HuLU: magyar nyelvű *benchmark* adatbázis kiépítése a neurális nyelvmodellek kiértékelése céljából

Ligeti-Nagy Noémi[1,2], Ferenczi Gergő[1], Héja Enikő[1], Jelencsik-Mátyus Kinga[1], Laki László János[1,2], Vadász Noémi[1], Yang Zijian Győző[1,2], Váradi Tamás[1]

[1]Nyelvtudományi Kutatóközpont
[2]MTA-PPKE Magyar Nyelvtechnológiai Kutatócsoport
vezeteknev.keresztnev@nytud.hu

Kivonat A cikkben bemutatjuk a neurális nyelvmodellek nyelvértésének mérésére, kiértékelésére és a modellek összehasonlítására létrehozott adatbázisunkat. Az angol példát (GLUE – Wang és mtsai, 2018 –, illetve SuperGLUE – Wang és mtsai, 2020 – *benchmarkok*) követve, de a redundanciát mellőzve kiválasztottunk 13 olyan specifikus feladatot és az ehhez kapcsolódó alkorpuszokat, amelyekkel a neurális modellek teljesítménye mérhető. Mivel az angol *benchmarkokban* szereplő korpuszok közül egyik sem létezik a magyarra, ezeket megtervezzük, megépítjük, majd egységes tesztelési keretbe foglalva közrebocsátjuk ezeket.
Kulcsszavak: *benchmarking*, neurális modellek, korpuszépítés, kiértékelés

1. Bevezetés

Az utóbbi években gombamód szaporodnak a neurális nyelvmodellek: évről évre egyre több, nagyobb, „okosabb” architektúrát mutatnak be. Ezeknek az összehasonlítására, kiértékelésére hozták létre a *benchmark* adatbázisokat, amelyek sokszor tulajdonképpen korpuszgyűjtemények, és változatos feladatokon mérik a modellek teljesítményét. 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 (Shavrina és mtsai, 2020) megfelelőjük, illetve az XGLUE, amely többnyelvű modellek kiértékelésére fókuszál (Liang és mtsai, 2020).

Kis késéssel ugyan, de megkezdődött a jelentősebb architektúrák magyar korpuszokon történő előtanítása (Nemeskey, 2021; Feldmann és mtsai, 2021). A jövőben várhatóan még több, magyarra tanított modell jelenik majd meg, melyeknek a nyelvértését ugyanúgy szükséges lesz mérni, összehasonlítani. Ezért döntöttünk úgy, hogy létrehozzuk a Hungarian Language Understanding Evaluation Benchmark Kit (HuLU) névre keresztelt adatbázis-gyűjteményünket.

A gépi tanulásban és a nyelvtechnológia területén a *benchmark* jellemzően egy vagy több adatbázisból, azokhoz tartozó megfelelő metrikákból és a teljesítmény

összesítésének módjából áll. A *benchmark* különböző rendszerek teljesítményének méréséhez biztosít olyan sztenderdet, melyben a szakmai közösség egyetért. Ez utóbbi kritériumnak történő megfelelés kényszere eredményezte például, hogy az elmúlt pár év nagy *benchmark* adatbázisai már meglévő feladatok korpuszaiból igyekeztek reprezentatív gyűjteményt összeállítani (pl. a GLUE, vagy az XTREME, Hu és mtsai, 2020), mások pedig kifejezetten a szakmai közösség ajánlásai alapján válogatták össze az adatbázis elemeit (pl. a SuperGLUE, vagy a BIG-Bench, Ghazal és mtsai, 2017). A *benchmarkok* jelentőségét mutatja, hogy például az AI Index Report 2021 a SuperGLUE és a SQuAD (Stanford Question Answering Dataset, Rajpurkar és mtsai, 2016) alapján számol be az NLP területének általános előrehaladásáról (Zhang és mtsai, 2021). Szintén komoly szakmai érdeklődésre tart számot az, ha valamely modellnek sikerül valamelyik *benchmarkon* a humán teljesítményhez hasonló eredményt elérnie.

A magyarra *benchmark* adatbázis még nem készült. Mi kiindulópontunknak a széles körben alkalmazott, *multi-task* jellegű, mérföldkőnek és meghatározó szerepű *benchmarknak* tartott GLUE-t, és utódját, a SuperGLUE-t választottuk. Tisztában vagyunk ugyanakkor ezeknek az adatbázisoknak a gyengeségeivel és hiányosságaival is (ezekről cikkünk összefoglaló fejezetében részletesen is szólunk), ezért az alapvetőnek tartott, GLUE-beli és SuperGLUE-beli korpuszok létrehozásán túl a jövőben célunk majd egy az eddigieknél jobb, nyelvészetileg és nyelvtechnológiailag is megalapozott, átgondolt, bővebb *benchmark* adatbázis létrehozása.

A GLUE, a SuperGLUE, sőt még néhány, más nyelvre összeállított *benchmark* adatbázis esetében is már meglévő korpuszokból válogathattak a kutatók. Magyarra azonban nem elérhetőek ilyen specifikus, adott feladatra fókuszáló, megfelelően annotált korpuszok. Éppen ezért az itt bemutatott korpuszépítési munka feladata kettős: i) célunk, hogy előállítsunk több kisebb, specifikus, jól annotált, megbízható korpuszt, amelyek a nyelvmodellek számára jellemzően kihívást jelentő nyelvértési feladatokat céloznak meg, illetve ii) ezekből összeállítsunk egy *benchmark* adatbázist, amelyen a nyelvmodellek teljesítménye mérhető, összehasonlítható.

