Approaches to Hungarian Named Entity Recognition

A magyar nyelvű tulajdonnév-felismerés módszerei

Tézisfüzet

Simon Eszter

Pszichológia Doktori Iskola – Kognitív Tudomány
Budapesti Műszaki és Gazdaságtudományi Egyetem

Budapest, 2013

Témavezető:
Kornai András

Bevezetés

A számítógépes nyelvészet egy interdiszciplináris tudományterület a számítástechnika és a nyelvészet határán. A kognitív tudományok közé tartozik, és átfedésben van a mesterségesintelligencia-kutatással, amelynek célja az emberi kogníció számítógépes modellálása. A számítógépes nyelvészet elméleti célja, hogy formális elméleteket és modelleket építsen a nyelvi tudásról, amely a nyelv létrehozásához és megértéséhez szükséges. Ugyanakkor van egy alkalmazott komponense is, amelyet más néven nyelvtechnológiának is hívnak, amely olyan számítógépes rendszerek létrehozására irányul, amelyek képesek feldolgozni vagy előállítani az emberi nyelv különböző formáit.

Az információkinyerés a számítógépes nyelvészet egyik fontos alterülete, amelynek célja, hogy a számítógép által olvasható, de strukturálatlan szövegből automatikusan információt nyerjünk ki. A feladatok igen tág köre tartozik alá: például megtalálni az összes cégnevet egy dokumentumban, vagy kideríteni egy szövegből, hogy milyen esemény történt, milyen szereplőkkel. A lényege, hogy hatalmas mennyiségű szöveg átnézése helyett csak a felhasználó számára releváns információt kapjuk meg.

A tulajdonnév-felismerés (Named Entity Recognition, NER) az információkinyerés egyik legtöbbet kutatott alfeladata, melynek során megnevezett entitások bizonyos előre kiválasztott típusait kell beazonosítani. Egy szöveg nyelvi elemzése általában azzal kezdődik, hogy a szöveg szavait főnévként, melléknévként, igeként stb. azonosítjuk szótárak segítségével. Viszont a legtöbb szöveg tartalmaz neveket, amelyeket nem tud értelmes nyelvi egységként azonosítani a rendszer. Így tehát a tulajdonnév-felismerés nélkülözhetetlen lépés a nyelvfeldolgozás további szintjei, így például az információ-visszakeresés vagy a gépi fordítás számára.

A tulajdonnevek definíciója

A NER két fő lépésből áll: először lokalizálni kell a szövegben a neveket, majd besorolni az előre definiált névosztályok valamelyikébe.

Az első és legfontosabb kérdés, hogy hogyan definiáljuk a felismerendő entitásokat. Ez a kérdés erősen összefügg a NER terén alkalmazott annotációs sémák által használt névosztályokkal. A NER feladatot 1995-ben vezették be a 6. Message Understanding Conference (MUC) [Grishman and Sundheim, 1996] keretein belül, és három alfeladatra osz-

tották: tulajdonnevek, idői és numerikus kifejezések felismerésére. A NER közösségen belül elfogadott tény, hogy ez utóbbiakat is a felismerendő nyelvi elemek közé sorolják, de a legtöbbet vizsgált típusok a személy-, hely- és intézménynevek. A negyedik névtípust (`Miscellaneous`) a Conference on Computational Natural Language Learning (CoNLL) keretein belül kezdték el alkalmazni 2002-ben [Tjong Kim Sang, 2002] és 2003-ban [Tjong Kim Sang and De Meulder, 2003]. Ez az osztály azokat a neveket tartalmazza, amelyek a felsorolt három klasszikus névtípuson kívül esnek. Az azóta eltelt időben a MUC és a CoNLL annotációs sémák és a velük együtt rendelkezésre bocsátott annotált korpuszok váltak a legfőbb szabványokká a NER területén.

Ezen szabványok annotációs útmutatói nem adnak egzakt definíciót az annotálandó entitásokról, hanem csak példákat és ellenpéldákat sorolnak fel. A következő leírás a MUC-7 NER feladatkiírásából származik [Chinchor, 1998]:

> „Ez a feladat a tulajdonnevekre, rövidítésekre és talán egyéb vegyes egyedi azonosítókra korlátozódik, amelyek a TYPE attribútumon belül a következő kategóriákba sorolhatók:
> ORGANIZATION: vállalatok, kormányzati vagy egyéb intézmények nevei
> PERSON: személy- vagy családnevek
> LOCATION: politikailag vagy fölrdajzilag meghatározott helyek (városok, megyék, országok, nemzetközi régiók, vizek, hegyek stb.) nevei"

Emellett a leírás mellett negatív példák (nem nevek) is fel vannak sorolva. Ahhoz, hogy egy szöveget tulajdonnévi annotációval lássunk el, ez a fajta meghatározás nem kielégítő. Ráadásul a fent említett annotációs útmutatók kizárólag angol példákat és ellenpéldákat tartalmaznak. Viszont más nyelvekben, mint például a magyarban, vannak olyan nyelvi elemek, amelyeket ezek szerint nevekként kellene annotálni, pedig nem nevek. A magyarra készített annotációs útmutató [11] kidolgozásakor ezekből a széles körben használt annotációs útmutatókból indultunk ki, így ezek gyenge pontjai gyorsan napvilágra kerültek. Ezekből a tapasztalatokból arra a következtetésre jutottunk, hogy a nevek annotálásához szigorúbb definícióra van szükség.

