XVI. Magyar Számítógépes Nyelvészeti Konferencia Szeged, 2020. január 23–24.

Elírások automatikus detektálása és javítása radiológiai leletek szövegében

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

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

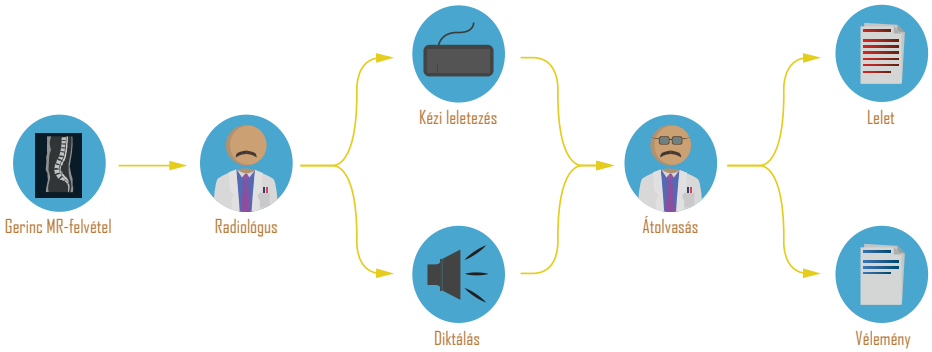
Kivonat A radiológiai leletezés közben gyakran előfordulhatnak szövegbéli hibák, melyek kézi javításra szorulnak. Ez időt von el a radiológustól, valamint a hibák sikertelen felismerése nyomán rontja a leletek minőségét és utólagos gépi feldolgozhatóságát is. Cikkünkben magyar nyelvű gerincleletek elírásainak automatikus kijavításával foglalkozunk. Ismertetjük az általunk felhasznált módszereket, és megmutatjuk, hogy az elírások automatikus javítása az értelmezést is nagymértékben javítja. Módszerünket 882 valós lelet kézi hibajavításával vetjük össze.
Kulcsszavak: radiológia, elírásjavítás, helyesírás, információkinyerés, NLP

1. Motiváció

A modern egészségügyben egyre nagyobb hangsúlyt kapnak a különböző számítógépes megoldások, ezen belül a mesterséges intelligencia is (Novák és Siklósi, 2015, 2016; Mykowiecka és Marciniak, 2007). Korszerű szoftverek képesek támogatni a műtéti tervezést, képi adatokon anomáliák keresését, és a klinikai leletezést is. A leletezésben például már egyáltalán nem ritka, a magyar egészségügyben sem, hogy az orvos gépelés helyett diktálja a leleteket egy szoftvernek, amely ezeket automatikusan rögzíti, igen jó minőségben, magyar nyelven is. Elmondható tehát, hogy rohamos fejlődés figyelhető meg az automatizálásban és az orvosok munkájának gépi segítésében (Kernighan és mtsai, 1990; Lai és mtsai, 2015).

A radiológiai vizsgálatokban is ugyanígy jelen van naprakész technológia. Jelen cikkben a diagnosztizálás ezen szegletével foglalkozunk, sőt ezen belül is csak a radiológiai gerinc vizsgálatokkal. Csak Magyarországon évente sok ezer gerinc Röntgen- és MR-felvétel készül. Ezek szöveges formában, magyar nyelven íródnak. A szakorvos elküldi a pácienst radiológiai vizsgálatra, a felvételek elkészülnek, majd ezeket egy radiológusnak továbbítják, aki szöveges lelet formájában rögzíti a képeken látott érdemleges információt. A lelet kulcsfontosságú része a radiológiai vélemény, mely összefoglalva tartalmazza a leletben tett legfontosabb megfigyeléseket, tömör orvosi szakzsargonnal, valamint latin rövidítésekkel

megfogalmazva. A lelet a radiológustól visszakerül a szakorvoshoz, aki az információ birtokában meghozza a páciens további kezelésére vonatkozó döntéseket. A leletezés folyamatát az 1. ábra illusztrálja.



1. ábra: A radiológus munkája a lelet rögzítése során

A leletezés során a radiológusnak már számos magyar klinikán is lehetősége nyílik diktáló szoftver használatára. Tradicionálisabb esetekben maga a radiológus gépel, vagy esetleg radiográfus asszisztens, szintén diktálás nyomán. Nem nehéz belátni, hogy a megfelelő minőségű diktálás értékes időt takaríthat meg. A magyar nyelvű klinikai diktálók minősége is igen fejlett már, bár valóban jó eredményeket leginkább egyéni tanítás után, és csak a speciális területen belül tudnak elérni. A radiológus számára tehát jelenleg adott a választás lehetősége, hogy időt szán a szoftver használatának betanulására, esetleg tanítására, vagy ragaszkodik a régi módszerekhez.

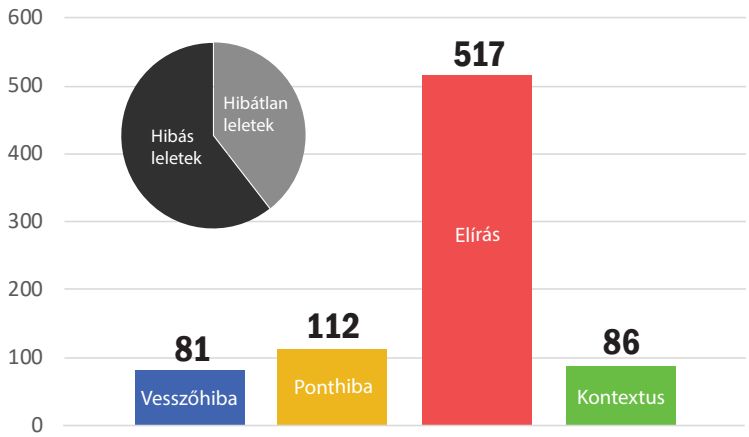
Bármelyik esetet is választja, az elkészült leletet valószínűleg át kell olvasnia a hibák kiküszöbölése érdekében. A hiba természetesen lehet kritikus, például ha kifelejtett valamit a leletből, viszont leggyakrabban csak elírásokról van szó. A diktáló szoftvereknél egyértelműen szükség van még erre a lépésre, ám pozitív oldaluk, hogy elírásokat, betűcseréket és értelmet nélkülöző szavakat sokkal kisebb valószínűséggel írnak a leletbe, hiszen értelmes szavakon tanították azokat.

