Koreferenciafeloldás magyar szövegeken BERT-tel

Vadász Noémi[1], Nyéki Bence*

[1]Nyelvtudományi Kutatóközpont
1068 Budapest, Benczúr u. 33.
vadasz.noemi@nytud.hu, nyeki.bence96@gmail.com

Kivonat A cikk egy kísérletet mutat be, amelyben összevonva használtunk két koreferenciakorpuszt a magyar BERT modell finomhangolásához, amivel magyar szövegeken lehet koreferenciafeloldást végezni. A cikk ismerteti a kísérletünk lépéseit a korpuszok előkészítésétől és felhasználásától a BERT modell finomhangolásán keresztül az eredmények kiértékeléséig. A koreferenciafeloldót szabadon hozzáférhetővé tettük.
Kulcsszavak: BERT, finomhangolás, koreferencia

1. Bevezetés

A neurális nyelvmodellek lehetővé teszik, hogy finomhangolás segítségével magasabb szintű nyelvfeldolgozási feladatokat oldjunk meg. A finomhangolás során egy kisebb méretű, speciális annotációval ellátott korpuszt használunk. Mivel a koreferenciafeloldás feladatának megoldását tűztük ki célként, a finomhangoláshoz koreferenciaannotációval ellátott korpuszra volt szükségünk. Abban a szerencsés helyzetben vagyunk, hogy magyarra már rendelkezésre állnak nagy méretű neurális nyelvmodellek és a kitűzött feladathoz alkalmazható koreferenciakorpuszok is.

Kísérletünk során a `huBERT` (Nemeskey, 2021) modellt finomhangoltuk a SzegedKoref (Vincze és mtsai, 2018) és a KorKor (Vadász, 2020) korpuszok felhasználásával. Az általunk készített koreferenciafeloldó szabadon hozzáférhető, a GitHub-on[1] elérhető.

2. Háttér

2.1. Anafora és koreferencia

Az információkinyerés területén belül számos olyan feladat van, amelyek megoldásához anafora- vagy koreferenciafeloldásra van szükség. Anaforikus kapcsolat egy visszautaló elem (pl. egy névmás) és a szövegbeli előzménye (antecedense) között áll fenn, a koreferencia pedig azt jelenti, hogy a szövegben két elem ugyanarra a valóságbeli entitásra refereál. Számítógépes nyelvészeti szempontból általában a koreferenciafeloldás feladata egy entitás összes szövegbeli említésének az

* A kutatás ideje alatt a szerző a Nyelvtudományi Kutatóközpont munkatársa volt.
[1] https://github.com/nytud/bert_coref_hu

azonosítása, míg az anaforafeloldás az anafora előzményének, az antecedensnek a megtalálása.

Az anafora- és koreferenciaannotációt tartalmazó korpuszok különböző annotációs elvek mentén készülnek, a különböző anaforikus- és koreferenciakapcsolatok aleseteit gyakran különböző osztálycímkékkel jelölve. A névmási anafora alesetei kategorizálhatók a névmás alapján (személyes, mutató, birtokos, kölcsönös, visszaható, vonatkozó névmások), míg a főnevek között fennálló koreferenciakapcsolat is többféle lehet (pl. ismétlés, szinonima, rész-egész kapcsolat, epitheton, appozíció stb.). Emellett megjelenhetnek a nem névszói kapcsolattípusok is pl. az igei és a határozói anaforák esetében, valamint akár képzett alakok között is. Nem egységes a deixis és a katafora kezelése sem az egyes korpuszokban, valamint az ún. pro-drop nyelvek esetén a zérónévmások kezelése sem.

2.2. Feloldók

A magyar nyelvre eddig több tudásalapú, szabályokkal operáló anafora- és koreferenciafeloldó megoldás is született. Lejtovicz és Kardkovács (2007) szabályalapú anaforafeloldója a BFP algoritmust (Brennan és mtsai, 1987) implementálta és szintaktikailag elemzett mondatokat kap bemenetként, tehát szintaktikai és morfológiai információra támaszkodik a feloldás során. Miháltz és mtsai (2007) és Miháltz (2012) megoldása a mondatok mély szintaktikai szerkezete mellett a kötéselmélet tételeire és pszicholingvisztikai megállapításokra támaszkodik, valamint a Magyar WordNet (Miháltz és mtsai, 2008) ontológiában tárolt nyelvi tudást is kiaknázza. Ez utóbbi a koreferenciakapcsolatok esetében segít, míg az előbbiek az anaforikus kapcsolatok felismerésében. (Vadász, 2017) és az annak alapján készült szabályalapú anaforafeloldó, amit a KorKor korpusz (Vadász, 2020) építéséhez használtak fel szintén szintaktikai és morfológiai információkra támaszkodnak (az előbbi még a főnevek [+/-] élő szemantikai jegyére is). Mindkét megoldás első lépésként azonosítja a szintaxisfákban az egyes grammatikai funkciókat, majd az ún. Pléh-Radics algoritmus (Pléh és Radics, 1976) szabályait alkalmazva keresi az egyes szereplők antecedensét.

