Az NYTK-NerKor több szempontú kiértékelése

Simon Eszter[1], Vadász Noémi[1], Lévai Dániel[2,3], Nemeskey Dávid[2,3], Orosz György[4], Szántó Zsolt[4]

[1]Nyelvtudományi Kutatóközpont
1068 Budapest, Benczúr u. 33.
[2]Eötvös Loránd Tudományegyetem Bölcsészettudományi Kar
TI Digitális Bölcsészet Tanszék
1088 Budapest, Múzeum krt. 6-8.
[3]Digitális Örökség Nemzeti Laboratórium
1088 Budapest, Múzeum krt. 6-8.
[4]Szegedi Tudományegyetem Informatikai Intézet
6725 Szeged, Tisza Lajos körút 103.
simon.eszterke@gmail.com, vadasz.noemi@nytud.hu
levai.daniel@btk.elte.hu, nemeskey.david@btk.elte.hu
gyorgy@orosz.link, szantozs@inf.u-szeged.hu

Kivonat Cikkünkben az NYTK-NerKor korpusz kiértékelését mutatjuk be több rendszer segítségével. Azt vizsgáljuk, hogy az egymillió tokent tartalmazó, műfajilag heterogén, szabadon elérhető gold standard adathalmaz mennyire használható magyar nyelvű tulajdonnév-felismerő rendszerek fejlesztéséhez. A kiértékeléshez négy különböző rendszert használtunk: a CRFsuite-ot, a magyar spaCy-t, a Stanzát és az `emBERT`-et. Cikkünkben ismertetjük az egyes rendszerek által elért eredményeket, melyeket össze is hasonlítunk. Az eredmények azt mutatják, hogy az NYTK-NerKor és a Szeged NER korpusz együttes használata még stabilabb modelleket eredményezhet, valamint hogy az NYTK-NerKoron tanítva a rendszerek nagyobb általánosító képességgel rendelkeznek, ami ahhoz kell, hogy egy azelőtt nem látott szövegben jól azonosítsák a neveket.
Kulcsszavak: tulajdonnév-felismerés, kiértékelés, korpusz

1. Bevezetés

A felügyelt gépi tanuláson alapuló statisztikai és neurális rendszerek nagy mennyiségű gold standard adatot igényelnek. Ahhoz, hogy egy adathalmaz gold standard korpusznak minősüljön, több feltételnek is meg kell felelnie, úgymint reprezentativitásra kell törekednie, elég nagynak kell lennie gépi tanuló rendszerek tanításához és teszteléséhez, valamint kézzel hozzáadott pontos nyelvi annotációt kell tartalmaznia. Az ilyen korpuszok építése viszont sok időt és hozzáértést igényel, ezért a gold standard adathalmazokból kevés van, és rendkívül értékesek a természetesnyelv-feldolgozásban (Natural Language Processing, NLP). A magyar nyelvű tulajdonnév-felismerés (Named Entity Recognition, NER) terén is

hasonló volt a helyzet, mivel a meglevő névannotált korpuszok erősen domainspecifikusak (jellemzően csak híreket tartalmaznak), és korlátozottak méretükben, illetve hozzáférhetőségükben.

A jelenleg elérhető magyar nyelvű gold standard tulajdonnév-annotált korpuszok közül a legismertebb a Szeged NER korpusz (Szarvas és mtsai, 2006b), amely kizárólag gazdasági rövidhíreket tartalmaz, és összesen kb. 225 000 tokenből áll. A CoNLL2003 shared task (Tjong Kim Sang és De Meulder, 2003) annotációs sémáját és címkekészletét követi. A korpusz szövege a Szeged Treebankből (Csendes és mtsai, 2005) lett válogatva, annak egy alkorpusza, így annak a licencét örökíti tovább, vagyis csak kutatási célokra lehet használni.

A másik a Criminal NE Korpusz[1], amely gazdasági bűncselekményekről szóló HVG-cikkekből áll, és kb. 560 000 tokent tartalmaz. Ez a korpusz a Magyar Nemzeti Szövegtár (Váradi, 2002) alkorpusza, ezért felhasználhatósága még inkább korlátozott. Ez is a CoNLL2003 címkekészletét követi, de azzal a különlegességgel, hogy az annotációnak két verziója van. Az egyik az ún. *tag-for-meaning*, a másik a *tag-for-tagging* elvet követi. Egyes nevek bizonyos kontextusokban metonimikusan viselkednek, ami számos érdekes kérdést vet fel már a korpusz címkézése során. Két megközelítés létezik ennek a jelenségnek a kezelésére. Az első szerint – ez a *tag-for-meaning* – a nevet az aktuális kontextusának megfelelően annotáljuk. Ebben az esetben abban a mondatban, hogy *Az esetek 90%-ában Brüsszel javára dönt ez a bíróság*, a 'Brüsszel' intézménynévként címkézendő, mivel itt egy jogi entitásként, egy cselekvő félként szerepel. A *tag-for-tagging* elv alapján ugyanez a név ugyanebben a mondatban földrajzi névként címkéződik, mivel az az elsődleges referenciája.