A leletekben különösen gyakorinak számítanak a felcserélt betűk, amelyek latin szavakban még kevésbé észrevehetők, illetve az elmaradó ékezetek. Van például olyan radiológus, aki ugyan ír ékezeteket, az "ő" és "ű" betűket konzisztens módon ékezetmentesíti. A hibák természetesen nem csak a gépi, hanem az emberi értelmezést is ronthatják. A kézi javítás során körülbelül minden negyvenedik leletben volt olyan mondat, ami egyszerűen félbemaradt, vagy szavai szórendje alapján nem lehetett megfelelően értelmezni.

Kutatásunkban radiológiai leletek automatikus értelmezését tűztük ki célul, a témában korábban már publikáltuk módszerünket (Kicsi és mtsai, 2019), amellyel radiológiai leletekben szereplő testrészeket, elváltozásokat és tulajdonságokat detektáltuk automatizált módon, gépi tanulási módszerekkel. A testrész az

emberi test pontos részének megnevezése (például "L.V. discus"), elváltozás lehet bármely kóros elváltozás ("előboltosulás"), aspektus ("magassága") vagy normális állapotot jelző kifejezés ("normális"). A tulajdonság egy elváltozást módosító, mértéket ("3 mm-es") vagy minőséget ("körkörös") leíró szó. A tanításhoz 487 leletből álló, radiológus által annotált tanítóhalmazt alkalmaztunk. A detektálás természetesen nem jelent teljes értelmezést, publikációnk óta már rendelkezünk kialakított módszerrel a testrészek és elváltozások azonosítására, és a különböző elemek kapcsolatainak megállapítására is. Ennek pontos módszerei nem képezik jelen cikk tárgyát, ám az itt megfogalmazott célok jelentős motivációt szolgáltatnak az itt leírt kísérleteinkhez, illetve értelmezési munkánk az itt ismertetett módszer kiértékeléséhez is hozzájárulnak.

A gépi értelmezés az emberi látásmódtól lényegesen különbözik, így érthető, ha a számítógép esetleg nem képes túllépni a leletekben előforduló elírásokon, azok további hibákat indukálhatnak, amely rontja a rendszer megbízhatóságát. Természetesen a tanítóhalmaz átnézhető és kijavítható utólag, kézi átnézéssel, ahogy ezt meg is tettük. Ám a való életben történő használathoz fel kell készülnünk arra, hogy ilyen hibákat a jövőben is fognak véteni a leletet gépelő személyek. Ezért egy olyan módszert dolgoztunk ki, amellyel az elírásokat nagy valószínűséggel detektálni, és automatikusan javítani tudjuk a gerinc leletekben.



2. ábra: Kézzel talált hibák a 487 leleten

A tanítóadatként felhasználni kívánt 487 leleten első körben kézi javítást végeztünk, mely során a leletekben található összes hibát feljegyeztük a hiba helyével és a hibás mondat szövegével együtt, valamint egy alternatív helyes mondatot is megadtunk a hibás mellé. A hibákat utólag kézi átnézéssel csoportosítottuk négy halmazba. Kézenfekvő hibák a vesszőhibák, melyeket gyakorta ejthetnek mind kézi gépeléssel, mind diktálással, ez lehet kimaradó, vagy felesleges vessző egy mondatban. A ponthibák nagyon hasonlóak, itt legtöbbször a lelet végére, vagy esetleg a szövegtörzs mondatainak végére felejtett el a radio-

lógus pontot tenni. Ez utóbbi a gépi feldolgozás szempontjából lényegesebb a vesszőknél, mivel sok nyelvi elemző egy mondaton belül elemez, ezt pedig az írásjelekből állapítja meg. Nagyon gyakori hiba az elírás, amely annyit jelent, hogy önmagában értelmetlen szó került a leletbe, ez történhet betűk hozzáadásával vagy kihagyásával, felcserélt betűkkel, vagy helytelen ékezetekkel. Latin szó magyar ragozásával megengedőek voltunk, ezek ugyanis nem kiforrott szabályok alapján történnek, viszont nagyon gyakoriak a leletekben. A negyedik hibatípus a kontextusfüggő hibák, amelyek önmagukban helyes szavak helytelen használatán alapulnak, ide tartozik a rossz ragozás, illetve olyan elírások is, amelyek mégis értelmes szót produkálnak. A 487 leletből 295 tartalmazott valamilyen hibát, a hibák nagyon gyakran tömegesen jelentkeztek. A felfedett hibákról statisztikát láthatunk a 2. ábrán.

