Az [origo] automatikus címkézési projekt tapasztalatai

Farkas Richárd

MTA-SZTE Mesterséges Intelligencia Kutatócsoport,
6720 Szeged, Tisza Lajos krt. 103. III. lépcsőház
rfarkas@inf.u-szeged.hu

Kivonat: A cikkben bemutatjuk az [origo] hírportál archívumának automatikus címkézésére irányuló projektet. Címkézés alatt azt az eljárását értjük, ami az egyes dokumentumokhoz egy olyan kifejezéshalmazt rendel, amely annak tartalmát jól reprezentálja. A cikkben bemutatásra kerülnek az újságarchívumok címkézésére vonatkozó irányelvek, az automatikus címkézési megoldásunk, az elért eredmények és tárgyalunk olyan nyitott számítógépes nyelvészeti problémákat, amelyek megoldása nagyban hozzájárulhat a címkézés sikerességéhez. Az [origo] archívumának automatikus címkézése manuális kiértékelés alapján a dokumentumok 77,5 százalékát megfelelőnek minősítette, ami meghaladta az eredeti célkitűzéseket.

1 Bevezetés

Az egyik legismertebb Web 2.0-ás technológia az úgynevezett *címkézés* (tagging), aminek keretében az internetes közösség tagjai címkéket rendelnek az elektronikus tartalmakhoz (blogbejegyzésekhez, képekhez, URL címekhez stb.) [1]. A címkék egy vagy néhány szavas természetes nyelvű kifejezések, amelyek célja általában az adott tartalom tömör leírása, jellemzése. Egy nagyméretű, címkézett adathalmazban a keresés, rendszerezés jóval hatékonyabbá válik mind a címkéző felhasználó, mind az egész közösség számára. Ezen felül az úgynevezett *címkefelhő* segítségével az egész adathalmaz mindenki számára azonnal értelmezhető tartalmi reprezentációja is megvalósítható. Napjainkban címke-hozzárendelések vagy címkefelhő(k) szinte minden közepes és nagy weboldalon megtalálhatók.

Az [origo] internetes hírportál 2009 márciusától vezette be cikkeinek manuális címkézését[1]. A portál ezt megelőzően már üzemeltette azt a szolgáltatást, amelyben a videókat[2] a felhasználók (látogatók) szabadon címkézhették. A videócímkézés legfőbb tapasztalatai azok voltak, hogy a címkék hasznosak ugyan, de mivel a felhasználók saját szemszögükből (szubjektív címkék, saját kategóriarendszer) rendelik hozzá a címkéket, azok gyakran nem alkalmasak az adott tartalom témájának azonosítására (hasonló következtetéseket von le [2] is). Ezen tapasztalatok alapján az [origo] híreinek címkézésére egy köztes megoldást vezetett be: a cikkeket a szerkesztők közössé-

[1] http://www.origo.hu/techbazis/internet/20090312-tagging-a-hirportalon-az-origo-bevezeti-a-cikkek-cimkezeset.html

[2] http://videa.hu

ge (körülbelül 50 fő) együttesen címkézi, a híreket annak szerzője látja el címkével. Azonban nincsen előre rögzített taxonómia, a címkézés teljesen szabad.

Egy hírportál számára több szempontból is igen hasznos a teljes híranyagának felcímkézése. Ezzel minden olyan témának, melyről gyakran írnak, önálló oldala lehet, tulajdonképpen automatikusan önálló rovatoldalak keletkeznek. Az archívum címkézése lehetőséget biztosít arra is, hogy az [origo]-ról amúgy eltűnt tartalmainkat újra elérhetővé tegyék és segít abban is, hogy a különböző oldalakon megjelenő, de tematikában megegyező tartalmak egy helyről legyenek elérhetőek, ezáltal növelve az egyes termékeink közötti keresztolvasottságot.