2. Az angol benchmarkok

A General Language Understanding Evaluation (GLUE) *benchmarkot* 2019-ben mutatták be. Az adatbázisba úgy válogatták a korpuszokat, hogy a neurális modellek nyelvértésének tesztelése a lehető legváltozatosabb, eltérő nehézségű és doménbe tartozó feladatokon váljon lehetővé. Törekedtek arra, hogy kifejezetten kevés tanítóadatot tartalmazó korpuszokat bocsássanak közre, ezáltal kedvezve a *transfer learningre* épülő modelleknek, mintegy ebbe az irányba orientálva a nyelvtechnológiai szakmai közösséget (Wang és mtsai, 2018).

A GLUE-ba kilenc, már meglévő adatbázist válogattak be, azok eredeti szerkezetét és az általuk képviselt feladatot néhol némileg módosítva. A kilenc alkorpusz a következő:

– A CoLA (Corpus of Linguistic Acceptability, Warstadt és mtsai, 2018) 10 657 angol mondatot tartalmaz, melyeket a nyelvészeti szakirodalomból gyűjtöttek. A bináris címkék a mondat elfogadhatóságát jelzik.
– Az SST (Stanford Sentiment Treebank, Socher és mtsai, 2013) az egyik legismertebb szentiment annotációt tartalmazó angol nyelvű korpusz. A korpusz összeállításához 10 662 mondatot gyűjtöttek a Rotten Tomatoes oldaláról. A mondatokat Stanford Parserrel elemezték, és az így kapott 215 154 frázist egyesével annotáltatták egy 25 fokú érzelmi skálán. A 25 fokú skálát az SST5-nek nevezett változatban 0-5-ig terjedő skálára konvertálták, az SST2-ben pedig binárissá. A GLUE részeként az SST2-t szerepeltetik a szerzők, és abból is hangsúlyosan csak az egész mondatokat, frázisszintű elemeket nem. A GLUE adatbázisából így több mint 70 000 „mondatot" és a hozzájuk tartozó címkét lehet letölteni. (A két szám – ti. a 70 000 és a 10 662 – közti különbségről ld. a magyar szentimentkorpusz létrehozásáról szóló alfejezetet.)
– Az MRPC (Microsoft Research Paraphrase Corpus, Dolan és Brockett, 2005) online hírportálok tartalmából automatikusan válogatott mondatpárok gyűjteménye, amelyekben humán annotátorok címkézték fel, hogy a két mondat szemantikailag ekvivalens-e.
– A QQP (Quora Question Pairs)[1] adathalmazon az MRPC-hez hasonlóan mondatpár-osztályozás a feladat, de ebben a Quora oldaláról nyert kérdések ekvivalenciájának megállapítása a feladat.
– Az STS (Semantic Textual Similarity Benchmark, Cer és mtsai, 2017) online hírekből, videó- és képfeliratokból, és más NLP-feladatokból nyert mondatpárokat tartalmaz, amelyeknél a két mondat közti jelentésbeli hasonlóság van címkézve (humán annotátorok által, 1-5 skálán).
– Az MNLI (Multi-Genre Natural Language Inference Corpus, Williams és mtsai, 2018) az SNLI (Bowman és mtsai, 2015) utódja, 10 különböző forrásból származó, humán annotációt tartalmazó mondatpárokból áll, melyeknél a feladat annak megállapítása, hogy a második mondat következik-e az elsőből, vagy ellentmond neki, esetleg semleges a viszonyuk.
– A QNLI adathalmaz a SQuAD következtetéses feladattá alakított változata a GLUE-ban: a mondatpárok létrehozásához minden kérdést párosítottak a releváns szövegrészlet minden egyes mondatával.
– 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.

[1] https://data.quora.com/First-Quora-Dataset-Release-Question-Pairs

– A WNLI korpusz (Winograd NLI) a Winograd Schema Challenge (Levesque és mtsai, 2012) adatainak átalakított változata. A Winograd sémák lényege, hogy egy adott mondatban szereplő névmás referensét kell kiválasztani egy listából. A feladatban szereplő példákat a GLUE-ban mondatpár-klasszifikálássá formálták oly módon, hogy a névmás helyére beillesztettek minden egyes lehetséges referenst, és az így létrejött mondatról (mint hipotézisről) kell eldönteni, hogy az eredeti, névmást tartalmazó mondatból (a premisszából) következik-e.

A GLUE benchmark ezeken kívül tartalmaz egy ú.n. diagnosztikai adathalmazt is. A kézzel ellenőrzött adatbázis célja, hogy a modellek teljesítményét aprólékosan, nyelvi jelenségek széles körén lehessen elemezni. Ehhez többszáz mondatpárt válogattak össze, amelyek következtetési viszonnyal vannak címkézve mindkét irányban (következés, ellentmondás, semleges viszony). Ezen kívül mindegyik mondatpár annotációja tartalmazza olyan nyelvi jelenségeknek a címkéit, amelyek a két mondat között fennálló következtetési viszonyt alátámasztják.