Ezen eredményekből látható, hogy a hibák nagy arányban olyan elírásokat tartalmaznak, amelyek kontextustól független módon javíthatók lennének. Természetesen egyszerű átírási szabályokkal a pontosan ilyen hibákat ezután is sikeresen tudnánk javítani, de a lehetséges elírási hibák száma óriási, így erre egy automatizált módszer kell. Cikkünkben az erre irányuló kutatásunk eredményeit mutatjuk be. Tehát jelenleg az értelmetlen szavakból és ékezethibákból adódó szavak detektálását és automatikus javítását tűztük ki célként. Az eredményeinket a leletértelmezési kutatásunkban kívánjuk felhasználni, ezért különösen fontos kérdés, hogy a hibajavítás mennyiben befolyásolhatja a névelemfelismerési és -normalizálási feladatokat. Bár a gerincleletek területe igen szűkös, mégis hisszük, hogy az jó esettanulmányként szolgálhat más detektálási vagy azonosítási feladatokhoz is. A kutatás során fejlesztett helyesírásjavító eszközt továbbá közvetlenül a leletezés folyamatában is fel lehetne használni, hogy a helyesírási hibák már keletkezésükkor javításra kerüljenek. Természetesen a hamis találatokat ehhez minimalizálni kell, azonban egy megfelelő automatikus javítás nagyban csökkentheti a leletek végső átolvasási idejét. Bevezetjük tehát az alábbi kutatási kérdéseket:

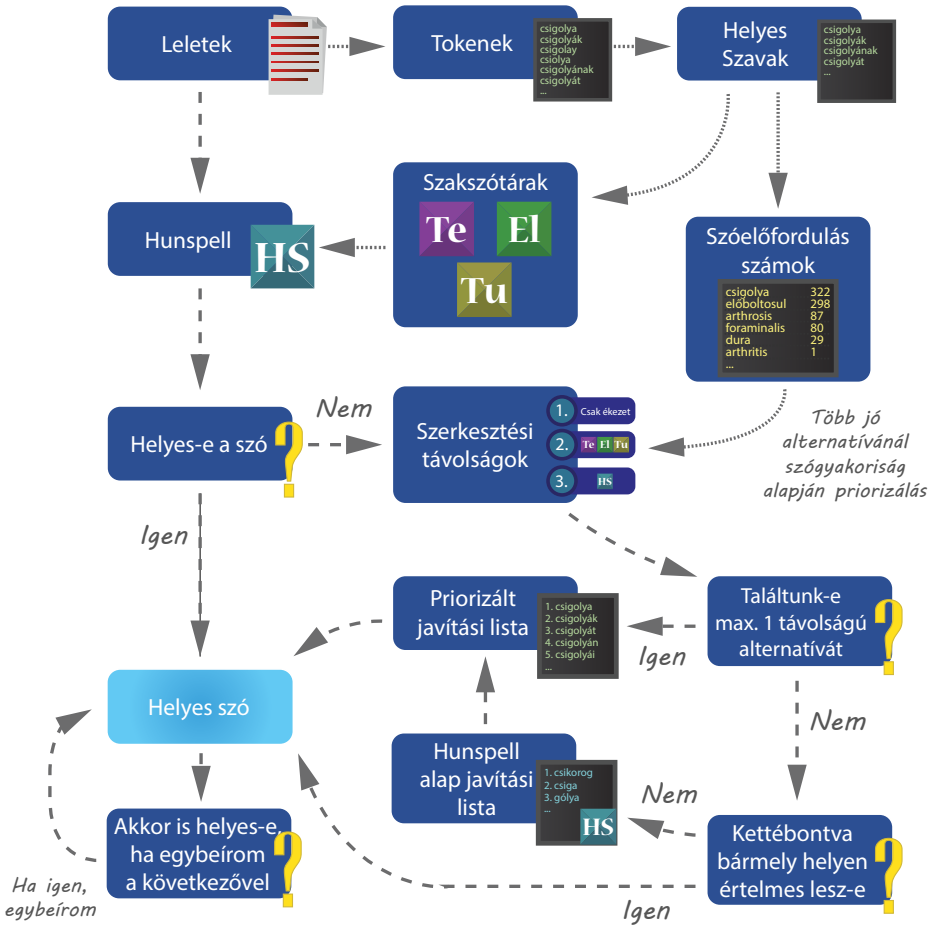
- **RQ1:** Az elírási hibák megfelelő javítása pozitívan befolyásolja-e a detektálási és azonosítási feladatok eredményét?
- **RQ2:** Mekkora részben lehet az emberi kiértékelés helyett az automatikus módszerre hagyatkozni?

2. Módszer

Jelen munkában a radiológiai gerincleletek szövegében törekszünk elírási hibák automatikus felismerésére és kijavítására. A feladatot a Hunspell[1] helyesírásellenőrző szoftver segítségével valósítottuk meg, amely nagy népszerűségnek örvend morfológiailag gazdag nyelvek szövegfeldolgozásában. A rendszer egy beépített szótárral dolgozik, és különböző ragozási és egyéb szabályokkal van ellátva, amelyeknek köszönhetően remekül szűr általános témájú szövegben. A Hunspellt önmagában kipróbálva, irreálisan nagyszámú hibát jelölt a leletekben, ugyanis

[1] http://hunspell.github.io

rengeteg latin, és egyéb orvosi szót hibásnak jelölt, mivel ezek nem voltak benne a szótárában. Természetesen ez nem róható fel a rendszernek, hiszen magyar nyelvű szövegre van kialakítva, a leleteknél pedig az egyszerű latin szavak mellett nagyon gyakran előfordulnak olyan, helyesnek tekintendő szavak is, ahol egy latin szó magyar ragozással szerepel, mint például „herniatiora” vagy „protrusiojának”. Ezen szavak a leletek értelmezhetőségét nem rontják, hiszen az orvosok ezeket konzisztensen és rendszeresen használják.



