Szaknyelvi annotációk javításának statisztikai alapú támogatása

Kicsi András[1], Pusztai Péter[1,2], Szabó Endre[3], Vidács László[1,2]

[1]Szegedi Tudományegyetem, Szoftverfejlesztés Tanszék
Szeged, Dugonics tér 13.
[2]MTA-SZTE Mesterséges Intelligencia Kutatócsoport
Szeged, Tisza Lajos körút 103.
[3]Szegedi Tudományegyetem
Szeged, Dugonics tér 13.
{akicsi,pusztaip,lac}@inf.u-szeged.hu, endrebacsi@gmail.com

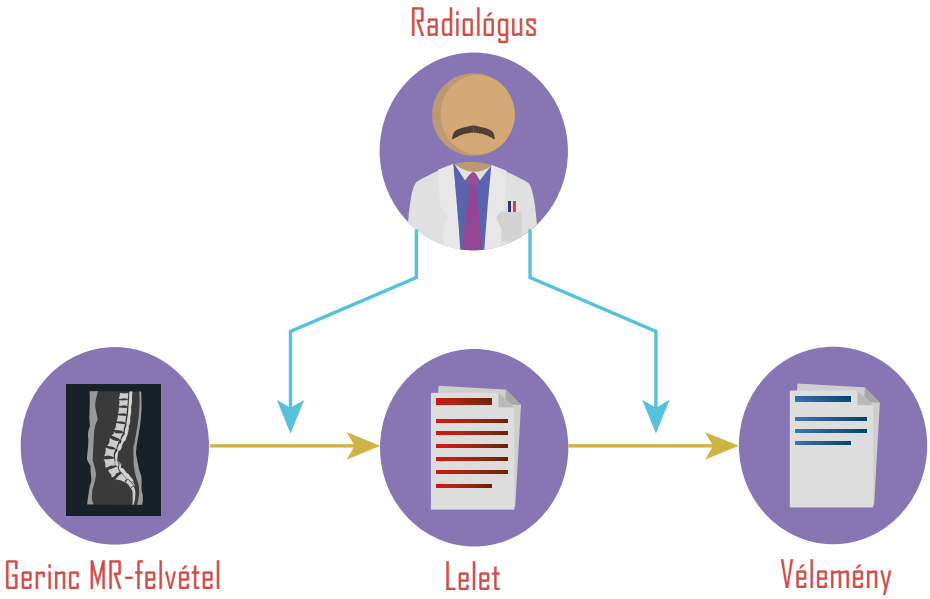
Kivonat A radiológiai leletezés komoly feladat, melynek automatizálása nagy jelentőséggel bír. A leletek gépi értelmezéséhez tanítópéldákra van szükség, amelyeknek megfelelő minőségben kell előállnia. Jelen munkában egy olyan módszert mutatunk be, amellyel az annotáció konzisztenciájának javítása érdekében, az újbóli átnézést statisztikai módszerekkel támogattuk, az inkonzisztenciákra az annotációs rendszer felületén hívva fel a figyelmet. Módszerünk eredményességét valós eredményekkel támasztjuk alá, amelyek nem csak a konzisztenciára, hanem a gépi tanulás sikerére is nagy mértékben kihatnak.
Kulcsszavak: radiológia, információkinyerés, nlp, annotáció

1. Motiváció

A klinikai leletezés az orvoslás jelentős területe, amelynek sikere nagyban hozzájárul a páciensek végső gyógyulásához. Ezen belül a radiológia területén végzett gerinc MR vizsgálatok is igen gyakoriak, csak Magyarországon évente sok ezer ilyen lelet készül. Ezeket általában természetes nyelvű szöveggel, magyar nyelven fogalmazzák meg a radiológusok. A vizsgálaton készített képeket szemlélve leírják orvosi szakértelmüknek megfelelően a látott elváltozásokat, így készülnek el a leletek, és a hozzájuk tartozó, tömörebb vélemények. Ezt a folyamatot láthatjuk az 1. ábrán.

A leletezés természetesen nem könnyű feladat, és folytonos odafigyelést igényel az orvos részéről. Ez azonban könnyíthető különböző automatizáló megoldásokkal, mint például lehetőség a leletek diktálására gépelés helyett. Amennyiben a leleteket automatizált módon értelmezni is tudnánk, rengeteg egyéb lehetőség nyílna meg a munka segítésére. Ezek felhasználhatók lennének mind a minőség biztosításában, mind a leletezés zökkenőmentesebbé és gyorsabbá tételében. Kutatásunkkal ezt tűztük ki célul, melyhez első lépésként a szövegben előforduló entitások detektálását tekintjük. Korábbi munkánkban (Kicsi és mtsai, 2019) már

publikáltuk a területen végzett annotációs módszerünket, illetve kezdeti eredményeinket is. Módszerünkben testrészeket, elváltozásokat és tulajdonságokat különböztettünk meg a leletek szövegében, melyeket automatikusan detektáltunk. Testrésznek tekintettük az emberi test egy pontosan megnevezett elemét, mint például „L.V. discus”, elváltozás minden kóros eltérés, mint például „előboltosulás”, de az aspektusok, például „magassága” és pozitív állapotot jelző szavak, mint például „ép” is ide tartoznak. Tulajdonság minden olyan mértéket vagy minőséget leíró kifejezés, amely elváltozást pontosít, mint például „3 mm-es” vagy „körkörös”. A megfelelő detektáláshoz tanulóadatokra van szükség, ezeket egy radiológus segítségével annotáltattuk, melyhez a Brat (Stenetorp és mtsai, 2012) annotációs szoftvert használtuk fel.



