A pragmatikai annotáció kontextusfüggősége nagy nyelvi modell esetében – felszólító alakok funkcióinak annotálása huBert modellel

Szécsényi Tibor[1], Virág Nándor[2]

[1] SZTE BTK Általános Nyelvészeti Tanszék,
MTA–SZTE–DE Elméleti Nyelvészeti és Informatikai Kutatócsoport
szecsenyi.tibor@szte.hu
[2] SZTE Nyelvtudományi Doktori Iskola,
MTA–SZTE–DE Elméleti Nyelvészeti és Informatikai Kutatócsoport
virag.nandor9910@gmail.com

Kivonat: Természetes nyelvi szövegekben a pragmatikai jellemzők annotálása során nagy kontextust kell figyelembe venni. Tanulmányunkban azt vizsgáljuk, hogy egy pragmatikai jellemző automatikus annotálása során milyen hatással van a rendelkezésre álló kontextus nagyságának a változása. A vizsgálat egy olyan korpuszon történt, melyen rendelkezésre állt egy manuálisan rögzített pragmatikai annotáció a felszólító alakok funkcióinak kategorizálására. A korpusz segítségével a huBert base nagy nyelvi modellt finomhangoltuk az annotációs feladat elvégzésére, és megvizsgáltuk, hogy a modell az annotálás során milyen mértékben támaszkodik a rendelkezésre álló kontextusra. A kontextuális hatást kétféleképpen vizsgáltuk: a tanítószekvenciák hosszának változtatásával hogyan befolyásolja a modell összesített pontosságát; illetve az annotálandó elemek szekvencián belüli elhelyezkedése milyen hatással van az annotáció pontosságára.

1 Bevezetés

Természetes nyelvi szövegekben az elemek nyelvi jellemzőinek meghatározása során a jellemzőtől függő mértékben kell különböző nagyságú kontextust figyelembe venni. Morfológiai jegyek vagy a POS-tagek meghatározásánál elegendő néhány szavas kontextus ismerete, szintaktikai jellemzőknél a mondatnyi kontextus ismeret lehet szükséges, szemantikai jegyeknél viszont már egy-két mondatnyi kontextusra és szükség lehet. Pragmatikai jellemzők esetében viszont sokszor a teljes szöveget figyelembe kell venni a nyelvi elemek helyes annotálásához. Tanulmányunkban azt vizsgáljuk, hogy egy pragmatikai jellemző automatikus annotálása során milyen hatással van a rendelkezésre álló kontextus nagyságának a változása.

Vizsgálatunkat a MedCollect Egészségügyi álhírkorpuszon (Németh T., 2023; Szécsényi és mtsai, 2024) végeztük, melyen rendelkezésre állt egy korábban kézi annotációval rögzített pragmatikai annotáció, a felszólító alakok funkcióinak azonosítása (Szécsényi és mtsai, 2024). A korpusz segítségével a huBert base nagy nyelvi modellt (Nemeskey, 2020; 2021) finomhangoltuk az annotációs feladat elvégzésére. A tanítás elsődleges célja az volt, hogy a manuális annotálást megközelítő pontosságú automa-

tikus annotáló eszközt hozzunk létre, közvetett, de jelen tanulmány szempontjából fontosabb célja pedig az, hogy megvizsgáljuk, hogy a modell az annotálás során milyen mértékben támaszkodik a rendelkezésre álló kontextusra. Az a kiinduló hipotézisünk, hogy a megfelelő megbízhatósággal működő gépi annotálás során is jobb eredmény érhető el, ha az elemző eszköz számára nagyobb kontextus hozzáférhető.

A kontextuális hatást kétféleképpen vizsgáltuk. Egyrészt azt néztük meg, hogy a tanítás és a teszt során használt szekvenciák hossza befolyásolja-e az annotálás pontosságát: elvárásunk szerint a hosszabb tanító és teszt szekvenciák esetében nagyobb az annotálás pontossága. Másrészt azt vizsgáltuk, hogy az annotálandó kifejezéseknek a szekvencián belüli elhelyezkedése miként befolyásolja a kifejezés annotációjának a pontosságát. Elvárásunk szerint a szekvenciák elején és végén kisebb a pontosság a szekvencia közepéhez viszonyítva, továbbá a szekvencia belsejében a szekvencia elejétől mért távolsággal növekszik a pontosság, vagyis a baloldali kontextus nagyobb hatással van a felszólító alakok funkcióinak a meghatározásánál.

A következő szakaszban bemutatjuk a korpusz annotációs sémáját, a manuális annotálás jellemzőit, valamint a manuális annotálás pontosságát befolyásoló tényezőket. A 3. szakaszban a huBert base finomhangolásának részleteit tárgyaljuk. A 4. szakasz különböző szekvenciahosszúságú tanítóadatokkal finomhangolt eszköz annotációjának a pontosságát, fedését és F1 értékeit hasonlítjuk össze, amit az 5. szakaszban a felszólító alakok teszt szekvencián belül elfoglalt pozíciójának a feltáró elemezés követ.

