MILQA kérdés-válasz benchmark adatbázis

Novák Attila[1,2], Novák Borbála[1,2]

[1]Pázmány Péter Katolikus Egyetem, Információs Technológiai és Bionikai Kar
[2]MTA-PPKE Magyar Nyelvtechnológiai Kutatócsoport
Budapest, Práter u. 50/a.
{vezetéknév.keresztnév}@itk.ppke.hu

Kivonat Jelen cikkben bemutatott kutatás során új kérdés-válasz benchmark adatbázist hoztunk létre magyar nyelvre. A MILQA[1] adatbázis létrehozásakor nagyrészt az angol nyelvű SQuAD 2.0 elveit követtük, azonban néhány a SQuAD-nál újabb keletű angol nyelvű kérdés-válasz adatbázishoz hasonlóan bevezettünk számos a SQuAD-on túlmutató újítást: többek között igen/nem válaszokkal megjelölt eldöntendő kérdéseket, több szakaszból álló listaszerű válaszokat, hosszú válaszokat, számolást igénylő és egyéb olyan kérdéseket, ahol nem másolható ki egyszerűen a válasz a szövegből.[2]
Kulcsszavak: kérdés-válasz rendszerek, benchmark adatbázis

1. Bevezetés

A számítógépet használó laikus felhasználók részéről természetes igényként jelenik meg, hogy természetes nyelven tehessenek fel kérdéseket a rendszernek, és arra értelmes és releváns természetes nyelvű válaszokat kapjanak. A mesterséges (gépi) intelligencia Alan Turing általi korai definíciója (a később Turing-teszt néven ismertté vált „imitációs játék”)[3] is ilyen jellegű interakció köré épül: javaslata szerint intelligensnek az embertől megkülönböztethetetlen módon társalgó gépet tekinthetjük. A neurális nyelvmodellek térhódításával ez az igény a korábbiaknál sokkal inkább kielégíthetővé vált. A gépi szövegértés tesztelésének egy módja (hasonlóképpen ahhoz, ahogy az emberi szövegértést tesztelni szokták), hogy a szöveg alapján a szöveghez kapcsolódó kérdésekre kell válaszolni, és a válaszok helyességét ellenőrizzük. Angol nyelvre viszonylag jól működő kérdés-válasz rendszerek léteznek, amelyeknek alapját az alapkutatás keretében (a legtöbb esetben azonban céges kutatólaborok által finanszírozott módon) létrehozott adatbázisok

[1] A korpusz nevét a lila tehenes csoki mellett a projektet támogató Mesterséges Intelligencia Nemzeti Labor kezdőbetűi ihlették.

[2] Az adatbázist a tesztkorpusz ellenőrzését követően publikáljuk: `https://github.com/ppke-nlpg/MILQA`.

[3] Az eredeti imitációs játékban két különböző nemű szereplő feladata az volt hogy a valódi nemüktől függetlenül adott nemű személynek adják ki magukat, és a harmadik szereplőnek ki kellett találnia pusztán a megnyilatkozások tartalma alapján, hogy melyikük az, aki csal.

képezik. Jelen kutatásban olyan magyar nyelvű kérdés-válasz adatbázis létrehozása volt a célunk, amely lehetővé teszi magyar nyelvű kérdés-válasz rendszerek betanítását, illetve teljesítményük gépi kiértékelését.

2. Előzmények

A korai kérdés-válasz adatbázisok vagy nagyon kis méretűek voltak, vagy nem tartalmaztak grammatikailag kérdő mondat formájú kérdéseket, hanem ún. cloze-típusú „kérdésekből" álltak: ezek olyan kijelentő mondatok, amelyeknek egy része ki van maszkolva, és azt kell kiegészíteni a szöveg alapján. Az utóbbi erőforrásokat géppel generálták, így könnyen létrehozhatóak voltak, azonban a maszkolt részt tartalmazó mondatok egyáltalán nem hasonlítanak a valódi kérdésekhez. Az egyik legfontosabb mérföldkő a kérdés-válasz tanító-adatbázisok sorában a Stanfordi Egyetemen létrehozott angol nyelvű SQuAD adatbázis (Stanford Question Answering Dataset) volt (Rajpurkar és mtsai, 2016). Ez egy a korábbiaknál jóval nagyobb méretű, első verziójában több mint 108000 kérdés-válasz párt tartalmazó adatbázis, amelyet később a második verzióban az adott szövegrészlet alapján nem megválaszolható kérdésekkel egészítettek ki (151000 kérdés, (Rajpurkar és mtsai, 2018)). A nyilvánosan elérhető tanító és a hangolókészlet 143000 (93000 megválaszolható és 50000 nem megválaszolható) kérdést tartalmaz. A méretén túl azért tekinthető mérföldkőnek ez az erőforrás, mert egyrészt ténylegesen jólformált kérdéseket tartalmazott a korábbi ún. cloze-típusú kérdéseket tartalmazó erőforrásokkal ellentétben (pl. CNN/Daily Mail (Hermann és mtsai, 2015)), illetve nem feleletválasztós volt (mint pl. az MCTest (Richardson és mtsai, 2013) vagy a WikiQA (Yang és mtsai, 2015)). Valamint mert óriási lökést adott a kérdés-válasz rendszerek fejlesztésének.