1. ábra: A radiológus munkája a vizsgálat után

Kezdeti detektálási kísérleteink után hamar nyilvánvalóvá vált, hogy a meglévő 487 lelet annotációja jelen minőségben nem elég valóban kiváló eredmények előállításához. Természetesen erre egy lehetséges módszer másik radiológus alkalmazása és a két annotáció összehasonlítása, mely munka azóta szintén megtörtént, ám kezdeti annotációink minőségét is javítani kívántuk, mivel számos inkonzisztenciát tapasztaltunk a jelölésekben. Noha az annotációs útmutatót igyekeztünk pontosan előállítani, mégis felmerült nagy mennyiségű egyéni döntés és dilemmás eset, amelyen a radiológus gyakran önmagával sem tudott konzisztens maradni.

Egy újbóli annotáció természetesen igen nagy feladat még akkor is, ha csak a hibás eseteket kell kijavítani. Arra sincs semmi garancia, hogy ezúttal fenntartható a folyamatos konzisztencia. Ezért automatizált módszerrel igyekeztünk ezt elősegíteni. Cikkünkben az erre kifejlesztett statisztikai módszerünket mutatjuk be, amely Brat rendszer által kimenetként adott .ann fájlok vizsgálata után tokenenként állapít meg konzisztenciát, az eredményeket pedig a Brat formátumának megfelelően rögzíti megjegyzésként. Ezzel felhasználóbarát módon hívja fel a figyelmet a kevéssé konzisztens jelölésekre, amelyek tudatában a radiológus ezután teljes mértékben saját elbírálása szerint járhat el.

Korábbi cikkünkben említettük, hogy testrészek, elváltozások és tulajdonságok mellett helyeket is jelöltünk, ezek a munka jelenlegi fázisában azonban komplex szerkezetűek, így velük itt nem foglalkozunk.

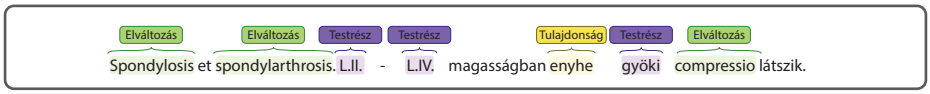
2. Folyamat

Munkánk jelenleg magyar nyelvű gerincleletek feldolgozását öleli fel. Cikkünkben a helyes klasszifikáció problémájával foglalkozunk, amelyben a leletek szövegének kifejezéseit három címkével igyekszünk ellátni jelentésüknek megfelelően, testrészeket, elváltozásokat és tulajdonságokat különböztetünk meg. A problémát gépi tanulási módszerekkel közelítettük meg. Mindkét módszer nagy mennyiségű tanulóadattal tud csak megfelelően működni, ezért ezt biztosítani kell. Erre a célra radiológus által annotált valós leleteket használtunk, 487 lelet annotációja készült el. Ehhez radiológusunk a Brat (Stenetorp és mtsai, 2012) annotációs rendszert használta, amelyet megfelelően konfiguráltunk a kívánt entitások jelölésére, így áttekinthető és felhasználóbarát környezetben végezhette a jelölést.

Ezen az annotált leletmennyiségen igyekeztünk javítani egy statisztikával támogatott újabb kézi elbírálással. Az annotációs útmutató számos esetet lefed, ám ezeket többszáz lelet átolvasása után már nem mindig idézi fel az annotátor helyesen. Vannak továbbá olyan különleges esetek, amelyek egyszerűen nem illenek semelyik, az útmutató által érintett problémakörbe. Ez utóbbiakat jobb esetben megbeszélés alapján kell kezelni, ám sok olyan eset adódik, amikor az annotátor eléggé biztosnak tart egy bizonyos helyzetet, és önállóan jelöli. Ilyenkor a legfontosabb, hogy önmagával konzisztens legyen a felmerülő hasonló dilemmás kérdéseket mindig egy irányelv mentén jelölje.