Végül a GLUE oldalán találunk egy dicsőségtáblát is, amelyen az egyes modellek eredménye látható, a 9 feladaton külön-külön, és összesítve is.[2]

A SuperGLUE (Wang és mtsai, 2020) létrehozását az motiválta, hogy a GLUE már túl könnyűnek bizonyult a nyelvmodellek számára. Így a szerzők igyekeztek nehezebb feladatokat célzó korpuszokat összeválogatni. Ennek eredményeképp a SuperGLUE a következő hat korpuszt tartalmazza (a GLUE-ban a legnehezebbnek bizonyuló, és ezért a SuperGLUE-ban is megtartott RTE és WNLI korpuszokon túl):

– QA feladatok
  - A BoolQ (Boolean Questions, Clark és mtsai, 2019) korpuszban egy szövegrészlet és egy eldöntendő kérdés alkot egy példát. A kérdéseket automatikusan gyűjtötték, és automatikusan rendelték hozzá Wikipédia-szócikkekhez.
  - A MultiRC korpusz (Multi-Sentence Reading Comprehension, Khashabi és mtsai, 2018) 871 bekezdéshez kapcsolódó 6 000, többmondatos (*multi-sentence*) kérdést tartalmaz. Többmondatos kérdés alatt azt értik a szerzők, hogy a kérdésre az információt több mondatból kell összegyűjteni. A kérdések feleletválasztós kérdések: több lehetőség közül kell kiválasztani a (több) helyes választ. A több lehetséges, de előre nem definiált darabszámú jó válasz arra kényszeríti a modelleket, hogy minden kérdés-válasz párt egyesével kiértékeljenek.
  - A ReCoRD (Reading Comprehension with Commonsense Reasoning Dataset; Zhang és mtsai, 2018) korpusz a szövegértési feladatok témakörén belül olyan kihívás elé állítja a modelleket, amelyben az emberi megértéshez hasonlóan kell értelmes következtetéseket levonni egy adott szövegből. Ehhez a CNN/Daily Mail[3] korpusz segítségével állítottak elő több

[2] A cikk írásakor (2021.10.06.) az első helyen az ERNIE modell áll (https://github.com/PaddlePaddle/ERNIE).

[3] https://github.com/abisee/cnndailymail

bekezdésből álló szövegeket, melyeknek a záró bekezdésében – mely az előző szövegrésznek egyfajta lezárása, konklúziója – egy már ismert tulajdonnevet elmaszkoltak: a feladat az elmaszkolt névelem kiválasztása egy listából. A korpusz több, mint 120 000 bejegyzést tartalmaz.

– jelentésegyértelműsítés
  - A WiC korpuszban (Word-in-Context, Pilevar és Camacho-Collados, 2019) mondatpárok bináris osztályozása a feladat: adott két mondat és egy poliszém szó[4], amely mindkét mondatban előfordul; a feladat pedig annak eldöntése, hogy az adott szó mind a két mondatban ugyanabban a jelentésében szerepel-e.
– következtetéses feladatok
  - A CB (CommitmentBank, 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. 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 ITA.
  - A CoPA (Choice of Plausible Alternatives, Roemmele és mtsai, 2011) ok-okozati összefüggésekre koncentrál: 1 000 kérdést tartalmaz, 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 két részletesen bemutatott *benchmark* mellett természetesen további, hasonló céllal készült adatbázisok is léteznek szép számmal, melyeknek a részletes bemutatása túlmutat a jelen tanulmány keretein.

3. A HuLU alkorpuszai

A két legnagyobb, *multitask benchmarkba* beválogatott 15 korpusz közül mi 13-nak a létrehozását tűztük ki célul. Egyrészt a redundancia csökkentése miatt nem tervezünk QQP korpuszt (hiszen a feladat megegyezik az MRCP korpusszal, és részben feldolgozásra kerül az STS-ben), illetve nem foglalkozunk a SQuAD magyar párjának létrehozásával sem, mivel annak elkészítésén egy másik magyar nyelvtechnológiai műhely már dolgozik.

A 13 kiválasztott korpusz két csoportra osztható: egyrészt vannak azok az adatbázisok, melyeket fordítással létrehozhatunk, mert a bennük fókuszban lévő feladat nem nyelvspecifikus, és az adathalmaz jól fordítható; másrészt vannak

[4] A kitétel, hogy minden esetben poliszém szóról van szó, a SuperGLUE-t bemutató cikkben szerepel csak (Wang és mtsai, 2020, 6). A WiC korpuszt bemutató cikkben leírt módszertanból ilyen kitétel nem derül ki – erre a WiC-kel kapcsolatos nehézségeket bemutató alfejezetben bővebben is kitértünk.

olyan adatbázisok, amelyeket nem tudunk fordítani a jelenség nyelvspecifikus volta vagy a korpuszban szereplő szövegek összetettsége miatt. A következőkben bemutatjuk i) a saját erőforrásokból, magyar szövegekből, annotátorokkal előállított korpuszainkat, illetve ii) az angol korpuszok gépi fordításával, majd fordításellenőrzéssel és annotátori munkával előállított korpuszainkat.

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.

3.1. Önállóan előállított korpuszok

HuCOLA A magyar COLA korpusz előállításához 9 944 példát gyűjtöttünk négy nagyobb, összefoglaló jellegű szakirodalmi tételből (Kiefer, 2015; Alberti és Laczkó, 2017a,b; É. Kiss és Hegedűs, 2021). A gyűjtés a következő szempontok szerint történt:

- Minden példamondatot kigyűjtöttünk a cikk írója által adott elfogadhatósági ítélet típusától függetlenül.
- A *Megnézzük (*a) Budapest hídjait.* típusú példákból két bejegyzést készítettünk: *Megnézzük Budapest hídjait.* és **Megnézzük a Budapest hídjait.*
- Ha egy mondat azért nem elfogadható, mert egy adott jelentést nem "jelenthet", azt nem gyűjtöttük, pl.: **Megver Péter.* 'Péter megveri Pétert' jelentésben (Kiefer, 2015, 49.)
- Ha egy mondatban értelmetlen szó szerepel, nem gyűjtöttük.
- Ha a fókusz helyzete miatt nem jó egy mondat, nem gyűjtöttük.
- Előíró szabályokat megsértő mondatokat nem gyűjtöttünk (*háttal* nem kezdünk mondatot).