A címkék automatikus tartalmakhoz rendelése csak az utóbbi években vált intenzíven kutatott témává, mind a számítógépes nyelvészet, mind egyéb tudományágak (képfeldolgozás, zene/videó címkézés stb.) területén. Az [origo] manuális címkézésével egyidejűleg a Szegedi Tudományegyetem Mesterséges Intelligencia Kutatócsoportjával közösen elindult az archívum automatikus felcímkézését célzó projekt is. Az origo.hu oldalon 1998. december 1-jén jelent meg az első hír, és a kézi címkézés elindulásáig 380 ezer cikk látott napvilágot. Az archívumcímkézési projekt célja az volt, hogy ezen hírekhez automatikusan, azok tartalmát jól reprezentáló címkéket rendeljünk, oly módon, hogy az egész dokumentumhalmazhoz tartozó címkekészlet koherens legyen. A következőkben bemutatásra kerülnek a címkézés irányelvei, a megoldási mód váza, illetve tárgyalunk olyan nyitott számítógépes nyelvészeti problémákat, amelyek megoldása nagyban hozzájárulhat a címkézési probléma megoldásához.

2 Címkézési útmutató

A projekt kezdetén az [origo] munkatársai elkészítettek egy címkézési útmutatót, ami elsősorban a szerkesztők (manuális címkézés) számára tartalmazott irányelveket, de ezeket az automatikus címkézésnél is követtük.

Címkeadásnál a közvetlen cél nem az, hogy a cikk témáját hogyan tudjuk kulcsszavak segítségével absztrahálni, hanem az, hogy megtaláljuk azokat a címkegyűjtőoldalakat, amelyek alatt szeretnék a felhasználók a cikket viszontlátni. Címkét négy kategóriában lehet megadni (*téma*, *személy*, *intézmény*, *földrajz*), ebből egyedül a téma kategóriánál kötelező megadni legalább egy címkét. Személynévnek kell tekinteni az állatneveket és fiktív élőlények neveit is. Ha több közszereplőre is vonatkozhat egy személynév, mindig meg kell toldani egy olyan kifejezéssel, mely egyértelműen csak rá vonatkozik, például: „*Csányi Sándor színész*”. A földrajzi nevek megkülönböztetése igen fontos, ezáltal lehetővé válik a hírek mellé megjeleníteni azok geográfiai pozícióját és az olvasónak lehetősége nyílik a helyi közösségéhez kapcsolódó hírek közt keresni. Földrajzi nevek közé tartoznak az egész univerzum egységei (bolygók stb.) is.

Az entitásokkal kapcsolatosan általánosságban követendő, hogy csak akkor vehetjük fel őket címkeként, ha azok nemcsak eseti alapon kerülnek bele a nemzetközi/hazai hírfolyamba, hanem viszonylag rendszeresen kerülnek szóba, és meghatározóak a cikkben elmondottakkal kapcsolatban.

A téma kategóriába kerülhetnek elvont fogalmak, jelenségek (pl. „*koalíciós válság*”), sportágak, ligák, események (pl. „*Sziget-fesztivál*”), tudományos fogalmak, szakkifejezések és egyéb entitások (pl. márkanevek) amelyek leírják a cikk főbb témáját. A téma kategória alá felvett fogalmaknál csak olyan címkék adhatóak meg, melyek önmagukban jól definiálnak egy korszakot, helyzetet, viszonyrendszert. A címkéknek jól kell definiálniuk egy területet, érdeklődési vagy fogalomkört, azaz vannak olyan emberek, akik kifejezetten csak arról a témáról akarnak majd olvasni.

A címkék megfelelő számát a cikk témája határozza meg. Koncentrált témájú cikkeknél általában 2-4 címke elegendő, általános elemzések, átfogó cikkek esetében több, maximum 8-10 címkét is adhatunk. A címkék csak főnévi szerkezetek lehetnek, és kerülendő a szleng, a zsargon, az átvitt értelmű szavak, a metafora, a humor és a parafrázis.

Ezek az irányelvek számos ponton eltérnek a szokásosnak tekinthető címkézési követelményektől. Ilyen például a cikk fontos szereplőinek (személyek, szervezetek, földrajzi helyek) kiemelt szerepe, legfeljebb három szó hosszúságú címkék, rokon értelmű címkék konzekvens használata stb.

