Program je zmožen popravljati določene napake, čeprav je njegova glavna pomanjkljivost trenutno, da požira prvih nekaj besed vsakega stavka. Morda problem v trainset podatkih ? Potrebno preveriti.

Lahko je tudi problem v tem, da je trainan na predolgih stavkih. Bi se omejili na max 10 besed v stavku in tudi ocr inpute na koncu formatiramo po tem ključu.

Ugotovljena potrebna formacija inputov: podatki so velikokrat včasih bili zapisani v dveh stolpcih in z deljenjem besed. To ga zmede, ker za inpute načeloma jemlje samo vrstice: POPRAVEK: zamenjamo »- \n« z »« in pazimo, da je št besed v vrstici največ 10 (Če je besed bistveno več (20+), začne model sam rezati input in obdela samo cca prvih 15 besed. Temu se hočemo izogniti.)

Morda namesto odstranitve [][] oklepajev, jih raje uporabimo za primerno mesto, da vnesemo ()() oklepaje, ki se pogosto pojavljajo tudi v originalnih tekstih iz razprav.

SKLEP 1: (9:20AM) nastavi trening z besedami: 20000 vrstic besed! (pazi na alignment…)—sledi uptrain na navadnem datasetu (po analizi je spredaj spuščena beseda le redko). Hočemo da sploh ne briše nič, ampak samo malo popravlja.

Base bistveno boljše kot small --- cer po analizi zgleda da več ne pada, ampak celo malo zraste…

KAJ NAREDIT:

Trening na besede – 2x

Trening na besede ---- še enkrat sprocesiraj trainset podatke da jih potem lahko [][]()() menjaš ---- max 15 besed na besede NE

* Najprej streniraj besedilo vs pokvarjen original, potem pa še ocr v splošnem

Trening 2000

Trening 2001

Vmesne rezultate redno pogledaj

MISTAKES\_01

Trenutno ok:

Inpute moramo spremeniti (odstranimo prelome vrstic in deljenje besed!) + naredim da je do 7 besed na vrstico

Trenutno trening rezultsi: k rok 😊

2.10.2023

Trening mistakes popravljeno () (nekej je nagajal batch size… traina 4 ure jokjok)

* Se traina sedaj – zaenkrat rezultati mistakes01 kar dobri, večinoma popravijo končni rezultat

(Ko bo to strainano več nismo daleč – trainaš male kose 2001 in pol vidimo kaj bo dalje)

Sprobamo kk to dela na real podatkih, če kaj zmanjšamo error rate.

FAZA 2: kako dela to na hrvaških, slovenskih podatkih??

Bi bilo treba naredit več trainerjev, ki potem pridejo v poštev v različnih instancah ali pa bo šlo vse na isti način?

* Najprej: damo dejanke primere čez, da vidimo, koliko je zgrešenega (če je približno enako kot za slovenske), sicer bo treba razmisliti o uptrainanju, oz posebnem modelu… gre se namreč za besedišče modela, ne toliko za slovnico (on popravlja autocorrect…)
* Preverila: že na prvi pogled obvezno trening več jezikov… poskus s križnim treningom??