver Joan, „Are Multidimensional Recurrent Layers Really Necessary for Handwritten Text Recognition?", in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 01, 2017, 67–72, https://doi.org/10.1109/ICDAR.2017.20.

5 „Public Models in Transkribus", READ-COOP, elérés 2021. május 26., https://readcoop.eu/transkribus/public-models/.

6 Vö.: „How To Use Existing Transcriptions to Train a HTR-Model with the TextToImage-Tool", READ-COOP, elérés 2021. június 7., https://readcoop.eu/transkribus/howto/how-to-use-existing-transcriptions-to-train-a-handwritten-text-recognition-model/.

felosztásának folyamata, célja a kép ábrázolásának egyszerűsítése.[7] A folyamat során olyan területekre osztjuk a képet, melyek meghatározott koordináták alapján felismerhetővé teszik a sorok és bekezdések helyzetét a neurális háló számára. Három réteget jelölünk ki a képen poligonok formájában (Text Region, Line és a Base Line). A szegmentálás manuálisan és automatikusan[8] (ill., a kettőt keverve) is elvégezhető. Míg az előbbi pontosabb, de időigényesebb, az utóbbinál szükség van az utólagos korrekcióra.

A szegmentálás után megkezdődhet a szövegek átírása is, ami ily módon a sorok szintjére lebontva eleve összeköttetésben van a hozzá tartozó képpel.[9]

A kézírás-felismertetés szempontjából fontos átíráskor a lehető legpontosabb, betű szerinti átírásra törekedni, és a karaktereket következetesen ugyanúgy megadni.[10] Az Unicode határain belül lehetőség van speciális karakterek beszúrására is.

HTR modell építése

Az átírással rendelkező képeket két csoportba osztjuk, melynek során egy Training set (TS) és egy Validation set (VS) jön létre. Lehetőség van ezek automatikus válogatására is, ami 2%, 5%, vagy 10% VS-eket hoz létre. A TS a rendelkezésre álló fájlok kb. 90%-át teszi ki.

A program a TS-be került fájlokon mesterséges intelligencia segítségével azonosítja a sorokban látható írásképet az átírt szöveggel. Egy tanulási ciklus (epoch) során végigmegy a TS összes fájlján és vizuális jegyek alapján kitalálja, hogy a szegmentált sorok milyen karakterekből állnak össze. Ezután ellenőrzi magát az általunk megadott átírásra hagyatkozva. A tanulási folyamat több ilyen ciklusból áll össze, és a következő epoch-ba már az előzőből elsajátított tudással kezd bele. Az epoch-ok száma a dokumentumok minősége szerint változtatható (max. 250).

Végül a másik csoporton, a VS-en teszteli le magát a mesterséges intelligencia. Ennél a csoportnál csak a TS-en elsajátítottakra és a képen látható vizuális jegyekre hagyatkozik. A VS-ben lévő fájlokon egyszer megy végig, majd ellenőrzi magát az emberi intelligencia által készített átírás alapján. Az itt kapott hibaérték (CER on Validation set) azt jelzi, hogyan teljesít a modell egy ismeretlen szövegen. Az eredményességet tovább lehet növelni Base Model beépítésével. Ilyenkor egy másik HTR modellből már elsajátított tudást építünk be.

7 Srinivasan G N, „Segmentation Techniques for Target Recognition" 1, 3 (2007): 7.

8 Elforgatott rájegyzések, szokatlan oldaltörések (pl. borítékok, képeslapok) esetében többet téveszt.

9 A Transkribus lehetőséget biztosít a szavak szintjén történő szegmentálásra is, de mivel a funkció használatát egy rövid tesztidőszakot követően elvetettük, a továbbiakban erről nem lesz szó

10 Például egy német szövegnél a sárfesz s-t egyféleképpen, ß karakterrel, vagy dupla s-sel jelöljük, nem keverve. A hiányzó szövegrészeket pedig jobb üresen hagyni.

Az eddigi legjobb eredményt a KEZ17_Kiss József kézírása_5 nevű modellel sikerült elérni, ahol a CER on Validation Set értéke 6,94% volt. Ez azt jelenti, hogy a modell által készített átírásban könnyen javítható hibák találhatók (pl. ékezetek, egy-egy betűtévesztés).