Munkácsy és Farkas (2016) mutatta be az első statisztikai módszert koreferenciafeloldásra magyar nyelvű szövegekben, amelyben a SzegedKoref (Vincze és mtsai, 2018) korpuszon tanították a HOTCoref (Björkelund és Kuhn, 2014) rendszert, amit aztán a magyar nyelv igényeihez igazítottak. Tudásszegény megoldást kínál még a magyar névmási anaforák antecedenskeresésére Kovács Viktória doktori disszertációja (Kovács, 2021), amelyben gépi tanulási kísérleteket végez a különböző névmástípusok esetében.

Neurális megoldás a legjobb tudásunk alapján magyarra eddig nem született a témában, az angol nyelv esetében viszont több példa is van rá, hogy neurális hálókat alkalmaztak koreferenciafeloldásra. Ezek közül kettőt emelünk ki, hiszen ezek a munkák inspiráltak bennünket a kísérletezéseink kezdetén.

A NeuralCoref[2] (Wolf, 2017) a SpaCy[3] (Honnibal, 2015) rendszer pipelinejába illeszthető koreferenciafeloldó eszköz, amit neurális háló segítségével készítettek és angol nyelvű szövegekre működik. Noha a NeuralCoref újratanítható akár más nyelvű korpuszokkal is, számos megszorítás vonatkozik a tanítóanyag formátumára és tartalmára. Emellett a rendszer összetettsége miatt bonyolult adaptálni azokra a nyelvekre, amelyeket a SpaCy nem támogat, vagy nem áll a rendelkezésre megfelelő tanítóanyag. Joshi és mtsai (2019) pedig egy magas szintű, transzformeralapú koreferenciafeloldó rendszert készítettek, amelyben a *c2f-coref* modellt(Lee és mtsai, 2018) fejlesztették tovább a BERT integrálásával.

A bemutatott angol nyelvre készült megoldásokkal szemben a kutatásunkban megpróbáltunk egy konceptuálisan egyszerűbb, könnyebben felépíthető megoldásra törekedni a finomhangolt `huBERT` modell utolsó rétege által kibocsátott kontextuális tokenbeágyazások klaszterezésével.

3. A felhasznált korpuszok

A kísérleteink során két korpuszt használtunk, amelyek többek között koreferenciaannotációt is tartalmaznak. A SzegedKoref (Vincze és mtsai, 2018) a Szeged Korpusz (Csendes és mtsai, 2005) egy részéből készült és iskolai fogalmazásokat és újságcikkeket tartalmaz. A korpusz a publikáció alapján 55 ezer tokent tartalmaz, ugyanakkor a méréseink alapján sokkal nagyobb, majdnem 124 ezer token. A másik korpusz, a KorKor[4] (Vadász, 2020) sokkal kisebb, mindössze 31 ezer tokenes (beleszámolva az írásjeleket és a zéró elemeket, mint a zéró létigék, elliptált igék, zérónévmások) és hírszövegeket és Wikipédia szócikkeket tartalmaz. Hogy minél nagyobb tanító- és tesztelőanyag álljon rendelkezésünkre, a két korpuszt egységes formátumra alakítva használtuk fel a kísérletünkhöz.

Mindkét korpusz tartalmazza a zérónévmásokat is, így azokat az anaforikus kapcsolatokat is, amelyekben a zérónévmások vesznek részt. A pro-drop nyelvek esetében nagyon fontos, hogy egy koreferenciakorpusz tartalmazza ezeket a zéró elemeket. A KorKor korpusz esetében például az összes névmás háromnegyede zérónévmás. A zérónévmások mellett minden nem testes elem, tehát a zéró létigék és elliptált igék is nagyon fontos szerepet játszanak a különböző információkinyerési feladatokban. Mindezek ellenére a jelenlegi kísérletben nem használtuk a zérónévmásokat. Ennek az az oka, hogy a koreferenciafeloldó program bemenete sem tartalmazza őket – hacsak nem egy előfeldolgozó lépés eredményeként. A jelenlegi munkában tehát csak a szövegekben testesen előforduló elemekkel dolgoztunk. A SzegedKoref elérhető egy olyan verzióban, ami nem tartalmazza ezeket a testetlen elemeket, a KorKor korpuszt pedig automatikus módszerekkel alakítottuk át ilyenné.

