NerKor annotálási útmutató

Simon Eszter

2021. szeptember 7.

1. Bevezetés

A számítógépes nyelvészet egy interdiszciplináris terület, amelynek célja az emberi nyelv szerkezetének gépi modellálása, valamint a természetes nyelvek számítógépes feldolgozása. Az információkinyerés a számítógépes nyelvészet egyik fontos alterülete; célja, hogy strukturálatlan szövegből automatikusan hozzájussunk a számunkra értékes információhoz. Mivel ezen információ nagy része tulajdonnevek formájában jelenik meg a szövegben, ezért a tulajdonnév-felismerésnek (named entity recognition, NER) kiemelt jelentősége van.

A NER során egy bemeneti tokensorozatban kell megnevezett entitást (named entity, NE) alkotó intervallumokat kijelölünk, ezeket véges sok kategóriába besorolva. Egy gépi tanuló algoritmus kiértékelése manuálisan annotált korpusszal való összevetés útján történik, és szokásosan maga az algoritmus is egy ilyen korpuszból tanulja meg a paramétereit automatikus módon. Ezért van szükség nagy méretű kézzel annotált korpuszokra. Ez a célja ennek a projektnek is, amelynek a tervezett kimenete a NerKor korpusz, egy egymillió tokenes kézzel NE-annotált korpusz.

A NE-k nem teljesen azonosak a tulajdonnevekkel, de a jelen annotációs sémában nem lépünk ki a nevek világán kívülre, így a 'NE' és a 'név' szavak szinonimaként fognak szerepelni a tárgyalási univerzumunkban.

A nevek egyedi referenciával bírnak, vagyis a világ egy egyedi entitására utalnak (pl. *London*). Ilyet a köznévi frázisok önmagukban állva (pl. *város*) nem tudnak csinálni, csak akkor, ha egyéb nyelvi elemekkel közösen egy olyan főnévi frázist alkotnak, amely már tényleg a világ egy egyedi entitására utal (pl. *az a város, ahol 9 millió brit lakik*). Mi itt csak a neveket annotáljuk, például:

(1) *Kosztolányi Dezső*

(2) *Szilas Menti Mezőgazdasági Termelőszövetkezet*

(3) *United Nations Educational, Scientific and Cultural Organization*

(4) *Déli-Shetland-szk.*

(5) *IBM*

(6) *Kiss János altábornagy utca*

(7) *Műegyetem*

(8) *The Coca-Cola Co.*

(9) *Kovács Pistike*

2. Az annotálás alapelvei

Fontos, az annotálás során végig szem előtt tartandó szabályok:

- Csak tulajdonneveket annotálunk. Nem annotálunk olyan frázisokat, amelyek ugyan a világnak valamely egyedi részére utalnak, de nem tulajdonnévvel. Például a *József Attila Gimnázium* annotálandó, de a szövegben szereplő az *a suli* frázis nem, hiába derül ki a szövegből, hogy melyik iskolára utal.

- A nevek nem kompozicionálisak. Mivel a nevek jelölete nem a név részeinek a jelöletéből áll össze, ezért a neveket nem bonthatjuk részekre az annotáláskor. Például a *Kossuth Lajos utca* egy földrajzi névként jelölendő, hiába van benne egy személynév. Ebből az következik, hogy mindig a leghosszabb nevet (a legkülsőbbet) jelöljük a jelölhetők közül.

- Nem annotálunk egymást átfedő vagy egymásba ágyazott neveket. Vagyis minden annotációnak be kell fejeződnie, mielőtt egy másik elkezdődik.

- A *tag-for-meaning* elvét követjük. Vagyis egy nevet mindig az aktuális kontextusnak megfelelő referenciája alapján annotálunk.

- Ha az azonosított név ragozott formában szerepel a szövegben, a raggal együtt, a teljes alakot annotáljuk.

- A nevek képzett alakjait nem jelöljük. Nem annotálandók tehát az olyanok, mint *magyarországi, fideszes, petőfieskedő*.

- Ha a név összetétel előtagja, és az összetétel alaptagja köznév, például *Horn-kormány, Tilos Rádió-hallgatók, TA-vezérigazgató*, akkor nem annotálandók névként.

- A névhez nem tartozik hozzá az esetleg előtte álló névelő. Kivétel az az eset, amikor a határozott névelő része a névnek, például *The Hague, The Times*.

- A név rövidítése (akronim, mozaikszó, monogram) is névként annotálandó.

