Bevezetés

Az összefoglaló ismerteti a szerző *Feature Engineering for Domain Independent Named Entity Recognition and Biomedical Text Mining Applications* című disszertációjának tartalmát, illetve főbb eredményeit. A disszertáció témakörét a Mesterséges Intelligencia két részterülete képezi, a Gépi Tanulás, illetve annak alkalmazása a Számítógépes Nyelvfeldolgozás (vagy Nyelvtechnológia) területén.

Az elektronikus formában tárolt szöveges tartalmak mennyisége napjainkban egyre gyorsuló mértékben növekszik. A szöveges tartalmakban való keresés egyre időigényesebbé és költségesebbé válik az élet számos területén (mint a gyógyászatban, a kutatásban vagy az üzleti életben), ami mindinkább elengedhetetlenné teszi a keresési folyamat automatizált, számítógépes megoldásokkal való támogatását. Emiatt az emberi munkát kiváltó vagy legalább hatékonyabbá, gyorsabbá tevő, intelligens szövegfeldolgozási megoldások iránti igény egyre intenzívebben megjelenik azon területeken, ahol nagy mennyiségű strukturálatlan (pl. szöveges) adat gyors és mélyreható elemzése szükséges. Az említett három terület emiatt a nyelvtechnológiai kutatás/fejlesztés legnagyobb "megrendelői" közé tartozik. Annak ellenére, hogy a számítógépes megoldások pontossága, megbízhatósága a legtöbb igazán összetett információkeresési feladat esetén elmarad az emberi feldolgozás eredményének precizitásától, a számítógépes rendszerek fölénye a feldolgozható adatmennyiség terén, illetve a feldolgozás relatív költségében mindenképp megmutatkozik. Emellett az egyszerűbb, jól definiált feladatok esetén a számítógépes megoldások minősége is összevethető az emberi feldolgozás eredményével (gondolunk itt pl. a világhálón elérhető tartalmakat indexelő webes keresőszolgáltatásokra).

A Nyelvtechnológia egyik dinamikusan fejlődő részterülete a Szövegbányászat, melynek célja nagy mennyiségű, strukturálatlan szöveges adat feldolgozása, a dokumentumokban szereplő (az adott alkalmazás szempontjából) releváns egységek/szövegrészek megtalálása és kategorizálása, melynek segítségével valamilyen strukturált adattár (pl. adatbázis) automatikusan feltölthető a dokumentumokban található *tudással*. Ez a feldolgozási folyamat az emberi információkeresés gyorsabbá, egyszerűbbé tételét célozza meg. A szövegfeldolgozási feladatok általában megkövetelik a szövegek bizonyos fokú megértését, hiszen az emberi információkeresés során felhasznált, sokszor bonyolult összefüggések automatikus felderítését célozzák meg. Így a szövegbányászati feladatok bonyolultságukat tekintve túlmutatnak az egyszerű kulcsszavas kereséssel megoldható indexelési/visszakeresési problémákon.

A disszertáció célja

A disszertációban több szövegbányászati problémát és az azokra a szerző és munkatársai által kifejlesztett megoldást ismertetünk, a jellemzőkinyerés feladatára koncentrálva. A tárgyalt feladatok széles skálát ölelnek fel a névelemek felismerésétől (tokensorozatok címkézése) a dokumentumosztályozási feladatokig, az alkalmazási területet tekintve pedig az üzleti hírektől az orvosi jegyzőkönyveken át a biológiai témájú tudományos publikációkig. A disszertáció szűkebb értelemben vett célja az, hogy megmutassuk, az adott konkrét feladatra kifejlesztett specifikus jellemzők segítségével olyan modellek konstruálhatók, melyek a gyakorlatban hasznosak, eredményük a szövegek túlnyomó részén az emberi feldolgozással összemérhető pontosságot mutat. A feladatspecifikus megoldások – egyes, egyszerűbb alapproblémák kivételével – általában azt eredményezik, hogy a kapott modellek más, hasonló feladatokra nem alkalmazhatók közvetlenül, azonban az emberi feldolgozás költségessége az egyes konkrét feladatokra adott egyedi megoldások fejlesztését is szükségessé és hasznossá teszi.

Az összefoglaló tematikája

Az összefoglaló a tézishez hasonló szerkezetet követ, azaz a disszertáció két fő témáját külön tárgyaljuk. Az első a Névelemfelismerés területével foglalkozik (a disszertáció 2-4. fejezetei), míg a második részben a Dokumentumosztályozással kapcsolatos téziseket ismertetjük (a disszertáció 5-7. fejezetei). Az összefoglaló végén áttekintjük az egyes fejezetekben ismertetett eredmények közül azokat, amelyeket a szerző saját eredményeinek tekint, majd a főbb eredményeket az egyes, a disszertációban hivatkozott cikkekre vonatkozóan is felsoroljuk.

I. rész – Névelem-felismerés

A szövegben található névelem-kifejezések (tulajdonnevek, nevek akronimjai, időt, mennyiséget jelölő kifejezések, azonosítók, e-mail címek, közigazgatási címek, telefonszámok, stb.) azonosítása és osztályozása a Szövegbányászat egyik legalapvetőbb feladata. Ezek a kifejezések úgynevezett merev jelölők [1], melyeknek a köznyelvi szavakkal ellentétben nincs jelentésük, hanem a világ valamely entitására vagy egy csoportra egyedi módon hivatkoznak (egyfajta referenciák). Ezeket a merev jelölőket a szakirodalomban *névelem*eknek is nevezik.

A névelemek legtöbbször fontos információval bírnak a dokumentum tartalmára nézve, és emiatt az emberi információkinyerés célpontjai gyakran névelemek (csak személynevek az internetes keresések mintegy 30 százalékában előfordulnak [2]). Emiatt a névelemek felismerése és automatikus kategorizálása a szövegbányászat egyik alapfeladata is [3]. A feladat a gépi fordítás területén is nagy jelentőséggel bír [4], hiszen a gépi fordítórendszerekben a tulajdonnevek más szabályok szerint fordítandók, mint a közszavak – általában a különböző típusú tulajdonneveket is változó módon fordítjuk (míg pl. a földrajzi neveknek általában van az adott nyelvre specifikus írásmódjuk, addig pl. a személyneveket tipikusan nem fordítjuk le egyik nyelvről a másikra, még akkor sem, amikor ez egyébként lehetséges lenne). A névelemek és az általuk hivatkozott entitás összerendelése (például a személynevek esetén az azonos nevű emberek közül a névelem által hivatkozott személy meghatározása) [5], és az ugyanarra az entitásra hivatkozó NE-k meghatározása, azaz a koreferenciafeloldás [6] a névelemek felismeréséhez kapcsolódó feladatok, ezek azonban a disszertáció témáján kívül esnek.