A kérdés-válasz adatbázisok, illetve rendszerek körében megkülönböztethetünk extraktív, illetve absztraktív/generatív megközelítést. Az előző esetében a választ egyszerűen a szövegből szó szerint kiemelt rész adja (mintha szövegkiemelővel dolgoznánk, ilyen a SQuAD), az utóbbi esetében pedig ténylegesen jólformált emberi nyelven van a válasz megfogalmazva (pl. MS MARCO (Nguyen és mtsai, 2016), NarrativeQA (Kočiský és mtsai, 2018)). Ezen túl a kérdés-válasz adatbázisok egy része olyan kérdéseket tartalmaz, amelyeknek megválaszolása többlépéses következtetési láncok végrehajtását követeli meg (multi-hop/többlépcsős kérdés-válasz feladatok). Ez nemcsak a mögöttes logikai levezetés nagyobb komplexitását jelenti, hanem ez a feladattípus így túlléphet az egyedi dokumentumok, illetve szövegrészletek szintjén is, amennyiben az adott kérdés csak több különböző dokumentumban vagy szövegrészben szereplő információk kombinálásával válaszolható meg (pl. HotpotQA (Yang és mtsai, 2018), NarrativeQA (Kočiský és mtsai, 2018)).

A többlépcsős kérdés-válasz feladatok és a SQuAD esetében is az adott szövegek alapján az annotátorok feladata volt a kérdések megfogalmazása. A nagy keresőmotorokat üzemeltető cégek ezzel szemben olyan erőforrásokat hoztak létre, amelyekben a keresőmotorba beírt gyakori kérdésekből kiindulva gyűjtöttek össze hozzájuk tartozó releváns dokumentumokat, és ezeket felhasználva jelölték

ki vagy fogalmazták meg az annotátorok a válaszokat. Az előbbi extraktív típusú adathalmaz a Google keresőbe beírt kérdéseken alapuló Natural Questions (NQ, Kwiatkowski és mtsai (2019)), ahol a kontextusként használt dokumentumok a SQuAD-hoz hasonlóan Wikipedia szócikkek voltak. Az utóbbi absztraktív/generatív típusba tartozik a Microsoft Bing-kérdéseken alapuló MS MARCO QnA adathalmaza (Nguyen és mtsai, 2016). Hasonlóképpen kereséssel jöttek létre a szintén eleve létező kvíz- és műveltségi kérdéshalmazokon alapuló erőforrások is (pl. TriviaQA (Joshi és mtsai, 2017)).

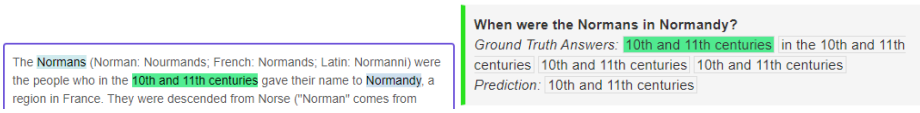
A SQuAD népszerűségének egyik forrása talán az, hogy egy olyan viszonylag leegyszerűsített modellt feltételez, amely szerint minden megválaszolható kérdésre egyetlen összefüggő szövegrész kijelölhető válaszként, és ez a SQuAD-alapú rendszerek implementációját jelentősen egyszerűsíti. Ez a megszorítás többékevésbé jól megvalósítható, ha az annotátorok eleve azt a feladatot kapják, hogy így megválaszolható kérdéseket tegyenek fel. A valóságban előforduló kérdések nem elhanyagolható része esetében azonban a válasz valamilyen lista, aminek az elemei nem feltétlenül összefüggő részt foglalnak el a szövegben. Ilyenkor az összes releváns részt magában foglaló összefüggő szakasz jelentős mennyiségű a válasz szempontjából irreleváns szövegrészt tartalmazhat. Ezért például a valóságos kérdéseken alapuló NQ adathalmazban nem korlátozták a válasz lehetőségét egyetlen összefüggő szakaszra, hanem ilyenek listája lehet a válasz (a Natural Questionsben a kérdések 6,9%-ánál nem egyelemű lista a válasz). A SQuAD esetében a kérdések kontextusa (a szövegrész, amiben a kérdésre a választ meg kell találni) viszonylag korlátozott hosszúságú: 150 és 4000 karakter közötti, átlagosan 740 karakter, ami szintén behatárolja a feladat komplexitását.

Az NQ-hoz hasonló ténylegesen feltett kérdésekből kiinduló adathalmazokban természetes módon előfordulnak eldöntendő kérdések is (Natural Questions: 2,5%). Ezekre önmagában tipikusan nem válasz a szöveg egy kijelölt részlete, hanem valamely releváns szövegrészből következik egy (bizonyos esetekben nem egyértelmű, csak valószínűsíthető) igen/nem válasz. Vannak kifejezetten csak eldöntendő kérdéseket tartalmazó adathalmazok is (pl. a szintén a Wikipedián alapuló BoolQ (Clark és mtsai, 2019), az Amazon termékértékeléssel kapcsolatos szövegeire épülő AmazonYesNo (Dzendzik és mtsai, 2019) vagy a orvosbiológiai cikkabsztraktokon alapuló PubMedQA (Jin és mtsai, 2019)). A BoolQ és az AmazonYesNo ugyanakkor jelentős átfedést mutat a NaturalQuestions, illetve az AmazonQA (Gupta és mtsai, 2019) adatbázisok eldöntendő kérdéseivel (az Amazonos erőforrások esetében lényegében részhalmaz viszony áll fenn).

Az NQ-hoz hasonlóan „természetes eredetű” orvosbiológiai témájú kérdéshalmazokban gyakran az NQ adatbázissal kapcsolatban korábban említettnél is jóval magasabb arányú a „nem SQuAD-kompatibilis” kérdések aránya. A ténylegesen praktizáló orvosok napi szakmai tevékenysége során megfogalmazott kérdéseket tartalmazó Clinical Questions Collection (CQC) adathalmaz (D'Alessandro és mtsai, 2004; Ely és mtsai, 1997, 1999) és a PubMed felhasználók által egy nap alatt megfogalmazott kérdésekből összeállított PubMed Query Log Dataset (Herskovic és mtsai, 2007) esetében például az eldöntendő kérdések aránya

28,1%-nak, a lista típusú válasszal rendelkezőké pedig 21,9%-nak adódik (Yoon és mtsai, 2022).

