Hibrid előfeldolgozó algoritmusok morfológiailag komplex nyelvek és erőforrás szegény domainek hatékony feldolgozására

PhD disszertáció tézisfüzete

Orosz György

Pázmány Péter Katolikus Egyetem
Információs Technológiai és Bionikai kar
Roska Tamás Műszaki és Természettudományi Doktori Iskola

Témavezető:
Prószéky Gábor, DSc

Budapest, 2015

1. Bevezetés

A modern nyelvtechnológiai alkalmazások szerves részét képezik mindennapi életünknek. Ilyen eszközök például azok, amik segítenek dokumentumaink helyesírásának vizsgálatában, idegen nyelvű források fordításában és megértésében, illetve az interneten található információk visszakeresésében. A szövegfeldolgozás a nyelvtechnológia azon ága, amelynek célja digitalizált szövegek automatikus elemzése. Az ilyen feladatokhoz szükséges nyelvi előfeldolgozást legtöbbször több lépésben oldhatjuk meg: először szó- és mondathatárok megállapítása szükséges, majd a szavak morfológiai elemzése és egyértelműsítése következhet, végül pedig a mondatok szintaktikai analízise, illetve a szöveg szemantikai értelmezése végezhető el. A gyakorlatban az egyes részfeladatokat végző komponensek egymásra épülnek. Bár ezen eszközök nem mindegyike érhető el minden nyelvre, két előfeldolgozó lépés mégis elengedhetetlen a szövegek magasabb szintű kezeléséhez. Először a tokenek és a mondatok azonosítása szükséges, hiszen ezek az entitások az alapegységei a nyelvtechnológiai elemző rendszereknek. Ezen túl, gyakran nélkülözhetetlen még a szavak szófaji címkéinek és szótöveinek meghatározása is.

A legtöbb esetben az első részfeladat megoldottnak tekinthető, mivel a létező alkalmazások nagy pontossággal képesek szövegek részekre bontására. Ennek ellenére, számos olyan nyelvterület létezik, amelyekre a jelenleg elérhető eszközök nem nyújtanak kielégítő eredményt.

Az utóbbi húsz év során számos nagy pontosságú szófaji egyértelműsítő eszköz készült, melyek a legtöbb élő nyelvre elérhetőek. A gyakorlatban azonban ezek legtöbbször adatvezérelt módszerekre épülnek, ezért teljesítményük nagyban függ a használt tanítóanyagtól. Egy másik probléma, hogy a terület kutatásában az élenjáró eljárások elsődleges célja mindig is az angol nyelv elemzése volt, ezért a létrejött algoritmusok sokszor nem képesek kezelni a morfológiailag komplex nyelvek által okozott nehézségeket. Így például az agglutináló nyelvek esetén (az angollal szemben) a szavak szófajainak megállapítása önmagában nem elégséges. Az elemzési lánc

további komponensei a teljes morfoszintaktikai címke illetve a szótő ismeretét is igénylik. Így megállapítható, hogy a mai nyelvtechnológia nem rendelkezik olyan algoritmusokkal, amelyek morfológiailag gazdag nyelvek esetén megfelelően működnének, továbbá kevés erőforrással rendelkező doménekre is hatékonyan alkalmazhatóak lennének.

Következésképpen, dolgozatunk célja kettős. Először olyan morfológiai egyértelműsítő eljárásokat vizsgáltunk, melyek megfelelően kezelik az agglutináló nyelvek okozta problémákat, mindazonáltal egyszerű domén adaptációs feladatok esetén is alkalmazhatóak. Másodsorban pedig olyan módszerekkel foglalkoztunk, melyek kevés erőforrással rendelkező domének elemzésére is alkalmasok.

Első lépésben azt vizsgáltuk, hogy **hogyan lehetséges létező szófaji címkéző eljárásokat a teljes morfológiai egyértelműsítés feladatára használni, úgy, hogy azok képesek legyenek kezelni az agglutináló nyelvek tipikus nehézségeit, továbbá alkalmazhatóak maradjanak egyszerűbb domén adaptációs feladatokra is**. Így létrehoztunk egy magas pontossággal rendelkező lemmatizáló algoritmust, amire épülve kifejlesztettünk egy teljes egyértelműsítő rendszert (PurePos). Ezt követően megvizsgáltuk még, hogy **miként lehet morfológiai annotáló rendszerek pontosságát növelni kombinációs sémák alkalmazásával**. Kifejlesztettünk egy olyan morfológiai egyértelműsítő rendszereket kombináló algoritmust, ami agglutináló nyelvekhez illeszkedő jellemzőket használ, így teljesítménye meghaladja más létező eljárásokét magyar nyelvű szövegen esetén.

A teljes egyértelműsítés feladatán túl, a címkéző rendszerek gyakorlati alkalmazása is kiemelkedő jelentőséggel bír. Ennélfogva kutatásunkban tanulmányoztuk még azt, hogy **miként lehetséges beszélt nyelvi átiratokhoz olyan morfológiai annotáló eszközt létrehozni, mely a szakterületi kutatók munkáját képes segíteni**. Bemutattunk egy olyan eljárást, mely a beszélt nyelvi lejegyzéseket nagy pontossággal képes morfológiai annotációval ellátni, illetve ismertettünk egy olyan módszert is, amely automatikusan képes megbecsülni gyermeknyelvi szövegek morfoszintaktikai komplexitását.

