Szegedi Tudományegyetem
Mesterséges Intelligencia kutatócsoport

Gépi tanulási módszerek az alkalmazott Infomráció-kinyerésben

PhD-értekezés tézisei

Farkas Richárd

Témavezető:
Dr. Csirik János

Szegedi Tudományegyetem
Természettudományi és Informatikai Kar
Informatika Doktori Iskola

2009

Bevezetés

Az összefoglaló ismerteti a szerző *Machine Learning techniques for applied Information Extraction* (Gépi tanulási technikák az alkalmazott Információ-kinyerésben) című disszertációjának tartalmát, illetve főbb eredményeit. A disszertáció témakörét a Mesterséges Intelligencia két részterülete képzi: a Gépi Tanulás és annak alkalmazása a Nyelvtechnológia (más néven Számítógépes Nyelvfeldolgozás) területén.

Az Internet gyors növekedésének és a globalizációs folyamatoknak köszönhetően az elérhető információ mennyisége soha nem látott ütemben nő. Az új adatok nagyobbik része szöveges formában van jelen, hiszen például a weboldalakat elsődlegesen emberek készítik a célból, hogy más emberek elolvassák azokat. A szöveges adatok mennyiségéből adódóan, azok feldolgozása pusztán emberi erővel lehetetlen, az informatika eszközeinek bevonása szükséges. A természetes nyelvű szövegek megértését és generálását célzó területet nevezzük *Nyelvtechnológiá*nak.

Az *Információ-kinyerés* a Nyelvtechnológia azon részterülete, melynek célja jól definiált információ kiemelése természetes nyelvi szövegekből, azaz weboldalakról, belső szervezeti dokumentumokból, újság hírekből, tudományos publikációkból, e-mailekből, blogokból stb. Az input tehát strukturálatlan adat (szöveg), az output pedig strukturált rekordok halmaza. Az Információ-kinyerésnek számos ipari alkalmazása létezik, többek közt üzleti információgyűjtés vállalkozásokról vagy fehérje interakcióban résztvevő párok kigyűjtése biológiai témájú szabadalmakból. Információ-kinyerési eszközökkel például az "*Eric Schmidt joined Google as chairman and chief executive officer in 2001.*" mondatból az alábbi információhármas emelhető ki:

{COMPANY=*Google*, CEO=*Eric Schmidt*, CEO_START_DATE=*2001*}

Az Információ-kinyerési problémák klasszikus megoldásában manuálisan épített döntési szabályrendszerekkel történik a célinformáció azonosítása. Ezeknek a szabályrendszereknek a kialakítása, majd karbantartása igen költséges, mert a döntési rendszer megkonstruálásakor nem csak az adott területhez, de a szabályrendszer felépítéséhez is értenie kell a készítőnek. *Gépi tanulás*i technikák segítségével ezeket a szabályokat automatikusan építi fel a gép a szakértő által adott példák általánosításával.

A disszertáció célja

A disszertációban bemutatunk számos gépi tanulási technikát, és azok valós életbeli Információkinyerési problémákban való alkalmazhatóságát vizsgáltuk. A gépi tanulási módszerek közt ritkán alkalmazott eljárásokkal és újszerű technikákkal is kísérleteztünk. A tárgyalt feladatok széles skálát ölelnek fel, a nyelv-független névelem (Named Entity) felismeréstől (token-sorozatok címkézése) kezdve, a névelem normalizáción át a vélemény-detekcióig.

Az összefoglaló tematikája

Az összefoglaló szerkezete a tézis felépítését követi, a disszertáció két fő témáját tárgyalja. Az első rész (3-5 fejezetek) a felügyelt gépi tanulási módszereket ismerteti, míg a második (6-8 fejezetek) a tanító adatbázison kívüli információk felhasználására, rendszerbe integrálási lehetőségeire mutat példát. 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 – Felügyelt gépi tanulási módszerek az Információ-kinyerésben

Az Információ-kinyerési problémák legelterjedtebb megoldásai az ún. felügyelt tanulási környezetet használják. Itt egy kézzel jelölt tanító adatbázis áll rendelkezésre, és a cél egy olyan modell építése, amely korábban nem látott egyedeken is jó döntést hoz.

Felügyelt gépi tanulási megközelítések

Minden gépi tanuló algoritmushoz adható olyan tanulási feladat, amelyiken más algoritmusok hatékonyabban teljesítenek [1], ezért érdemes a tanuló algoritmust feladat-specifikusan megválasztani. Az Információ-kinyerésben alkalmazott felügyelt tanulási módszerek két különböző megközelítést követhetnek:

A **token-alapú modellek** a szöveg egyes tokenjeire egymástól függetlenül hozzák meg a döntést. A szavak sorrendiségét, azok egymásra gyakorolt hatását a jellemzőtér foglalja magában. Azaz általában egy token jellemzőkkel való leírásakor a környezetében szereplő szavak jellemzőit is felvesszük (pl.: *a megelőző szó szerepel X listában*). Az ilyen modellekkel lehetőségünk nyílik szavanként mintavételezni, ami például kiegyensúlyozott tanító halmazok létrehozásához elengedhetetlen.

