Balázs Ágnes – Sebők Miklós

II.2. NÉVELEM-FELISMERÉS

A fejezet bemutatja a *névelem-felismerést*, mint az egyik legfontosabb szövegbányászati feladatot. Segítségével kinyerhetők egy adott korpuszon belül előforduló névelemek, s ezen belül a tulajdonnevek (személynevek, helyek, szervezetek és egyéb tulajdonnevek). A fejezetben meghatározzuk a névelem-felismeréshez kapcsolódó legfontosabb fogalmakat, valamint a módszer típusfeladatait és buktatóit. Ezt követően három konkrét példán mutatjuk be a módszer gyakorlati használatát, majd a hazai és nemzetközi alkalmazásokból ismertetünk néhányat.

A *névelem-felismerés* (NER) az egyik legfontosabb speciális szövegbányászati feladat. Legegyszerűbb formájában az *információ-visszakeresés* területéhez tartozik, komplexebb megoldásai ugyanakkor már a *szövegbányászathoz* tartoznak (ld. II.1. fejezet). *Információ-visszakeresés* esetén célunk az, hogy a már strukturált korpuszból visszakeressük a számunkra releváns információt (Russel – Norvig, 2005: 742; Vázsonyi – Tikk, 2007: 63). A szövegbányászat esetén már lehetőség van a kifejezések közötti kapcsolatok elemzésére, tendenciák és minták felismerésére, és az információk összekapcsolása révén új információk létrehozására (Hearst, 1999: 3–4; Szarvas – Farkas, 2007: 81).

A névelem-felismerés módszere az 1990-es években született meg, lényege, hogy egy program felismeri a korpuszban felbukkanó tulajdonneveket, azokat kigyűjti, és típusonként (pl. földrajzi név, márkanév, jogi személy stb.) csoportosítja (ennyiben egy hibrid információkinyerési és kategorizálási feladatként is tekinthetünk rá). A standardizált megoldások lehetővé teszik akár telefonszámok vagy időpontok kigyűjtését is, ennyiben tehát túlmutatnak a tulajdonnevek körén. A NER legáltalánosabb annotációs sémáit a Dokumentummegértési Konferenciák (MUC) és a Természetesnyelv-tanulási Konferenciák (CoNNL) fejlesztették ki (Jiang, 2012: 15–16). A fejezetben bemutatjuk e kutatási irány alapfogalmait és alkalmazásait.

A névelem-felismerés

A névelem-felismerés az információkinyerés egyik legalapvetőbb formája, amely igen fontos a szövegek összetettebb módszerek alkalmazására való előkészítésében (i. m.: 15–16). E módszer kapcsán fontos a *névelem* és a *tulajdonnév* fogalmának megkülönböztetése. Névelem minden olyan szóalaksorozat (*token*sorozat),[1] amely a világ valamely egyedi létezőjére utal. A tulajdonnév tehát csupán a névelem fogalmának egy részhalmaza, hiszen a tulajdonnév mellett névelemnek tekinthető például valamely azonosító, telefonszám, pénznem vagy e-mail cím is (i. m.: 15–16; Szarvas – Farkas, 2007: 91). Ezzel együtt a névelem-felismerés leggyakrabban a személynevek, helyek, szervezetek és egyéb tulajdonnevek felismerésére irányul, így a továbbiakban a két fogalmat szinonimaként használjuk.