4. Tapasztalataink

4.1 Első fázis

Egy modell kialakításakor fontos, hogy a korpuszban jobbára egy kéz által írt szövegek legyenek, melyek korban közel keletkeztek egymáshoz, illetve, hogy a forrás típusa is egyező legyen (pl. napló, számadáskönyv, levelezés stb.).[11] Először létrehoztunk egy csak Kiss József kézírásából és egy mindenki máséból álló korpuszt, ahova folyamatosan kerültek be újabb átírt anyagok a neurális háló számára.

Eleinte kevés kapaszkodót találtunk az eredmények értékelésére. A modellek finomhangolására nézve saját magunknak kellett egy viszonyrendszert kialakítani. Először is felmértük a rendelkezésre álló anyag összetételét. Erre különböző csoportokat alakítottunk ki, felcímkézve és táblázatos formában nyilvántartva a levelek nyelvét, a kézírás eszközét (ceruza vagy toll), illetve, hogy géppel írt vagy nyomtatott a dokumentum. Úgy véltük ezek olyan faktorok, amelyek befolyásolhatják a kézírás-felismertetést végrehajtó neurális háló viselkedését. A gépi feldolgozás során az algoritmus figyelembe veszi az adott nyelvben előkerülő karakterek számát és azok egymáshoz viszonyított előfordulási gyakoriságát is. Másrészt a nyomtatott és géppel írt betűk más vizuális jegyekkel rendelkezek, mint a kézzel írtak.

homokozó, próba 2				kiss_jozsef_masolat, Kiss József kézírása				kj 2				kj3			
validation set	7	training set	61	validation set	22	training set	194	validation set	15	training set	137	validation set	22	training set	229
CER	9,23%	CER	0,95%	CER	12,66%	CER	1,99%	CER	17,29%	CER	1,12%	CER	9,72%	CER	1,89%
0	átüt a tinta	2	átüt a tinta	0	átüt a tinta	2	átüt a tinta	0	átüt a tinta	0	átüt a tinta	0	átüt a tinta	2	átüt a tinta
0	boríték	1	boríték	0	boríték	27	boríték	0	boríték	1	boríték	3	boríték	32	boríték
3	ceruza	23	ceruza	5	ceruza	63	ceruza	4	ceruza	54	ceruza	4	ceruza	72	ceruza
0	címzés	1	címzés	4	címzés	6	címzés	0	címzés	0	címzés	1	címzés	15	címzés
0	firka	1	firka	0	firka	0	firka	0	firka	0	firka	0	firka	2	firka
0	gépirat	1	gépirat	0	gépirat	1	gépirat	0	gépirat	0	gépirat	0	gépirat	0	gépirat
0	nem KJ kézírás	0	nem KJ kézírás	0	nem KJ kézírás	1	nem KJ kézírás	0	nem KJ kézírás	0	nem KJ kézírás	0	nem KJ kézírás	0	nem KJ kézírás
0	kép. firka.	0	kép. firka.	0	kép. firka.	0	kép. firka.	0	kép. firka.	0	kép. firka.	0	kép. firka.	10	kép. firka.
1	német	1	német	0	német	0	német	0	német	0	német	0	német	1	német
0	nyomtatott kézírás	0	nyomtatott kézírás	1	nyomtatott kézírás	0	nyomtatott kézírás	0	nyomtatott kézírás	0	nyomtatott kézírás	0	nyomtatott kézírás	0	nyomtatott kézírás
3	toll	31	toll	12	toll	94	toll	11	toll	82	toll	14	toll	95	toll

1. ábra Az első modellek

Az első modellépítések eredményei a folyamatosan bővülő szövegkorpusz miatt nem voltak teljes mértékben összehasonlíthatóak, de rengeteg tapasztalatot szereztünk, és kijelölték a további tesztelések irányát.

4.2 Második fázis

Ekkorra a Kiss József által írt levelek nagy részét feldolgoztuk, ez kb. 370 felhasználható oldalt, több mint 20.000 szót jelentett. Az elképzelés szerint egyféle anyagon többféleképpen teszteltük a modelleket.