A két korpusz formátuma és annotációs sémája több ponton is eltér, ezért mindenekelőtt egységesítenünk kellett őket. A KorKor esetében a dependenciaelemzés mintájára a koreferenciakapcsolatban résztvevő elem feje mellett szere-

[2] https://github.com/huggingface/neuralcoref
[3] https://spacy.io/
[4] https://github.com/vadno/korkor_pilot

pel az antecedens vagy a szövegben korábban előforduló koreferens elem indexszáma, míg a SzegedKorefben a konstituenselemzéshez hasonlóan vannak összezárójelezve az összetartozó elemek. Az utóbbi a szerencsésebb megoldás, mert így az egymásba ágyazott vagy mellérendelt főnévi csoportok esetében is egyértelműen kiderül, hogy melyik főnévi csoport szerepel a koreferenciakapcsolatban. A SzegedKoref esetében ez azt jelenti, hogy egy token mellett akár több indexszám is szerepelhet. Az egységesítés azonban sajnos csak úgy volt automatikusan megoldható, ha a KorKor-ban használt módszert alkalmazzuk mindkét korpuszra. Ez azt jelenti, hogy a SzegedKoref esetében néhány koreferenciakapcsolatot elveszítettünk. Ha egy token mellett több koreferens elem is meg van jelölve, akkor csak a legbelső főnévi csoporthoz tartozó index-számot tartottuk meg.

Mindkét korpusz rövid szövegekből áll (KorKor: 112 és 652 token között, SzegedKoref: 11 és 1496 token között), az előkészítés során mégis tovább kellett darabolnunk őket. A `huBERT` bemenetét ugyanis subword tokenekre kell bontani, a modell pedig csak olyan szekvenciákat dolgoz fel, amelyek legfeljebb 512 subwordöt tartalmaznak. A feldarabolás után 600 darab fájlt kaptunk.

A két korpusz méretének és a bennük szereplő szövegek műfajának az arányát megtartva tanító, fejlesztő és teszthalmazra bontottuk őket. Az 1. táblázat mutatja az egyes halmazok méretét.

forrás		train	devel	test	TOTAL
SzegedKoref		98 774	12 131	12 521	12 3426
	HVG	7 779	772	1 163	9 714
	iskolai fogalmazás (8.o.)	73 891	9 353	9 215	92 459
	iskolai fogalmazás (10.o.)	17 104	2 006	2 143	21 253
KorKor		22 222	2 837	2 817	27 876
	hír	6 970	887	803	8 660
	Wikipedia	15 252	1 950	2 014	19 216
TOTAL		120 996	14 968	15 338	151 302

1. táblázat. A két korpuszból elkülönített tanító, fejlesztő és tesztanyag mérete tokenben kifejezve, a központozási jeleket is beleszámítva.

A 2. táblázatban a korpuszokban jelölt anaforikus és koreferenciakapcsolatok száma látható. Azt is szerettük volna megjeleníteni, hogy a korpuszokban összesen hány koreferenciacsoport van. Ez nem azonos a szövegekben előforduló entitások számával, hiszen csak azokat az entitásokat számítottuk bele, amiknek legalább két említése volt egy szövegben. Ennek az az oka, hogy a csupán egyetlen említéssel rendelkező entitások nincsenek megjelölve az általunk használt korpuszokban (lásd a 4. fejezetben). Az összevont korpusz fájljai átlagosan 11-12 koreferenciacsoportot tartalmaznak.

	KorKor		SzegedKoref	
	kapcsolat	csoport	kapcsolat	csoport
tanítóadat	2 521	796	10 290	3 522
fejlesztőadat	364	113	1 294	432
tesztelőadat	406	121	1 211	431

2. táblázat. A kapcsolatok és a koreferenciacsoportok száma a korpuszokban.

4. Szófajalapú előszűrés

Az egyes koreferenciakorpuszok között különbségek lehetnek atekintetben, hogy hogyan jelölik azokat a referáló elemeket, amelyek nem állnak semmivel koreferenciakapcsolatban. (A szakirodalomban ezekre *singleton*-ként hivatkoznak, a cikkben a továbbiakban szingletonnak nevezzük.) Az általunk használt korpuszokban a szingletonok nem voltak annotálva, ezért nem állt a rendelkezésünkre tanítóadat ehhez az alfeladathoz.

A nemzetközi szakirodalom a *mention* vagy *markable* kifejezéseket használja az anaforikus- vagy koreferenciakapcsolatokban potenciálisan részt vevő elemek megnevezésére, a cikkben a továbbiakban mi jelölteknek nevezzük ezeket. A lehetséges jelöltek kiválogatása egy külön előfeldolgozási lépésként értelmezhető a tényleges koreferenciafeloldás előtt. Megoldásunk azonban más irányból közelíti meg ezt a feladatot: a jelöltek kiválogatása helyett megpróbáljuk kizárni azokat a tokeneket, amelyeknek valószínűleg nincs antecedense vagy nem vesz részt koreferenciakapcsolatban. Szerencsére egyszerű szabályok alapján könnyen szűkíthetjük a lehetséges jelöltek körét. Ehhez szófajalapú előszűrést alkalmaztunk.