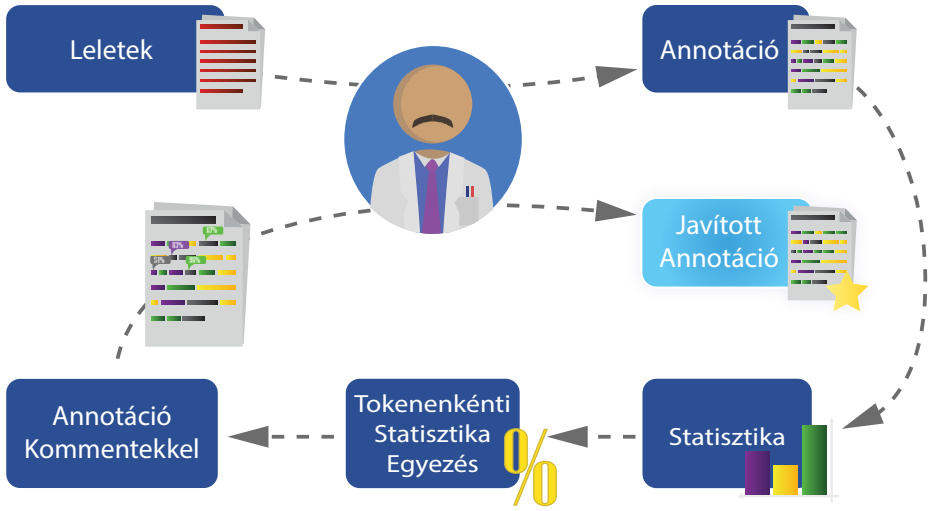
Az annotátor önmagától való inkonzisztenciája egyszerű statisztikai módszerekkel könnyen mérhető, megtekinthetjük, hogy egy adott kifejezést általában ugyanazon címkével látta-e el. Ez persze rengeteg esetben indokolt kilengés, mint például munkánkban a „jobb” szó esetében, ahol ez lehet tulajdonság része, egy testrész helyének pontosítása, vagy akár annak leírása, hogy egyik csigolya a másiknál jobb állapotban van, amely elváltozás lenne. Tehát az emberi elbírálás semmiképpen sem nélkülözhető.

A statisztikák azonban segíthetnek a kézi ellenőrzésben, nagyban felgyorsíthatják azt, és felhívhatják figyelmet olyan dilemmás esetekre, amikre az emberi személő esetleg nem is figyelt volna fel. Tekintsük a 2. ábrán látható példát. Itt különböző dilemmás esetek merülnek fel. Először is a „spondylosis et spondyl-



2. ábra: Egy több dilemmás esetet tartalmazó példa

arthrosis" szövegrész nagyon sokszor ugyanígy fordul elő a leletekben, ugyanis leggyakrabban a két elváltozás együtt jelentkezik. Ez kísértést jelent az annotátor számára, hogy egyben jelölje őket. Az „et" szó továbbá, mivel latinul van, kevésbé intuitívan tagol elváltozásokat, mint például az „és" szó. Másrészt láthatjuk, hogy ahogy a leletek többségében, itt is vannak intervallummal megadott testrészek. Ilyenkor megegyezés és az annotációs útmutató szerint ezeket külön-külön be kell jelölni. Efölött azonban könnyű átsiklani, hiszen tömörek, egyértelműen testrészt jelölnek, és szinte teljesen egyben vannak, még szóköz sincs közöttük az eredeti szövegben. Ezért rengeteg hasonló típusú hiba volt a kezdeti annotációban, amely a testrészek inkonzisztens detektálását idézte elő egyes esetekben. A „gyöki compressio" kifejezést is gyakran egyben jelölte a radiológus, hiszen úgy tűnik, hogy ez az elváltozás teljes megnevezése. A „gyök" azonban egy testrész és más megfogalmazásban így is jelölné a radiológus is, már igen apró változtatás után is, például „a kilépő gyökök compressioja látható" formában. Az ilyen típusú hibák szintén nagyon gyakoriak.