A fent leírt gold standard korpuszok mellett létezik egy silver standard korpusz is. A hunNERwiki korpusz (Simon és Nemeskey, 2012) automatikusan lett generálva a magyar Wikipédiából, ugyanazt az annotációs sémát követi, mint a Szeged NER korpusz, de több mint 19 millió tokenből áll. A Wikipédia licencét követve ez szabadon felhasználható CC-BY-SA 3.0 licenc alatt.

Az NYTK-NerKor korpusz (Simon és Vadász, 2021) a fentiektől több paraméterében is különbözik. Méretét tekintve egy nagyságrenddel nagyobb, mint az eddigiek, ugyanis 1 millió tokenből áll. Gold standard korpusz, vagyis a címkézés kézzel lett ellenőrizve. Kiegyensúlyozott válogatást nyújt többféle domainből: tartalmaz szépirodalmi, jogi és vegyes webes szövegeket, híreket, valamint Wikipédia cikkeket is. Egy kb. 200 000 tokennyi alkorpusz gold standard morfológiai címkézést is kapott, hogy a klasszikus statisztikai gépi tanuláson alapuló rendszerek morfológiai jegyei is biztosítva legyenek. Az adatformátum követi a nemzetközi sztenderdeket, ugyanis a széles körben ismert és alkalmazott CoNLL-U Plus[2] formátumban van, a névannotáció a CoNLL2002 (Tjong Kim Sang, 2002) címkézési szabványt követi, és a morfológiai információ a Universal Dependencies[3] v2 szófajkódjaival és jegy–érték párjaival van kódolva. Ezenfelül az NYTK-NerKor újdonsága, hogy CC-BY-SA 4.0 licenc alatt fel-

[1] `https://rgai.inf.u-szeged.hu/node/130`

[2] `https://universaldependencies.org/ext-format.html`

[3] `https://universaldependencies.org/`

használható bármilyen célra, és szabadon elérhető a GitHub repozitóriumából: `https://github.com/nytud/NYTK-NerKor`.

A META-NET Fehér könyvek sorozatának magyar nyelvről szóló kiadványa (Simon és mtsai, 2012) alapján a magyar az erőforrásokkal közepesen jól ellátott nyelvek közé tartozik, vagyis feltételezhetjük, hogy a helyzet a tulajdonnév-annotált korpuszok tekintetében is hasonló. Több olyan aggregátor weboldal is létezik, amely különféle nyelvekre elérhető erőforrásokat listáznak, mint például a CLARIN tudástára[4]. Ezt az oldalt áttanulmányozva sok névannotált korpuszt találunk a különféle európai nyelvekre, 46 000-től 1 millió tokenig terjedő méretben, amibe az NYTK-NerKor kiválóan illeszkedik. Az erőforrásokkal leginkább ellátott angol nyelvre a legismertebb adathalmaz az OntoNotes 5.0[5], amelynek az angol nyelvű része kb. 1,5 millió tokent tartalmaz.

Összefoglalva az NYTK-NerKor jelenleg a legnagyobb magyar gold standard tulajdonnév-annotált korpusz, amilyen azelőtt nem volt, annak ellenére, hogy igény lett volna rá. Ezt az is jól mutatja, hogy megjelenése után nem sokkal többen használatba is vették. A korpusz GitHub repozitóriumának kérdései alapján az ELTE RC2S2 kutatócsoportja[6] és a Stanford NLP Group[7] használják. Cikkünkben a korpusz alapos kiértékelését mutatjuk be több tulajdonnév-felismerő rendszert használva tanításra és tesztelésre. A 2. fejezetben először az NYTK-NerKor releváns tulajdonságait ismertetjük, majd a 3. fejezetben a kiértékelések eredményeit írjuk le. A cikket összegzés zárja a 4. fejezetben.

2. A korpusz ismertetése

Ebben a fejezetben az NYTK-NerKor korpusznak csak azon tulajdonságait ismertetjük, amelyek relevánsak a kiértékelés szempontjából. További részletekért a korpuszt bemutató cikkhez (Simon és Vadász, 2021) utaljuk az olvasót.

Az NYTK-NerKor korpusz 5 műfajból tartalmaz egyenletes szövegválogatást: szépirodalom, jogi szövegek, hírek, vegyes webes szövegek és Wikipédia. Az egymillió token egyenletesen oszlik el a műfajok között, vagyis minden műfaj kb. 200 000 tokent tartalmaz.

A korpusz fő annotációhalmaza a named entity (NE). A 2002-es és 2003-as CoNLL shared taskok sztenderd címkekészletét használja, ami 4 fő névkategóriát különít el: `PER`, `ORG`, `LOC`, `MISC`. A címkeprefixek tekintetében a CoNLL2002 annotációs formátumát, az ún. `IOB2` formátumot követi, miszerint minden név első eleme '`B-`' prefixet, míg minden nem első elem '`I-`' prefixet kap. A nem neveket `O` betű jelöli.

