Hogyan tanítsunk HTR modellt a Transkribusban[1]

Bevezetés

- A **“HTR modell”** automatikus átírást készít egy dokumentum gyűjteményről, ehhez a modellt meg kell tanítani rá, hogy felismerjen bizonyos kézírási stílusokat azáltal, hogy a képet és a hozzá tartozó pontos átírást megmutatjuk neki.
- Min. 5000–15000 szót tartalmazó átírt kéziratos dokumentum kell hozzá, a nyomtatott szövegek esetén kevesebb is elengedő lehet.
- A modell tanítás funkciója, illetve gombja nem alapértelmezetten van jelen a Transkribus asztali alkalmazásban, ezt külön igényelni kell az email@transkribus.eu címen.

Előkészületek

- A forrás típusától függően (kéziratos vagy nyomtatott) ajánlott min. 5000–15000 szót tartalmazó átírt dokumentummal kezdeni a tanítást
- A HTR technológiában lévő neurális hálózat gyorsan tanul és annál jobb eredményeket produkál, minél több adat érhető el számára
- A HTR számára képek feltöltésével és átírt szöveg prezentálásával lehet adatot szolgáltatni.
- Már meglévő szövegkorpusz esetén a képekkel együtt azt is lehet modell tanításra használni.

[1] A kivonat alapja: https://transkribus.eu/wiki/images/3/34/HowToTranscribe_Train_A_Model.pdf

Betanítás

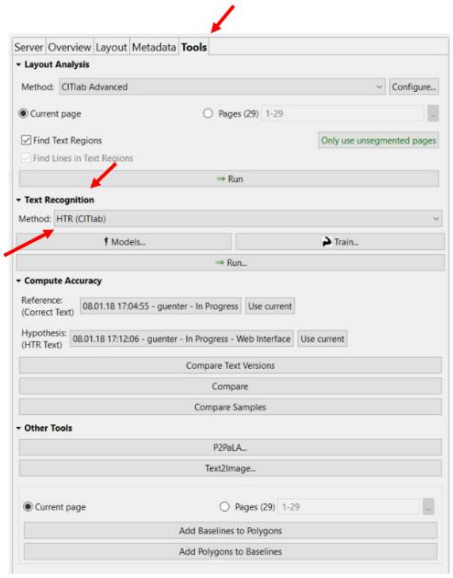

Figure 1 Where to find the tools for the training

- A modell betanításának főbb lehetőségei a **“Tools”** fülön találhatók a **“Text Recognition”** szakasz alatt.
- Módszerként (**Method**) a “**HTR (CITlab)**” a legmegfelelőbb a választható lehetőségek közül
- A **“Models”** gombra kattintva látható, hogy mely modellek érhetőek el és milyen dokumentumokon tanították be őket.
- A **“Train”** gomb megnyomásával lehet eljutni a modell betanításához

HTR+ Tanítás beállításai

- A **“Tools”** fülön lévő **“Train”** gomb megnyomásával jeleníthető meg az ezt irányító ablak
- Az ablak felső részében lehet megadni a leendő modell adatait
    - név, nyelv, leírás
    - “nr. of Epochs” arra utal, hogy hányszor lett értékelve a tanításhoz használt adat. Minél nagyobb az epoch-ok száma, az betanítás folyamata annál hosszabb ideig tart.
    - Epoch-ok szerepéről: https://rechtsprechung-im-ostseeraum.archiv.uni-greifswald.de/?s=epochs
    - Service and Tool Integration https://readcoop.eu/wp-content/uploads/2019/08/D4.6_service_integration_P3.pdf

Base Model / Alapmodell

- hozzá lehet adni egy alapmodellt a betanítási folyamathoz. Ennek hasonló írásnak kell lennie, hogy felgyorsítsa a tanítás folyamatát.
- az alapmodell lehetővé teszi azt is, hogy kisebb korpusszal is elkezdődhessen a tanítási folyamat.
- kiválasztania a “Choose” gombra kattintva lehet

Training Set / Tanuló készlet

- ki kell választani azokat az oldalakat, amelyek a készlet részét képezik majd
- a +Training gomb megnyomásával adhatók hozzá a kiválasztott gyűjtemények

Validation Set / Tesztkészlet

- néhány oldalt félretesz a program, melyeket nem használják fel a modell tanítás során. Ezekkel lehet később tesztelni a pontosságát a modellnek.
- ajánlott 50-100 oldalanként legalább egyet választani
- A választott oldalaknak reprezentatívnak kell lenniük a gyűjteményre nézve.
- Minél több oldal van a tesztkészletben, annál hosszabb ideig tart a tanítás folyamata
- a kiválasztás után a **“+Testing”** gombra kattintva lehet teszt oldalakat kivenni a gyűjteményből.

