Entitások azonosítása és szemantikai kapcsolatok feltárása radiológiai leletekben

Kicsi András[1], Szabó Ledenyi Klaudia[1], Pusztai Péter[1,2], Németh Péter[1], Vidács László[1,2]

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

Kivonat Cikkünkben magyar nyelvű radiológiai leletek szövegében automatizáltan azonosítjuk az előforduló testrészeket és elváltozásokat, valamint megállapítjuk a szöveg testrészeinek, elváltozásainak és tulajdonságainak kapcsolatát. Ismertetjük módszereinket, amelyekkel felállítottunk egy megfelelő azonosítóhalmazt, valamint elvégeztük ezek különböző szóalakokhoz való rendelését. Az egyszerű kapcsolatokon kívül az intervallummal vagy utalással megadott speciális eseteket és a tagadásokat is figyelembe vesszük. 487 valós leleten mért eredményeinket közöljük.
Kulcsszavak: radiológia, információkinyerés, NLP, azonosítás, konstituens

1. Motiváció

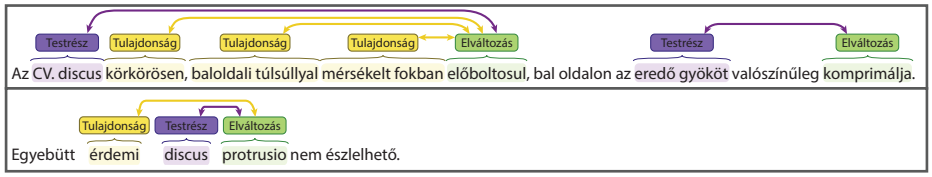
A radiológiai vizsgálatok után az eredmények kezdetben képi formában állnak rendelkezésre. Ezeknek az adatoknak a feldolgozását a mai orvoslásban radiológus végzi orvosi szaktudására támaszkodva. Ez elengedhetetlen a megfelelő értelmezéshez. A radiológus a megtekintett képi információt áttekinti, és szöveges formában rögzíti, általában saját anyanyelvén. Nem csak a képen látott információt írja le, hanem véleményt is alkot, mely a leletben megfogalmazott jelentősebb tények alapján megállapításokat tartalmaz a rögzített elváltozásokról. A leletet és véleményt a radiológus ezután a vizsgálatot kérő szakorvosnak továbbítja, aki ezek figyelembe vételével meghozza a páciens jövőbeli kezelésére irányuló döntéseket. A leletek archiválásra kerülnek, későbbi vizsgálatoknál a radiológus számára elérhetőek, ezzel levonhatóvá válik a következtetés egy korábbi kezelés sikerességéről is, mely szintén kritikus lehet a további lépésekhez.

A leletek tehát rengeteg információt hordoznak, és a radiológus munkájának és szaktudásának gyümölcsét jelképezik. Gépi értelmezésük ezért számos felhasználási lehetőséggel kecsegtet, mint például statisztikák leszűrése, automatikus összehasonlítás a korábbi leletekkel, vélemények automatikus generálása, vagy leletek gyors, vázlatpontos áttekintése. Ezek a jövőben egyúttal tehetnék hatékonyabbá és könnyebbé a radiológus munkáját és szolgálhatnának eszközként a magas szolgáltatási színvonal fenntartása érdekében.

A helyes gépi értelmezéshez azonban feltétlenül szükség van a szöveg elemeinek, entitásainak pontos beazonosítására, tudnunk kell egy testrész leírásáról, hogy pontosan melyik testrészt jelöli, és el kell tudnunk igazodni a megannyi különböző szóalak és szinonima között mind a testrészek, mind a megállapított elváltozások esetében.

2. Áttekintés

Jelen munkában a radiológiai mágneses rezonancia (MR) gerincleletek gépi értelmezésére vonatkozó azonosítási módszereink eredményeit ismertetjük, amelyet a magyarlánc (Zsibrita és mtsai, 2013) nyelvi elemző rendszerrel való feldolgozás alapján alakítottunk ki, és esettanulmányként szolgálhat a hasonló, szakkifejezésekre erősen támaszkodó, szűkös szókincsű természetes nyelvű szövegek gépi értelmezéséhez.



1. ábra: Példamondatok egymásra utalással és viszonylag komplex szerkezettel

A munka során építünk korábbi munkánkra (Kicsi és mtsai, 2019), amely szintén leletek szövegét dolgozza fel. Ebben a szövegben előforduló testrészek, elváltozások és tulajdonságok detektálása volt a célunk. Testrésznek az emberi test egy pontját tekintettük, amely egy viszonylag szűkös, a szöveg alapján alkotóelemekre már nem bonható helyet jelöl. Elváltozásnak tekintettünk minden olyan kifejezést, amely megállapítást fogalmaz meg egy adott testrész állapotáról, illetve annak változásáról. Ide tartoztak a különböző aspektusok is, mint például *„víztartalom”*, amely önmagában nem megállapítás, de a *„víztartalom csökkent”* kifejezés részeként mégis egy elváltozás részét képezi. Szintén ide tartozott a normális állapot megállapítása is, ugyanis a radiológus általában ezen információt is rögzíti, hiszen a károsodás hiánya is értékes információt hordoz egy vizsgálat során. Tulajdonságnak olyan kifejezéseket tekintettünk, amelyek egy elváltozás fokozatát, mértékét vagy egyéb, az elváltozás megnevezéséből nem egyértelmű jellemzőjét írják le, mint például *„3 mm-es”* vagy *„körkörös”*. Ugyanezen nevezéktannal dolgozunk jelen írásban is. A szövegben előforduló tagadásokat a detektálás fázisában nem kezeltük. Detektálási módszerünk gépi tanuláson alapult, melynek során 487 lelet kézzel annotált szövegén tanított Bi-LSTM (Hochreiter és Schmidhuber, 1997) segítségével címkéztük a kifejezéseket, kielégítő eredményekkel.