2 A felszólító alakok funkcióinak kézi annotálása a MedCollect korpuszban

A felszólítás (direktíva) a nyelvi manipuláció eszköze, amellyel a beszédpartnert vagy a szöveg olvasóját próbáljuk meg rávenni egy olyan tevékenység jövőbeni elvégzésére, amelyet egyébként nem feltétlenül tenne meg. A felszólítás legközvetlenebbül felszólító alakú igét tartalmazó megnyilatkozásokkal történhet, de megvalósulhat indirekt módon is, például kérdő mondattal: *Ide tudnád adni a sót?*

A felszólító alakoknak azonban nem csak a felszólítás kifejezése lehet a funkciójuk. A MedCollect korpusz annotációja során a felszólító alakok különböző funkcióban való használatának azonosítása volt a cél, az annotálás során a felszólító alakokhoz egy-egy funkció lett rendelve értékként. Vegyük sorra ezeket az értékeket a korpuszban való előfordulásuk gyakorisága szerint!

`Nodirectiva` – olyan felszólító alakok, amelyek nem hajtanak végre felszólítást, és a többi nem felszólító funkciót sem kapják meg. Ezek leggyakrabban kötőmódban álló igék, pl. *Belefáradtam, hogy állandóan maszkot **viseljek**.*

`Saját hangú` – olyan felszólítás, amelyben a szöveg alkotója szólítja fel a szöveg olvasóját valamire, pl. *Mindenki **viseljen** maszkot!*

`Közvetített` – olyan felszólítás, amely a szöveg olvasójára irányul, de a felszólítás forrása nem közvetlenül a szövegalkotó, hanem valaki más, a szövegalkotó csak közvetíti azt, de egyet is ért vele. Pl. *A szakértők szerint is **viseljünk** maszkot*. A saját hangú és a közvetített felszólítások valódi felszólítások.

`Meta` – olyan felszólítás, amit valaki intézett valaki felé, de az nem a szöveg olvasója (*Az USA elnöke felszólította a New York-iakat, hogy **viseljenek** maszkot.*), vagy ha igen, akkor a szövegalkotó nem ért vele egyet (*Az oltásellenesek azt akarják, hogy ne **viseljünk** maszkot.*). Ezekben az esetekben a szöveg szerzője csak beszámol egy felszólításról.

`Szövegszervező` – lehet ugyan valamilyen felszólító erejük, elsődleges funkciójuk azonban a figyelemirányítás, a szöveg koherenciájának megteremtése, pl. ***Lásd** a következő példát!*

`Interakciós` – olyan felszólító alak, amelyek a beszélőnek vagy a hallgatónak a szöveg megértését segítő tevékenységgel, vagy a beszélő és a hallgató közötti viszonnyal kapcsolatos, pl. ***Gondoljunk** csak bele! Hogy őszinte **legyek**...*

`Ambiguous` – olyan felszólító alak, amely a fentebb megadottak közül több funkcióval is rendelkezhet.

A korpusz maga 707 darab online elérhető nyilvános weboldalról származó hírszerű egészségügyi témájú szöveget tartalmazott, a korpusz összterjedelme kb. 370 000 token. A kézi annotálás során 2664 felszólító alak került címkézésre. Az annotálásban 22 annotátor és 2 kurátor vett részt, az annotátorok összmunkaideje kb. 400 óra volt. A kézi annotálás során a felszólító alakok funkcióinak megállapításán kívül még két másik jegy is jelölve lett (`source`: a felszólítás forrása; `target`: a felszólítás célzottja), de ezek a gépi tanulási kísérletek során nem lettek figyelembe véve.

A felszólító alakok funkcióinak meghatározása során az annotátoroknak olyan döntést kellett végrehajtaniuk, ami nem kizárólag az annotálandó elemek és közvetlen szövegkörnyezetük formai jellemzőire támaszkodik, hanem sokszor a tágabb kontextust és az olvasói háttértudást (világtudás) is felhasználó következtetéseket is figyelembe kellett venni (l. pl. Archer és mtsai, 2009; Milà-Garcia, 2018). A tágabb kontextus hatása például akkor jelentkezik, amikor egy hosszabb szövegben a felszólító alak megjelenési helyénél akár bekezdésekkel korábban jelzi a szöveg, hogy valakinek a véleményét idézi (például egy interjúban), és ezért a felszólító alakot `meta` vagy `közvetített` címkével kell ellátni. Annak eldöntéséhez pedig, hogy a két címke közül melyik a helyes, az olvasó (annotátor) háttértudására alapuló következtetés szükséges, ti. hogy a szöveg alkotója vajon egyetért-e a felszólítással (`közvetített`), vagy sem (`meta`). A felszólító alakok funkcióinak meghatározása során további tényezőként jelenik még meg az annotátorok egyéni különbségei is: a felszólítás erősségének, vagy akár a felszólítás meglétének a megítélése is eltérő lehet annotátoronként.