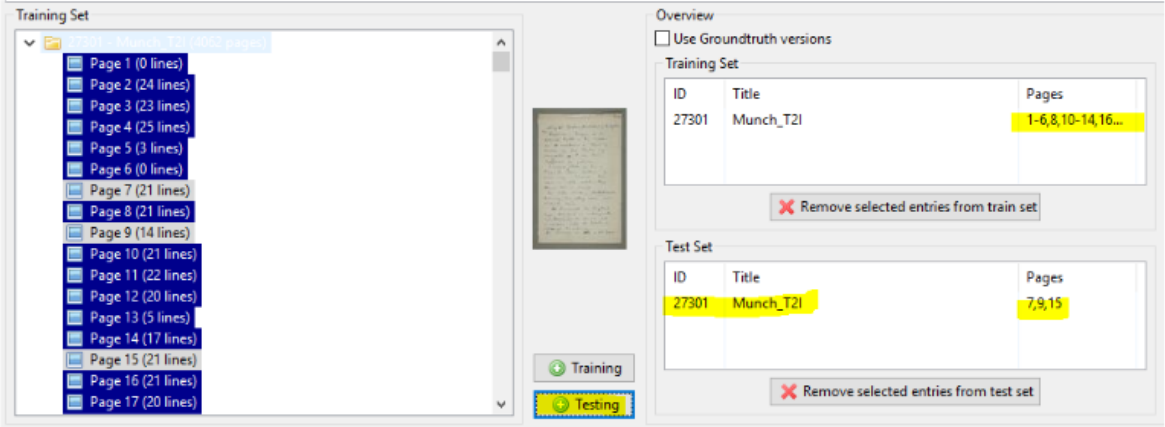

Figure 6 Adding pages to the Test set

- eltávolítani a **“x Remove…”** gombbal lehet.

A folyamat ellenőrzése

- a **“Jobs”** gombra 2a **“Server”** fülön lehet követni az eseményeket
- minden epocha befejezése és a tanítási folyamat vége is látszódni fog a “**Jobs on server**” ablakban
- A HTR+ modell tanítási folyamata néhány napot is igénybe vesz. Ezalatt más munkákat lehet végezni a Transkribus-szal és be is lehet zárni az alkalmazást a folyamat alatt.

A tanítási folyamat után

- miután a tanítás befejeződött a modell látszódni fog a gyűjtemények között
- ezt a **“Models”** gombra kattintva lehet elérni a **“Tools”** fülön.
- Az ablak bal oldalán áttekinthetőek az elérhető modelleket.
- Ablak jobb felső sarkában láthatóak a modell részletei.

- A jobb alsó sarokban látható a modell tanulási görbéje. A statisztikákkal kapcsolatos további információk az alábbiakban találhatók.

Statisztikák

- A tanulási görbe a modell pontosságát jelzi.
- CER = Character Error Rate, azaz azon karakterek százalékos aránya, amelyeket a HTR hibásan írt át.
- *Y tengely*en: **“Accuracy in CER”** a pontosságát adja meg.
- mindig 100%-ról indul a görbe és a tanulási folyamat előrehaladtával folyamatosan javul a hibák százalékos aránya.
- *X tengely* az Epoch-okat jelöli – ennek a számát a tanítási folyamat elkezdésekor lehet megadni (minél több ilyen van, annál pontosabb lehet az eredmény, de annál hosszabb a folyamat is). Illetve előfordulhat a túltanulás jelensége is.