A **szekvencia-alapú modellek** egy egész szekvenciát (általában egy mondatot) egyben vizsgálnak, és céljuk a legvalószínűbb címke-sorozat megtalálása. Ezek a modellek elvetik a szavak függetlenségére tett feltételezést, ennek ára, hogy a tanítás időigénye nagyságrendekkel megnő a token-alapú módszerekhez viszonyítva.

E két megközelítési módot empirikusan összehasonlítottuk és számos osztályozó algoritmus hatékonyságát teszteltük többségében *névelem felismerés*i (Named Entity Recognition, NER) [2; 3] adathalmazokon. Ezen osztályozási problémáknak speciális tulajdonságai a nagy dimenziós (általában több tízezres) jellemzőtér, illetve a ritka és diszkrét jellemzők.

A szövegben található névelem-kifejezések (tulajdonnevek, nevek akronimjai, időt, mennyiséget jelölő kifejezések, azonosítók, e-mail címek, közigazgatási címek, telefonszámok stb.) azonosítása és osztályozása az Információ-kinyerés egyik alapvető feladata. A névelemek legtöbbször fontos információval bírnak a dokumentum tartalmára nézve, és emiatt az emberi Információ-kinyerés célpontjai gyakran névelemek. A feladat a gépi fordítás területén is nagy jelentőséggel bír, hiszen a gépi fordítórendszerekben a tulajdonnevek más szabályok szerint fordítandók, mint a köznevek. Néha maga a NER lehet önálló végalkalmazás is, erre jó példa az anonimizálás [4], ahol a névelemek felismerése után azok eltávolítását vagy lecserélését kell csak végrehajtani, hogy a személyes adatoktól megtisztítsuk a dokumentumot. A NERben nem csak az egyedekre referáló frázisok szövegbeli azonosítása a cél, hanem azok szemantikai osztályokba (például *személy*, *szervezet*, *földrajzi név* stb.) sorolása is.

Meta-tanulás

A *meta-tanuló*k különböző tanulók (különböző tanuló algoritmusok vagy ugyanazon algoritmus paraméterezett változatai) együttes alkalmazásával jönnek létre. Több tanuló-kombinációs módszer ismert, melyek általában jobb modellt eredményeznek, mint az alapjául szolgáló algoritmusok. Ezen hibrid módszerek sikerének kulcsa, hogy az alap-tanulók a tanító adatbázis különböző aspektusait ragadják meg.

A disszertációban számos meta-tanulóval végzett kísérletet mutatunk be angol és magyar NER korpuszokon. Az alkalmazott kombinációs módszerek szignifikánsan jobb eredményeket értek el, mint a legjobb egyéni alap-tanulók, és az eredmények alátámasztják, hogy a "gyengébb" tanulóknak is van hozzáadott értéke meta-tanulás során.

Az ismert meta-tanulók alkalmazásával elért eredmények mellett bemutatásra került egy újszerű eljárás is [3]. Ez a megközelítésünk néhány kisebb, átfedő jellemzőhalmazt választ ki az eredetei jellemzőtérből, majd az ezekkel tanított modelleket ötvözi.

Meta-tanulók felhasználásával egy komplex statisztikai NER rendszert állítottunk össze [3]. Ebben a rendszerben a fent röviden bemutatott jellemzőteret felosztó majd újrakombináló módszer is egy-egy boostingolt (AdaBoostM1) döntési fát épít, és kiegészül néhány utó-feldolgozási lépéssel.

Ez a komplex rendszer több adatbázison versenyképes eredményt ért el (a legjobb publikált rendszerekhez hasonlítva), míg eltérő elvi alapokon nyugszik. Ez utóbbi tény különösen alkalmassá teszi más külső NER rendszerekkel való kombinálásra.

Információ-kinyerő rendszerek adaptálhatósága

A felügyelt gépi tanulási módszerek általában jó pontosságot érnek el az automatikus címkézési feladatokon ismeretlen szövegek esetén, ha a tesztszöveg karakterisztikája megegyezik a tanító adatbáziséval. Azonban, ha a célszövegek jellemzői megváltoznak (például gazdasági hírekről orvosi zárójelentésekre térünk át), akkor új tanító adatbázisra van szükség.