11 Vö.: https://readcoop.eu/transkribus/resources/how-to-guides

	KJ_keze_valogatas_1	KJ_keze_valogatas_2
Base Model	kj3	nincs
Train set	aut_ts_2	aut_ts_2
pages	302	302
lines	4170	4170
words	18941	18941
Validation set	aut_vs_2	aut_vs_2
pages	33	33
lines	558	558
words	2681	2681
Eredmények		
CER on Train Set	2,81%	6,11%
CER on Validation Set	6,13%	11,49%
Nr. of Epochs	50	50
Összes fájl	335	335

	KJ_keze_osszes_1	KJ_keze_osszes_2
Base Model	kj3	nincs
Train set	aut_ts_1	aut_ts_1
pages	330	330
lines	4399	4399
words	20039	20039
Validation set	aut_vs_1	aut_vs_1
pages	36	36
lines	459	459
words	2120	2120
Eredmények		
CER on Train Set	3,05%	6,54%
CER on Validation Set	5,82%	10,70%
Nr. of Epochs	50	50
Összes fájl	366	366

2. ábra A Válogatás gyűjtemény eredményei az Összes gyűjteményhez képest

Először megbizonyosodtunk arról, hogy a tollal, ceruzával és nyomtatottan írt Kiss József levelek külön kezelését megszüntethetjük. A *Válogatás* gyűjteményben csak Kiss József kézzel írt, magyar nyelvű levelei szerepeltek, ezzel párhuzamosan egy másikban az összes általa írt levél. A válogatott gyűjteményen futtatott modell eredménye nem jobban, hanem 1–2%-kal rosszabbul teljesített a másikhoz képest (Vö.: 1. ábra CER on Validation Set sorai). A válogatás minősége fontos, mindkét set-nek megfelelően kell reprezentálnia a korpusz egészének összetételét.

Model name	Selection	Selection type	Base model	SUM	Train Set (pages)	Train Set lines	Train Set words	Validation set (pages)	Validation set lines	Validation set words	Epoch	CER on Train Set	CER on Validation Set	Collection ID	Collection name
KJ_keze_osszes_1	aut1	minden átírt oldal (10%)	Kiss József kézírása_3	366	330	4399	20039	36	459	2120	50	3,05%	5,82%	79360	KJ_keze_osszes_1
KJ_keze_osszes_2	aut1	minden átírt oldal (10%)	nincs	366	330	4399	20039	36	459	2120	50	6,54%	10,70%	79360	KJ_keze_osszes_1
KJ_keze_osszes_3	manu1	minden átírt oldal	KJ_keze_osszes_1	369	336	4413	20014	33	475	2264	50	2,26%	4,26%	79360	KJ_keze_osszes_1
KJ_keze_osszes_4	manu1	minden átírt oldal	Kiss József kézírása_3	369	336	4413	20014	33	475	2264	50	2,58%	8,38%	79360	KJ_keze_osszes_1
KJ_keze_osszes_5	manu1	minden átírt oldal	KJ_keze_osszes_1	369	336	4413	20014	33	475	2264	100	1,77%	6,08%	79360	KJ_keze_osszes_1
KJ_keze_osszes_6	aut1	minden átírt oldal (10%)	KJ_keze_osszes_1	366	330	4399	20039	36	459	2120	50	2,13%	6,59%	79360	KJ_keze_osszes_1

3. ábra Második fázisban keletkezett HTR modellek

Ebben a fázisban a Base Model (BM) használatával is kísérleteztünk, két modell között mindig csak egy dolgot megváltoztatva. A KJ_keze_osszes_4 modellnél az automatikus besorolást felváltotta egy manuális, szélesebb körű válogatás. Az eredményeken mégis az látszik, hogy az összes próbálkozás közül ennek a legmagasabb a karakter hiba értéke. Egy másik alapmodellel ugyanez az összeállítás nagyon jó eredményt hozott. Az epoch-ok számát duplázva, a KJ_keze_osszes_5 modellnél szintén rosszabb eredményeket kaptunk. És végül, a hatos modellnél az automatikus TS és VS válogatást alapul véve, egy másik alapmodell szintén rosszabb eredményeket hozott. Ezek alapján az számít, hogy melyik alapmodellt melyik korpuszon használjuk; a legjobb eredményt a hasonló adatokon tanult BM adja.