Ezért Kripke [Kripke, 2000] elméletét tanulmányoztuk, amely szerint a tulajdonnevek merev jelölők. Kripke szakít Frege [Frege, 2000] és Russell [Russell, 2000] leíráselméletével, amely szerint a tulajdonnevek egyenlőek lennének határozott leírásokkal. A dolgozat 2. fejezetében a tulajdonnevek

elméletének nyelvfilozófiai és nyelvészeti megközelítéseit ismertetjük. Az elméleti háttér bemutatása után az eredményeket a NER feladatra alkalmazzuk.

1. tézis. *A tulajdonnevekre vonatkozó nyelvfilozófiai és nyelvészeti elméletek vizsgálata után arra a következtetésre jutottunk, hogy ahhoz, hogy a nevekre egy használható definíciót tudjunk adni, a klasszikus arisztoteliánus megközelítés, amely szerint létezik egy differentia specifica, amely alapján valamit egy csoporthoz sorolhatunk, más valamit pedig kizárhatunk, nem kielégítő. A mi céljainkra a prototípus-elmélet jobban alkalmazhatónak tűnik, amely szerint a tulajdonnevek egy kontinuumot alkotnak a legprototipikusabb nevektől (személy- és földrajzi nevek) a nem tipikus nevekig (termék- és nyelvnevek). Végül a NER alkalmazás célja lesz az, amely lehatárolja a felismerendő entitások körét.*

A szerző hozzájárulása. A szerző részt vett a HunNer korpusz építési munkálataiban, amelynek célja egy kellően nagy méretű, tematikusan heterogén, konzisztens annotálási szabályzaton alapuló, manuálisan névannotált magyar nyelvű korpusz létrehozása volt. A szerző felelős az annotációs séma kidolgozásáért és az annotációs útmutató megírásáért. A korpusz a szerző hatáskörén kívül eső okok miatt nem készült el, de az útmutatót későbbi projektekben, például a Criminal NE korpusz[1] építéséhez használták. Az eredmények a következő publikációkban olvashatók: [10] és [11], az annotációs útmutató pedig elérhető az alábbi URL-en keresztül: `http://krusovice.mokk.bme.hu/~eszter/utmutato.pdf`.

Metonimikusan viselkedő tulajdonnevek

Akkor beszélünk metonímiáról, amikor egy kifejezést egy másik, vele kapcsolatban álló kifejezés helyett használunk bizonyos kontextusban [Lakoff and Johnson, 1980]. A köznevek mellett a tulajdonnevek is rendszeresen metonimikusan viselkednek, ahogy az látható az 1. és a2. példákban. (Az itt szereplő magyar nyelvű példák a szerző cikkéből [10] vagy az internetről származnak, nem önálló intuitív kitalálmányok. A példákban a releváns részek dőlt betűvel vannak szedve.)

(1) A férfi egy hajtásra megitta az egész *üveget*.

(2) A hölgy *Bachot* játszik.

[1] http://www.inf.u-szeged.hu/rgai/nlp?lang=en&page=corpus_ne

Szó szerinti értelemben a fenti két mondat egyike sem igaz. Az 1. példában a férfi nem magát az üveget itta meg, hanem a benne levő folyadékot. A 2. példában a hölgy pedig nem a személyt játszotta, akinek a neve Bach, hanem a zenét, amit Bach komponált.

Ez a fajta referenciaátvitel szisztematikus, vagyis bármilyen személynévvel működik, abban az esetben, ha a diskurzus résztvevői tisztában vannak vele, hogy az említett személy egy művész, és egy művet tudnak társítani hozzá. A nyelvészeti szakirodalom [Lakoff and Johnson, 1980, Fass, 1988] konvencionális metonímiáknak hívja az ilyen eseteket, amelyek szemantikai osztályok (itt: személy-, hely- és intézménynevek) között valósítanak meg konceptuális leképezést. Néhány példa a konvencionális metonímiákra (a metonímiatípusok nevét a nemzetközileg elfogadott jelölési módnak megfelelően kiskapitális betűkkel szedve közlöm a példamondat után zárójelben):

(3) A *Manchester* ma a *Münchennel* játszik. (A HELY AZ EMBEREK HELYETT)

(4) Federer idén is meghódította *Wimbledont*. (A HELY AZ ESEMÉNY HELYETT)

(5) Az *IBM* ma jelentette be új technológiáját. (A SZERVEZET A TAGOK HELYETT)

(6) Egy *Volvo* kormánya mögött érezhető igazán a kényelem és a dinamika. (A SZERVEZET A TERMÉK HELYETT)

A szabályos, sémákba rendezhető referenciaátvitel mellett újszerű, egyedi darabokat is létrehozhatunk: a 7. példában az 'egyeske' az egyes ágyon fekvő betegre utal. Markert és Nissim [Markert and Nissim, 2007a] az ilyen eseteket nem konvencionális metonímiáknak hívja.