Mindkét általunk felhasznált korpusz tartalmaz kézzel ellenőrzött minőségű nyelvi annotációt, így a szófajt is – igaz, eltérő annotációs sémákkal és címkekészletekkel. Az egységesítés kedvéért ezért mindkét korpuszt leelemeztük az e-magyar (Váradi és mtsai, 2018; Indig és mtsai, 2019) elemzővel. Ez azt jelenti, hogy a gold standard minőségű kézi annotációt feláldoztuk az egységesítés oltárán, ugyanakkor azt is szem előtt kell tartani, hogy az e-magyar nagyon jó minőségű morfológiai elemzést bocsát ki (Simon és mtsai, 2020).

A nyelvi jelenségek ismeretében és a korpuszok segítségével nagy biztonsággal megállapítható, hogy mely szófajokat érdemes kizárni. Ez az egyszerű előszűrési lépés nagymértékben megkönnyíti a koreferenciafeloldó munkáját, még ha feltételezzük is, hogy az előszűrés során használt szófaji címkéző esetleges hibái begyűrűzhetnek a következő feldolgozási lépésekre, így rontva a koreferenciafeloldó teljesítményét. Az előszűrés során az igéket, igekötőket, csak jelzőként használt mellékneveket, határozószókat, névelőket, névutókat, kötőszavakat, indulat- és mondatszavakat és az írásjeleket zártuk ki a jelöltek közül.

5. Finomhangolás

A kísérletünkben a `huBERT` (Nemeskey, 2021) által kibocsátott kontextuális tokenbeágyazásokat finomhangoltuk. A célunk az volt, hogy a tokenbeágyazásokat úgy módosítsuk, hogy megragadják a koreferenciakapcsolatokat, tehát azt szerettük volna elérni, hogy az azonos koreferenciacsoporthoz tartozó tokenek reprezentációi a lehető legközelebb legyenek egymáshoz és a lehető legtávolabb legyenek a többi token beágyazásától.[5] Ehhez a reprezentációtanulásban gyakran használt[6] ún. *triplet loss* célfüggvényt alkalmaztuk. A függvény definíciója

$$\mathcal{L}(A,P,N) = \sum_{i}^{N} max(||f(A^{(i)}) - f(P^{(i)})|| - ||f(A^{(i)}) - f(N^{(i)})|| + \epsilon, 0)$$

ahol $A^{(i)}$ a viszonyítási alap (*anchor*),[7] $P^{(i)}$ egy *pozitív* példa (egy olyan beágyazás, amelytől azt várjuk, hogy közel legyen a viszonyítási alaphoz), $N^{(i)}$ pedig egy *negatív* példa (egy olyan beágyazás, amelynek távol kell lennie a viszonyítási alaptól). $||\cdot||$ egy távolságmetrika, ϵ pedig egy nemnegatív valós szám, amit toleranciának (*margin*) nevezhetünk.

Ezt a reprezentációtanulásban használt eljárást a saját feladatunkhoz igazítottuk. Jelöljük a tanítóanyag i-edik adatpontját $X^{(i)}$-vel, amely egy koreferenciajelölésekkel annotált szöveg. Legyen n_i azon tokenek száma $X^{(i)}$-ben, amelyekre igaz, hogy valamelyik koreferenciacsoporthoz tartoznak. $X^{(i)}$-t n_i-szer másoltuk le. Az annotáció módosítása után ezeket a másolatokat használtuk tanítóadatként.

A módosítás a következőképpen zajlott: Minden új $X^{(i,j)}$ ($j \in \{1,\ldots,n_i\}$) adatpontban viszonyítási alapként jelöltük meg a j-edik olyan tokent, amely valamelyik koreferenciakapcsolathoz tartozott. Azokat a tokeneket, amelyek ugyanahhoz a koreferenciakapcsolathoz tartoztak, mint a viszonyítási alap, pozitívként címkéztük meg. Minden egyéb olyan tokent, amelyet nem zárt ki a szófaji szűrőnk, negatívnak jelöltünk, eltekintve egy kivételtől: a tanítóanyagban azokat a szingletonokat is figyelmen kívül hagytuk (azaz a szófaji alapon kiszűrt tokenekkel azonos módon kezeltük), amelyek szintaktikailag valamilyen nem szingleton tokentől függtek. Ennek az az oka, hogy nem feltétlenül akartuk megakadályozni, hogy a szingletonok beágyazásai távol kerüljenek a szintaktikai szüleik beágyazásától, ez ugyanis megnehezíthette volna a tanulást. Ez ugyan némi zajt eredményezhet a klaszterezés kimenetében (lásd a 6. fejezetben), amely azonban szintaktikai szabályokkal csökkenthető.

