A méret a lényeg? Morfológiailag annotált korpuszok összehasonlító kiértékelése

Dömötör Andrea[1], Indig Balázs[1,2], Nemeskey Dávid Márk[1]

[1]Digitális Örökség Nemzeti Laboratórium
[2]ELTE Informatikai Kar
{domotor.andrea,nemeskey.david}@btk.elte.hu
indig.balazs@inf.elte.hu

Kivonat A korpuszok egyik legjellemzőbb tulajdonsága a méretük, második szempontként pedig az annotációk minőségét szokás említeni. Minőség alatt elsősorban konzisztenciát értünk: azt mérjük, hogyan teljesít a korpusz egy részén betanított modell a korpusz egy másik részén. Cikkünkben különböző méretű morfológiailag annotált korpuszokon (ELTE DH gold standard korpusz, NYTK-NerKor, Szeged Treebank) vizsgáltuk, hogy mekkora az *elég nagy* korpusz. Mivel a vizsgált korpuszok standard címkekészleteket követnek, kézenfekvőnek tűnt a kombinálásukkal is kísérletezni, így ugyanis a nagyobb méret mellett nagyobb műfaji változatosságot is elérhetünk. Eredményeink szerint nem a korpuszok mérete a volt döntő szempont a teljesítményben, a különböző korpuszok vegyítése pedig még rontott is az eredményeken az annotációs sémák eltérő értelmezései miatt.
Kulcsszavak: morfológia, korpuszannotáció, korpuszkiértékelés

1. Bevezetés

Már szinte közhely, hogy a gépi tanulásra épülő természetesnyelv-feldolgozó algoritumusoknak nagy mennyiségű tanítóadatra van szükségük. De mit jelent pontosan a *nagy mennyiség*? Ez nyilván nagyban függ a feladattól. Kutatásunkban a szótövesítés és a morfológiai címkézés feladatát választottuk a kérdés megvizsgálásához, mert ehhez több különböző méretű gold standard korpusz áll rendelkezésre. A korpuszok teljesítményének összehasonlításával elsősorban arra kerestük a választ, hogy mekkora korpusz elég ahhoz, hogy elfogadható eredményeket kapjunk a kérdéses feladatokon. Azaz alapvetően a méret és a teljesítmény összefüggését vizsgáltuk, nem volt célunk a korpuszok minőségének értékelése.

A korpuszméret kapcsán felmerül a különböző korpuszok együttes használatának lehetősége is. Ez korábban a névelem-felismerésben jól működő módszernek bizonyult (Simon és mtsai, 2022), így a morfológiai elemzés tanításához is megpróbáltuk kombinálni az azonos címkekészletet használó korpuszokat. Mivel ezek a címkekészletek (emMorph és UD) erősen szabványosítottak, így elvileg az ezeket használó korpuszok annotációi kompatibilisek egymással. Nem szabad azonban figyelmen kívül hagyni, hogy mind a szótövesítés, mind a morfológiai

elemzés jóval szubjektívebb feladat a névelem-annotálásnál. Emiatt erősen kérdéses, hogy az elvileg kompatibilis korpuszok vajon valóban kompatibilisek-e (és egyáltalán van-e rá szükség, hogy azok legyenek).

2. Kapcsolódó munkák

A mieinkhez hasonló kérdéseket vizsgált Straka és Straková (2017) a UDPipe 1.1-es verziójának kiértékelése kapcsán. Mivel számos nyelvre többféle UD korpusz is létezik, ezekkel külön-külön és vegyítve is végeztek tanításokat. Általánosságban arra jutottak, hogy a modellek jobb eredményt értek el, ha a tanításhoz csak egy korpuszt használtak fel, a különböző korpuszok vegyítése rontott a teljesítményen. A kisebb korpuszok esetén részletesebb kísérleteket is végeztek a tanítóadatok más korpuszokból való dúsításának lehetőségeiről. Az eredményekben csak azokat a korpuszokat jelenítették meg, ahol a más korpusz adataival való bővítés jobb teljesítményt eredményezett a függőségi annotációban. Ez összesen 12 korpuszt jelent ógörög, cseh, angol, francia, olasz, latin, szlovén és svéd nyelveken. Ebből a POS-tagelés 6, a morfológiai jegyek meghatározása 4, a szótövesítés pedig 7 esetben volt sikeresebb a bővített korpuszok esetén. A korpuszméret más forrásból való növelése tehát még a kis korpuszoknál sem bizonyult egyértelműen jó ötletnek. A szerzők ezt a különböző korpuszok annotációinak inkonzisztenciájával magyarázzák (*"the Universal Dependencies are yet not so universal as everyone would like"*).