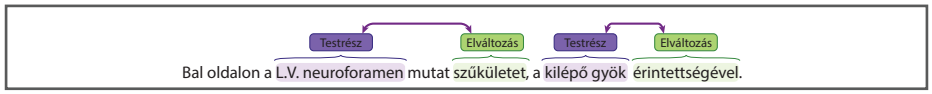
Egy lelet általában állítást fogalmaz meg egy bizonyos testrésszel kapcsolatban. Erre láthatunk két példát az 1. ábrán. Az ábrán jelölésre kerültek a módszerünk által detektált testrészek, elváltozások és tulajdonságok. Azt tehát jól látjuk, és a számítógép számára is egyértelmű már, hogy például a *„CV. discus"* egy testrészt, míg a *„előboltosul"* egy elváltozást jelöl. Az viszont továbbra sem egyértelmű, hogy az emberi test melyik részéről ejtettünk szót, illetve hogy az elváltozásunk pontosan melyik ismert elváltozás, és említése pozitív vagy negatív színezetű-e. Az is megfigyelhető, hogy a tagadás teljes egészében a látókörünkön kívül kerül. A különböző egységek detektálása tehát megtörtént, ám az azonosítás egyáltalán nem, és a mondatok szemantikai jelentése ismeretlen marad.

A fenti problémák tisztán gépi tanulással történő orvoslására nehéz feladat, mely nagy mennyiségű (milliós, vagy milliárdos nagyságrendű), megfelelő minőségű tanítóadat rendelkezésre állása esetén ugyan kivitelezhető lenne, ilyen adatbázisok sajnos még angol nyelvre is nehezen hozzáférhetőek, magyarul pedig még kevésbé. További segítséget nyújthatnának a területspecifikus ontológiák. A jó minőségű angol nyelvű ontológiák nem szabad hozzáférésűek, a szabadon használhatóak pedig egyelőre elmaradnak minőségben a zárt hozzáférésű társaiktól. Ennél is szomorúbb tény, hogy magyar nyelvre tudomásunk szerint átfogó orvosi témájú ontológia nem is létezik. A nagy mennyiségű tanítóadat és a magyar nyelven elérhető ontológiák hiánya miatt az azonosítási és értelmezési feladatainkat nyelvi jellemzők és szabály alapú módszerek alapján végeztük. Jelen írásban ezen megoldásokat tárgyaljuk. Célunk a detektált testrészek és elváltozások azonosítása, kapcsolataik megállapítása, és szemantikai függőségeik feloldása.

3. Módszer

Azonosítási módszerünk egy nyelvi modellen alapul, amelyhez a magyarlánc (Zsibrita és mtsai, 2013) elemző segítségével nyerünk ki bizonyos jellemzőket, majd szabály alapú módszerekkel rendelünk azonosítókat az egyes detektált entitásokhoz. Ide tartozik a szinonimák feloldása is, csakúgy mint az összetartozó latin és magyar szóalakok egymáshoz rendelése. A testrészekhez és elváltozásokhoz egyedi azonosítókat készítettünk, amelyek radiológus által is átnézésre kerültek. Olyan azonosítóhalmazt sem magyar, sem angol nyelvű kapcsolódó kutatásokban, sem nyilvánosan elérhető adatbázisokban nem találtunk, amely elegendő mélységig tartalmazná a gerinc területén található testrészeket és lehetséges elváltozásokat. Az ilyen adatok és ontológiák sajnos még angol nyelvre is kevéssé rendszerezettek, számos kívánnivalót hagynak maguk után az általunk tekintett mélységben. A tulajdonságok azonosításával jelen fázisban nem foglalkozunk, ezek ugyanis általában bonyolultabb szemantikai tartalmat fogalmaznak meg, amely nem feltétlenül írható le előre megalkotott azonosítókkal.