A NER rendszerünkkel magyar [5] és angol nyelvű [6] szövegeken is végeztünk kísérleteket, és azt tapasztaltuk, hogy annak viselkedése nem változik, az egyes komponensek hozzáadott értéke megegyezik a két nyelven. Ezek az eredmények nem kézenfekvőek figyelembe véve, hogy a magyar nyelv számos speciális tulajdonsággal rendelkezik (pl.: szabad szórend, agglutináció).

2007-ben az újsághírekre kifejlesztett NER modellünket adaptáltuk orvosi szövegek anonimizálására, ahol a cél Személyes Egészségügyi Információk (mint például a páciensek és orvosok nevei, azonosító számok, telefonszámok, kórháznevek, dátumok stb.) azonosítása zárójelentésekben. Apró módosítások után a rendszer igen jó eredményeket ért el egy nyílt, nemzetközi versenyen [4].

Később a magyar és angol NER modellünket teszteltük olyan szövegeken is, amik témájukban különböznek azok tanító adatbázisától [7]. Ezen kísérletekben a kiértékelést egy mondat-párhuzamosító eszközön keresztül, indirekt módon végeztük el. A mondat-párhuzamosítás célja, hogy kétnyelvű párhuzamos korpuszokban automatikusan megtalálja, mely kiindulási nyelvi mondat(ok)nak mely célnyelvi mondat(ok) a fordítása(i). Az összerendeléshez gyakran használják a mondatokban előforduló nagybetűs szavak számát, mint indikátort. Ennél jóval pontosabb képet kapunk, ha a tulajdonnevek számából indulunk ki. Ennek az ötletnek a felhasználásával sikerült a mondat-párhuzamosító algoritmusunkon javítani.

Az automatikus modell-adaptáció irányába tett első lépésként megvizsgáltuk különböző domainek

statisztikai különbségeit a tagadás és feltételes mód detektálási feladatokhoz. Ezek a részfeladatok igen fontosak számos Információ-kinyerési alkalmazásban, például orvosi zárójelentések feldolgozása során jelentőssséggel bír, hogy egy tünet megléte tagadva van, feltételes módban áll vagy tényként közli a dokumentum. A két nyelvi jelenség automatikus azonosítására alkalmas módszerek fejlesztése és kiértékelése céljából létrehoztuk a BioScope korpuszt [8]. A korpusz három részből áll: orvosi zárójelentésekből, biológiai cikkek absztraktjaiból és teljes biológiai cikkekből. A szövegeken manuálisan annotálásra került a tagadást és feltételes módot kifejező kulcsszavak és azok nyelvi hatókörei. A részkorpuszok statisztikai elemzéséből az derül ki, hogy ugyanazon feladat minimálisan különböző témájú szövegeken (biológiai absztraktok és teljes cikkek), illetve nagyon hasonló feladatok ugyanazon szövegeken (tagadás és feltételes mód) nagyon eltérő tulajdonságokkal rendelkeznek, ezért az automatikus modell-adaptáció komplikált feladat.

II. rész – Külső információs források kiaknázása Információ-kinyerési feladatokban

A felügyelt tanulásnál minden feladathoz szükség van egy megfelelő méretű tanító adatbázisra, még akkor is, ha a feladatok csak kis mértékben térnek el egymástól (pl.: NER gazdasági és sporthíreken). A dolgozat II. részében különböző külső információs források felhasználási lehetőségeit vizsgáltuk Információ-kinyerési problémák megoldására. Ezen kísérletek célja, hogy a szükséges tanító példák számát minimalizáljuk, így az újabb problémákra történő adaptáció gyorsabbá és költséghatékonyabbá válik.

A WWW, mint jelöletlen korpusz

A *részben felügyelt tanulás* során a jelölt tanító adatbázis mellett jelöletlen adatból is próbálunk hasznos információt kinyerni. A Nyelvtechnológiai problémák egyedi tulajdonságai speciális részben felügyelt tanulási technikák alkalmazását követelik meg, illetve teszik lehetővé. Ezen Nyelvtechnológiai feladatok megoldása során kihasználhatjuk, hogy (elsősorban az Internetnek köszönhetően) jelöletlen adat (szöveg) szinte végtelen mennyiségben rendelkezésre áll. Ezen felül a WWW-t, mint jelöletlen korpuszt folyamatosan bejáró alkalmazások naprakészek tudnak maradni.

