Hát te mekkorát nőttél! – A HuLU első életéve új adatbázisokkal és webszolgáltatással

Ligeti-Nagy Noémi[1], Héja Enikő[1], Laki László János[1], Takács Dávid[2], Yang Zijian Győző[1], Váradi Tamás[1]

[1]Nyelvtudományi kutatóközpont
1068 Budapest, Benczúr u. 33.
{vezeteknev.keresztnev}@nytud.hu
[2]takdavid@nytud.hu

Kivonat Az egyre nagyobb számban megjelenő neurális nyelvmodellek jelentőségének felbecsüléséhez elengedhetetlen valamilyen eszköz, amellyel ezek mérhetők és összehasonlíthatók. A nyelvtechnológiában ez a mérőeszköz a *benchmark* adatbázis, amelyből számtalan elérhető már az angol nyelvre, és több más nyelvre is. A magyarra 2022-ben jelent meg a HuLU *benchmark* adatbázis-készlet (Ligeti-Nagy és mtsai, 2022), melynek célja, hogy egyre bővülő gyűjteményként a nyelvmodellek magyarnyelvtudásának mérésére szolgáljon. Jelen cikkünkben bemutatunk két olyan új adatbázist, amelyekkel a HuLU-t bővítettük: egyrészt a Recognizing Textual Entailment (RTE) *shared taskok* anyagából válogatott RTE korpusz magyar fordítását, másrészt a CommitmentBank mintájára létrehozott magyar HuCommitmentBank korpuszt. Ezen kívül bemutatjuk a HuLU használatát megkönnyítő és kiteljesítő webszolgáltatást, amely lehetővé teszi a modellek eredményeinek gyors kiértékelését, és a feltöltött eredmények látványos összehasonlítását. Ezt kihasználva tanulmányunkban bemutatjuk néhány magyarul is tudó nyelvmodellnek a HuLU korpuszain elért eredményeit.
Kulcsszavak: *benchmarking*, NLI, nyelvmodellek, kiértékelés

1. Bevezetés

Az elmúlt évtized a neurális nyelvmodellek térnyeréséről szólt a nyelvtechnológiában: évről évre egyre több és egyre nagyobb architektúra kerül bemutatásra, melyeknek a teljesítményét a szintén egyre szélesebb körben elérhető, és egyre változatosabb *benchmark* adatbázisokon hasonlítják össze. Az angol GLUE és SuperGLUE *benchmarkokat* (Wang és mtsai, 2018, 2020) hamarosan követte a francia (FLUE, Le és mtsai, 2020), a spanyol (GLUES, Cañete és mtsai, 2020), vagy az orosz (Russian SuperGLUE, Shavrina és mtsai, 2020) megfelelőjük, illetve az XGLUE, amely többnyelvű modellek kiértékelésére fókuszál (Liang és mtsai, 2020). Mindeközben természetesen fokozatosan fény derült ezeknek a statikus *benchmarkoknak* a hiányosságaira is: hamar devalválódnak, a modellek hamar túltanulnak rajtuk, gyakran hibás annotálást tartalmaznak, és az értékelési metrikák sok esetben tökéletlenek, nem egyértelműek. A Dynabench (Kiela

és mtsai, 2021) célja, hogy dinamikusan, a modellek eredményeit és az emberi munkát fokozatosan beépítve hozzon létre *benchmark* adatbázisokat, amelyekkel a statikus elődök hibái kiküszöbölhetők.

Kis késéssel ugyan, de megkezdődött a jelentősebb neurális architektúrák magyar korpuszokon történő előtanítása is (Nemeskey, 2021; Feldmann és mtsai, 2021; Yang, 2022). A jövőben várhatóan még több, magyarra tanított modell jelenik majd meg, melyeknek a nyelvértését ugyanúgy szükséges mérni, összehasonlítani. Ezért időszerűvé vált egy magyar korpuszgyűjtemény, amely megfelel a *benchmarking* céljaira. Ebből a megfontolásból jött létre a HuLU (Ligeti-Nagy és mtsai, 2022), amelynek a kiindulópontját a *multi-task* jellegű GLUE, és a SuperGLUE jelentette. Mivel magyarra nem álltak rendelkezésre olyan adatbázisok, amelyeket be lehet emelni egy *benchmark-gyűjteménybe*, a HuLU összeállításához több új, specifikus, egy adott feladatra koncentráló korpuszt készítettünk.

Jelen cikkben a HuLU részeként prezentálunk két új adatbázist, illetve bemutatjuk a HuLUhoz készített webszolgáltatást, amellyel könnyen kiértékelhető és összehasonlítható a különböző, magyarul (is) tudó nyelvmodellek teljesítménye.

2. Előzmények

A HuLU 2022-ben bemutatott állapotában négy adatbázisból állt (Ligeti-Nagy és mtsai, 2022). A HuCOLA, az elfogadhatósági ítéletek korpusza 9 944 példát tartalmaz, melyeket manuálisan gyűjtöttünk magyar nyelvészeti szakirodalmi munkákból, majd 4-4 annotátor annotálása után a többségi döntés alapján címkéztek.[1].