A UD korpuszok annotációinak eltéréseit vizsgálja Wisniewski és Yvon (2019), elsősorban angol és francia nyelvű treebankeken (mivel ezekből elég sok van). A korpuszok közötti eltérések detektálására Boyd és mtsai (2008) módszerét használták, mely szerint ha két azonos szekvencia máshogy van annotálva, akkor valamelyik szekvencia vélhetően inkonzisztens. Inkonzisztenciák természetesen előfordulhatnak (és elő is fordulnak) egy korpuszon belül is, Wisniewski és Yvon (2019) eredményei szerint azonban az eltérő annotációk aránya minden esetben nagyobb volt, ha a korpuszt egy másik korpusszal hasonlították össze, mint ha saját magával. A kopruszok közötti eltérések jellemzésére egy másik kísérletet is végeztek a szerzők. Ebben egy bináris osztályozót tanítottak be, amelynek az volt a feladata, hogy eldöntse egy mondatról, hogy két korpusz közül melyikhez tartozik. Intuitív feltételezés szerint minél nagyobb lesz ennek az osztályozónak a hibaaránya, annál hasonlóbb a két korpusz. Az osztályozót betanították csak szavakra, csak POS-tagekre és szó + POS-tag kombinációkra is. A legsikeresebb osztályozást ez utóbbinál érték el, ami arra utal, hogy az azonos szavak (vagy szószekvenciák) eltérő annotációi jól jellemzik a korpuszokat, azaz a korpuszok közötti eltérések szisztematikusak.

A korpuszméret és a vele betanított modell sikerességének kapcsolatához érdemes még megemlíteni Martin és mtsai (2020) munkáját, akik CamemBERT nevű BERT alapú francia modelljüket (és összehasonlításképpen más modelleket) tanították be és értékelték ki négy különböző korpuszon. Ezek közül a legnagyobb 389363 tokent tartalmazott, a szófaji címkézésben a legjobb eredmé-

nyeket azonban a nála jóval kisebb, 68615 tokenből álló Sequoia korpusszal érték el a szerzők.

Egyrészt elmondható tehát, hogy az annotációs sémák eltérése a különböző korpuszokban más nyelveknél is ismert probléma. Másrészt feltehető az is, hogy a morfológiai címkéző modellek eredményességének nem feltétlenül az óriási korpuszméret a kulcsa.

3. Felhasznált korpuszok és eszközök

Kísérleteinkhez 3 különböző méretű kézzel annotált korpuszt használtunk fel. Ezek közül a legnagyobb a HuSpacy tanítóanyagául is szolgáló Szeged Treebank (Vincze és mtsai, 2010). Ennek teljes mérete 1 362 505 token. Az eredeti annotáció nagyrészt automatikusan lett a Universal Dependencies szabványra konvertálva[1], a UD annotációk csak a korpusz egy kis részén (42 032 token) kaptak kézi ellenőrzést.[2]

A második legnagyobb felhasznált korpusz az NYTK-NerKor (Simon és Vadász, 2021)[3], amely összesen 1 017 340 tokent tartalmaz. A korpuszhoz gold standard morfológiai annotáció készült az emMorph címkekészletével, majd ez az emtsv emmorph2ud2 (Vadász és Simon, 2019) eszközével lett UD szabványra konvertálva. A konvertált UD annotációk automatikus és kézi javításokon is átestek.

A legkisebb vizsgált korpusz az ELTE DH gold standard korpusza (K. Molnár és Dömötör, 2023)[4], amely 496 060 tokenből áll, és az NYTK-NerKorhoz hasonló módszertannal készült. Mindkét korpuszt az emtsv (Indig és mtsai, 2019) morfológiai elemző moduljaival előelemezték, majd az elemző által adott annotációkat javították kézzel az annotátorok. Az eredeti, kézzel ellenőrzött annotációk ennek megfelelően itt is az emMorph címkekészletét követik, majd ezek is az emmorph2ud2 eszközzel lettek UD annotációval kiegészítve. Ez utóbbi annotációs réteg nem kapott kézi ellenőrzést.