3. ábra: Automatikus elírásjavító módszerünk működése

Kifejlesztett módszerünk tehát a Hunspell rendszerre támaszkodik, ám szakszótárakat és priorizálási szabályokat kellett kiépítenünk a gerincleletek megfelelő kezeléséhez. Módszerünk teljes sémáját a 3. ábrán mutatjuk be. 5649 lelet állt

összesen rendelkezésünkre, ebből 487 radiológus által beannotált, tanulóadatnak is alkalmas lelet. A helyesírás fejlesztéséhez azonban további leleteket is kiértékeltünk, így összesen 882 leletet használtunk kiterjedt szótárak létrehozásának reményében. A Hunspell eredeti szótára sajnos nehezen birkózik meg az orvosi, nem ritkán latinul írt szakkifejezésekkel. Az ábrán elsőként a pontozott vonalak mentén indulunk el. Az 5649 leletekből első lépésben kinyertük az összes testrészként, elváltozásként vagy tulajdonságként annotált tokent (a Bi-LSTM (Hochreiter és Schmidhuber, 1997) technikával történő detektálás a szövegkörnyezet alapján, olyan szavakat is képes helyesen felcímkézni, amelyeket még nem látott). A Bi-LSTM rendszer által felismert névelemeket lexikografikus rendezés után egyesével átnéztük, a hibás szavakat pedig kézzel szűrtük. Igyekeztünk körültekintően eljárni. Mivel nem rendelkezünk a szükséges orvosi háttértudással minden szó pontos értelmezéséhez, így külön gyűjtöttük az általunk dilemmásnak ítélt tokeneket, amelyekre gyanakodtunk, hogy tartalmazhatnak elírást, ám orvosi szaknyelvhez tartoznak. Ezen dilemmás eseteket a radiológus kézi ellenőrzésével oldottuk fel. A helyes szavakat ezután három szakszótárba soroltuk (testrész, elváltozás és tulajdonság), melyek összesen 5723 szót tartalmaznak jelen állapotban. Ugyanezen helyes szavakat nem csak a szakszótárakban használjuk fel, hanem később, a javítások priorizálásánál is. Ezért az összes szóhoz megállapítunk egy gyakoriság listát, amely aszerint van rendezve, hogy hányszor fordultak elő az összes leletben, ennek felhasználását később látjuk. A szakszótárakat a Hunspell rendelkezésére bocsájtjuk, amely az adatok birtokában bírálja el egy adott új szó helyességét.

Amennyiben helyes volt a szó, akkor nincs igazán teendőnk. Ilyenkor kísérletet teszünk két szó összevonására, mivel gyakori például a „csigolyatest” szó leletbéli különírása „csigolya” és „test” szavakként, amik ugyan külön is értelmesek, de a radiológus valószínűleg egyben szándékozott leírni. Tehát minden helyes szóra leellenőrizzük azt is, hogy ha a következő szóval egybeírnánk, akkor is értelmes szót kapnánk-e. Ha igen, akkor összeolvasztjuk a két szót.

Ha a szó helytelen volt, azaz nem található meg sem a Hunspell saját szótárába, sem pedig a szakszótárakban, akkor megpróbáljuk kijavítani azt. A Hunspell alapértelmezetten egy szerkesztési távolságon alapuló priorizált listát ad a lehetséges javításokról. Ez azonban nem volt megfelelő, szintén a speciális szavak miatt. A priorizálást átalakítottuk olyan módon, hogy először az ékezethibákat tartsa valószínűnek, például az „eloboltosulás” szónál az „előboltosulás” előrébb kerül a listában mint az „elboltosulás”. Amennyiben ilyen alternatívát nem talált, akkor az egyszerű szerkesztési távolságot veszi alapul. Először a szakszótárak szavai közt, majd a saját beépített szótárának szavai között keres javítási lehetőséget. Ezen belül tovább rangsorol a szógyakoriságok alapján, amelynek listáját korábban állítottuk elő. Ez alkalmas az olyan esetekre, amikor két értelmes alternatíva is van, amelyet néha még kézi vizsgálattal is csak nagy odafigyeléssel, vagy tudással lehet javítani. Ilyen például az „arthrotis” szó, amely egyaránt kijavítható „arthritis” és „arthrosis” irányba. Ilyenkor a gyakoribbat választjuk, ami a gerinc területén egyértelműen az „arthrosis” lenne.

Ezután ellenőrizzük, hogy találtunk-e olyan javítást, ami maximum (vagyis valójában pontosan) 1 szerkesztési távolságra van az eredeti szótól. Ezek a legvalószínűbb elírások, ugyanis itt csak egy betű hiánya vagy félreütése jelentkezett. Ha találtunk ilyet, akkor azok egyenesen mennek tovább a priorizálásukat megtartva a javítási listába, majd közülük a legvalószínűbb lesz az ajánlott automatikus javítás. Ha nem volt ilyen alternatíva, akkor ahelyett, hogy valami nagyon különböző szót ajánlanánk, először megpróbáljuk a szót felbontani két, vagy ha ez nem sikeres, akkor akár három részre, ahogy például az „előboltosulódiscus" hibás szó felbontható „előboltosuló" és „discus" helyes szavakra. Ha ezután a szavak mindegyike értelmes, akkor automatikusan azt tekintjük helyes javításnak. Ez természetesen hordoz veszélyeket például szóeleji vagy szóvégi „a" betűknél, de a tapasztalatok alapján nem okoz komoly hibákat. Ha a felbontás sem sikeres, akkor a szerkesztési távolságokkal előállított priorizált listát adjuk tovább, ahogy azt pár lépéssel korábban előállítottuk Hunspell segítségével.

Az esetek többségében tehát egy valamilyen módon priorizált listát kapunk. Az elkészült módszer tapasztalataink szerint szinte soha nem ad hamis riasztást, illetve az ajánlások igen pontosak, az algoritmust leginkább csak a ragozás tévesztheti meg. A megvalósított rendszer futásideje igazán rövid, tulajdonképpen észrevehetetlen, bőven egy másodpercen belül teljesít egy egész leletre, így alkalmas lehet akár írás közbeni ellenőrzésre és javításra is.

3. Eredmények

Motivációnk bevezetésekor két kérdést tettünk fel. Ezek azt vizsgálják mennyiben javítja elírásjavító módszerünk a leletek automatikus értelmezésének minőségét, illetve hogy mennyivel rosszabb, vagy esetleg jobb elírásjavítást kapunk így, mintha emberi munkával, kézileg néznénk át a leleteket.