A fogalmi alapokhoz tisztázni kell a strukturált és a strukturálatlan szöveg fogalmát is. A strukturálatlan szöveg alatt a hétköznapi értelemben vett szövegeket értjük. A strukturált szövegek olyan szövegek, amelyek valamilyen szempont szerint már feldolgozásra kerültek. A strukturált szöveg rendelkezik *metaadatokkal,* azaz az adatállományhoz kapcsolódó adatokkal, például kódokkal, *címkékkel*. Míg az *adatbányászat*i és az információ-visszakeresési módszereket összefüggések strukturált adatokból való kinyerésére használjuk, a *szövegbányászat* eszközeit strukturálatlan szöveges halmazok feldolgozásánál alkalmazzuk (Tikk, 2006: 344–346; Tikk, 2007a: 20–21). Ez utóbbi viszont külön *szövegelőkészítést* igényel, melynek módszerei közé tartozik a *tiltólistás szavak* alkalmazása, a *lemmatizálás* és a *toldaléklevágás* (Hu – Liu, 2012: 388–389 – minderről bővebben ld. II.1. fejezet).

A következő lépések már a természetesnyelv-feldolgozás területére vezetnek (amely szinte elválaszthatatlan az információkinyerés fogalmától). Mint az előző fejezetben láttuk, ahhoz, hogy a természetes nyelvből adatokat kapjunk, fel kell dolgoznunk a vizsgálandó szöveget, strukturált adatállományokat kell létrehoznunk (Aggarwal – Zhai, 2012: 3, Markov – Larose, 2007: 13). A webes információkinyerő rendszerek *(wrapperek)* alkalmazására azért van szükség, mert a honlapok gyakran tartalmaznak olyan strukturált vagy félig strukturált szövegeket, mint a táblázatok vagy a felsorolások, amelyek sokkal inkább állnak HTML vagy egyéb kódokat tartalmazó elemekből, mint természetes nyelvi elemekből (Jiang, 2012: 14–15).

[1] „*Token*nek nevezzük egy karaktersorozat konkrét dokumentumbeli előfordulását, míg típusnak hívjuk az azonos karaktersorozatot tartalmazó tokenek osztályát. A típusok összessége alapján állítjuk össze […] a szótárat…" (Tikk – Kovács, 2007: 39).

Az információkinyerésen belül a névelem-felismerés mellett a *kapcsolatbányászat* említhető fontos részterületként. Kapcsolatbányászatról akkor beszélhetünk, amikor egy adott szövegben található elemek között a szemantikus, azaz jelentésbeli kapcsolatok is felfedezésre és jellemzésre kerülnek (Jiang, 2012: 22).[2] Különösen fontos itt az elemzéshez felhasználandó szótár megfelelő kialakítása, hiszen a nem megfelelően kialakított szótár sokat torzíthat az eredményeken. Problémát jelenthet például, ha a különböző *névelemosztályok* között átfedés található (például valaki vezetékneve egyúttal földrajzi név is, ld. pl. Balaton Károly), vagy ha egy névelemen belül egy másik is található (pl. Budapest Főváros Kormányhivatala – Tikk et al., 2006: 29–30).

A névelem-felismerés körében két alapvető módszer alkalmazására van lehetőség. Az egyik a szabályalapú módszer, amikor előre megadott adatok alapján kerül kinyerésre az információ, (ilyen szabály lehet a mondatközi nagybetű mint a tulajdonnév kezdete). Mivel ezek a szabályok ütközhetnek egymással, szükséges közöttük hierarchiát felállítani. A másik módszer a statisztikai tanulás, amikor a gép alkot szabályokat a kutató előzetes mintakódolása alapján (Jiang, 2012: 16–22 – minderről bővebben ld. III.1. és III.3. fejezet).

Végezetül egy adott névelem-felismerő rendszer által produkált eredmények *hitelességének* ellenőrzésére többféle mutató létezik. A pontosság (megbízhatóság) azt mutatja meg, hogy az adott korpuszban a rendszer által felismert névelemek közül mekkora a helyes megoldások aránya. A *felidézés (fedés, teljesség)* pedig azt jelenti, hogy a találatok között hány darab lelhető fel az összes releváns dokumentum közül, mekkora az adott korpuszban talált névelemek aránya a szöveg többi részéhez képest. Ezek segítségével kiszámítható azok harmonikus közepét jelentő mutató, a széleskörűen használt F-mérték (Tjong Kim Sang – De Meulder, 2003: 143–144; Hobbs et al., 1991: 2; Vázsonyi – Tikk, 2007: 69–70).