(7) az *egyeske* az ajtó mellett fekszik

A tulajdonnevek metonimikus használata meglehetősen gyakori, ennek ellenére a jelenlegi névfelismerő rendszerek nem különböztetik meg a metonimikus használatot a szó szerintitől. Pedig a tulajdonnevek metonimikus használatának detektálása minden olyan nyelvfeldolgozó alkalmazás teljesítményét javítná, amely használ névfelismerést. A metonímiafeloldás fontosságát már több nyelvtechnológiai feladat esetében kimutatták, mint például a gépi fordításban [Kamei and Wakao, 1992], a kérdésmegválaszoló rendszerekben [Stallard, 1993], valamint az anaforafeloldásban [Harabagiu, 1998, Markert and Hahn, 2002].

A literális és a metonimikus használat közötti különbségtétel és a keresett referens azonosítása egyfajta klasszifikációs feladatként kezelhető. Markert és Nissim [Markert and Nissim, 2002] a metonímiafeloldási feladatot a szójelentés-egyértelműsítéshez hasonlítja, és azt állítja, hogy az ott megszokott módszerek használhatóak erre a feladatra is. Ebből a feltételezésből kiindulva rendezték meg a SemEval-2007 fórumon a metonímiafeloldási versenyt [Markert and Nissim, 2007b], amelyben az volt a feladat, hogy fel kellett ismerni és be kellett kategorizálni a hely- és intézménynevek literális, vegyes és metonimikus használatát. Építettünk egy maximum entrópia alapú rendszert [5], amely a legjobb eredményt érte el ezen a versenyen. A dolgozat 3. fejezete a metonímiatípusok leírását adja, majd ismertetjük a rendszert.

2. tézis. *Mivel a metonímiában részt vevő, egymással kapcsolatban álló referensek közötti konceptuális leképezés nincsen konkrét nyelvi formához kötve, a metonimikusan viselkedő tulajdonnevek felismerése nem triviális feladat. Bizonyos felszíni és szintaktikai jegyek, valamint szemantikai általánosítási módszerek használatával viszont javíthatjuk a metonímiafelismerő rendszerek teljesítményét. Bemutatunk egy felügyelt gépi tanuláson alapuló rendszert, amely a legjobb eredményt érte le a SemEval-2007 metonímiadetektáló versenyében. Az eredményeink azt mutatják, hogy a fő határvonal nem a konvencionális és nem konvencionális metonímiák, hanem a literális és metonimikus nyelvhasználat között húzódik.*

A szerző hozzájárulása. A metonímiafeloldó rendszer építése közös munka volt a társszerzőkkel: Farkas Richárddal, Szarvas Györggyel és Varga Dániellel. A szerző feladata volt az elméleti háttér feltárása és a szakirodalom tanulmányozása. Emellett a szerző munkája egyes szemantikai általánosítási módszerek kidolgozása, különösen a Levin igeosztályok beépítése és a trigger szavak gyűjtése. A szerző közreműködött a jegyek kitalálásában és kimérésében, valamint az eredmények kiértékelésében. A rendszerleírás megtalálható a SemEval-2007 konferenciakiadványában [5], míg az elméleti háttér egy külön cikkben jelent meg [10].

Gold és silver standard korpuszok a tulajdonnévfelismerésben

A felügyelt statisztikai módszerek alkalmazásához nagyméretű strukturált szöveggyűjteményekre, vagyis korpuszokra van szükség. A korpuszokat különböző kritériumok alapján csoportosíthatjuk: vannak általános

és témaspecifikus, egy- és többnyelvű, címkézett és címkézetlen korpuszok. Ahhoz, hogy egy korpuszt gold standardnek minősítsünk, többféle követelménynek is meg kell felelnie, például teljes mértékben le kell fednie egy nyelvváltozatot, vagy ha ez lehetetlen, akkor reprezentativitásra kell törekednie; elég nagynak kell lennie ahhoz, hogy felügyelt statisztikai rendszereket lehessen rajta tanítani és kiértékelni; valamint kézzel hozzáadott pontos nyelvi annotációt kell tartalmaznia.

A NER területén alkalmazott gold standard korpuszok általában témaspecifikusak (jellemzően csak sajtóhíreket tartalmaznak) és korlátozott méretűek. Kellően robusztus tulajdonnév-felismerő rendszerek építéséhez viszont nagyméretű, a téma tekintetében heterogén korpuszokra van szükség. A kézi annotálás rendkívül idő-, erőforrás- és szakértelemigényes feladat, ezért az elmúlt időkben különösen nagy hangsúly került az annotációs költségek csökkentésére.

Ez a cél többféleképpen is elérhető. Az egyik megközelítés, hogy felügyelet nélküli módszereket használunk, amelyekhez nem kellenek nagy méretű kézzel annotált korpuszok. A másik lehetőség, ha automatikusan állítunk elő erőforrásokat, de legalábbis olyan eszközöket használunk, amelyek elég jók ahhoz, hogy automatikus annotáláshoz lehessen használni őket. Még egy további lehetőség az olyan webes közösségi tartalmak felhasználása korpuszépítéshez, mint például a Wikipédia vagy a DBpedia. A dolgozatban egy olyan megközelítést mutatunk be, mely ezen lehetőségeket kombinálja: automatikus eszközökkel tulajdonnév-annotált korpuszokat építettünk Wikipédia-szócikkekből.

Egy automatikusan generált vagy silver standard korpusz a gold standard korpuszok alternatívájaként szolgál. Az ilyen korpuszok jól használhatók tulajdonnév-felismerő rendszerek teljesítményének növelésére több módon is.