A leletek értelmezésével foglalkozó kutatásunk (Kicsi és mtsai, 2019) kapcsán képesek vagyunk testrészek, elváltozások és tulajdonságok detektálására, és pontos azonosítására is. Ezen eredmények nem képezik jelen cikk tárgyát, ám illusztrálhatják, hogy a gépi értelmezésre milyen mértékben és irányban hathat ki helyesírásjavító módszerünk. A detektálás Bi-LSTM technikával történt, a modellt 487 annotált leleten tanítottuk, melyből 70% képezte a tanító, 10% a validációs, a maradék 20% pedig a teszthalmazt. A rendszer nyelvi jellemzőket is felhasznál, amelyeket a magyarlánc (Zsibrita és mtsai, 2013) nyelvi elemző eszközzel nyertünk ki a szövegből. Azonosítási módszerünk szintén nyelvi elemzést, valamint szabály alapú módszereket használ testrészek és elváltozások azonosítására egy saját azonosítóhalmaz alapján. A tulajdonságokkal, azok óriási változatossága miatt nem foglalkoztunk.

Felmerül a kérdés, hogy a detektálási eredményekre kihathatnak-e az elírások. Természetesen kihathatnak, hiszen az elírások azon felül, hogy más szóalakot eredményeznek, a magyarlánc elemzését is megzavarják. Több esetben megfigyeltük, hogy az elírt szavakhoz a magyarlánc nem a megfelelő nyelvi jellemzőket rendeli és mivel ezek a Bi-LSTM modell tanítóadatának szerves részét képezik jelentős mértékben torzíthatják a tanítóadatot. A lemmatizálás lényege,

hogy egységes alakra hozza a ragozott szavakat, mely jelentősen tudja javítani a tanítási eredményeket. Ugyanezen gondolatmenetet alkalmazva, az ékezethibákból fakadó számos különböző szóalak egységesítése (javítás által) szintén segítheti a tanítás eredményességét. Detektálási eredményeinket egy mikro-átlaggal összegezve kezdetben 91,09 F1-mérték eredményt kaptunk. Az elírások javítására kidolgozott automatikus módszerünk futtatása és újabb tanítás után az eredményünk már 91,52 F1-mérték, amely csaknem fél százalékos javulás.

Az azonosítás eredményeire még valószínűbb, hogy kihat a javítás, hiszen itt szavakat vagy szótöveket próbálunk szabály alapon illeszteni a szövegre, ha pedig a szövegben ezek nem megszokott formában szerepelnek, akkor valószínűleg egyáltalán nem tudjuk automatikusan azonosítani őket. Itt szembetűnő minőségjelző szám lehet azoknak a testrészeknek és elváltozásoknak a száma, amelyeket egyáltalán nem tudtunk azonosítani. Az eredeti adathalmazban ezek az azonosíthatatlan elemek 6358 testrészből 505 testrész, és 7794 elváltozás közül 521 elváltozás volt. Az elírások javítását követően ezek az értékek 488 testrészre és 332 elváltozásra módosultak. Látható tehát, hogy javításunk valóban nagymértékben kihatott az azonosítási feladat sikerére.

RQ1 válasz: Vizsgálataink alapján a detektálást közepesen jelentős mértékben (saját rendszerünknél 0,43%-ban) pozitívan, míg az azonosítást jelentős mértékben (saját rendszerünknél több mint 20%-ban) pozitívan befolyásolja az általunk felvázolt automatikus elírásjavító módszer.

Második kérdésünk megválaszolásához a kézi adatokra támaszkodtunk. Ezeket, ahogy azt a motivációnál leírtuk, a 487 lelet kézi átnézésével, hibák és javításaik rögzítésével és utólagos klasszifikációjával állítottuk elő. A pontosabb kiértékeléshez ezt bővítjük további 395 lelet ugyanilyen módszerű kézi átnézési eredményeivel, így kiértékelésünket összesen 882 lelet szövegén végezzük. A további leleteket véletlenszerűen válogattuk. Természetesen továbbra is csak az elírásokat tűzzük ki célul, így a vesszőhibákkal nem foglalkozunk. A kézi hibajavítás során összesen 908 esetben történt elírásnak címkézett javítás. A gépi módszernél ez a szám 1248 volt. A kézi hibakeresést minden esetben ugyanaz a személy végezte, a találatok pedig másik személy által kerültek ellenőrzésre és klasszifikációra. Már első ránézésre látható, hogy a gépi hibakeresés jóval több javítást produkál. Kézzel kiértékeltük a két módszer különbségét.

A kézi adathalmazon 36 olyan javítás volt, amelyet nem jelölt a rendszer, ezek többsége értelmes szót adott a hibával, vagy csak hiányzó pont vagy kötőjel volt benne („iv”,„kb 3 mm”). Több esetben volt felesleges javítás is, mint például a „folyadéktartalmú” szó szeparálása. Az ellenőrzés során 5 olyan esetet találtunk, amely valóban hibás volt, és rendszerünk nem fedte fel, ezek a szótár szűrésének tökéletlenségéből eredhetnek. 17 esetben ugyanarra a szóra különböző javítást adott a kézi keresés és rendszerünk, ebből több valójában ekvivalens („levő”,”lévő” vagy „normál”,”normális”), és itt is volt több rossz emberi javítás vagy kontextusfüggő javítás, amelyet nem lehet kitalálni a mondatkörnyezet nélkül, például ragozási forma. 7 olyan eltérés volt, amelyet valóban hibás javításnak ítéltünk meg. Ezek leginkább olyan esetek, amikor a névelő egyben volt a szóval, a javításunkban pedig teljesen kimaradt („Abal”).