- A szöveghez a névannotálás során nem nyúlunk hozzá, vagyis nem javítjuk ki a helyesírási hibákat, nem vonunk egybe különírt szavakat, és nem választunk szét egybeírtakat. Ha valamilyen éktelen hibát látunk az aktuálisan címkézendő névvel kapcsolatban, akkor azt külön fel kell jegyezni

az erre rendszeresített google táblázatban. Ez a szépirodalmi szövegekre nem vonatkozik: ott az az elv, hogy amit a szerző leírt, az sérthetetlen, vagyis nem kell kijavítani, és így a táblázatba sem kell felvenni.

3. NE-típusok

A következő típusokat annotáljuk:

PERSON: Valós és kitalált személyek neve, becenevek, művésznevek, álnevek. Ide tartoznak a kisebb, kevésbé strukturált embercsoportok, közösségek is.

ORGANIZATION: Olyan csoportok nevei, amelyek valamilyen szervezett struktúrával rendelkeznek, mint például intézmények, vállalatok, kormányzati hivatalok, sportcsapatok, múzeumok, egyetemek.

LOCATION: Földrajzilag vagy politikailag definiált helyek nevei, úgymint városok, országok, hegyek, völgyek stb. Idetartoznak az emberalkotta építmények is, mint a repterek, utak, gyárak, épületek stb.

MISC: A felsorolt típusok egyikébe sem tartozó nevek.

Az útmutatóban a NE-ket [szögletes zárójelek] közé tesszük. A példáknál csak az olyan típusú NE-ket jelöljük, amelyikről éppen szó van. A példákban a személyneveket a PER, a szervezetneveket az ORG, a helyneveket a LOC, a be nem sorolhatókat pedig MISC rövidítésekkel jelöljük.

3.1. Személynevek (PERSON)

Személyekre utalhatnak teljes személynevek, becenevek, művésznevek, álnevek, rövidítések. A kitalált személyek, úgymint mozihősök, mesefigurák, mitológiai alakok, illetve a szentek, bibliai alakok nevei is személynévként annotálandók, például:

(10) [Ady_{PER}] *írói álneve* [Ida_{PER}]

(11) *a legkisebb gyerek, aki gyakran játszik* [$Mikrobival_{PER}$]

(12) *zenéjével meglágyította* [$Hádész_{PER}$] *és* [$Perszephoné_{PER}$] *szívét*

(13) [$Isten_{PER}$] *szellem, legfelsőbb, láthatatlan és halhatatlan lény*

Bizonyos bibliai alakok neve jajgatáskor, panaszkodáskor vagy káromkodáskor is felmerül – az ilyen esetekben nem kell névként annotálni:

(14) *Úristen, milyen emberekkel élünk egy országban?*

A családnevek, az uralkodóházak nevei is személyekre, egészen pontosan személyek csoportjára referálnak, ezért azokat is személynévként kell megjelölni, például:

(15) *a* [*Széchényi*$_{PER}$] *család Nógrád megyéből származik*

(16) *a* [*Károlyiak*$_{PER}$] *Apáti nevű faluját felgyújtották*

A zenekarok, együttesek neve is személynévnek minősül, ha az adott kontextusban személyek csoportjaként van rájuk utalva, például:

(17) *A nagyszínpadon fellép többek közt a* [*Wellhello*$_{PER}$]

Ha viszont kifejezetten egy szervezett csoportként jelenik meg egy együttes a szövegben, aminek tagjai vannak, akkor inkább `ORG`-ként annotálandó, például:

(18) *1993-ban Erik a* [*Morbid Angel*$_{ORG}$] *tagja lett.*

Ha emberek bármilyen csoportjáról úgy vélekedünk, hogy az nem vagy kevésbé rendelkezik szervezett struktúrával, inkább informális és kicsi, akkor legyen `PER`-rel jelölve. De ha inkább érezzük dominánsnak a szervezetet, akkor legyen `ORG`. Ez a szabály vonatkozik minden embercsoportra, legyen az együttes, sportcsapat vagy valamilyen közös elv mentén összejött csoport.