A leírt gyűjtés során felvételre kerültek teljes mondatok és nem egész mondatok, frázisok, tagmondatok is. Mivel a megcélzott feladat mondatosztályozás, a nem teljes mondatnyi példákat egész mondatokra egészítettük ki.[5]

Az angol korpuszban a szerzők a gyűjtés után a leggyakoribb 100 000 angol szóra szűrték a korpuszt, és az ennél ritkább szavakat lecserélték. Mi nem alkalmaztunk ilyen szűrést a korpuszunkon, mert a *subword* alapú tokenizálás ezt már nem teszi szükségessé a mai nyelvmodellek esetében.

Minden egyes mondatot négy annotátor[6] címkézett fel. Az útmutató alapján azt kellett eldönteniük, hogy az adott mondat elfogadható-e, jó magyar mondatnak hangzik-e.

[5] A mondatkiegészítés alapelveit ld. az annotálási útmutatóban: https://github.com/nytud/HuCOLA.

[6] A feladathoz komolyabb nyelvészeti ismeretekkel nem rendelkező, nem nyelvészet szakon tanuló, vagy ott végzett annotátorokat választottunk. Összesen 12 annotátor dolgozott a korpuszon.

A mondatokat a gyűjtés során a bennük található nyelvi jelenségek alapján is felcímkéztük.

Bár a CoLA angol előzményénél a mondatok címkéi az eredetileg a nyelvész szerzők által meghatározott címkék voltak, mi a mondataink „eredő" címkéit kivettük az elemzésből. Ezzel garantáltuk, hogy a gyűjtés során hibásan feljegyzett címkék, vagy a nyomdahibák nem befolyásolják az adatok minőségét. A mondatok 69,2%-ában (6883 mondat) a négy annotátor ugyanazt a címkét rendelte a mondathoz. 22,2%-ban (2213 mondat) 3:1 arányban címkézték a mondatot. A 2:2 arányban annotált mondatokat (8,5%, 848 mondat) félretettük, ezek nem képezik részét az adatbázisnak. Ugyanakkor elérhetővé tesszük őket, mert értékes nyelvészeti kutatási anyagot jelentenek.

A mondatok végső címkéje a 3:1 arányú annotálás esetében a többség döntése alapján lett meghatározva. A GLUE-ban található arányokat követve az adatokat 80-10-10% arányban tanító-, validációs és teszthalmazra osztva adjuk közre.[7]

HuRC Az angol nyelvű ReCoRD alapján állítottuk elő a magyar nyelvű HuRC korpuszt. Zhang és mtsai (2018) automatikus módszerrel állították elő a ReCoRD-ot: több mint 120 000 példát nyertek ki a CNN/Daily News[8] korpuszból. A napi híreket több részre bontották (lásd 1. ábra bal oldali példa): főszöveg (*passage*), kérdés – az utolsó bekezdésben kimaszkolt tulajdonnév (*cloze-style query*), referenciaválasz (*reference answer*). A főszöveg a cikk első néhány bekezdéséből áll. A cikk utolsó bekezdésében, ami egyfajta cikklezáró passzus, szerepelnie kell egy olyan tulajdonnévnek, ami a főszövegben is előfordul. Ez a tulajdonnév a referenciaválasz. A konkrét szövegértési feladat során ezt a tulajdonnevet kimaszkolják, és a modellnek a megfelelő referenciaválaszt kell kiválasztania egy listából.

A magyar anyag előállításához a Népszabadság Online[9] napi cikkeit vettük alapul, ezek közül is azt a 396 886 cikket, amelyeknek volt címe, szövege és összefoglalója (*lead*-je) egyaránt. Ha valamelyik összetevő hiányzott egy cikkből, azt nem használtuk. Ezután kiválogattuk a 3-6 bekezdésből álló cikkeket. Fontos kritérium volt, hogy mind a főszöveg, mind a kérdés (az utolsó bekezdés) tartalmazzon tulajdonnevet.

A tulajdonnév felismeréséhez saját névelemfelismerő modellt tanítottunk a huBERT (Nemeskey, 2021) segítségével. A NER modell finomhangoláshoz a NerKor (Simon és Vadász, 2021) korpusz hivatalos tanító-validációs-teszt adathalmazait használtuk fel, valamint a Huggingface által nyújtott tokenszintű osztályozó könyvtárat.[10] NER modellünk 90,18 F-mértéket ért el a tesztanyagon.

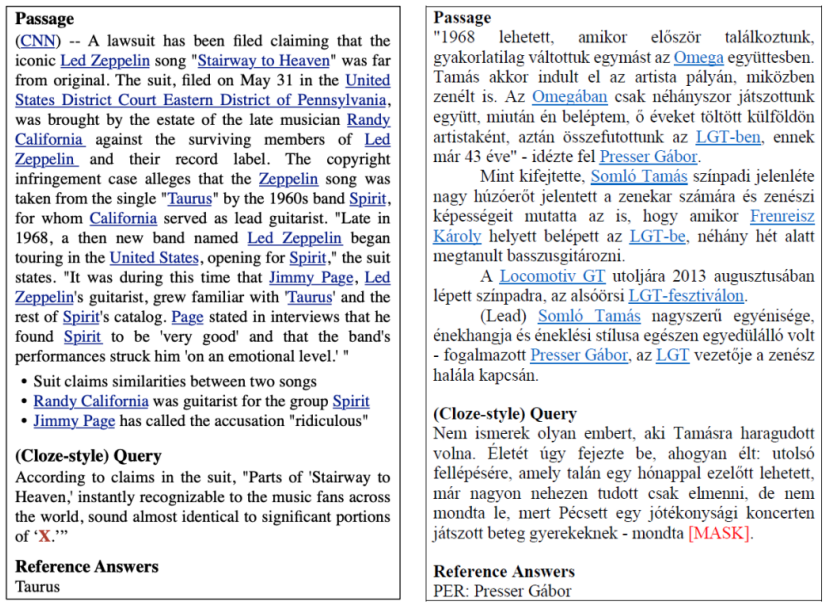
Utolsó lépésben megkerestük azokat a tulajdonnévpárokat, amelyek a főszövegben és a kérdésben is egyaránt szerepeltek. Egy cikkben több tulajdonnévpár

[7] https://github.com/nytud/HuCOLA, illetve a korpusz elérhető Huggingface-en, a *dataset card* linkje: https://huggingface.co/datasets/NYTK/HuCOLA.