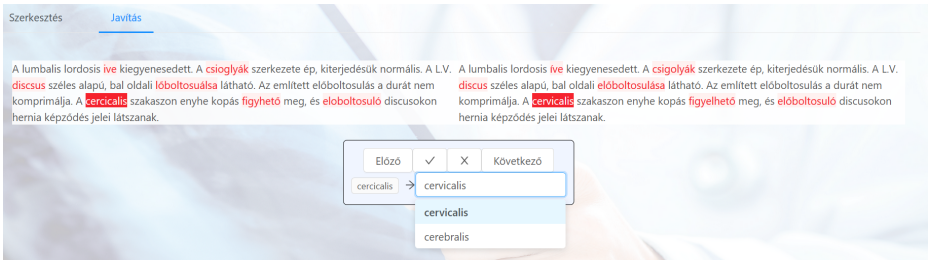
További nagyon lényeges kérdés, hogy mennyire valós a 328 hiba, amit az automatikus keresés felfedett, az emberi javító pedig nem. A hibajavítás monoton feladat, és betűcserék felett könnyen elsiklik az emberi szemlélő, így elképzelhető ekkora kihagyás. Ennek vizsgálata során a következőket állapítottuk meg: a detektált elemekből összesen 4 javítás volt indokolatlan, ezek önmagukban nem szokványos, de valószínűleg szándékosan így írt szavak („VA”,”VB”, ”radici”,”gerinccsatorna- és”). További 3 szó indokoltan volt jelölve, ezeket azonban rossz javítással láttuk el. Tehát megállapíthatjuk, hogy a 328 detektált hibából 324 emberi szemlélő számára is indokolt hibajavításnak tekinthető. Belátjuk tehát, hogy rendszerünk valóban képes emberi javítással összemérhető, sőt, túl is szárnyaló hibajavításra. A számok közti jelentős különbség alapján felmerülhet, hogy ha 1244 közeli javítás valóban indokolt, mennyire tekinthetjük a kézi kiértékelés eredményét egy megbízható összehasonlítási alapnak a nagyszámú kihagyott javításával. A kézi kiértékelés természetesen tökéletlen, de ez pusztán az emberi figyelem, és nem a kiértékelő személy hiányosságait tükrözi. Az is ezt mutatja, hogy egy orvos a jelen leleteket egyszer már át kellett, hogy olvassa, az említett számos hiba azonban így is előfordul.

Mivel a kézi javítás eredményeit felhasználva készítettük módszerünket, nem meglepő, hogy ezeken jól teljesít, hiszen rájuk optimalizáltunk. Ezért további 300, nem optimalizált lelettel is kipróbáltuk ugyanezt, az eredményeket kézzel vizsgáltuk át ismét. A kísérletben 441 hibariasztást kaptunk. Ezekből az átvizsgálás során azt szűrtük le, hogy 11 esetben történt indokolatlan jelölés, és további 6 esetet indokoltan, de rosszul javított. A szótárak bővítésével természetesen ezen hibás találatok és hibák nagy része is kiküszöbölhető lenne.

RQ2 válasz: Szűk területünkön belül az automatikus javítás jelentősen, több mint 30%-al több valós hibát tárt fel a kézi ellenőrzésnél, a javítás minőségének különösebb romlása nélkül.

Végezetül módszerünk használhatóságát egy felületen is demonstráljuk. Ennek képernyőképe látható a 4. ábrán. A felületen bal oldalon egy kitalált lelet számos elírást tartalmazó szövege látható. Ezeket, amint az ábrán is látható, módszerünk helyesen detektálta, piros színnel jelölte ki. A jobb oldali ablakban láthatjuk a módszerünk által összeállított automatikus javítást, az ábrán ezek mindegyike helyes javítás. Amennyiben hibás javítást tapasztalnánk, a javításokat az alsó panelen tudjuk elbírálni. Egy javítást elfogadhatunk, elvethetünk amennyiben a hiba nem volt valós, vagy akár módosíthatunk is. Ez utóbbihoz nagyban hozzájárul, hogy módszerünk egy priorizált listát szolgáltat, amelyre szintén láthatunk példát az ábrán. Amennyiben a listában sem találjuk meg a kívánt javítást, abban az esetben újat is beírhatunk a szövegdobozba. Az ábrán szereplő javaslatok módosítás nélkül, automatikus javításként álltak elő.

A felület jelenleg a módszer hibáinak javítása szempontjából hasznos, hiszen könnyen kipróbálhatók benne különleges megfogalmazású szóalakok is. A jövőben azonban akár hibák gyors kézi annotációjában is segíthet, elég adatot kinyerve megtalálhatjuk a tipikusan nem kezelt hibákat, bővíthetjük a szótárakat, illetve akár még egy esetleges gépi tanulási módszer bemenetét is alkothatják.



4. ábra: Automatikus helyesírásjavító felületünk radiológiai gerinc leletek ellenőrzésére: Bal oldal: eredeti szöveg, Jobb oldal: javítások, Alsó panel: kezelés további lehetséges találatokkal

Fontos megjegyezni, hogy módszerünk kizárólag magyar nyelvű radiológiai gerinc leletek elírásjavítására lett optimalizálva. Az nem jelenthető ki, hogy más területeken is garantáltak lennének a hasonló eredmények, vagy a módszer sikere. Jelen területünk viszonylag szűkös szókészlettel dolgozó orvosi szakszöveg. Feltételezzük, hogy hasonló, behatárolható orvosi területeken új szótárakkal hasonló eredmények produkálhatók, ám az általános szöveg elemzésében már sokkal rosszabbul teljesít megközelítésünk, hiszen szakszavakat priorizál. Az általános elírásjavítókkal szemben azonban nagy előny, hogy valóban támaszkodhatunk a terület szokásaira, ezzel sokkal pontosabb eredményeket adva a területen. Motiváló példaként kipróbáltuk, hogy egy ismert szövegszerkesztő magyar nyelvű ellenőrzője a 4. ábra szövegében ugyan detektál minden hibásan írt szót, de ezek egyikére sem adja meg a helyes javítást, valamint 5 további szót is hibásnak jelöl, amelyek a leletekben azonban nagyon gyakran előfordulnak.