A HuCoPA ok-okozati összefüggések feltárását célozza. Az angolból fordított korpuszban 1000 példa található, amelyekben egy premisszához két alternatíva közül kell kiválasztani azt, amelyik valószínűbben ok-okozati viszonyban áll a premisszával. A kérdések egy részében az okot, egy másik részében az okozatot kell az alternatívák közül kiválasztani. A CoPA 1000 kérdésének gépi fordítóval történő lefordítása után annotátorok ellenőrizték és javították a fordítást, a fluenciára törekedve. Egy-egy annotátor megjelölte a helyes választ a kérdésre. Ha az annotátor döntése eltért az eredeti címkétől, akkor kézzel lett ellenőrizve és javítva az adott példa.[2]

A HuSST szentimentkorpusz az ún. SST-5 adathalmaz fordításával és újraannotálásával készült. 11 683 mondatot tartalmaz, melyek negatív-semleges-pozitív címkékkel vannak ellátva.[3].

A magyar Winograd-sémák a referencia-feloldás feladatát célozzák. A Winograd-séma olyan mondatpár, amely csak egy vagy két szóban különbözik egymástól, és olyan kétértelműséget tartalmaz, amely a két mondatban ellentétesen oldódik fel, és amelynek feloldásához világismeret szükséges (Levesque és mtsai,

[1] `https://github.com/nytud/HuCOLA`
[2] `https://github.com/nytud/HuCoPA`
[3] `https://github.com/nytud/HuSST`

2012). Az eredeti adathalmazt 150 mondatpár alkotta, melyekből a fordítás és a validálás után 122 magyar séma lett (Vadász és Ligeti-Nagy, 2022).[4]

3. Új adatbázisok

A HuLU céljaink szerint egy folyamatosan bővülő korpuszkészlet, amelybe újabb és újabb, *benchmarkolás* céljára megfelelő korpuszok kerülhetnek be. Első, nyilvánosságra hozott formájában négy elkészült, annotált, publikált adatbázist tartalmazott (HuCoLA, HuCOPA, HuSST és a Winograd-sémák).

Mind a GLUE, mind a SuperGLUE jelentős részét képezik a *natural language inference* (NLI) feladatot képviselő adatbázisok. Az NLI-ben a feladat annak meghatározása, hogy egy adott premissza alapján a hipotézis igaz (azaz következik a premisszából), hamis (azaz ellentmond a kettő egymásnak), vagy pedig semleges-e (a premissza alapján nem dönthető el, hogy a hipotézis igaz-e). A két legjelentősebb és legtöbbet idézett NLI adatbázis az SNLI (Bowman és mtsai, 2015) és a MultiNLI (Williams és mtsai, 2018). Az SNLI 550 000 példát tartalmaz, a MultiNLI pedig 433 000 példát, 10 különböző domainből. A MultiNLI a GLUE része. Ezen kívül a nagy *benchmark* gyűjteményekben következtetési feladatként fogalmaznak meg több, eredetileg más formájú feladatot is: így a GLUE-ban következtetési feladatként jelenik meg a SQuAD (Rajpurkar és mtsai, 2016) valamint a Winográd-sémák (Levesque és mtsai, 2012), a SuperGLUE-ban pedig a CommitmentBank (de Marneffe és mtsai, 2019) is. A HuLU első formájában igazi következtetési feladatot nem tartalmazott. Ezt egyrészt a Winograd-sémák NLI formátumra hozásával, másrészt két új adatbázis elkészítésével orvosoljuk.

3.1. HuWNLI

A HuLU jelenleg a Winograd-sémákból konvertált HuWNLI adatbázist tartalmazza koreferencia-feloldást tesztelő korpuszként, melyben az anaforafeloldás következtetési feladatként fogalmazódik meg. A HuLU részét képező NLI-formátumot úgy hoztuk létre, hogy a kétértelmű névmást a sémákban szereplő minden lehetséges referenssel helyettesítettük (a módszert a GLUE-t bemutató tanulmány írja le, ld. Wang és mtsai, 2018). A sémákból származtatott mondatpárok halmazát kibővítettük azoknak a mondatpároknak a fordításával, amelyek – a Winograd-sémák mondataival együtt – a GLUE WNLI-adatbázisát alkotják. A korpuszt három részre osztva adjuk közre (zárójelben az adott halmaz elemszáma): tanítóhalmaz (562), validáló halmaz (59) és teszthalmaz (134). A felosztások követik a GLUE WNLI felosztását, de kevesebb példát tartalmaznak, mivel több mondatpárt el kellett vetnünk, mert magyarra nem voltak lefordíthatók. A teszthalmaz mondatpárjai mind a GLUE WNLI teszthalmazának lefordított példái.

[4] `https://github.com/nytud/HuWSC`

3.2. HuRTE

A GLUE-ba beválogatták a Recognizing Textual Entailment (RTE) kihívás adathalmazainak egy részét: az RTE1-ből (Dagan és mtsai, 2006), az RTE2-ből (Bar-Haim és mtsai, 2006), az RTE3-ból (Giampiccolo és mtsai, 2007) és az RTE5-ből (Bentivogli és mtsai, 2009) gyűjtöttek hírszövegekből és Wikipedia szócikkekből származó példákat. Ezekben egy (néha többmondatos) premisszáról és egy egymondatos hipotézisről kell eldönteni, hogy ez utóbbi következik-e az elsőből vagy sem. A feladat bináris címkézés, így az eredetileg hármas osztályozású példáknál a *semleges* és *ellentmondás* címkéket összevonták a konzisztencia érdekében. Az RTE adathalmaz a SuperGLUE-nak is része maradt.