Az első két esetben a Base Modellel rendelkező modell kb. kétszer jobban teljesített. A többinél azonban már nem ennyire egyértelmű a kimenetel. Ennek oka a BM használatban rejlik. A folyamatos dokumentumfeldolgozás miatt a Base Model és a Training data (VS + TS) között átfedések keletkeztek, ami hamis pozitív eredményt és túltanulást is okozott ezeknél a modelleknél. Tehát, a második fázis legjobb eredményű modellje (4,26%) egy teljesen új próbaszövegen lefuttatva valójában rosszabbul (8–10%-os pontossággal) teljesített.

4.3 Harmadik fázis

A befolyásoltság (bias)[12] kizárására a harmadik fázisban került sor. Három, az alábbi ábrán zöld árnyalatokkal jelölt szövegkorpuszt hoztunk létre, és a Base Modell használatát egyelőre megszüntettük, azt a teljes vegyes kézírásmodell tanításakor alkalmazzuk újra.

Model name	Selection	Selection type	Train Set	Validation set	Epoch	CER on Validation Set	CER on Train Set	Collection ID	Collection name
KEZ_12_KJ_keze_osszes_7	6. adag kivételével)	manu	434	48	25	11,71%	15,09%	93264	KEZ12 Kilevéliró (6. adag kivételével)
KEZ_12_KJ_keze_osszes_8	6. adag kivételével)	manu	434	48	50	8,73%	6,62%	93264	KEZ12 Kilevéliró (6. adag kivételével)
KEZ_12_KJ_keze_osszes_9	6. adag kivételével)	manu	434	48	75	8,19%	4,38%	93264	KEZ12 Kilevéliró (6. adag kivételével)
KEZ_12_KJ_keze_osszes_10	6. adag kivételével)	manu	434	48	100	8,21%	3,45%	93264	KEZ12 Kilevéliró (6. adag kivételével)
KEZ_15_friss_kj_keziras_1	külön tesztkorpusz	manu	406	42	75	7,45%	4,35%	95196	KEZ15 Friss KJ kézírása modell
KEZ_15_friss_kj_keziras_2	külön tesztkorpusz	manu	406	42	100	7,76%	3,87%	95196	KEZ15 Friss KJ kézírása modell
KEZ_15_friss_kj_keziras_3	külön tesztkorpusz	aut	405	45	75	14,19%	3,81%	95196	KEZ15 Friss KJ kézírása modell
KEZ_15_friss_kj_keziras_4	külön tesztkorpusz	aut	405	45	75	13,81%	4,01%	95196	KEZ15 Friss KJ kézírása modell
KEZ_15_friss_kj_keziras_5	külön tesztkorpusz	aut	405	45	75	12,16%	4,24%	95196	KEZ15 Friss KJ kézírása modell
KEZ_15_friss_kj_keziras_6	külön tesztkorpusz	aut	405	45	100	13,65%	2,78%	95196	KEZ15 Friss KJ kézírása modell
KEZ17_Kiss József kézírása_1	teljes KJ kézírás (TK nélkül)	manu	455	42	75	7,39%	4,75%	95196	KEZ15 Friss KJ kézírása modell
KEZ17_Kiss József kézírása_2	teljes KJ kézírás (TK nélkül)	manu	455	42	100	7,15%	3,81%	95196	KEZ15 Friss KJ kézírása modell
KEZ17_Kiss József kézírása_3	teljes KJ kézírás (TK nélkül)	manu	455	42	125	7,11%	2,99%	95196	KEZ15 Friss KJ kézírása modell
KEZ17_Kiss József kézírása_4	teljes KJ kézírás (TK nélkül)	manu	455	42	150	7,01%	2,58%	95196	KEZ15 Friss KJ kézírása modell
KEZ17_Kiss József kézírása_5 / József Kiss' handwriting 19–20th century	teljes KJ kézírás (TK nélkül)	manu	455	42	200	6,94%	2,13%	95196	KEZ15 Friss KJ kézírása modell