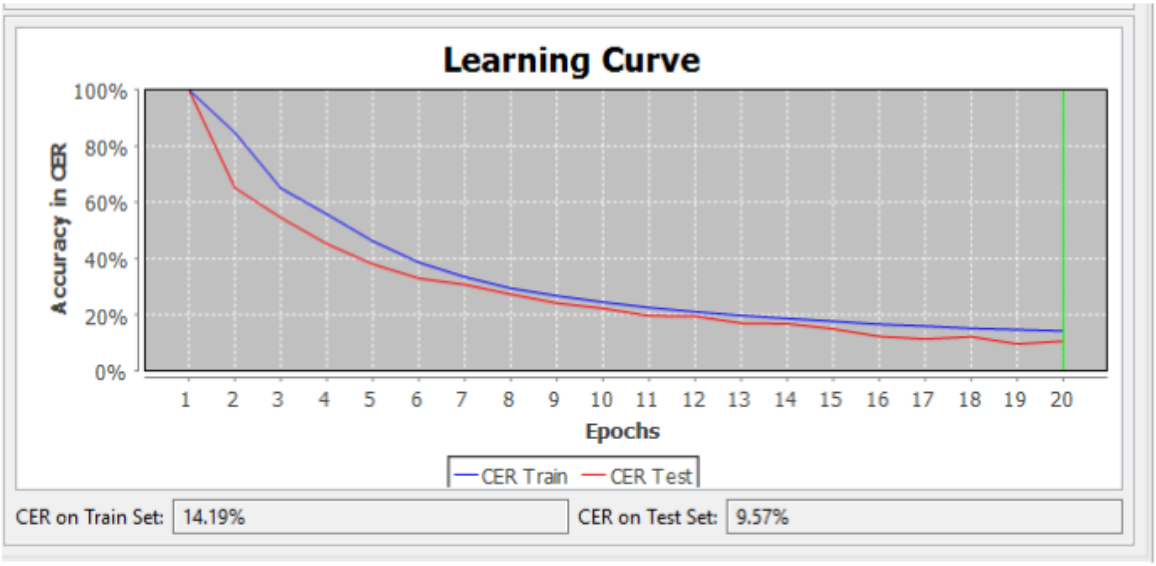

Figure 12 “Learning Curve” of your model

- **A kék vonal**: a tanulás előrehaladtát jelzi
- **A piros vonal** az értékelés előrehaladását mutatja a tesztkészleten
- A program először a tanuló készleten próbálja ki magát, ezután tér át a tesztkészletkre
- A grafikon alatt két százalékos érték látható az tanuló készlet és a tesztkészlet CER-értékére vonatkozóan.
- A 12. ábrán a modell 14,19% CER-rel teljesít az tanuló készletnél és 9,57% CER-rel a tesztkészletnél.
- A tesztkészlet jelentősége nagyobb, mert ez mutatja meg hogy teljesít a modell teljesen ismeretlen szövegen
- Eredmények:
    - A 10% -os vagy annál alacsonyabb CER érték: jónak számít
    - A 20-30% -os CER-értékkel rendelkező eredmények elegendőek a Keyword Spotting technology alkalmazásához

HTR átiratok generálása

- ha már van megfelelő modell, akkor lehet automatikusan generált átiratokat készíteni vele
- először fel kell tölteni a dokumentumot a Transkribusra
- másodszor szegmentálni kell a dokumentumot text region-okban, line és base line-okkal együtt.
- a modellt a **“Tools”** fülecskén lehet elérni a **“Text Recognition”** szekcióban.
- egy vagy akár több oldalt is ki lehet választani a HTR átirat létrehozásánál
- Run gomb megnyomásával lehet elindítani
- ha elkészült az automatikus átirat, akkor az a szövegszerkesztőben jelenik majd meg

Modell megosztása

- a HTR modell megosztható más Transkribus gyűjteményekkel is (saját vagy idegen gyűjteményekkel egyaránt)
- ez utóbbit csak az teheti meg, aki az adott gyűjtemény birtokosa /owner/
- jobb kattintás a modell névre → **“Share a model”** opció
- utána az ablakban ki kell választani a gyűjteményt és OK

Kimenet

- miután a modell tanítás befejeződött, bármelyik történelmi szövegen ki lehet próbálni azt
- meg lehet osztani másokkal is
- megismételhető a tanulási folyamat több adattal is, hogy hatékonyabb eredményeket adjon a modell
- a pontosságát a “Compute Accuracy” funkcióval lehet mérni
- a HTR eredményei attól is függenek, hogy mennyire hasonló és világos a kézírás a másik dokumentumban
- A Transkribus csapata jelenleg olyan algoritmuson dolgozik, amely lehetővé teszi bármilyen dokumentum automatikus átírását anélkül, hogy képzési adatokat kellene készítenie. A technológia a Transkribus-ban feldolgozott összes képzési adatból tanul.
- minél több adat van ehhez, annál eredményesebb lesz a végeredmény!

Egyéb

- Combining Models
- Trug und Schein: A Correspondence – A Critical Engagement with Everyday Life in the Second World War – Write with Us!