A fentiek miatt a legtöbb szövegbányászati probléma esetén a legelső lépés a feladat szempontjából fontosnak ítélt típusba tartozó névelemek megtalálása. Ezeket a feladatokat, ahol a célunk a szövegben olyan tokeneknek (vagy tokenek egymást követő sorozatainak) a megtalálása, mely(ek) egy merev jelölő frázist alkotnak, majd a megtalált frázisok pontos kategorizálása, egységesen névelem-felismerési feladatoknak (Named Entity Recognition, NER) nevezi az irodalom. A kategorizáció mindig az adott alkalmazásra jellemző, hiszen más típusú entitások lényegesek az egyes feladatoknál. A gyakorlatban a szövegben megtalált névelemek további feldolgozás alapját képezik, azaz felismerésük egy köztes lépés a feldolgozási folyamatban, azonban néha maga a NER is lehet önálló végalkalmazás. Ilyen példa az anonimizálás, ahol a névelemek felismerése után csak azok eltávolítása vagy lecserélése történik, hogy a személyes adatoktól megtisztítsuk a dokumentumot.

A névelem-felismerési feladatok hatékonyan megoldhatók címkézett korpusz (olyan szöveges adatbázis, melyben a névelemek előzetesen be lettek jelölve), valamint statisztikai módszerek segítségével. Ezek a kézzel bejelölt példák alapján olyan NE-jelölő szabályokat állítanak elő, melyek később, ismeretlen szövegekben is alkalmasak a hasonló típusú névelemek felismerésére. Mivel viszonylag jól definiált és egyszerűen megoldható feladatról van szó, számos nyelvre és doménre készültek névelem-felismerő modellek. A nyelvek skálája, melyekre nagyobb NER-kiértékelési rendezvény is megszervezésre került, az angoltól [7] a németen [8], hollandon, spanyolon [9] át a kínaiig [10] vagy a japán [11] nyelvig terjed. A főbb domének, melyekre a NER-t intenzíven kutatják: gazdasági, politikai és sporthírek [8], orvosi szövegek [12], kémiai [13] és biológiai szövegek [14], valamint katonai/hírszerzési dokumentumok [15]. A disszertációban magyar és angol újságszövegek és angol orvosi dokumentumok névelem-felismerési problémáival foglalkozunk.

Annak ellenére, hogy a felismerendő névelemek gyakran különböznek az egyes alkalmazási területeken, a feladat megoldására megadhatók olyan rendszerek, amelyek, ha csak korlátozott értelemben is, nyelv- [8] [16] és doménfüggetlenek [17] [18]. A nyelv- vagy doménfüggetlenség NER-alkalmazásoknál az a követelmény, hogy ugyanaz az algoritmus oldja meg a feladatot az alkalmazási területtől, illetve a célnyelvtől függetlenül. Természetesen ehhez megköveteljük, hogy a különböző doménekre és célnyelvekre rendelkezésre álljon címkézett tanítóadatbázis, illetve a felismerendő objektumok típusai valamelyest hasonlók legyenek. A disszertáció témáját elsősorban a tulajdonnevek felismerése és kategorizálása (személynevek, szervezetek nevei, helynevek) képezi, magyar és angol nyelvű újsághírekben, illetve angol orvosi szövegekben[1].

[1] Az orvosi szövegek esetén néhány nem-tulajdonnév kategória, mint pl. telefonszámok, dátumok, stb. felismerésével is foglalkoztunk.

Példák névelem-felismerési feladatokra

Néhány konkrét példán keresztül igyekszünk mélyebb betekintést nyújtani a névelem-felismerési feladatokba. Az egyszerűség és követhetőség kedvéért a példák pontosan azokat az alkalmazásokat fedik le, amelyekkel a disszertáció részletesen is foglalkozik.

NER angol újsághírekben

Angol nyelvű tulajdonnevek automatikus felismerésére a Computational Natural Language Learning (CoNLL) 2003 [8] konferencia korpuszát használtuk. Ezen a konferencián egy nagyszabású NER kiértékelési versenyt rendeztek, amelyhez a szervezők elkészítettek egy statisztikai modellek tanítására alkalmas adatbázist, majd a beküldött rendszereket egy független példahalmazon kiértékelték, összehasonlították. A verseny célja személyek, szervezetek nevei, valamint helynevek felismerése volt a Reuters Inc. hírügynökség 1996-ból származó újsághíreiben (illetve használtak egy ún. *egyéb (MISC)* kategóriát azokra a tulajdonnevekre, amelyek az előző három, kiemelt osztályba nem tartoztak bele). Példa angol nyelvű NER-re:

- [U.N.]$_{ORGANISATION}$ official [Rolf Ekeus]$_{PERSON}$ heads for [Baghdad]$_{LOCATION}$.

NER magyar újsághírekben

Magyarra [19] egy, a CoNLL 2003 konferencia szabványát követő feladatot vizsgáltunk. A cél tehát itt is személyek, szervezetek nevei, helynevek és egyéb tulajdonnevek felismerése és osztályozása volt. A feladathoz a Szeged TreeBank [20] az MTI online oldalról gyűjtött rövidhíreit használtuk fel.
Példa magyar nyelvű NER-re:

- A pénzügyi kockázatok kezeléséről kétnapos nemzetközi konferenciát tartanak csütörtökön és pénteken [Budapesten]$_{LOCATION}$ - mondta [Kondor Imre]$_{PERSON}$, a [Magyarországi Kockázatkezelők Egyesületének]$_{ORGANISATION}$ elnöke szerdán [Budapesten]$_{LOCATION}$ a sajtótájékoztatón.

NER angol orvosi jelentésekben

Gyógyászati területen a névelem-felismerés fontos alkalmazása az orvosi leletek anonimizálása, mely előmozdítja a gyógyászati tevékenységet végző szervezetek (pl. kórházak) közötti információcserét azzal, hogy a betegek személyiségi jogainak védelmében eltávolít minden egyedi, személyes információt a dokumentumokból. A Egyesült Államokban a Health Information Portability and Accountability Act (HIPAA) előírása, 17 különböző típusú személyes információt (Personal Health Information, PHI) különböztet meg. Ezek egyike sem szerepelhet a dokumentumokban ahhoz, hogy azok megoszthatóvá váljanak a beteg közvetlen gyógyításán túlmutató (pl. kutatási) célokra. Az általunk használt dokumentumokban a 17 kategóriából csak 8 fordult elő: betegek és hozzátartozóik kereszt- és családnevei, a betegek életkora (amennyiben 90 éves vagy idősebb betegről van szó), orvosok / kórházi személyzet nevei, kórháznevek, dátumok, azonosítók, személyhívó- és telefonszámok és helynevek. A disszertációban is ismertetett kísérleteinkhez az I2B2 Workshop on Challenges in Natural Language Processing for Clinical Data [12] orvosidokumentum-anonimizáló rendszerek tanítására alkalmas korpuszát használtuk.
Példa angol orvosi dokumentumok anonimizálására:

- Mr. [Cornea]$_{PATIENT}$ underwent an ECHO and endoscopy at [Ingree and Ot of Weamanshy Medical Center]$_{HOSPITAL}$ on [April 28]$_{DATE}$.

Az általunk kifejlesztett statisztikai NER rendszerek

A fentebb ismertetett feladatok megoldására kifejlesztettünk egy statisztikai módszereken alapuló NER-rendszert, amely több nyelven (angol és magyar), illetve alkalmazási területen (újsághírek és orvosi dokumentumok) jó eredményt ért el. A rendszer egy-egy újabb problémára való alkalmazása csak apróbb szükségszerű módosításokat/kiterjesztéseket igényelt – pl. az orvosi területen való alkalmazás néhány új jellemző kifejlesztését igényelte.