A dolgozat 4. fejezetében egy általános korpusznyelvészeti bevezető után bemutatjuk a NER terén használt gold standard korpuszokat, majd ismertetjük az általunk kidolgozott új módszert.

3. tézis. *Bemutatunk egy új módszert, amellyel közelebb kerülünk a NER egyik fő céljához, a korpuszépítés annotációs költségeinek csökkentéséhez. Automatikus eszközökkel tulajdonnév-annotált magyar és angol korpuszokat építettünk a Wikipédiából. Magyar nyelvre ez az első silver standard névannotált korpusz. Angolra is csak egy szabadon elérhető silver standard korpusz létezik, a Semantically Annotated Snapshot of Wikipedia, de az ő módszerük nem alkalmazható kevés erőforrással rendlkező nyelvekre. Mivel a mi módszerünk szinte teljesen nyelvfüggetlen, minden Wikipédiával rendelkező nyelvre alkalmazható.*

4. tézis. *Megmutattuk, hogy az automatikusan előállított silver standard korpuszok jól használhatók tulajdonnév-felismerő rendszerek teljesítményének növelésére több módon is: (a) a kevés erőforrással rendelkező nyelvek esetében tanítókorpuszként tudnak szolgálni; (b) kiegészítő vagy önálló tanítókorpuszként használhatók a klasszikus sajtóhírektől eltérő műfajokra; (c) forrásai lehetnek nagyméretű névlistáknak; és (d) a rajtuk tanított rendszer kimenete jegyként felhasználható más névfelismerő rendszerekben.*

A szerző hozzájárulása. A szerző több korpuszépítési projektben is részt vett.

A Magyar Generatív Történeti Szintaxis projektben a szerző feladata egy olyan korpusz építése, amely tartalmazza az összes ómagyar szövegemléket, és válogatást nyújt a középmagyar korból. A korpusz elérhető egy online keresőfelületen keresztül: `http://rmk.nytud.hu/`. Kapcsolódó publikációk: [14, 13, 8].

Az ABSTRACT projekt vizsgálati témája az volt, hogy az ember hogyan sajátítja el és dolgozza fel az absztrakt fogalmakat. A projekten belül a szerző volt a felelős a korpuszépítési munkákért, amelynek során metaforikus kifejezéseket annotáltunk fel félig automatikus eszközökkel. Maga a korpusz és az eredmények a következő cikkekben vannak bemutatva: [2, 1, 4].

A HunNer korpusz építése során a szerző feladata volt az annotációs séma előkészítése és az annotációs útmutató megírása. A korpusz leírása a [11] cikkben olvasható.

A magyar és angol nyelvű silver standard korpuszok építése közös munka volt a társszerzővel, Nemeskey Dáviddal. A szerző feladata volt az elméleti háttér feltárása és a szakirodalom tanulmányozása. Továbbá a szerző részt vett a DBpedia ontológiai osztályok CoNLL névosztályokra való leképezésében, és a hibatípusok elemzésében és kiértékelésében. Az újonnan létrehozott adathalmazok kiértékelése a szerző munkája. A módszer és a korpuszok leírása elérhető a következő publikációkban: [12, 6].

A tulajdonnév-felismerés módszerei

A NER, mint minden más nyelvfeldolgozási feladat, kétféleképpen közelíthető meg: manuális munkával létrehozott szabályokkal, vagy statisztikai alapú gépi tanuló algoritmusokkal. Ez a kettősség jellemző az egész számítógépes nyelvészetre, és az 1950-es évekig vezethető vissza, amikor Chomsky publikálta nagy hatású írását Skin-

ner *Verbal Behavior* című munkájáról [Chomsky, 1959]. A véges állapotú és valószínűségi modellek, amelyek előtte elterjedtek voltak, vesztettek népszerűségükből ebben az időben, és a számítógépes nyelvészet két egymástól élesen elkülönülő paradigmára bomlott szét: az elméletorientált vagy szabályalapú és az adatvezérelt vagy sztochasztikus paradigmákra. A beszédtechnológiában sikeresen alkalmazott statisztikai módszerek az 1990-es években terjedtek el a számítógépes nyelvészet más ágaiban is. Ezt az időszakot az empirizmus visszatéréseként szokták emlegetni. A filozófiai hátterük alapján a két paradigmát racionalista és empirikus megközelítéseknek is nevezik. Az 5.1. alfejezetben a filozófiai háttérről adunk áttekintést, valamint bemutatjuk a két tábor történetét, egészen mostanáig, amikor a két megközelítés elkezdett közeledni egymás felé, és a kutatók olyan hibrid rendszerekkel kísérleteznek, amelyek a két metodológia előnyeit ötvözik.

Egy szabályalapú NER rendszerhez olyan mintákat kell definiálni, amelyek a nevek belső szerkezetét írják le, valamint olyan környezetfüggő szabályokat kell írni, amelyek a nevek klasszifikálásához adnak támpontot. Az 5.2. alfejezetben különféle mintákat mutatunk be, amelyek belső és külső bizonyítékokként szolgálhatnak a NER-hez, és bemutatunk egy szabályalapú rendszert, amellyel egy magyar enciklopédia szövegéből nyertünk ki neveket. Rámutatunk a szabályalapú rendszerek hátrányaira is, amelyek alapján azt a következtetést vonhatjuk le, hogy a statisztikai rendszerek jobban teljesítenek a NER feladatban.