[8] https://github.com/abisee/cnndailymail

[9] http://nol.hu

[10] https://github.com/huggingface/transformers/tree/master/examples/pytorch/token-classification



1. ábra: Egy ReCoRD (Zhang és mtsai, 2018) és egy HuRC példa

is előfordulhatott. A példánkban (lásd 1. ábra jobb oldali példa) a *Presser Gábor* mellett a *Tamás* szerepel még mind a kérdésben, mind a főszövegben. Ilyen esetekben egy adott cikket többször is belevettük az adatbázisba, más-más tulajdonnév párral. Így összesen 49 782 különböző típusú cikk (*type*) került kiválasztásra, amelyekből összesen 88 655 instancia alkotja az adathalmazunkat a több tulajdonnévpár jelensége miatt. Az automatikus módszerekkel előállított korpuszunk kvantitatív tulajdonságai a következők: cikkek száma: 88 655, különböző cikkek száma (type): 49 782, token: 27 703 631, type: 1 115 260, szövegrész átlagos hossza (token): 249,42 (medián: 229), kérdés átlagos hossza (token): 63,07 (medián: 56).

Az előállított korpuszunkon ezután végeztünk néhány apró javítást. Az így létrejött, javított adathalmazunkat 100-as egységenként egy-egy annotátorral ellenőriztettük. Az annotáláshoz saját magunk által készített annotálófelületet biztosítottunk. Az automatikus maszkolást kellett validálni a következő szempontok alapján: i) jó-e a névelem-felismerés és -maszkolás (tehát *Ferenc pápa* lett maszkolva, és nem csak *Ferenc*, illetve a *Gödöllőre* az nem [MASK]re, hanem [MASK]), továbbá ii) szerepel-e a cikk korábbi részeiben is az elmaszkolt névelem.[11] Az ellenőrzés eredményeképpen 80 587, automatikusan előállított, kézzel validált szövegegység szerepel az adatbázisban.[12]

[11] Összesen 12 annotátor dolgozott a korpuszon.
[12] https://github.com/nytud/HuRC, https://huggingface.co/datasets/NYTK/HuRC

A HuCommitmentBank A HuCommitmentBank-et a feladat nyelvspecifikus volta miatt magyar nyelvű példák gyűjtésével lehet csak előállítani. Az angol korpusz 1 200 diskurzusszegmenst tartalmaz. Mindegyik szegmens egy alárendelt mellékmondatot tartalmazó mondatból és az azt megelőző 2-3 mondatos kontextusból áll. A célmondatban a mellékmondatot vonzó főige egy *entailment canceling operator* alá van beágyazva (ezek az operátorok a tagadás, a kérdés, a feltételes mód és a modális módosítók). A példákat automatikus módszerekkel gyűjtötték, majd kézzel validálták.

A magyar gyűjtést több irányban indítottuk el, az MNSZ2 (Oravecz és mtsai, 2014) beszélt nyelvi alkorpuszán: egyrészt mintázatillesztéssel próbálkoztunk a felszíni alakokon, másrészt az angol módszertanhoz hasonlóan szintaktikai elemzéssel kerestünk megfelelő jelölteket. 4 annotátor kétheti munkájával előállt kicsivel több mint 1 000 példa. Ezek egy részének[13] a validálása már megtörtént. A kutatás következő fázisában a példákat annotátoroknak osztjuk ki, hogy megcímkézzék a célmondatokat.

HuWiC A jelentés-egyértelműsítési feladat bináris osztályozási feladattá egyszerűsítve jelenik meg Pilevar és Camacho-Collados (2019) korpuszában. Itt azt kell eldönteni egy célszóról és az azt tartalmazó mondatpárról, hogy a szóelőfordulások ugyanazt jelentik-e a két különböző kontextusban, vagy sem. A feladat, látszólagos egyszerűsége ellenére, meglehetősen nehéznek bizonyult (vö. Véronis, 2001).

Az angol WiC korpusz építése során már meglévő jelentéstárakban (WordNet, VerbNet, Wiktionary) található példamondatokra támaszkodtak.[14] Mivel a WordNet-en kívül nem állnak rendelkezésünkre a magyarra hasonló adatbázisok, illetve Véronis (2001) kísérletei rávilágítottak a feladat komplexitására, úgy döntöttünk, az angol WiC korpusz létrehozásának módszertanától eltérő módon építjük meg a saját adatbázisunkat. Első lépésben a legegyszerűbbnek tűnő feladattal, az egyjelentésű főnevek csoportjával kezdtünk. Azokat a főneveket tekintettük egyjelentésűnek, amelyek a HuWN-ben (Miháltz és mtsai, 2008) egy jelentéssel szerepeltek, és a hozzájuk rendelt ÉKSz. (Pusztai, 2003) link a címszó alatti 1.1-es jelentésre mutatott.[15] Így egy 5 981 elemű lista állt elő.