Az általunk kifejlesztett rendszer a névelem-felismerési feladatot tokenszintű osztályozási problémaként kezelte. Mintegy 200000 szövegszó méretű címkézett adatbázisokat felhasználva a gépi tanulás területén jó ismert módszereket, döntésifa osztályozót (C4.5 [21]) és boosting algoritmust (AdaBoostM1 [22]) alkalmaztunk.
A feladatok megoldására ugyanazt a tanuló-modellt és megegyező (vagy csak alig eltérő) jellemzőket használtunk. Természetesen az azonos jellemzők értékeit különböző külső források (pl. listák, gyakorisági adatok, stb.) felhasználásával számítottuk ki, azaz magyar NER-re magyar nyelvű listákat, angol újsághírekre angol szövegekből gyűjtött listákat, az orvosi feladatra az adott doménnek megfelelő listákat használtunk. Az általunk kifejlesztett osztályozó modell jellemzők széles skáláját használja fel a névelemek felismerésére és osztályozására (bővebb leírás a jellemzőkről a disszertáció megfelelő fejezeteiben található):

- **Névtárak**, melyek a tanító adatbázisból az egyértelmű tulajdonneveket tartalmazták. Minden olyan névelem, amely legalább ötször előfordult a tanító adatbázisban, valamint az esetek legalább 90 százalékában ugyanolyan címkét kapott, bekerült a névtárakba.

- **Szótárak**, melyek keresztneveket, cégformák neveit, sportklubok neveit, földrajzihely-típusokat, stb. tartalmaztak. A hasznosnak gondolt tematikus listákat az internetről gyűjtöttük össze.

- **Helyesírási / Felszíni jellemzők**: kis vagy nagy kezdőbetűs alak, szóhossz, egyszerű felszíni információk a szóalakról (tartalmaz-e számjegyet, van-e nagybetű a szó belsejében, reguláris kifejezések, stb.), illetve a különböző típusú NE-kben előforduló legjellemzőbb karakter bi- és trigramokat összegyűjtöttük.

- **Gyakorisági információ**: a szóalak gyakorisága, kisbetűs és nagybetűs előfordulások aránya, mondateleji nagybetűs és összes nagybetűs előfordulások aránya.

- **Frázisszintű információ**: a mondattani egység kódja amelyhez a szóalak tartozik, az előző szóalakokhoz hozzárendelt címkék (ehhez online kiértékelést használtunk).

- **Szövegkörnyezetet leíró információk**: Szófaji kód, mondatpozíció, dokumentumzónák (cím vagy szövegtörzs), témakód, kulcsszavak (a leggyakoribb olyan kulcsszavak, melyek csak egyfajta névelem előtt/után állhatnak a tanítókorpusz statisztikái alapján), illetve a szóalak zárójelek vagy idézőjelek között van-e.

A döntésifa-tanulás kedvező tulajdonságai miatt, illetve az általunk használt kompakt jellemzőtér-reprezentációnak köszönhetően (átlagosan körülbelül 200 jellemzőt használtunk a NER probléma megoldásához), a modelljeink gyorsan taníthatóak és tesztelhetőek, valamint a szabvány kiértékelési adatbázisokon jó pontosságot adtak.
A domén- és nyelvfüggetlen NER-modellünk a következő eredményeket érte el:

- 89.02% F-mérték a CoNLL 2003 NER verseny kiértékelési adatbázisán

- 94.76% F-mérték a Magyar NER korpuszon

- 94.34% F-mérték az I2B2 workshop anonimizálási kiértékelő adatbázisán.

A doménspecifikus kiterjesztések a modell pontosságát orvosi szövegeken 97.64%-ra növelték. Az eredmények mind frázisszintű F mértéket jelentenek, mely a CoNLL 2003-as NER verseny szabvány kiértékelési metrikája.

II. rész – biológiai és orvosi dokumentumosztályozás

A szöveges adatok (rendszer logok, orvosi jelentések, újságcikkek, fogyasztói visszajelzések, stb.) emberi feldolgozása munkaigényes és költséges feladat, amely a szöveges információ növekedtével egyre nehezebben oldható meg. Egyre növekszik az igény olyan megoldások iránt, amelyek automatizálják vagy felgyorsíthatják a most még sokszor emberek által végzett adatelemző, információkereső tevékenységet. Emiatt a természetes nyelvi szövegek automatikus kategorizálása/osztályozása napjainkra a Szövegbányászat egyik legfontosabb feladatává vált.

Sok szövegfeldolgozási feladat felírható a gépi tanulás területén közismert ún. osztályozási feladatként, ami lehetővé teszi azok gépi tanulási algoritmusok segítségével történő, eredményes megoldását [23]. Ezek a megoldások képesek a folyó szövegben megtalálható rejtett szabályszerűségek, struktúra felfedezésére, amennyiben rendelkezésünkre állnak címkézett dokumentumok, melyek segítségével a rendszerek taníthatók. A dokumentumosztályozási feladatok esetén a rendszertől elvárt kimenet minden esetben használható tárgyi tudás (tényszerű információ), nem pedig egy döntés, hogy a dokumentum tartalmaz-e számunkra érdekes információt vagy sem. Emiatt ezek a megoldások általában túlmutatnak az egyszerű kulcsszavas információkeresési technikákon (kulcsszavas keresés és a találatok rangsorolása), hiszen a feladatok szükségessé teszik a szövegek bizonyos szintű *megértését*. A dokumentumosztályozó rendszereknek általában kezelniük kell a különböző írott alakok, a szinonímia, vagy pl. a tagadás, érzelmi töltet, bizonytalanság, illetve az időbeliség okozta nehézségeket [24].

A szövegbányászati megoldások legnagyobb alkalmazási területei közé tartozik a Biológia és a Gyógyászat [25]. Az ezeken a területen dolgozó szakemberek, kutatók általában nagy mennyiségű szöveges dokumentummal dolgoznak mindennapi munkájuk során a kutatásban (tudományos publikációkat, szabadalmakat, vagy a témához kapcsolódó korábbi kísérletek beszámolóit olvassák) vagy a döntéshozásban (pl. korábbi, hasonló tünetekkel vagy diagnózissal kezelt betegek kórtörténetét elemzik).

Annak ellenére, hogy a célnyelvtől és alkalmazási területtől független, hatékony eszközökre itt is nagy szükség lenne, az ilyen általános rendszerek napjainkban még nem igazán megvalósíthatók. Így a gazdasági megfontolások speciálisabb, konkrét egyedi feladatok megoldására alkalmas rendszerek fejlesztését is indokolttá teszik. A disszertációban biológiai és orvosi szövegfeldolgozási feladatokra koncentrálunk.

Példák dokumentumosztályozási feladatokra

Néhány konkrét példán keresztül igyekszünk mélyebb betekintést nyújtani a dokumentumosztályozási feladatokba. Az egyszerűség és követhetőség kedvéért a példák pontosan azokat az alkalmazásokat fedik le, amelyekkel a disszertáció részletesen is foglalkozik.