A lista típusú válaszok és az eldöntendő kérdések (valamint a korábban említett többlépcsős következtetést igénylő válaszok) hiánya miatt a SQuAD-típusú adatbázisokon betanítható egyetlen szövegszakaszt válaszként visszaadó extraktív rendszerek viszonylag hamar megbuknának a korábban említett Turing-teszten még abban az esetben is, ha egyébként tökéletes teljesítményt nyújtanának a maguk műfajában, mert ilyen típusú kérdésekre nem tudnának meggyőzően válaszolni. Ugyanakkor további probléma adódik abból a tényből, hogy egy adott szöveggel kapcsolatban az igazán érdekes kérdések nem azokat a szavakat használják, amelyek az eredeti szövegkörnyezetben szerepeltek (hiszen ellenkező esetben gyakran egyszerű lexikai kereséssel meg lehetne találni a választ). A SQuAD összeállítása során is törekedtek arra, hogy az annotátorok a kérdések megfogalmazásakor parafrazeálják a kérdés szövegkörnyezethez horgonyozó részét, és ne egyszerűen lemásolják azt. Ez önmagában nem feltétlenül jelent komoly problémát a kurrens előtanított nyelvmodelleken alapuló neurális modellek számára, hiszen ezek általában megfelelően absztrakt belső szemantikai reprezentációkkal bírnak ahhoz, hogy a parafrázisok ne zavarják meg őket (vagy legalábbis sokkal kevésbé, mint egy lexikai mintaillesztésen alapuló rendszert). A probléma azonban abból adódik, hogy ha a kérdésben például az eredeti kontextusban szereplőtől eltérő igét használunk, akkor sok esetben ez más vonzatkerettel jár, ami azt jelenti, hogy sokszor más esetben kellene szerepelnie a válaszban az adott kifejezésnek, mint amilyen formában az eredeti szövegben állt. A SQuAD esetében erre a problémára az volt a megoldás, hogy a prepozíciókat általában nem tették a válasz részévé, hanem csak a minimálisan szükséges lexikális tartalmat (erre utasították az annotátorokat, l. 1. ábra).



1. ábra. A SQuAD-ban az egyébként lényeges prepozíciók általában nem képezik részét a válaszoknak

Ez ugyan a lényegében izoláló nyelvnek tekinthető angol esetében az esetek nagy részében megoldja a fenti problémát, de azon az áron, hogy a rendszer többnyire „foghegyről” válaszol, és nem a pragmatikailag megfelelő alakban adja vissza/jeleníti meg a választ. Az emberek a megfelelő prepozíciót is beillesztenék a válaszba, tehát ezen a ponton is „kilóg a lóláb” a Turing-teszt szempontjából. Morfológiailag komplex (pl. agglutináló) nyelvek esetében pedig ez a megoldás nyilvánvalóan nem működik, ilyen esetekben egy SQuAD-típusú extraktív rendszer mindenképpen pragmatikailag inadekvát választ fog adni, mert az eredetileg a szövegben szereplő esetraggal, esetleg valamilyen képzős formában adja vissza a választ. Ugyanakkor ez a szövegkiemelős megjelenítés esetén mégsem jelent igazi

problémát, hiszen ebben az esetben a felhasználó nem úgy értékeli, hogy a gép „furcsán válaszolt”, hanem hogy „jól kiemelte a választ a szövegből”. Amennyiben ugyanakkor a válasz valóban válaszként jelenik meg, akkor mindenképpen célszerű továbblépni valamilyen generatív/absztraktív megoldás irányába.

Van néhány további kérdéstípus: a megszámlálást, valamilyen aritmetikai műveletek végrehajtását, illetve összehasonlítást igénylő kérdés-válasz párok *(hány, mennyi, melyik a leg... stb.),* amelyek már a minimális iskolázottságú embereknek sem jelentenek problémát, de a modelleket speciálisan fel kell készíteni az ilyen jellegű feladatok végrehajtására annak érdekében, hogy a gép ne bukjon el csúfosan. Elsősorban ilyen kérdésekre fókuszál a DROP (Dua és mtsai, 2019) kérdés-válasz adatbázis.

Egyes erőforrások valamilyen dialógus kontextusában közelítik meg a kérdés-megválaszolás problémáját. A kérdések sokszor többértelműek vagy hiányosak, és kiegészítő információra van szükség a megválaszolásukhoz. Ilyen szituációk modellezését tűzik ki célul az olyan adathalmazok, mint pl. a ShARC (Saeidi és mtsai, 2018).

2.1. Nem angol nyelvű erőforrások

A korábban említett minden kérdés-válasz adatbázis (illetve ezek mellett még számtalan másik) angol nyelvű. Ugyanakkor számos más nyelvre is adaptálták a bemutatott módszereket, illetve többnyelvű kérdés-válasz adatbázisok is születtek.

Viszonylag sok és nagy méretű kínai nyelvű adathalmaz született. A legismertebb a DuReader (He és mtsai, 2018). Ez és számos későbbi hozzá kapcsolódó erőforrás az NQ-hoz és az MS MARCO-hoz hasonlóan a kínai Baidu keresőmotor-kérdéseken és a magyar viszonylatban talán leginkább a gyakorikerdesek.hu-hoz hasonlítható, de annál sokkal kiterjedtebb, és hivatalos entitások, például cégek által szolgáltatott tartalmakat is tartalmazó Baidu Zhidaón alapuló kérdés-válasz adatbázisa.

A SQuAD módszertana alapján készült francia (FQuAD 2.0, (Heinrich és mtsai, 2022), majdnem 80000 kérdés), koreai (KorQuAD 2.0 (Youngmin Kim, 2020) 100000 kérdés), orosz (SberQuAD (Efimov és mtsai, 2020)) és német (GermanQuAD (Möller és mtsai, 2021) kb. 14000 kérdés) erőforrás is. Az XQuAD (Artetxe és mtsai, 2019) a SQuAD 1.1 hangolóhalmazából (dev. set) 240 bekezdéshez kapcsolódó 1190 kérdés-válasz pár professzionális fordítók által készített fordítását tartalmazza 10 nyelven.