A NE annotáció a teljes szöveget lefedi. Ez praktikusan azt jelenti, hogy minden tokenhez tartozó cella ki van töltve. Minden ugyanolyan értékű tokennek számít, az írásjelek is. A folyó szöveg tokenekre bontásakor bizonyos tapadó

[4] `https://www.clarin.eu/resource-families/manually-annotated-cor-pora#Named%20Entity%20recognition`

[5] `https://catalog.ldc.upenn.edu/LDC2013T19`

[6] `https://rc2s2.elte.hu/`

[7] `https://nlp.stanford.edu/`

írásjeleket a tokenizáló leválaszt az előtte-utána álló szóról, így azok is önálló tokenné válnak. Minden, ami nem név, `O` annotációt kap, így az írásjelek is. A mondathatárt jelölő üres sorok viszont ténylegesen üres sorok.

Mivel a klasszikus gépi tanuló rendszerek számára a morfológiai információk hasznos jegyeket jelentenek, a korpusz egy részére morfológiai annotáció is került. A Szeged NER korpusz esetében a kb. 200 000 tokennyi tanítóanyag elégnek bizonyult 96% feletti F-mérték eléréséhez, ezért az NYTK-NerKor esetében is kb. ennyi szöveganyag kapott kézzel ellenőrzött egyértelműsített morfológiai elemzést.

A korpusz egyes alkorpuszai különféle előfeldolgozási lépéseken estek át, és került rájuk automatikus morfológiai és NE annotáció is. A szépirodalmi és jogi szövegek, valamint a hírek az `emtsv`-vel (Indig és mtsai, 2019) lettek elemezve, a mondatra bontástól a morfológiai egyértelműsítésig. A NE előelemzéshez szintén az `emtsv` volt használva, annak az `emBERT-NER`, illetve az `emNER` moduljai. A kézi annotáció így tehát az automatikus előannotáció ellenőrzésévé és javításává egyszerűsödött. A Wikipédia alkorpusz szövegei két forrásból származnak: egyrészt a KorKorpusz (Vadász, 2020) morfológiailag annotált szócikkeiből, másrészt a hunNERwiki korpuszból (Simon és Nemeskey, 2012). A KorKorpuszból érkezett szövegek eleve fel voltak dolgozva, csak NE előcímkézést igényeltek, amihez szintén az `emBERT-NER` lett használva. A hunNERwikiből érkező szövegek esetében sem volt szükség előfeldolgozásra, hiszen az ott szereplő szövegek eleve mondatokra és tokenekre voltak bontva. NE előcímkézésre sem volt szükség, hiszen a hunNERwiki korpusz tartalmaz silver standard névannotációt, ráadásul ugyanazokat a névkategóriákat használja, mint amiket az NYTK-NerKor. A webes szövegek forrása teljes egészében a Magyar Webkorpusz 2.0 (Nemeskey, 2020b), ami eleve elemezve volt az `emtsv`-vel, így csak NE előcímkézést igényelt, ami az `emBERT-NER`-rel készült.

A korpusz kétféle morfológiai annotációt is tartalmaz. Egyrészt az `emtsv` `emMorph` (Novák és mtsai, 2016) modulja által kiadott elemzést, másrészt a Universal Dependencies v2 szófajkódjait és morfoszintaktikai jegy–érték párjait.

A korpuszhoz készült egy hivatalos train–devel–test vágás is. Az egyes halmazok nagyjából aránya: 80%–10%–10%. A vágás törekszik a kiegyensúlyozottságra, vagyis minden műfaj, forrás és morfológiai annotáltság ugyanilyen arányban van képviseltetve. A vágásnál figyelembe lettek véve a dokumentumhatárok, vagyis csak egész fájlok kerültek az egyes halmazokba.

3. Kiértékelések

A kiértékeléshez négy különböző rendszert használtunk: a CRFsuite-ot (Okazaki, 2007), a HuSpaCy-t (Orosz és mtsai, 2022), a Stanzát (Qi és mtsai, 2020) és az `emBERT`-et (Nemeskey, 2020a). Az első rendszer klasszikus gépi tanulást valósít meg, míg az utóbbi három rendszer neurális architektúrát alkalmaz. A spaCy-nek és az `emBERT`-nek közös vonása, hogy már az NYTK-NerKor korpusz létrejötte előtt is volt magyar nyelvű tulajdonnév-felismerője, míg a Stanza fejlesztői csak az NYTK-NerKoron tanítva készítettek magyar NER modult. Ez

utóbbi esetben a kiértékelést az indokolta, hogy képet kapjunk ennek az új rendszernek a teljesítményéről, míg az előbbi rendszerek esetében azt vártuk, hogy a nagyobb és heterogénebb korpuszon tanítva jobb teljesítményt fognak nyújtani, mint az eddig elérhető, de jóval kisebb és domainspecifikusabb korpuszokon. A CRFsuite-ot kimondottan baseline-ként használtuk, hogy összehasonlítsuk a neurális rendszerek teljesítményét egy klasszikus gépi tanuláson alapuló rendszerével.

Mindegyik rendszer esetében ugyanazokat a méréseket végeztük el: egyrészt tanítottuk és teszteltük az NYTK-NerKor hivatalos vágásán, ugyanezt megtettük a Szeged NER Korpuszon is, illetve ezeket keresztbe is mértük, továbbá a két korpuszt együtt is használtuk tanításhoz és teszteléshez. Ezenfelül az NYTK-NerKor egyes alkorpuszain is végeztünk méréseket.