Egy tipikus alkalmazás

A névelem-felismerés tipikus alkalmazásai az egyes névelemjellemzők vagy *tématerületek (szövegtartományok, domainek)* sajátosságaihoz kapcsolódnak. Ennek oka az, hogy a korábbi kutatások szerint például egy földrajzinév-kigyűjtésre kialakított technika csak korlátozottan hasznos eredményeket hoz például intézménynevek kapcsán. Igen fejlett névelem-felismerési módszereket alkalmaznak napjainkban például az orvostudomány, a honvédelem és a nem-

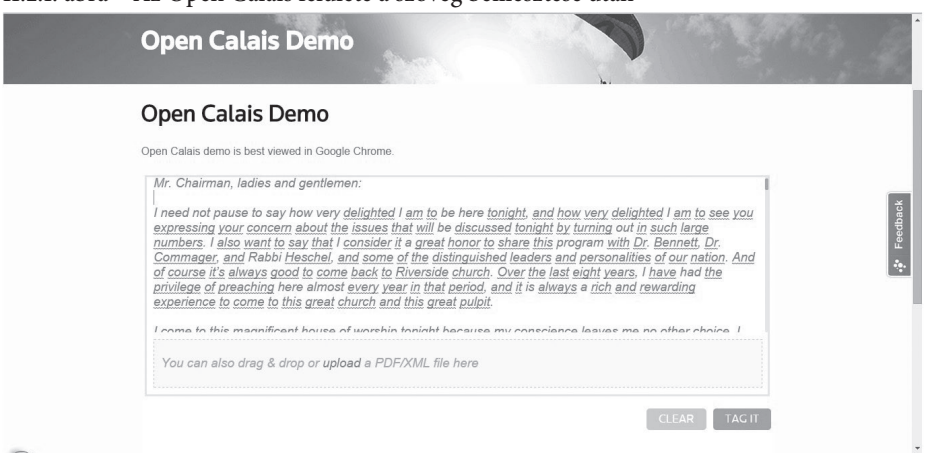
[2] A szemantika vagy jelentéstan legszűkebb értelmében a nyelv különböző összetevőinek (szavak, kifejezések) jelentésével foglalkozó, a nyelvészeten belüli tudományterület (Munk, 2014: 312). Szintaktikusan helyesnek kell lennie a mondatnak ahhoz, hogy szemantikájáról beszélhessünk (Bach, 2005: 20).

zetbiztonság (hírszerzés, felderítés, terrorizmusmegelőzés) terén (Neri et al., 2011: 393).

A következőkben egy konkrét politikatudományi alkalmazási lehetőséget mutatunk be. A feladat Martin Luther King, a vietnami háborút ellenző és a társadalmi igazságosságról szóló, 1967. április 4-i beszédének[3] vizsgálata a szemantikai információk strukturálatlan webes szövegekből való kinyerésére létrehozott Open Calais[4] online program segítségével.

Elsőként be kell illeszteni a vizsgált szöveget, majd a program egyetlen kattintásra elvégzi az elemzést (ld. II.2.1. ábra).

II.2.1. ábra – Az Open Calais felülete a szöveg beillesztése után



Mint látjuk (II.2.2–II.2.4. ábra) a program képes felismerni témaköröket, városokat, kontinenseket, országokat, ipari kifejezéseket, filmeket, szervezeteket, személyneveket, pozíciókat, államokat és tartományokat, régiókat, illetve különböző eseményeket és tényeket, akár idézeteket is. Azonban a felismerés során hibák is léphetnek fel.