3 Kapcsolódó munkák

A hírarchívum-címkézési probléma több szempontból is újszerű. Egyrészt nem illeszthető a létező automatikus címkézési megközelítések közé, mert azok vagy egyetlen dokumentum kulcskifejezéseinek megtalálására törekednek (keyphrase extraction) [3, 4], vagy hasonló dokumentumok címkéit emelik át (tag recommendation) [5, 6]. Előbbi megközelítés a dokumentumból kiemeli a potenciális kulcsszavakat, majd azok közül kiválaszt néhányat úgy, hogy azok a dokumentum tartalmát lefedjék, de ne tartalmazzon redundáns elemet. Ez a megközelítés nem alkalmazható a mi esetünkben, mert csak egyetlen dokumentumra fókuszál, az archívum címkézésénél pedig kiemelt szempont az egész dokumentumhalmazon vett konzisztens címkézés.

A másik megközelítésben rendelkezésre áll egy kielégítő méretű címkézett dokumentumhalmaz, és a fókusz egy címkézetlen dokumentumhoz hasonló dokumentumok megtalálására irányul, ui. a hipotézis az, hogy a hasonló témájú tartalmakról a címkék egy az egyben átemelhetőek. Az ilyen rendszereket általában címkeajánlásra használják blogbejegyzésekhez, ahol rendelkezésre áll nagyszámú címkézett dokumentum, a blog szájt összes korábbi bejegyzése [5]. Ez a megközelítés sem alkalmazható közvetlenül a mi esetünkben, habár hozzáfértünk a szerkesztők által 2009 márciusa és májusa közt címkézett hírekhez, azok nem tartalmazhatnak minden hozzárendelendő címkét (a 2009-es hírek nem kapnak például „*Tocsik-ügy*” címkét).

Az archívumcímkézés folyamán a két bemutatott módszert ötvözve kell alkalmazni, ahol alapvetően a címkéket a szövegekből kell származtatni (ezáltal biztosítani az új témák, entitások felismerését), de tekintettel kell lenni az egész dokumentumhalmaz koherens címkézésére is (témájukban megegyező hírek kapjanak közös címkét).

Amellett, hogy – legjobb tudomásunk szerint – ez az első munka, ami egy hírportál automatikus címkézését célozta meg (annak specialitásaival), ez az első megoldás magyar nyelvű automatikus címkézésre is.

4 Automatikus címkézés

Az automatikus címkézés során azt a soros feldolgozást választottuk, hogy első lépésben kiemeljük a szövegben egzaktul előforduló potenciális címkéket, majd ezeket megpróbáljuk absztrahálni, ami újabb címkék felvételét eredményezi. Végül a címke jelöltek halmazát leszűkítjük egy megfelelő méretűre, és ezt tekintjük végleges címkézésnek.

4.1 Szövegbeli címkejelöltek gyűjtése

A címkézési útmutató alapján csak főnévi csoportok szerepelhetnek címkeként. Három különböző módon gyűjtöttünk főnévi csoportokat: automatikus tulajdonnév-felismeréssel, szófaji kódok alapján derivációval és szótárillesztéssel.

A tulajdonnevek automatikus felismerése és szemantikai kategorizálása (személynév, földrajzi név, szervezetnév, egyéb) felügyelt tanulási keretben tulajdonképpen megoldottnak tekintett (habár bizonyos esetekben a módszerek pontossága a 70%-ot sem éri el [7]). Azonban ha nem áll rendelkezésre megfelelő méretű, karakterisztikájában a jelölendő szöveggel megegyező tanító adatbázis, a pontosság drasztikusan csökken. Az [origo] hírei témájukat, karakterisztikájukat tekintve igen diverzek, ezért manuálisan annotálásra került az *autós*, *itthon*, *nagyvilág*, *sport*, *szórakozás* és *techbázis* kategóriák körülbelül 200-200 híre. Ezeken az adatbázisokon Conditional Random Fields-et[3] tanítottunk a korábban gazdasági hírekre kidolgozott jellemzőkészlet [8] felhasználásával. A felsorolt főkategóriákon kívüli hírek esetében az egész annotált dokumentumhalmazon tanított modell predikcióját használtuk fel.