A magyarlánc nyelvi elemző (Zsibrita és mtsai, 2013) a magyar nyelvű szöveg morfológiai, konstituens és dependencia elemzését támogató szoftver, amelyet számos, magyar nyelvű szövegekkel foglalkozó kutatás nagy sikerrel felhasznált már. Az általa biztosított nagyszámú lehetőség közül munkánk során legfőképp a konstituens elemzésre támaszkodtunk, illetve a morfológiai elemzés során feltárt



2. ábra: Egyszerűbb példamondat testrészekkel és elváltozásokkal

tagadásokra. A konstituens elemzés egy fa struktúrát ad, amelyben a mondatok alkotóelemei figyelhetők meg, elkülöníthetők belőle a tagmondatok, amelyek már általában egyetlen szemantikai tartalomra fókuszálnak a nagyobb, összetett mondatok esetében is. Ezt rendkívül hasznosnak találtuk, ugyanis a leletekben szereplő mondatok (például 1. ábra) túlnyomó része egy testrészre mond ki egy elváltozást. A tagadószavak is általában a velük egy tagmondatban lévő entitásokra vonatkoznak, mégpedig pontosan az elváltozásokat leíró szavakra, ahogy a példa második mondatában a *„protrusio”* kerül tagadásra. Ezzel pedig mind a detektált entitások kapcsolata, mind a tagadás tárgya igen könnyen felfedhető. A feladat természetesen nem ennyire egyszerű, rengeteg kivétel felmutatható, ám ezen feltételezések kiindulópontnak mégis több mint elegendők.

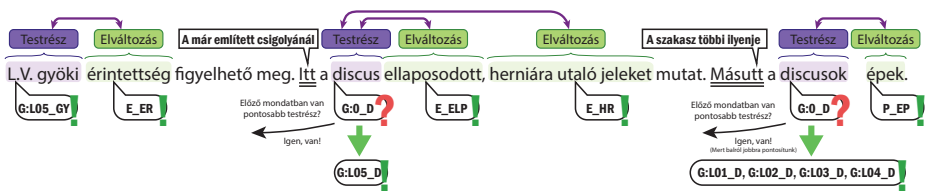
Kézenfekvő megoldás lehetne a dependenciákra támaszkodni a konstituensek helyett, ám a fejlesztés és kísérleteink során azt tapasztaltuk, hogy a konstitunsekre való építés elegendő mélységig tárja fel a szavak egymáshoz tartozását, míg a dependencia a jelenleg kezelt szövegen hibákkal terhelt. Ennek fő oka a speciális nyelvezet lehet, amelyben magyar és latin szavak keverednek. A mondatok azonban itt általában egyszerűbb szerkezettel rendelkeznek egy általános magyar szövegnél, így a konstituensek jó eredményt adnak.

A testrészek és elváltozások azonosítását tehát szabály alapú módszerekkel végezzük. Ehhez felhasználtuk a már meglévő leleteinkben általában szereplő, a detektálás során feltárt testrészeket és elváltozásokat. A szavakat lemmatizáltuk magyarlánc segítségével, ám az eredményeket mindenképpen kézzel kellett javítani, ugyanis itt gyakran speciális szavak fordulnak elő, amelyek a magyarlánc szótárában egyáltalán nem lelhetők fel, és bár az megpróbálja a ragozást így is kimutatni, mégis sokszor problémákba ütközik. Erre lehet példa a *„myelon”* szó, amelyet az elemző egy „myel” szó helyhatározóraggal ellátott változatának tekint. A latin szavak ezen kívül sokszor magyar ragozással fordulnak elő a szövegben (mint például *„herniálódott”* vagy *„degeneratiora”*). Az azonosítók megalkotásánál ezeket először előfordulási szám szerint rendeztük sorrendbe, majd azokat a szavakat tekintettük át, amelyek legalább 10-szer előfordultak a 7648 leletben. Ezek közül lexikografikus listázást követően, kézzel szűrtük ki a helytelenül leírt kifejezéseket. Az összegyűjtött szavak kézzel kerültek csoportosításra, a szóhoz megítélt szótő alapján. Az előálló halmazokhoz azonosítókat rendeltünk. Ezen szóhalmazok közül radiológus segítségével jónéhányat egyesítettünk szinonimák alapján, így például a *„sérv”* és *„hernia”* egy halmazba kerültek. Tapasztalataink szerint az elváltozások azonosításához több orvosi szaktudás volt szükséges, míg a testrészek nehézsége, hogy több, hasonló alakban jelenhetnek meg. Később a listát radiológussal való egyeztetés alapján még kézzel bővítettük.

3.1. Testrészek

A testrészek különlegessége, hogy kapcsolatban állhatnak egymással, amely jelentősen kihat értelmezésükre. A mondat leírhat egy porckorongot *„L. V. discus”*-ként, de mondhatja azt is, hogy *„Az L. V. szerkezete ép. A porckorong apróbb előboltosulása látszik.”*. Ezért nem elég pusztán egymás melletti tokenek sorozatának tekintenünk őket. Utóbbi szerkezetre a megoldás az, ha külön tudjuk detektálni a testrészt, jelen esetben porckorongot, és külön a helyét pontosító másik testrészemlítést, itt az L.V. csigolyát. A testrészeket két részre bontottuk, a csigolyákhoz nem rendelhető és a csigolyákhoz rendelhető testrészekre, utóbbiak pontos helye keresendő. Amikor egy pontosítandó testrészt találunk, akkor egy általánosabb azonosítót rendelünk hozzá, mint például G:0_D, míg ha egy pontos testrészt találunk, akkor egy informatívabb azonosítót kaphat, mint például G:L05. Az általánosabb azonosítóval ellátott testrészeket utólag próbáljuk meg pontosítani. Itt problémát jelent a koreferencia, ahol az *„egyebütt”*, *„máshol”*, *„itt”*, *„többi”* típusú szavak utalnak a testrészre, ahogy az az 1. ábrán is látható. Az itt említett elváltozásokat így egy előző mondatbéli testrészhez kellene vonatkoztatni. Az ilyen szavak detektálásra kerülnek. Az utalások feloldását a 3. ábra szemlélteti. Az ábrán a zöld felkiáltójeles szövegdobozok azt jelzik, hogy a kifejezés megkapta a hozzá illeszkedő, pontos és kellően részletes azonosítót. A piros kérdőjel azt jelenti, hogy csak egy általánosabb azonosítót kapott, ezt próbáljuk feloldani. Detektáltuk az *„itt”* és *„másutt”* szavakat, amelyekhez előre definiált jelentés tartozik. Az „itt” szó egy korábbi testrész csigolyáját, vagy legalább szakaszát jelöli, míg a „másutt” szót úgy tekintettük, hogy egy korábbi csigolya szakaszában a megnevezett csigolyákon kívüli magasságokat jelöli. Ha ezután találunk megfelelően pontosított testrészt az előző tagmondatban, vagy esetleg az előző mondatban, akkor ezek alapján pontosíthatjuk a bizonytalan testrészt. Mivel balról jobbra oldjuk fel az ilyen utalásokat, a példa mindkét dilemmás esetét helyesen fel tudjuk oldani.