A statisztikai gépi tanuló algoritmusok az alapján osztályozhatók, hogy milyen típusú bemenő adatot igényelnek. A felügyelet nélküli tanulók alkalmazásához nincs szükség nyelvi annotációval ellátott adatra, vagyis az a feladat, hogy a címkézetlen adatban találjunk rejtett összefüggéseket. A félig felügyelt rendszerek annotált és sima szövegeket is használnak a címkézési feladatokhoz. A felügyelt rendszerek pedig annotált korpuszokból kiindulva építenek modellt az adatokból megtanult szabályszerűségek alapján.

Egy felügyelt NER rendszer építéséhez először is szükség van egy nyelvi információval ellátott gold standard korpuszra. Az algoritmus ebből tanulja meg a paramétereket, és a rendszer kiértékelése is ez alapján történik. Ehhez a korpuszt egy tanító és egy kiértékelő adathalmazra kell osztani. A következő lépés a jegykinyerés, amelynek során olyan jegyeket definiálunk, amelyek fontosak lehetnek a feladat szempontjából, majd ezeket hozzárendeljük az egyes adatpontokhoz. Ezek a jegyek szolgálnak bemenetül a tanuló algoritmusnak, amely egy modellt épít az adatokban talált szabályszerűségek alapján. Végül a kiértékelő korpuszt felcímkézzük a modell alapján a legvalószínűbb címkékkel, és ezt a kime-

netet hasonlítjuk össze a gold standard címkékkel. Az 5.3.1. alfejezet a felügyelt NER rendszerek alkalmazásának teljes folyamatleírását adja.

A nagyobb nyelvekre sok olyan NER rendszert építettek már, amelyek valamilyen felügyelt tanulási módszert használnak. Ezeknek általában nincs sok nyelvfüggő komponense, ennek ellenére magyarra ez előtt csak egy ilyen rendszer épült [Szarvas et al., 2006]. Bemutatunk egy statisztikai NER rendszert, a `hunner` névfelismerőt, amely a legmagasabb F-mértéket érte el magyar nyelvre. Az 5.3.2. alfejezetben részletes rendszerleírást adunk.

5. tézis. *A tulajdonnév-felismerés, hasonlóan más nyelvfeldolgozási feladatokhoz, kézzel definiált szabályokkal és gépi tanuló algoritmusokkal is megoldható. Bemutatunk egy szabályalapú rendszert, amellyel magyar nyelvű enciklopédiaszövegekből nyertünk ki neveket, valamint egy felügyelt gépi tanuláson alapuló rendszert, amely magyarra a legjobb eredményt adja. Az eredményeink azt mutatják, hogy statisztikai algoritmusok használatával robusztusabb és jobban teljesítő rendszert lehet létrehozni.*

A szerző hozzájárulása. A szerző több olyan munkában is részt vett, amely racionalista vagy empirikus módszereket alkalmazott a nyelvelsajátítás kutatásában és különféle nyelvfeldolgozási feladatokban egyaránt.

A szerző részt vett az 'Analogikus általánosítási folyamatok a nyelvelsajátításban' című projektben, amelynek célja a nyelvelsajátítás mechanizmusainak modellálása volt, különös tekintettel arra, hogy a gyerekek hogyan tanulják meg a vonzatkereteket a rendelkezésükre álló nyelvi inputból. Különböző statisztikai modelleket alkalmaztunk a vonzatkeretek automatikus kinyerésére, és arra az eredményre jutottunk, hogy a gyakoriság és a nyelvi input mennyisége fontos paraméterek mind a pszicholingvisztikában, mind a gépi tanulásban. Az eredményeket a következő cikkek mutatják be: [9, 15, 3].

A Magyar Generatív Történeti Szintaxis projekt keretein belül kifejlesztettünk egy félig automatikus szövegnormalizáló rendszert ómagyar szövegekre. A történeti dokumentumok normalizálása jellemzően manuálisan összeállított szabályok alkalmazásával történik. Ezzel szemben mi egy olyan automatikus rendszert építettünk, amely a zajos csatorna modellen alapul. Ebben az esetben a kézi munka a szabályírás helyett a tanuló adatok előállítására tolódott, amely a szerző munkája. Az automatikus normalizálás előnye, hogy a kézi munka lecsökken arra, hogy a lehetséges megoldások listájából ki kell választani a megfelelőt, vagyis nagyban segíti az annotátor munkáját. Az alkalmazott mdószerek és az eredmények bemutatása a következő cikkekben olvasható: [8, 7].

A szabályalapú névfelismerő rendszert a Magyar Nagylexikon Kiadó Zrt. megbízásából fejlesztettük, így a titoktartási kötelezettség miatt nem publikáltunk eredményeket. A rendszer közös munka volt a munkatársakkal: Gyepesi Györggyel, Incze Lajossal, Czinkos Zsolttal és Kiss Árpáddal. A szerző részt vett a névtövelő kifejlesztésében, a névátírási szabályok megalkotásában, a névlisták összeállításában, és a nevek felismerésében belső és külső bizonyítékokként szolgáló reguláis kifejezések írásában.