Írásunk harmadik részében elektronikus orvosi feljegyzések előfeldolgozásával foglalkoztunk. Mindenekelőtt azt vizsgáltuk, hogy **miként lehetséges megfelelő szó- és mondatrabontó eljárásokat létrehozni létező algoritmusok továbbfejlesztésével**. Bemutattunk egy olyan hibrid eszközt, mely szabályalapú komponenseken túl felügyelet nélküli gépi tanulásra építve azonosítja a klinikai rekordok szavait és mondatait. Ezt követően tanulmányoztuk még a morfológiai egyértelműsítés kérdését az orvosi szövegek tekintetében. Megvizsgáltuk, hogy **milyen tényezők okozhatják egy klinikai dokumentumokat feldolgozó morfológiai annotáló alkalmazás legfőbb nehézségeit** illetve, hogy **a PurePos rendszer miként alkalmazható a doménre**. A klinikai szövegek speciális tulajdonságait kihasználva, számos olyan domén adaptációs eljárást készítettünk, amelyek jelentős mértékben képesek javítani a felhasznált alaprendszer hibáin.

2. Felhasznált módszerek

A morfológiai egyértelműsítő eljárásokat a Szeged Korpuszt [22] használva alkottuk meg, de az algoritmusok egy részét az ó- és közép-magyar szövegeken [12] is kiértékeltük. A beszélt nyelvi átiratokhoz készült alkalmazások létrehozásához a MONYEK korpuszt [2] használtuk, míg a klinikai rekordokat kezelő eljárásokhoz nem létezett etalon szöveggyűjtemény, így azt mi készítettük el.

A bemutatott rendszerek legtöbbször hibrid eljárásokat használnak. Így egyfelől építettünk morfológiai elemzők kimenetére, míg másrészről gépi tanulást használó algoritmusokat is alkalmaztunk.

A morfológiai elemző rendszerek közül legtöbbször a Humort [23, 24, 25] (vagy annak valamely adaptál verzióját [12, 26, 16]) használtuk, de a `magyarlanc` [27] megfelelő komponensét is alkalmaztuk.

Gépi tanulást használó eljárásaink legtöbbször rejtett Markov modellezést [28, 29] használnak, mindazonáltal különösen építettünk két közismert szófaji egyértelműsítő (a HunPos [30] és a TnT [31]) módszereire. Ezeken kívül

alkalmaztunk még szuffixum fákat, n-gram modellezést, illetve általános interpolációs technikákat is. Az ismertetett kombinációs algoritmust példány alapú tanulásra [32] építettük, amit a Weka [33] eszköz implementációjában értünk el. Végezetül a klinikai rekordokat mondatrabontó algoritmus a Dunning által bemutatott [34] majd Kiss és Strunk által továbbfejlesztett [35] kollokációs metrikára épül.

Az eszközök teljesítményét a tudományágban bevett, sztenderd módszerekkel mértük. Elsősorban szó- és mondat-szintű pontosságot számoltuk a szófaji és morfológiai egyértelműsítő rendszerek esetén. Azonban ezeket néhány esetben úgy módosítottuk, hogy a kiértékelés során nem vettük figyelembe a központozást jelölő tokenek annotációját. Az így mért pontossági értékek között számos alkalommal hibaráta csökkenést is számoltunk. Vizsgáltuk még az egyes rendszerek pontossága közti eltérések statisztikai szignifikanciáját is, amihez a Wilcoxon féle előjeles rangszámösszeg próba a SciPy [36] eszközben elérhető implementációját alkalmaztuk.

Az egyszerű osztályozási problémákhoz osztály szintű pontosságot, fedést és F-értékeket számoltunk, illetve figyelembe vettük még a módszerek teljes címkekészletre vetített pontosságát is. Végezetül a számszerű értékek összevetését átlagos relatív hibaráta [37] illetve Pearson korrelációs együtthatója [37] számolásával végeztük el.

3. Új tudományos eredmények

I. Hatékony morfológiai egyértelműsítő algoritmusok

A morfológiai egyértelműsítés egy olyan összetett feladat, amely a szófaji címkék meghatározásából és a szavak töveinek azonosításából áll. Míg az első részfeladatot a szakirodalom megoldottnak tekinti, addig az utóbbi területen sokkal kevesebb fejleményről számolhatunk be.

I.1. TÉZIS. Kidolgoztam egy olyan metódust, ami agglutináló nyelvek, így magyar esetén is nagy pontossággal képes szavak lemmáit azonosítani. Az eljárás a tanítóanyagban látott szavakon túl az ún. ismeretlen szóalakokat is képes hatékonyan kezelni, amihez a morfológiai elemző lehetséges elemzésein kívül a tanítóanyagból épített statisztikai modellekre is épít. Mérésekkel kimutattam, hogy a módszer magyar nyelv esetén kimagasló pontossággal bír.

A tézishez kapcsolódó publikációk: [18, 17, 10, 8]

A létrehozott algoritmus két fázisban végzi a szótövesítést. Első lépésben meghatározza a szavak lehetséges lemmáinak halmazát, amihez felhasználja a morfológiai elemző javaslatait illetve egy ismeretlenszó-elemző kimenetét is. Ezt követően a jelölteket (l) rangsorolja a szóalak (w) és az előzetesen kalkulált morfoszintaktikai címke (t) függvényében számolt valószínűségi értékek alapján:

$$S(l|w,t) = P(l)^{\lambda_1} P(l,t|w)^{\lambda_2} \tag{1}$$

Az így kapott algoritmus egy egyszerű szótő gyakorisági eloszlást és egy szóvég-alapú valószínűségi modellt kombinálva határozza meg a legmegfelelőbb lemmát. A bemutatott eljárás az egyes összetevők javaslatainak helyességét vizsgálva hangolja a λ_i paraméterek értékét.