A NER rendszerünk hibáinak egy jelentős része abból származik, hogy a gép nincs birtokában az általános emberi tudásnak. Ha birtokában lenne, akkor nem követne el olyan hibákat, mint hogy a *'Budapesttől Szeged'* kifejezést egyetlen névelemnek jelölni vagy a *'Real Madrid'*-ot földrajzi helynek címkézni. Az ilyen hibák elkerülésére több WWW-alapú utófeldolgozó eljárást dolgoztunk ki [9]:

- bizonytalan esetekben megkeressük a frázis Interneten leggyakrabban használt szemantikai osztályát,
- a frázis határait a szövegben kiterjesztjük, ha az gyakoriságok alapján indokolható,
- szétvágunk hosszú frázisokat, ha a részek Internet-gyakoriságuk alapján egymást követő névelemeknek tekinthetőek.

A névelemek azonosítása után, azok szótövének megtalálása is feladat, hiszen a tárolás és keresés csak a normalizált formán hajtható végre. A köznevek szótövesítése viszonylag egyszerű feladat, hiszen a lehetséges szótövek felsorolhatóak [10]. A tulajdonnevek esetében ez nem lehetséges, ráadásul az agglutináló nyelvekben (mint a magyar is) a főnevekhez igen sok fajta rag adható. Vannak azonban olyan szótövek is, amelyek úgy végződnek, mint egy potenciális rag (magyarban például *Pannon* a szótő és nem *Pann*, angolban *Adidas* és *McDonald's* és nem *Adida* vagy *McDonald*), ezért nem lehet automatikusan eltávolítani azokat.

A szótő megtalálására a következő eljárást javasoltuk [11]: minden végződésre (potenciális ragra) megnézzük a teljes alak és a szótő-jelölt gyakoriságát az Interneten. A hipotézisünk az, hogy a gyakoriságok aránya alapján eldönthető, hogy a szótő-jelölt valóban egy ragozatlan névelem-e.

Szakértői rendszerek integrálása gépi tanulási modellekbe

Igen hasznos (de kevésbé vizsgált) külső információforrásnak bizonyulnak az emberi erőforrással épített taxonómiák, döntési szabályrendszerek, mint például a klinikai információ-kinyerés során alkalmazható, ezer évek tudását magába foglaló orvosi enciklopédiák. A dolgozatban bemutatunk több módszert, amelyekben ilyen szabályrendszereket használtunk fel gépi tanulási modellek támogatására az automatikus BNO (Betegségek Nemzetközi Osztályozása) kódolási [12] és betegség-azonosítási feladatokban [13].

A *BNO kódolás* tünetek, betegségek és kezelések egységes kódolását teszi lehetővé orvosi dokumentumokban. Ez az alapja a világ legtöbb országában az egészségügyi intézmények és biztosítók közti elszámolásnak. A BNO kódolást napjainkban manuálisan végzik, annak éves költségét az USA-ban 25 milliárd dollárra becslik [14], így annak automatizálása igen komoly gazdasági potenciált rejt.

Megterveztünk és implementáltunk egy BNO kódoló rendszert, amellyel részt vettünk egy nemzetközi nyílt versenyen [15]. A rendszer felismerte a tagadásban, illetve feltételes módban álló szövegrészeket, azonosította a szakkifejezéseket, majd meghozta a több-címkés osztályozási döntést. A szakkifejezések azonosítására (aminek központi szerepe volt) három különböző módszert fejlesztettünk ki, amelyek a tanító adatbázisból nyert összefüggések és az enciklopédiából származó tudás (szabály-alapú rendszer) együttes kiaknázását valósították meg [12]:

Szabály-alapú rendszer kiterjesztése:

A szabály-alapú rendszerből hiányzó szinonimákat, eltérő alakokat a következő módon azonosítottuk. Azokból a dokumentumokból, amikre a tanító adatbázison a szabályok nem aktivizálódtak (hibás negatívok), egy statisztikai rendszert tanítottunk, és az általa megtalált kifejezéseket hozzáadtuk a szabályhalmazhoz.

Statisztikai rendszer kiterjesztése:

A szabály-alapú rendszer predikcióit beépítettük az osztályozó módszer jellemzőterébe. Ily módon a statisztikai rendszer a szabályrendszer és a tanító szöveghalmaz szabályosságait egyszerre volt képes megtanulni.

Hibrid modell:

Több-címkés környezetben kézenfekvő kombinálási stratégia a két rendszer által predikált címkehalmaz uniójának vagy metszetének választása. A mi esetünkben az unió szolgáltatta a jobb

eredményeket. Ebben az esetben a szabály-alapú rendszer és a gépi tanuló rendszer egymástól függetlenül hozta meg a döntést.