A személynévvel együtt csak azok a prefixek vagy szuffixok annotálandók, amelyek hivatalosan is a névhez tartoznak, például magyarban a 'Dr.' prefix vagy a 'PhD' szuffix az igazolványba is bekerülhet, illetve a Sir', Lord' vagy Lady' is hivatalosan a névhez tartozik az angoloknál, valamint a 'Jr.' az amerikaiaknál. De az egyéb rangot, címet, beosztást vagy rokoni viszonyt jelölő köznévi szavak nem tartoznak hozzá, például:

(19) *Lemondott* [*Heinz-Christian Strache*$_{PER}$] *osztrák alkancellár*

(20) *Érvcsata* [*Dr. Boldogkői Zsolt*$_{PER}$] *molekuláris biológus és* [*Dr. Gődény György*$_{PER}$] *gyógyszerész között*

(21) [*Sir Winston Leonard Spencer Churchill*$_{PER}$] *brit politikus, miniszterelnök.*

(22) [*Gyuri*$_{PER}$] *bácsi gyógynövénytúrái*

(23) *Mr.* [*Joyce*$_{PER}$]*, jöjjön be!*

A köznévi megszólítások nem jelölendők névként, például:

(24) *Nagymama, miért ilyen nagyok a szemeid?*

A @-cal jelölt felhasználónevek is személynévnek minősülnek, így `PER`-rel jelölendők, de a @ nélkül. A # után következő token pedig csak akkor név, ha # nélkül az, és persze jelölni is anélkül kell.

(25) *@* [*MazenMahdi*$_{PER}$]*: A 2 idős férfi, akit letartóztattak tegnap #* [*karzakanban*$_{LOC}$]

3.2. Szervezetnevek (ORGANIZATION)

Azok a tulajdonnevek, amelyek egy szervezett struktúrával rendelkező csoportra referálnak, amelyek aktorként szerepelnek az adott szövegkontextusban, és olyanokat tudnak csinálni, mint döntést hozni, árat emelni, nyilatkozni valamiről stb., szervezetnévként annotálandók. A következők mind ilyenek:

- Cégek, vállalatok

  (26) *a [SERCO Kft.$_{ORG}$] az eltelt évek során jelentős fejlődésen ment keresztül*

  (27) *1878-ban Grosvernor Lowry-val létrehozzák az [Edison Electric Light Co.-t$_{ORG}$]*

- Multinacionális szervezetek

  (28) *az [Európai Unió$_{ORG}$] ezen a néven 1992-ben jött létre*

- Politikai pártok

  (29) *bántalmazták a [Fidesz$_{ORG}$] egyik ajánlószelvényeket gyűjtő aktivistáját*

- Sportcsapatok

  (30) *A [Budapest Black Knights$_{ORG}$] csapata fölényesen legyőzte a [Szolnok Soldiers$_{ORG}$] csapatát.*

- Katonai szervezetek

  (31) *Az [Észak-atlanti Szerződés Szervezete$_{ORG}$] székhelye Brüsszelben van.*

- Kórházak, egészségügyi intézmények

  (32) *A [Péterfy Kórház-Rendelőintézet Országos Traumatológiai Intézet$_{ORG}$] a Főváros egyik legnagyobb egészségügyi intézménye*

- Hotelek

  (33) *A paksiakat is várja az [Erzsébet Nagy Szálloda$_{ORG}$]*

- Színházak, múzeumok

  (34) *A [Szépművészeti Múzeum$_{ORG}$] az egyetemes és a magyar művészet emlékeit mutatja be az ókortól a 18. század végéig.*

- Egyetemek

(35) *A [Kossuth Lajos Tudományegyetem$_{ORG}$] Honoris Causa Doktorai.*

- Kormányzati hivatalok

(36) *A [Honvédelmi Minisztérium$_{ORG}$] hivatalos Facebook oldala*

Az online vagy nyomtatott sajtótermékek esetében szervezetként annotáljuk a nevet, amikor aktívan cselekvő szervezetként lépnek fel, például megjelentetnek valamit, vagy kiállnak valami mellett vagy ellen, például:

(37) *Mi, akik a [Telexnél$_{ORG}$] dolgozunk, vállaljuk, hogy*

(38) *Hétfőn a [Sabr Online$_{ORG}$] újság arról számolt be, hogy*

Ha viszont ugyanezt a sajtóterméket úgy említik, mint amit olvasnak, vagy mint amin/amiben megjelent egy cikk, tanulmány, írás, akkor `MISC`-ként annotálandó, lásd a 3.4. fejezetben.

Az általános intézményneveket, mint *rendőrség* vagy *kormány* nem annotáljuk, mert ezek csak egy főnévi frázis részeként tudnak egy egyedi entitást jelölni, nem egyedi jelölők.