A beteg dohányzási státuszának azonosítása a zárójelentése alapján

Az orvosi jelentések számítógépes feldolgozásának egyik fő célja, hogy segítse az orvosok, kutatók munkáját pl. a gyógyszerkutatások során. Ennek egyik módja lehet pl. statisztikailag releváns adatok gyűjtése a kutatók számára, további elemzések céljára. Ilyen elemzés lehet pl. a betegségek lefolyásának és a különböző szenvedélyeknek (pl. dohányzás, drogfogyasztás) egymásra gyakorolt hatásának vizsgálata [26]. A különféle betegségek, valamint a káros szenvedélyek közötti hatások, összefüggések feltárása kritikus fontosságú a megelőzési és gyógyászati folyamatban.

Ilyen összefüggések automatikusan kinyerhetőek statisztikai eszközök és nagyméretű dokumentumadatbázisok felhasználásával. A disszertációban ismertetett eredményekhez az I2B2 Workshop on Challenges in Natural Language Processing for Clinical Data [27] dohányzási státusz azonosítására készített szabvány adatbázisát használtuk. Az említett verseny során a feladat a betegeknek az alábbi öt kategória valamelyikébe történő besorolása volt:

- nemdohányzó: a betegnek nincs dohányos múltja, azaz nem dohányzik és korábban sem dohányzott.
- Aktív dohányos: a beteg jelenleg is dohányzik, vagy a kezelést megelőző 1 éven belül dohányzott.
- Exdohányos: a beteg legalább egy éve nem dohányzik.
- Dohányos: amikor a beteg vagy aktív, vagy exdohányos, de ez nem eldönthető a dokumentum alapján.
- Ismeretlen: a dokumentum nem tartalmaz semmilyen információt a beteg dohányzási szokására vonatkozóan.

Egy példamondat a beteg dohányzási szokásáról a zárójelentésben:

- *The patient is a 60 yo right handed gentleman with a 20-years history of heavy smoking. Agreed to participate in a smoking cessation program. (aktív dohányos)*

Bizonytalan tartalom detektálása mondatokban

Bizonyos, gyakran használt nyelvi jelenségeknek, mint amilyen a spekuláció [28] [29], tagadás vagy a múlt idő használata, a pontos felismerése alapkövetelmény a biológiai/orvosi dokumentumok hatékony feldolgozásához. A legtöbb szövegbányászati alkalmazásban azok az információk, melyek spekulatív kontextusban találhatók meg a szövegben, hibás pozítív elemekként jelennek meg (a rendszernek nem szabad ilyen információt tényként kigyűjtenie). Így a spekulációk detektálásakor egyfajta tartalmi szűrését végezzük el a szövegnek – a cél a tényszerű információk elválasztása a bizonytalan, spekulatív elemektől. A biológiai tudományos szövegeken végzett kísérleteinkhez a Medlock és Briscoe [30] által közzétett adatbázist használtuk, valamint egy másik, általunk felcímkézett adatbázist is felhasználtunk arra a célra, hogy megbecsüljük, mennyire jól alkalmazhatók az egyes modellek más forrásból származó (de hasonló témájú) szövegekre. Az orvosi dokumentumokon elvégzett vizsgálatokhoz az International Challenge on Classifying Clinical Free Text Using Natural Language Processing konferencia klinikai dokumentumok betegség- és tünetkódokkal való címkézésére készített adatbázisát használtuk fel.
Két spekulatív mondat biológiai tudományos szövegekből:

- *Thus, the D-mib wing phenotype may result from defective N inductive signaling at the D-V boundary.*
- *A similar role of Croquemort has not yet been tested, but seems likely since the crq mutant used in this study (crqKG01679) is lethal in pupae.*

Két spekulatív mondat radiológiai jelentésekből:

- *Findings suggesting viral or reactive airway disease with right lower lobe atelectasis or pneumonia.*
- *Right middle lobe infiltrate and/or atelectasis.*

Radiológiai jelentések automatikus BNO-kódolása

Az orvosi jelentésekhez hozzárendelt International Classification of Diseases, 9th Revision, Clinical Modification (magyarul Betegségek Nemzetközi Osztályozása, BNO kódrendszer) címkék szolgálnak az egyes kezelések könyvelésére, számlázásra. Tehát (az USA-ban) a biztosító társaságok a kórházaknak a térítéseket a dokumentumokhoz rendelt kódok szerint fizetik meg. A megfelelő kódok hozzárendelése a jegyzőkönyvekhez, illetve az ezzel kapcsolatos hibák javítása megközelítőleg 25 milliárd dollár költséggel jár éves szinten, csak az Egyesült Államokban [31].

Mivel a dokumentumokhoz rendelt BNO-kódok elsődlegesen számlázási célokra szolgálnak, a kódolás során elkövetett hibáknak közvetlen anyagi vonzata van: a hamis negatív kódok (azaz olyan címkék, amelyeket hozzá kellett volna rendelni a dokumentumhoz, de ez elmaradt) bevételkieséssel járnak az egészségügyi intézet számára, míg a hamis pozitív kódok (azaz amelyeket tévesen rendelnek hozzá egy dokumentumhoz, túlkódolás) miatt büntetésként az azzal keresett összeg háromszorosát róják ki az intézményre. Ez utóbbi szélsőséges esetben jogi következményekkel is járhat (csalás). Emiatt a minél precízebb automatikus BNO-kódoló eljárások fejlesztése iránti igény igen nagy. A klinikai kódolás területen végzett kísérleteinkhez az International Challenge on Classifying Clinical Free Text Using Natural Language Processing [32] verseny keretében közzétett radiológiai jelentésekből álló korpuszt használtuk.
Két példa radiológiai jelentések BNO-kódolására:

- *CODES: 486; 511.9*
  HISTORY: Right lower lobe pneumonia, cough, followup.
  IMPRESSION: Persistent right lower lobe opacity with pleural effusion present, slightly improved since prior radiograph of two days previous.

- *CODES: 593.89; V13.02*
  HISTORY: 14-year - old male with history of a single afebrile urinary tract infection in January with gross hematuria for a week. The patient was treated with antibiotics.
  IMPRESSION: Mild left pyelectasis and ureterectasis. Otherwise normal renal ultrasound. The bladder appears normal although there is a small to moderate post void residual.

Az általunk kifejlesztett statisztikai mondat / dokumentum-osztályozó modellek

A fenti feladatok mindegyikére különböző modellt fejlesztettünk ki. Mivel a címkézett korpuszok előállításának költsége igen nagy, a névelem-felismerési feladatoknál használt méretű jelölt adatbázis ezeknél a problémáknál nem volt kivitelezhető – emiatt olyan jellemzőket kellett használnunk, melyek segítségével az alkalmazott gépi tanulási modellek néhány száz példa használatával is képesek voltak a fontos összefüggések megtalálására, és így a feladat megoldására. Egy alternatív lehetőség a címkézett példák részben vagy teljeset felügyelet nélküli előállítása (részben felügyelt vagy felügyelet nélküli tanulás), ebben az esetben azonban az automatikus címkézésben jelen lévő zajjal kellett megbirkózni annak érdekében, hogy megbízható modelleket kapjunk eredményül. Annak ellenére, hogy az egyes konkrét feladatokra viszonylag jó eredményeket sikerült elérnünk, a tanult modellek általánosítóképessége, pont az említett kis számú tanító minta miatt, korlátozott maradt.