Mindhárom korpusz műfajilag heterogén, tartalmaznak egymással átfedő és egyedi szövegtípusokat is, így a korpuszok kombinálásával nem csak nagyobb méretet, hanem nagyobb műfaji változatosságot is elérhetünk. A korpuszokban található műfajokat az 1. táblázatban foglaltuk össze.

[1] https://github.com/huspacy/huspacy-resources/tree/master/data/processed/szeged-corpus
[2] https://github.com/UniversalDependencies/UD_Hungarian-Szeged/
[3] https://github.com/nytud/NYTK-NerKor/
[4] https://github.com/ELTE-DH/gold-standard

	ELTE DH	NYTK-NerKor	Szeged Treebank
Szépirodalom	✓	✓	✓
Tudományos-ismeretterjesztő (cikkek)	✓	(wikipedia) ✓	
Blog	✓		
Jogi	✓	✓	✓
Hírek		✓	✓
Web		✓	
Diákok fogalmazásai			✓
Informatikai			✓

1. táblázat. A korpuszok műfajai

A korpuszok teszteléséhez elsősorban a HuSpaCy (Orosz és mtsai, 2023) lemmatizáló és morfológiai elemző moduljait használtuk, mivel azonban a három vizsgált korpuszból kettő eredetileg emMorph címkekészlettel készült, így ezek összehasonlításához a PurePos (Orosz és Novák, 2013) tanításával is végeztünk kísérleteket. Ez utóbbinál lehetőség van előelemzést is végeztetni a szabályalapú emMorph (Novák és mtsai, 2016) modul segítségével.

A korpuszok train-dev-test felosztásához a HuSpaCy eredeti (Szeged Treebankből származó) tanítóanyagának arányait használtuk. A vágásnál szempont volt, hogy minden alkorpuszból nagyjából azonos arányban kerüljön szöveg a train, dev és test halmazokba, illetve hogy a bekerülő szövegek teljes mondatok legyenek. A korpuszokat először külön-külön használtuk tanításra és tesztelésre, majd megpróbáltuk kettesével kombinálni őket.

A HuSpaCy eredményei összehasonlíthatók a spaCy (Honnibal és mtsai, 2020) más nyelvű modelljeivel. A hivatalos weboldalon[5] 24 nyelv összesen 82 modelljének eredményei érhetők el. Ebből 16 nyelv esetén van adat mind a UD POS-tag, a morfológiai jegyek és a lemmatizálás kiértékelésére. A modellek átlagos teljesítménye az egyes feladatokon a 2. táblázatban látható.

POS	Morph	Lemma
0,966	0,944	0,940

2. táblázat. A spaCy különböző nyelvű modelljeinek átlagos accuracy értékei

A legjobb eredményt a spanyol, a katalán és a német modellek érték el, ezek mindhárom feladaton az átlagosnál jobban teljesítettek. A leggyengébbek (mindhárom feladaton átlag alatti) teljesítményt a finn, a litván és a görög modellek adták. Ezen kívül még az olasz és a norvég modellek eredményei mondhatók kiegyensúlyozottnak: mindkettő átlagos eredményt nyújt POS-tagelésben, a má-

[5] `https://spacy.io/models`

sik két feladaton pedig átlagnál jobbat. A többi modell sikeressége eltérő volt az egyes feladatoknál.

A 3. táblázat az előbb említett modellek eredményeit és a modell alapjául szolgáló korpuszok méretét mutatja. Az egyes nyelveknél a legjobb eredményt elérő modellt vettük figyelembe. A korpuszméret a modell tanításához és kiértékeléséhez használt teljes korpusz tokenszámát jelenti.

Nyelv	Korpuszméret	POS	Morph	Lemma
német	900 000	0,99	0,97	0,98
spanyol	500 000	0,99	0,99	0,97
katalán	500 000	0,99	0,99	0,98
norvég	311 277	0,97	0,96	0,97
olasz	278 429	0,97	0,97	0,98
finn	201 948	0,96	0,92	0,86
litván	70 047	0,95	0,89	0,86
görög	61 673	0,96	0,91	0,90

3. táblázat. A spaCy legjobb és legrosszabb eredményt elért modelljei

Itt az látszik, hogy a gyengébb teljesítményű modellek tanítókorpusza valóban jellemzően kisebb, egy bizonyos korpuszméret fölött (a 278 ezres olasz korpusztól kezdve) azonban már nincs lényeges különbség az eredmények között. Bár az eredmények pontos értelmezéséhez nyilvánvalóan figyelembe kellene venni az egyes nyelvek (morfológiai gazdagság) és korpuszok (műfaji sokszínűség) különbségeit, az általános tendencia mégis arra utal, hogy nem a méret a lényeg.