Az eredeti `hunner` rendszer fejlesztése közös munka a társszerzővel, Varga Dániellel. A szerző részt vett a jegyek definiálásában és implementálásában, a névlisták összeállításában és a kiértékelésben. A rendszer reimplementációja nem a szerző munkája, de az azóta eltelt időben új jegyeket implementált és értékelt ki, valamint új névlistákat állított össze. Az eredeti rendszer a következő cikkekben lett publikálva: [19, 20].

Jegykinyerés

A jegyek a szövegben található adatpontok (jelen esetben tokenek) jellemző tulajdonságait írják le. A tokenalapú klasszifikációs feladatokban minden tokenhez jegyvektorokat rendelünk, ahol a vektorok egy vagy több jegyet is tartalmazhatnak. A NER területén általánosan használt jegyek két csoportra oszthatók az alapján, hogy milyen értéket vehetnek fel, így megkülönböztetünk sztringértékű és bináris jegyeket. Például ha egy token nagybetűvel kezdődik, akkor megkapja azt a jegyet, hogy `iscap=1`. A jegyvektorok alkalmazása egyfajta absztrakciót valósít meg a szöveg fölött. A gépi tanuló algoritmus feladata pedig az, hogy ebben a nagy mennyiségű információban találjon szabályszerűségeket, amelyek relevánsak a névfelismerés szempontjából.

A jegyek definiálása manuális munka, hasonlóan ahhoz, ahogy a szabályalapú rendszerekhez mintákat írunk. A különbség az, hogy a statisztikai metodológiában a nyelvész nem mond semmit az egyes jegyek erősségéről, hanem azt az algoritmus tanulja meg a korpusz alapján. Az emberi kogníció hajlamos csak a kiugró eseteket észrevenni, vagyis fontosnak ítél olyan tulajdonságokat, amelyekről a korpuszadatok alapján kiderül, hogy mégsem azok. Ezért minden jegy erősségét ki kell mérni igazi adatokon, mielőtt beépítenénk a rendszerbe.

Ebből a célból virtuális rendszereket építettünk új jegyek egyenkénti hozzáadásával, párhuzamosan magyarra és angolra. A mérésekhez a `hunner` rendszer reimplementált változatát használtuk. A dolgozat 6. fejezetében ismertetjük a tulajdonnév-felismerésben általában használt je-

gyeket és azok diszkriminatív erejét. A jegyeket az alapján csoportosítottuk, hogy milyen típusú információt szolgáltatnak: felszíni tulajdonságokat, számmintákat, morfológiai vagy szintaktikai információt, vagy névlistába való tartozást. Ez utóbbiak esetében azt is kimértük, hogy a névlisták méretének mekkora hatása van a statisztikai névfelismerő rendszerek teljesítményére.

6. tézis. *Bemutatunk egy módszert, amellyel megállapítjuk a tulajdonnév-felismerésben általában alkalmazott jegyek erősségét. Arra az eredményre jutottunk, hogy azok a sztringértékű jegyek a legerősebbek, amelyek a token felszíni szerkezetéről szolgáltatnak információt. Azok a jegyek, amelyek azt mutatják meg, hogy a token nagybetűvel kezdődik-e, illetve hogy a mondat elején található-e, meglepetésre nem javítanak a teljesítményen. Továbbá az olyan jegyek, amelyek valamilyen külső erőforrás (morfológiai elemző, sekély mondattani elemző) kimenetét használják, szintén nem feltétlenül szükségesek a nevek felismeréséhez.*

7. tézis. *Összehasonlítottuk egy maximum entrópia alapú tulajdonnév-felismerő rendszer teljesítményét különböző méretű névlisták használata mellett, és arra jutottunk, hogy a statisztikai névfelismerő rendszerek teljesítményére nincs számottevő hatása a névlisták méretének. Ha nagyméretű névlisták állnak rendelkezésre, nincs akadálya a használatuknak, de a hiányuk nem okoz nagy problémát a névfelismerő rendszerek fejlesztésében.*

A szerző hozzájárulása. A 6. fejezetben használt jegyek nagy részének definiálása, erősségük kimérése és az eredmények kiértékelése a szerző saját munkája. A mérésekhez használt magyar és angol adatok előfeldolgozása, valamint megfelelő nyelvi információval való ellátása szintén a szerző önálló munkája. (Egy kivétellel: a Szeged Treebank szintaktikai információinak leképezése a Szeged NER korpuszra Zséder Attila és Ács Judit munkája.) Az itt alkalmazott névlisták összegyűjtése és feldolgozása szintén a szerző munkája.

Az itt felhasznált korpuszok morfológiai elemzése a `morphdb` lexikai adatbázis és nyelvtan alkalmazásával történt, amelynek a fejlesztésében a szerző is részt vett. A morfológiai adatbázis a következő cikkekben lett publikálva: [17, 16, 18].