A beteg dohányzási státuszának azonosítása a zárójelentése alapján

A dohányzási státusz megállapításához mondatszintű osztályozó modellt használtunk, vektortérmodelles jellemzőtér-reprezentáció mellett. Az egyszerű szószintű reprezentációt összetett jellemzőkkel egészítettük ki. A komplex jellemzők, melyeket kifejlesztettünk a mondatokban található kétszavas kifejezések csoportosításával (jelentésük alapján) álltak elő, illetve felhasználtunk egyszerű mondattani jellemzőket és a tagadást is kezeltük a mondatban, a nemdohányos osztály pontosabb azonosítása érdekében. Az elvégzett kísérletek azt mutatták, hogy az általunk kifejlesztett jellemzők hasznosabbak az osztályozómodellek tanításához, mint a korábbi munkákban használt egyszerű szószintű reprezentáció (különböző automatikus jellemzőkiválasztó módszerek többnyire ezeket választották az egyszerű kulcsszavas jellemzők helyett). Az osztályozási feladat megoldására több tanuló algoritmust teszteltünk (C4.5 döntési fa, Mesterséges Neuronháló, Support Vector Machine osztályozó), a végső modellünk ezek kombinációjával (többségi szavazással) állt elő. Az általunk kifejlesztett modell különösen hatékonynak bizonyult az *ismeretlen* osztályba tartozó dokumentumok megtalálásában, valamint a nemdohányos osztályba is jó pontossággal sorolta be a megfelelő dokumentumokat. Ezáltal a

rendszerünk kiváló előfeldolgozó eszköz lehet az emberi feldolgozáshoz (az irreleváns dokumentumok, illetve a nemdohányzók leválogatásával). A rendszer 86.54% pontosságot ért el a hivatalos kiértékelő adatbázison, amely megközelítette a legjobb beküldött rendszerek pontosságát.

Bizonytalan tartalom detektálása mondatokban

Ennél a feladatnál egy részben felügyelt modellt használtunk biológiai szövegek és egy felügyelet nélküli módszert orvosi dokumentumok esetén a címkézett tanítóadatbázis előállítására. A mondatok spekulatív/nem-spekulatív osztályozására Maximum Entrópia osztályozót [33], valamint vektorteres jellemzőtér-reprezentációt alkalmaztunk. A felhasznált jellemzőkészletet szó uni- bi- és trigramok uniója alkotta, melyből automatikus módszerrel választottuk ki a leghasznosabb jellemzőket. Az általunk kifejlesztett jellemzőkiválasztás egy iteratív szűrés/újrarangsoroláson alapuló eljárás volt, ahol az újrarangsorolásnál olyan módon számítottuk az egyes jellemzők prediktív erejét (a spekulatív osztályra vett osztályfeltételes valószínűségét), hogy minden pozitív példánál csak a két, az előző iterációban legjobbnak értékelt jellemzőt vettünk figyelembe (míg negatív példáknál minden jellemzőt figyelembe vettünk). Ez a szűrési módszer azon alapul, hogy egy mondaton belül legtöbbször egy, vagy legfeljebb kettő elem spekulatív értelmű, azaz célszerű mindig csak a legígéretesebb egy-két jellemzőt figyelembe venni. Ennek a szűrésnek, illetve az egyes jellemzőket kiértékelve a MaxEnt modell által adott $P(spec)$ valószínűségeket felhasználva a zajos adatbázis segítségével is képesek voltunk kinyerni a szövegekből a legfontosabb spekulatív kulcsszavakat. A kiválasztott jellemzőket felhasználva a részben vagy teljesen automatikus módon címkézett tanulóadatokat felhasználva az általunk fejlesztett rendszer 85.08%-os $F_{\beta=1}(spec)$ értéket ért el biológiai tudományos cikkek mondatain, illetve 82.07%-os $F_{\beta=1}(spec)$ értéket produkált orvosi dokumentumok mondatain.

Azt tapasztaltuk, hogy a legfontosabb kulcsszavak kiválasztása után a tanult modellek már egyszerű kulcsszó-alapú osztályozuásra egyszerűsödtek, azaz nem voltak képesek egy-egy jellemző spekulatív illetve nem-spekulatív használatának elkülönítésére. Ennek megoldására nagyobb méretű címkézett adatbázisra, illetve a kiválasztott kulcsszavak nem-spekulatív használatait megragadó kifejezések gyűjtésére lenne szükség.

Orvosi jelentések automatikus BNO-kódolása

Az orvosi dokumentumok automatikus betegség- és tünet-kódolására egy hibrid, statisztikai és szabályalapú komponenseket ötvöző rendszert fejlesztettünk ki. A feladat jellegzetessége volt, hogy az interneten szabadon hozzáférhetőek olyan kódolási útmutatók, melyek a szakemberek számára írják le a kódolás során alkalmazandó protokollt. Ezek az online elérhető útmutatók lehetőséget adtak a kódolást elvégző egyszerű, szabályalapú modellek gyors kifejlesztésére (melynek szabályai éppen az útmutatóban található előírások félig automatikus formalizálásával álltak elő). Emiatt tehát kísérleteink során azt vizsgáltuk, hogyan lehet a rendelkezésre álló címkézett adatokat az előbb említett szabályalapú rendszerek javítására, továbbfejlesztésére felhasználni.

Első lépésben előállítottunk egy egyszerű szabályrendszert, mely egy online hozzáférhető kódolási útmutatóban megtalált előírások alapján végezte el az inputként kapott dokumentumok BNO-kódolását. Ezután statisztikai módszerek és a rendelkezésre álló címkézett adatok segítségével modelleztük a különböző betegség-tünet relációkat, melyek a szabályalapú rendszer számára rejtve maradtak, majd pedig hasonló megközelítéssel a szabályalapú rendszer szótáraiból hiányzó szinonímákat kerestünk (mindkét esetben a szabályalapú rendszer hibás címkézéseit tekintve a tanulás során pozitív példaként).

A disszertációban ismertetett hibrid, szabályalapú és statisztikai modellek kombinációján alapuló modellel kiemelkedő pontosságot sikerült elérni (a rendszer teljesítménye közelíti az emberi címkézés pontosságát), viszonylag alacsony fejlesztési idő mellett. Ez utóbbit amiatt tartjuk lényeges eredménynek, mert ezáltal lehetővé válhat a rendszer kifejlesztése a kísérleteinkben használt számúnál lényegesen nagyobb címkehalmaz esetén is (míg a teljes egészében kézi szabályokon alapuló rendszereket több száz, vagy több ezer kód esetén már problémássá válhat kifejleszteni, illetve karbantartani).

Összegzés

Fejezetenkénti áttekintés

Ebben a részben összefoglaljuk a disszertáció egyes fejezeteinek főbb eredményeit, különös tekintettel azokra, melyeket a szerző saját eredményének tekint. Az eredmények és a disszertációban hivatkozott cikkek viszonyát is megadjuk a disszertáció fejezetei szerint.