4. Eredmények

4.1. HuSpaCy

A 4. táblázatban a HuSpaCy eredményei láthatók a különböző korpuszokon és azok kombinációin tanítva. A szófaj meghatározásában (POS), a NerKor érte el a legjobb eredményt. A lemmatizálásnál nagyjából a mérettel arányosan változott a teljesítmény az egyes korpuszoknál. A morfológiai jegyek azonosításában (Feats) a Szeged Treebank jelentősen alulteljesít a másik két korpuszhoz képest. A legnagyobb eltérés a ritka jegyeknél (`Reflex, Aspect, Number[psed]`), a melléknevek fokozásánál (`Degree`) és a számjegyeknél (`NumType`) mutatkozott. Összességében azonban elmondható, hogy mindhárom korpusz eredménye eléri, sőt meghaladja a spaCy más nyelvű modelljeinek előző fejezetben ismertetett átlagos teljesítményét, és közelítik a hasonló korpuszmérettel bíró német, spanyol és katalán modellek eredményeit.

A táblázat alsó részében azt láthatjuk, hogy a különböző korpuszok kombinálása szinte minden esetben lényegesen rontott az eredményeken; egyedül az

ELTE–Szeged párosítás lemma eredménye nem rosszabb mindkét korpusz önálló teljesítményénél. A legrosszabb eredményt éppen a két nagyobb korpusz, a NerKor és a Szeged Treebank párosítása adta.

Korpusz	train	dev	test	POS	Lemma	Feats
ELTE DH	485 525	5250	5285	0,982	0,975	0,977
NerKor	997 002	10 167	10 148	0,986	0,982	0,979
Szeged	1 340 639	11 418	10 448	0,983	0,987	0,969
ELTE DH + Szeged	1 826 164	16 668	15 733	0,976	0,979	0,954
NerKor + Szeged	2 337 641	21 585	20 596	0,914	0,918	0,897

4. táblázat. A HuSpaCy eredményei különböző korpuszokon tanítva

Az eddigiek szerint tehát a kisebb korpuszok nem teljesítenek (sokkal) rosszabbul a nagyobbaknál, így azt is mértük, mi az a minimális korpuszméret, amivel még értékelhető eredményt lehet elérni. Ehhez a legkisebb (ELTE-DH) korpusz train és dev alkorpuszait vágtuk fokozatosan kisebb szeletekre (75-50-25-10-5%). A teszthalmazon nem változtattunk. Ennek eredményei az 5. táblázatban láthatók.

train	dev	POS	Lemma	Feats
485 525	5250	0,982	0,975	0,977
351 723	4101	0,977	0,970	0,975
237 287	3019	0,975	0,967	0,972
122 254	1426	0,976	0,962	0,973
47 971	877	0,965	0,930	0,958
23 540	610	0,233	0,640	0,079

5. táblázat. A HuSpaCy teljesítménye kisméretű korpuszokon

A POS-tageknél a korpusz 75%-ra (350 ezer token) csökkentése 0.5% visszaesést eredményez, a következő vágásoknál (50 és 25%) azonban nem tapasztalunk ekkora változást. A 25%-os (120 ezres) korpusz mind a POS, mind a feats esetében egy kicsivel még jobb eredményt is ad, mint az 50%-os (237 ezres) korpusz. A lemmák pontossága 25%-ig folyamatos csökkenést mutat. Nagyobb romlást mindhárom feladatnál a 10%-os (48 ezres) korpusznál látunk először, a még ennél is kisebb tanítókorpusz esetén pedig már értékelhetetlen eredményt kapunk. (A 64%-os lemma accuracy értelmezéséhez megjegyezzük, hogy a tesztkorpusz szavainak 63,4%-ára igaz, hogy a szótő megegyezik a szóalakkal.)