Az RTE adathalmazok GLUE-ba beválogatott részét a HuLU többi korpuszának létrehozásához is használt gépi fordítóval fordíttattuk magyarra.[5] Ezt követően az így előállt 5 797 példát (amely nagyjából 18 000 mondatot jelent) annotátorok ellenőrizték és javították, a fluenciára törekedve.[6] Mivel az eredeti benchmarkban a teszthalmazhoz nem voltak megadva a helyes címkék, ezek fordításait címkéztetnünk kellett. Így az 5 797 példából 3000-et (a teszthalmaz példáit) 3-3 annotátor címkézett. Az egyetértés a 3000 példán 0.61 (Fleiss's κ). A HuRTE teszthalmazába csak azokat az eseteket válogattuk be, ahol teljes volt az egyetértés az annotátorok között (2123 példa).[7] Ezzel a lépéssel a tesztanyag tisztaságát kívánjuk biztosítani, hogy olyan példákon mérjük a nyelvmodellek tudását, amelyeknél nagyobb bizonyossággal tudjuk a válaszát helyesnek vagy helytelennek ítélni. Bár a nyelvmodellek kiértékelésének módszertanáról jelen tanulmány keretein belül nincs lehetőségünk értekezni, a kérdés komplexitásáról, a következtetéses feladatok nehézségéről és a kiértékelés lehetséges újfajta módszereiről lásd pl. Baan és mtsai (2022); Plank (2022).

Az eredeti címkével együtt elérhető angol példákat pedig a fordítás és fluenciaellenőrzés után egy-egy annotátor címkézte. Ez lehetővé tette, hogy kiderüljön, ha a fordítás során elveszett a mondatpárok közti eredeti következtetési viszony. Ebből a 2 797 példából az eredeti tanítóhalmaz 357, illetve a validációs halmaz 35 példájánál ütközik egymással az eredeti, angol adatbázisból nyert címke és a fordítás után, magyar anyanyelvű annotátortól nyert címke. Ezt a 392 példát egyelőre nem szerepeltetjük a HuRTE adatai között. Így végül a HuRTE korpuszt tanító-, validációs és teszthalmazra bontva, 2131-242-2123 példával adjuk

[5] A gépi fordításhoz az OPUS (Tiedemann, 2012) korpusztárból magunk építettünk egy angol-magyar párhuzamos korpuszt. A felhasznált alkorpuszok a következők: ParaCrawl, OpenSubtitles, Tatoeba, WikiMatrix, EUbookshop, PHP manual, TED2020, KDEdoc, KDE4. A párhuzamos korpuszból a Marian NMT (Junczys-Dowmunt és mtsai, 2018) nevű keretrendszerrel építettünk egy *transformer encoder-decoder* architektúrájú neurális fordítórendszert. A betanított modell paraméterei: 6 réteg enkóder és 6 réteg dekóder; 16 figyelmi fej; 1024 szóbeágyazás dimenzió; 1024 bemeneti hossz; előre csatolt háló méret: 4096.

[6] 14 annotátor végezte a fordítás-, 7 a fluenciaellenőrzést. 13 annotátor címkézte a teszthalmaz példáit, 4 pedig az eredeti címkével is rendelkező példákat.

[7] Bentivogli és mtsai (2009) szintén ezt a módszert alkalmazták az RTE 5 adatbázis összeállításánál: csak azokat a példákat tartották meg, amelyeknél háromból három annotátor egyetértett a címkén.

közre.[8] [9] Az 1. példa a HuRTE tanítóanyagának egyik példáját mutatja. Ezen is látszik, hogy az RTE (és elsősorban az RTE5) anyaga nem egy-egy mondatos premisszákkal és hipotézisekkel operál, hanem sokszor kettő, vagy akár három mondatos premisszákat gyűjtöttek az adatbázis összeállítói.

(1) *premise*: A veszettség vírusa megfertőzi a központi idegrendszert, agyvelőbántalmat és végül halált okoz. A veszettség korai tünetei az embernél nem jellegzetesek: láz, fejfájás és általános rossz közérzet.
hyphothesis: A veszettség az emberre nézve halálos kimenetelű.
label: *entailment*

A HuRTE kiértékeléséhez választott metrikánk, tekintettel a kiegyensúlyozatlan teszthalmazra, az MCC (Matthews, 1975).

3.3. HuCommitmentBank

A CommitmentBank (CB, de Marneffe és mtsai, 2019) olyan rövid szövegrészletekből áll, amelyekben legalább az egyik mondat tartalmaz egy alárendelő mellékmondatot. Mindegyik mellékmondat meg van címkézve azzal, hogy a szöveg írója milyen mértékben elkötelezett a mellékmondat igazsága mellett (2. példa). A SuperGLUE-ban a feladatot egy hármas osztályozású következtetési feladattá alakították: a premissza a teljes szövegrészlet, a hipotézis pedig a beágyazott tagmondat. Csak azt a részét használták a korpusznak, ahol 80% fölötti volt a mért IAA. Lényeges tulajdonsága a korpuszban szereplő példáknak, hogy minden vizsgált beágyazott mondat egy logikai következtetést semlegesítő operátor alá tartozik szintaktikailag. Logikai következtetést semlegesítő operátornak számít a kérdés, a tagadás, a feltételes mód és a modálissal való bővítés.