[5] A tokenek reprezentciója alatt a token első subwordjének a reprezentációját értjük, a többi subword reprezentációját nem vettük figyelembe a tanítás és a klaszterezés során.

[6] A triplet loss célfüggvényt használták többek között a FaceNet (Schroff és mtsai, 2015) és az SBERT (Reimers és Gurevych, 2019) tanításakor.

[7] A továbbiakban így nevezzük a viszonyítási alapnak választott tokent és annak beágyazását is, mivel a kontextusból mindig világos lesz, hogy adott esetben melyikről van szó.

A triplet loss formulát úgy módosítottuk, hogy $||f(A^{(i)}) - f(P^{(i)})||$-t a viszonyítási alapnak a pozitív tokenek beágyazásaitól mért átlagos euklidészi távolságával cseréltük ki. A $||f(A^{(i)}) - f(N^{(i)})||$ tag helyett pedig euklidészi távolságot számoltunk a viszonyítási alap és a hozzá legközelebbi negatív token beágyazása között.

A fentebb ismertetett triplet loss célfüggvény segítségével dúsított tanítóadattal finomhangoltuk a `huBERT` modellt. AdamW optimalizálót használtunk, a tanulási rátát pedig koszinuszos ütemezővel csökkentettük a tanítás alatt. A hiperparamétereket Bayes-i keresés segítségével állítottuk be a validációs veszteség minimalizálásával.

5.1. Vizualizáció

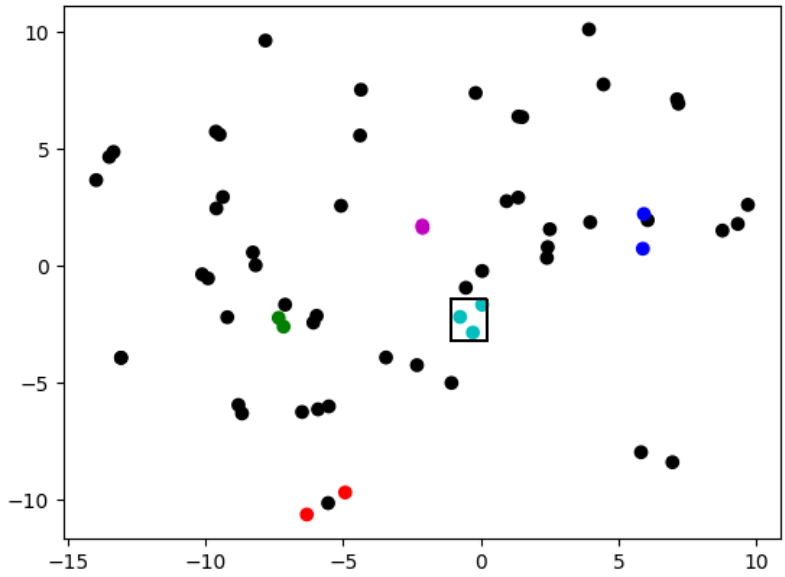
A klaszterező algoritmus bemeneteként szolgáló tokenbeágyazásokat vizualizáltuk is olyan módon, hogy a 768 dimenziós beágyazást két dimenzióra vetítettük le főkomponens-analízis (PCA) és t-elosztott sztochasztikus szomszédbeágyazás (t-SNE) segítségével.

Az 1. ábrán egy tesztfájl eredményének a vizualizációja látható. A pontok színe jelöli az egy koreferenciacsoportba tartozó elemeket a gold standard alapján, tehát minden szín egy entitást jelöl. A fekete pontok a szingleton elemek, tehát azok, amelyeket nem zárt ki a szófajalapú előszűrés, de mégsem szerepelnek koreferenciakapcsolatban. Az ábrán látható, hogy az egyes koreferenciacsoportok tagjai valóban közel vannak egymáshoz.

6. Klaszterezés

Azt vártuk, hogy az egyes koreferenciacsoportokba tartozó tokenreprezentációk klaszterekbe rendeződnek. A klaszterek felismerésére egy egyszerű algoritmust alkalmaztunk. Minden t tokenhez meghatároztunk egy s_t halmazt, ami azokat a tokeneket tartalmazza, amelyek beágyazása közelebb áll a t token beágyazásához, mint egy előre meghatározott δ küszöbérték. Minden (t, t') tokenpárra érvényes, hogy t-hez és t'-höz akkor és csakis akkor rendeltünk azonos klaszter címkét, ha $s_t = s_{t'}$ fennállt. Úgy hangoltuk δ-t, hogy az maximalizálja a fejlesztő halmaz koreferenciacímkéi és a klasztercímkék között mért normalizált kölcsönös információ (NMI) értékét. A legjobb eredményt úgy értük el, hogy a δ értékét 10-re állítottuk.