Az orvosi dokumentumok osztályozása aszerint, hogy egy adott betegség a szóban forgó páciensnél megállapításra került-e, a BNO kódoláshoz nagyon hasonló feladat. 2008-ban került megrendezésre az *Obesity Challenge*[1] elnevezésű verseny, ahol a feladat olyan módszerek kidolgozása volt, amelyeknek a túlsúlyosság és ahhoz kapcsolódó rendellenességek szövegbeli detektálását kellett megoldani. A versenynek két részfeladata volt, egyrészt a szövegben egyértelműen jelen lévő bizonyítékok alapján kellet döntést hozni, másrészt a dokumentum egésze alapján orvos szakértők intuitív osztályozását kellett reprodukálni.

A megközelítésünk egy kiterjesztett lista-alapú kereső eljárás volt, ami az illesztésnél figyelembe veszi a dokumentum struktúráját és a megtalált kulcsszó (szakkifejezés) mondatkörnyezetét is. A lista felépítéséhez először statisztikai módszerekkel kigyűjtöttük a tanító adatbázisból az adott rendellenességre legjobban jellemző szakkifejezéseket, rövidítéseket és azok különböző írásmódjait, majd a MedlinePlus enciklopédia[2] segítségével tovább bővítettük azt [13].

Dokumentumközi információk kiaknázása

Alkalmazott Információ-kinyerési problémáknál a feldolgozandó dokumentumok általában nem függetlenek egymástól. A dokumentumok közti, gráf jellegű, külső információk alkalmazását ismerteti a dolgozat utolsó fejezete.

A biológiai publikációk nagy mennyiségben tartalmaznak hasznos információt génekről, fehérjékről és azok különböző szituációbeli viselkedéséről. A biológiai NER génnevek és vegyi anyagok neveinek szövegbeli azonosítását célozza meg. A feldolgozás következő lépése ezen nevek egyedi azonosítóhoz rendelése (*normalizáció*) [16], ugyanis egy egyedre több néven is hivatkozhatunk és ugyanazzal a névvel több gént is illethetnek (például az *IL-21* vonatkozhat a 27189, 50616 vagy 59067 Entrez-Gene azonosítójú egyedekre is). Ez utóbbi probléma (azaz egy adott génnév szövegkörnyezet-függő normalizálását) megoldását *génnév-egyértelműsítés*nek nevezzük.

A fő észrevételünk az volt, hogy a dokumentum szerzője konzekvensen használja a génneveket, azaz egy névvel mindig pontosan egy gént illet publikációiban. Általánosítva ezt megvizsgáltuk, hogy ugyanez igaz-e az adott szerző társszerzőire, illetve azok társszerzőire. Olyan módszereket dolgoztunk ki, amelyeknél az ún. szerzőségi gráfból [17] és magukból a publikációk szövegeiből nyert információt együttesen aknáztuk ki [18].

Hasonló gráf-jellegű információkat használtunk fel egy vélemény-detekciós [19] feladat megoldása során is. A *vélemény-detekció* célja különböző irányultságú vélemények összegyűjtése, érzelmi töltetek azonosítása folyó szövegekből. Ez a terület egy évtizede került az akadémiai és ipari kutatások középpontjába, ugyanis az Internet elterjedésével a felhasználók gyakran fejezik ki véleményüket bárki által hozzáférhető platformokon (elsősorban blogokon, fórumokon) és ezekből a forrásokból hasznos információ nyerhető ki. Az információ segíthet például vállalkozásokat abban, hogy visszajelzést kapjanak arról, mit gondolnak vásárlóik termékeikről vagy politikai pártoknak, hogy megszervezzék kampányukat.

A hipotézisünk itt az volt, hogy a különböző véleményen lévő hozzászólók gyakran reagálnak egymás hozzászólásaira, ezért megkonstruáltuk a reagálási gráfot, aminek csúcspontjai a hozzászólókat

[1] az Informatics for Integrating Biology and the Bedside (I2B2) szervezésében. www.i2b2.org/NLP/

[2] http://www.nlm.nih.gov/medlineplus/encyclopedia.html

reprezentálják, míg akkor fut irányított él A és B pontok közt, ha A legalább egyszer reagált B valamely hozzászólására. Az élek súlya a reagálások száma. A megközelítésünk ismét együttesen használja fel a gráfból és a szövegekből nyert információkat. A módszert empirikusan teszteltük egy magyar nyelvű fórumon, ahol a cél annak automatikus eldöntése volt, hogy a fórum témájában az adott hozzászóló milyen véleményen van [20].

Eredmények fejezetenként

Ebben a szakaszban a tézis minden fejezetének eredményét összegezzük, és áttekintjük, mely publikációk támasztják alá a tézis egyes fejezeteinek eredményeit.