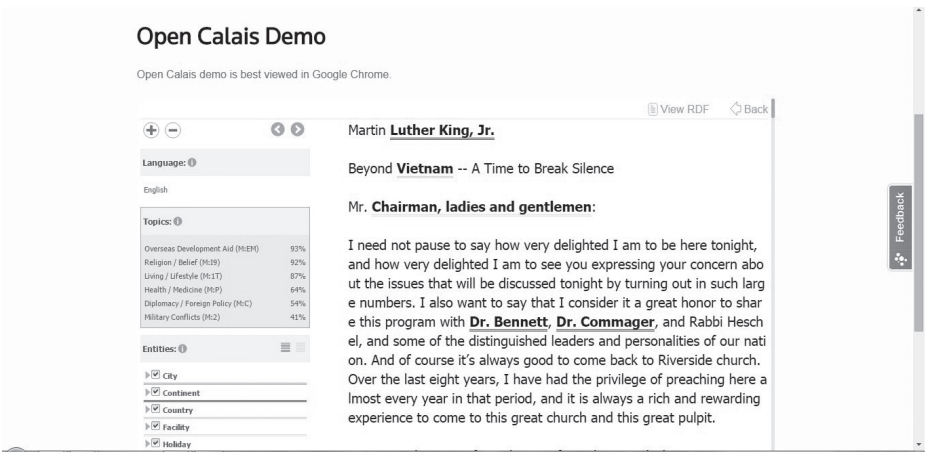
A II.2.2. ábrán láthatjuk például, hogy a program szerint igen jelentős mértékben vallási témájú szövegről van szó, illetve hogy volt olyan névelem, például Herschel rabbi neve, amelyet nem ismert fel. További hiba, hogy az applikáció a francia nemzetközösséget ünnepnapnak tekintette, hogy a kollektív megoldást az ipari kifejezésekhez sorolta be, illetve hogy a Dien Bien Phu nevet filmként ismerte fel (II.2.3. ábra). Láthatjuk továbbá, hogy Martin Luther

[3] Martin Luther King, Jr. Beyond Vietnam – A Time to Break Silence http://www.americanrhetoric.com/speeches/mlkatimetobreaksilence.htm (Letöltés ideje: 2015. október 30.)

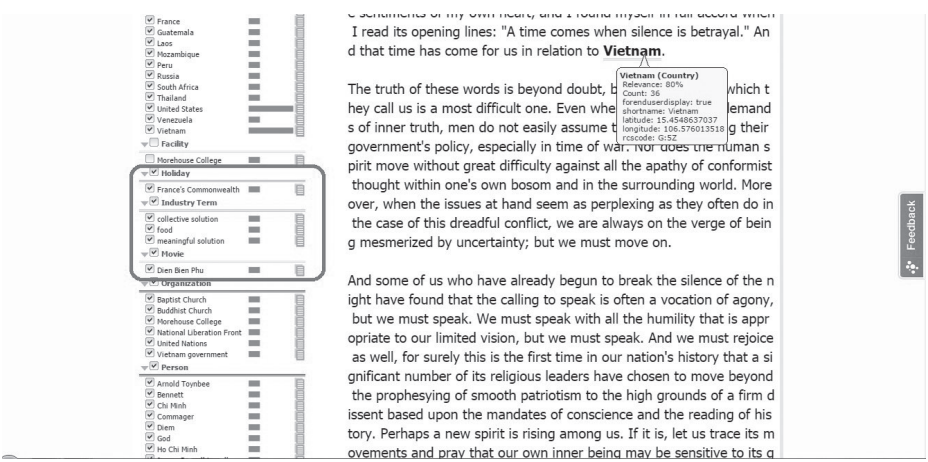
[4] Thomson Reuters – Open Calais Demo http://www.opencalais.com/opencalais-demo/

King vezetéknevét egy alkalommal pozícióként (királyként) ismerte fel, a Genfi Egyezményt pedig Genfnek tekintette (II.2.4. ábra).