A szerző részt vett egy olyan rendszer építésében, amely metaforikus kifejezéseket azonosít be különféle szövegekben. A rendszer különböző méretű és eltérő megközelítésekkel létrehozott listákat alkalmaz. Ebben a projektben a szerző feladata volt a listák feldolgozása, a felismerő rendszer teljes munkafolyamatának és szoftverkörnyezetének kialakítása, valamint a korpuszok létrehozása, amelyeken a módszereket kiértékeltük. A

munkálatok egyik legfontosabb eredménye az, hogy a szisztematikusan, kézzel válogatott listák használata adja a legjobb eredményt a metaforikus kifejezések felismerésében, amely egybevág a névfelismerés terén tapasztaltakkal. Az eredményeket a következő cikkekben publikáltuk: [2, 1, 4].

A szerző több tulajdonnév-felismerő rendszerben használt jegyek definiálásában és kiértékelésében is közreműködött: metonimikusan viselkedő neveket detektáló rendszerben angol szövegekben (vö. 3. fejezet) és az eredeti `hunner` rendszer építésében (vö. 5. fejezet). Ezek az eredmények a következő cikkekben lettek publikálva: [5, 19, 20].

A tézisekhez rendelt publikációk

[1] Babarczy Anna, Bencze Ildikó, Fekete István, és Simon Eszter. A metaforikus nyelvhasználat egy korpuszalapú elemzése. In Tanács Attila és Vincze Veronika, szerk., *VII. Magyar Számítógépes Nyelvészeti Konferencia*, 145–156, Szeged, 2010.

[2] Babarczy Anna, Bencze Ildikó, Fekete István, és Simon Eszter. The Automatic Identification of Conceptual Metaphors in Hungarian Texts: A Corpus-based Analysis. In *Proceedings of the LREC 2010 Workshop on Methods for the automatic acquisition of Language Resources and their evaluation methods*, 31–36, Malta, 2010.

[3] Babarczy Anna, Serény András, és Simon Eszter. Magyar igei vonzatkeretek gépi tanulása. In Tanács Attila, Szauter Dávid, és Vincze Veronika, szerk., *VI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2009)*, 333–342, Szeged, 2009. SZTE.

[4] Babarczy Anna és Simon Eszter. A fogalmi metaforák és a szövegstatisztika szerepe a metaforák felismerésében. In Prószéky Gábor és Váradi Tamás, szerk., *Általános Nyelvészeti Tanulmányok XXIV. Nyelvtechnológiai kutatások*, 223–241. Akadémiai Kiadó, Budapest, 2012.

[5] Farkas Richárd, Simon Eszter, Szarvas György, és Varga Dániel. GYDER: maxent metonymy resolution. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, 161–164, Prague, June 2007. Association for Computational Linguistics.

[6] Nemeskey Dávid Márk és Simon Eszter. Automatikus korpuszépítés tulajdonnév-felismerés céljára. In Tanács Attila és Vincze Veronika, szerk., *IX. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2013)*, 106–117, Szeged, 2012.

[7] Oravecz Csaba, Sass Bálint, és Simon Eszter. Gépi tanulási módszerek ómagyar kori szövegek normalizálására. In Tanács Attila, Sza-

uter Dávid, és Vincze Veronika, szerk., *VI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2009)*, 317–324, Szeged, 2009. SZTE.

[8] Oravecz Csaba, Sass Bálint, és Simon Eszter. Semi-automatic Normalization of Old Hungarian Codices. In *Proceedings of the ECAI 2010 Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2010)*, 55–60, Lisbon, Portugal, 2010. Faculty of Science, University of Lisbon.

[9] Serény András, Simon Eszter, és Babarczy Anna. Automatic acquisition of Hungarian subcategorization frames. In *Proceedings of the 9th International Symposium of Hungarian Researchers on Computational Intelligence*, 2009.

[10] Simon Eszter. Nyelvészeti problémák a tulajdonnév-felismerés területén. In Sinkovics Balázs, szerk., *LingDok 7. Nyelvész-doktoranduszok dolgozatai*, 181–196. Szegedi Tudományegyetem Nyelvtudományi Doktori Iskola, Szeged, 2008.

[11] Simon Eszter, Farkas Richárd, Halácsy Péter, Sass Bálint, Szarvas György, és Varga Dániel. A HunNER korpusz. In Alexin Zoltán és Csendes Dóra, szerk., *IV. Magyar Számítógépes Nyelvészeti Konferencia*, Szeged, 2006.

[12] Simon Eszter és Nemeskey Dávid Márk. Automatically generated NE tagged corpora for English and Hungarian. In *Proceedings of the 4th Named Entity Workshop (NEWS) 2012*, 38–46, Jeju, Korea, July 2012. Association for Computational Linguistics.

[13] Simon Eszter és Sass Bálint. Nyelvtechnológia és kulturális örökség, avagy korpuszépítés ómagyar kódexekből. In Prószéky Gábor és Váradi Tamás, szerk., *Általános Nyelvészeti Tanulmányok XXIV. Nyelvtechnológiai kutatások*, 243–264. Akadémiai Kiadó, Budapest, 2012.

[14] Simon Eszter, Sass Bálint, és Mittelholcz Iván. Korpuszépítés ómagyar kódexekből. In Tanács Attila és Vincze Veronika, szerk., *VIII. Magyar Számítógépes Nyelvészeti Konferencia*, 81–89, Szeged, 2011. SZTE.

[15] Simon Eszter, Serény András, és Babarczy Anna. Automatic Acquisition of Hungarian Subcategorization Frames. In *Proceedings of the LREC 2010 Workshop on Methods for the automatic acquisition of Language Resources and their evaluation methods*, 7–11, Malta, 2010.