További problémát jelenthetnek az intervallummal megadott testrészek, mint például *„L.II.-L. V. discus”*, ahol az intervallum összes eleméről beszél a lelet. Itt a kötőjeleket, az *„és”* és a *„közti”* szavakat keressük, és előfordulásuk esetén átértelmezzük az érintett testrészeket. Ez viszonylag jól automatizálható, ám figyelni kell olyan esetekre is, mint például *„Th.XII.-L.II.”*, ahol a gerincszakaszok közötti váltást is be kellett építeni szabályként.



3. ábra: Példa koreferencia feloldására

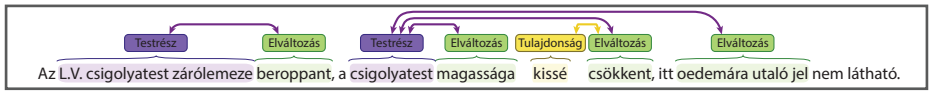
3.2. Elváltozások

Az elváltozások esetében nem igazán fontos két elváltozás kapcsolatát meghatározni. Mivel az aspektusokat a detektálásnál egyben jelöltük az elváltozással, hogy annak valóban értelme is legyen, mint például *„víztartalma csökkent"* esetén, ezért ez az akadály itt nem olyan jelentős. Nagyobb problémát okoz azonban annak értelmezése, hogy pozitív vagy negatív-e az említett elváltozás, tehát az orvos csak megjegyezte, hogy normális állapotot lát, vagy egy valódi degeneratív elváltozást figyelt meg. Ezen megkülönböztetés szintén kézzel került definiálásra. Ezt leginkább a leletek szűkös szókészletének köszönhetően sikerült megfelelő minőségben megtenni. Ez az elváltozások azonosítójában is megjelenik, külön jelöljük a degeneratív elváltozásokat (mint például *„hernia"* - *E_HR*), pozitív állításokat (*„normális"* - *P_NORM*) és az önmagukban polaritással nem rendelkező aspektusokat (*„magassága"* - *ASP_MGS*). Ezeket ismert alakjaiknak megfelelően és magyarlánc segítségével végzett lemmatizálással keressük ki. Mivel a szinonimák már rendelkezésünkre állnak, így ezek tetszőleges szövegben feloldásra kerülnek.

3.3. Kapcsolatok

Bár korábbi elképzeléseink arra irányultak, hogy az entitások közötti kapcsolatokat esetleg gépi tanulási módszerrel állapítanánk meg, úgy találtuk, hogy ezek kézi annotációjára a jelenlegi keretek között nincs feltétlenül szükség. A kapcsolatokat ehelyett a tagmondatokra alapoztuk. Kétféle kapcsolatot kerestünk, testrész és elváltozás, valamint elváltozás és tulajdonság közötti relációkat. A szövegben természetesen előfordulhatnak jelzők a testrészekre is, de ezek az esetek túlnyomó többségében valójában nem is tulajdonságok, hanem elváltozások, mint például *„az előboltosuló discus"* esetében. További megszorító feltételezés, hogy az elváltozások általában egy testrészre, vagy egy testrészek által megadott intervallumra vonatkoznak. Ez szintén helytálló a leletek nagy többségénél, és hasznunkra válik, hiszen így egy elváltozáshoz csak egyetlen testrész keresésére van szükségünk, amelyet a koreferenciák feloldásához nagyon hasonló, priorizált szabály alapú módszerrel valósítottunk meg. A szabályok azonban figyelnek arra, hogy ha *„és"*, *„vagy"* és hasonló szavak választanak el testrészeket, ott minden tagra vonatkozzon az elváltozás.