A kontextus hatása, a háttértudás felhasználásának szükségessége és az annotátorok egyéni különbségei miatt a pragmatikai jegyek annotálása során sokkal kisebb az annotátorok közötti egyetértés mértéke, mint például a morfológiai vagy szintaktikai jellemzők megállapítása során. Azonban az annotátoroknak a kurátor által elfogadottól eltérő annotációja nem tekinthető feltétlenül annotálási hibának: az eltérések egy része tényleges hiba (pl. figyelmetlenség), más része viszont a lehetséges eltérő megítélésből fakad.

A felszólító alakok funkcióinak annotálása során az annotátorok átlagosan 0,824 pontossággal és 0,846 fedéssel dolgoztak, az F1 értékek átlaga pedig 0,830 volt. A megbízhatósági értékek számításánál a kurátor által elfogadott annotációkat tekintet-

tük helyesnek. Az 1. Táblázat azt mutatja, hogy az annotátorok átlagosan milyen pontossággal, fedéssel és F1 értékkel tudták az egyes jegyeket azonosítani.

1. Táblázat: A manuális felszólításannotálás megbízhatósági értékei jegyenként és összesítve

	Pontosság	**Fedés**	**F1**
nodirectiva	0,83	0,88	0,85
sajat_hangu	0,91	0,90	0,90
kozvetitett	0,63	0,60	0,60
meta	0,61	0,58	0,58
szovegszervezo	0,75	0,82	0,76
interakcios	0,53	0,58	0,55
ambiguous	0,32	0,24	0,27
összesítve	0,82	0,85	0,83

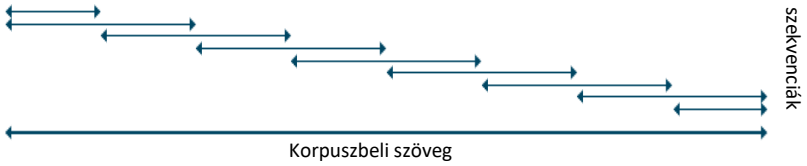
A rendelkezésre álló kézzel annotált korpusz segítségével megpróbáltunk egy nagy nyelvi modellt finomhangolni az annotációs feladat elvégzésére. Célkitűzésünk az volt, hogy egy olyan eszközt hozzunk létre, amely a manuális annotátorokhoz közelítő megbízhatósággal működik. Mivel a manuális annotáció megbízhatóságát befolyásoló egyik tényező a kontextus figyelembevétele volt, ezért az eszközön megvizsgáltuk a kontextus hatását is, ezt mutatjuk be a 4. és az 5. részben.

3 A huBert base finomhangolása

Az automatikus annotáció megoldásának fontossága kettős: egyrészt a tanítókorpuszként használt annotált korpusz bővítése során alkalmazható előannotálásra, vagy az annotátorok kiváltására (a kurátori döntést meghagyva), másrészt pedig közvetlenül használhatjuk a felszólítás funkcióit azonosítani képes modellt olyan alkalmazásokban, amelyekben szükség van ilyen információra. Ha az erre a feladatra finomhangolt nagy nyelvi modellt be akarjuk vetni éles terepen, számot kell adni azokról a potenciális problémákról, amelyekkel a humán annotátorok is szembesültek, konkrétan a kontextuális hatásról és a háttértudásról az automatikus annotáció során.

A kettő közül a háttértudás szerepének értelmezése egyértelműbb. Egy nagy nyelvi modell nem rendelkezik azzal a világról alkotott mögöttes tudással, amivel a humán elemző. Ennek a hiánya azonban nem feltétlenül jelentős, hiszen az emberi befogadó számára se feltétlenül szükséges a világról alkotott kép bevonása a felszólítások értelmezésében, mindössze támpont lehet a helyes kategorizáláshoz. Sokkal fontosabb a kontextuális tényezők felmérésére való képesség, ahogy azt korábban is említettük. A felszólítások funkcióinak azonosítása során a humán befogadók támaszkodnak a szövegkörnyezetre is, ezért felmerül a kérdés a nagy nyelvi modellek alkalmazása esetében, hogy „figyelembe veszik”-e megfelelő módon a rendelkezésre álló kontextust, ahhoz, hogy ez segítsen az annotálási feladatban. Ha a válasz igen, akkor pedig mennyire komoly szerepe van a kontextus nagyságának a modell megbízhatóságának mértékében?