Az NYTK-NerKornak van hivatalos vágása (lásd a 2. fejezetet) – ott azt használtuk. A Szeged NER korpusznak ugyan nincs hivatalos vágása, de a magyar NLP közösség hagyományosan azt a vágást használja tanításhoz–teszteléshez, amelyet Szarvas és mtsai (2006a) használtak a rendszerük építéséhez. Mivel a Szeged NER korpusz a Szeged Treebank egy alkorpusza, annak a morfológiai elemzését tartalmazta eredetileg, amit át kellett konvertálni az `emmorph` címkekészletére. Ezt a konvertált verziót használtuk minden kiértékelésnél.

Címkekészlet tekintetében a kiértékeléshez használt Szeged NER korpusz és az NYTK-NerKor némileg eltérnek. A fő névkategóriák (`PER`, `ORG`, `LOC`, `MISC`) megegyeznek, de a címkeprefixek terén van eltérés. A Szeged NER korpusz a `BIE1` címkeprefixálást követi, amely megkülönbözteti a kezdő (`B-`), a közbülső (`I-`), a záró (`E-`) és az egyelemű (`1-`) névelemeket. A Stanza és a spaCy is ezt használja, csak más névvel illeti: a Stanzánál ugyanez `BIOES` (Beginning, Inside, Outside, Ending, Single), a spaCy-nél `BILOU` (Beginning, Inside, Last, Outside, Unit-length). Ezzel szemben az NYTK-NerKor a CoNLL2002-es `IOB2` címkeprefixálási formátumot alkalmazza (lásd a 2. fejezetet). A két formátum információvesztés nélkül átjárható – a kiértékelésnél a korpuszokat természetesen átkonvertáltuk a kellő formátumra.

Az egyes rendszerek teljesítményét a 3.1., a 3.2., a 3.3. és a 3.4. fejezetekben ismertetjük, összehasonlító összegzést pedig a 3.5. fejezetben adunk. A rendszerek teljesítményét mindenhol a tulajdonnév-felismerésben szokásosan entitás szinten számoltuk, a szintén szokásos F-mértékkel, amelyet százalékos formában szerepeltetünk a táblázatokban. A tanítás fő paramétereit az 1. táblázat tartalmazza.

3.1. CRFsuite

A neurális rendszerekkel szemben baseline-ként egy manuálisan kinyert jellemzőkre építő CRF (Lafferty és mtsai, 2001) alapú megoldást választottunk. Ehhez a gyors implementációval rendelkező, C nyelven íródott CRFsuite könyvtárat használtuk fel.

Jellemzőként a szó kisbetűs változata mellett annak kettő és három hosszú szuffixe, valamint felszíni jellemzői (nagybetűs-e, nagybetűvel kezdődik-e, szám-e) voltak felhasználva, illetve ugyanezen tulajdonságok az azt megelőző és rákö-

Rendszer	Architektúra	CPU/GPU	Tanítás		Tesztelés	
			Batch	Idő (min:s)	Batch	Idő (s:ms)
CRFsuite	CRF	Xeon 5218	1	1:25	1	1:18
`emBERT`	BERT	1db A100	10	40:20	16	15
HuSpaCy	CNN	1db A100	1024	120	1024	1:47
Stanza	Flair	1db A100	4096	360	4096	7:13

1. táblázat. A négy rendszer főbb tanítási paraméterei. A 'Tanítás' oszlop a teljes tanítás idejére vonatkozik, a 'Tesztelés' az NYTK-NerKor tesztkorpuszának egyszeri annotálására.

vetkező szóra. Tanításhoz az L-BFGS (Nocedal, 1980) algoritmust használtuk, 0,1-es L1 és L2 regularizációs paraméterekkel, és az iterációk számát 100-ban limitáltuk. Az így kapott eredmények a 2. táblázatban láthatók.

	NerKor	SzegedNER	Együtt
NerKor	75,22	67,81	73,65
SzegedNER	45,95	93,42	56,44
Együtt	75,12	92,89	79,04

2. táblázat. A CRFsuite teljesítménye különböző korpuszokon tanítva és kiértékelve. A sor mutatja a tanító-, az oszlop a tesztkorpuszt.

3.2. HuSpaCy

A természetesnyelv-feldolgozó keretrendszerek közül az egyik legelterjedtebb a spaCy. Népszerűségét a relatív korai indulásának, könnyű felhasználhatóságának és erőforrásbarát felépítésének köszönheti. Jelenleg 64 nyelvet támogat alap szinten, ami elsősorban nyelvspecifikus tokenizálást és stopszószűrést jelent. Ezen belül 19 nyelven érhetők el hivatalos modellek olyan magasabb szintű feladatokra, mint a szófaji egyértelműsítés, szintaktikai elemzés vagy esetünkben a tulajdonnév-felismerés. A magyar nem tartozik ezek közé, ugyanis hivatalos modellek nem léteznek rá, csak külső fejlesztések, amelyek között viszont elérhető a spaCy 3. főverziójához egy tulajdonnév-felismerő modul is. Mi a jelen kísérletekhez ezt az új spaCy-re épülő eszközt, a HuSpaCy-t használtuk fel (Orosz és mtsai, 2022).