Méréseimmel megmutattam, hogy az új szótövező algoritmus kiemelkedő pontossággal bír magyar nyelv esetén.

—•—

I.2. TÉZIS. Létrehoztam egy olyan hibrid morfológiai egyértelműsítő eszközt (PurePos[1]), mely hatékonyan alkalmazható morfológiailag komplex és nyelvi erőforrásokban szegény nyelvek esetén. Az algoritmus statisztikai eljárásokra támaszkodva, morfológiai elemző

[1] A bemutatott rendszer szabadon elérhető a https://github.com/ppke-nlpg/purepos címen.

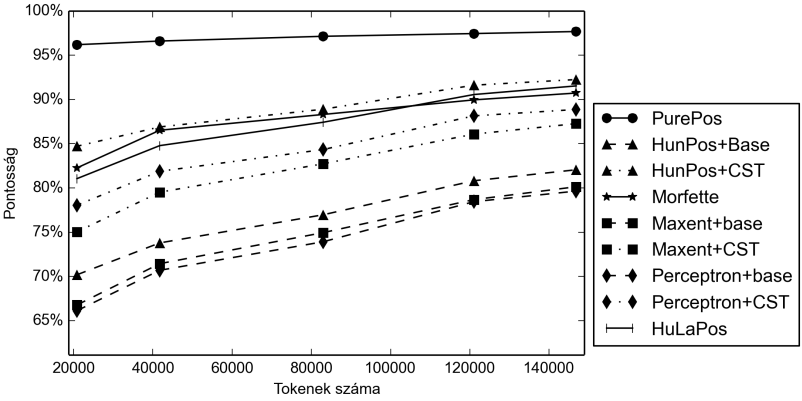
integrált alkalmazásával és szabály alapú komponensek használatával hatékony egyértelműsítést tesz lehetővé. Az eszköz a szavak lemmáinak meghatározását a I.1 tézisben ismertetett módszerrel végzi. Megmutattam, hogy az eljárás magyar nyelv esetén state-of-the-art teljesítménnyel rendelkezik. Ismertettem, hogy a rendszer architektúrája lehetőséget nyújt domén specifikus szabályok hatékony alkalmazására, illetve méréseimmel alátámasztottam, hogy a létrehozott algoritmus kiemelkedő pontossággal bír kevés tanítóanyag használata esetén is.

A tézishez kapcsolódó publikációk: [18, 17, 10, 8]

1. ábra A hibrid morfológiai egyértelműsítő algoritmus architektúrája

A rendszer architektúráját (ld. 1. ábra) úgy alakítottam ki, hogy a statisztikai modulokon túl szimbolikus komponensekkel is hatékonyan tudjon együttműködni. Ily módon a címkék és szótövek azonosítása több lépésben történik: az egyértelműsítés alapját egy morfológiai elemző képezi, melyet több lépésben sztochasztikus algoritmusok követnek. A felhasznált trigram-alapú metódust úgy adaptáltam, hogy azok hatékonyan működjenek morfológiailag komplex nyelvek esetén is. A szófaji címkézést követően, az elemzés utolsó fázisában történik a lemmák meghatározása a I.1 tézisben bemutatott módon.

Méréseimmel megmutattam, hogy a bemutatott egyértelműsítő algoritmus kimagasló pontossággal bír magyar nyelv esetén. Ehhez a PurePos rendszert a Szeged Korpuszon tanítottam és teszteltem. A rendszer ezen szövegek esetén 96,27%-os szószintű pontosságot nyújt, mely meghaladja más szabadon elérhető eszközök teljesítményét. Vizsgáltam még az eszköz alkalmazhatóságát olyan esetekben, amikor csak kevés tanítóanyag áll rendelkezésre. Kimutattam, hogy a PurePos ilyenkor is nagy



2. ábra Morfológiai egyértelműsítő algoritmusok tanulási görbéje

pontossággal alkalmazható (vö. 2. ábra). Ismertettem még az eljárás egy olyan alkalmazását, ahol a hibrid komponenseinek köszönhetően jelentős mértékben sikerült javítani az eredeti elemzőlánc pontosságán, így gyorsítva a manuális annotációs folyamatot.

—●—

Bár a I.2 tézisben bemutatott algoritmusok magas pontossággal rendelkeznek, megmutattam, hogy ezek teljesítménye tovább növelhető más rendszerekkel való kombinációval.

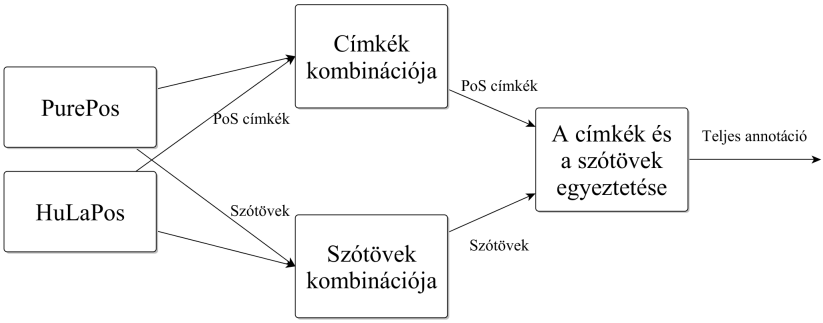
I.3. TÉZIS. Létrehoztam egy olyan módszert, mely morfológiai egyértelműsítő rendszerek kombinációjával hatékony növeli a címkézés pontosságát magyar nyelv esetén. A kidolgozott eljárás újdonsága, hogy külön modulban végzi a lemmák és morfoszintaktikai címkék azonosítását, majd azok kimenetét egyesítve határozza meg a morfológiai annotációt. A módszer példány alapú tanulásra épül és az egyes alrendszereket keresztvalidáción keresztül tanítja.