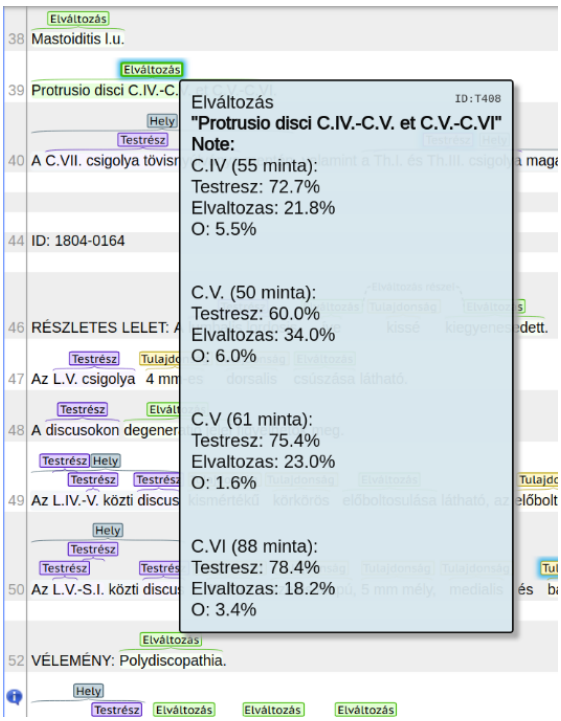


3. ábra: A javasolt javítási módszer áttekintése

Természetesen ezek nagy odafigyeléssel kijavíthatók egy újbóli átnézés során, de ennél sokkal jobb módszer lehet, ha megpróbálunk ezekre automatikusan rávilágítani. Az erre kidolgozott módszerünk látható a 3. ábrán. A 487 lelet összes szavát tokenek szintjén listába gyűjtöttük. Ezután minden egyes tokenhez meghatároztuk, hogy hány esetben voltak testrészként, elváltozásként és tulajdonságként jelölve, tehát előállítottuk a szükséges statisztikát. Ennek a listának a birtokában az összes leleten végigiterálva és tokenizálva minden egyes token előforduláshoz meghatároztuk, hogy a többségi címkéjükben vannak-e, három címke esetén is a legnagyobbhoz viszonyítottunk. Egyenlőség esetén mindkét címkét többséginek vettük. Amennyiben egy token nem a többségi címkéje részeként szerepelt egy jelölésben, akkor ezt a Brat rendszer által kimenetként adott .ann fájlban jelöltük. Az annotációs rendszer megengedi megjegyzések beszúrását egyes címkézett elemekhez. Ilyenkor praktikus módon a címke fejléce ragyogó körvonalat kap, szembetűnően felhívva magára a figyelmet. A fájlba tehát a Brat rendszer formátumával teljesen megegyező új sorokat szúrtunk be automatizáltan, amelyben leírtuk, hogy a token nem a többségi címkéjében fordult elő. Egy jelölt kifejezéshez több ilyen megjegyzés is tartozhat, hiszen sokszor több token szerepel egymás mellett, ilyen esetekben az összes megjegyzés egymás alá kerül. Miután az algoritmus az összes leletet átnézte, és előálltak a módosított .ann fájlok, a Brat rendszerrel megnyitva a leleteket már láthatjuk a kék színnel ragyogó címkéket, amelyek remekül felhívják a figyelmet a dilemmás helyzetekre. Erre látható példa a 4. ábrán, amely egy képernyőkép módszerünk kimenetéről. További előnyt jelent, hogy a radiológusnak sem kell új rendszerrel megismerkedni, a már megszokott környezetben végezheti a leletek átvizsgálását.

Az ábrán látható példán az annotátor a diagnózis egy teljes mondatát egy elváltozásként jelölte, a rendszer pedig felhívja a figyelmet arra, hogy a csigolyák megnevezése általában testrész szokott lenni. A számok nyomán egyébként gyanítható lenne, hogy máskor is csinált már ilyet. Az "O" címke azt jelöli, hogy nem volt jelölve egyik címkével sem. A 2. ábrán látható példában a megnevezett rossz jelölések esetén módszerünk ugyanígy szólna például, hogy az „et” szó és a kötőjel általában nem szokott jelölésre kerülni ha külön tokenként fordul elő, a „gyöki” szó pedig általában testrész.

Biztosítottunk továbbá egy megértést segítő modult is a statisztikák mellett. Ez egy egyszerű programkód, amely szöveget vár bemenetként, és az összes leletet végigpásztázva megadja, hogy milyen címkével, hol, és milyen szövegkörnyezetben volt jelölve az adott kifejezés. Ez azokra az esetekre alkalmazható, ha esetleg nem értjük, hogy az adott szó mi alapján került egy kisebbségi címkébe, hiszen jelenleg jó jelölést látunk rá. Ezután akár a többi ilyen eset, vagy esetleg hibásan kiosztott többségi címke is könnyen megkereshető és javítható. A statisztikát kiszámító és kommenteket beszúró programkód is a radiológus rendelkezésére állt, amelyet bármikor újrafuttathatott, hiszen ezek nem aktualizálják magukat automatikusan.



4. ábra: Képernyőkép kimenetünk Brat-ban való megjelenítéséről

3. Kísérletek

Kísérleteink célja olyan eszköz fejlesztése volt, mellyel segíteni tudjuk a radiológust a tekintetben, hogy az általa készített annotáció egyrészt önmagával, másrészt az annotálási útmutatóval is minél inkább konzisztens maradjon. Első lépésben 487 gerinc MR leletet annotáltattunk a radiológus kollégával az előre meghatározott útmutató szerint. Már a folyamat közben is kimutatható volt, hogy az annotáció nem teljesen konzisztens, amit a radiológus kolléga is alátámasztott, azzal a megfigyelésével, hogy sokszor nem hogy az útmutatóval, de még önmagával sem tudja tartani a konzisztenciát egy hosszabb annotálás során. Az annotációban tapasztalható inkozisztenciák felméréséhez minden egyedi tokenhez statisztikát készítettünk, melyben kimutattuk a különböző címkék tokenhez rendelésének százalékos eloszlását. A 487 leletben 2760 egyedi tokent találtunk, melyből 2082 tokenhez kizárólag egyféle címke lett rendelve. A maradék 678 tokent a radiológus minimum kétféleképpen annotálta. Az inkonzisztencia sok esetben adódott a kifejezés különböző szövegkörnyezetben történő előfordulásaiból, aminek következtében a radiológus eltért az útmutatótól és saját legjobb belátása szerint annotált. Ezekhez az annotációs esetekhez, mivel statisztikailag gyakran kisebbségben voltak, egy figyelmeztető megjegyzést rendeltünk, melyet az annotáló szoftverben jelenítünk meg. Ezek alapján a radiológus belátása sze-

rint döntött az eset javítása, vagy változatlanul hagyása mellett, természetesen az annotációs útmutató által lefektetett alapelvekkel továbbra is összhangban maradva.