4.2. PurePos

Hasonló kísérleteket végeztünk a PurePos tanításával is a két emMorph annotációt tartalmazó korpusszal (ELTE DH és NerKor). Az elemzőt először az emMorph modul kikapcsolásával tanítottuk, azaz előelemzés nélkül, csak a korpuszokból kellett megtanulnia a címkekészletet. Itt is először külön-külön, majd együtt is tanítottunk a két korpusszal. A tanítási kísérleteket elvégeztük az ELTE DH korpusz kisebbre vágott változataival is. Ezek eredményei a 6. és 7. táblázatban láthatók. A 6. táblázatban azért szerepel külön oszlopban a UD és az emMorph lemma, mert a NerKorban kétféle lemma érték is tartozik a szavakhoz: az eredeti (emMorph) lemmákat a UD-re konvertálás során a UD sémájára igazították a korpusz készítői. Kísérleteinkben ezért mindkét lemmával elvégeztük a tanítást.

Korpusz	**train**	**test**	**Tag**	**Lemma (UD)**	**Lemma (emMorph)**
ELTE DH	485 525	10 535	0,948	0,925	0,925
NYTK-NerKor	997 002	20 315	0,948	0,940	0,925
ELTE + NerKor	1 482 527	30 850	0,942	0,923	0,919

6. táblázat. A PurePos eredményei különböző korpuszokon tanítva emMorph nélkül

A két korpusz eredményeit összehasonlítva azt látjuk, hogy a címkézési feladatnál különböző méretük ellenére is azonos teljesítményt nyújtott a két korpusz. A lemmatizálásnál a NerKor UD lemmái könnyebben tanulhatónak bizonyultak, mint az eredeti lemmák, ez utóbbiak szintén azonos eredményt mutatnak az ELTE DH korpusszal. Itt is igaz, hogy a két korpusz kombinálása nemhogy nem javított, de még rontott is az eredményeken.

Train méret	**Tag accuracy**	**Lemma accuracy**
485 525	0,948	0,925
351 723	0,947	0,927
237 287	0,946	0,927
122 254	0,946	0,938
47 971	0,925	0,935
23 540	0,894	0,912
11 951	0,889	0,908
5863	0,836	0,872

7. táblázat. A PurePos teljesítménye kisméretű korpuszokon emMorph nélkül

A 7. táblázatban látható eredmények szerint a PurePos a HuSpaCy-nél kevésbé érzékeny a korpuszméretre. (Ami nem meglepő, hiszen a PurePosnak ki-

fejezett célja volt, hogy kis korpuszokon is képes legyen viszonylag megbízható eredményt adni.) A lemmatizálásban még egészen kicsi, alig 12 ezer token méretű tanítókorpusszal is elérte a 90%-ot. Érdekes, hogy a 120 és a 48 ezres korpuszoknál a lemmatizálás eredményessége átmenetileg még javult is, ez valószínűleg a részkorpuszok sajátosságai miatt lehet. A tagelésnél itt is a 48 ezres korpusznál látjuk az első nagyobb visszaesést, de ennek az eredményei is sokáig 90% közelében maradnak.

A 8. és 9. táblázatban ugyanezeket a tanításokat látjuk, de ezúttal használtuk a beépített emMorph előelemző modult, azaz a korpuszokból betanított modellnek már csak az egyértelműsítés volt a feladata. Ehhez referenciának érdemes megvizsgálni, hogy a szavak mekkora része eleve egyértelmű. Ez legkönnyebben az ELTE DH korpusz xml változatán volt mérhető, ez ugyanis tartalmazza az összes alternatív emtsv elemzést. Eszerint a szavak közel felénél (45,7%) mind a szótő, mind az elemzés egyértelmű, így nem meglepőek az előzőeknél még erősebb eredmények. A legjobban az ELTE DH korpusza teljesített. A NerKor alulmaradása valószínűleg annak tudható be, hogy a korpusz készítői vélhetően többször adtak meg saját elemzést az emtsv által felajánlottak helyett.

Korpusz	**train**	**test**	**Tag**	**Lemma (UD)**	**Lemma (emMorph)**
ELTE DH	485 525	10 535	0,963	0,982	0,982
NYTK-NerKor	997 002	20 315	0,936	0,948	0,954
ELTE + NerKor	1 482 527	30 850	0,942	0,958	0,958