A spaCy egy moduláris mély neuronhálós architektúrát alkalmaz, aminek a legalsó szintjén egy szóbeágyazási réteg található. A mi esetünkben ez a réteg két részre osztható: egy 300 dimenziós előre tanított szóbeágyazásra, ami a

CBOW (Mikolov és mtsai, 2013) algoritmus segítségével a Magyar Webkorpuszon (Halácsy és mtsai, 2004) és a magyar Wikipédián lett tanítva; továbbá egy 256 dimenziós, szóalakokra építő alrétegre, amely a szó mellett annak prefixét, szuffixét és alakját is kódolja 64–64 dimenzióban. Egy négyrétegű CNN-alapú encoder épül a beágyazások fölé. A hálózat legtetején pedig a névcímkék prediktálásáért egy átmenetalapú elemző felel. Bár a spaCy képes beolvasni mind az `IOB2`, mind a `BILUO` formátumokat, a háttérben a `BILUO` címkerendszert alkalmazza.

A 3. táblázat mutatja a HuSpaCy teljesítményét a különböző korpuszokon tanítva és tesztelve.

	NerKor	SzegedNER	Együtt
NerKor	80,75	79,13	80,39
SzegedNER	58,80	95,31	66,61
Együtt	80,59	93,86	83,46

3. táblázat. A HuSpaCy teljesítménye különböző korpuszokon tanítva és kiértékelve. A sor mutatja a tanító-, az oszlop a tesztkorpuszt.

3.3. Stanza

A Stanza a Stanford NLP Group egyik Python-alapú szövegfeldolgozó eszköze. A teljes Stanza pipeline neurális hálókon alapul. 66 nyelvet támogat jelen pillanatban, de könnyen hozzá lehet adni újabb nyelveket is, ugyanis teljesen a Universal Dependencies formátumán alapul minden eszköze.

A Stanza NER modulja kontextualizált sztringreprezentáción alapuló szekvenciataggert (Akbik és mtsai, 2018) használ. Egy előre feltanított karakterszintű LSTM-modell (Hochreiter és Schmidhuber, 1997) már be van építve a Stanzába, amit a Ginter és mtsai (2017) adathalmazon feltanítottak, és címkézéskor konkatenáljuk a reprezentációkat szavanként mindkét irányból az előre feltanított szóbeágyazást használva. A reprezentáció ezután egy Bi-LSTM szekvenciataggerbe kerül be, aminek a kimenetét egy CRF-alapú (Sutton és McCallum, 2007) dekóder alakítja névelemcímkékké.

Mivel a Stanza minden nyelvhez ugyanazt az architektúrát használja, egyszerűen lehet a tanítási folyamatot elindítani. A fejlesztők adnak receptet a NER modellek tanításához – mi is az előre megadott receptet használtuk. A modellek alapértelmezett bemeneti formátuma JSON, amiben egy-egy szóhoz meg van adva a *text* és a *ner* mező. A Stanza belső címkeprefixálási formátuma a `BIOES`, amire a konvertálást automatikusan el lehet végezni egy beépített függvénnyel[8] a `stanza` Python csomagból.

[8] `stanza.utils.datasets.ner.prepare_ner_file.process_dataset`

A Bi-LSTM rétegek számát és méretét lehet változtatni; adathalmazonként 3 modellt tanítottunk: egyet az alapbeállítással, azaz 1 Bi-LSTM réteggel, mely 256 egységből áll, egyet 2 réteggel és rétegenként 512 egységgel, és egyet 4 réteggel, szintén 512 egységgel rétegenként. A tanítást NVIDIA A100-as GPU-kon végeztük. A modellek kezdetben 1.0-es tanulási rátával tanultak, a minimum tanulási ráta 0.01 volt, amint ezt az ütemezővel elérték, leállt a tanulás. A tanítási folyamatok 6-24 órát vettek igénybe, mérettől függően. Az eredményeket a 4. táblázat tartalmazza.

	Méret	NerKor	SzegedNER	Együtt
NerKor	1*256	80,53	75,23	79,25
NerKor	2*512	79,42	73,06	77,90
NerKor	4*512	79,25	72,84	77,66
SzegedNER	1*256	49,18	91,78	60,57
SzegedNER	2*512	44,54	89,52	57,25
SzegedNER	4*512	50,85	90,32	61,15
Együtt	1*256	**80,66**	**92,90**	**83,75**
Együtt	2*512	80,07	91,65	82,97
Együtt	4*512	79,55	90,62	82,27

4. táblázat. A Stanza teljesítménye különböző korpuszokon tanítva és kiértékelve. A sor mutatja a tanító-, az oszlop a tesztkorpuszt.

Jól láthatóan a modellek által elért F-mérték független a modell méretétől. Erre a legvalószínűbb magyarázat az, hogy egyrészt a Stanza viszonylag egyszerű architektúrát használ osztályozásra, másrészt pedig a prediktált címke legfeljebb olyan jó, mint a beérkező reprezentáció.

3.4. `emBERT`

Az `emBERT` az `emtsv` egyik modulja, amely lehetővé teszi BERT (Devlin és mtsai, 2019) alapú tokenszintű osztályozók integrálását a szövegfeldolgozó láncba. A modul maga nyelvfüggetlen, azonban az `emtsv` részeként csak magyar tulajdonnév- és főnévicsoport-felismerésre lett feltanítva. Magyar nyelvre mindkét feladaton az `emBERT` számít a legjobb teljesítményű rendszernek (Nemeskey, 2021).