Mivel a HuWN példamondatai nem tükrözik a természetes nyelvhasználatot, ezért a korpuszunkhoz a példamondatokat az MNSZ2-ből (Oravecz és mtsai, 2014) nyertük ki automatikus módszerekkel.

[13] A cikk leadásakor a példák mintegy 90%-a lett már validálva.

[14] Fontos megjegyezni, hogy a WiC korpusz létrehozásának leírásánál nem esik szó arról, hogy a gyűjtés során kifejezetten poliszém szavakra fókuszáltak volna; megfogalmazásuk szerint első lépésben mindent gyűjtöttek. Éppen ezért nem világos, hogy Wang és mtsai (2020) miért írják, hogy a korpusz poliszém célszók viselkedését vizsgálja (ld. a bevezetőben részletezett leírást.)

[15] Ezzel a HuWN-ben egyjelentésűként feltüntetett főnevek közül kiszűrtük azokat, amik az ÉKSz.-ben biztosan többjelentésűek. Az 1.1-es jelentésre való utalás azonban nem zárja ki egyértelműen, hogy az adott főnévnek az ÉKSz.-ben van más jelentése is.

A mondatpárok gyűjtésénél a GDEX (Good Dictionary EXamples, Kilgarriff és mtsai, 2008) leírásból indultunk ki. Az abban megfogalmazott kritériumok lehetővé teszik, hogy kvantitatív módon megragadhatóak legyenek a jó példamondatok. Összefoglalva tehát, első lépésben az volt a célunk, hogy olyan mondatpárokat gyűjtünk automatikusan, amelyekben a célszó garantáltan ugyanazzal a jelentéssel szerepel.

125 célszóhoz tartozó 5-5 példamondat kvalitatív kiértékelése azt mutatta, hogy a gondos gyűjtés ellenére a célszavak jelentős része eltérő jelentéssel szerepel a példamondatokban. Szerencsére ezek a többjelentésű esetek csoportokba sorolhatók (szisztematikus metonímia, deverbális főnevek, kollokációk), így egy részletes annotálási útmutatóval a feladat jól definiálttá tehető. A korpusz előállításához fontosnak tartjuk a megfelelő annotálási alapelvek kidolgozását, elsősorban formai, szintaktikai kritériumok alapján, amennyiben ez lehetséges.

3.2. Fordított korpuszok

HuCoPA A CoPA 1000 kérdését először gépi fordítóval fordíttattuk, ennek a kimenetét annotátorok ellenőrizték és javították, a fluenciára törekedve. Egy-egy annotátor pedig megjelölte a helyes választ a kérdésre. Ha az annotátor döntése és az eredeti címke között eltérést találtunk, akkor kézzel ellenőriztük az adott példát. Az annotálás harmadik lépésében derült fény például a CoPA egyik hibás címkéjére (a *training set* 380-as id-jú kérdése). Így végül előállt az 1 000 egységből álló HuCoPA korpusz.[16] Ebből az eredeti felosztást megtartva 400-at tanító-, 100-at validációs, 500-at pedig tesztanyagnak különítettünk el. Ahhoz, hogy a tesztanyag hibátlanságát biztosítsuk, minden, a tesztanyagban szereplő mondatot 4-4 annotátornak adunk, így egyszerre mérjük a humán teljesítményt a korpuszon és validáljuk is a címkéket az ITA vizsgálatával. Ez a lépés, ti. a tesztanyag annotálása még hátravan.

A HuSST A szentimentkorpusz lefordításakor annak a GLUE-ban szereplő formája helyett az ún. SST-5 adathalmazt vettük alapul. A GLUE alkorpuszai közt található SST-2 adatok letöltésekor szembesültünk azzal, hogy bár a GLUE-cikk szerzői kizárólag a teljes mondatok használatát jelzik (és mondatosztályozási feladatként fogalmazzák meg a kérdést),[17] a fájlokban rengeteg frázis található (néhány példa a tanítóanyagból: *of saucy*, *in world cinema*, *a doa*). Ez okozza tehát az SST bemutatásakor említett 10 662 mondat és a 70 600-as GLUE-s korpusz közti méretbeli különbséget. Megjegyzendő, hogy itt 11 855 mondatot találtunk, amiket gépifordítóval magyarra fordíttattunk. Ezt követték a HuCoPA korpusznál ismertetett ellenőrzési lépések. Végül minden magyar mondatot három-három annotátor címkézett a szentiment alapján egy hármas skálán. A szentimentcímkéket egy kurátor nézte át, aki végleges címkével látta

[16] https://github.com/nytud/HuCoPA; illetve a korpusz elérhető Huggingface-en, a *dataset card* linkje: https://huggingface.co/datasets/NYTK/HuCoPA

[17] „We use the two-way (positive/negative) class split, and use only sentence-level labels.”, (Wang és mtsai, 2018, 3).

el a mondatokat.[18] 7064 mondatnál (59.6%) teljes volt az egyetértés a három annotátor között, 4619 (38,96%) esetben pedig 2:1 arányban címkéztek. A végső címke minden esetben a kurátor döntése.[19] 172 mondatot nem használunk az adatbázisban, ezeknél a három annotátor három különböző címkével látta el a mondatot.