Az olyan hosszú, többtagú intézményneveket, amelyek tartalmaznak köznévi elemet is, teljes egészében névnek kell jelölni, például:

(39) *Az [Állami Privatizációs és Vagyonkezelő Rt.$_{ORG}$] zártkörű részvénytársaságként működő állami vállalat volt.*

Ha egy intézménynév után zárójelbe téve szerepel a rövidítése is, akkor a teljes név és a rövidítés külön nevekként kezelendők, a zárójel pedig nem a név része, például:

(40) *Az autóipari óriás [DaimlerChrysler AG$_{ORG}$] ([DC$_{ORG}$]) amerikai részlege*

3.3. Helynevek (LOCATION)

A helynévnek annotálandó entitások közé tartoznak többek között a kontinensek, az országok, a régiók, a városok, a települések, a repterek, az utak, a gyárak, az óceánok, a tengerek, a folyók, a szigetek, a tavak, a nemzeti parkok, a hegyek és a mitikus helyek, például:

(41) [*Franciaországot$_{LOC}$*] *kilenc ország határolja.*

(42) [*Szihalom$_{LOC}$*] *község* [*Heves megye$_{LOC}$*] [*Füzesabonyi kistérségében$_{LOC}$*].

(43) *A [Bükki Nemzeti Park$_{LOC}$] mintegy 95 százalékát erdő borítja.*

(44) [*Gatwick$_{LOC}$*] *délre,* [*Stansted$_{LOC}$*] *észak-keletre,* [*Luton$_{LOC}$*] *északnyugatra fekszik* [*Londontól$_{LOC}$*].

(45) *Platón dialógusaiban részletesen szól* [*Atlantisz*$_{LOC}$] *szigetéről.*

Az utak nagy részének csak számozása van, de nincs neve – ezeket nem jelöljük. Névként jelöljük viszont azt, amikor szám helyett vagy mellett neve van az útnak, például:

(46) *Ünnepélyes keretek között tartották szerdán az 54. számú főút* [*M5*$_{LOC}$] *autópálya és az 5. számú főút közötti szakasz kapacitásbővítésének átadását*

Ha azonban az út neve után *-s* melléknévképző áll, akkor nem annotálandó névként, például:

(47) *Kisteherautó égett az M5-ösön*

A szervezeteknek sajátjuk, hogy van székhelyük, és előfordul, hogy a szervezet nevét helymegjelölésként használjuk. Ilyenkor a *tag-for-meaning* elve alapján a kontextusnak megfelelően helynévként annotáljuk őket, például:

(48) *tűz ütött ki a* [*Kapos Hotelben*$_{LOC}$]

(49) *elbarikádozták magukat az* [*SZFE-n*$_{LOC}$]

Ha egy geopolitikai entitás (birodalom, ország, megye, város stb.) aktívan cselekszik, valamilyen politikai cselekedeteket hajt végre, mint például megtámad egy másikat, vagy döntést hoz, elítél, nyilatkozik stb., akkor `ORG`-ként annotálandó. Ide tartoznak az ország- vagy városnévvel említett sportcsapatok is, például:

(50) *a* [*Spanyolország*$_{ORG}$]–[*Németország*$_{ORG}$] *összecsapást koordinálta*

(51) *a* [*Habsburg Birodalom*$_{ORG}$] *semleges maradt*

(52) *Az esetek 90%-ában* [*Brüsszel*$_{ORG}$] *javára dönt ez a bíróság.*

Ha viszont inkább földrajzi a kontextus, vagy egy geopolitikai entitás területén történnek az események, amiket az passzívan elszenved, akkor ezek is `LOC`-ként annotálandók:

(53) *a* [*Római Birodalom*$_{LOC}$] *legvégső határára bukkantak*

(54) [*Magyarország*$_{LOC}$] *közepe* [*Pusztavacs*$_{LOC}$] *község területén van.*

(55) *benyomultak a rend területére és elfoglalták* [*Graudenzt*$_{LOC}$]

Ha egy földrajzi név után más nyelven, más írásmóddal is szerepel a név, akkor mindkettőt jelöljük `LOC`-ként, de külön névként, és a köztük levő zárójelet és egyéb részeket kihagyva, például:

(56) [*Bécs*$_{LOC}$] *(németül:* [*Wien*$_{LOC}$]*, bajor nyelvjárás szerint:* [*Wean*$_{LOC}$]*)*

3.3.1. Összetett kifejezések

Az olyan összetett kifejezésekben, ahol földrajzi nevek vesszővel elválasztva szerepelnek, és a második név nagyobb helyre referál, tehát egyfajta pontosító funkciót tölt be, a neveket külön annotáljuk, például:

(57) [*Los Angeles*$_{LOC}$], [*Calfornia*$_{LOC}$]

(58) [*Budapest*$_{LOC}$], [*Magyarország*$_{LOC}$]

3.3.2. Köznévi tagok

Vannak olyan földrajzi nevek, melyek köznévi utótagot tartalmaznak. A közvetlenül a földrajzi név után álló köznévi frázisok, melyek nélkül a név nem ugyanaz lenne, a névvel együtt annotálandók, mint például az alábbiak:

