ReCoRD annotálási útmutató

Vadász Noémi & Ligeti-Nagy Noémi

2021. október 28.

1. Bevezetés

A feladat célja a ReCoRD nevű korpusz magyar párjának előállítása.

A felület itt érhető el: `juniper.nytud.hu/demo/record`

2. A feladat

A feladat a következő: a cikk alján látunk egy Reference Answers részt, ahol a névelem típusa után kettősponttal elválasztva jön az aktuális bekezdésben elmaszkolt entitásunk. Két dolgot kell ellenőrizni: i) jól van-e maszkolva a szó, tehát két szóköz közötti rész van elfedve (például ha azt látjuk, hogy "ELKH," mint elmaszkolt entitás, az nem jó, csak az "ELKH" kéne ott legyen, vessző nélkül; vagy ha kimarad egy betű a maszkolásból, és [MASK]t marad a szövegben, az sem jó). Ha jó a maszkolás, akkor nincs teendő, ha rossz, akkor a jobb oldali kis ablakba be kell írni, hogy mi a baj. Ez esetben az *else* gombra kell kattintani. Ezen túl ii) ellenőrizni kell, hogy az elmaszkolt névelem egyébként szerepel-e a szövegben máshol is, ez nagyon fontos kritérium. Csak egy gyors átfutással megkeressük. Ha nem, akkor megint ezt is be kell írni a kis dobozba, hogy mi történt, és *drop*. Reményeink szerint az esetek nagy részében csak *keep* és mehet minden tovább. Ha olyat látunk, hogy Budapest-Gödöllő-[MASK] vasútvonal, vagy egyéb ilyen hosszú szörnyűség, amiből egy elem van maszkolva, akkor *drop*. Egyébként ha Fidesz-[MASK] van, és a KDNP a maszkolt elem, akkor javítjuk az egészet [MASK]-ra. Ragozással nem törődünk, az is megy bele a maszkba. (Tehát Gödöllőre az [MASK], nem pedig [MASK]re.) A fő szabály tehát, hogy szóköztől szóközig maszkolunk.

A következő linken elérhető drive dokumentumban követjük, hogy ki melyik id-jú cikkeket nézte már át: `https://docs.google.com/spreadsheets/d/1ZJvYdtp7B3LBF8Oc_FlhjpM1DOQxSrnZAZtBvHxphIg/edit?usp=sharing`. Az annotálós felületen fölül van egy kis ablak, alapértelmezetten az van beleírva, hogy 0. Ide kell beírni 0 és 88 000 között a 100 valamely többszörösét, majd a kis letöltős gombra kattintani. Ilyen módon a felületen az adott id-jú cikk jelenik meg, és onnantól egyszerre 100-at tölt be nekünk a rendszer (amelyek közt a nyilakkal lépkedhetünk). Ha végeztünk a 100 cikkel, nem fog újat adni magától, hanem egy másik id-t kell beírnunk a kis ablakba, és onnan indítanunk

újra egy *batch*-et. Mielőtt elkezdünk dolgozni, a drive dokumentumba jegyezzük fel, hogy melyik számot választjuk, nehogy ketten dolgozzanak ugyanazokon a cikkeken.