A disszertáció két fő részből áll, az első rész névelem-felismerési feladatokkal, a második rész dokumentumosztályozási feladatokkal foglalkozik.

- NER fejezetek

  1. Magyar nyelvű névelem-felismerés fejezet

     A szerző részt vett az első magyar nyelvű névelem-felismerési referenciakorpusz kialakításában, mely lehetővé tette magyar nyelvű statisztikai alapú névelem-felismerő rendszerek kutatását, fejlesztését. Ez közös és oszthatatlan eredménye a [19] cikk szerzőinek, illetve a nyelvész kollégáknak, akik a korpusz annotációs munkáit végezték.

     Szerzőtársaival a disszertáció szerzője megtervezett és implementált egy névelem-felismerési feladatok megoldására alkalmas keretrendszert, mely 94.76%-os frázis-szintű F-mérték felismerési pontosságot ért el a Szeged Korpusz gazdasági hírekből álló szövegein. Ebben a munkában a szerző hozzájárulása volt meghatározó a felhasznált jellemzőtér-reprezentáció megtervezésében és kialakításában.

     Ezek az eredmények az alábbi cikkekben kerültek közlésre: [19], [34] és [35].

  2. Angol nyelvű névelem-felismerés fejezet

     A szerző részt vett az eredetileg magyarra tervezett rendszer egy új célnyelvre történő átalakításában (nyelvfüggetlen NER-rendszer kialakításában). Az átalakított rendszer 89.02%-os F-mérték felismerési pontosságot ért el a CoNLL 2003 konferencia szabvány kiértékelési adatbázisán. Az angol nyelvű rendszer fejlesztése során a szerző hozzájárulása volt meghatározó a jellemzőtér-reprezentáció testreszabásában, illetve kibővítésében.

     A szerző szintén részt vett egy a SemEval-2007 konferencia metonímia-feloldási versenyére [36] készített, MaxEnt modellen alapuló névelemek metonimikus használatát detektáló rendszer fejlesztésében. A kifejlesztett rendszerben a szerző egy webes heurisztikán alapuló modell kifejlesztését végezte, amelyet eredményesen használtak fel a névelemek toldalékainak leválasztására (jellemzőként felhasználtuk a névelemek egyes, illetve többes számú előfordulását elkülönítő heurisztika eredményét).

     A többes számú alakokat elkülönítő heurisztika eredményességéből kiindulva a szerző olyan korpuszgyakoriságon alapuló eljárásokat dolgozott ki, amelyek képesek a NER-rendszerek bizonyos típushibáit kijavítani (egymást követő, azonos típusú frázisok szétvágása wikipédiás tartalmak elemzésével). Ezeket a heurisztikákat később egy egységes keretben mint a névelemek szótövezésére alkalmas rendszert dolgozta ki. A rendszer (melynek megvalósításában szerzőtársai segítettek) alkalmas a névelemek normalizált alakjának előállítására korpuszgyakoriságon alapuló heurisztikák segítségével. A probléma (mely angolra is megjelenik a többes számú, illetve birtokos esetű névelemeknél, valamint kimondottan gyakori agglutinatív nyelvekben, amilyen a magyar) fontosságát az adja, hogy a különböző morfológiai elemző rendszerek általában szótövek listáját használva működnek, azaz nem adnak jó eredményt névelemek szótövezése esetén (hiszen a névelemek kimerítő listáját nem lehetséges felhasználni a feladat megoldásához). A szerző hozzájárulása ezekhez az eredményekhez az elméleti modell és a módszerek megtervezése volt.

     A felsorolt eredmények az alábbi cikkekben kerültek közlésre: [35], [37], [38] és részben [39].

3. NER angol orvosi jelentésekben fejezet

   A szerző és kollégái részt vettek a 2006-os I2B2 shared task challenge on medical record de-identification versenyen, mely anonimizáló alkalmazások fejlesztését és kiértékelését tűzte ki célul. A meglévő NER-keretrendszer testreszabását, illetve az elért eredményeket közös eredménynek tekintjük. Az elért eredmények (a kifejlesztett rendszer a 2. legjobb pontosságot érte el) alapján elmondható, hogy a disszertációban ismertetett keretrendszer (így a disszertáció eredményeként felsorolt jellemzőtér-reprezentáció) alkalmas a NER-jellegű feladatok hatékony megoldására a célnyelvtől és az alkalmazási területtől függetlenül.

   A szerző hozzájárulása volt döntő a jellemzőtér-reprezentáció kialakításában, illetve kibővítésében (speciálisan az orvosi szövegeken hasznos jellemzők fejlesztésében). A kifejlesztett újszerű jellemzők hozzájárultak az elért kiemelkedő eredményhez (a 2. legjobb szószintű pontosságot és a legjobb frázisszintű pontosságot a szerző és társai által fejlesztett rendszer adta).

   Ezeket az eredményeket a [40] cikkben foglaltuk össze.

- Dokumentumosztályozási fejezetek

1. A beteg dohányzási státuszának azonosításával foglalkozó fejezet

   A szerző és kollégái részt vettek a 2006-os I2B2 shared task challenge on patient smoking status classification versenyen, mely a beteg dohányzási szokásainak zárójelentések alapján történő azonosítását tűzte ki célul. A kifejlesztett rendszert és a rendszerszintű eredményeket itt is közös eredménynek tekintjük.

   A szerző hozzájárulása volt döntő a jellemzőtér-reprezentáció kialakításában, mely a korábbi megközelítésekben használt jellemzőtér kifejlesztéséből, illetve újszerű jellemzők megtervezéséből és kifejlesztéséből állt. Az újszerű jellemzők közös vonása, hogy több, szintaktikai vagy szemantikai szempontból hasonló mondatrészt egyetlen komplex jellemzőbe vont össze. A komplex, több elemi jellemző összevonásával kapott jellemzők fejlesztését a szerző elsősorban a kis mintaméret hatásának enyhítése és a túltanulás elkerülése érdekében dolgozta ki; a jellemzők hasznosnak bizonyultak a rendszer értékelése szempontjából, hiszen több különböző szelekciós eljárás is a legmagasabban rangsorolt jellemzők közé helyezte azokat.

   Ezeket az eredményeket a [41] cikkben foglaltuk össze.

2. Bizonytalan tartalom detektálása fejezet

   Minden, a fejezetben ismertetett eredmény a szerző munkája. A fejezet főbb eredményei közé a komplex jellemző-rangsorolási és szelekciós módszert tekintjük, amely a valódi spekulatív kulcsszavakat sikeresen elkülöníti az egyéb szavaktól a szövegben.

   Két különböző környezetben, orvosi és biológiai szövegek esetén is megmutattuk, hogy minimális felügyelet mellett is jó pontosságú spekulációdetektáló modellek készíthetők félig vagy teljesen automatikus módon előállított tanítóadatbázis használata esetén is.

   A korábbi megközelítésekhez képest felhasználtunk 2-3 szó hosszúságú frázisokat is mint jellemzőket, valamint megmutattuk, hogy ezzel javulás érhető el a modellek pontosságában.