Az eredmények javítására egy függőségi információra támaszkodó heurisztikát használtunk. Ha egy c klaszterben a tokenek függőségi szüleinek a halmaza nem tartalmazott legalább két olyan tokent, amely nem volt része c-nek, akkor a c klasztert eltávolítottuk. Azért választottuk ezt a technikát, mert a szintaktikai fejek és jelzőik beágyazásai gyakran közel kerültek egymáshoz.



1. ábra: Egy tesztfájl eredményének 2D vizualizációval ábrázolva. A bekeretezett három pont által jelölt csoportba az alábbi szóalakok tartoztak: *itt, gazdaságok, ságazdaságok*.

7. Kiértékelés

A koreferenciafeloldó teljesítményét több szempontból is szerettük volna megítélni, ezért – és a feladat összetettsége miatt – a megoldásunkat két lépésben értékeltük ki.

7.1. A szófajalapú előszűrés kiértékelése

A koreferenciafeloldó megoldásunk első lépése a szófajalapú előszűrés, aminek a minősége nagyban befolyásolja a tényleges koreferenciafeloldás teljesítményét, így elsőként az előszűrés minőségét értékeltük ki. A szófajalapú előszűrés kimenetét összevetettük a tesztanyaggal és megvizsgáltuk, hogy hány olyan elemet zárt ki a lehetséges jelöltek közül, amit nem kellett volna, és fordítva.

Az előszűrés kiértékeléséhez mindenképpen szükség van a szingletonokra is a tesztanyagban, hiszen az, hogy egy elem nem vesz részt koreferenciakapcsolatban, még nem jelenti azt, hogy nem jó lehetséges jelölt. Ennek érdekében a tesztanyagban kézzel bejelöltük a lehetséges jelölteket.

Az összehasonlítás során a következő találati kategóriákat állítottuk fel minden tokenre:

– TP: az előszűrő kizárta és a tesztanyagban sem jelölt
– TN: az előszűrő nem zárta ki és a tesztanyagban is jelölt
– FP: az előszűrő kizárta, de a tesztanyagban jelölt
– FN: az előszűrő nem zárta ki, de a tesztanyagban nem jelölt

A fenti találati kategóriákat összesítettük, majd pontosságot, fedést és F-mértéket számoltunk. A eredmények a 3. táblázatban láthatók.

metrika	eredmény
pontosság	0,9900
fedés	0,6798
F-mérték	0,8061

3. táblázat. A szófajalapú előszűrés kiértékelése.

Az eredménynek értékelésekor figyelembe kell venni, hogy a szófajalapú előszűrés a szófaji egyértelműsítés kimenetére támaszkodik, tehát előfordulhat, hogy a POS-tagger által vétett hiba továbbgyűrűzik és rontja a szófajalapú előszűrés és a rá épülő koreferenciafeloldó teljesítményét is.

7.2. A koreferenciafeloldás kiértékelése

Ahhoz, hogy jobban felmérhessük a feloldónk teljesítményét, összevetettük egy baseline megoldással. Sajnos nem találtunk olyan elérhető, könnyen beüzemelhető koreferenciafeloldót magyarra, ami megfelelt volna az elvárásainknak. A 2. fejezetben ismertetett megoldások nagy része nem biztosított hozzáférést a forráskódhoz, így nem tudtuk reprodukálni a működésüket. Vadász (2020) szabályalapú szkriptje pedig hozzáférhető[8] ugyan, de csak a személyes névmások antecedenskeresését végzi.[9]

A fentiek miatt tehát egy saját baseline megoldást készítettünk el. A baseline megoldásban fájlonként összekapcsoltuk azokat a főneveket, amelyeknek ugyanaz volt a lemmája. A baseline tehát pusztán egy bizonyos főnévi koreferenciakapcsolatot, az ismétlést próbálja megragadni, viszont az ismétlés egy nagyon gyakori koreferenciatípus. A SzegedKoref anyagában az összes kapcsolat 23,44%-a ismétlés! Ennél csak a névmási anaforakapcsolatokból van több (az összes kapcsolat 34,37%-a), de ezek a különböző névmások eltérő viselkedésének köszönhetően

[8] https://github.com/vadno/korkor_pilot/blob/master/scripts/anafora.py
[9] Mivel a zérónévmásokat kihagytuk a megoldásunkból és így a tesztanyagból is, ezért nagyon kevés személyes névmás van a tesztanyagban. A tesztanyag 132 személyes névmásából 4 zérónévmás volt.

nehezebben megragadhatóak. Az ismétlés tehát elég gyakori, és felszíni jegyek alapján könnyen azonosítható, így jól megfelelt baseline megoldásnak.