- I. rész – **Felügyelt gépi tanulási módszerek az Információ-kinyerésben**
  - 3. fejezet: **Felügyelt modellek az Információ-kinyerésben**

    A szerző és társai megterveztek és kifejlesztettek egy gépi tanulási módszereken alapuló *névelem-felismerő keretrendszer*t, amely több nemzetközi referencia adatbázison is kiemelkedő eredményt ért el [2; 3]. A rendszer tervezésének egyik alappillére volt gépi tanulási algoritmusok viselkedésének empirikus vizsgálata. A szerző főbb megállapításai a következők:

    * Az Információ-kinyerési feladatoknál a jellemzőtér nagy mennyiségű diszkrét jellemzőt tartalmaz, ami döntési fák és generatív modellek használatát implikálja.
    * A döntési fa tanítási ideje nagyságrendekkel kisebb, mint a többi vizsgált tanulóé.
    * A döntési fa modellje közvetlenül értelmezhető az ember számára is.
    * A generatív modellek (pl.: Logisztikus Regresszió [21], Feltételes Valószínűségi Mezők [22]) kimente címkék feletti valószínűségi eloszlás, ami alkalmazások számára igen fontos lehet, mint megbízhatósági mérték.

    Összességében a szerző döntési fák alkalmazását javasolja a fejlesztés, kísérletezés folyamán, annak gyors tanítási ideje és a tanult modell interpretálhatósága miatt. A generatív modellek általában néhány százalékkal jobban teljesítenek (hosszabb tanítási idő árán), így azok végső modellként való alkalmazása ajánlott.

  - 4. fejezet: **Tanuló-kombinációs módszerek**

    A szerző empirikus vizsgálatokkal bizonyította, hogy egyszerű *tanuló-kombinációs séma*k is szignifikáns javulást eredményeznek. Ezen kombinációs sémák egyszerű, gyors alaptanulókat használva is általában jobb eredményeket képesek elérni, mint a szofisztikáltabb, időigényesebb tanulók önmagukban. Ezt a szerző empirikusan validálta különböző metatanulókkal magyar és angol nyelvű NER adatbázisokon [2; 3; 23]. A disszertáció 4. fejezetben bemutatásra kerül a szerző újszerű kombinációs algoritmusa, amely a jellemzőtér felosztásán és a tanult modellek kombinációján alapul [3].

    A szerző és társai által kidolgozott, komplex NER rendszer is meta-tanuló algoritmusok alkalmazására épül. A módszer több adatbázison versenyképes eredményt ért el a legjobb, publikált rendszerekhez hasonlítva (89,2% és 94,77% F-mérték az angol, illetve magyar adatbázisokon), míg azoktól különböző elméleti alapokon nyugszik. Ez utóbbi tény különösen alkalmassá teszi más külső NER rendszerekkel való kombinálásra.

– 5. fejezet: **Információ-kinyerő rendszerek adaptálhatósága**

A szerző és társai a NER rendszert eredményesen alkalmazták orvosi zárójelentések szövegein is (angol nyelvű kórházi dokumentumokban *betegek, orvosok neveit, a beteg életkorát, telefonszámokat, azonosítókat, helyneveket, kórházneveket és dátumokat* azonosítottak). Ez az orvosi dokumentumokon működő rendszer a második legjobb eredményt érte el egy anonimizáló rendszerek kiértékelésére szolgáló adatbázison [4].

A [7] munkában a szerző hozzájárulása a NER rendszerek alkalmazásának ötlete a mondatpárhuzamosító eljárásban, míg a BioScope korpusszal kapcsolatosan [8] a szerző részkorpuszok statisztikai összehasonlítását végezte el.

A fejezet eredményei első pillantásra meglepőek lehetnek. Azt találtuk, hogy az Információkinyerés középső rétegeiben – ahol a nyelv-specifikus információk (morfológiai jegyek, POS kódok) már a jellemzőtérbe vannak kódolva – ugyanazon statisztikai rendszer tulajdonképpen változtatás nélkül több nyelvre is ugyanolyan pontossággal működik. A szöveg domainjének megváltozása esetén a rendszeren nagyobb adaptációs lépéseket kell elvégezni.

- II. rész – **Külső információs források kiaknázása Információ-kinyerési feladatokban**

  – 6. fejezet: **A WWW, mint jelöletlen korpusz**

  A szerző társaival több Internet-gyakoriság-alapú NER utófeldolgozó algoritmust dolgozott ki, amelynek hatékonyságát empirikusan validálta referencia adatbázisokon [9]. Hasonló statisztikai módszer segítségével egy tulajdonnév-lemmatizálási eljárás is kidolgozásra került [11]. Ezen eredmények igazolják, hogy habár a WWW igen zajos, a redundanciát kihasználva hasznos információval szolgálhat különböző Információ-kinyerési problémák megoldásához.