Méréseimmel alátámasztottam, hogy az ismertetett módszer jelentős mértékben képes növelni a címkézési feladat pontosságát.

A tézishez kapcsolódó publikációk: [20, 9, 5]

Első lépésként, kidolgoztam egy új metrikát (OER), amellyel egyértelműsítő rendszerek hibáinak különbözőségét vizsgáltam. Ezt felhasználva megmutattam, hogy a HuLaPos rendszer tipikus hibái számottevően eltérnek a PurePos-étól.

Ezt követően olyan jellemzőhalmazokat hoztam létre, melyek morfológiailag komplex nyelvek esetében magas pontossággal használhatóak. Majd kidolgoztam egy olyan eljárást (3. ábra) melyben két külön komponens választja ki a szófaji címkéket, illetve a lemmákat. A kombinációs rendszer moduljai keresztvalidáció segítségével tanítják az első szintű osztályozókat, amik példány alapú tanulásra épülnek.



3. ábra A kombinációs rendszer működése

Méréseimmel megmutattam, hogy az új algoritmus a PurePos hibáinak mintegy 28,90%-át javítja.

II. Morfoszintaktikai komplexitás automatikus becslése morfológiai egyértelműsítő algoritmusok alkalmazásával

A morfológiai komplexitás mérése fontos eszköze a nyelvfejlődést mérő nyelvészeti kutatásoknak. Ezt agglutináló nyelvek esetén a megnyilatkozások átlagos morfémában mért hosszával (MLUm) számolják. Míg angolra és más morfológiailag nem összetett nyelvre léteznek automatikus algoritmusok a feladat megoldására, addig magyarra (és egyéb agglutináló nyelvekre) ezek közvetlenül nem alkalmazhatóak. Ezekben az esetekben a megnyilatkozások hosszának mérése csak időigényes manuális számolással végezhető el.

Dolgozatomban megmutattam[2], hogy a PurePos rendszer egy megfelelő morfológiai elemzővel kiegészülve adekvát alapja egy automatikus morfémaszám-becslő eljárásnak.

II.1. TÉZIS. Létrehoztam egy hibrid morfológiai egyértelműsítő láncot magyar gyermeknyelvi beszédátiratok nagy pontosságú elemzésére. Az algoritmus alapját az I.2 tézisben ismertetett rendszer képezi, amelyet a beszélt nyelv címkézéséhez szükséges szabályokkal adaptáltam. Méréseimmel igazoltam, hogy a létrejött elemzési lánc teljesítménye megközelíti az általános nyelvi címkézők eredményességét.

A tézishez kapcsolódó publikációk: [2, 4]

Mivel a bemutatott morfológiai egyértelműsítő rendszer a Humor elemzőre épül, azt a beszélt nyelvben tipikus jelenségekkel egészítettem ki. Ezt követően a PurePos rendszert további szabály alapú eljárásokkal adaptáltam a doménhez.

[2] A morfológiai komplexitás becslésének feladatát Jelencsik-Mátyus Kingával együtt végeztem. A korpusz manuális címkézése, az annotálás útmutató kidolgozása közös munka eredménye. Az MLUm becslés nyelvészeti alapvetései a társszerző érdeme, míg a folyamat algoritmizálása önálló eredmény.

A lánc pontosságának méréséhez létrehoztunk egy 1000 megnyilatkozásból álló etalon korpuszt, mely a MONYEK [2] adatbázis részét képezik. Az annotáció folyamatához kidolgoztuk egy az eddigiektől eltérő, beszélt nyelvre adaptált címkekészletet, majd létrehoztunk egy annotálási útmutatót is.

A bemutatott szabály-alapú és sztochasztikus technikák alkalmazásával 96%-os szószintű pontosságot értem el, mely megközelíti az általános nyelvi egyértelműsítőkét. Vizsgálataimmal alátámasztottam, hogy a PurePos rendszer nagy pontossággal használható magyar nyelvű beszédátiratok elemzésére.

— • —

II.2. TÉZIS. Kifejlesztettem egy olyan új eljárást, amely magyar nyelvű beszédátiratok morfoszintaktikai összetettségét képes automatikusan becsülni. Az algoritmus a II.1 tézisben bemutatott elemzőláncra épülve számolja a megnyilatkozások morfémában mért hosszát. Méréseimmel kimutattam, hogy a módszer megfelelően képes helyettesíteni az időigényes manuális számolást.

A tézishez kapcsolódó publikációk: [2, 4]

Az algoritmus a szavak morfoszintaktikai annotációjára épülve összegzi a megnyilatkozások morfémáit. A Humor elemző által ismert szavakat annak használatával morfémákra bontja, míg az ismeretlen szóalakokhoz a morfoszintaktikai címke alapján készít becslést.