A nagyméretű dokumentumhalmaz lehetőséget biztosított az automatikus tulajdonnév-felismerés hibáinak javítására (utófeldolgozására) és normalizálására. Hibajavítás alatt az automatikusan jelölt tulajdonnévi frázisok határainak korrekcióját értjük (azaz összeragadt tulajdonnevek szétbontását, hozzáragadt tokenek eltávolítását, illetve a határok kiterjesztését). A normalizáció elsődleges célja a tulajdonnevek szótövesítése volt – ami nem oldható meg a standard morfológiai elemzők segítségével, hiszen itt a lehetséges szótövek felsorolása nem lehetséges – (például a „*Pannon”* szótöve „*Pann”*?). Emellett egyszerű szabályokkal kísérletet tettünk a rövidítések feloldására és az egyes kifejezések egységes szemantikai kategorizálására (leggyakoribb szerep) is. Ez utóbbi csak egyes főkategóriákon belül értelmes, hiszen például a *Kecskemét* a sporthíreken belül általában szervezetként szerepel (mint egy csapat), míg a belföldi hírek esetében földrajzi entitásként. Ezen utófeldolgozási lépéseket a korpusz automatikus jelöléséből nyert statisztikák alapján végeztük el a [9]-ben bemutatott eljáráshoz hasonlóan. Itt a fő hipotézisünk az volt, hogy egy tulajdonnév ragozatlan alakjának gyakorisága szignifikánsan nagyobb, mint bármely ragozott alakjáé.

A cikkek témájának felismeréséhez főnévi csoportokat (NP) is gyűjtöttünk a szövegből. Ehhez kísérleteztünk a hunpars-szal [10], de azt találtuk, hogy egy POS-tagger eredményeit felhasználva egyszerűbben és kevésbé zajosan tudunk NP-ket kiemelni. A megoldás során NP-nek tekintettük az egyszavas főneveket, melléknév-

[3] implementácó: http://mallet.cs.umass.edu/

főnév párokat, főnévi birtokos szerkezeteket és az igéből és melléknévből képzett főneveket. Az igékből és melléknevekből történő főnévképzésre, valamint a főnévi birtokos szerkezetek összetett szavakká alakítására (például „*üzemanyagok ára*"-ból „*üzemanyagár*") egyszerű átírási szabályokat alkalmaztunk.

A tulajdonnevek és NP-k azonosítása távol van a tökéletestől, ezért külső tudásbázisok is beépítésre kerültek a rendszerbe. Külső tudásbázisnak használtuk a Wikipédia[4] szócikkeinek címeit és annak gyűjtőoldalait (amelyek címe „listája"-ra végződik). Az így nyert listákat illesztettük a szövegre a ragozási és tőhangváltási lehetőségek figyelembevételével.

A tulajdonnév-kinyerés, főnévicsoport-azonosítás és listaillesztés eredményeit a következőképpen aggregáltuk: az azonos helyről érkező – pontosabban átfedő – találatokat (például egy azonosított tulajdonnév a szótárban is szerepelhet) elhagytuk, hiszen az, hogy két módszer is azonosította, nem implikálja, hogy kétszeres súlyt kapjon. Végül a halmazt egy paraméterezett tfidf metrika felhasználásával sorba rendeztük. A metrika figyelembe vette azt is, hogy a vizsgált találat a dokumentum melyik zónájából érkezett (cím, összefoglaló, képaláírás stb.), illetve vonatkozik-e rá formázási információ (például dőlt, kiemelt). A tfidf optimális paraméterezését a rendelkezésünkre álló kézzel jelölt cikkek alapján határoztuk meg.

4.2 Absztrakt címkézés

A szövegben egzaktul előforduló címkejelölteken felül általában szükség van ún. absztrakt címkék felvételére is, amik a dokumentum tartalmát általánosabb módon írják le (például a „*kékfény*" szó általában nem szerepel a cikkekben). Az ilyen jellegű absztrakciók elvégzésére két módszert dolgoztunk ki. Az első módszer a Wikipédia linkstruktúrájának kiaknázásával, a potenciáliscímke-halmaz alapján gyűjt össze absztrakt címkéket (ez a megközelítés általánosságban kerül bemutatásra a [11] publikációban).