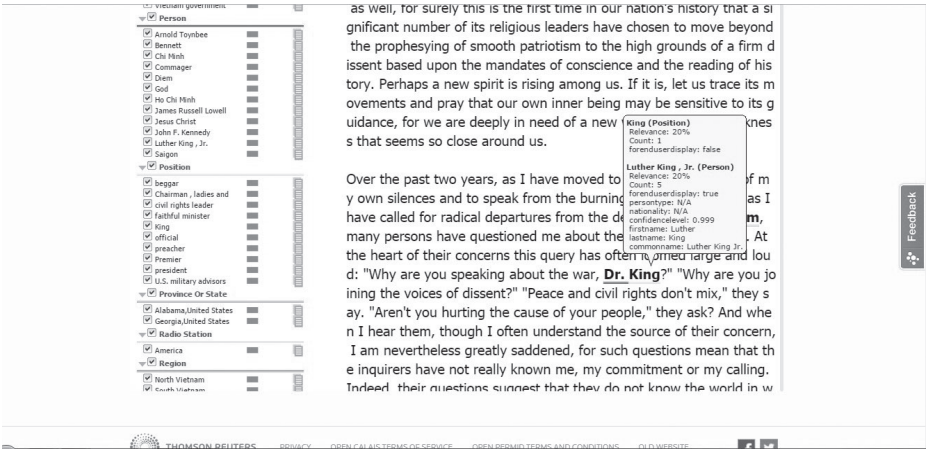
II.2.2. ábra – Elemzés az Open Calais programmal