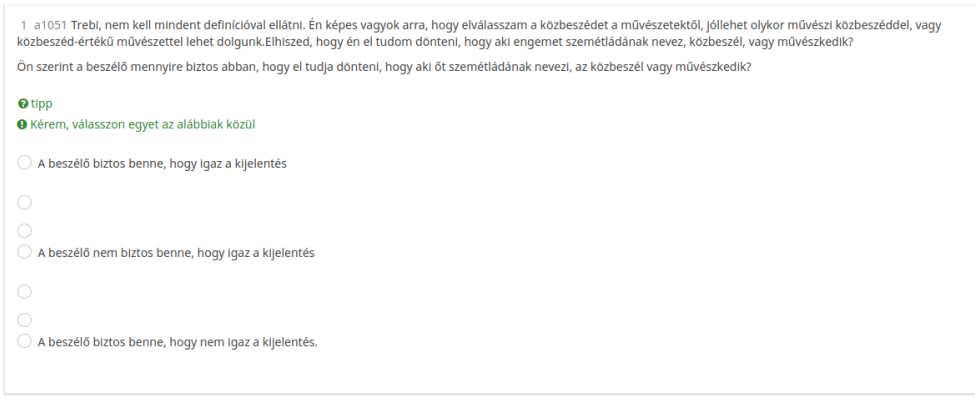
(2) A kontextus: *Polly had to think quickly.* 'Pollynak gyorsan kellett döntenie.'
A vizsgált mondat: *They were still close enough to shore for him to return her to the police if she admitted she was not an experienced ocean sailor.* 'Még elég közel voltak a parthoz ahhoz, hogy a férfi mégis átadja a rendőrségnek, ha bevallja, hogy nem egy tapasztalt tengerész.'
Az alárendelt tagmondat: Polly was not an experienced ocean sailor. 'Polly nem egy tapasztalt tengerész.'
A címke: 2 (−3-tól 3-ig terjedő skálán, ahol a 3 azt jelenti, hogy a beszélő egészen biztos az alárendelt tagmondat igazságában.)

A magyar nyelvű korpusz előállításához az angol példák 30 gyakori mátrixigéjének / -kifejezésének magyar megfelelőivel kerestünk mondatokat a korpuszban. A szövegek forrásai egyrészt az MNSZ2 (Oravecz és mtsai, 2014) beszéltnyelvi alkorpusza, másrészt a szépirodalmi alkorpusz néhány szövege, és online

[8] A teszthalmaznak a tanítóanyaghoz viszonyítva meglepően nagy aránya a GLUE és a SuperGLUE több adatbázisánál megfigyelhető. Mi egyelőre ezeknek a mintáját követve választjuk meg a tanító-, validációs és teszthalmaz méretét.

[9] A HuRTE korpusz elérhető a githubon: `https://github.com/nytud/HuRTE`.

fórumok hozzászólásainak szövegei. Az így kinyert példákon először ellenőrizni kellett, hogy található-e bennük következtetést semlegesítő operátor. 4 annotátor összesen 1100 valid szövegrészletet gyűjtött össze. A gyűjtött szövegeket az annotátorok egymás között is validálták. Az 1100 példát végül 5-5 annotátor címkézte egy 7 fokú likert-skálán (−3 és 3 között, ahol a 0 jelentette azt, hogy a beszélő nem tudja, hogy az alárendelt tagmondat igaz-e vagy hamis). Az annotálás a LimeSurvey felület segítségével történt. Összesen 9, magyar anyanyelvű annotátor dolgozott a korpuszon. Óránként 50 példa elolvasása, megértése és megítélése volt a feladatuk. Munkájuk megkönnyítése érdekében mind az 1100 példát elláttuk egy kérdéssel is, amelynek megválaszolása segíti a döntésüket. A magyar nyelvű CB előállításának részletesebb bemutatását ld. Hatvani (2022). Egy példa az annotálási felületről az 1. ábrán látható.



1. ábra: Példa a magyar nyelvű CommitmentBank előállításához használt annotálási felületről. Forrás: Hatvani (2022)

A korpusz példái a következőképpen épülnek fel: a *kontextus* a célmondat előtti 1-2 mondatot jelenti; a *célmondat* az a mondat, amely tartalmazza a következtetést törlő operátort és az alárendelt tagmondatot; minden példához külön feltüntetjük az *igét*, illetve az *alárendelt tagmondatot*; a *marker* az adott példában található következtetést semlegesítő operátor típusa. A korpuszban összesen 10 féle marker figyelhető meg: modális, feltételes, kérdés, tagadás, modális tagadás, modális kérdés, modális feltételes, feltételes kérdés, retorikai kérdés, tagadott kérdés. A 3. alatt a magyar korpusz egy példáját láthatjuk.

(3) Kontextus: *Vivien:- Tudja mit csinálnék? Bekopognék, megkérdezem elnézést van a háznak biztosítása? Igen van. Akkor itt estem fenékre. De azt hogyan bizonyítom? Mert ugye nagyon elmésen ma már egy kátyút fotóval kell bizonyítani ami szerintem egy mulatságos dolog. A fenékre esést hogyan*

lehet?
Célmondat: *Hogyan bizonyítom hogy itt történt a baleset?*
Mátrixige: *bizonyít*
Marker: kérdés
Alárendelt tagmondat: *Itt történt a baleset.*
Címke: −2

A SuperGLUE-ba a CB-nek egy olyan részhalmazát válogatták csak be, ahol az annotátorok közti egyetértés legalább 80% volt.[10] Így az eredeti 1200 példából egy 250 példát tartalmazó tanító-, egy 57 példás validáló-, és egy 250 példás teszthalmazt választottak. A feladatot hármas osztályozású következtetési feladatként fogalmazták meg, ld. a 2. példából levezetett 4. példa.