Az általunk fejlesztett segédeszközökkel felvértezett radiológust ezután egy javítóannotációra kértük. A visszakapott leletekben a 2760 egyedi tokenből ezután már 2276 token rendelkezet kizárólag egyféle annotációval a többi minimum kétféle címkét kapott. Már ez a szám is mutatja, hogy a korábbi annotációhoz képest konzisztensebb eredményt kaptunk a javítást követően. A többféle címkével annotált tokenek pontos eloszlását az első és második körös annotáció után az 1. táblázat szemlélteti. A táblázatban is jól látható az annotációk javításának eredményessége. Javítás után a három- és négyféle címkét kapott tokenek mennyisége a felére, míg a kétféle címkét kapott tokenek mennyisége az eredeti ötödével csökkent.

1. táblázat. A többféle annotációt kapott tokenek eloszlása annotációjavítás előtt és után.

Annotáció	Eredeti	Javított
Egyféle	2080	2276
Kétféle	493	392
Háromféle	156	77
Négyféle	31	15
Összesen	2760	2760

Következő lépésben intra-annotátor egyezést mértünk a radiológusunk eredeti és javított annotációja között. Az egyezés minőségének megítéléséhez a Cohen kappa mutatót, valamint mikroátlag F1-mérték metrikákat alkalmaztunk, ahol a referenciának a javított annotációt vettük. Cohen kappára 0,9278-as értéket, míg F1-mértékre 0,9350-es értéket kaptunk. A szakirodalom szerint a 0,8-as érték feletti Cohen kappa jó egyezésre utal, valamint a magas F1-mérték is azt sugallja, hogy az eredeti és javított annotáció egymással konzisztensnek számít.

A fentiek alapján azt gondolhatnánk, hogy az annotációjavításnak nem sok jelentősége volt, azonban érdekesebb eredményeket kapunk, ha megvizsgáljuk az eredeti és javított annotációval tanított modellek teszthalmazon mutatott teljesítményét. Demonstrációs céllal, a kísérleteinkben referenciaként használt, IOB címkéket nem tartalmazó osztálycímkékkel tanított Bi-LSTM (Hochreiter és Schmidhuber, 1997) eredményeit mutatjuk be a 2 és 3. táblázatokon.

A tanítási eredmények jól mutatják, hogy az annotációk javítása jelentős mértékben javított a modellünk teljesítményén. A testrész és tulajdonság esetében több, mint 3%-os, míg az elváltozás tekintetében megközelítőleg 2%-os javulást értünk el az F1-mértéket tekintve. Érdemes kiemelni, hogy a testrészek felismerési pontossága majdnem 5%-kal nőtt, ez is tükrözi, mennyire jellemző volt az annotációban a testrészek fent bemutatott tipikus inkonzisztens jelölése.

2. táblázat. Az eredeti annotáción tanított Bi-LSTM modell teljesítménye a három fő névelemtípus felismerésében.

	Pontosság	Fedés	F1-mérték
Elváltozás	0,9143	0,9285	0,9213
Testrész	0,9112	0,9519	0,9311
Tulajdonság	0,8856	0,8610	0,8731
Mikroátlag	0,9083	0,9253	0,9167

3. táblázat. A javított annotáción tanított Bi-LSTM modell teljesítménye a három fő névelemtípus felismerésében.

	Pontosság	Fedés	F1-mérték
Elváltozás	0,9380	0,9432	0,9406
Testrész	0,9598	0,9698	0,9648
Tulajdonság	0,9108	0,9020	0,9064
Mikroátlag	0,9420	0,9464	0,9442

Kísérleteink jól szemléltetik, hogy az elsőre kiemelkedően jónak tűnő intra-annotátor egyezés megtévesztő lehet, a számok mögé tekintve, a modelleket a tényleges adatokon tesztelve láthatjuk, hogy az annotáció konzisztenciájának javítása jelentős javulást eredményezhet a modellek működését illetően.

4. Kapcsolódó kutatások

Ugyan próbálkozások történtek már a strukturált leletezés egészségügyi szektorba történő bevezetésére, a gyakorlat napjainkig azt mutatja, hogy a szakorvosok és radiológusok is előnyben részesítik a szabad megfogalmazású leletek készítését a strukturált leletezéssel szemben. Ez egyfelől lehetőséget ad a természetes nyelv komplexitásának kihasználására és a leletek szabatos megfogalmazásra, másfelől megnehezíti a leletekből történő információkinyerést, szövegértelmezést, illetve a leletezési folyamat minőségbiztosítását. Éppen ennek a kihívásnak köszönhetően a terület kiváló kutatási lehetőséget biztosít a számítógépes nyelvészet számára, mely során újfajta természetesnyelv feldolgozási módszerek, illetve az egészségügyi szakembereket segítő alkalmazások egyaránt napvilágot láthatnak.