   Emellett kísérletekkel igazoltuk, hogy a spekulatív tartalom kifejezési módja nagyban változhat, ha különböző forrásból származó szövegekkel dolgozunk. Ez jelentős visszaesést eredményez a rendszerek pontosságát tekintve, amennyiben egy kifejlesztett modellt mindenféle átalakítás nélkül alkalmazunk egy új területre.

   Ezeket az eredményeket a [42] és részben a [43] cikkben foglaltuk össze.

3. Radiológiai jelentések automatikus BNO-kódolása fejezet

   A szerző és kollégája részt vett a 2007-es CMC shared task challenge on automated ICD-9-CM coding of medical free texts using Natural Language Processing versenyen, mely

orvosi dokumentumok automatikus BNO-kódolását tűzte ki célul. A kifejlesztett rendszert és a rendszerszintű eredményeket itt is közös eredménynek tekintjük.

A szerző hozzájárulása volt meghatározó a kezdeti és a teljes egészében szakértői szabályokon alapuló rendszer kifejlesztésében; a komplex annotátor-egyetértési ráta elemzésben, valamint a gépi tanulási modell és a jellemzők kifejlesztésében, amelyet a címkeközi összefüggések felderítésére használt fel.

Ezeket az eredményeket a [44] cikkben foglaltuk össze.

	HunNER	EngNER	DE-ID	SMOKER	HEDGE	ICD-9
LREC[19]	•					
ACTA[34]	•					
DS2006[35]	•	•				
SEMEVAL[39]		•				
ICDM2007[37]		•				
TSD2008[38]		•				
JAMIA[40]			•			
WSEAS[41]				•		
ACL[42]					•	
BIONLP[43]					•	
LBM2007[44]						•

1. táblázat. A disszertáció fejezeteinek és a hivatkozott publikációk viszonya.

Publikációk szerinti összefoglalás

Az alábbiakban felsoroljuk az egyes publikációkban szereplő fontosabb eredményeket, amelyeket a szerző a *saját* eredményeinek tekint. Itt is megemlítjük, hogy a rendszerszintű eredményeket, értékeléseket minden esetben közös eredménynek tekintjük, mivel lehetetlen számszerűsíteni, hogy az elért pontosságértékekben az egyes rendszerelemek hozzájárulása milyen mértékű. Ez alól az egyetlen kivételt a [42] cikkben ismertetett eredmények képezik, melyek a szerző egyedüli munkáját képezik.

A [19] cikket kihagytuk a felsorolásból, hiszen annak minden eredményét a szerzők közös munkájának tekintjük (illetve nem tekintjük a disszertáció szerves részének). A [39] cikkben összefoglalt eredményekhez a szerző csak kis mértékben járult hozzá, így ezeket sem soroljuk a disszertáció főbb eredményei közé.

- ACTA[34]
  - Jellemzőtér-reprezentáció kidolgozása magyar NER-re.
  - Tömör jellemző-reprezentáció.
  - Gyakoriság-alapú jellemzők kidolgozása.
- DS2006[35]
  - A jellemzőtér-reprezentáció átalakítása és kiterjesztése angol NER-re.
- SEMEVAL[39]
  - A "többes szám" jellemző a metonímia-feloldáshoz.

- ICDM2007[37]

  – Webes gyakoriságalapú heurisztikák kidolgozása az egymást követő, azonos típusú névelemek szeparálására.

- TSD2008[38]

  – A webes heurisztikákon alapuló NE normalizálás (szótövesítés és inflexiók előállítása) általános modelljének, illetve az alkalmazott módszereknek a megtervezése.

- JAMIA[40]

  – A jellemzőtér-reprezentáció átalakítása és kiterjesztése orvosi szövegekre.

  – Az iteratív tanítás-jellemzőgenerálás módszer kidolgozása.

- WSEAS[41]

  – Szó bi- és trigramok használata jellemzőként.

  – Komplex jellemzők kidolgozása (jelentésük alapján csoportosított kifejezések, mondattani információk, tagadás).

  – Jellemzőkiválasztási módszerek használata a leghatékonyabb reprezentáció megtalálására.

- ACL[42]

  – A cikkben ismertetett összes eredmény.

- BIONLP[43]

  – A negáció, spekuláció és hatókörük annotációs munkáinak néhány alapvető elvét a szerző dolgozta ki.

- LBM2007[44]

  – Részletes rendszer- és annotátor-egyetértés elemzés.

  – Komplex jellemzők és gépi tanulási modell kidolgozása a címkeközi összefüggések felderítésére.

  – Egy egyszerű, szabályalapú BNO-kódoló alkalmazás kifejlesztése, mely a további kutatási / fejlesztési munkák alapját képezte, illetve egy teljes egészében szakértői szabályokon alapuló rendszer készítése (mely összehasonlításul szolgált).

Hivatkozások

[1] Kripke S: *Naming and Necessity*. Harvard University Press 1972.

[2] Guha RV, Garg A: **Disambiguating People in Search**. In *Proceedings of the 13th World Wide Web Conference (WWW 2004)*, ACM Press 2004.

[3] Jurafsky D, Martin JH: *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. Prentice Hall, second edition 2008, [http://www.amazon.de/exec/obidos/redirect?tag=citeulike01-21\&path=ASIN/013122798X].

[4] Babych B, Hartley A: **Improving Machine Translation Quality with Automatic Named Entity Recognition**. In *Proceedings of the 7th International EAMT workshop at EACL-2003*, Budapest, Hungary: Association for Computational Linguistics 2003:18–25.

[5] Cucerzan S: **Large-Scale Named Entity Disambiguation Based on Wikipedia Data**. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)* 2007:708–716.

[6] Nicolae C, Nicolae G: **BESTCUT: A Graph Algorithm for Coreference Resolution**. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, Sydney, Australia: Association for Computational Linguistics 2006:275–283, [http://www.aclweb.org/anthology/W/W06/W06-1633].

[7] Chinchor NA: **Overview of MUC-7/MET-2.** In *Proceedings of the Seventh Message Understanding Conference (MUC-7)* 1998.

[8] Tjong Kim Sang EF, De Meulder F: **Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition**. In *Proceedings of CoNLL-2003*. Edited by Daelemans W, Osborne M, Edmonton, Canada 2003:142–147.

[9] Tjong Kim Sang EF: **Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition**. In *Proceedings of CoNLL-2002*, Taipei, Taiwan 2002:155–158.

[10] Tou Ng H, Kwong OOY (Eds): *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, Sydney, Australia: Association for Computational Linguistics 2006.

[11] Sekine S, Isahara H: **IREX: IR and IE evaluation project in Japanese** 2000, [citeseer.ist.psu.edu/sekine00irex.html].

[12] Uzuner O, Luo Y, Szolovits P: **Evaluating the State-of-the-Art in Automatic De-identification**. *J Am Med Inform Assoc* 2007, **14**(5):550–563, [http://www.jamia.org/cgi/content/abstract/14/5/550].

[13] Corbett P, Batchelor C, Teufel S: **Annotation of Chemical Named Entities**. In *Biological, translational, and clinical language processing*, Prague, Czech Republic: Association for Computational Linguistics 2007[http://www.aclweb.org/anthology/W/W07/W07-1008].