Az módszer tökéletesítéséhez létrehoztunk egy morfémaszámokat is tartalmazó etalon korpuszt. Megmutattam, hogy az automatikus módszer ezen az adathalmazon 0,9901 korrelációs értékkel bír, míg az algoritmus átlagos relatív eltérése is csupán 4,49%. Méréseimmel bebizonyítottam, hogy az eljárás alkalmas az időigényes manuális morfémaszámolás kiváltására.

III. Hatékony előfeldolgozó algoritmusok egy kevés erőforrással rendelkező zajos doménhez

Napjainkban egyre több elektronikusan rögzített dokumentum keletkezik klinikai környezetben, melyek nagy mennyiségű eddig el nem érhető közvetett tudást reprezentálnak. Mivel létrehozásuk során nem fordítanak kellő figyelmet a szövegek struktúrájának kialakítására és a helyesírási normák betartására, így azok feldolgozása gyakran nem lehetséges létező eszközök közvetlen alkalmazásával. Bár angol nyelvre számtalan megoldás született az évek során, a magyar (és más morfológiailag összetett) nyelvű orvosbiológiai szövegek elemzése egy alig vizsgált terület.

III.1. TÉZIS. Létrehoztam egy olyan hibrid eljárást, mely magyar nyelvű klinikai rekordokat képes magas pontossággal mondatokra és szavakra bontani. A módszer alapját egy szabály-alapú szegmentáló algoritmus képezi, amelyet felügyelet nélküli gépi tanulással egészítettem ki. Méréseimmel alátámasztottam, hogy a hibrid rendszer által azonosított mondat- és szóhatárok kellően pontosak a gyakorlati alkalmazhatósághoz. Ezen túl kimutattam még, hogy a magyar nyelvre elérhető algoritmusok közül sem a szabályalapú, sem a gépi tanulást használó rendszerek nem alkalmasak orvosbiológiai szövegek tokenizálására és mondatokra bontására.

A tézishez kapcsolódó publikációk: [5, 14]

4. ábra A szegmentáló algoritmus részei

A bemutatott szegmentáló algoritmus alapját szabály-alapú, mintaillesztést használó algoritmusok képezik. Ezek bár magas pontossággal bírnak, fedésük alacsony, így ezt további heurisztikus eljárásokkal bővítettem. Megmutattam, hogy a módosított $\log \lambda$ szűrő nagy mértékben képes növelni a rendszer fedését. A teljesítmény javításához a szóalakok felszíni jegyein túl, a Humor morfológiai elemzéseit is felhasználtam.

A mérésekhez létrehoztam egy manuálisan javított korpuszt, amin az egyes rendszerek szó- és mondat-szintű pontosságát, fedését és a kombinált F-értéket is vizsgáltam. Kutatásomban a speciálisan magyar nyelvre fejlesztett rendszereket, illetve egy maximum entrópia módszeren alapuló általánosan használt eszközt is kiértékeltem.

Mérésekkel megmutattam, hogy a mondatrabontás feladatában a legtöbb elérhető rendszer 50%-os F-érték alatt teljesít. Ezzel szemben méréseimmel azt is alátámasztottam, hogy az új hibrid algoritmus mind a mondatokra mind pedig a szavakra bontás tekintetében 90% feletti F értéket produkál, így az alkalmas magyar nyelvű klinikai dokumentumok szegmentálására.

— • —

III.2. TÉZIS. Megmutattam, hogy az I.2 tézisben ismertetett rendszer, megfelelő adaptációs technikákkal kombinálva alkalmas orvosbiológiai szövegek elfogadható minőségű morfológiai egyértelműsítésére. Méréseimmel kimutattam, hogy az ismertetett szabály-alapú és statisztikai doménadaptációs módszerek jelentős mértékben javítanak a teljes elemzési lánc pontosságán.

A tézishez kapcsolódó publikációk: [16, 1]

Az eljárás a Humor morfológiai elemző egy bővített változatára és a PurePos egyértelműsítő rendszerre épül. Dolgozatomban feltártam az így kapott alaprendszer tipikus hibáit és az algoritmus számos hiányosságát orvosoltam doménspecifikus szabályok alkalmazásával.

A mérések elvégzéséhez létrehoztam egy etalon korpuszt, melynek morfológiai annotációját manuálisan javítottam. Megmutattam, hogy a közreadott rendszer szószintű pontossága (93,73%) jelentősen meghaladja az alapjául szolgáló eredeti rendszer teljesítményét (88,09%).

Ismertettem még, hogy a bemutatott klinikai dokumentumokat szegmentáló és egyértelműsítő eljárások hibái a rövidítések kezelésének nehézségeiből fakadtak. Így a jövőben indokolt lehet egy olyan módszer kidolgozása, mely a két feladatot egyszerre célozza meg.

4. Alkalmazási területek

Az ismertetett módszerek nyelvtechnológiai alapfeladatokra adnak megoldást, így ezek komplex feldolgozóláncok alapját képezhetik. A morfológiai címkéző algoritmusok (ld. I téziscsoport) széles körben használhatóak információkinyerési és szövegbányászati alkalmazásokban úgy mint névelemek azonosítása, kulcsszavak kinyerése vagy dokumentumok osztályozása. Ezeken túl, az egyértelműsítő eljárások alábbi gyakorlati alkalmazásairól van tudomásom:

1. Laki és tsai. átrendezés-alapú angol-magyar gépifordító-rendszert [7] építettek a PurePos rendszer használatával,

2. Novák és tsai. ó- és középmagyar szövegek morfológiai annotációjának elkészítéséhez használta [12] a közreadott címkéző eszközt,

3. Endrédy és tsai. [38] magyar nyelvre főnévi csoportokat azonosító algoritmust készített, mely tartalmazza az ismertetett morfológiai egyértelműsítő algoritmusokat,

4. Indig és Prószéky [39] egy kötegelt helyesírás-ellenőrző programban alkalmazta az eljárást, míg

5. Prószéky és tsai. egy pszicholingvisztikai indíttatású elemzőben hasznosították [40] a PurePos rendszer egyes komponenseit.