4. ábra A harmadik fázisban létrehozott HTR modellek

A korpuszokat a szöveg mennyisége és a válogatás folyamata különböztette meg egymástól. Az elsőben (kez_12) a második fázisban alkalmazott felosztást és szövegkorpuszt meghagytuk, de új modellépítés történt (base modell nélkül). A legjobb hibaszázalék-eredmény 8% körüli lett. A második csoporton belül (kez_15) elkülönítettünk egy tesztkorpuszt, és a fennmaradó anyagból manuálisan alakítottunk ki TS-t és VS-t. A megfelelő mennyiségű epoch után 8% alá ment a CER értéke. Az automatikus válogatás a második csoporton csak rosszabb eredményeket hozott. A harmadik csoport (kez_17) abban tér el az eddigiektől, hogy az eredeti felosztás mellett a tesztkorpusz anyaga is bekerült a végleges TS-be. Itt értük el a legjobb, 6,94%-os eredményt, bias nélkül.

5. Javítási lehetőségek

Minél több adat áll a neurális háló rendelkezésére, annál jobb eredményt fogunk kapni. Egy bias nélküli Base Model használatával felgyorsul a training process, és kevesebb training data is elég igen jó eredmények eléréséhez. Fontos, hogy a BM szöveganyaga hasonlítson a felismertetésre váró szöveghez, és hogy minden típusú forrás reprezentálva legyen.

Szótárak beépítésével a szövegben előforduló specifikus kifejezéseket lehet megtanítani a MI segítségével, mellyel tovább lehet javítani a modell eredményességén. Ehhez fel kell venni a kapcsolatot a Transkribus csapatával.

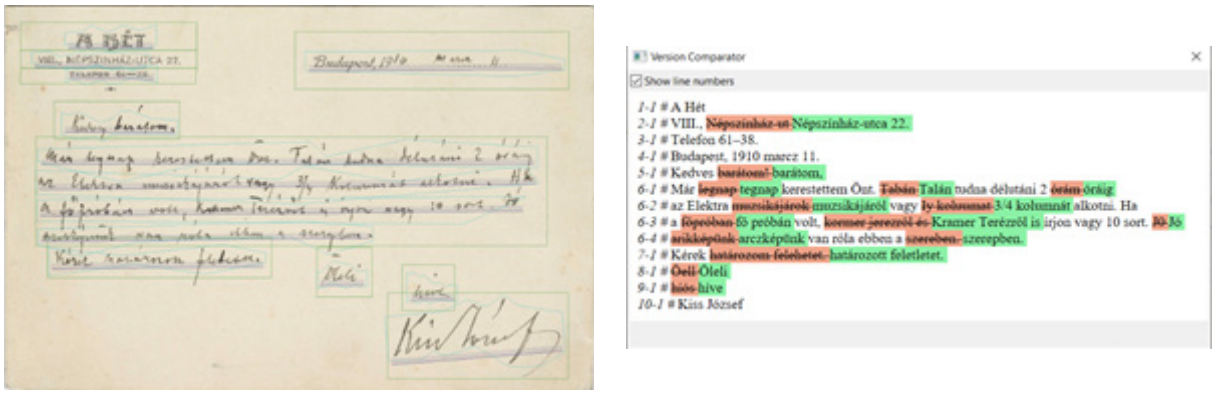
12 A statisztikai bias során az eredmények várható értéke eltér a becslés alapjául szolgáló valódi kvantitatív paramétertől. A minta torzított/elfogult, ha nem reprezentatív, ha a vizsgált tulajdonság megoszlása a mintában nem ugyanaz, mint az alapsokaságban. Ennek három oka lehet: (1) a mintavételi eljárás elfogult, (2) a minta elégtelen, (3) túl messzire megy az általánosítás. Margitay Tihamér, *Az érvelés mestersége: érvelések elemzése, értékelése és kritikája* (Budapest: Typotex, 2007) (11.5. Statisztikus következtetések és hibáik).

A túltanulás elkerülése végett az epoch-ok számát a tanulási görbe változásának figyelembevételével azon a ponton kell megállítani, ahol az a legalacsonyabban van.

Problémákat okozhat a HTR modell tanulása során a szegmentálás pontatlansága is. Érdemes tehát ellenőrizni, és manuálisan javítani a szegmentálást, figyelembe véve, hogy a rendszer a baseline és a kijelölt sor alapján próbálja megtalálni a szöveges információkat.