[14] Kim J, Ohta T, Tsuruoka Y, Tateisi Y, Collier N: **Introduction to the bio-entity recognition task at JNLPBA**. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA), Geneva, Switzerland*. Edited by Collier N, Ruch P, Nazarenko A 2004:70–75.

[15] Grishman R, Sundheim B: **Message Understanding Conference-6: a brief history**. In *Proceedings of the 16th conference on Computational linguistics*, Morristown, NJ, USA: Association for Computational Linguistics 1996:466–471.

[16] Cucerzan S, Yarowsky D: **Language Independent Named Entity Recognition Combining Morphological and Contextual Evidence**. In *Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, College Park, MD, USA: Association for Computational Linguistics 1999:90–99.

[17] Kozareva Z: **Bootstrapping Named Entity Recognition with Automatically Generated Gazetteer Lists**. In *Proceedings of the Student Research Workshop at 11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy: Association for Computational Linguistics 2006:15–21.

[18] Lee HS, Park SJ, Jang H, Lim J, Park SH: **Domain Independent Named Entity Recognition from Biological Literature**. In *Proceedings of The 15th International Conference on Genome Informatics*, Yokohama, Japan 2004.

[19] Szarvas Gy, Farkas R, Felföldi L, Kocsor A, Csirik J: **A highly accurate Named Entity corpus for Hungarian**. In *Proceedings of Language Resources and Evaluation Conference* 2006.

[20] Csendes D, Csirik J, Gyimóthy T, Kocsor A: **The Szeged Treebank**. In *TSD* 2005:123–131.

[21] Quinlan JR: *C4.5: Programs for Machine Learning*. Morgan Kaufmann 1993.

[22] Schapire R: **The boosting approach to machine learning: An overview**. In *Proceedings of MSRI Workshop on Nonlinear Estimation and Classification*, Berkeley, CA, USA 2001.

[23] Sebastiani F: **Machine learning in automated text categorization**. *ACM Comput. Surv.* 2002, **34**:1–47, [http://portal.acm.org/citation.cfm?id=505282.505283].

[24] Shanahan JG, Qu Y, Wiebe J: *Computing Attitude and Affect in Text: Theory and Applications (The Information Retrieval Series)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc. 2005.

[25] Ananiadou S, Mcnaught J: *Text Mining for Biology And Biomedicine*. Norwood, MA, USA: Artech House, Inc. 2005.

[26] Zeng Q, Goryachev S, Weiss S, Sordo M, Murphy S, Lazarus R: **Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system**. *BMC Medical Informatics and Decision Making* 2006, **6**:30, [http://www.biomedcentral.com/1472-6947/6/30].

[27] Uzuner O, Goldstein I, Luo Y, Kohane I: **Identifying Patient Smoking Status from Medical Discharge Records**. *J Am Med Inform Assoc* 2008, **15**:14–24, [http://www.jamia.org/cgi/content/abstract/15/1/14].

[28] Light M, Qiu XY, Srinivasan P: **The Language of Bioscience: Facts, Speculations, and Statements In Between**. In *HLT-NAACL 2004 Workshop: BioLINK 2004, Linking Biological Literature, Ontologies and Databases*. Edited by Hirschman L, Pustejovsky J, Boston, Massachusetts, USA: Association for Computational Linguistics 2004:17–24.

[29] Hyland K: **Hedging in Academic Writing and EAP Textbooks**. *English for Specific Purposes* 1994, **13**(3):239–256.

[30] Medlock B, Briscoe T: **Weakly Supervised Learning for Hedge Classification in Scientific Literature**. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, Prague, Czech Republic: Association for Computational Linguistics 2007:992–999, [http://www.aclweb.org/anthology/P/P07/P07-1125].

[31] Lang D: **Consultant Report - Natural Language Processing in the Health Care Industry**. *PhD thesis*, Cincinnati Children's Hospital Medical Center 2007.

[32] Pestian JP, Brew C, Matykiewicz P, Hovermale D, Johnson N, Cohen KB, Duch W: **A shared task involving multi-label classification of clinical free text**. In *Biological, translational, and clinical language processing*, Prague, Czech Republic: Association for Computational Linguistics 2007:97–104, [http://www.aclweb.org/anthology/W/W07/W07-1013].

[33] Berger AL, Pietra SD, Pietra VJD: **A Maximum Entropy Approach to Natural Language Processing**. *Computational Linguistics* 1996, **22**:39–71, [citeseer.ist.psu.edu/berger96maximum.html].

[34] Farkas R, Szarvas Gy, Kocsor A: **Named entity recognition for Hungarian using various machine learning algorithms**. *Acta Cybern.* 2006, **17**(3):633–646.

[35] Szarvas Gy, Farkas R, Kocsor A: **A Multilingual Named Entity Recognition System Using Boosting and C4.5 Decision Tree Learning Algorithms**. In *Discovery Science* 2006:267–278.

[36] Markert K, Nissim M: **SemEval-2007 Task 08: Metonymy Resolution at SemEval-2007**. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, Prague, Czech Republic: Association for Computational Linguistics 2007:36–41, [http://www.aclweb.org/anthology/W/W07/W07-2007].

[37] Farkas R, Szarvas Gy, Ormándi R: **Improving a State-of-the-Art Named Entity Recognition System Using the World Wide Web**. In *Industrial Conference on Data Mining* 2007:163–172.

[38] Farkas R, Vincze V, Nagy I, Ormándi R, Szarvas Gy, Almási A: **Web based lemmatisation of Named Entities**. In *Accepted for 11th International Conference on Text, Speech and Dialogue* 2008.

[39] Farkas R, Simon E, Szarvas Gy, Varga D: **GYDER: Maxent Metonymy Resolution**. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, Prague, Czech Republic: Association for Computational Linguistics 2007:161–164, [http://www.aclweb.org/anthology/W/W07/W07-2033].

[40] Szarvas Gy, Farkas R, Busa-Fekete R: **State-of-the-art anonymisation of medical records using an iterative machine learning framework**. *J Am Med Inform Assoc* 2007, **14**(5):574–580, [http://www.jamia.org/cgi/content/abstract/M2441v1].

[41] Szarvas Gy, Iván S, Bánhalmi A, Csirik J: **Automatic Extraction of Semantic Content from Medical Discharge Records**. *WSEAS Transaction on Systems and Control* 2006, **1**(2):312–317.

[42] Szarvas Gy: **Hedge classification in biomedical texts with a weakly supervised selection of keywords**. In *Accepted for the 45th Annual Meeting of the Association of Computational Linguistics*, Columbus, Ohio, United States of America: Association for Computational Linguistics 2008.

[43] Szarvas Gy, Vincze V, Farkas R, Csirik J: **The BioScope corpus: annotation for negation, uncertainty and their scope in biomedical texts**. In *Accepted for Biological, translational, and clinical language processing (BioNLP Workshop of ACL)*, Columbus, Ohio, United States of America: Association for Computational Linguistics 2008.

[44] Farkas R, Szarvas Gy: **Automatic construction of rule-based ICD-9-CM coding systems**. *BMC Bioinformatics* 2008, **9**(3), [http://www.biomedcentral.com/1471-2105/9/S3/S10].