A leletekben szereplő mondatok tipikusan úgy épülnek fel, hogy először egy testrészt említenek, majd megnevezik a testrész elváltozását, az elváltozás előtt vagy után pedig felsorolják annak tulajdonságait. Ezt megfigyelhetjük például az 1. és 2. ábrán. A mondatok állítmánya gyakran egyik címkéhez sem illeszkedik, mint például *„látszik"* vagy *„észlelhető"*. Természetesen kivételek ez alól a szokás alól gyakran adódnak, ám ezen egyszerű mondatoknál nem nehéz belátni, hogy a kapcsolatok feltérképezése nem komplex feladat. Módszerünk jelen cikk összes példájában szereplő összes kapcsolatot megtalálja. Problémák valójában csak egzotikus megfogalmazás esetén valószínűek, ekkor a kapcsolatot nem sikerül detektálnunk (például *„Mindkét említett discus előboltosul"*). A kapcsolatok hibái általában az entitás detektálás hiányosságaiból erednek.



4. ábra: Példamondat tagadással és több kapcsolattal

3.4. Tagadás

A leletekben gyakran előfordulnak tagadó mondatok is, amik sokszor egy degeneratív elváltozás hiányát írják le. Erre az 1. és 4. ábrán láthatóak példák. Az ismertebb tagadószavakat a magyarlánc is felismeri. Ezek pontos tárgyát is gyakran megadja a dependencia, ezen specifikus szövegeknél azonban azt tapasztaltuk, hogy sokkal jobb eredményeket kapunk, ha ebben is a konstituensekre hagyatkozunk. Ezért itt az előzőekben leírt módszer egy nagyon egyszerű változatát alkalmaztuk, a tagadást tagmondatonként értelmeztük, és amennyiben egy tagmondatban szerepel tagadószó, akkor az egész tagmondatot tagadónak tekintettük. A tagadószó detektálásra a magyarlánc morfológiai elemzését használtuk, ezt azonban ki kellett egészítenünk a *„nincs”*, *„nincsenek”*, *„sincs”* és *„sincsenek”* szavakkal. Egy tagmondat tagadása általában a benne szereplő egyetlen elváltozás jelenlétének tagadása, amely a lelet értelmezése szempontjából kritikus fontosságú. Tapasztalataink alapján így megfelelő eredmények érhetők el a tagadás detektálásában jelenlegi szűk területünk tekintetében.

4. Eredmények

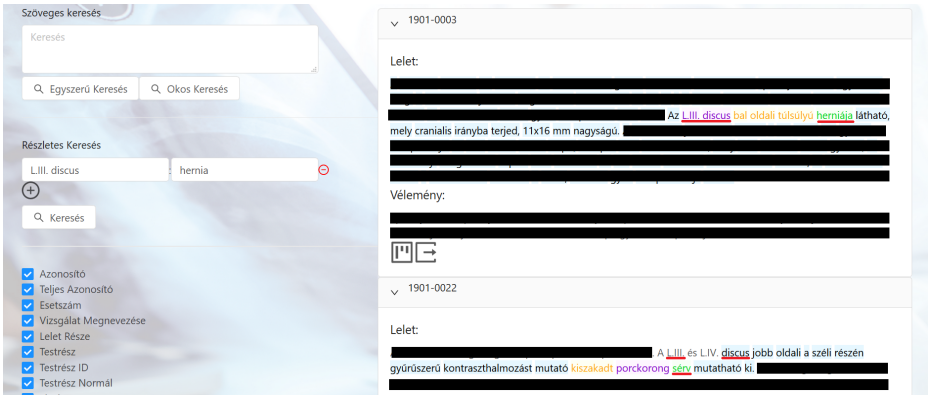
Szabályaink és azonosítóink megalkotása során 5649 lelet adataival dolgozunk, az eredményeinknél közölt számokat pedig a detektálás tanításához is használt 487 leleten végeztük. Módszerünk meghatározott szabályok alapján törekszik testrészek és elváltozások azonosítására és szemantikai kapcsolataik detektálására. Ehhez először is azonosítók szükségesek, amelyeket 5649 lelet adatai alapján alkottunk meg. 39 különböző testrészt különböztetünk meg a csigolyák számait nem tekintve. Ebből 20 testrésznek lehet csigolyaszámozása, tehát ezek mindegyikéhez az általános formán kívül tartozik még 29 pontosítás, ez összesen 629 testrész azonosító. Az elváltozásoknál 214 kóros elváltozást, 8 pozitív jelentésű elváltozást és 20 aspektust különítettünk el, ez összesen 242 elváltozás azonosító.

A 487 leletben 6359 testrész és 7785 elváltozás címke volt. A **testrész** azonosítás során 10258 testrész azonosítót osztottunk ki (ez több, mint az összes detektált testrész, leginkább az intervallumok és a koreferenciák miatt van), ebből 355 különböző. 488 testrészhez nem tudtunk azonosítót rendelni. Az **elváltozás** azonosítás során 9171 elváltozás azonosítót osztottunk ki (itt a többlet az aspektusokból ered, amelyek részei az elváltozásnak), ebből 177 különböző. 332 elváltozáshoz nem tudtunk azonosítót rendelni. Az azonosíthatatlan testrészek és elváltozások természetesen hibát képeznek, ezek nagy valószínűséggel kevésbé gyakori elemek vagy megfogalmazások, ezek javítását a jövőben szintén

kézzel kell megtenni. Szintén ide tartozik a viszonylag nagy mennyiségű, elírások által rontott szóalak, ezek automatikus javítása is utat engedhet a további helyes azonosításhoz. Elmondható azonban, hogy jelenleg a detektált adatokból a testrészek 92,3%-át és az elváltozások 95,7%-át azonosítani tudtuk.