A koreferenciafeloldás kiértékelése összetett feladat és többféle szempontból is megközelíthető. A minőségbe beleszámít az, hogy a feloldó megtalálta-e a koreferenciakapcsolatokban résztvevő elemeket, összekapcsolta-e az összetartozó elemeket, megtalálja-e egy entitás összes említését a szövegben stb. A szakirodalomban ezért többféle kiértékelési módszerrel és mérőszámmal találkozhatunk, a módszerek és a mérőszámok megbízhatósága pedig gyakran megkérdőjelezhető[10]. Az elterjedt kiértékelő metrikákhoz azonban szükség lenne a szingletonokra is, ez azonban nem állt a rendelkezésünkre.

A fentiek miatt jelen kísérletünkben nem egy standard kiértékelő metrikát alkalmaztunk, hanem egy olyan módszert, ami több szempontból is megmutatja a koreferenciafeloldó megoldásunk teljesítményét. Mivel klaszterezést alkalmaztunk az egymáshoz közel álló elemek reprezentációjára, ezért a klasztereket értékeltük ki és vetettük össze a gold standard koreferenciacímkékkel. Azokat a tokeneket nem vesszük figyelembe, amelyeket a szófajalapú előszűrés során kizártunk a jelöltek közül, valamint azokat sem, amelyeket a feloldó helyesen jelölt szingletonnak.

Két metrikát alkalmaztunk a kiértékelésre: a tisztságot (*purity*) és a normalizált kölcsönös információt (*normalized mutual information*, NMI). Az előbbi azt méri, mennyire homogének az osztálycímkék (esetünkben a korpusz annotációjában megadott címkék) a klasztereken belül, következésképpen a tisztaság könnyen növelhető a klaszterszám emelésével. Az NMI arról ad képet, hogy mennyivel csökken az osztálycímkék entrópiája a klasztercímkék ismeretében. Mindkét metrika 0 és 1 között vesz fel értékeket (a tisztaság rosszabb esetben megközelíti a nullát). Az eredmények a 4. táblázatban láthatók. A koreferenciafeloldó mindkét metrika alapján jobban teljesített a baseline megoldásnál.

metrika	baseline	**eredmény**
purity	0,7245	0,7619
NMI	0.6272	0,6794

4. táblázat. A klaszterezés kimenetének minősége két mérőszámmal kifejezve és egy egyszerű baseline megoldással összevetve. Az eredmények az összes tesztfájl átlagát mutatják.

A tesztfájlokat közelebbről megnézve kiderül, hogy a koreferenciakapcsolatok közül az ismétlés, az anaforikus kapcsolatok közül pedig a vonatkozó névmások előzményét találta meg sikeresen a feloldó.

[10] A gyakran használt koreferenciakiértékelő metrikák tulajdonságait, előnyeit és hátrányait Moosavi és Strube (2016) vette sorra, valamint bevezettek egy újat is (*LEA*), ami bekerült a CoNLL-scorerjébe is.

8. Diszkusszió

A cikkben egy kísérletet mutattunk be, amiben egy neurális nyelvmodellt, a `huBERT`-et finomhangoltunk két egységesített koreferenciakorpusz segítségével a koreferenciafeloldás feladatára. Összetett és bonyolult felépítésű rendszerek felállítása helyett igyekeztünk konceptuálisan átlátható megoldásra törekedni, amit kiegészítettünk egy egyszerű szabályalapú előszűréssel. A programot szabadon hozzáférhetővé tettük és törekedtünk a reprodukálhatóságra is.

Az eredmények azt mutatják, hogy a koreferenciafeloldónk bár jobban teljesített a baseline megoldásnál, van még tér a fejlődésre, ezért a későbbiekben szeretnénk még jobb eredményt elérni. Emellett a jelenlegi megoldásunkból szándékosan kihagytuk a magyarban egyébként olyan fontos zérónévmásokat, azonban a továbbiakban szeretnénk ezekkel is kísérletezni.

Hivatkozások

Björkelund, A., Kuhn, J.: Learning Structured Perceptrons for Coreference Resolution with Latent Antecedents and Non-local Features. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 47–57. Association for Computational Linguistics, Baltimore, Maryland (Jun 2014), https://aclanthology.org/P14-1005

Brennan, S.E., Friedman, M.W., Pollard, C.J.: A Centering Approach to Pronouns. In: Proceedings of the 25th Meeting of the Association for Computional Linguistics. pp. 155—162 (1987)

Csendes, D., Csirik, J., Gyimóthy, T., Kocsor, A.: The Szeged Treebank. In: Proceedings of the 8th International Conference, TSD 2005. pp. 123–131. Springer, Karlovy Vary, Czech Republic (2005)