A feladat elvégzéséhez a huBert base modellt (Nemeskey, 2020; 2021) használtuk, amely egy, a Common Crawl adatbázis magyar részkorpuszán és a Wikipédia magyar nyelvű szövegein tanított Bert modell. Ennek a finomhangolásához használtuk fel a MedCollect korpusz szövegeit. Ahhoz, hogy a kontextus méretéből adódó különbségeket tesztelni lehessen, eltérő tokenhosszúságú (64, 128, 256 és 512) szekvenciákra szeletelve kapta meg a modell az egyes verziók tanítása során. Mivel a pragmatikai annotálásnál releváns kontextuális információkat találhatunk a teljes szövegben is, ezért a modellt nem mondatokra szegmentált tanító adatokkal láttuk el, hanem a korpuszban található teljes szövegek adott tokenhosszúságú részeivel. Annak érdekében, hogy tesztelhessük a szekvencián belüli pozíció jelentőségét is, meg kellett oldanunk, hogy minden annotálandó szónak legyen kellő nagyságú bal- és jobboldali kontextusa, ezért a szegmentumokat átfedéssel, eltolással állítottuk elő. Ennek sematikus ábrázolása látható az 1. ábrán.



1. ábra. A korpusz szövegeinek szekvenciákra osztása 50%-os átfedéssel.

Az egymást követő szekvenciák közötti átfedés 50 százalékos volt, tehát például, ha 128 tokenes szegmentumokkal dolgoztunk, akkor az első szekvencia második 64 eleme megegyezett a következő szekvencia első 64 elemével. Ezáltal, ha egy annotálandó elem a szekvenciájának a legvégén helyezkedett el, és nem volt mellette kellő nagyságú jobboldali kontextus, a következő szekvenciában középtájon jelent meg, így megfelelő méretű kontextus esetén is látta a tanulás során a modell. Annak érdekében hogy minden szó ugyanannyiszor (kétszer) szerepeljen a tanító szekvenciákban, az első szekvencia első felét, és az utolsó második felét külön szekvenciaként is szerepeltettük.

A tokenhosszúságot leszámítva az egyes verziók minden másban megegyeztek egymással, hogy a kontextusméretet azonos körülmények között tudjuk vizsgálni. A modellt 32-es batch-mérettel 4 epochban tanítottuk 5e-5 tanulási rátával. Emellett tízszeres keresztvalidálást alkalmaztunk.

A modellt megbízhatóságát a humán annotátorokéhoz hasonlóan értékeltük: a kurátori verziót tekintjük a szövegek helyes annotálásának, és ahhoz viszonyítottuk, hogy mennyire tudja megbízhatóan megjósolni az annotálandó szavak címkéjét a seqeval (Nakayama, 2018) Python modul használatával. A kiértékelésnél az egyes jegyekre vonatkozó pontosság, fedés és F1 érték megállapításán felül a modell teljes megbízhatóságának jellemzésére az összes jegyre vonatkozó súlyozott átlagát használtuk. A következő szakaszban az egyes verziókra vonatkozó adatokat ismertetjük, illetve az ezekből levonható következtetéseket részletezzük.

4 A különböző szekvenciahosszúságú adatokkal tanított huBert pontossági adatainak összevetése

A szekvenciahosszúság, azaz az egyes annotációs döntések meghozásához rendelkezésre álló kontextus nagyságának hatásának vizsgálatára négy verziót készítettünk a finomhangolt huBert base modellből. A verziókra vonatkozó megbízhatósági értékek közül az egyes kategóriákra vonatkozó F1 értékeket, valamint ezek súlyozott átlagát az 2. Táblázatban láthatjuk.

2. Táblázat: A különböző tokenhosszúságú szekvenciákkal tanított modellek F1 értékei kategóriánként és összesítve

szekvenciahossz	64	128	256	512	support
nodirectiva	0,87	0,88	0,91	0,90	2828
sajat_hangu	0,86	0,86	0,85	0,89	2036
kozvetitett	0,46	0,45	0,44	0,47	402
meta	0,37	0,42	0,45	0,38	334
szovegszervezo	0,74	0,78	0,76	0,78	189
interakcios	0,47	0,55	0,56	0,56	114
ambiguous	0,00	0,00	0,00	0,00	24
súlyozott átlag	0,80	0,81	0,82	0,83	5927

A verziók súlyozott átlagát tekintve láthatjuk, hogy az elérhető kontextus növelésével javult az automatikus annotáció megbízhatósága. A javulás azonban kis mértékű, 0,80-ról 0,83-ra emelkedett csak az F1 értékek súlyozott átlaga a 64-es maximális tokenhosszúságú verziótól az 512-es verzióig. A modell megbízhatóságáról azonban elmondható minden verzió esetében, hogy összehasonlítható a humán annotátorok által elért megbízhatóságával (F1 = 0,830), tehát a munka elsődleges célját elértnek tekinthetjük.