A másik módszerben felügyelt tanulási problémaként fogalmaztuk meg egyes címkék felvételének lehetőségét. Ehhez statisztikai jellemzők és szemrevételezés útján kiválasztottunk 243 darab absztrakt témát jelölő címkét és 243 osztályozási modellt építettünk, ami a dokumentumhoz rendelt potenciális címkék alapján (azokat használva jellemzőkészletként) hivatott eldönteni, hogy a szóban forgó címkével kelle bővítenünk a címkejelöltek halmazát. Első pillantásra ezeknek a nagyon absztrakt témákat jelölő címkéknek egyszerűen következniük kellene az adott dokumentum kategóriájából (pl. „*foci*" kategória). Azonban annak ellenére, hogy az [origo] kategóriahierarchiája több mint ezer elemet tartalmaz, ezek nem egyenszilárdságúak (vannak köztük, amelyek több tízezer hírt tartalmaznak) és ráadásul a hierarchia időben evolválódott. Például a *kosárlabda* kategória 2001-ben került bevezetésre, az 1998 és 2001 közt a *kosárlabdával* foglalkozó hírek a *csapatsport* kategóriába kerültek. Ezért úgy döntöttünk, hogy az általunk fontosnak ítélt magas szintű absztrakciót képviselő címkéket gépi tanulási módon keressük meg.

A tanításhoz pozitív példaként egy magasabb szintű kategórián belül azokat a dokumentumokat használtuk, amelyeknél a kérdéses címke szerepelt a potenciáliscím-

[4] http://hu.wikipedia.org

ke-halmazban. Negatív példaként az ugyanezen időszakból származó kategórián kívüli dokumentumok szolgáltak. A kategóriabeli megkötésre például azért volt szükség, mert a „*Manchester*” és „*Liverpool*” potenciális címkék csak a sporthíreken belül implikálhatják a „*Premier League*” absztrakt címkét.

4.3 A címkehalmaz szűrése

A szövegből kiemelt tulajdonnevek, főnévi csoportok, szótárillesztések és az absztrakt címkék után előálló potenciális címkék halmazának átlagos mérete túl magas (17,3 az elvárt 4-5-tel szemben), ezért a legfontosabbak kiválogatását meg kellett oldanunk. Ehhez figyelembe vettük a 4.1 fejezetben röviden bemutatott címkerangsort – vegyük észre, hogy az absztrakt címkékre nem értelmezett a tfidf alapú rangsoroló metrika –, a címke forrását (pl. listaillesztés vagy Wikipédia-alapú absztrakt), a cikk fő kategóriájára vonatkozó specialitásokat és az útmutató egyéb megkötéseit (például legalább egy *téma* címkének mindig szerepelnie kell, és csak olyan címkék használhatóak, amelyek legalább három dokumentumhoz hozzá lettek rendelve).

Ezen jellemzők alapján manuálisan konstruáltunk döntési szabályokat arra vonatkozólag, hogy mely címkék szerepeljenek a dokumentum végső címkehalmazában. Ezek a szabályok csak a felsorolt szintaktikai jellemzőkre épültek. A legfontosabb jövőbeli kutatási iránynak a szemantikai információk felhasználását tekintjük ebben a szűrésben. Ehhez a címkejelöltek közt páronként tervezzük a szemantikai kapcsolat numerikus értékkel történő jellemzését (például a Wikipédia-alapú heurisztikák felhasználásával [11]) majd az így kialakuló súlyozott teljes gráf elemzésével (például hubok vagy communityk azonosítása) kialakítható egy reprezentatív, de koherens szűrt címkehalmaz.

5 Kiértékelés

Az archívum végső címkézésben 59.364 különböző címke került felhasználásra, ami összesen 1.885.427 címke-cikk összerendelést eredményezett (átlagosan 4,98 címke hírenként). A címkék átlagos hossza 1,45 token.