A II. téziscsoport magyar nyelvű beszédátiratok feldolgozásával foglalkozik. Az itt bemutatott módszer egy olyan speciális alkalmazása a PurePos rendszernek, mellyel ennek a doménnek a további vizsgálatát teszi lehetővé. A II.2. tézis morfémabecslő eljárása jól használható gyermeknyelvi szövegek esetén morfoszintaktikai komplexitás automatikus mérésére, kiváltva így az időigényes manuális kalkulációt. Továbbá Jelenics-Mátyus [41] egy a gyermekek nyelvi fejlődését vizsgáló kutatásban alkalmazta a közreadott eljárásokat.

Az utolsó téziscsoportban ismertetett algoritmusok zajos klinikai szövegek hatékony előfeldolgozását teszik lehetővé. A III.1. tézisben bemutatott eljárások hatékonyan képesek szavakra és mondatokra bontani magyar nyelvű orvosi szövegeket, ezáltal lehetővé téve az azokban kódolt információ kinyerését. Ezen kívül a III.2. algoritmusai elfogadható minőségű morfológiai annotációt készítenek a klinikai rekordokhoz, így azok mélyebb elemzését teszik lehetővé. A fenti eredményeimet egy folyamatban lévő projekt [42, 43] hasznosítja, mely klinikai dokumentumokban rejlő rejtett összefüggések feltárását célozza meg.

Köszönetnyilvánítás

„Az Úr ad nekem segítséget, aki az eget és a földet alkotta”
– Zsolt 121,2

Először is szeretném megköszönni témavezetőmnek, Prószéky Gábornak az évek során nyújtott megannyi segítséget és a folyamatos támogatását. Ugyanakkor, hálával tartozom Novák Attilának a konzultációkért és az értékes javaslatokért. Köszönöm Wenszky Nórának, hogy fáradhatatlanul igyekezett csiszolni az angol íráskészségemet és közreműködött e dolgozat hibáinak javításában.

Köszönöm még munkatársaimnak a mindig jó hangulatú és inspiráló szakmai beszélgetéseket, továbbá hálás vagyok nekik a feledhetetlen ebéd- és kávészünetekért is. Köszönet illeti a közös munkáért Laki Lászlót, Siklósi Borbálát, Indig Balázst és Mátyus Kingát. Hálás vagyok még a 314-es szoba lakóinak: Endrédy Istvánnak, Yang Zijian Győzőnek, Sass Bálintnak, Miháltz Mártonnak, Simonyi Andrásnak és Varasdi Károlynak a közös munkáért és a derűs légkörért.

Köszönettel tartozom a Pázmány Péter Katolikus Egyetemnek és az MTA-PPKE Magyar Nyelvtechnológiai Kutatócsoportnak hogy befogadtak és munkámat végig támogatták. Hála illeti az egyetem korábbi és jelenlegi vezetőit, Roska Tamás, Nyékyné Gaizler Judit és Szolgay Péter, akik lehetőséget biztosítottak tanulmányaim elvégzéséhez. Szeretnék még köszönetet mondani Hubay Katalinnak és Adorján Líviának, akik mindig türelemmel és megértéssel szervezték utazásaink problémás ügyeit.

A közreadott eredmények a TÁMOP 4.2.1.B – 11/2/KMR-2011–0002 és 4.2.2/B – 10/1–2010–0014 projektek részleges támogatásával jöttek létre.

Végül, de nem utolsó sorban szeretnék köszönetet mondani családomnak. Köszönöm feleségemnek, Jucusnak szüntelen türelmét, támogatását és a sok-sok bátorítást. Hálás vagyok szüleimnek és öcsémnek, Tominak, hogy tanulmányaim során mindig mellettem voltak és minden lehetséges módon segítettek.

Irodalomjegyzék

A szerző folyóirat publikációi

[1] **György Orosz**, Attila Novák, and Gábor Prószéky. Lessons learned from tagging clinical Hungarian. *International Journal of Computational Linguistics and Applications*, 5(1):159–176, 2014. ISSN: 0976-0962.

[2] Kinga Mátyus and **György Orosz**. MONYEK: morfológiailag egyértelműsített óvodai nyelvi korpusz. *Beszédkutatás – 2014*:237–245, 2014. ISSN: 1218-8727.

[3] Borbála Siklósi, Attila Novák, **György Orosz**, and Gábor Prószéky. Processing noisy texts in Hungarian: a showcase from the clinical domain. *Jedlik Laboratories Reports*, II(3):5–62, 2014. Péter Szolgay, editor. ISSN: 2064-3942.

A szerző könyvfejezet publikációi

[4] **György Orosz** and Kinga Mátyus. An MLU Estimation Method for Hungarian Transcripts. English. In Petr Sojka, Aleš Horák, Ivan Kopeček, and Karel Pala, editors, *Text, Speech and Dialogue*. Volume 8655, in Lecture Notes in Computer Science, pages 173–180. Springer International Publishing. ISBN: 978-3-319-10815-5.