Az `emBERT` tanításkor egy kész BERT modellt vesz alapul, aminek a kimenetére egy softmax osztályozót köt. A tanítás folyamán a két komponenst együtt finomhangolja. Mi alapmodellnek a `huBERT`-et választottuk, amely a Magyar Webkorpusz 2.0-n lett előtanítva, és jelentősen jobb eredményt lehet vele elérni, mint a többnyelvű BERT-tel (Nemeskey, 2021). A jósolt címkeszekvencia konzisztenciáját egy Viterbi-algoritmus biztosítja, ami csak az érvényes címkeátmeneteket engedélyezi. Az `emBERT` mind az `IOB2`, mind a `BIOES` formátumot támogatja. Az

NYTK-NerKoron belüli méréseket a korpusz natív `IOB2` formátumán végeztük. A korpuszok közötti keresztmérésekhez az NYTK-NerKort `BIE1` formátumba konvertáltuk, hogy a címkekészlete megegyezzen a Szeged NER-ével.

Hogy teljesebb képet kapjunk a korpuszról, többféle kiértékelést csináltunk. Először a teljes tanítóanyagon tanítottunk, majd az egyes műfajoknak megfelelő teszthalmazokon teszteltünk. Ennek eredményét mutatja az 5. táblázat. Másodszor az egyes műfajokhoz tartozó alkorpuszok tanítóhalmazán tanítottunk és teszthalmazán teszteltünk – ennek eredményeit a 6. táblázat mutatja. A tanításhoz egy NVIDIA A100 GPU-t használtunk. Minden modell négy iterációt tanult; a futási idő az alkorpuszokon 5–12, a teljes korpuszon 40 perc volt.

	fiction	legal	news	web	wikipedia	**teljes**
`LOC`	90,32	86,62	92,01	89,72	96,11	93,42
`MISC`	52,94	85,46	80,54	68,34	76,83	76,57
`ORG`	62,50	95,87	87,89	82,30	86,80	90,72
`PER`	97,25	95,24	98,06	87,08	95,99	96,14
átlag	93,88	93,50	91,13	80,88	92,85	91,44

5. táblázat. A teljes NYTK-Nerkoron tanított `emBERT` modell teljesítménye a különböző alkorpuszokon.

	fiction	legal	news	web	wikipedia
`LOC`	90,67	84,97	89,83	85,97	95,96
`MISC`	51,61	83,54	81,85	59,39	82,20
`ORG`	57,14	94,89	87,40	80,99	87,10
`PER`	96,74	100,00	97,88	88,64	96,00
átlag	93,29	92,40	90,64	77,57	93,37

6. táblázat. Az `emBERT` teljesítménye a különböző alkorpuszokon tanítva és tesztelve.

A táblázatok számaiból több következtetés is levonható. Az egyik legszembetűnőbb, hogy a `MISC` kategória felismerése a legnehezebb. Ez indokolható egyfelől a kategória heterogenitásával, másfelől a többinél alacsonyabb előfordulási gyakorisággal: a teljes tanítóadatban összesen 4 604 `MISC` típusú entitás található, ami kevesebb, mint a fele a második legritkább entitástípus, az `ORG` számának (9 657). Szintén a gyakoriság (jobban mondva annak hiánya) magyarázhatja a

szépirodalmi alkorpuszban a MISC és az ORG kiugróan alacsony, 50-60%-os F-mértékét. Itt a két entitástípus 134-szer, illetve 112-szer fordul elő, míg LOC típusú névelemből 695, PER-ből pedig 3 674 van.

Az alkorpuszok között a legrosszabb átlag F-mértéket a webes szövegeken mértük. Ez egyrészt meglepő lehet annak fényében, hogy a huBERT a Magyar Webkorpusz 2.0-n készült, másrészt bizonyítja a BERT architektúra általánosító képességét. Ez az eredmény feltehetően annak köszönhető, hogy a másik négy alkorpusszal ellentétben ezek a szövegek nem szerkesztettek, és a minőségük hullámzó. Az egyes kategóriák közül itt is negatív irányba lóg ki a MISC, annak ellenére, hogy a webes alkorpuszon belül ez a második leggyakoribb entitástípus. A rendszer a Wikipédia alkorpuszon teljesített a legjobban, amit az magyarázhat, hogy az enciklopédia szövegek meglehetősen kötött nyelvezetűek.

Az NYTK-NerKor és a Szeged NER korpusz összehasonlítása céljából keresztméréseket is végeztünk a két korpusz, illetve az uniójuk között. A 7. táblázat mutatja az eredményeket. Az átlót megvizsgálva több dolog is feltűnik. Az NYTK-NerKor esetén elért F-mérték 0,65 százalékponttal meghaladja a 6. táblázatban lévőt, ami arra utal, hogy a BIE1 címkekészleten könnyebb tanulni. Ez valószínűleg annak köszönhető, hogy a részletesebb címkekészlettel a rendszer könnyebben meg tudja tanulni az egyes névelemekre jellemző tulajdonságokat. A Szeged NER-en mért érték hibahatáron belül van a Nemeskey (2021)-ben publikált értékhez (97,62) képest, ami felfogható egy „sanity check”-nek. Végül az unió F-mértéke nagyjából a két érték relatív korpuszmérettel súlyozott átlaga körül van.