[16] Trón Viktor, Halácsy Péter, Rebrus Péter, Rung András, Simon Eszter, és Vajda Péter. morphdb.hu: magyar morfológiai nyelvtan és szótári adatbázis. In *III. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2005)*, 169–179, Szeged, December 2005.

[17] Trón Viktor, Halácsy Péter, Rebrus Péter, Rung András, Vajda Péter, és Simon Eszter. Morphdb.hu: Hungarian lexical database and morphological grammar. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, 1670–1673, 2006.

[18] Trón Viktor, Halácsy Péter, Rebrus Péter, Rung András, Vajda Péter, és Simon Eszter. Morphdb.hu: Hungarian lexical database and morphological grammar. In S. Nagy Katalin és Szakadát István, szerk., *Média és társadalom. Válogatás a Szociológia és Kommunikáció Tanszék Média Oktató és Kutató Központ munkatársainak legújabb munkáiból*, 283–290. Műegyetemi Kiadó, 2006.

[19] Varga Dániel és Simon Eszter. Magyar nyelvű tulajdonnév-felismerés maximum entrópia módszerrel. In Alexin Zoltán és Csendes Dóra, szerk., *IV. Magyar Számítógépes Nyelvészeti Konferencia*, 32–38, Szeged, 2006.

[20] Varga Dániel és Simon Eszter. Hungarian named entity recognition with a maximum entropy approach. *Acta Cybernetica*, 18:293–301, 2007.

Hivatkozások

[Chinchor, 1998] Chinchor, N. (1998). MUC-7 Named Entity Task Definition Version 3.5. In *Proceedings of the 7th Message Understanding Conference (MUC-7)*.

[Chomsky, 1959] Chomsky, N. (1959). A review of B. F. Skinner's Verbal Behavior. *Language*, 35(1):26–58.

[Fass, 1988] Fass, D. (1988). Metonymy and Metaphor: What's the Difference? In *Proceedings of the 12th Conference on Computational linguistics – Volume 1*, COLING '88, pages 177–181, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Frege, 2000] Frege, G. (2000). Ueber Sinn und Bedeutung (On Sense and Reference). In Stainton, R. J., editor, *Perspectives in the Philosophy of Language – A concise anthology*. Broadview Press.

[Grishman and Sundheim, 1996] Grishman, R. and Sundheim, B. (1996). Message Understanding Conference – 6: A Brief History. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING)*, pages 466–471, Kopenhagen.

[Harabagiu, 1998] Harabagiu, S. (1998). Deriving Metonymic Coercions from WordNet. In *Workshop on the Usage of WordNet in Natural Language Processing Systems, COLING ACL*, pages 142–148.

[Kamei and Wakao, 1992] Kamei, S. and Wakao, T. (1992). Metonymy: Reassessment, survey of acceptability and its treatment in machine translation systems. In *Proceedings of ACL*, pages 309–311.

[Kripke, 2000] Kripke, S. (2000). Naming and Necessity. In Stainton, R. J., editor, *Perspectives in the Philosophy of Language – A concise anthology*. Broadview Press.

[Lakoff and Johnson, 1980] Lakoff, G. and Johnson, M. (1980). *Metaphors We Live By*. Chicago University Press, London.

[Markert and Hahn, 2002] Markert, K. and Hahn, U. (2002). Understanding Metonymies in Discourse. *Artificial Intelligence*, 135(1/2):145–198.

[Markert and Nissim, 2002] Markert, K. and Nissim, M. (2002). Metonymy Resolution as a Classification Task. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 204–213, Philadelphia. Association for Computational Linguistics.

[Markert and Nissim, 2007a] Markert, K. and Nissim, M. (2007a). Metonymic Proper Names: A Corpus-based Account. In Stefanowitsch, A. and Gries, S. T., editors, *Corpus-Based Approaches to Metaphor and Metonymy*, pages 152–174. Mouton de Gruyter.

[Markert and Nissim, 2007b] Markert, K. and Nissim, M. (2007b). SemEval-2007 Task 08: Metonymy Resolution at SemEval-2007. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pages 36–41, Prague. Association for Computational Linguistics.

[Russell, 2000] Russell, B. (2000). Descriptions. In Stainton, R. J., editor, *Perspectives in the Philosophy of Language – A Concise Anthology*. Broadview Press.

[Stallard, 1993] Stallard, D. (1993). Two kinds of metonymy. In *Proceedings of ACL*, pages 87–94.

[Szarvas et al., 2006] Szarvas, Gy., Farkas, R., and Kocsor, A. (2006). A Multilingual Named Entity Recognition System Using Boosting and C4.5 Decision Tree Learning Algorithms. In *Proceedings of Discovery Science 2006*, pages 267–278. Springer Verlag.

[Tjong Kim Sang, 2002] Tjong Kim Sang, E. F. (2002). Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition. In Roth, D. and van den Bosch, A., editors, *Proceedings of CoNLL-2002*, pages 155–158. Taipei, Taiwan.

[Tjong Kim Sang and De Meulder, 2003] Tjong Kim Sang, E. F. and De Meulder, F. (2003). Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In Daelemans, W. and Osborne, M., editors, *Proceedings of CoNLL-2003*. Edmonton, Canada.