II.2.3. ábra – Félreértelmezett kifejezések a szövegben

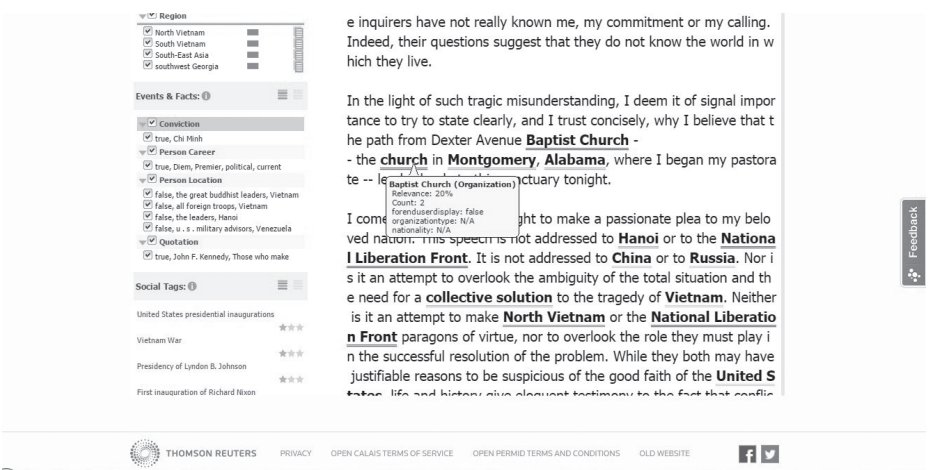


II.2.4. ábra – Viszonyok felismerése és elemzési hiba

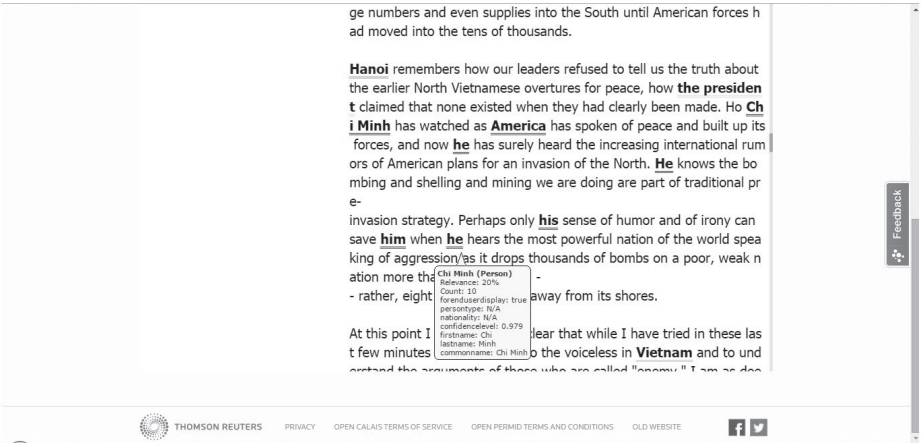


A program képes összekapcsolni a kifejezéseket, felfedezni a viszonyokat, így felismerni azt is, ha később csupán egy személyes névmással vagy rövidebb kifejezéssel utalunk egy személyre vagy szervezetre (II.2.5–II.2.6. ábra).

II.2.5. ábra – Egyes szavak, kifejezések közötti kapcsolatok felismerése 1.



II.2.6. ábra – Egyes szavak, kifejezések közötti kapcsolatok felismerése 2.



Láthatjuk tehát, hogy a névelem-felismerés jelentős mértékben képes segíteni a kutatómunkát, ám azt is, hogy az eredményeket óvatosan kell kezelni. További probléma, hogy mindeddig kevés magyar nyelvű szótár, illetve névelem-felismerő rendszer készült, amelyekre alább hozunk példákat.

Nemzetközi és hazai politikatudományi kutatások és egyéb alkalmazások

A politikatudományban gyakori alkalmazás az egyes közösségi oldalakon történő interakciók vizsgálata, ám például mikroblogok adatait egyéb politikatudományi kutatásokban is felhasználták. Számos tanulmány született Twitter-bejegyzések kapcsán (ld. pl. Nebhi, 2012). Tumasjan és társai (2010) azt vizsgálták, hogy a 2009. szeptemberi németországi választások eredményei előre jelezhetők-e Twitter-bejegyzések alapján. Az LIWC szövegelemző program segítségével több mint 100 000 bejegyzésben vizsgálták meg, hogy említésre kerül-e benne valamely német politikus vagy politikai párt. Arra jutottak, hogy a választás közeledtével növekedett a bejegyzések száma, s hogy a Twitter mára a politikai viták egyik színterévé vált. A bejegyzések vizsgálata során hasonló eredményre jutottak egyes pártok támogatottsága tekintetében, mint a közvélemény-kutatások. A névelem-felismerés az ilyen kutatások mellett alkalmas társadalmi hálózatok, így terroristacsoportok felderítésére is (ld. pl. Diesner – Carley, 2005).

Magyarországon egy érdekes alkalmazás a Szervezett Bűnözés Elleni Koordinációs Központ és a Szegedi Tudományegyetem közös projektjében elkészült

magyar nyelvű, bűnügyi névelem-felismerő rendszer, amelyről Molnár és társai (2010a; 2010b) írtak, illetve tartottak előadást a 2010. évi Magyar Számítógépes Nyelvészeti Konferencián. Ez a program egy, a Szegedi Tudományegyetemen korábban már kifejlesztett rendszer továbbfejlesztett változata. A rendszer megalkotásának célja a rendőrségi zárójelentések gyors feldolgozásának, a lényeges adatok kinyerésének és statisztikai adatok előállításának biztosítása volt. Ezen alkalmazás különlegessége, hogy nem a szokásos négy névelemosztályt (földrajzi név, személynév, szervezetnév, egyéb tulajdonnév), hanem összesen tizenhárom szemantikai osztályt különböztettek meg elkészítése során (például irányítószám, város, kerület, utca, házszám). Mivel a projekt során a névelemek egyes típusainak (például vezetéknév és keresztnév) megkülönböztetésére is szükség volt, kétszintű predikciós módszert alkalmaztak.