(59) [*Váci utca*$_{LOC}$]

(60) [*Erzsébet híd*$_{LOC}$]

(61) [*Baranya megye*$_{LOC}$]

(62) [*Duna–Tisza köze*$_{LOC}$]

Nem tartoznak viszont a földrajzi névhez a magyarázó, pontosító funkciójú elemek, illetve az alkalmi jelzők sem, például:

(63) [*Kent*$_{LOC}$] *grófság*

(64) [*New York*$_{LOC}$] *állam*

(65) [*Gyöngyös*$_{LOC}$] *város*

(66) [*Mátra*$_{LOC}$] *hegység*

(67) [*Duna*$_{LOC}$] *folyó*

(68) *az olasz* [*Alpok*$_{LOC}$]

(69) *a lengyel* [*Magas-Tátra*$_{LOC}$]

(70) *a gyönyörű* [*Alpok*$_{LOC}$]

(71) *„Mit nekem te zordon* [*Kárpátoknak*$_{LOC}$]*...”*

3.4. Egyebek (MISC)

Ebbe a kategóriába kerülnek azok, amelyek NE-k, de a felsorolt kategóriák egyikébe sem illenek bele, mint a könyvcímek, újságnevek, konferencianevek, márkanevek, tőzsdeindexek nevei, programozási nyelvek, díjak, kampányok, kereskedelmi útvonalak, gázvezetékek, rádiócsatornák nevei, például:

(72) *A [Le Monde$_{MISC}$] francia napilapban jelent meg, hogy*

(73) *Az eset azonnal heves vitákat váltott ki a [Sina Weibón$_{MISC}$], a meghatározó kínai közösségi portálon.*

(74) *fedezze fel a [Fiat$_{MISC}$] modelleket*

(75) *Érdekel a [Python$_{MISC}$] programozás?*

(76) *Hitler megtiltotta, hogy német állampolgárok elfogadják a [Nobel-díjat$_{MISC}$].*

(77) *Újraindult a [Bringázz a munkába!$_{MISC}$] kampány*

(78) *így kereskedtek az ókorban a [Selyemúton$_{MISC}$]*

(79) *a [Déli Áramlat$_{MISC}$] gázvezeték éves kapacitása 63 milliárd köbméter*

(80) *a [Bartók Rádió$_{MISC}$] szeptemberi koncertjei*

A törvények, rendeletek, irányelvek stb. esetében, ha csak számmal és/vagy köznévi elemekkel vannak jelölve, akkor nem jelöljük őket. De ha kaptak valamilyen egyedi nevet, akkor azt névként jelöljük, például:

(81) *A [TEN-T$_{MISC}$] rendelet (1315/2013 rendelet) 2013 végi hatályba lépését követően*

(82) *ismételten változott a [Munka törvénykönyve$_{MISC}$] (2012. évi I. tv.)*

Annak eldöntésében, hogy mi micsoda az EU-s jogi szövegekben, hasznos segítség lehet ez a glosszárium.

Ha egy újságcím úgy van használva, hogy arra referál, amit olvasunk, akkor az `MISC`. De ha a szerkesztőségre vagy a lapvezetésre utal, mint akik megírnak vagy kiadnak valamit, akkor az `ORG`-ként jelölendő.

Az együttesek, zenekarok, zeneszerzők nevét sokszor használjuk az általuk létrehozott mű helyett – ilyenkor a nevüket `MISC`-ként annotáljuk, például:

(83) *[Bachot$_{MISC}$] hallgatunk.*

Az események, bajnokságok, csaták `MISC`-ként jelölendők, ha tulajdonnévvel referálunk rájuk, például:

(84) *[Waterloo$_{MISC}$] még ma is ott él mindannyiunkban*

(85) *A* [*Forma-1*$_{MISC}$] *legfrissebb hírei*

Ha viszont eseményekre köznevekkel van utalva, akkor azokat nem jelöljük névként, például:

(86) *2021-re halasztanák a 2020-as tokiói nyári olimpia megrendezését*

Ha egy nevet metanyelvi szinten használunk, vagyis magáról a névről szól a szöveg, akkor `MISC`-ként annotálandó, például:

(87) *Az elnevezés továbbra is* [*Tokió 2020*$_{MISC}$] *lesz.*

(88) *Nekem rettentően tetszik fiúnévként az* [*Arion*$_{MISC}$]

(89) *ugyanebben az évben kezdték használni a* [*Cryptopsy*$_{MISC}$] *nevet*