[5] **György Orosz**, Attila Novák, and Gábor Prószéky. Hybrid text segmentation for Hungarian clinical records. In Félix Castro, Alexander Gelbukh, and Miguel González, editors, *Advances in Artificial Intelligence and Its Applications*. Volume 8265, in Lecture Notes in Computer Science, pages 306–317. Springer, Berlin Heidelberg, 2013. ISBN: 978-3-642-45114-0.

[6] **György Orosz**, László János Laki, Attila Novák, and Borbála Siklósi. Improved Hungarian Morphological Disambiguation with Tagger Combination. In Ivan Habernal and Václav Matousek, editors, *Text, Speech, and Dialogue*. Volume 8082, in Lecture Notes in Computer Science, pages 280–287. Springer, Berlin, Heidelberg, 2013. ISBN: 978-3-642-40584-6.

[7] László János Laki, **György Orosz**, and Attila Novák. HuLaPos 2.0 – Decoding Morphology. In Félix Castro, Alexander Gelbukh, and Miguel González, editors, *Advances in Artificial Intelligence and Its Applications*. Volume 8265, in Lecture Notes in Computer Science, pages 294–305. Springer Berlin Heidelberg, 2013. ISBN: 978-3-642-45113-3.

A szerző angol nyelvű konferencia publikációi

[8] **György Orosz** and Attila Novák. PurePos 2.0: a hybrid tool for morphological disambiguation. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2013)*. INCOMA Ltd. Shoumen, Hissar, Bulgaria, 2013, pages 539–545.

[9] **György Orosz**, László János Laki, Attila Novák, and Borbála Siklósi. Combining Language Independent Part-of-Speech Tagging Tools. In *2nd Symposium on Languages, Applications and Technologies*. José Paulo Leal, Ricardo Rocha, and Alberto Simões, editors. In OpenAccess Series in Informatics (OASIcs). Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Porto, 2013, pages 249–257.

[10] **György Orosz** and Attila Novák. PurePos – an open source morphological disambiguator. In *Proceedings of the 9th International Workshop on Natural Language Processing and Cognitive Science*. Bernadette Sharp and Michael Zock, editors. Wroclaw, 2012, pages 53–63.

[11] László Laki and **György Orosz**. An Efficient Language Independent Toolkit for Complete Morphological Disambiguation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. European Language Resources Association (ELRA), Reykjavik, Iceland, 2014, pages 26–31.

[12] Attila Novák, **György Orosz**, and Nóra Wenszky. Morphological annotation of Old and Middle Hungarian corpora. In *Proceedings of the ACL 2013 workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*. Piroska Lendvai and Kalliopi Zervanou, editors. Association for Computational Linguistics, Sofia, Bulgaria, 2013, pages 43–48.

[13] Borbála Siklósi, **György Orosz**, Attila Novák, and Gábor Prószéky. 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*. Guy De Pauw, Gilles-Maurice de Schryver, Mike L. Forcada, Francis M. Tyers, and Peter Waiganjo Wagacha, editors. Istanbul, 2012, pages 29–34.

A szerző egyéb konferencia publikációi

[14] **György Orosz** and Gábor Prószéky. Hol a határ? Mondatok, szavak, klinikák. In *X. Magyar Számítógépes Nyelvészeti Konferencia*. Attila Tanács, Viktor Varga, and Veronika Vincze, editors. Szegedi Tudományegyetem, Szeged, 2014, pages 177–187.

[15] **György Orosz** and Attila Novák. PurePos 2.0: egy hibrid morfológiai egyértelműsítő rendszer. In *X. Magyar Számítógépes Nyelvészeti Konferencia*. Attila Tanács, Viktor Varga, and Veronika Vincze, editors. Szegedi Tudományegyetem, Szeged, 2014, pages 373–377.

[16] **György Orosz**, Attila Novák, and Gábor Prószéky. Magyar nyelvű klinikai rekordok morfológiai egyértelműsítése. In *IX. Magyar Számítógépes Nyelvészeti Konferencia*. Attila Tanács and Veronika Vincze, editors. Szegedi Tudományegyetem, Szeged, 2013, pages 159–169.

[17] **György Orosz**. PurePos: hatékony morfológiai egyértelműsítő. In *VI. Alkalmazott Nyelvészeti Doktoranduszkonferencia*. Tamás Váradi, editor. Budapest, 2012, pages 134–139.

[18] **György Orosz**, Attila Novák, and Balázs Indig. Javában taggelünk. In *VIII. Magyar Számítógépes Nyelvészeti Konferencia*. Attila Tanács and Veronika Vincze, editors. Szeged, 2011, pages 336–340.

[19] László János Laki and **György Orosz**. HuLaPos2 – Fordítsunk morfológiát. In *X. Magyar Számítógépes Nyelvészeti Konferencia*. Attila Tanács, Viktor Varga, and Veronika Vincze, editors. Szegedi Tudományegyetem, Szeged, 2014, pages 41–49.

[20] László János Laki and **György Orosz**. Morfológiai egyértelműsítés nyelvfüggetlen annotáló módszerek kombinálásával. In *IX. Magyar Számítógépes Nyelvészeti Konferencia*. Attila Tanács and Veronika Vincze, editors. Szegedi Tudományegyetem, Szeged, 2013, pages 331–337.