Egy címkézés kiértékelése igen nehéz (és főképp szubjektív feladat), mert meg kell ítélni a kiválasztott címkék megfelelő számát, azok relevanciáját és koherenciáját. Ez nem végezhető el automatikus módon (ahhoz az egyes fogalmak közt ismernünk kellene a pontos szemantikai kapcsolatot, aminek birtokában tulajdonképpen az egész címkézési probléma sem lenne nyitott) csak manuális szemrevételezéssel. A projekt végén az [origo] munkatársai 1000 véletlenszerűen választott cikk automatikus címkézését manuálisan ellenőrizték. A véletlen választás biztosította, hogy a kiértékelő halmaz mind időben, mind cikk-kategóriában kövesse azok valós eloszlását.

A végső kiértékelési metrika dokumentumszintű volt, azaz minden dokumentumról született egy bináris – jó/rossz – döntés. A cikkhez automatikusan rendelt címkehalmazt manuálisan öt különböző szempont szerint értékelték:

- helyesen kiválasztott, valid címkék száma (súly +1),
- olyan címkék száma, amelyek nem kapcsolódnak szorosan a cikk témájához (súly -1),
- olyan címkék száma, amelyek ugyan kapcsolódnak a témához, de valamilyen egyéb szempontból érvénytelenek, például túl absztrakt, túl szűk fogalmak, elírások, összeragadt entitások (súly -0,2),
- a szerkesztő által hiányzónak ítélt címkék száma (súly -0,7)
- helytelen típusba sorolások száma (pl. személynév helyett földrajzi kategória) (súly -0,5).

Egy dokumentumot akkor tekintünk jónak, ha a fenti pontszámok súlyozott összege pozitív. Az egyes típusok súlyai a kiértékelés előtt rögzítésre kerültek és az Origo Zrt. elvárásainak figyelembevételével lettek kialakítva.

A kiértékelés alapján a dokumentumok 77,5 százalékának címkézése megfelelő minőségű lett, ami az eredeti célkitűzéseket meghaladja.

6 Konklúzió, nyitott kérdések

A cikkben bemutattuk, hogy egy újság archívumának automatikus címkézése kielégítő eredményt képes elérni. Címkézési módszerünk számos számítógépes nyelvészeti és statisztikai megoldást használt fel. A problémát több részproblémára bontottuk fel. Ezen részmodulok közül néhány már eléri a már jónak tekinthető szintet (pl. tulajdonnevek azonosítása, dokumentumzónák súlyozása), azonban van számos, amire idő és magyar nyelvtechnológiai erőforrások hiányában csak egy alapmegoldást adtunk. Ezeken a jövőben tovább dolgozunk.

Végezetül felsoroljuk azokat a szükséges számítógépes nyelvészeti módszereket, amelyek megléte a címkézés szempontjából nagy jelentőséggel bírna:

- A tulajdonnevek (ill. minden szótárban fel nem sorolt frázis) szótövesítése a morfológiai elemzési (guessing) megközelítések [12] és a korpuszstatisztikai módszerek [9] kombinációjaként kellene, hogy működjön.
- A főnevek képzése melléknevekből, igékből igen fontos lépés. A jelenlegi egyszerű átalakítási szabályok helyett szükség lenne egy morfológiailag megalapozott derivációra. A címkézés keretében elégséges lenne azt megvizsgálni, hogy egy adott kiindulási szóból lehetséges-e képezni egy szótár valamely elemét (azaz feltehetjük, hogy ismerjük a lehetséges címkék halmazát). Megjegyezzük, hogy a morphdb.hu természetesen már tartalmazza ezeket az átalakítási szabályokat, valószínűleg azok kiegészítése és invertálása lenne a célravezető.
- Jelenleg a szövegből kiemelt potenciális címkék szövegkörnyezetét nem vizsgáljuk. Ha a címke rangsorolásnál figyelembe vennénk például a címkék és az igék közötti viszonyt (vagy csak magát a vonatkozó igét) egy jóval szofisztikáltabb módszert kapnánk. Az igei vonzatkeretek és egyéb függőségi viszonyok automatikus azonosításában nagy előrelépést eredményezhet a Szeged TreeBank függőségi nyelvtan változatának elkészülése [13].