A két leggyakoribb címke a tanítókorpuszban a `nodirectiva` és a `saját hangú` voltak. Ennek köszönhető lehet, hogy a modell minden tokenhosszúságnál ezeknek az annotálását tudta a leginkább elsajátítani, és átlagon felüli, 0,9 körüli F-értéket elérni. A többi funkcióból egy nagyságrenddel kevesebb tanító adat állt rendelkezésre, így nem meglepő, hogy ezeket nem volt képes olyan jól elsajátítani. Inkább meglepő ilyen szempontból, hogy a `szövegszervező` funkciót milyen magas megbízhatósággal sajátította el, tekintve, hogy ez volt az egyik legkisebb előfordulással álló kategória. Ennek az lehet az oka, hogy ez egy könnyen sémába rendezhető, véges szókészletet használó funkció, amelynek a mintázatát néhány példa alapján is be lehet azonosítani. Az egyes funkciók nem mutatják a súlyozott átlagon megfigyelhető monoton növekedést, de alapjaiban véve elmondható, hogy a nagyobb szekvenciahosszúság pontosabb annotációhoz vezetett összesítésben és jegyenként is. A modell egyszer sem jelölt az `ambiguous` címkével, minden alkalommal, amikor a tanító adatok között azt látta, választott neki egy olyan funkciót, amit felismert.

Annak ismeretében, hogy a rendelkezésre álló kontextusméret azonos körülmények között ugyan javítja a modell megbízhatóságát, azonban ez a javulás kis mértékű, felmerülhet a kérdés, hogy szükséges-e a nagyobb tokenhosszúságú verziók használata, vagy megelégedhetünk például a 128-as szekvenciahosszúsággal finomhangolt

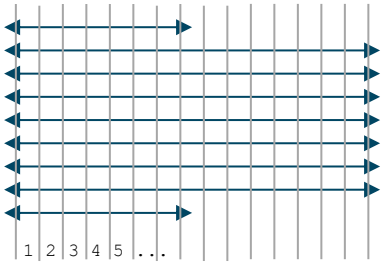
modellel is. A kísérlet további paramétereinek egyenlőségénél azt mondhatjuk, hogy igen, azonban a tanítás során megfigyeltük, hogy a kisebb maximális tokenhosszúságú verziók a tanulás végére az egyes lépésekben kevéssel javultak már (64-es tokenhosszúságnál a cumulative loss a harmadik epoch végén 0,0103, a negyedik epoch végén 0,0082), míg a nagyobb tokenhosszúságúak még lépésenként nagyobb javulást produkáltak (512-es tokenhosszúságnál a cumulative loss a harmadik epoch végén 0,0235, a negyedik epoch végén 0,0174). Emiatt felmerülhet a kérdés, hogy más tanítási feltételek mellett egy nagyobb modell megbízhatóbb eredményeket mutatott volna-e. Gondolhatunk itt az epochok számának növelésére, amellyel egy sokkal időigényesebb tanulási folyamat állna elő, de több lépés állna a modell rendelkezésére a nagyobb megbízhatóság elérésére, vagy eltérő átfedés alkalmazására, hogy több tanulható adatpont szerepeljen a tanítókorpuszban. A legkézenfekvőbb megoldás ezzel kapcsolatban pedig természetesen a tanítóadatok számának növelése lenne, ez azonban a jelenlegi helyzetben nem lett volna megoldható.

Összességében elmondható, hogy az első számú célunkat, az annotátorok eredményeihez hasonló megbízhatósági értékeket produkálni képes automatikus annotáló eszközt el tudtuk érni. A kontextus méretének vizsgálatában pedig arra jutottunk, hogy mutat ugyan javulást a modell az elérhető kontextus növelésével, ez a javulás nem nagy mértékű, így azonos feltételek mellett azt mondhatjuk, hogy a kontextus méretének hatása az automatikus annotálás megbízhatóságára nem nagy.