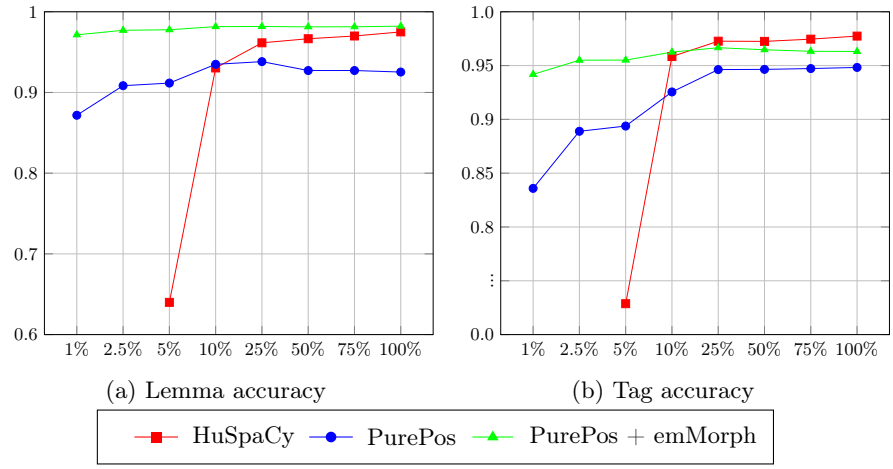
8. táblázat. A PurePos eredményei különböző korpuszokon emMorph előelemzéssel

A korpuszméret csökkentése itt – nem meglepő módon – még kevésbé befolyásolta az eredményeket: a korpusz csupán 10%-át (48 ezer tokent) használva is ugyanolyan eredményeket kaptunk, mint a teljes korpuszon tanítva.

Train méret	**Tag accuracy**	**Lemma accuracy**
485 525	0,963	0,982
351 723	0,963	0,981
237 287	0,965	0,981
122 254	0,967	0,982
47 971	0,963	0,982
23 540	0,955	0,978
11 951	0,955	0,977
5863	0,942	0,971

9. táblázat. A PurePos teljesítménye kisméretű korpuszokon emMorph előelemzéssel

Az 1. ábrán még jobban látszik a neurális (HuSpaCy) és a HMM (PurePos) rendszerek eltérő karakterisztikája. A HuSpaCy eredménye mind a szótövesítésben, mind a tagelésben meredeken zuhan a 10%-osra redukált korpuszméret alatt, míg a PurePos teljesítménye jóval visszafogottabb ütemben csökken. (A HuSpaCy esetén az (1b) ábrán a feats eredményeket jelenítettük meg, mert nehézségben ez a leginkább összevethető az emMorph címkézéssel.) Az emMorph előelemzővel az esés mindkét feladaton még lassabb, a szótövesítés pontossága pedig végig a HuSpaCy-é felett marad.



(a) Lemma accuracy (b) Tag accuracy

1. ábra. A lemma- és POS-tag pontosság alakulása a korpuszméret függvényében

5. Diszkusszió

Az eredmények tükrében visszatérhetünk az eredeti kérdéshez: mennyire számít a méret? Úgy tűnik, hogy a neurális háló alapú rendszereknél, amilyen a HuSpaCy, számít, de ott sem mindenek felett. A kisebb korpuszok nemcsak vesenyképesnek bizonyultak a legnagyobb mérettel bíró Szeged Treebankkel szemben, de bizonyos feladatokban meg is előzték. Azt mondhatjuk tehát, hogy nagyjából félmillió tokentől (ami az ELTE DH korpusz teljes mérete) kezdve már nem feltétlenül számít a méret. Az ennél kisebb részkorpuszokon azonban egyértelmű volt a teljesítménycsökkenés, különösen a lemmatizálásnál. A POS-tagek és a morfológiai jegyek pontossága viszont meglepően kis korpuszméretnél (120 ezer token) is a "vállalható" tartományban maradt.

Összességében tehát úgy tűnik, hogy mindhárom vizsgált korpusz mérete megfelelő a HuSpaCy tanításához, a PurePosnak pedig még kisebb is elég lenne, különösen ha a címkekészlet lehetővé teszi az emMorph előelemzésének használatát. Ez azt is megmutatja, hogy a neurális hálók és a mesterséges intelligencia

térhódítása mellett is lehet helye a szabályalapú moduloknak: az emMorph és a PurePos éppen a HuSpaCy számára leginkább kihívást jelentő lemmatizálásban volt különösen sikeres a kisebb korpuszokon. Az, hogy a tagelési feladatban viszont (bár abban is jó) nem éri el a lemmatizáló szintjét amiatt lehet, hogy a szófaj kérdése sokszor a szakértő emberi annotátorok számára sem egyértelmű. Mind a NerKorban, mind az ELTE DH korpuszban gyakori tévesztés volt a melléknevek és főnevek keverése, ami a nyelvészeknek is "szürke zónának" számít.