(4) premissza: *Polly had to think quickly. They were still close enough to shore for him to return her to the police if she admitted she was not an experienced ocean sailor.* 'Pollynak gyorsan kellett döntenie. Még elég közel voltak a parthoz ahhoz, hogy a férfi mégis átadja a rendőrségnek, ha bevallja, hogy nem egy tapasztalt tengerész.'
hipotézis: *Polly was not an experienced ocean sailor.* 'Polly nem egy tapasztalt tengerész.'
címke: entailment

A SuperGLUE-hoz hasonlóan a HuLU-ban is következtetési feladatként szerepel ez a korpusz, de a 7 fokú címkét hármas osztályozásra cseréltük: az eredeti címkék −1, 0 és 1 értékei a *semleges*, a −3 és −2 az *ellentmondás*, a 2 és 3 pedig a *következik* kategóriákba lettek tömörítve (5).

(5) premissza: *Vivien:- Tudja mit csinálnék? Bekopognék, megkérdezem elnézést van a háznak biztosítása? Igen van. Akkor itt estem fenékre. De azt hogyan bizonyítom? Mert ugye nagyon elmésen ma már egy kátyút fotóval kell bizonyítani ami szerintem egy mulatságos dolog. A fenékre esést hogyan lehet? Hogyan bizonyítom hogy itt történt a baleset?*[11]
hipotézis: *A beszélő szerint itt történt a baleset.* [12]
címke: *ellentmondás*

A korpuszon az IAÁ-t többféleképpen mértük (a Likert-skálára intervallum skálaként tekintve, ennek vitatottságáról ld. pl. Wu és Leung, 2017). A Krippendorf's α 60,5%. A szórás (SD) 1,01. Csak azokat a példákat vettük fel a HuLU-ba, amelyeknél $SD < 1$. Így végül 604 példából egy 250-250 példás tanító, illetve

[10] Az nem derül ki, hogy hogyan mérték az egyetértést a korpuszon.

[11] A szöveget módosítások nélkül emeltük át az MNSZ2-ből, a helyesírási hibák, elütések javítása nélkül.

[12] Mivel eredetileg is a beszélőnek a tagmondat igazságértékével kapcsolatos meggyőződését (*commitmentjét*) vizsgálta a korpusz, nem értünk egyet azzal, hogy a SuperGLUE-ban a megfogalmazott hipotézis a tagmondat tartalma maga. Az eredeti annotáció hipotézisbe forgatva helyesebb úgy, ha a beszélő véleményeként fogalmazzuk meg.

teszthalmazt, és egy 103 példát tartalmazó validációs halmazt hoztunk létre.[13] A teszthalmaz példáinál $SD < 0,5$.

A kiértékelés metrikája, a teszthalmazban lévő címkék arányának kiegyensúlyozatlanságát is figyelembe véve, a *súlyozott F1-mérték*.

4. A webszolgáltatás bemutatása

Az új adatbázisokon túl a HuLU-t egy webszolgáltatással is kiegészítettük, amely a hulu.nytud.hu címen elérhető, regisztráció után minden funkciója bárki számára szabadon használható. A webszolgáltatás lehetővé teszi a nyelvmodellek egységes módon történő gyors és kényelmes kiértékelését, valamint az eredmények közzétételét. Így az egyes modellek teljesítményét mindenki egy kattintással megtekintheti.

A magyarul és angolul is elérhető webszolgáltatás megjelenésében és funkcióiban is követi a GLUE és SuperGLUE felületeket. Felépítésével a teljes modellfejlesztési folyamat során végig támogatja a felhasználót, a bevezető magyarázattal kezdve a publikációk elérésén át a *benchmark* feladatok megértéséig és a gyarkorlati teendők megismeréséig.

Az oldalról közvetlenül letölthetők a *benchmark* korpuszok (jelenleg: HuCOLA, HuCommitmentBank, HuCoPA, HuRTE, HuSST, HuWNLI), amelyek a szokásos tanító-, validációs- és teszthalmazokra vannak bontva. A tesztanyag címkéit nem adjuk közre, azok csak a HuLU szerverén futó kiértékelő modul számára hozzáférhetők.

A specifikált formátumban megcímkézett teszthalmaz feltöltése után a webszolgáltatás az eredményeket az alábbi metrikák szerint értékeli ki:

- HuCOLA, HuCoPA, HuRTE: Matthew's Correlation Coefficient (MCC). Bár a GLUE/SuperGLUE az utóbbi kettő esetében abszolút pontosságot (*accuracy*) számol, mi a címkék tesztanyagbeli eloszlásának egyenlőtlensége miatt az MCC mellett döntöttünk ezeknél is.
- HuSST, HuWNLI: abszolút pontosság (*accuracy*) a GLUE/SuperGLUE mintájára.
- HuCommitmentBank: súlyozott F1-mérték, amely kiválóan alkalmas az eltérő méretű osztályok kezelésére többosztályos klasszifikáció esetében.