5 A szekvencián belüli pozíció hatása

A pragmatikai annotáció során a kontextusnak az annotáció megbízhatóságára gyakorolt hatását más módon is vizsgálhatjuk. Megnézhetjük azt is, hogy az annotáció során az egyes nyelvi elemeknek az annotációjának a pontossága/fedése hogyan változik a szekvencián belül elfoglalt helyétől. A vizsgálat során megállapíthatjuk, hogy a szekvenciák elején, közepén vagy a végén található elemek címkézése történt-e megbízhatóbb módon, illetve hogy ez a megbízhatóságváltozás szimmetrikus-e, azaz ugyanolyan hatása van-e a bal és a jobb oldali kontextusnak. Várakozásunk szerint a szekvenciák közepén nagyobb megbízhatósággal osztályoz az eszköz, mint a szélein, illetve a maximális megbízhatóság a szekvencia közepétől jobbra lesz megfigyelhető, vagyis nagyobb bal oldali kontextus szükséges a helyes osztályozáshoz, mint jobb oldali. A várható aszimmetria oka az, hogy a természetes nyelv használata során lineárisan, balról jobbra produkáljuk és dolgozzuk fel a szövegeket, és a feldolgozhatóság, azaz a megértés szempontjából előnyösebb, ha a bal oldali kontextus hordozza azokat az információkat, amelyek a megértéshez szükségesek. Például a felszólító alakok megértése szempontjából azt, hogy egy felszólító alak `nodirectiva`, azaz kötőmódú-e, legtöbbször a felszólító alakot tartalmazó tagmondatot megelőző tagmondat igéje határozza meg, pl. ***Belefáradtam****, hogy állandóan maszkot* ***viseljek***. Ugyanakkor a jobb oldali kontextus is fontos lehet bizonyos esetekben, például idézések esetében gyakran az idézett szövegrész után jelenik meg az idézést kifejező szövegrészlet: *Mindenki* ***viseljen*** *maszkot –* ***mondta az államtitkár***. A kérdés az, hogy a nyelvben megfigyelhető aszimmetria jelentkezik-e az automatikus annotáló eszköz működése során is.

Az 1. ábrán látható volt, hogy a modell tanítása során hogyan lettek kialakítva egy szöveg szavaiból 50 százalékos átfedéssel az egyes, tanításra és tesztelésre használt szekvenciák. A szekvencián belüli pozíció hatásának vizsgálatához pozíciónként összehasonlítottuk az egyes szekvenciáknak a tanításhoz használt változatán megfigyelhető címkéit ugyanezen szekvenciának a tesztelés során megjósolt címkékkel, azaz gyakorlatilag a szekvenciákat a 2. ábrán látható módon egymás alá toltuk, és az első pozícióban megfigyelt és megjósolt címkék alapján számoltuk az első pozíció pontosságát, fedését és F-értékét, és így tovább: minden pozícióhoz meghatároztuk ezeket a megbízhatósági értékeket.



2. ábra. A nyelvi elemek címkézésének megbízhatóságmérése pozíciónként.

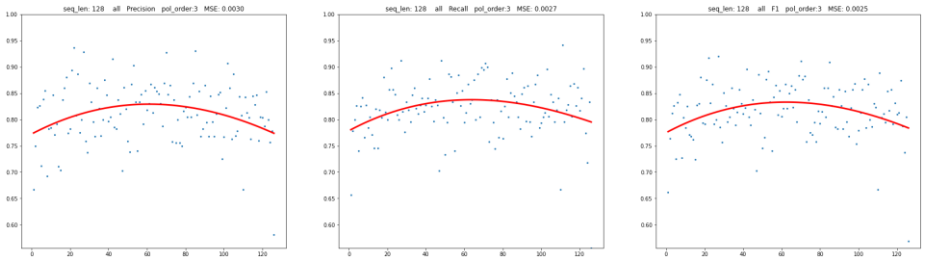
A felszólító alakok funkcióinak meghatározása során szavakhoz lett címke rendelve, a szekvenciák viszont tokenekből állnak, ezért a vizsgált címkék a szó első tokenjének a pozíciójában lettek figyelembe véve.

A megbízhatósági értékek funkciónként külön-külön és összesítve is meg lettek határozva. A funkciónkénti számolásnál true pozitív lett, ha a megfigyelt és a megjósolt címke is az adott funkciócímke volt; false pozitív esetén a megjósolt címke az adott címke volt, de a tényleges címke ettől eltérő (vagy 0 címkéjű); false negatív esetében pedig e megfigyelt címke egyezett meg a vizsgált címkével, a jósolt pedig nem. Az összesített megbízhatóság esetében TP-nek az azonos (de nem nulla) címkézés, FP-nek a különböző (de nem 0 jósolt) címkézés, FN-nek pedig a különböző (de nem 0 tanított) címkézés számított, azaz az osztályozás mikroátlagot számítottuk. A megbízhatósági értékeket 64, 128, 256 és 512 szekvenciahosszra is meghatároztuk, ezek grafikonos ábrázolása a https://github.com/szecsenyi/MSZNY2025 githubon található.

A pozíciónkénti megbízhatósági értékeknek igen nagy a szórása. Ennek leginkább az az oka, hogy az egyes pozícióknál aránylag kevés felszólító alak található. Az 1. táblázatban is látható, hogy összesen mintegy 6000 felszólító alakot kell felcímkézni, ez 64 tokenhosszúságú szekvenciák esetében átlagosan 100 címkét jelent, 512 szekvenciahossznál azonban csak 10-et. Még kevesebb átlagos címkézés jut egy-egy pozícióra, ha csak az egyes értékek megbízhatóságát vizsgáljuk: nodirectiva és saját hangú címkékből 2000-3000 található, de a többiből már csak 500 alatti mennyiségű. Ez utóbbi esetekben 64, illetve 512 szekvenciahossznál átlagosan már csak 10, ill. 1 alatti mennyiség jut az egyes pozíciókra, ezek az eredmények így már valójában értékelhetetlenek.