4. Kapcsolódó kutatások

Az egészségügyi rendszerben a leletezés túlnyomórészt szabad megfogalmazású formában történik. Ez egyfelől hasznos a beteg állapotának szabatos és pontos leírása szempontjából, másfelől viszont jelentősen megnehezíti a leletekből történő információkinyerést. A szabad megfogalmazású leletek használatának egy további hátránya a bonyolult információkinyerésen kívül, hogy ez a leletezési forma sokkal több hibalehetőséget rejt magában a strukturált leletezéssel szemben. Hibák alatt itt elsősorban elgépelésből, vagy helytelen hangfelismerésből származó elírási hibákra kell gondolni. Egy-két elírt szó a leletben nem tűnik nagy problémának, azonban ha az az elírás pont egy kulcsfontosságú részen történik, a diagnózist vagy esetleg a kezelési formát érintve, akkor az komoly következményekkel járhat a páciens egészsége szempontjából. Mindezek mellett információkinyerés szempontjából is jelentőséggel bírnak az elírt szavak, ugyanis több rendszernél is bevett eljárás a szavak ontológiákhoz történő hozzárendelése és kódolása, ez pedig pontos szóegyezéses alapon működik elsősorban, így egy elírt szó információvesztést jelent (Tolentino és mtsai, 2007).

Kukich hibajavításokkal kapcsolatos átfogó tanulmányában (Kukich, 1992) három fő elírási kategóriát állapított meg, melyekkel a hibajavító rendszereknek meg kell birkózniuk: nem létező szó felismerése (nonword error detection), izolált szójavítás (isolated word error correction), kontextusfűggő hibajavítás. A nem létező szó felismerési technikák két fő csoportja az n-gram analízisen alapuló módszer, mely során szokatlan karaktermintázatok alapján becsülik a hiba lehetőségét, illetve a szótár alapú módszer, mely során egy nagy szótárral való egyezéskeresés történik. Az izolált szójavító algoritmusok többsége valamilyen szerkesztési távolság alapján ajánl javításokat. Egy tanulmány szerint az elírások 80%-át a következő hibák teszik ki: szóhoz hozzáírt betű, szóból kihagyott betű, a szó egy betűjének más betűvel történő helyettesítése, illetve a szó két betűjének felcserélése (Damerau, 1964). A kontextusfüggő hibák esetén egy helyesen írt szó egy másik helyesen írt, nem oda illő szóval van helyettesítve. Az ilyen hibák szűrésére statisztikai alapú nyelvi modelleket alkalmaznak elsősorban.

A nemzetközi szakirodalomban több ízben is találhatunk példákat orvosi szöveg javítására, így például Tolentino és szerzőtársai vakcinabiztonságról szóló jelentésekben végeztek hibajavítást, mely során szótárat készítettek általános angol szavak és területspecifikus szavak gyűjteményéből, majd egy szerkesztési távolságon alapuló módszert alkalmaztak a hibák szűrésére (Tolentino és mtsai, 2007). Crowell és munkatársai egy szabad forrású szoftvert alkalmaztak egy orvosi portálhoz intézett lekérdezésekben található elírások szűrésére, szerintük a szavak között statisztikailag kimutatható, hogy melyiket milyen gyakorisággal használják keresésre a felhasználók. Ennek alapján újrarendezték a hibajavító szoftver javaslatait a szavak keresőmotorba írásának gyakorisága alapján, ennek hatására jelentős javulást tapasztaltak módszerük pontosságában (Crowell és mtsai, 2004). Mykowiecka és Marciniak bigram alapú nyelvi modellt használtak lengyel nyelvű mammográfiai leletek automatikus helyesírás-javítására (Mykowiecka és Marciniak, 2007). A szerzők saját maguk építettek szótárt, az elírt szavak több, mint 90%-át pontosan tudták javítani, azonban a szótárukban nem szereplő, helyesen írt szavak több, mint felét az algoritmus helytelenül módosította. Patrick és szerzőtársai egy trigram nyelvi modellel rangsorolt, szerkesztési távolságon alapuló rendszert dolgoztak ki kórházi leletekben található elírások javítására. A modellhez a szótárat külső forrásokból, illetve javított leletek szövegéből építették (Patrick és mtsai, 2010). Ruch és szerzőtársai névelemfelismerésen alapuló rendszert fejlesztettek francia nyelvű leletekben található elírások javítására (Ruch és mtsai, 2003). Megvalósításukban a névelemfelismerő rendszer megtalálta a szövegben előforduló neveket, melyeket a hibajavító algoritmus figyelmen kívül hagyott. Módszerükkel a szerzők a hibás pozitív detektálásokat jelentős mértékben csökkentették (~23%-ról ~3%-ra). Kenneth és szerzőtársai noisy channel módszeren (Kernighan és mtsai, 1990) alapuló hibajavító algoritmust alkalmaztak orvosi leletekben, receptekben és allergiára vonatkozó bejegyzésekben található elírások javítására. A programhoz kiterjedt szótárat építettek változatos forrásokból, valamint névelemfelismerő rendszer segítségével a helyesen írt nevek hibás javítását is kiküszöbölték (Lai és mtsai, 2015).

A magyar nyelvű klinikai szövegekben tapasztalható elírások kezelésére korábban már folytak kutatások. Siklósi és szerzőtársai több ízben fejlesztettek klinikai szövegek rendezésére, illetve a szövegekben tapasztalható elírások automatikus javítására rendszert, mely a javítás során figyelembe vette a szövegkörnyezetet, valamint képes volt az egybeírások kezelésére is (Siklósi és mtsai, 2013a,b, 2012). A szerzők egy másik kutatásukban ékezet nélküli szavak ékezetesítésére fejlesztettek statisztikai gépi fordításon alapuló rendszert, mely az esetek 99%-ában helyes ékezetes alternatívát javasolt (Novák és Siklósi, 2015, 2016).