A kiértékelés eredményét visszajelezzük a feltöltő felhasználónak, de azok nem válnak automatikusan nyilvánossá a weboldalon, ahhoz az NYTK munkatársainak jóváhagyása szükséges.

A már nyilvánossá tett eredmények többféleképpen jelennek meg: a modellek teljesítménye időrend szerint grafikonon és teljesítmény szerint rendezett táblázatos formában is megtekinthető.

A HuLU webszolgáltatás lehetővé teszi a modellek egy-egy feladaton történő kiértékelését tetszőleges időpontban és sorrendben, míg a GLUE és SuperGLUE oldalakon a modelleket csak az összes feladaton egyszerre értékelhetjük ki.

[13] A HuCommitmentBank korpusz elérhető a githubon: `https://github.com/nytud/HuCommitmentBank`.

4.1. A webszolgáltatás implementációja

A hulu.nytud.hu weboldal a NYTK szerverein fut Docker konténerekben, Portainer platformon. A kiértékelő modul és a webes backend is Pythonban lett implementálva, Django keretrendszer felett. Az eredmények tetszetős megjelenítéséhez a Streamlit[14] rendszert használtuk, amely önmagában is alkalmas dinamikus adatok publikációjára, de, mint esetünkben, weboldalba is beágyazható.

A kiértékelésre feltölendő tesztanyagok elvárt formátuma megegyezik a letölthető tanító- és tesztanyag formátumával. Mindegyiket *json* formátumban kell feltölteni, szerepelnie kell benne az aktuális entitás azonosítójának, valamint a tesztanyaggal megegyező nevű eredménymezőnek. A kiértékelő program minden feltöltésnél ellenőrzi az elvárt formátumot, valamint ellenőrzi, hogy minden elemre érkezett-e predikció. A teljes rendszerhez hasonlóan a kiértékelő rendszert is modulárisan építettük fel, hogy könnyen tudjunk integrálni új metrikákat, valamint felvenni új *benchmark* teszteket a rendszerbe.

5. Eredmények

A HuLU 5 korpuszán kiértékeltük néhány magyarul (is) tudó nyelvmodell teljesítményét (1. táblázat). Az eredmények alapján a legnehezebb feladatnak a következtetési feladat, a HuRTE bizonyul, amelynél a legjobb eredményt a PULI BERT-Large (Yang és mtsai, 2023) érte el (MCC: 51,7; viszonyításképpen a BERT Devlin és mtsai, 2019 eredménye a GLUE-ban az RTE adathalmazon 70,1% abszolút pontosság). Jobban teljesítettek a modellek a HuCOLA és a HuSST feladatokon (mindkettőn a PULI BERT-Large a legeredményesebb, a HuCOLA MCC: 71,1, illetve abszolút pontosság a HuSST-n: 79,9%). Az eredményekből látható, hogy a huBERT (Nemeskey, 2021) alig marad el a PULI BERT-Large-tól, sőt, a HuCoPA feladaton, amely szintén nagyon nehéznek bizonyul a modellek számára, jobban is teljesít annál (MCC: 56,1; a BERT eredménye a CoPÁ-n 70,6% abszolút pontosság). A GPT modellt nem sikerült a HuCoPA adatbázisokra finomhangolni, illetve a megfelelő promptot megtalálni a tanításához.

	HuCOLA (MCC)	HuCoPA (MCC)	HuRTE (MCC)	HuSST (Pontosság)	HuWNLI (Pontosság)
huBERT	70,9	**56,1**	48,7	79,4%	64,9%
XLM-R base[15]	55,9	3,2	33,3	66,1%	63,4%
PULI GPT-2[16]	49,9		41,5	72,8%	61,9%
PULI BERT-Large	**71,1**	41,4	**51,7**	**79,9%**	**65,7%**

1. táblázat. Néhány magyar nyelvmodell eredménye a HuLU egyes korpuszain. Az oszlopok az egyes korpuszokat és az azoknál alkalmazott kiértékelési metrikát, a sorok a modelleket jelölik.

[14] https://streamlit.io

Az 1. táblázatban látható számok a legelső eredmények a HuLU adatbázisain. Valójában új kutatási irányt jelent, hogy az egyes feladatokon hogyan tudunk az angol SOTA eredményekhez hasonlóakat elérni a magyar modellekkel, hiszen minden modell és feladat más-más hiperparaméter-beállítást igényel. Például egy nagyobb modell kisebb tanulási rátát igényel, de a *batch* méret is hatással van a teljesítményre. A HuSST és a HuCOLA megoldható egyszerű mondatszintű osztályozás módszerével. A HuCoPA egy többválasztásos (*multiple choice*) feladat. A HuRTE és a HuWNLI korpuszokat egy entitásorientált szentimentelemzéshez hasonló megoldással érdemes megközelíteni, ahol a két vizsgált mondat viszonyát szeretnénk megtudni.