Az eddig fejlesztett alkalmazások köre a kinyert információ típusától függően széles spektrumon változik. Többek között beszélhetünk diagnózissegéd (Pham és mtsai, 2014; Rink és mtsai, 2013; Solti és mtsai, 2009), diagnosztikai minőségbiztosítást (Raja és mtsai, 2012; Ip és mtsai, 2011; Sistrom és mtsai, 2009; Dang és mtsai, 2008), a leletek automatikus BNO kódolását végző (Farkas és Szarvas, 2007), a nem várt elváltozásokra adott válaszlépéseket (Dutta és mtsai, 2013), vagy a további vizsgálatokra vonatkozó ajánlásokat figyelő (Yetisgen-Yildiz és mtsai, 2011), illetve a páciens egészségi állapotát nyomon követő al-

kalmazásokról (Cheng és mtsai, 2010). A közelmúltban több olyan összefoglaló cikk is megjelent, mely jól bemutatja az elmúlt egy évtizedben történt fontosabb előrelépéseket (Wang és mtsai, 2018; Pons és mtsai, 2016; Ford és mtsai, 2016; Cai és mtsai, 2016; Yim és mtsai, 2016; Meystre és mtsai, 2008).

A leletekből történő információkinyerés első lépése továbbra is a szöveg, előre meghatározott útmutató alapján, szakember által végzett, pontos annotálása. Az annotálást minden esetben minimum két annotátor egymástól függetlenül végzi. A nem egyértelmű esetek eldöntése kettőnél több annotátor esetében többségi szavazással történik, míg két annotátor esetében vagy megegyezéses alapon, vagy egy harmadik, szenior kolléga döntése alapján oldják fel az ellentétet. Az egyezés mérésére, az annotátorok számának függvényében többféle metrikát is alkalmaznak, azonban az egyik legelterjedtebb ilyen mérőszám a Cohen kappa (Artstein és Poesio, 2008). A mérőszám interpretálása a szakirodalomban vita tárgyát képezi. Általánosságban elmondhatjuk, hogy 0,8-as érték felett az egyezés megalapozottnak, az annotáció minősége pedig jónak mondható, ennek hitelességét azonban egyes kutatók megkérdőjelezik (Klebanov és Beigman, 2009). Szerintük ugyanis a magas kappa érték főleg két annotátor esetében nem feltétlen jelent jó minőségű annotációt, csakúgy, mint ahogy az alacsony kappa érték, öt annotátor esetében nem feltétlen jelent rossz minőségű annotációt.

Egy modell maximum annyira lehet jó, mint amennyire jó az adat, amin tanították. Az annotáció minőségének javítása ezért komoly kihívás a számítógépes nyelvészek számára. Ennek elérése érdekében több megközelítést is alkalmaznak. Az egyik legkézenfekvőbb módszer az adat előannotálása, majd az automatikusan létrehozott annotációk szakemberrel történő hitelesítése, illetve javítása. Az előannotáció történhet egy már meglévő adatbázis, vagy az annotátor korábbi annotációja alaján (pl. az annotátor annotálja az adatok felét, majd ezek alapján megtörténik az adatok másik felének automatikus annotációja, amit az annotátor jóváhagy, vagy javít (Ganchev és mtsai, 2007)). Ilyen támogatás több annotáló szoftverben is megtalálható. Egy másik lehetőség az annotációk minőségének javítására, ha annotáció közben ajánlásokat teszünk az annotátornak. Ez annyiból kifinomoltabb, mint az előannotálás, hogy ebben az esetben a korábban többféleképpen annotált esetekre egy megbízhatósági értéket is biztosítunk. Vagyis minden egyes szóra az ajánlást az adott szóhoz korábban hozzárendelt osztálycímkék háttérben kiértékelt statisztikája alapján hozzuk létre. Az ajánlás tehát több osztálycímkét is tartalmaz egy százalékos megbízhatósági érték kíséretében (Oliveira és mtsai, 2017; Morton és LaCivita, 2003). Az MIT fejlesztése a Story Workbench szoftver, mely automatikus annotálási funkcióval is el van látva. Ez annyiban különbözik az előannotálástól, hogy itt az annotációk az annotálás során, a módosításokat figyelembe véve, valós időben keletkeznek (Finlayson, 2011). A WebAnno egy másik félautomata annotációs eszköz, melyben az annotációs javaslatot egy külön ablakban jelenítik meg, az éles szövegen csak a már elfogadott javaslatok, illetve az annotátor által kézzel készített annotációk láthatóak. Ez a konstrukció a szerzők szerint arra ösztönzi az annotátort, hogy minden egyes javaslatot jóváhagyjon, mielőtt az az éles szövegbe kerülne. A program egyébként többrétegű ajánlási rendszert alkalmaz, melynek egyik rétege egy