Az angol mellett hat másik nyelvet lefedő MLQA benchmark adatbázis (Lewis és mtsai (2020); angolra 12 ezer körüli, a többi nyelvre 5-6 ezer kérdés-válasz pár) kváziekvivalens Wikipédia-mondatok köré épül, amelyekhez a kérdéseket fordítók fordították angolról. A SQuAD-ot számos nyelvre (pl. koreai, hindi, japán, spanyol, cseh, francia, illetve az MLQA adathalmazban szereplő nyelvekre) gépifordították.

11 tipológiai szempontból változatos nyelvet fed le a TyDi QA adathalmaz (Clark és mtsai (2020); összesen 200 ezer kérdés-válasz pár), amely szintén a Wikipédián alapul. A kérdéseket csak a szócikkek bevezető szakasza ismeretében

fogalmazták meg, de bármit lehetett a témával kapcsolatban kérdezni. Így TyDi QA-ben megfogalmazott kérdések nagyobb részére nincs válasz, ahol viszont van, ott a módszer jól garantálja, hogy a kérdés másképp van megfogalmazva, mint a válasz.

2.2. Az első magyar extraktív kérdés-válasz-adatbázis

A SQuAD mintájára készült az Arcanum Adatbázis Kiadó fejlesztésében az AQuAD erőforrás, amelyben elsősorban a SQuAD-ban szereplő szócikkek egy részének magyar megfelelőjéhez írtak kérdéseket, illetve ezt kiegészítették 28 – harmadrészt magyar tematikájú – további magyar Wikipédia szócikkel (összesen 308 szócikket dolgoztak fel). A módszertanból kifolyólag az erőforrásban szereplő szövegek egy része fordítás. A szócikkekből csak a legalább ezer karakter hosszú szakaszokat tartották meg (a szócikket a belső címek osztják szakaszokra). Az eredeti SQuAD-dal ellentétben az egységek itt nem bekezdések, hanem teljes szakaszok, ebből kifolyólag az annotált egységek hosszának eloszlása más, mint a SQuAD-ban (2. ábra). Az annotációt három annotátor végezte. Az AQuAD-ban 8463 megválaszolható kérdés szerepel. Emellett az eredeti adatbázis 8368 nem megválaszolható kérdést is tartalmaz, amelyeknek többségét úgy hozták létre, hogy a megválaszolható kérdésekben valamelyik kifejezést lecserélték egy másikra, amely az esetek többségében egyébként a szövegben nem szerepel. Így a nem megválaszolható kérdések viszonylag könnyen felismerhetőek. Az AQuAD eredetileg az Arcanum belső adatbázisa volt, nem publikálták és publikáció sem készült róla. A jelen munkát magába foglaló projekt keretében azonban a megválaszolható kérdéseket tartalmazó részét a konzorcium megvásárolta, és így ez az erőforrás is szélesebb körben hozzáférhetővé válik a kutatás számára.

2.3. Egyéb magyar nyelvű előzmények

Tágabb értelemben az itt bemutatott kutatás előzményének tekinthető egyrészt a HuLU benchmark (Ligeti-Nagy és mtsai, 2022), amely jelenleg 5 az angol GLUE, illetve SuperGLUE benchmarkok által ihletett vagy azokból fordított magyar nyelvű gépi szövegértési feladatot tartalmaz, másrészt a Novák és mtsai (2019) cikkben bemutatott magyar nyelvű kérdezőrendszer.

3. A MILQA magyar kérdés-válasz adatbázis

Jelen cikkben bemutatott kutatás során egy újabb extraktív kérdés-válasz benchmark adatbázist hoztunk létre magyar nyelvre. Az adatbázis létrehozásakor nagyrészt a SQuAD 2.0 elveit követtük, azonban hasonlóan néhány a 2. szakaszban említett a SQuAD-nál újabb keletű angol nyelvű kérdés-válasz adatbázishoz (Natural Questions, MS MARCO, DROP) bevezettünk számos olyan a SQuAD-hoz képest új kérdés-válasz típust, amely nehezebb, de realisztikusabb feladatokat tartalmaz.

A SQuAD 2.0 nyomán az alábbiak jellemzik a korpuszt:

- jó minőségű Wikipedia szócikkek szolgálnak a kérdések kontextusaként,
- tényszerű (nem vélemény jellegű) kérdések szerepelnek benne,
- olyan kérdéseket is tartalmaz, amelyekre nincs válasz az adott szövegben,
- az eredeti szövegben jelöltük az adott kérdésre vonatkozó lehető legrövidebb választ (ha van),
- a kérdések megfogalmazásakor parafrazeáltuk az eredeti szöveget, így lexikális kereséssel a legtöbb esetben nem található meg a válasz,
- a kérdések nem csak az adott szöveg kontextusában értelmezhetőek, hanem önállóan is megállják a helyüket (pl. nem tartalmaznak nem kötött névmásokat).