A Winograd-sémák A Winograd-sémák a referencia-feloldás feladatát célozzák. A feladat nem triviális: *The man couldn't lift his son because he was so (weak/heavy).* A Winograd-sémákat már más nyelvekre is lefordították (japán, francia, portugál, kínai, héber). Az angol eredetit először a már ismertetett gépi fordító segítségével lefordítottuk, majd ennek kimenetét két ember validálta. Bizonyos sémákat elvetettünk, mert nem tudtuk őket úgy lefordítani, hogy megőrizzük bennük a Winograd-sémák jellemzőit (pl. *Lily spoke to Donna, breaking her (silence/concentration)*: a két angol kifejezés nem fordítható le magyarra úgy, hogy pusztán egy szóban térjen el a két mondat, de mindkettőben megmaradjon a birtokos szerkezet). Más esetekben kis módosítással ültettük át magyarra az eredeti sémát. A Winograd-sémák önálló kutatási témát jelentenek, amelynek részletes kifejtése túlmutat a jelen tanulmány keretein.

Az eredeti adathalmaz 150 mondatpárt tartalmaz. A fordítás és a validálás után 122 magyar séma lett.[20]

HuRTE Az RTE adathalmazok GLUE-ba beválogatott részét a már ismertetett gépi fordítóval fordíttattuk magyarra. Ezt követően az így előállt 9 000 példát (amely nagyjából 18 000 mondatot jelent) az SST-2 magyar párjának előállításánál ismertetett módon annotátorokkal ellenőriztettük, javíttattuk. Következő lépésként minden példát annotátorokkal címkéztetünk, hogy kiderüljön, ha a fordítás során elveszett a mondatpárok közti eredeti következtetési viszony. Az így validált adathalmazt a többi korpusznál ismertetett módon közzétesszük.

3.3. További alkorpuszok

A magyar *benchmark* korpuszba a fentieken kívül egy *multi-sentence reasoning* korpuszt is készítünk majd. Az Oktatási Hivatallal történő együttműködés keretében rendelkezésünkre állnak majd a PISA-felmérések szövegértési feladatai az elmúlt 21 év felméréseiből. Ezek a szövegértési feladatok jellegükből adódóan éppen a MultiRC korpusz céljainak megfelelő szövegek: a kérdések megválaszolásához több bekezdésen átívelő, komplex szövegértésre van szükség. Ezen felül kézzel validált, válogatott szövegekről van szó. A feladatok kérdései közül pedig a korpusz céljainak megfelelően a feleletválasztósakat építjük majd be az adatbázisba.

[18] A fordításellenőrzést 12 annotátor végezte, a fluenciajavítást 8. 11-en végezték a szentimentannotálást, és 4 fő foglalkozott a kurátori feladattal.

[19] Az adatbázis elérhetősége: https://github.com/nytud/HuSST, illetve https://huggingface.co/datasets/NYTK/HuSST.

[20] https://github.com/nytud/HuWSC, https://huggingface.co/datasets/NYTK/HuWSC

A BoolQ korpusz magyar párját nem fordítással, hanem az angol korpusz létrehozásának módszertanát követve, magyar nyelvű Wikipédia-szócikkekből és annotátori munka segítségével állítjuk majd elő. A két nagy korpuszt, a szemantikai hasonlóságra fókuszáló STS-t és a következtetéses feladatot képviselő MNLI-t csak komolyan nyelvészeti kutatómunka után tudjuk elkezdeni megtervezni és előállítani, így ezek még a jövő feladatai közé tartoznak.

4. Összegzés

A tény, hogy a magyar modellek kiértékelésére létrehozott *benchmark* 1-2 év késéssel követi angol elődjét, az előnyünkre fordítható: több tanulmány is vizsgálta időközben a *benchmarkok* hiányosságait. Moradi és Samwald (2021) a BERT, az XLNet, a RoBERTa és az ELMo teljesítményét vizsgálták különböző feladatokon, elsősorban arra fókuszálva, hogy a modellek hogyan teljesítenek zajos inputon. Eredményeik azt mutatták, hogy az említett nyelvmodellek kifejezetten érzékenyek a bemeneti szövegek minőségére, és teljesítményük már a legapróbb változtatástól is jelentősen romlani kezd. Arra is rávilágítanak, hogy a jelenleg használt *benchmarkok* nem mérik megfelelően a modellek robosztusságát. Érvelésük szerint a jelenleg használt *benchmarkokat* ki kéne egészíteni a zajos inputokon történő kiértékeléssel, hogy az NLP rendszerek robosztusságáról reálisabb képet kapjunk.

A feladatok kiválasztásán, és a feladathoz készített tanító- és tesztanyag megfelelő összeállításán túl természetesen az annotáció minősége is kulcsfontosságú. Mivel a *benchmarkok* előállításának folyamata egy vagy több ponton is tartalmaz automatizált lépéseket, a hiba lehetősége adott. Több tanulmány is vizsgálta, hogy a tanítóanyagban lévő hibás címkék milyen hatással vannak a modellekre. Általában az az álláspontja a kutatóknak, hogy valójában nem jelent nagy problémát a hibás tanítócímke, a neurális modellek elég robosztusak a hibás címkékkel kapcsolatban (erről ld. például Mahajan és mtsai, 2018; Rolnick és mtsai, 2017). Ezzel szemben a tesztadatok közt előforduló hibás címke komoly gondot okoz: Northcutt és mtsai (2021) 10 nagyobb adatbázis tesztanyagát vizsgálva 3.3%-osra becsülte a hibás címkék arányát – ami értelemszerűen túl sok, ha a tesztanyagon mért teljesítmény (például pontosság) alapján ítélünk megbízhatónak egy-egy nyelvmodellt.

Következő lépésként az előzőekben bemutatott korpuszokon szeretnénk humán teljesítményt mérni, és eközben a korpuszok tesztanyagnak elkülönített részét validálni, hogy megbízhatóságukról meggyőződve adhassuk azokat közre. Hosszabb távú célunk pedig az, hogy a magyar nyelvű nyelvmodellek előállításának és tesztelésének egyik hátráltató tényezőjét, a *benchmarkok* hiányát szisztematikusan felszámoljuk azzal, hogy a GLUE-t és SuperGLUE-t alkotó alkorpuszok mellett más, újabb, nehezebb feladatot megcélzó, robusztusabb modellek létrehozását ösztönző és karbantartott *benchmark* készlet álljon elő műhelyünkben.