A **koreferencia feloldásban** a vizsgált 487 leletben a következő utalószavak fordultak elő: máshol(376), ebben a magasságban(254), többi(138), itt (122), egyebütt(38), ezen magasságban(34), vizsgált magasságban(20), ugyanebben a magasságban(14), másutt(5), és ugyanitt(3).

A vizsgált 487 leletben összesen 10306 **kapcsolatot** tártunk fel, ebből 6924 elváltozás és testrész, míg 3382 elváltozás és tulajdonság kapcsolata. 843 testrész és 129 tulajdonság volt, amelyhez nem tudott módszerünk elváltozást rendelni. Az elváltozások nélküli testrészek szinte mindegyike abból ered, hogy az egyik testrészt egy másik pontosította, mint például *„L.V. magasságban a discus”*, ilyenkor csak a pontosabb testrészhez kötöttük az elváltozást. A szabadon maradt tulajdonságok nagyrészt a detektálás hibáiból, vagy furcsa megfogalmazásból erednek. 1131 elváltozáshoz nem volt megadva, vagy nem sikerült detektálni egy testrészt sem, itt gyakran mélyebb szemantikai értelmezés lenne szükséges, illetve jó néhány esetben még olvasva sem egyértelmű, hogy milyen testrészhez kötődik egy adott elváltozás. Olyan elváltozások is léteznek, amelyek önmagukban már az érintett testrészre is utalnak. 774 elváltozáshoz nem találtunk egy tulajdonságot sem.



5. ábra: A keresőfelület képernyőképe valós leletekkel. A teljes leletet titkosítottuk, ám a keresés szempontjából lényeges mondatokat meghagytuk

A 487 leletben a magyarlánc konstituens elemzése 6694 tagmondatot tárt fel, ebből módszerünk 1157 tagmondatot tekint **tagadónak**. Nem találtunk olyan valós példát, amelyen a tagadás detektálás hibás eredményt adna, az itt előforduló hibák korábbi feladat hibáiból eredtek minden esetben, mint például az elváltozások detektálásából, vagy a tagmondatokra bontás hiányosságaiból.

Mesterséges példákkal szintén előállíthatók tagadási hibák, külön tagadószavak nélküli megfogalmazásokkal, ám ezeket tapasztalataink szerint nem használják a leletezésben.

Azonosítási módszerünk jelen fázisban már számos felhasználási lehetőséggel bír. Az egyik ilyen lehetőség lehet a leletek intelligens keresése testrészek vagy elváltozások alapján. A módszerre épülő kereső képernyőképe az 5. ábrán látható. A keresőbe beírható keresendő szöveg, ahogy egy hagyományos keresőnél is. Ezen felül azonban testrészek és elváltozások is megadhatók, amelyet kész lehetőségek közül választhatunk, vagy akár sajátot is beírhatunk. Ha a keresődobozra kattintunk, megkapjuk az összes testrész vagy elváltozás listáját, amelyben minden elem csak egyszer (tehát szinonimák nélkül) szerepel. Amennyiben azonban mégis például sérvre szeretnénk keresni hernia helyett, akkor ezt is megtehetjük, ugyanis módszerünkkel ez a keresőszó is kap azonosítót, amely ugyebár megegyezik a hernia azonosítójával. Több testrész és elváltozás is megadható, illetve amennyiben egymás melletti dobozban választjuk őket, a két megadott elem kapcsolatára is szűrünk. Az ábrán jobb oldalon látható két lelet, amelyeken látható, hogy valóban tartalmazzák a keresőszavakat valamilyen formában, és ezek említései kapcsolatban is vannak. Az ábrán látható keresésre egyébként 165 találat volt a 6748 leletből, ezek véletlenszerű sorrendben jelennek meg.

5. Kapcsolódó kutatások

A klinikai szövegek természetesnyelv feldolgozási folyamatában egy fontos lépés, hogy a szavakat kategorizálni tudjuk bizonyos szempontok szerint. Az egyik legalapvetőbb osztályozási forma, amikor a szavakat előre meghatározott névelem osztályokba soroljuk, mint például testrész, betegség, kezelési forma stb. Névelemfelismerés során ugyan meghatározzuk, hogy a mondatban melyik szó milyen osztályba tartozik, ez azonban csak az első lépés a szavak értelmezésének irányába. Az értelmezés vagy azonosítás során az egyik probléma, amivel szembesülhetünk, hogy két ugyanúgy írt szó különböző jelentéssel bír. Ebben az esetben névelemegyértelműsítés segítségével tudjuk feloldani a szavak jelentésbeli különbözőségét. A másik eset, amikor egy jelentéshez, fogalomhoz több különböző szóalak is rendelhető, ilyen esetben névelemnormalizálást alkalmazva, a különböző szóalakokat egy közös fogalomhoz, vagy azonosítóhoz rendelve a probléma feloldható. A nemzetközi szakirodalomban a gyakorlat, hogy a szavakat valamilyen ontológia fogalmaihoz rendelik. Ilyen ontológia például a MeSH (Medical Subject Headings), az RxNorm, a UMLS (Unified Medical Language System) és a SNOMED CT (Systematized Nomenclature of Medicine—Clinical Terms), mely a UMLS részét képezi.