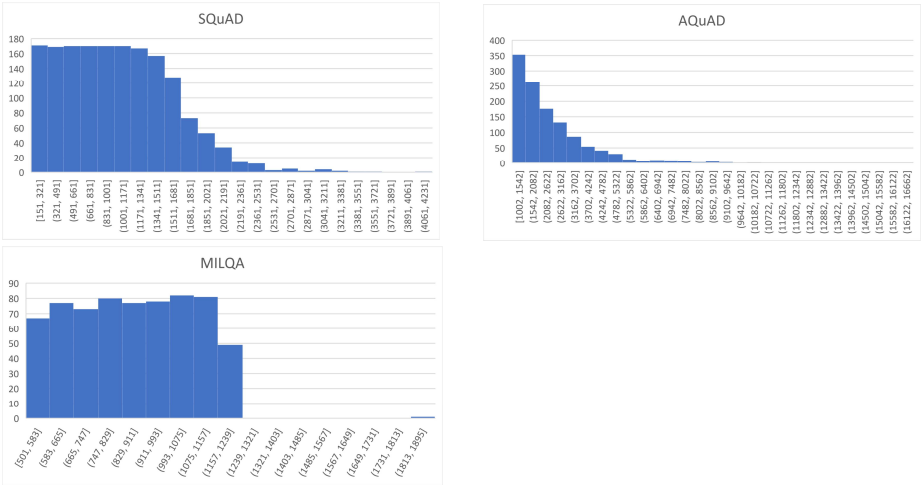
A SQuAD-hoz képest az alábbi újításokat vezettük be. A speciális kérdéstípusok az adatbázisban meg vannak jelölve.

- Egynél több rövid válasz is lehet az adott szövegben az adott kérdésre (a megválaszolt kérdések kb. 8,5%-a).
- A rövid válasz mellett hosszú választ is adtunk, amely a kérdés megválaszolásához lényeges összes körülményt magában foglalja (min. 1 tagmondat, sokszor több mondat).
- Tartalmaz eldöntendő kérdéseket (kb. 9%). Itt a lényeges körülményeket tartalmazó hosszú válasz mellett igen/nem válasz is meg van adva (ha van egyértelmű válasz).
- A „nincs válasz”-kérdések (a kérdések kb. 28,3%-a) az adott témához kapcsolódó releváns kérdések, nem „van válasz”-típusú kérdésekből helyettesítéssel generált kérdések.
- Csak megszámlálás vagy aritmetikai műveletek elvégzése után megválaszolható kérdések is vannak (hasonlóan a DROP adatbázishoz, ezek nehéz feladatot jelentenek a jelenlegi modellek számára). A számítások felsorolt elemek összeszámolására, dátumok, időtartamok és egyéb mennyiségek egyszerű aritmetikai műveletekkel való kiszámítására vonatkoznak.
- A „nincs válasz”-kérdések egy része „beugratós” kérdés, ahol az anyanyelvi beszélők nagy része kiolvasna egy választ a szövegből gyakran téves default feltételezések alapján. Ezeket az eseteket külön megjelöltük, a feltételezhető választ is megadva.
- Ha a szövegben szereplő kifejezés formailag nem felel meg annak, ahogy az adott kérdésre válaszolni kellene (pl. nem megfelelő az esetrag), az annotátorok megadták a válasznak a kérdés kontextusában megfelelő formáját.

3.1. A korpusz létrehozása

Az adatbázis alapját képező adathalmaz létrehozásához a magyar Wikipédiából kiválasztottuk a kiemeltnek, illetve jónak minősített szócikkeket, és a hozzájuk tartozó 2016 eleje és 2021 vége közötti megtekintési statisztika alapján sorba rendeztük őket. Ezek közül az AQuAD-ban nem szereplő szócikkeket vettük be a listába, amelyből az annotátorok választották ki az annotálandó szócikkeket. Minden szócikk első szakaszát, és ezen kívül maximum 10 legalább 500

karakteres véletlenszerűen kiválasztott szakaszt használtunk fel. A SQuAD-hoz hasonlóan bekezdések voltak az egységek, de az 500 karakternél rövidebb bekezdéseket összevontuk, és a kb. 1200 karakternél hosszabbakat kihagytuk (ekkora szövegrészeket lehetett jól áttekinthetően megjeleníteni az annotálófelületen). A szöveghosszak eloszlását l. a 2 ábrán.

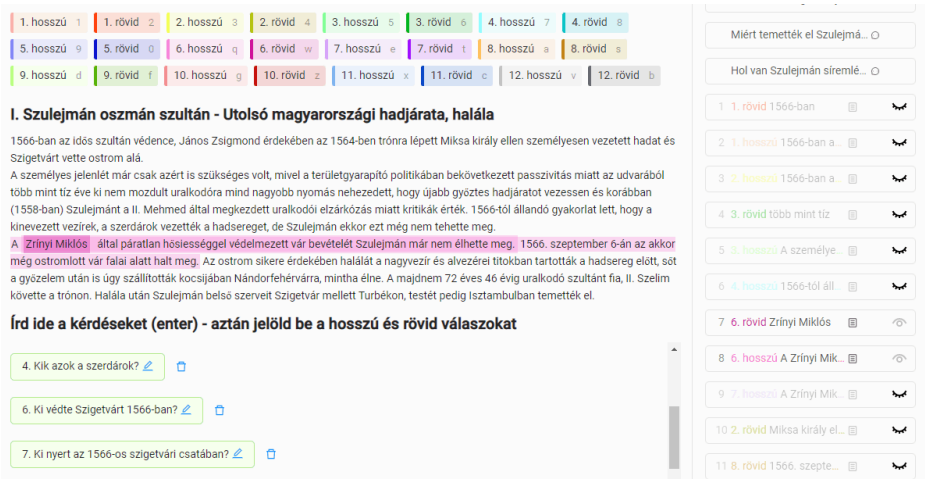


2. ábra. A szöveghosszak eloszlása a SQuAD, AQuAD és MILQA kérdés-válasz korpuszokban.