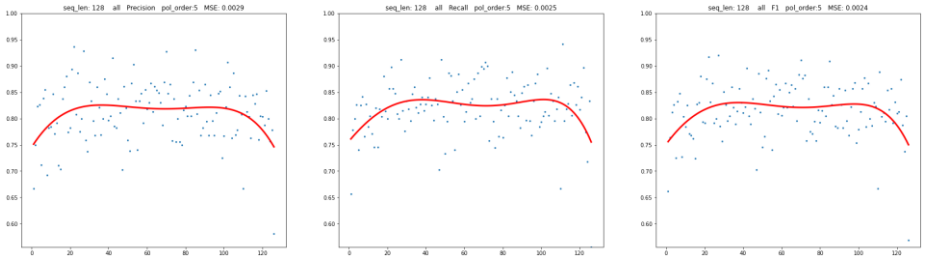
A megbízhatósági értékek nagy szórása miatt a pozíció és a megbízhatóság közötti összefüggést a ponthalmazra illesztett görbékkel lehet jobban szemléltetni. Másodfo-

kú polinom illesztése esetén a kapott görbe a maximális értékhez viszonyítva szimmetrikus, ezért a bal és jobboldali kontextus hatásának a különbségét nem mutatja. Harmadfokú polinomnál már látható ez a különbség is, ezért a grafikonoknál ezt alkalmaztuk. Ötödfokú polinom esetében elég nagy pozíciónkénti címkeszámnál még jobban megfigyelhető az összefüggés, ezért ilyen közelítéseket is ábrázoltunk. A polinomok illesztését a `numpy` python library `polyfit` függvényével végeztük. A polinomillesztés metrikájaként a polinom átlagos négyzetes eltérését (MSE) számítottuk, ami az ábrák feliratában is megjelenik.



3. ábra. 128 tokenhosszúságú szekvenciáknál az egyes pozíciókban mért P, R és F1 értékek az összes címkét figyelembe véve, illetve a pontokhoz illesztett harmadfokú polinom.

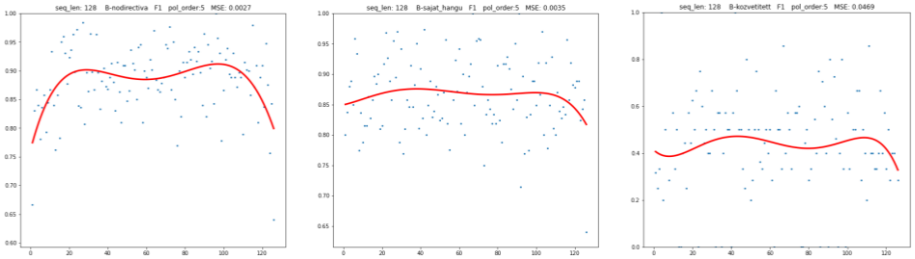
Az 3. ábrán látható, hogy mindhárom megbízhatósági érték esetében a széleken alacsonyabb értékek vannak, mint a szekvencia közepén. A regressziós görbék továbbá aszimmetriát is mutatnak, mivel a maximális érték nem az szekvencia közepére esik, és két oldalán nem ugyanolyan mértékű az értékek csökkenése. Az aszimmetria még jobban megfigyelhető ötödfokú polinom illesztése esetén (4. ábra): a szekvencia közepén található aránylag egyenletes szakasz bal oldalán hosszabb, de laposabb emelkedés van, a jobb oldalán rövidebb, de meredekebb csökkenés.



4. ábra. 128 tokenhosszúságú szekvenciáknál az egyes pozíciókban mért P, R és F1 értékek az összes címkét figyelembe véve, illetve a pontokhoz illesztett ötödfokú polinom.

A grafikonok alapján a felszólító alakok funkcióinak megállapításakor a huBert az annotált elemek bal oldalán nagyobb kontextust vesz figyelembe, mint a jobb oldalán, de a két (különböző méretű) kontextus hatása nagyjából megegyezik. Az 4. ábra grafikonjai alapján a releváns baloldali kontextus mérete kb. 25–30 token hosszúságú, a jobb oldal pedig 15–20.

A kontextusfüggőség az egyes funkcióknál is megfigyelhető, bár különböző mértékben (5. ábra). A `nodirectiva` értéknél nagyobb a kontextus hatása, mint a `saját hangú` értéknél, de a figyelembe vett kontextusok nagysága nem különböző. A `közvetített` funkció grafikonján észrevehető, hogy az F1 értékek sokkal jobban szórnak, mint a másik funkciónál, és a regressziós görbe is jóval alacsonyabban van. Ennek oka az, hogy a korpuszban sokkal kevesebb ilyen címkéjű elem szerepel; ennek tudhatjuk be azt is, hogy ebben a grafikonban a pontok vízszintes csíkokba rendeződnek.