A különböző modellek egymásba építésé esetén pedig a TS és VS adatai nem szabad, hogy keveredjenek egymással. Így elkerülhető a fals pozitív eredmény a karakterhiba-értékben.[13]

6. Legújabb eredmények – 6,94%



5. ábra A Ground Truth-ként elmentett átírás és a HTR által elvégzett átirat közötti különbségek

A Transkribus-ban összehasonlítható az egyes szövegek különböző verziója. Látható, hogy a 6,94%-os CER esetében a tévesztések nem értelemzavaróak, többnyire gyorsan javítható ékezetekről, írásjelekről, kis-nagybetű tévesztésről van szó. A későbbiekben az 5% alatti eredmények elérése a cél.

7. Következő lépések

A Transkribus szerves részét képezi a PIM-ben őrzött kéziratos hagyatékok, eddig publikálatlan kéziratok digitális feldolgozása. Ahogy jelenleg az átírást, a jövőben az automatikus kézírás-felismertetést is szeretnénk a workflow részévé tenni. A Kiss József-projekt ideális az igények felmérésére, a fejlesztések és eszközök tesztelésére. Az egyik kimeneti formátum, melyet használunk a kétrétegű PDF. Előnye, hogy a Transkribusból közvetlenül állítható elő és azonnal publikálható a PIM online katalógus felületén (PIM OPAC).[14] A másik, még tesztelés alatt álló formátum a TEI XML, amely egy következő publikálási stádiumban lesz elérhető. Az átírás végén ez szintén exportálással nyerhető

13 Ha olyan fájlok kerülnek a VS-be, melyeken a mesterséges intelligencia a TS-en tanult, akkor A CER on Validation Set jó százalékos eredményt mutat, de az nem egyezik a modell valós teljesítményével teljesen ismeretlen szövegen.

14 A PIM Kéziratgyűjteményében található Kiss József-levelezés: https://resolver.pim.hu/gyujtemeny/levelek/"Kiss József 1843-1921".

ki a Transkribusból, viszont megfelelő megjelenítő felület fejlesztése szükséges hozzá, ahol fontos, hogy jól érvényesüljön a szöveg és a fakszimilék sor szintű összeköttetése, s az átírt szöveget olvasva látszik a kontextus is.

Megkezdődött egy vegyes kézíráson alapuló modell építése. Ezután elérjük azt a szövegmennyiséget, hogy megtörténhet az első magyar nyelvű kézírásmodell publikálása is. Hosszútávon a PIM Kézirattárában lévő egyéb írók kézírásából álló modellek kialakításával, és ezek egymásba építésével egy általánosabb modellt szeretnénk megalkotni, amely a 19–20. századi magyar kézírások automatikus felismertetését teszi majd lehetővé.

Bibliográfia

Dietrich, Felix. „OCR vs. HTR or "What Is AI, Actually?"". *READ-COOP* (blog). Elérés 2021. június 2. https://readcoop.eu/insights/ocr-vs-htr/ .

„How To Use Existing Transcriptions to Train a HTR-Model with the TextToImage-Tool". READ-COOP. Elérés 2021. június 7. https://readcoop.eu/transkribus/howto/how-to-use-existing-transcriptions-to-train-a-handwritten-text-recognition-model/ .

Katuščak, Dušan. „Automated Transcription of Handwritten Text: READ and TRANSKRIBUS (An Experiment with Transcribing Letters of Andrej Kmeť)", 2019. október 20.

Margitay, Tihamér. *Az érvelés mestersége: érvelések elemzése, értékelése és kritikája*. Budapest: Typotex, 2007.

„Public Models in Transkribus". READ-COOP. Elérés 2021. május 26. https://readcoop.eu/transkribus/public-models/.

Puigcerver, Joan. „Are Multidimensional Recurrent Layers Really Necessary for Handwritten Text Recognition?" In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 01:67–72, 2017. https://doi.org/10.1109/ICDAR.2017.20 .

Srinivasan, G. N., Shobha, G. „Segmentation Techniques for Target Recognition" INTERNATIONAL JOURNAL OF COMPUTERS AND COMMUNICATIONS, 1, 3 (2007): 75.