  A szerző hozzájárulása a gyakoriság-alapú NER utófeldolgozó módszerek közül a leggyakoribb szerep és frázis kiterjesztési heurisztikákban volt meghatározó [9]. A tulajdonnévlemmatizálási eredmények elérésének érdekében a szerző tervezte meg a jellemzőteret és hajtotta végre a gépi tanulási kísérletek többségét [11].

  – 7. fejezet: **Szakértői rendszerek integrálása gépi tanulási modellekbe**

  A szerző és társa egy kórházi leletek betegségkódokkal, illetve BNO-kódokkal való automatikus címkézésére alkalmas rendszert fejlesztett ki. Ez a rendszer egy automatikus klinikai kódoló rendszerek kiértékelésére szervezett versenyen a legjobb pontosságot érte el [15]. A verseny tapasztalatai alapján a szerző és társa egy szakértői és statisztikai rendszerek kombinációján alapuló modellt dolgozott ki, mely képes a rendelkezésre álló szabályalapú rendszereket címkézett példák felhasználásával tovább pontosítani [12; 13]. Az ide kapcsolódó kísérletek során a szerző hozzájárulása a gépi tanulási modellek kiválasztásában és implementálásában, a szabály-alapú rendszerek integrálásában és kivitelezésében volt meghatározó.

  A disszertációban bemutatott kísérleti eredmények demonstrálják a javasolt módszer valós életbeli alkalmazhatóságát és azt, hogy csak felületi nyelvi elemzést használó rendszerek kielégítő pontosságot képesek elérni klinikai Információ-kinyerési feladatokon.

  – 8. fejezet: **Dokumentumközi kapcsolatok kiaknázása**

  A szerző újszerű módon kombinálta a dokumentumközi kapcsolatokból és a szövegből nyerhető információkat. Az ún. szerzőségi gráf felhasználásával igen jó eredményeket

			Fejezet					
			3	4	5	6	7	8
ACTA	2006	[2]	•	•				
LREC	2006	[5]	•					
SEMEVAL	2007	[24]	•					
DS	2006	[3]		•				
JAMIA	2007	[4]			•			
BIONLP	2008	[8]			•			
ICDM	2007	[9]				•		
TSD	2008	[11]				•		
BMC	2007	[12]					•	
JAMIA	2009	[13]					•	
BMC	2008	[18]						•
DMIIP	2008	[20]						•

Table 1: A fejezetek (tézis pontok) és publikációk közti kapcsolat.

ért el a génnév-egyértelműsítési feladaton (a végső rendszer az egyértelműsítési döntés 97,22%-os pontossággal hozza meg 100%-os fedés mellett) [18]. Hasonló módszer segítségével, a szerző társaival egy vélemény-detekciós probléma megoldása során a fórumozók válaszadási-gráfjából nyert ki információt [20].

A [18] publikáció eredményei teljes egészében a szerző munkája. A [20]-ban a szerző ötlete volt a reagálási gráf használata a vélemény-detekciós feladatban, a kísérleteket a társszerző végezte el.

Eredmények publikációnként

Az alábbiakban felsoroljuk a fontosabb publikációkban[3] szereplő eredményeket, amelyeket a szerző a *saját* eredményeinek tekint. Megemlítjük, hogy a rendszer-szintű eredményeket, értékeléseket minden esetben közös eredménynek tekintjük, mivel lehetetlen számszerűsíteni, hogy az elért pontosságértékekben az egyes rendszerelemek a hozzájárulása milyen mértékű.

A felsorolásból kihagytuk a [18] cikket, mert az abban ismertetett eredmények a szerző egyedüli munkáját képezik, illetva a [5] cikket, mert annak minden eredményét a szerzők közös munkájának tekintjük (illetve nem tekintjük a disszertáció szerves részének). A [24] cikkben összefoglalt eredményekhez a szerző csak kis mértékben járult hozzá, így ezeket sem soroljuk a disszertáció főbb eredményei közé.

- ACTA 2006 [2]
  - Felügyelt gépi tanulási algoritmusok empirikus összehasonlítása.
  - Kísérletek a szavaztatási algoritmussal.

[3] A szerző teljes publikációs listája elérhető a http://www.inf.u-szeged.hu/~rfarkas/publications.html oldalon.