A különböző feladatok esetében a hiperparaméterek mellett különböző bemeneti adatstruktúrákkal, úgynevezett promptokkal (Shin és mtsai, 2020) is érdemes kísérletezni, amelyek hatással lehetnek a teljesítményre.[17]

6. Összegzés

A nyelvtechnológia különböző feladatain a korábbi szabályalapú, vagy hagyományos statisztikai alapú modelleket rendre túlteljesítő neurális nyelvmodellek teljesítményének megfelelő mérésére és összehasonlítására szolgálnak az ún. *benchmark* adatbázisok és adatbázis-gyűjtemények. A HuLU-t az előtanítás során magyar szöveget is látó, és így magyarul (is) tudó nyelvmodellek kiértékelésére indítottuk útjára 2022-ben. Jelen cikkünkben bemutattunk két új adatbázist, amelyek kifejezetten nehéz, következtetési feladatot képviselő adatbázisként a HuLU részei lehetnek, tovább színesítve a HuLU-ban elérhető korpuszok széles palettáját. Bemutattuk a HuLU-hoz készített webszolgáltatást, amellyel a modellek eredményeinek kiértékelése könnyen elérhetővé válik akár csak egy-egy adatbázison is. Az eredmények összehasonlíthatóságát pedig egy dicsőségtábla segíti. A HuLU eddigi korpuszain végzett kiértékeléseink azt mutatják, hogy

[15] Conneau és mtsai (2019)

[16] Yang (2022)

[17] Azoknál a feladatoknál, ahol két mondat viszonyát szeretnénk megtanítani a modelleknek (például HuCoPA, HuRTE vagy HuWNLI), kísérleti irány lehet a vizsgálandó szövegek szeparátor címkével (`SEP`) való elválasztása, vagy egy egységes folyószöveggé való átalakítása. Egy-egy példa a két típusra a HuCoPA egy példájával:

- Eredeti:
  - *premise*: A bár bezárt.
  - *choice1*: Zsúfolt volt.
  - *choice2*: Hajnali három volt.
  - *question*: *cause*
- Szeparátor címkével: A bár bezárt. [SEP] Zsúfolt volt.
- Szöveggé alakítva: A bár bezárt. Mert zsúfolt volt.

A fenti példában a szeparátor címke lehet akár `[CLS]` vagy a GPT modelleknél `</s>` stb. A folyószöveggé való átalakítás során a fenti példában két mondatot látunk, de lehet egy mondatba is foglalni őket vesszővel elválasztva stb.

a magyar nyelvmodelleken még bőven van mit csiszolni, hogy az angolul tudó társaik angolnyelv-tudásához hasonló szintre érjenek a magyarnyelv-tudásukkal.

Hivatkozások

Baan, J., Aziz, W., Plank, B., Fernández, R.: Stop Measuring Calibration When Humans Disagree. CoRR abs/2210.16133 (2022)

Bar-Haim, R., Dagan, I., Dolan, B., Ferro, L., Giampiccolo, D., Magnini, B., Szpektor, I.: The Second PASCAL Recognising Textual Entailment Challenge. In: Proceedings of the Second PASCAL Challenges Workshop on Recognizing Textual Entailment, Venice, Italy. pp. 1–9 (2006)

Bentivogli, L., Dagan, I., Dang, H.T., Giampiccolo, D., Magnini, B.: The Fifth PASCAL Recognizing Textual Entailment Challenge. In: Proceedings of the TAC Workshop (2009)

Bowman, S.R., Angeli, G., Potts, C., Manning, C.D.: A large annotated corpus for learning natural language inference. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics (2015)

Cañete, J., Chaperon, G., Fuentes, R., Ho, J.H., Kang, H., Pérez, J.: Spanish Pre-Trained BERT Model and Evaluation Data. In: PML4DC at ICLR 2020 (2020)

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale (2019), `https://arxiv.org/abs/1911.02116`

Dagan, I., Glickman, O., Magnini, B.: The PASCAL Recognising Textual Entailment Challenge. In: Quiñonero-Candela, J., Dagan, I., Magnini, B., d'Alché Buc, F. (szerk.) Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment. pp. 177–190. Springer Berlin Heidelberg, Berlin, Heidelberg (2006)

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 (Jun 2019), `https://aclanthology.org/N19-1423`

Feldmann, Á., Hajdu, R., Indig, B., Sass, B., Makrai, M., Mittelholcz, I., Halász, D., Yang, Z.Gy., Váradi, T.: HILBERT, magyar nyelvű BERT-large modell tanítása felhő környezetben. In: XVII. Magyar Számítógépes Nyelvészeti Konferencia. pp. 29–36. Szegedi Tudományegyetem, Informatikai Intézet, Szeged, Magyarország (2021)

Giampiccolo, D., Magnini, B., Dagan, I., Dolan, B.: The Third PASCAL Recognizing Textual Entailment Challenge. In: Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing. p. 1–9. RTE '07 (2007)

Hatvani, P.: A Corpus to Investigate Projection Methods: The Hungarian Commitment Bank. Szakdolgozat, Pázmány Péter Katolikus Egyetem, Bölcsészet- és Társadalomtudományi Kar, Angol-Amerikai Intézet, Elméleti Nyelvészet Tanszék (2022)

Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, T., Seide, F., Germann, U., Fikri Aji, A., Bogoychev, N., Martins, A.F.T., Birch, A.: Marian: Fast Neural Machine Translation in C++. In: Proceedings of ACL 2018, System Demonstrations. pp. 116–121. Association for Computational Linguistics, Melbourne, Australia (July 2018)

Kiela, D., Bartolo, M., Nie, Y., Kaushik, D., Geiger, A., Wu, Z., Vidgen, B., Prasad, G., Singh, A., Ringshia, P., Ma, Z., Thrush, T., Riedel, S., Waseem, Z., Stenetorp, P., Jia, R., Bansal, M., Potts, C., Williams, A.: Dynabench: Rethinking benchmarking in NLP. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 4110–4124. Association for Computational Linguistics, Online (Jun 2021), https://aclanthology.org/2021.naacl-main.324