Az annotálófelületet a Label Studio nyílt forráskódú webes annotálóplatform 1.4 verziójának testreszabásával hoztuk létre. Viszonylag komplex feladatot jelentett a felület alkalmassá tétele arra, hogy mind kérdések feltevését, mind az azokhoz tartozó válaszok bejelölését, mind a speciális kérdés- és választípusok megjelölését áttekinthető módon el lehessen végezni, de sikerült az annotátorok számára viszonylag jól használható felhasználói felületet és workflow-t összeállítani (3. ábra). Az annotálórendszer az előrehaladásról folyamatos statisztikát biztosított az annotátorok számára, és az adott annotált szövegrészből generált kérdés-válasz lista formájában is ellenőrizhették, hogy a válaszok bejelölése valóban a szándékuknak megfelelően sikerült-e. Az annotációt öt annotátor készítette. A később újraszerkesztett problémásabb esetektől eltekintve a Label Studio által tárolt szerkesztési idő alapján jól becsülhető a munka időigénye: kérdésenként nagyjából 85 másodpercet vett igénybe a kérdések megfogalmazása és a válaszok bejelölése.

A korpusz egy 36 szócikkből álló 2391 kérdést (köztük 1751 megválaszolható kérdést) tartalmazó részét (nagyjából a korpusz 10%-át) teszt/hangolókészlet számára leválasztottuk, és ehhez a részhez két független annotáció készült. A kérdésírást nem igénylő annotációs munka gyorsabban haladt: átlagosan 46 má-

sodpercet vett igénybe a hosszú és a rövid válaszok bejelölése. A tesztkorpusz ellenőrzése jelenleg folyamatban van.



3. ábra. A korpusz annotálásához használt webes felület

3.2. A korpusz jellemzői

Az adatbázis összesen 23700 (17000 megválaszolható és 6700 nem megválaszolható) kérdést tartalmaz. 142 Wikipédia-szócikkhez születtek kérdések. Az 1. táblázatban összefoglaltuk a speciális kérdés-, illetve választípusok előfordulását a korpuszban.

9,20% eldöntendő kérdés. Ezeknek 7,65%-a esetében nincs egyértelmű igen/nem válasz, hanem a szövegből az derül ki, hogy az adott kérdéssel kapcsolatban megoszlanak a vélemények, vegyes az eredmény, vagy bizonytalanság van. Ugyanakkor ez nem az egyértelmű *Nincs válasz* esete, amikor a szöveg egyáltalán nem válaszol a kérdésre, hanem a szövegből konkrétan az derül ki, hogy a világ az adott kérdés szempontjából nem fehér-fekete. Az eldöntendő kérdések esetében a korpuszban szereplő annotáció azért lényeges, mert a megjelölt szövegrészekből következik a válasz. Az eldöntendő kérdés, mint kérdéstípus önmagában egyébként nem SQuAD-inkompatibilis: az eredeti SQuAD-ban vannak eldöntendő kérdések, amelyeket úgy fogalmaztak meg, hogy szép extraktív választ lehessen rájuk adni. A MILQA-ban újdonság, hogy explicit jelölés szerepel benne a kérdéstípusra, és a válasz igen/nem voltára. A „nincs válasz” típusú kérdéseknek is mintegy 9%-a eldöntendő kérdés.

Érdekes, hogy miután az annotációs környezet lehetővé tette, hogy a SQuAD-ban szokásos megszorításoktól (hogy csak egy összefüggő extraktív válasz lehet) mentesen dolgozzanak az annotátorok, azoknak a kérdéseknek, amelyekre van

Típus	eset	arány
Van válasz	**16992**	71,67%
. Eldöntendő	1621	9,20%
. . Igen	859	52,99%
. . Nem	638	39,36%
. . Homályos	124	7,65%
. Nem extraktív a válasz	4452	26,20%
. Aritmetika	427	2,51%
. Lista	1455	8,56%
. Nem SQuAD-komp.	5203	30,62%
Nincs válasz	**6716**	28,33%
. Trükkös nincs válasz	629	9,37%
Összesen	**23708**	100

1. táblázat. A kérdés-/választípusok megoszlása a MILQA korpuszban. A beljebb kezdve szedett altípusok esetében az arány oszlop az adott főtípuson belüli arányt adja meg.

válasz a szövegben, több mint 30%-a nem SQuAD-kompatibilisnak adódott (ha az eldöntendő kérdéseket nem soroljuk ide).

A (megválaszolt) kérdések 26,2%-a nem extraktív: az adott kérdésre más lenne a válasz természetes formája, mint ami a szövegben szerepel (pl. más esetben kellene állnia az adott kifejezésnek ahhoz, hogy adekvát válasz legyen a kérdésre). 2,51%-uk megválaszolásához számolni kell (a DROP adatbázisban szereplőkhöz hasonló jellegű kérdések; értelemszerűen ezeknél sem lehet szövegből kimásolni a választ, tehát ezek benne vannak az előbbi 26,2%-ban). 8,56% esetében pedig a SQuAD „egyetlen összefüggő szakasz a válasz” feltételezése nem teljesül (ezen belül persze szintén lehetnek még ragozási problémák is).

A „nincs válasz” kérdések 9,37%-a trükkös. Ezeknél valamilyen ökölszabály jellegű feltételezések mentén hajlamosak lehetünk levezetni egy választ (akár még számolást is végezve), aminek eredményéről viszont könnyen bebizonyosodhatna, hogy téves. Például a Normandiai partraszállásról szóló szócikk egy adott bekezdésében abból, hogy a partraszállásban résztvevő flottaegységeknek három parancsnoka volt, arra következtethetnénk, hogy három flottaegység volt (valójában azonban csak kettő, és volt egy főparancsnok).

A 2. táblázatban a leggyakoribb száz kérdőszavas kifejezést és ezek eloszlását mutatjuk be. A leggyakoribb kérdések időpontokra, okokra, mennyiségekre, helyekre és az alanyra kérdeznek rá.