További érdekes és hasznos tanulság, hogy bármennyire is csábító gondolat a nagyobb méret elérése érdekében együtt használni a meglévő korpuszokat, ez sajnos minden esetben rossz ötletnek bizonyult. Úgy tűnik, hogy a lemmatizálás és morfológiai elemzés túlságosan sokrétű feladat ahhoz, hogy bármely két műhely egységesen tudjon annotálni, még akkor is, ha azonos sémát próbálnak követni. A korpuszok közötti eltérések részletes vizsgálata egy külön tanulmányt érdemelne, itt csak néhány különbséget említünk meg, amelyek vagy már a korpuszok leírásából kiderülnek, vagy az általunk végzett mérések eredményeiből nagyon szembetűnőek voltak.

A NerKor például sajátos elemzést alkalmaz a többtagú tulajdonnevekre. Ezek belső tagjai nem kapnak esetragot, csak sima `[/N]` taggel annotálják őket, a szótő pedig azonos a szóalakkal. Ez a UD konverzióba nem került át, így csak a NerKor és az ELTE DH kompatibilitását érinti (amennyiben az emMorph annotációkat használjuk). További különbség a két kisebb korpusz között, hogy a NerKor megkülönbözteti a melléknévi igeneveket a melléknevektől, előbbiekre a `[/V][_ImpfPtcp/Adj]`, `[/V][_PerfPtcp/Adj]` és `[/V][_ModPtcp/Adj]` címkéket használja, míg az ELTE DH korpuszban ez a különbség csak a részletes elemzésben jelenik meg, az egyszerű címke minden esetben `[/Adj]`. Ez érinti a UD változatot is, a NerKor a melléknévi igeneveknél használ egy extra `VerbForm` feature-t, ami se az ELTE DH, se a Szeged Korpuszban nem szerepel, ezekben a melléknévi igenevek annotációja megegyezik a sima melléknevekével. Egyébként pont ezek az eltérések okozzák a NerKor gyengébb teljesítményét a PurePos+emMorph tanításában, az emMorph ugyanis nem ad sima `[/N]` illetve igenévi címkéket, így ezek okozták a legtöbb tévesztést ennél a modellnél.

Az ELTE DH korpusz sajátossága, hogy a ragozott névmások (pl. *róla*) esetén a szótőben különbséget tesz a személyek (*ő*) és a nem személyek (*az*) között, míg a másik két korpusz nem használja az *az* szótövet ezekben az esetekben. A NerKor dokumentációja továbbá megjegyzi, hogy a Szeged Treebank nem mindig következetes a névmások szótövesítésében, és az sem könnyítette meg a szabályok egyértelmű meghatározását, hogy néhány ragozott alakra nincs példa a korpuszban.

Végül meg kell említeni a segédigék kérdéskörét, amit szintén eltérően kezelnek a korpuszok. A Szeged Treebankben `AUX` címkét kap a múlt idejű feltételes mód (*volna*) és a jövő idő (*fog*) segédigéje, valamint a kopula. Az emMorph címkekészletben azonban nincs segédige kategória, így az emmorph2ud2 konvertáló eszközbe sem került bele. A NerKor a sima `[/V]` címkével különbözteti meg a segédigét a többi igétől, ezt azonban csak a *volna* esetében használják (ami érthető, hiszen a jövő időt jelző *fog*-nak és a kopuláknak van ragozása). Az ELTE DH

korpusz UD változatában viszont (mivel az az emmorph2ud2 eszközzel készült) egyáltalán nem szerepel az `AUX` címke.

6. Összegzés

Összefoglalva, a morfológiai annotálás tanítása a tapasztalataink szerint nem feltétlenül igényel óriási tanítókorpuszt. A méretnél lényegesebb szempont az annotációk konzisztenciája. A korpuszok kombinálásával kapott eredmények azt mutatták, hogy már az annotációs sémák kis eltérései is komoly nehézséget tudnak okozni a tanuló algoritumusoknak.

Mindebből jogosan következne az igény a különböző korpuszok annotációs sémáinak egységesítésére, különösen egy olyan nemzetközi szabvány esetén, mint a Universal Dependencies. Ugyanakkor azt is láttuk, hogy nincs feltétlenül szükség a meglévőknél nagyobb annotált korpuszokra, így az is elfogadható irány lehet, ha hagyjuk, hogy az egyes korpuszok megőrizzék az "egyéniségüket", aminek az emMorph elemzési réteg jó tere lehet.