Le, H., Vial, L., Frej, J., Segonne, V., Coavoux, M., Lecouteux, B., Allauzen, A., Crabbé, B., Besacier, L., Schwab, D.: FlauBERT: Unsupervised language model pre-training for French. In: Proceedings of the 12th Language Resources and Evaluation Conference. pp. 2479–2490. European Language Resources Association, Marseille, France (May 2020), https://aclanthology.org/2020.lrec-1.302

Levesque, H.J., Davis, E., Morgenstern, L.: The Winograd Schema Challenge. In: Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning. p. 552–561. KR'12, AAAI Press (2012)

Liang, Y., Duan, N., Gong, Y., Wu, N., Guo, F., Qi, W., Gong, M., Shou, L., Jiang, D., Cao, G., Fan, X., Zhang, R., Agrawal, R., Cui, E., Wei, S., Bharti, T., Qiao, Y., Chen, J.H., Wu, W., Liu, S., Yang, F., Campos, D., Majumder, R., Zhou, M.: XGLUE: A New Benchmark Dataset for Cross-lingual Pretraining, Understanding and Generation. arXiv abs/2004.01401 (2020)

Ligeti-Nagy, N., Ferenczi, G., Héja, E., Jelencsik-Mátyus, K., Laki, L.J., Vadász, N., Yang, Z.Gy., Váradi, T.: HuLU: magyar nyelvű benchmark adatbázis kiépítése a neurális nyelvmodellek kiértékelése céljából. In: XVIII. Magyar Számítógépes Nyelvészeti Konferencia. pp. 431–446. Szegedi Tudományegyetem, Informatikai Intézet, Szeged (2022)

de Marneffe, M.C., Simons, M., Tonhauser, J.: The CommitmentBank: Investigating projection in naturally occurring discourse. Proceedings of Sinn und Bedeutung 23(2), 107–124 (Jul 2019), https://ojs.ub.uni-konstanz.de/sub/index.php/sub/article/view/601

Matthews, B.: Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta (BBA) - Protein Structure 405(2), 442–451 (1975), https://www.sciencedirect.com/science/article/pii/0005279575901099

Nemeskey, D.M.: Introducing huBERT. In: XVII. Magyar Számítógépes Nyelvészeti Konferencia. pp. 3–14. Szegedi Tudományegyetem, Informatikai Intézet, Szeged, Magyarország (2021)

Oravecz, Cs., Váradi, T., Sass, B.: The Hungarian Gigaword Corpus. In: Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (szerk.) Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). European Language Resources Association (ELRA), Reykjavik, Iceland (2014)

Plank, B.: The „Problem" of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation. Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP (2022)

Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: SQuAD: 100,000+ Questions for Machine Comprehension of Text. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. pp. 2383–2392. Association for Computational Linguistics, Austin, Texas (Nov 2016), https://aclanthology.org/D16-1264

Shavrina, T., Fenogenova, A., Emelyanov, A., Shevelev, D., Artemova, E., Malykh, V., Mikhailov, V., Tikhonova, M., Chertok, A., Evlampiev, A.: RussianSuperGLUE: A Russian Language Understanding Evaluation Benchmark. arXiv preprint arXiv:2010.15925 (2020)

Shin, T., Razeghi, Y., Logan IV, R.L., Wallace, E., Singh, S.: AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 4222–4235. Association for Computational Linguistics, Online (Nov 2020), https://aclanthology.org/2020.emnlp-main.346

Tiedemann, J.: Parallel Data, Tools and Interfaces in OPUS. In: Calzolari, N., Choukri, K., Declerck, T., Dogan, M.U., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S. (szerk.) Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12). European Language Resources Association (ELRA), Istanbul, Turkey (2012)

Vadász, N., Ligeti-Nagy, N.: Winograd schemata and other datasets for anaphora resolution in hungarian. Acta Linguistica Academica 69(4) (2022), in press.

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.R.: SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems (2020)

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.: GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In: Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. pp. 353–355. Association for Computational Linguistics, Brussels, Belgium (Nov 2018), https://aclanthology.org/W18-5446

Williams, A., Nangia, N., Bowman, S.: A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers).

pp. 1112–1122. Association for Computational Linguistics, New Orleans, Louisiana (Jun 2018), https://aclanthology.org/N18-1101

Wu, H., Leung, S.O.: Can Likert Scales be Treated as Interval Scales?—A Simulation Study. Journal of Social Service Research 43(4), 527–532 (2017), https://doi.org/10.1080/01488376.2017.1329775

Yang, Z.Gy.: "Az invazív medvék nem tolerálják a suzukis agressziót" - Magyar GPT-2 kísérleti modell. In: XVIII. Magyar Számítógépes Nyelvészeti Konferencia. pp. 463–476. Szegedi Tudományegyetem, Informatikai Intézet, Szeged, Magyarország (2022)

Yang, Z.Gy., Dodé, R., Ferenczi, G., Héja, E., Kőrös, Á., Laki, L.J., Ligeti-Nagy, N., Jelencsik-Mátyus, K., Vadász, N., Váradi, T.: Jönnek a nagyok! BERT-Large, GPT-2 és GPT-3 nyelvmodellek magyar nyelvre. In: XIX. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged, Hungary (2023)