Az általunk fejlesztett hibajavító algoritmus a javaslatok kialakításához több alegységre támaszkodik. Az egyik ilyen egység a szabadforrású Hunspell helyesírásjavító szoftver, melyet számos népszerű nemzetközi IT cég használ rendszereiben (Firefox, Chrome, Libre Office, Photoshop, Eclipse), valamint akadémiai berkeken belül is előszeretettel alkalmazzák, változatos nyelvekre adaptálva, mint például angol (Bettenburg és mtsai, 2011; Al-Hussaini, 2017), arab (Zerrouki és Balla, 2009), vagy esperanto (Blahuš, 2009).

Kifejezetten radiológiai gerincleletekre optimalizált elírásjavító tudomásunk szerint nem született még, sem magyar, sem külföldi forrásoktól. Módszerünk a terület szűkös szókészletét és specifikusságát kihasználva javít elírásokat. Munkánk esettanulmányként szolgálhat más hasonló területek javításai számára is, nem kizárólag klinikai környezetben.

5. Összegzés

Írásunkban elírásjavító módszerünket mutattuk be, amely a Hunspell elemzővel és radiológus részvételével valós leletekből összeállított szakszótárak segítségével azonosít elírásokat magyar nyelvű radiológiai gerincleletek szövegében, és automatikus javítást képes megvalósítani. Bemutattuk technikánk lépéseit, majd beláttuk, hogy az automatikus elírásjavítás jelentős mértékben pozitívan befolyásloja a szöveg gépi értelmezhetőségét, saját detektálási módszerünkön 0.43% javulást, azonosításunkban pedig a korábban azonosíthatatlan entitások több mint 20%-át sikerült azonosítani. Eredményeinket 882 leleten értékeltük ki, kiértékelésünk alapján módszerünk 38%-al több hibát tártunk fel, javításunk pedig csak néhány esetben volt téves. A módszer kézenfekvő jövőbeli felhasználása a leletek gépi értelmezésének javítása, de a rendszer akár a leletezés során, valós idejű javítások megvalósítására is alkalmas lehet, ezzel gyorsítva a leletek átnézési idejét és javítva azok végső minőségét. Bemutattunk egy felületet, amely alkalmas jellemző hiányosságok feltérképezésére is, csakúgy mint a hibák gyors annotációjára. Ezzel a jövőben a technika tovább finomítható.

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

Al-Hussaini, L.: Experience: Insights into the benchmarking data of hunspell and aspell spell checkers. J. Data and Information Quality 8(3-4), 13:1–13:10 (jun 2017)

Bettenburg, N., Adams, B., Hassan, A.E., Smidt, M.: A lightweight approach to uncover technical artifacts in unstructured data. In: 2011 IEEE 19th International Conference on Program Comprehension. pp. 185–188 (June 2011)

Blahuš, M.: Morphology-aware spell-checking dictionary for esperanto. In: Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2009. pp. 3–8. Masaryk University, Brno (2009)

Crowell, J., Zeng-Treitler, Q., Ngo, L., Lacroix, E.M.: A frequency-based technique to improve the spelling suggestion rank in medical queries. Journal of the American Medical Informatics Association : JAMIA 11, 179–85 (05 2004)

Damerau, F.: A technique for computer detection and correction of spelling errors. Commun. ACM 7, 171–176 (03 1964)

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

Kernighan, M., Church, K., Gale, W.: A spelling correction program based on a noisy channel model. pp. 205–210 (01 1990)

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)

Kukich, K.: Techniques for automatically correcting words in text. ACM Comput. Surv. 24, 377–439 (12 1992)

Lai, K., Topaz, M., Goss, F., Zhou, L.: Automated misspelling detection and correction in clinical free-text records. Journal of Biomedical Informatics 55 (04 2015)

Mykowiecka, A., Marciniak, M.: Domain–Driven Automatic Spelling Correction for Mammography Reports, vol. 35, pp. 521–530 (04 2007)

Novák, A., Siklósi, B.: Automatic diacritics restoration for hungarian. In: EMNLP. p. 2286–2291. The Association for Computational Linguistics, The Association for Computational Linguistics (2015)

Novák, A., Siklósi, B.: Ékezetek automatikus helyreállítása magyar nyelvű szövegekben, p. 49–58. Szegedi Tudományegyetem, Szeged (2016)

Patrick, J., Sabbagh, M., Jain, S., Zheng, H.: Spelling correction in clinical notes with emphasis on first suggestion accuracy. In: 2nd Workshop on Building and Evaluating Resources for Biomedical Text Mining. pp. 1–8 (2010)

Ruch, P., Baud, R., Geissbühler, A.: Using lexical disambiguation and named-entity recognition to improve spelling correction in the electronic patient record. Artificial intelligence in medicine 29, 169–84 (09 2003)

Siklósi, B., Novák, A., Prószéky, G.: Context-Aware Correction of Spelling Errors in Hungarian Medical Documents, p. 248–259. No. Lecture Notes in Computer Science 7978, Springer Berlin Heidelberg (2013a)

Siklósi, B., Novák, A., Prószéky, G.: Helyesírási hibák automatikus javítása orvosi szövegekben a szövegkörnyezet figyelembevételével. p. 148–158. Szegedi Tudományegyetem, Szeged (2013b)

Siklósi, B., Orosz, G., Novák, A., Prószéky, G.: Automatic structuring and correction suggestion system for hungarian clinical records. In: 8th SaLTMiL Workshop on Creation and Use of Basic Lexical Resources for Less-resourced Languages. p. 29–34 (2012)

Tolentino, H., Matters, M., Walop, W., Law, B., Tong, W., Liu, F., Fontelo, P., Kohl, K., Payne, D.: A umls-based spell checker for natural language processing in vaccine safety. BMC medical informatics and decision making 7, 3 (02 2007)

Zerrouki, T., Balla, A.: Implementation of infixes and circumfixes in the spellcheckers. In: In Proceedings of the Second International Conference on Arabic Language Resources and Tools. pp. 61–65 (2009)

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)