adott szó korábbi annotációinak későbbi esetekhez rendelése egyszerű szöveges egyezés alapján (Muhie és mtsai, 2014). A GoNTogle egy szemantikus annotációt ellátó eszköz, mely teljes dokumentumok vagy dokumentum részek automatikus annotálására is képes. Az automatikus annotációhoz egy súlyozott kNN osztályozót használ, mely a szöveges információt és az annotátor korábbi annotációit egyaránt felhasználja az annotálási javaslatok kialakításához (Bikakis és mtsai, 2010). Az eddigiektől eltérően a Widlöcher és munkatársai (Widlöcher és Mathet, 2012) által fejlesztett Glozz eszköz nem automatikus annotálás segítségével, hanem a meglévő annotációk folyamatos monitorozási lehetőségével támogatja a konzisztens, jó minőségű annotációk készítését. Ehhez a fejlesztők egy GlozzQL-re keresztelt lekérdező nyelvet is készítettek.

A magyar nyelvű számítógépes nyelvészeti szakma, követve a nemzetközi gyakorlatot elsősorban annotátorok közötti egyezésmérést alkalmaz az annotáció minőségének ellenőrzésére. Ugyan a magyar szakirodalomban is találhatunk példát annotációk minőségének javítását célzó tanulmányokra (Novák, 2016), vagy már meglévő annotációk automatikus javítására (Kalivoda, 2017), a fentiekben bemutatott javaslattevő és annotációkat monitorozó alkalmazások használata tudomásunk szerint nem bevett gyakorlat.

5. Összegzés

Munkánk során bemutattuk, hogy az annotációk minősége jelentős mértékben befolyásolja a gépi tanuló algoritmusok teljesítményét, valamint javaslatot tettünk egy általunk fejlesztett annotációk konzisztenciájának fenntartását szolgáló eszköz alkalmazására. Kísérleteinkben egyetlen radiológus javítás előtti és utáni annotációja között mértünk intra-annotátor egyezést. A magas Cohen kappa és F1-mérték értékek arra utaltak, két annotáció jó egyezést mutat, azonban a modellünket a javítás előtt és utáni adatokon tanítva szembetűnő különbségeket tapasztaltunk. Az annotáció konzisztensebbé tételével 2-3%-os F1-mértékben tapasztalható javulást sikerült elérnünk az egyes névelemek esetén. Kísérleteink jó alapot szolgáltatnak egy későbbi, összetettebb rendszer fejlesztéséhez.

Köszönetnyilvánítás

Jelen kutatás az Innovációs és Technológiai Minisztérium ÚNKP-19-3 kódszámú Új Nemzeti Kiválóság Programjának támogatásával készült. A kutatást részben az Emberi Erőforrások Minisztériuma támogatta (TUDFO/47138-1/2019-ITM).

Hivatkozások

Artstein, R., Poesio, M.: Inter-coder agreement for computational linguistics. Computational Linguistics 34, 555–596 (12 2008)

Bikakis, N., Giannopoulos, G., Dalamagas, T., Sellis, T.: Integrating keywords and semantics on document annotation and search. pp. 921–938 (01 2010)

Cai, T., Giannopoulos, A.A., Yu, S., Kelil, T., Ripley, B., Kumamaru, K.K., Rybicki, F.J., Mitsouras, D.: Natural Language Processing Technologies in Radiology Research and Clinical Applications. RadioGraphics 36(1), 176–191 (jan 2016)

Cheng, L.T.E., Zheng, J., Savova, G.K., Erickson, B.J.: Discerning Tumor Status from Unstructured MRI Reports-Completeness of Information in Existing Reports and Utility of Automated Natural Language Processing. Journal of Digital Imaging 23(2), 119–132 (apr 2010)

Dang, P.A., Kalra, M.K., Blake, M.A., Schultz, T.J., Stout, M., Lemay, P.R., Freshman, D.J., Halpern, E.F., Dreyer, K.J.: Natural Language Processing Using Online Analytic Processing for Assessing Recommendations in Radiology Reports. Journal of the American College of Radiology 5(3), 197–204 (mar 2008)

Dutta, S., Long, W.J., Brown, D.F., Reisner, A.T.: Automated Detection Using Natural Language Processing of Radiologists Recommendations for Additional Imaging of Incidental Findings. Annals of Emergency Medicine 62(2), 162–169 (aug 2013)

Farkas, R., Szarvas, Gy.: Eljárás radiológiai leletek automatikus BNO kódolására. In: V. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2007). p. 149–157. Szegedi Tudományegyetem Informatikai Tanszékcsoport, Szegedi Tudományegyetem Informatikai Tanszékcsoport, Szeged (2007)

Finlayson, M.A.: The story workbench: An extensible semi-automatic text annotation tool. In: Proceedings of the 4th Workshop on Intelligent Narrative Technologies. pp. 21–24 (2011)

Ford, E., Carroll, J.A., Smith, H.E., Scott, D., Cassell, J.A.: Extracting Information from the Text of Electronic Medical Records to Improve Case Detection: A Systematic Review. Journal of the American Medical Informatics Association 23(5), 1007–1015 (2016)