4. Összefoglalás

Cikkünkben egy új magyar nyelvű kérdés-válasz benchmark adatbázist mutattunk be. A MILQA több szempontból túlmutat a SQuAD-ból ismert egy rövid összefüggő extraktív választ tartalmazó kérdésekre szorítkozó adatbázisokon. Igen/nem válaszokkal megjelölt eldöntendő kérdéseket, több szakaszból álló

mikor	9,78%	hová	0,37%	milyen szerepet	0,18%	mikorra	0,10%	milyen néven	0,08%
miért	7,81%	mi történt	0,36%	milyen a ...	0,18%	milyen viszonyban	0,10%	mit csinált	0,08%
mi	7,46%	kit	0,35%	mi az oka	0,17%	melyek voltak	0,10%	mit tett	0,08%
hány	7,17%	miben	0,35%	mennyire	0,17%	mikortól	0,10%	melyik	0,08%
ki	6,92%	hányadik	0,33%	milyen anyagból	0,17%	hogy nevezték	0,09%	melyik csapat	0,08%
hol	5,79%	melyik országban	0,33%	milyen volt	0,17%	hogyan lehet	0,09%	mennyibe	0,08%
mit	3,16%	melyik a ...	0,32%	mennyi idő	0,16%	kiknek	0,09%	hogyan kell	0,07%
kik	1,77%	melyik városban	0,31%	melyek a ...	0,15%	milyen stílusban	0,09%	melyik az ...	0,07%
honnan	0,94%	hányan	0,28%	kitől	0,14%	melyik magyar	0,09%	melyik napon	0,07%
mivel	0,85%	meddig	0,28%	kivel	0,14%	mennyit	0,09%	milyen célból	0,07%
melyik évben	0,76%	mekkora a ...	0,27%	milyen magas	0,14%	milyen módon	0,09%	hogy mondják	0,07%
kinek	0,68%	milyen színű	0,24%	hogy néz ki	0,13%	milyen típusú	0,09%	melyik cég	0,07%
hogy hívják	0,64%	mennyi ideig	0,22%	hogyan halt meg	0,12%	hova	0,08%	melyik város	0,07%
mire	0,62%	melyik volt	0,22%	mennyivel	0,12%	melyik ország	0,08%	melyik évszázadban	0,07%
minek	0,62%	milyen nemzetiségű	0,22%	milyen nyelven	0,11%	mi volt az oka	0,08%	milyen gyakran	0,07%
miből	0,60%	mi történik	0,21%	milyen állatok	0,11%	milyen funkciót	0,08%	melyik folyó	0,06%
mik	0,58%	mekkora volt	0,20%	milyen pozíciót	0,11%	milyen messze	0,08%	mennyi pénzt	0,06%
hogy hívták	0,56%	hányszor	0,19%	milyen stílusú	0,11%	milyen vallású	0,08%	mennyi volt	0,06%
miről	0,56%	milyen hosszú	0,19%	kiket	0,10%	milyen	0,08%	merre	0,06%
mióta	0,38%	mitől	0,19%	kiről	0,10%	milyen alakú	0,08%	milyen az ...	0,06%

2. táblázat. A leggyakoribb száz kérdőszavas kérdéstípus a MILQA korpuszban (az összes kérdés arányában).

listaszerű válaszokat, hosszú válaszokat, számolást igénylő és egyéb olyan kérdéseket is tartalmaz, ahol nem másolható ki egyszerűen a válasz a szövegből. Az adatbázist a tesztkorpusz ellenőrzése után közzétesszük.

Köszönetnyilvánítás

Jelen kutatás a Nemzeti Kutatási Fejlesztési és Innovációs Alap által biztosított támogatással az FK 125217 és a PD 125216 számú projekt keretében az FK 17 és a PD 17 pályázati program valamint az Innovációs és Technológiai Minisztérium és a Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal támogatásával a Mesterséges Intelligencia Nemzeti Laboratórium keretében valósult meg.

Hivatkozások

Artetxe, M., Ruder, S., Yogatama, D.: On the cross-lingual transferability of monolingual representations. CoRR abs/1910.11856 (2019)

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`

Clark, J.H., Choi, E., Collins, M., Garrette, D., Kwiatkowski, T., Nikolaev, V., Palomaki, J.: TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. Transactions of the Association for Computational Linguistics 8, 454–470 (2020), `https://aclanthology.org/2020.tacl-1.30`

Dua, D., Wang, Y., Dasigi, P., Stanovsky, G., Singh, S., Gardner, M.: DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. 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. 2368–2378. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019), https://aclanthology.org/N19-1246

Dzendzik, D., Vogel, C., Foster, J.: Is it dish washer safe? Automatically answering „yes/no" questions using customer reviews. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop. pp. 1–6. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019), https://aclanthology.org/N19-3001

D'Alessandro, D.M., Kreiter, C.D., Peterson, M.W.: An evaluation of information-seeking behaviors of general pediatricians. Pediatrics 113(1), 64–69 (2004)

Efimov, P., Chertok, A., Boytsov, L., Braslavski, P.: SberQuAD – Russian reading comprehension dataset: Description and analysis. In: Arampatzis, A., Kanoulas, E., Tsikrika, T., Vrochidis, S., Joho, H., Lioma, C., Eickhoff, C., Névéol, A., Cappellato, L., Ferro, N. (szerk.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. pp. 3–15. Springer International Publishing, Cham (2020)

Ely, J.W., Osheroff, J.A., Ebell, M.H., Bergus, G.R., Levy, B.T., Chambliss, M.L., Evans, E.R.: Analysis of questions asked by family doctors regarding patient care. Bmj 319(7206), 358–361 (1999)

Ely, J.W., Osheroff, J.A., Ferguson, K.J., Chambliss, M.L., Vinson, D.C., Moore, J.L.: Lifelong self-directed learning using a computer database of clinical questions. Journal of family practice 45(5), 382–390 (1997)

Gupta, M., Kulkarni, N., Chanda, R., Rayasam, A., Lipton, Z.C.: AmazonQA: A review-based question answering task. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19. pp. 4996–5002. International Joint Conferences on Artificial Intelligence Organization (7 2019), https://doi.org/10.24963/ijcai.2019/694