Hivatkozások

Alberti, G., Laczkó, T. (szerk.): Syntax of Hungarian. Nouns and Noun Phrases, Volume 1. Comprehensive Grammar Resources, Amsterdam University Press (2017a)

Alberti, G., Laczkó, T. (szerk.): Syntax of Hungarian. Nouns and Noun Phrases, Volume 2. Comprehensive Grammar Resources, Amsterdam University Press (2017b)

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)

Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., Specia, L.: SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). pp. 1–14. Association for Computational Linguistics, Vancouver, Canada (Aug 2017), https://aclanthology.org/S17-2001

Clark, C., Lee, K., Chang, M.W., Kwiatkowski, T., Collins, M., Toutanova, K.: BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions. 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. 2924–2936. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019), https://aclanthology.org/N19-1300

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)

Dolan, W.B., Brockett, C.: Automatically Constructing a Corpus of Sentential Paraphrases. In: Proceedings of the Third International Workshop on Paraphrasing (IWP2005) (2005), https://aclanthology.org/I05-5002

É. Kiss, K., Hegedűs, V. (szerk.): Syntax of Hungarian. Postpositions and Postpositional Phrases. Comprehensive Grammar Resources, Amsterdam University Press (2021)

Feldmann, A., Hajdu, R., Indig, B., Sass, B., Makrai, M., Mittelholcz, I., Halász, D., Yang, Z.G., 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)

Ghazal, A., Ivanov, T., Kostamaa, P., Crolotte, A., Voong, R., Al-Kateb, M., Ghazal, W., Zicari, R.V.: BigBench V2: The New and Improved BigBench. In: 2017 IEEE 33rd International Conference on Data Engineering (ICDE). pp. 1225–1236 (2017)

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)

Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., Johnson, M.: XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalization (2020)

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)

Khashabi, D., Chaturvedi, S., Roth, M., Upadhyay, S., Roth, D.: Looking Beyond the Surface: A Challenge Set for Reading Comprehension over Multiple Sentences. In: NAACL (2018)

Kiefer, F. (szerk.): Strukturális magyar nyelvtan 1. Mondattan. Akadémiai Kiadó, Budapest (2015)

Kilgarriff, A., Husák, M., McAdam, K., Rundell, M., Rychlý, P.: GDEX: Automatically Finding Good Dictionary Examples in a Corpus. In: Elisenda Bernal, J.D. (szerk.) Proceedings of the 13th EURALEX International Congress. pp. 425–432. Institut Universitari de Linguistica Aplicada, Universitat Pompeu Fabra, Barcelona, Spain (jul 2008)

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)

Mahajan, D., Girshick, R., Ramanathan, V., He, K., Paluri, M., Li, Y., Bharambe, A., van der Maaten, L.: Exploring the Limits of Weakly Supervised Pretraining (2018)

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

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 GWC 2008. pp. 310–320 (2008)

Moradi, M., Samwald, M.: Evaluating the Robustness of Neural Language Models to Input Perturbations. Computer Science – Computation and Language (2021)

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

Northcutt, C.G., Athalye, A., Mueller, J.: Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks. ArXiv abs/2103.14749 (2021)

Oravecz, C., Váradi, T., Sass, B.: The Hungarian Gigaword corpus. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). pp. 1719–1723. European Language Resources Association (ELRA), Reykjavik, Iceland (May 2014), http://www.lrec-conf.org/proceedings/lrec2014/pdf/681_Paper.pdf

Pilevar, M.T., Camacho-Collados, J.: WiC: the Word-in-Context Dataset for Evaluating Context-Sensitive Meaning Representations. In: NAACL 2019 (Minneapolis, USA) (2019)

Pusztai, F. (szerk.): Magyar értelmező kéziszótár. Akadémiai Kiadó, Budapest (2003)

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

Roemmele, M., Bejan, C., Gordon, A.: Choice of Plausible Alternatives: An Evaluation of Commonsense Causal Reasoning. AAAI Spring Symposium - Technical Report (01 2011)

Rolnick, D., Veit, A., Belongie, S.J., Shavit, N.: Deep Learning is Robust to Massive Label Noise. ArXiv abs/1705.10694 (2017)

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)

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)

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C.D., Ng, A., Potts, C.: Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. pp. 1631–1642. Association for Computational Linguistics, Seattle, Washington, USA (Oct 2013), https://aclanthology.org/D13-1170

Tiedemann, J.: Parallel Data, Tools and Interfaces in OPUS. In: Chair), N.C.C., 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 (may 2012)

Véronis, J.: Sense Tagging: Does It Make Sense? In: Corpus Linguistics Conference (2001), http://www.up.univ-mrs.fr/ veronis/pdf/2001-

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

Warstadt, A., Singh, A., Bowman, S.R.: Neural Network Acceptability Judgments. arXiv preprint arXiv:1805.12471 (2018)

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

Zhang, D., Mishra, S., Brynjolfsson, E., Etchemendy, J., Ganguli, D., Grosz, B., Lyons, T., Manyika, J., Niebles, J.C., Sellitto, M., Shoham, Y., Clark, J., Perrault, R.: The AI Index 2021 Annual Report (2021)

Zhang, S., Liu, X., Liu, J., Gao, J., Duh, K., Durme, B.V.: ReCoRD: Bridging the Gap between Human and Machine Commonsense Reading Comprehension (2018)