	NerKor	SzegedNER	Együtt
NerKor	92,09	91,61	91,97
SzegedNER	81,01	97,40	84,39
Együtt	91,84	97,25	92,99

7. táblázat. Az emBERT teljesítménye különböző korpuszokon tanítva és kiértékelve. A sor mutatja a tanító-, az oszlop a tesztkorpuszt.

A korpuszok közötti keresztmérések az előre sejthető eredménnyel zárultak. Bár minden modell a saját korpuszán teljesít a legjobban, az NYTK-NerKor viszonylag jó pontszámot ért el a Szeged NER-en, míg fordítva ez nem áll. Az együtt tanított modell pedig, a két korpuszon kiértékelve, 0,15–0,25 százalékpontra megközelíti azok saját modelljeit, mutatva, hogy az NYTK-NerKor és a Szeged NER korpusz együttes használata még stabilabb modelleket eredményezhet.

3.5. Diszkusszió

A könnyebb összehasonlíthatóság kedvéért egy közös táblázatba rendeztük az egyes rendszerek által elért összehasonlítható eredményeket. A 8. táblázat soraiban a tanító–teszt korpuszok, míg az oszlopaiban az egyes rendszerek szerepelnek. A Stanza esetében az alapbeállítással (1*256 egység) elért eredményeket másoltuk ide, mert azok bizonyultak a legjobbnak.

	CRFsuite	spaCy	Stanza	`emBERT`
NerKor–NerKor	75,22	80,75	80,53	92,09
NerKor–SzegedNER	67,81	79,13	75,23	91,61
NerKor–Együtt	73,65	80,39	79,25	91,97
SzegedNER–NerKor	45,95	58,80	49,18	81,01
SzegedNER–SzegedNER	93,42	95,31	91,78	97,40
SzegedNER–Együtt	56,44	66,61	60,57	84,39
Együtt–NerKor	75,12	80,59	80,66	91,84
Együtt–SzegedNER	92,89	93,86	92,90	97,25
Együtt–Együtt	79,04	83,46	83,75	92,99

8. táblázat. Az egyes rendszerek teljesítménye az egyes korpuszokon.

A rendszerek összehasonlításából minden mérés esetében az `emBERT` kerül ki győztesen. Ez feltehetően köszönhető egyrészt a modell nagyobb kapacitásának (300 helyett 768 dimenziós beágyazás, a Stanza és a spaCy 2-4 rétege helyett 12, egyenként 3072 dimenzióban stb.), másrészt annak, hogy az attention modell az LSTM-mel és a CNN-nel szemben minden egyes tokennél az egész mondatra közvetlen rálátást biztosít. Az `emBERT`-től elmaradva, de szinte minden értékben a spaCy a második legjobb, a Stanza a harmadik, míg a CRFsuite produkálta a legalacsonyabb F-mértékeket. Ez rímel az angol tulajdonnév-felismerés tapasztalataira: az NYTK-NerKorral méretben nagyjából összevethető OntoNotes 5.0 eredménylistáját is egy BERT modell vezeti 92%-kal[9], míg a spaCy 85%[10], a Stanza pedig 89% körül teljesít (Qi és mtsai, 2020).

A korpuszok közötti mérések esetében az eredmények rendre az elvárásoknak megfelelően alakultak, vagyis a saját korpuszon való tanítás és tesztelés adta a legjobb eredményeket, és a korpuszok közötti keresztmérések adták a legrosszabbakat. A legalacsonyabb teljesítményt akkor kaptuk, amikor a Szeged NER korpuszon tanítottunk, és az NYTK-NerKoron teszteltünk, ami azzal magyarázható, hogy a Szeged NER korpusz kicsi és domainspecifikus – nem igazán ad lehetőséget arra, hogy a rendszer más műfajú szövegekre is általánosítson.

Bár a spaCy és a Stanza által elért eredmények az `emBERT` teljesítményéhez és az eddigi magyar NER eredményekhez hasonlítva alacsonynak tűnhetnek, de

[9] `https://paperswithcode.com/sota/named-entity-recognition-ner-on-ontonotes-v5`
[10] `https://spacy.io/models/en`

igazából hozzák a többi nyelvre is szokásosan elért átlagot. Az eddigi magas F-mértékek elsősorban annak voltak köszönhetőek, hogy kicsi és specifikus korpuszon lettek tanítva és kiértékelve a rendszerek, amelyen könnyebb magas pontszámokat elérni, mint egy nagyobb, de heterogén szövegen. Az NYTK-NerKoron elért alacsonyabb számok így tehát nem azt jelentik, hogy alacsonyabb lenne a korpusz minősége vagy hogy rosszabb lenne az adott rendszerek teljesítménye, hanem, hogy jobb általánosító képességgel rendelkeznek azelőtt nem látott szövegeken is.