He, W., Liu, K., Liu, J., Lyu, Y., Zhao, S., Xiao, X., Liu, Y., Wang, Y., Wu, H., She, Q., Liu, X., Wu, T., Wang, H.: DuReader: a Chinese machine reading comprehension dataset from real-world applications. In: Proceedings of the Workshop on Machine Reading for Question Answering. pp. 37–46. Association for Computational Linguistics, Melbourne, Australia (Jul 2018), https://aclanthology.org/W18-2605

Heinrich, Q., Viaud, G., Belblidia, W.: FQuAD2.0: French question answering and learning when you don't know. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference. pp. 2205–2214. European Language Resources Association, Marseille, France (Jun 2022), https://aclanthology.org/2022.lrec-1.237

Hermann, K.M., Kociský, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., Blunsom, P.: Teaching machines to read and comprehend. In: Cortes, C.,

Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R. (szerk.) Advances in Neural Information Processing Systems. pp. 1693–1701 (2015), http://papers.nips.cc/paper/5945-teaching-machines-to-read-and-comprehend

Herskovic, J.R., Tanaka, L.Y., Hersh, W., Bernstam, E.V.: A day in the life of PubMed: analysis of a typical day's query log. Journal of the American Medical Informatics Association 14(2), 212–220 (2007)

Jin, Q., Dhingra, B., Liu, Z., Cohen, W., Lu, X.: PubMedQA: A dataset for biomedical research question answering. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 2567–2577. Association for Computational Linguistics, Hong Kong, China (Nov 2019), https://aclanthology.org/D19-1259

Joshi, M., Choi, E., Weld, D., Zettlemoyer, L.: TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1601–1611. Association for Computational Linguistics, Vancouver, Canada (Jul 2017), https://aclanthology.org/P17-1147

Kočiský, T., Schwarz, J., Blunsom, P., Dyer, C., Hermann, K.M., Melis, G., Grefenstette, E.: The NarrativeQA reading comprehension challenge. Transactions of the Association for Computational Linguistics 6, 317–328 (2018), https://aclanthology.org/Q18-1023

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., Toutanova, K., Jones, L., Kelcey, M., Chang, M.W., Dai, A.M., Uszkoreit, J., Le, Q., Petrov, S.: Natural Questions: A benchmark for question answering research. Transactions of the Association for Computational Linguistics 7, 452–466 (2019), https://aclanthology.org/Q19-1026

Lewis, P., Oguz, B., Rinott, R., Riedel, S., Schwenk, H.: MLQA: Evaluating cross-lingual extractive question answering. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 7315–7330. Association for Computational Linguistics, Online (Jul 2020), https://aclanthology.org/2020.acl-main.653

Ligeti-Nagy, N., Ferenczi, G., Héja, E., Jelencsik-Mátyus, K., Laki, L.J., Vadász, N., Yang, Z.G., 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)

Möller, T., Risch, J., Pietsch, M.: GermanQuAD and GermanDPR: Improving non-English question answering and passage retrieval. In: Proceedings of the 3rd Workshop on Machine Reading for Question Answering. pp. 42–50. Association for Computational Linguistics, Punta Cana, Dominican Republic (Nov 2021), https://aclanthology.org/2021.mrqa-1.4

Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., Deng, L.: MS MARCO: A human generated machine reading comprehension dataset. In: Besold, T.R., Bordes, A., d'Avila Garcez, A.S., Wayne, G. (szerk.) Proceedings of the Workshop on Cognitive Computation: Integrating neural

and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016. CEUR Workshop Proceedings, vol. 1773. CEUR-WS.org (2016), http://ceur-ws.org/Vol-1773/CoCoNIPS_2016_paper9.pdf

Novák, A., Laki, L.J., Novák, B., Dömötör, A., Ligeti-Nagy, N., Kalivoda, Á.: Egy magyar nyelvű kérdezőrendszer. In: XV. Magyar Számítógépes Nyelvészeti Konferencia. pp. 83–95. Szegedi Tudományegyetem, Informatikai Intézet, Szeged (2019)

Rajpurkar, P., Jia, R., Liang, P.: Know what you don't know: Unanswerable questions for SQuAD. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 784–789. Association for Computational Linguistics, Melbourne, Australia (Jul 2018), https://aclanthology.org/P18-2124

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

Richardson, M., Burges, C.J., Renshaw, E.: MCTest: A challenge dataset for the open-domain machine comprehension of text. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. pp. 193–203. Association for Computational Linguistics, Seattle, Washington, USA (Oct 2013), https://aclanthology.org/D13-1020

Saeidi, M., Bartolo, M., Lewis, P., Singh, S., Rocktäschel, T., Sheldon, M., Bouchard, G., Riedel, S.: Interpretation of natural language rules in conversational machine reading. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 2087–2097. Association for Computational Linguistics, Brussels, Belgium (Oct-Nov 2018), https://aclanthology.org/D18-1233

Yang, Y., Yih, W.t., Meek, C.: WikiQA: A challenge dataset for open-domain question answering. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. pp. 2013–2018. Association for Computational Linguistics, Lisbon, Portugal (Sep 2015), https://aclanthology.org/D15-1237

Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W., Salakhutdinov, R., Manning, C.D.: HotpotQA: A dataset for diverse, explainable multi-hop question answering. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 2369–2380. Association for Computational Linguistics, Brussels, Belgium (Oct-Nov 2018), https://aclanthology.org/D18-1259

Yoon, W., Jackson, R., Lagerberg, A., Kang, J.: Sequence tagging for biomedical extractive question answering. Bioinformatics 38(15), 3794–3801 (06 2022), https://doi.org/10.1093/bioinformatics/btac397

Youngmin Kim, S.L.L.P.K.: KorQuAD 2.0: Korean QA Dataset for Web Document Machine Comprehension. Journal of KIISE 47, 577–586 (2020), http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE09353166