A névelemnormalizálás kérdése nagyjából egy idősnek tekinthető a névelemfelismerés problémájával, klinikai szövegeken az első hivatalos megmérettetést a 2013-as ShARe/CLEF eHealth Evaluation Lab Task 1 kihívás jelentette, mely során klinikai szövegekből kellett betegségneveket felismerni és normalizálni (Pradhan és mtsai, 2014). A verseny célja az akkoriban legmodernebbnek számító megvalósítások összegyűjtése és egyben egy alapvonal meghúzása volt ezen a

területen. Az angol nyelvű ontológiák kihasználása érdekében számos eszközt fejlesztettek már, melyek a klinikai szövegekben található releváns kifejezéseket rendelik az ontológiákban található fogalmakhoz. Az ontológiához való hozzárendelést a korai programok javarészt még szabályalapú algoritmusokkal végezték, az elmúlt években azonban folyamatosan jelennek meg az egyre szofisztikáltabb, gépi tanulást alkalmazó modellek. A korai modellek közül néhány említésre méltó példa:

- MedLEE (Friedman, 2000): Szabályalapú eszköz, melyet eredetileg mellkasröntgen leletek feldolgozására fejlesztettek, azóta azonban kiterjesztették a felhasználhatóságát egyéb területekre is.
- MetaMap (Aronson, 2001; Aronson és Lang, 2010): Tudás-intenzív eljárást, azaz természetesnyelv feldolgozási és számítógépes nyelvészeti eljárások egyvelegét alkalmazva rendeli tudományos orvosbiológiai szövegek szavait az UMLS fogalmaihoz.
- cTAKES (Savova és mtsai, 2010): Egy inforációkinyerésre alkalmazható, szabad forrású szoftver, mely többek között használható orvosi szövegekben előforduló kifejezések UMLS fogalmakhoz történő hozzárendelésére is.
- YTEX (Garla és Brandt, 2012): Egy sor kiegészítő modul cTAKES-hez, mely egy általános keretrendszert biztosít szavak ontológiákhoz történő hozzárendeléséhez.
- DNorm (Leaman és mtsai, 2013): Gépi tanulást alkalmazó eszköz, mely hasonlóságot számít a klinikai szövegekben előforduló kifejezések és az ontológia fogalmai között.

Rohit algoritmusa az UMLS-ben található kifejezéseket alapul véve, szerkesztési távolságon alapuló mintázatokat tanult meg, majd ezeket a mintázatokat általánosította korábban nem látott esetek normalizálására (Kate, 2015). Jonnagaddala és szerzőtársai orvosbiológiai szövegekben található betegségnevek felismerésére fejlesztettek CRF (feltételes valószínűségi mezők) alapú névelemfelismerő rendszert, valamint vizsgálták a szótári egyezéskereséses módszerek továbbfejlesztési lehetőségeit, pontosabb névelemnormalizálási eredmények elérése érdekében (Jonnagaddala és mtsai, 2016). Leaman és szerzőtársai a DNorm eszközön alapuló, klinikai szövegekre optimalizált rendszert fejlesztettek, melyet DNorm-C névre kereszteltek (Leaman és mtsai, 2015). A rendszer normalizáláson kívül névelemfelismerést is végez, a klinikai szövegben előforduló kifejezések és az ontológia fogalmai között pedig direkt módon tanul párossával hasonlósági függvényeket. A szerzők állítása szerint a párokban történő tanítás segíti a névelemfelismerő rendszer teljesítményét változatos kifejezéseket tartalmazó szövegek feldolgozásában, valamint a módszer kiterjeszthető más területekre is. A szerzők egy későbbi tanulmányukban elsőként mutatnak be egy semi-Markov modellen alapuló rendszert, mely névelemfelismerést és normalizálást egyidőben végez, mind tanítás, mind pedig kiértékelés közben (Leaman és Lu, 2016). A TaggerOne névre keresztelt rendszer ráadásul szabad forrású. Wang és szerzőtársai saját, kizárólag testrészekből álló ontológiát építettek az UMLS fogalmai alapján, gépi tanuláson alapuló névelemnormalizáló algoritmusuk teljesítményét pedig a Wikipédia tudásbázisára támaszkodó pontozó algoritmussal fejlesztették

tovább (Wang és mtsai, 2019). Az elmúlt évek újabb technológiáit a névelemnormalizálás területén is próbálják alkalmazni, így nem régiben Li és szerzőtársai állítottak fel orvosbiológiai és klinikai szövegek normalizálása terén state-of-the-art eredményeket BERT alapú rendszerükkel (Li és mtsai, 2019).

A magyar nyelvű klinikai szövegeken végzett ide vonatkozó kutatások közül, mindenképp említésre méltó Siklósi és Novák munkája, melyet az orvosi szövegekben található rövidítések megtalálása és feloldása terén végeztek (Siklósi és Novák, 2013; Siklósi és mtsai, 2014; Siklósi és Novák, 2014).