- A szemantikai kapcsolatok felderítése területén a legfrissebb kutatások a részben strukturált, hatalmas méretű nyers korpuszok (elsősorban Wikipédia) kiaknázására építenek. A magyar nyelvtechnológia szempontjából igen kedvezőtlen, hogy a magyar Wikipédia mérete mindössze 4%-a az angolénak, így az onnan kinyerhető információ is kevesebb. Véleményünk szerint itt nem lesz elégséges az angolra bevált módszerek alkalmazása, hanem újszerű megközelítésekre lesz szükség, amelyek képesek szemantikai kapcsolatokat kinyerni ilyen jellegű erőforrásokból.

Köszönetnyilvánítás

Szeretnék köszönetet mondani az Origo Zrt. munkatársainak (Krich Balázs, Kárpáti András, Cserti Gergely) – akik nélkül ez a valós életbeli kutatási projekt el sem indulhatott volna – a konstruktív és inspiráló eszmecserékért, valamint a projektben részvevő kollégáknak (Almási Attila, Berend Gábor, Hegedűs István, Vincze Veronika) áldozatos munkájukért.

A kutatást – részben – a TEXTREND projekt (Jedlik Ányos program) keretében az NKTH támogatta.

Hivatkozások

1. Golder, S. A., Huberman, B. A.: Usage patterns of collaborative tagging systems. Journal of Information Science, Vol. 32, No. 2 (2006) 198-208
2. Kipp, M. E.I.: Tagging for Time, Task and Emotion. In: Proceedings of the 8th Information Architecture Summit, Las Vegas (2007)
3. Liu, Z., Li, P., Zheng, Y., Sun, M.: Clustering to Find Exemplar Terms for Keyphrase Extraction. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (2009) 257-266
4. Mihalcea R., Tarau, P.: Textrank: Bringing order into texts. In: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (2004)
5. Sood, S. C., Owsley, S. H., Hammond, K. J., Birnbaum, L.: TagAssist: Automatic tag suggestion for blog posts. In: Proceedings of the International Conference on Weblogs and Social Media (2007)
6. Tatu, M., Srikanth, M., D'Silva, T.: RSDC'08: Tag Recommendations using Bookmark Content. In: Proceedings of the ECML PKDD Discovery Challenge (2008)
7. Hasan, K. S., Rahman, A., Ng, V.: Learning-Based Named Entity Recognition for Morphologically-Rich, Resource-Scarce Languages. In: Proceedings of the 12th Conference of the European Chapter of the ACL (2009) 354-362
8. Szarvas, Gy., Farkas, R., Kocsor, A.: A multilingual named entity recognition system using boosting and c4.5 decision tree learning algorithms. In: DS2006, LNAI 4265 (2006) 267-278
9. Farkas, R., Vincze, V., Nagy, I., Ormándi, R., Szarvas, Gy., Almási, A.: Web-based lemmatisation of Named Entities In: TSD2008 LNCS Volume 5246 (2008) 53-60
10. Babarczy, A., Gabor, B., Hamp, G., Rung, A.: Hunpars: a rule-based sentence parser for Hungarian. In: Proceedings of the 6th International Symposium on Computational Intelligence (2005)

11. Berend G., Farkas R.: A Wikipédia felhasználása az absztrakt címkézési feladatban. In: Tanács A., Szauter D., Vincze V. (szerk.): VI. Magyar Számítógépes Nyelvészeti Konferencia (2009) 93-103
12. Trón, V., Németh, L., Halácsy, P., Kornai, A., Gyepesi, Gy., Varga, D.: Hunmorph: open source word analysis. In: Proceeding of ACL (2005)
13. Vincze V., Szauter D., Almási A., Móra Gy., Alexin Z., Csirik J.: A Szeged Treebank függőségi fa formátumban. In: Tanács A., Szauter D., Vincze V. (szerk.): VI. Magyar Számítógépes Nyelvészeti Konferencia (2009) 127-138