4. Összegzés

Cikkünkben az NYTK-NerKor korpusz alapos kiértékelését adtuk. A magyar tulajdonnév-felismerésben eddig elérhető gold standard korpuszok közös tulajdonsága, hogy kicsik és domainspecifikusak, ezért a domainen kívüli szövegek elemzéséhez kevéssé voltak jól használhatóak a rajtuk tanított rendszerek. Ugyanezen korpuszok teszthalmazán végezve a kiértékelést torz képet kaphattunk a rendszerünk teljesítményéről, hiszen a rendszer valószínűleg nem rendelkezett azzal az általánosító képességgel, ami ahhoz kell, hogy egy azelőtt nem látott szövegben jól azonosítsa a neveket. Az NYTK-NerKor kellően nagy és heterogén ahhoz, hogy ezt kiküszöbölje, amit a méréseink is alátámasztanak.

Hivatkozások

Akbik, A., Blythe, D., Vollgraf, R.: Contextual string embeddings for sequence labeling. In: Proceedings of the 27th International Conference on Computational Linguistics. pp. 1638–1649. Association for Computational Linguistics, Santa Fe, New Mexico, USA (Aug 2018)

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

Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: 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. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (6 2019)

Ginter, F., Hajič, J., Luotolahti, J., Straka, M., Zeman, D.: CoNLL 2017 shared task - automatically annotated raw texts and word embeddings (2017), `http://hdl.handle.net/11234/1-1989`, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University

Halácsy, P., Kornai, A., Németh, L., Rung, A., Szakadát, I., Trón, V.: Creating open language resources for Hungarian. In: Proceedings of the 4th international conference on Language Resources and Evaluation (LREC2004) (2004)

Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation 9, 1735–80 (12 1997)

Indig, B., Sass, B., Simon, E., Mittelholcz, I., Kundráth, P., Vadász, N.: `emtsv` — egy formátum mind felett. In: Berend, G., Gosztolya, G., Vincze, V. (szerk.) XV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2019). pp. 235–247. Szegedi Tudományegyetem Informatikai Tanszékcsoport, Szeged (2019)

Lafferty, J.D., McCallum, A., Pereira, F.C.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the Eighteenth International Conference on Machine Learning. pp. 282–289 (2001)

Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space (2013)

Nemeskey, D.M.: Egy `emBERT` próbáló feladat. In: XVI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2020). pp. 409–418. Szeged (2020a)

Nemeskey, D.M.: Natural Language Processing methods for Language Modeling. Ph.D.-értekezés, Eötvös Loránd University (2020b)

Nemeskey, D.M.: Introducing `huBERT`. In: XVII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2021). pp. 3–14. Szeged (2021)

Nocedal, J.: Updating quasi-Newton matrices with limited storage. Mathematics of computation 35(151), 773–782 (1980)

Novák, A., Siklósi, B., Oravecz, Cs.: A new integrated open-source morphological analyzer for Hungarian. In: Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S. (szerk.) Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). European Language Resources Association (ELRA), Paris, France (May 2016)

Okazaki, N.: CRFsuite: a fast implementation of Conditional Random Fields (CRFs) (2007), `http://www.chokkan.org/software/crfsuite/`

Orosz, Gy., Szántó, Zs., Berkecz, P., Szabó, G., Farkas, R.: HuSpaCy: an industrial-strength Hungarian natural language processing toolkit. In: XVIII. Magyar Számítógépes Nyelvészeti Konferencia (2022)

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., Manning, C.D.: Stanza: A Python natural language processing toolkit for many human languages. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations (2020)

Simon, E., Lendvai, P., Németh, G., Olaszy, G., Vicsi, K.: A magyar nyelv a digitális korban – The Hungarian Language in the Digital Age. Georg Rehm and Hans Uszkoreit (Series Editors): META-NET White Paper Series, Springer (2012)

Simon, E., Nemeskey, D.M.: Automatically generated NE tagged corpora for English and Hungarian. In: Proceedings of the 4th Named Entity Workshop (NEWS) 2012. pp. 38–46. Association for Computational Linguistics, Jeju, Korea (July 2012)

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)

Sutton, C., McCallum, A.: An Introduction to Conditional Random Fields for Relational Learning (01 2007)

Szarvas, Gy., Farkas, R., Kocsor, A.: A Multilingual Named Entity Recognition System Using Boosting and C4.5 Decision Tree Learning Algorithms. In: Proceedings of Discovery Science 2006. pp. 267–278. Springer Verlag (2006a)

Szarvas, Gy., Farkas, R., Felföldi, L., Kocsor, A., Csirik, J.: A highly accurate Named Entity corpus for Hungarian. In: Electronic Proceedings of the 5th International Conference on Language Resources and Evaluation (May 2006b)

Tjong Kim Sang, E.F.: Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition. In: Roth, D., van den Bosch, A. (szerk.) Proceedings of CoNLL-2002. pp. 155–158. Taipei, Taiwan (2002)

Tjong Kim Sang, E.F., De Meulder, F.: Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In: Daelemans, W., Osborne, M. (szerk.) Proceedings of CoNLL-2003. Edmonton, Canada (2003)

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

Váradi, T.: The Hungarian National Corpus. In: Proceedings of the Third International Conference on Language Resources and Evaluation (LREC-2002). pp. 385–389. European Language Resources Association, Las Palmas de Gran Canaria (2002)