Köszönetnyilvánítás

A kutatást a Digitális Örökség Nemzeti Laboratórium támogatta.

Hivatkozások

Boyd, A., Dickinson, M., Meurers, D.: On detecting errors in dependency treebanks. Research on Language and Computation 6, 113–137 (10 2008)

Honnibal, M., Montani, I., Van Landeghem, S., Boyd, A.: spaCy: Industrial-strength Natural Language Processing in Python (2020)

Indig, B., Sass, B., Simon, E., Mittelholcz, I., Kundráth, P., Vadász, N.: emtsv – Egy formátum mind felett. In: XV. Magyar Számítógépes Nyelvészeti Konferencia. pp. 235–247. Szegedi Tudományegyetem Informatikai Intézet, Szeged (2019)

K. Molnár, E., Dömötör, A.: Gondolatok a gondola-tokról. Morfológiai annotációt javító módszerek tesztelése gold standard korpuszon. In: Berend, G., Gosztolya, G., Vincze, V. (szerk.) XIX. Magyar Számítógépes Nyelvészeti Konferencia. pp. 341–356. Szeged (2023)

Martin, L., Muller, B., Ortiz Suárez, P.J., Dupont, Y., Romary, L., de la Clergerie, É., Seddah, D., Sagot, B.: CamemBERT: a tasty French language model. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J. (szerk.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 7203–7219. Association for Computational Linguistics, Online (Jul 2020), https://aclanthology.org/2020.acl-main.645/

Novák, A., Siklósi, B., Oravecz, Cs.: A New Integrated Open-source Morphological Analyzer for Hungarian. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). pp. 1315—-1322. Portorož (2016)

Orosz, Gy., Szabó, G., Berkecz, P., Szántó, Zs., Farkas, R.: Advancing Hungarian Text Processing with HuSpaCy: Efficient and Accurate NLP Pipelines. In: Ekštein, K., Pártl, F., Konopík, M. (szerk.) Text, Speech, and Dialogue. pp. 58–69. Springer Nature Switzerland, Cham (2023)

Orosz, Gy., Novák, A.: PurePos 2.0: a hybrid tool for morphological disambiguation. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2013). pp. 539–545. INCOMA Ltd. Shoumen, Hissar, Bulgaria (2013)

Simon, E., Vadász, N.: Introducing NYTK-NerKor, A Gold Standard Hungarian Named Entity Annotated Corpus. In: Ekstein, K., Pártl, F., Konopík, M. (szerk.) Text, Speech, and Dialogue - 24th International Conference, TSD 2021, Olomouc, Czech Republic, September 6-9, 2021, Proceedings. Lecture Notes in Computer Science, vol. 12848, pp. 222–234. Springer (2021)

Simon, E., Vadász, N., Lévai, D., Nemeskey, D.M., Orosz, Gy., Szántó, Zs.: Az NYTK-NerKor több szempontú kiértékelése. In: Berend, G., Gosztolya, G., Vincze, V. (szerk.) XXVIII. Magyar Számítógépes Nyelvészeti Konferencia. pp. 403–416. Szegedi Tudományegyetem TTIK, Informatikai Intézet, Szeged (2022)

Straka, M., Straková, J.: Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In: Hajič, J., Zeman, D. (szerk.) Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. pp. 88–99. Association for Computational Linguistics, Vancouver, Canada (Aug 2017), https://aclanthology.org/K17-3009/

Vadász, N., Simon, E.: Konverterek magyar morfológiai címkekészletek között. In: Berend, G., Gosztolya, G., Vincze, V. (szerk.) XV. Magyar Számítógépes Nyelvészeti Konferencia. pp. 99–111. Szegedi Tudományegyetem, Informatikai Intézet, Szeged (2019)

Vincze, V., Szauter, D., Almási, A., Móra, Gy., Alexin, Z., Csirik, J.: Hungarian Dependency Treebank. In: Proceedings of LREC 2010. ELRA, Valletta, Malta (May 2010)

Wisniewski, G., Yvon, F.: How Bad are PoS Tagger in Cross-Corpora Settings? Evaluating Annotation Divergence in the UD Project. In: Burstein, J., Doran, C., Solorio, T. (szerk.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 218–227. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019), https://aclanthology.org/N19-1019/