Ganchev, K., Pereira, F., Mandel, M., Carroll, S., White, P.: Semi-automated named entity annotation pp. 53–56 (06 2007)

Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. 9(8), 1735–1780 (Nov 1997)

Ip, I.K., Mortele, K.J., Prevedello, L.M., Khorasani, R.: Focal Cystic Pancreatic Lesions: Assessing Variation in Radiologists' Management Recommendations. Radiology 259(1), 136–41 (apr 2011)

Kalivoda, Á.: Az igekötök gépi annotálásának problémái. In: Doktoranduszok tanulmányai az alkalmazott nyelvészet köréből. pp. 100–108 (2017)

Kicsi, A., Pusztai, P., Szabó Ledenyi, K., Szabó, E., Berend, G., Vincze, V., Vidács, L.: Információkinyerés magyar nyelvű gerinc mr leletekből. In: XV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2019). p. 177–186. Szeged (2019)

Klebanov, B., Beigman, E.: From annotator agreement to noise models. Computational Linguistics 35, 495–503 (12 2009)

Meystre, S.M., Savova, G.K., Kipper-Schuler, K.C., Hurdle, J.F.: Extracting Information from Textual Documents in the Electronic Health Record: A Review of Recent Research. Yearbook of Medical Informatics pp. 44–128 (2008)

Morton, T., LaCivita, J.: Wordfreak: An open tool for linguistic annotation. (01 2003)

Muhie, S., Biemann, C., Eckart de Castilho, R., Gurevych, I.: Automatic annotation suggestions and custom annotation layers in webanno. pp. 91–96 (01 2014)

Novák, A.: Improving corpus annotation quality using word embedding models. Polibits 53, 49–53 (2016)

Oliveira, L., Gebeluca, C., Silva, A., Moro, C., Hasan, S., Farri, D.: A statistics and umls-based tool for assisted semantic annotation of brazilian clinical documents. pp. 1072–1078 (11 2017)

Pham, A.D., Névéol, A., Lavergne, T., Yasunaga, D., Clément, O., Meyer, G., Morello, R., Burgun, A.: Natural Language Processing of Radiology Reports for the Detection of Thromboembolic Diseases and Clinically Relevant Incidental Findings. BMC Bioinformatics 15(1), 266 (aug 2014)

Pons, E., Braun, L.M., Hunink, M.G., Kors, J.A.: Natural Language Processing in Radiology: A Systematic Review. Radiology 279(2), 329–343 (may 2016)

Raja, A.S., Ip, I.K., Prevedello, L.M., Sodickson, A.D., Farkas, C., Zane, R.D., Hanson, R., Goldhaber, S.Z., Gill, R.R., Khorasani, R.: Effect of Computerized Clinical Decision Support on the Use and Yield of CT Pulmonary Angiography in the Emergency Department. Radiology 262(2), 468–474 (feb 2012)

Rink, B., Roberts, K., Harabagiu, S., Scheuermann, R.H., Toomay, S., Browning, T., Bosler, T., Peshock, R.: Extracting Actionable Findings of Appendicitis from Radiology Reports Using Natural Language Processing. AMIA Joint Summits on Translational Science Proceedings. AMIA Joint Summits on Translational Science p. 221 (2013)

Sistrom, C.L., Dreyer, K.J., Dang, P.P., Weilburg, J.B., Boland, G.W., Rosenthal, D.I., Thrall, J.H.: Recommendations for Additional Imaging in Radiology Reports: Multifactorial Analysis of 5.9 Million Examinations. Radiology 253(2), 453–61 (nov 2009)

Solti, I., Cooke, C.R., Xia, F., Wurfel, M.M.: Automated Classification of Radiology Reports for Acute Lung Injury: Comparison of Keyword and Machine Learning Based Natural Language Processing Approaches. In: Proceedings - 2009 IEEE International Conference on Bioinformatics and Biomedicine Workshops, BIBMW 2009. vol. 2009, pp. 314–319. NIH Public Access (nov 2009)

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., Tsujii, J.: brat: A Web-based Tool for NLP-Assisted Text Annotation. In: Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics. pp. 102–107. Association for Computational Linguistics, Avignon, France (April 2012)

Wang, Y., Wang, L., Rastegar-Mojarad, M., Moon, S., Shen, F., Afzal, N., Liu, S., Zeng, Y., Mehrabi, S., Sohn, S., Liu, H.: Clinical Information Extraction Applications: A Literature Review (jan 2018)

Widlöcher, A., Mathet, Y.: The glozz platform: a corpus annotation and mining tool. DocEng 2012 - Proceedings of the 2012 ACM Symposium on Document Engineering (09 2012)

Yetisgen-Yildiz, M., Gunn, M.L., Xia, F., Payne, T.H.: Automatic Identification of Critical Follow-Up Recommendation Sentences in Radiology Reports. AMIA Symposium pp. 1593–602 (2011)

Yim, W.w., Yetisgen, M., Harris, W.P., Kwan, S.W.: Natural Language Processing in Oncology. JAMA Oncology 2(6), 797 (jun 2016)