Rendszerünk sajátossága, hogy kevés rendelkezésre álló lelet mellett, valamint magyar nyelvű, területspecifikus ontológia hiányában is képes megfelelő pontosságú névelemazonosítást végezni. Az azonosítás szabályalapon történik, melyhez a leletek szövegét felhasználva egy saját kezdetleges ontológiát is építettünk. Elért eredményeink alapot szolgáltatnak, további kutatások, valamint összetettebb ontológiafejlesztés számára.

6. Összegzés

Cikkünkben azonosítási és információkinyerési feladatokat végeztünk radiológiai leleteken. Korábbi munkánk detektálására is építve azonosítottunk testrészeket és elváltozásokat, amelyekhez saját azonosítóhalmaz definiálására volt szükség. A testrészek, elváltozások és tulajdonságok kapcsolatait is feltártuk, ehhez leginkább konstituens elemzés eredményeire támaszkodva.

Értelmeztük továbbá az intervallumokat, az elváltozások kórosságát, a tagadásokat és utalásokat is. Bemutattuk a módszerrel előállított eredményeinket 487 valós leletre. Munkánknak számos jövőbeli felhasználása lehet a leletek értelmezésében, ilyenek lehetnek az intelligens szemléltetés, strukturált leletek készítése, automatikus véleménygenerálás, vagy, ahogy azt be is mutattuk, intelligens keresés.

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). Készült az EFOP-3.6.3-VEKOP-16-2017-00002 támogatásával.

Hivatkozások

Aronson, A.: Effective mapping of biomedical text to the umls metathesaurus: The metamap program. Proceedings / AMIA Symposium pp. 17–21 (02 2001)

Aronson, A., Lang, F.M.: An overview of metamap: Historical perspective and recent advances. Journal of the American Medical Informatics Association : JAMIA 17, 229–36 (05 2010)

Friedman, C.: A broad coverage natural language processing system. AMIA Symposium pp. 270–4 (02 2000)

Garla, V., Brandt, C.: Knowledge-based biomedical word sense disambiguation: An evaluation and application to clinical document classification. Journal of the American Medical Informatics Association : JAMIA 20 (10 2012)

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

Jonnagaddala, J., Jue, T.R., Chang, N.W., Dai, H.J.: Improving the dictionary lookup approach for disease normalization using enhanced dictionary and query expansion. Database (08 2016)

Kate, R.: Normalizing clinical terms using learned edit distance patterns. Journal of the American Medical Informatics Association : JAMIA 23 (07 2015)

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)

Leaman, R., Dogan, R., Lu, Z.: Dnorm: Disease name normalization with pairwise learning to rank. Bioinformatics (Oxford, England) 29 (08 2013)

Leaman, R., Khare, R., Lu, Z.: Challenges in clinical natural language processing for automated disorder normalization. Journal of biomedical informatics 57 (07 2015)

Leaman, R., Lu, Z.: TaggerOne: Joint Named Entity Recognition and Normalization with Semi-Markov Models. Bioinformatics 32 (06 2016)

Li, F., Jin, Y., Liu, W., Rawat, B., Cai, P., Yu, H.: Fine-tuning bidirectional encoder representations from transformers (bert)–based models on large-scale electronic health record notes: An empirical study. JMIR Medical Informatics 7 (09 2019)

Pradhan, S., Elhadad, N., South, B., Martinez, D., Christensen, L., Vogel, A., Suominen, H., Chapman, W., Savova, G.: Evaluating the state of the art in disorder recognition and normalization of the clinical narrative. Journal of the American Medical Informatics Association : JAMIA 22 (08 2014)

Savova, G., Masanz, J., Ogren, P., Zheng, J., Sohn, S., Kipper-Schuler, K., Chute, C.: Mayo clinical text analysis and knowledge extraction system (ctakes): Architecture, component evaluation and applications. Journal of the American Medical Informatics Association : JAMIA 17, 507–13 (09 2010)

Siklósi, B., Novák, A.: Detection and Expansion of Abbreviations in Hungarian Clinical Notes, Lecture Notes in Artificial Intelligence, vol. 8265, p. 318–328. Springer-Verlag, Heidelberg (2013)

Siklósi, B., Novák, A.: Rec. et exp. aut. Abbr. mnyelv. KLIN. szöv-ben – Rövidítések Automatikus Felismerése és Feloldása Magyar Nyelvű Klinikai Szövegekben. In: X. Magyar Számítógépes Nyelvészeti Konferencia. p. 167–176. Szegedi Tudományegyetem, Informatikai Tanszékcsoport, Szeged (2014)

Siklósi, B., Novák, A., Prószéky, G.: Resolving abbreviations in clinical texts without pre-existing structured resources. In: Fourth Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing (BioTxtM 2014). Reykjavík (2014)

Wang, Y., Fan, X., Chen, L., Chang, E., Ananiadou, S., Tsujii, J., Xu, Y.: Mapping anatomical related entities to human body parts based on Wikipedia in discharge summaries. BMC Bioinformatics 20, 430 (08 2019)

Zsibrita, J., Vincze, V., Farkas, R.: magyarlanc: A Toolkit for Morphological and Dependency Parsing of Hungarian. http://rgai.inf.u-szeged.hu/index.php?lang=en&page=magyarlanc (2013)