5. ábra. 128 tokenhosszúságú szekvenciáknál az egyes pozíciókban mért F1 értékek az a `nodirectiva`, `saját hangú` és a `közvetített` funkcióknál, illetve a pontokhoz illesztett ötödfokú polinom.

6 Összegzés

A tanulmány azt vizsgálta, hogy a pragmatikai jegyek nagy nyelvi modellel történő automatikus annotációja során az eszköz mennyire veszi figyelembe a rendelkezésre álló kontextust. A vizsgálat a MedCollect egészségügyi álhírkorpuszban található felszólító alakok különböző pragmatikai funkcióinak kézzel annotált változatának segítségével történt.

A kontextuális hatás kétfajta módszerrel lett elemezve. Az egyiknél azt vizsgáltuk, hogy a tanításnál és tesztelésnél használt 50%-os átfedéssel kialakított szekvenciák hossza (64, 128, 256, 512 token) befolyásolja-e az eszköz megbízhatósági értékeit (P, R, F1). Azt tapasztaltuk, hogy a szekvenciahossz növelésével együtt nőtt az F1 érték is (0,80, 0,81, 0,82 és 0,83). A másik módszernél azt vizsgáltuk, hogy a felszólító alakok a tesztszekvencián belüli pozíciója hogyan befolyásolja a funkció meghatározásának megbízhatóságát. Azt tapasztaltuk, hogy a szekvenciák széléhez közeli pozíciókban alacsonyabbak voltak a megbízhatósági értékek, mint a szekvenciák közepén, vagyis a rendelkezésre álló kontextusnak volt hatása, továbbá hogy a kontextusfüggőség aszimmetrikus, vagyis nagyobb bal oldali kontextust (25–30 token) vesz figyelembe az eszköz, mint jobb oldalit (15–20 token). A figyelembe vett kontextus nagysága megmagyarázza azt is, hogy a szekvenciák hosszának növelése miért csak ilyen kis mértékben növeli a megbízhatóságot: ha a funkciók megállapításához csak ±25 token kontextust vesz figyelembe az eszköz, akkor a szekvenciák közepén, ahol ez a kontextus rendelkezésre áll, állandó megbízhatóságú a kategorizá-

lás, amit csak a szekvenciák végeinek megbízhatósága ront le. Minél hosszabb a szekvencia, annál nagyobb a középső rész hossza a végekhez viszonyítva.

A korpuszban levő felszólító alakok kis száma miatt a kis előfordulású funkciók pozíciónkénti megbízhatóságának mérése nem értelmezhető. Ez javítható lenne azzal, ha a szekvenciák átfedését megnövelnénk 75 vagy 90 százalékra.

A nagy nyelvi modell figyelembe veszik az automatikus pragmatikai annotáció során a rendelkezésre álló kontextust. A figyelmbe vett kontextus azonban sokkal kisebb, mint a humán annotátorok esetében feltételezett.

Bibliográfia

Archer, D., Culpeper, J., Davies, M.: Pragmatic annotation. In: Lüdeling, A., Kytö, M. (szerk.) Corpus Linguistics: An International Handbook. pp. 613–542. Walter de Gruyter, Berlin; New York (2009)

Milà-Garcia, A.: Pragmatic annotation for a multi-layered analysis of speech acts: A methodological proposal. Corpus Pragmatics 2, 265–287 (2018) https://doi.org/10.1007/s41701-018-0037-z

Nakayama, H.: seqeval: A Python framework for sequence labeling evaluation. (2018)

Nemeskey, D. M.: Introducing huBERT. In: Berend, G., Gosztolya, G., Vincze, V. (szerk.) XVII. Magyar Számítógépes Nyelvészeti Konferencia. pp. 3–14. Szegedi Tudományegyetem, Informatikai Intézet, Szeged (2021)

Nemeskey, D. M.: Natural Language Processing Methods for Language Modeling (PhD Thesis). Eötvös Loránd University (2020)

Németh T. E.: Álhírek, általtudományos nézetek nyelvészeti azonosítása. Magyar Nyelv 119, 490–496 (2023) https://doi.org/10.18349/MagyarNyelv.2023.4.490

Szécsényi T., Nagy C. K., Németh T. E.: Felszólításannotálás a MedCollect egészségügyi álhírkorpuszban. In: Berend G., Gosztolya G., Vincze V. (szerk.) XX. Magyar Számítógépes Nyelvészeti Konferencia. pp. 159–170. Szegedi Tudományegyetem, Informatikai Intézet, Szeged (2024)