- SEMEVAL 2007 [24]
  - A C4.5 és Logisztikus Regresszió összehasonlító elemzése.
- DS 2006 [3]
  - A komplex NER modell architektúrája.
  - A jellemzőtér felosztási és újra-kombinációs módszer.
  - Boosting kísérletek.
  - Utófeldolgozási szabályok.
- JAMIA 2007 [4]
  - Trigger-alapú modell-kombinációs megközelítés.
  - Standardizációs fázis.
- BIONLP 2008 [8]
  - Részkorpuszok statisztikai összehasonlítása.
- ICDM 2007 [9]
  - Frázis-kiterjesztési heurisztika.
  - A leggyakoribb szerep heurisztika.
- TSD 2008 [11]
  - Jellemzőtér kialakítása a tulajdonnév-lemmatizálási feladathoz.
  - Gépi tanulási kísérletek.
- BMC 2007 [12]
  - Szakértői rendszereket és statisztikai módszereket kombináló eljárások.
  - Negációt és feltételes módot detektáló modul.
  - Több-címkés osztályozási megközelítések.
  - Kísérletek gépi tanulási módszerekkel.
- JAMIA 2009 [13]
  - Statisztikai eljárások a kulcsszó azonosításra.
  - Statisztikai eljárások a szövegkörnyezet azonosításra.
- DMIIP 2008 [20]
  - A reagálási gráf alkalmazásának ötlete vélemény-detekciós feladatoknál.

References

[1] Wolpert DH, Waters R: **The relationship between PAC, the statistical physics framework, the Bayesian framework, and the VC framework**. Tech. rep., The Santa Fe Institute 1994.

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

[3] Szarvas Gy, Farkas R, Kocsor A: **A Multilingual Named Entity Recognition System Using Boosting and C4.5 Decision Tree Learning Algorithms**. *DS2006, LNAI* 2006, **4265**:267–278.

[4] Szarvas Gy, Farkas R, Busa-Fekete R: **State-of-the-art anonymisation of medical records using an iterative machine learning framework**. *Journal of the American Medical Informatics Association* 2007, **14**(5):574–580, [http://www.jamia.org/cgi/content/abstract/M2441v1].

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

[6] Sang TK, F E, 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.

[7] Tóth K, Farkas R, Kocsor A: **Sentence Alignment of Hungarian-English Parallel Corpora Using a Hybrid Algorithm.** *Acta Cybernetica* 2008, **18**(3):463–478.

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

[9] Farkas R, Szarvas Gy, Ormándi R: **Improving a State-of-the-Art Named Entity Recognition System Using the World Wide Web**. *ICDM2007, LNCS* 2007, **4597**:163–172.

[10] Halácsy P, Trón V: **Benefits of Resource-Based Stemming in Hungarian Information Retrieval**. In *CLEF* 2006:99–106.

[11] Farkas R, Vincze V, Nagy I, Ormándi R, Szarvas Gy, Almási A: **Web based lemmatisation of Named Entities.** In *Proceedings of the 11th International Conference on Text, Speech and Dialogue* 2008:53–60.

[12] 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].

[13] Farkas R, Szarvas Gy, Hegedűs I, Almási A, Vincze V, Ormándi R, Busa-Fekete R: **Semi-automated construction of decision rules to predict morbidities from clinical texts.** *Journal of the American Medical Informatics Association* 2009, **accepted for publication**.

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

[15] 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].

[16] Hirschman L, Colosimo M, Morgan A, Yeh A: **Overview of BioCreAtIvE task 1B: normalized gene lists**. *BMC Bioinformatics* 2005, **6**(Suppl 1):S11.

[17] Barabasi AL, Jeong H, Neda Z, Ravasz E, Schubert A, Vicsek T: **Evolution of the social network of scientific collaborations**. *Physica A: Statistical Mechanics and its Applications* 2002, **311**(3-4):590–614.

[18] Farkas R: **The strength of co-authorship in gene name disambiguation**. *BMC Bioinformatics* 2008, **9**, [http://dx.doi.org/10.1186/1471-2105-9-69].

[19] Ghose A, Ipeirotis PG, Sundararajan A: **Opinion Mining using Econometrics: A Case Study on Reputation Systems**. In *ACL* 2007.

[20] Berend G, Farkas R: **Opinion Mining in Hungarian based on textual and graphical clues.** In *Proceedings of the 4th International Symposium on Data Mining and Intelligent Informaion Processing* 2008.

[21] le Cessie S, van Houwelingen J: **Ridge Estimators in Logistic Regression.** *Applied Statistics* 1992, **41**:191–201.

[22] Sutton C, Mccallum A: **Introduction to Conditional Random Fields for Relational Learning.** In *Introduction to Statistical Relational Learning*. Edited by Getoor L, Taskar B, MIT Press 2006.

[23] Farkas R, Szarvas Gy, Csirik J: **Special Semi-Supervised Techniques for Natural Language Processing Tasks.** In *Proceedings of the 6th International Conference on Computational Intelligence, Man-Machine Systems and Cybernetics* 2007:360–365.

[24] 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].