[21] Borbála Siklósi, **György Orosz**, and Attila Novák. Magyar nyelvű klinikai dokumentumok előfeldolgozása. In *VIII. Magyar Számítógépes Nyelvészeti Konferencia*. Attila Tanács and Veronika Vincze, editors. Szeged, 2011, pages 143–154.

Egyéb hivatkozások

[22] Dóra Csendes, János Csirik, and Tibor Gyimóthy. The Szeged Corpus: A POS tagged and syntactically annotated Hungarian natural language corpus. In *Proceedings of the 5th International Workshop on Linguistically Interpreted Corpora LINC 2004 at The 20th International Conference on Computational Linguistics COLING 2004*, 2004, pages 19–23.

[23] Gábor Prószéky and Balázs Kis. A Unification-based Approach to Morpho-syntactic Parsing of Agglutinative and Other (Highly) Inflectional Languages. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*. Robert Dale and Kenneth Ward Church, editors. ACL, College Park, Maryland, 1999, pages 261–268.

[24] Attila Novák. Milyen a jó humor? In *Magyar Számítógépes Nyelvészeti Konferencia 2003*. Szeged, 2003, pages 138–145.

[25] Gábor Prószéky and Attila Novák. Computational Morphologies for Small Uralic Languages. In *Inquiries into Words, Constraints and Contexts.* Stanford, California, 2005, pages 150–157.

[26] Attila Novák and Nóra Wenszky. Ó- és középmagyar szóalaktani elemző. In *IX. Magyar Számítógépes Nyelvészeti Konferencia*. Attila Tanács and Veronika Vincze, editors. Szegedi Tudományegyetem, Informatikai Tanszékcsoport, Szeged, 2013, pages 170–181.

[27] János Zsibrita, Veronika Vincze, and Richárd Farkas. magyarlanc: A Toolkit for Morphological and Dependency Parsing of Hungarian. In *Proceedings of Recent Advances in Natural Language Provessing 2013*. Association for Computational Linguistics. Hissar, Bulgaria, 2013, pages 763–771.

[28] Lawrence Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[29] Christer Samuelsson. Morphological tagging based entirely on Bayesian inference. In *9th Nordic Conference on Computational Linguistics NODALIDA-93*. Stockholm University, Stockholm, Sweden, 1993.

[30] Péter Halácsy, András Kornai, and Csaba Oravecz. HunPos: an open source trigram tagger. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*. Association for Computational Linguistics, Prague, Czech Republic, 2007, pages 209–212.

[31] Thorsten Brants. TnT – A Statistical Part-of-Speech Tagger. In *Proceedings of the Sixth Conference on Applied Natural Language Processing*. Universität des Saarlandes, Computational Linguistics. Association for Computational Linguistics, 2000, pages 224–231.

[32] David W. Aha, Dennis Kibler, and Marc K. Albert. Instance-based learning algorithms. *Machine Learning*, 6(1):37–66, 1991.

[33] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA data mining software. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, 2009. ISSN: 19310145.

[34] Ted Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational linguistics*, 19(1):61–74, 1993.

[35] Tibor Kiss and Jan Strunk. Unsupervised multilingual sentence boundary detection. *Computational Linguistics*, 32(4):485–525, 2006.

[36] Eric Jones, Travis Oliphant, Pearu Peterson, et al. SciPy: open source scientific tools for Python. [Online; accessed 2014-11-26]. 2001–. URL: `http://www.scipy.org/`.

[37] Ian H. Witten, Eibe Frank, and Mark A. Hall. *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd edition, 2011, page 629. ISBN: 978-0-12-374856-0.

[38] István Endrédy. Corpus driven research: ideas and attempts. In *PhD Proceedings Annual Issues of the Doctoral School - 2014*. Faculty of Information Technology and Bionics, Pázmány Péter Catholic University., Budapest, Hungary, 2014, 137–140.

[39] Balázs Indig and Gábor Prószéky. Ismeretlen szavak helyes kezelése kötegelt helyesírás-ellenőrző programmal. In *X. Magyar Számítógépes Nyelvészeti Konferencia*. Attila Tanács, Viktor Varga, and Veronika Vincze, editors. Szegedi Tudományegyetem, Szeged, 2013, pages 310–317.

[40] Gábor Prószéky, Balázs Indig, Márton Miháltz, and Bálint Sass. Egy pszicholingvisztikai indíttatású számítógépes nyelvfeldolgozási modell felé. In *X. Magyar Számítógépes Nyelvészeti Konferencia*. Attila Tanács, Viktor Varga, and Veronika Vincze, editors. Szegedi Tudományegyetem, Szeged, 2014, pages 79–87.

[41] Kinga Jelencsik-Mátyus. A szociolingvisztikai stílus: Stratégiák a gyermek-felnőtt diskurzusban. PhD thesis. Szeged, Hungary: University of Szeged, 2015, page 192.

[42] Borbála Siklósi and Attila Novák. Identifying and clustering relevant terms in clinical records using unsupervised methods. In Laurent Besacier, Adrian-Horia Dediu, and Carlos Martín-Vide, editors, *Statistical language and speech processing*. Volume 8791, in Lecture Notes in Computer Science, pages 233–243. Springer International Publishing, 2014. ISBN: 978-3-319-11396-8.

[43] Borbála Siklósi and Attila Novák. A magyar beteg. In *X. Magyar Számítógépes Nyelvészeti Konferencia*. Attila Tanács, Viktor Varga, and Veronika Vincze, editors. Szegedi Tudományegyetem, Szeged, 2014, pages 188–198.