Honnibal, M.: (2015), https://explosion.ai/blog/introducing-spacy

Indig, B., Sass, B., Simon, E., Mittelholcz, I., Vadász, N., Makrai, M.: One format to rule them all – The `emtsv` pipeline for Hungarian. In: Proceedings of the 13th Linguistic Annotation Workshop. pp. 155–165. Association for Computational Linguistics, Florence, Italy (2019)

Joshi, M., Levy, O., Zettlemoyer, L., Weld, D.: BERT for Coreference Resolution: Baselines and Analysis. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 5803–5808. Association for Computational Linguistics, Hong Kong, China (Nov 2019), https://aclanthology.org/D19-1588

Kovács, V.: Névmási anaforafeloldási kísérletek a magyar nyelvben. Ph.D.-értekezés, Szegedi Tudományegyetem (2021)

Lee, K., He, L., Zettlemoyer, L.: Higher-Order Coreference Resolution with Coarse-to-Fine Inference. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). pp. 687–692. Association for Computational Linguistics, New Orleans, Louisiana (Jun 2018), https://aclanthology.org/N18-2108

Lejtovicz, K., Kardkovács, Zs.T.: Anaphora Resolution. In: Proceedings of the 8th International Symposium of Hungarian Researchers on Computational Intelligence. Budapest (2007)

Miháltz, M.: Tudásalapú koreferencia- és birtokviszony-feloldás magyar szövegekben. Általános Nyelvészeti Tanulmányok 24, 151–166 (2012)

Miháltz, M., Hatvani, C., Kuti, J., Szarvas, G., Csirik, J., Prószéky, G., Váradi, T.: Methods and Results of the Hungarian WordNet Project. In: Proceedings of The Fourth Global WordNet Conference. pp. 311–321 (2008)

Miháltz, M., Naszódi, M., Vajda, P., Varasdi, K.: NP-koreferenciák feloldása magyar szövegekben a Magyar WordNet ontológia segítségével. In: Tanács, A., Csendes, D. (szerk.) V. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2007). pp. 138–146. Szegedi Tudományegyetem, Informatikai Tanszékcsoport, Szeged (2007)

Moosavi, N.S., Strube, M.: Which Coreference Evaluation Metric Do You Trust? A Proposal for a Link-based Entity Aware Metric. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 632–642. Association for Computational Linguistics, Berlin, Germany (Aug 2016), https://aclanthology.org/P16-1060

Munkácsy, G., Farkas, R.: Statisztikai koreferenciafeloldó rendszer magyar nyelvre — első eredmények. In: Tanács, A., Varga, V., Vincze, V. (szerk.) XII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2016). pp. 295–297. Szegedi Tudományegyetem, Informatikai Tanszékcsoport, Szeged (2016)

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

Pléh, Cs., Radics, K.: „Hiányos mondat", pronominalizáció és a szöveg. Általános Nyelvészeti Tanulmányok 11(1), 261–277 (1976)

Reimers, N., Gurevych, I.: Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics (11 2019), https://arxiv.org/abs/1908.10084

Schroff, F., Kalenichenko, D., Philbin, J.: FaceNet: A Unified Embedding for Face Recognition and Clustering. CoRR abs/1503.03832 (2015), http://arxiv.org/abs/1503.03832

Simon, E., Indig, B., Kalivoda, , Mittelholcz, I., Sass, B., Vadász, N.: Újabb fejlemények az `e-magyar` háza táján. In: Berend, G., Gosztolya, G., Vincze, V. (szerk.) XVI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2020). pp. 29–42. Szegedi Tudományegyetem Informatikai Tanszékcsoport, Szeged (2020)

Vadász, N.: Anaforafeloldás menet közben – névmások egy pszicholingvisztikailag motivált elemzőben. In: Ludányi, Z. (szerk.) Doktoranduszok tanulmányai az alkalmazott nyelvészet köréből 2017: XI. Alkalmazott Nyelvészeti Doktoranduszkonferencia, pp. 192–205. MTA Nyelvtudományi Intézet, Budapest (2017)

Vadász, N.: KorKorpusz: kézzel annotált, többrétegű pilotkorpusz építése. In: XVI. Magyar Számítógépes Nyelvészeti Konferencia. pp. 141–154. Szegedi Tudományegyetem (2020)

Vincze, V., Hegedűs, K., Sliz-Nagy, A., Farkas, R.: SzegedKoref: A Hungarian coreference corpus. In: Proceedings of the 11th Language Resources and Evaluation Conference. European Language Resource Association, Miyazaki, Japan (2018)

Váradi, T., Simon, E., Sass, B., Mittelholcz, I., Novák, A., Indig, B., Farkas, R., Vincze, V.: E-magyar – A Digital Language Processing System. In: Calzolari, N., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., Tokunaga, T. (szerk.) Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), Miyazaki, Japan (2018)

Wolf, T.: (2017), https://medium.com/huggingface/state-of-the-art-neural-coreference-resolution-for-chatbots-3302365dcf30