Nehézséget okozott a különböző névelemosztályok közötti gyakori átfedés (például egy adott szó településnév és vezetéknév is lehet). A *tanítóhalmazt* és a *teszthalmazt* kétszáz anonimizált dokumentumból állították össze, így a személynevek hiánya miatt a teszteredmények nem pontosak. Bár a valós adatokon készített modell egyes névelemosztályok tekintetében jobb eredményt mutatott, tesztelése összességében elmaradt az anonimizált dokumentumokon végzett vizsgálattól. A rendszer ugyanakkor lehetőséget nyújt az egyes név- és címadatok kiszűrésére (Molnár et al., 2010a: 366–369; Molnár et al., 2010b). Az ilyen üzleti és kormányzati alkalmazások egyre elterjedtebbek: Solt Illés és társai (2010) névelem-felismerést alkalmaztak általános egészségügyi problémák, gyógykezelések és vizsgálatok kórházi zárójelentésekben való azonosítására.

II.2.1. példa

A névelem-felismerésre érdekes példát nyújtanak az 1991-ben és 1992-ben megrendezett Dokumentummegértési Konferenciák (MUC-3 és MUC-4) témái. Mindkettő a Latin-Amerikában folytatott terroristacselekményekre vonatkozó adatok kinyerését tűzte ki célul. A projektben különböző, Latin-Amerikában elkövetett terroristacselekményekre vonatkozó hírösszefoglalókból hozták létre a mintegy ezerháromszáz szövegből álló korpuszt. Mindezt a hat modullal rendelkező, úgynevezett TACITUS rendszerrel végezték el, lehetővé téve a releváns események automatikus felfedezését és különválogatását (Hobbs et al., 1991). E két konferencia célja – csakúgy, mint a többié – az volt, hogy sor kerülhessen az információkinyerés kapcsán a szerzett tapasztalatok cseréjére és a terület további fejlesztésére, s ezzel továbbá a kormányzati igények kielégítésére.[5]

[5] Message Understanding Conference Proceedings http://www-nlpir.nist.gov/related_projects/muc/proceedings/proceedings_index.html (Letöltés ideje: 2015. december 3.)

Így jelentős eredménynek volt tekinthető, hogy a MUC-4 során kinyert adatok validitása magasabb szintű, a felidézés, a pontosság és az F-mérték is jobb, sokkal konzisztensebb az eredmény (Sundheim, 1992: 9–21).[6]

II.2.2. példa

A Twitter NewsCloud Alkalmazás segítségével megfigyelhetjük, hogy egy adott napszakban, tízperces bontásban egy adott hírfolyamban milyen gyakran és hányszor kerülnek említésre egyes névelemek (személyek, helyek, szervezetek).[7] Fehér Katalin írásából tájékozódhatunk a GeoX Kft. és a Zetema Ltd. által közösen kifejlesztett internetes tartalomelemző rendszer egy konkrét alkalmazásáról. Steve Ballmer, a Microsoft akkori vezérigazgatója 2013. augusztus 23-án jelentette be a cég éléről egy éven belül történő távozását. Az OpinHU webes véleményelemző rendszer segítségével megfigyelésre kerültek a Twitteren ezzel kapcsolatosan megjelenő aktivitások. Ennek keretében Steve Ballmer és a Microsoft említését vizsgálták 2013. augusztus 1-je és 29-e között, a Twitter NewsCloud alkalmazással. Ebből kiderült, hogy az első pár tíz percben a két tulajdonnév együtt jelent meg a hírfolyamban, a Microsoft említése ugrásszerűen növekedett, ahogyan Steve Ballmeré is, akinek a neve említésre sem került az esemény előtt. A hír hirtelen felfutása után az említések száma zuhanásszerűen le is csökkent.[8] Az OpinHu internetes tartalomelemző rendszer nyelvtechnológiai hátteréről bővebben Miháltz (2010) írásából tájékozódhatunk. Kitűnően alkalmazhatjuk tehát a névelem-felismerést akkor, ha egy meghatározott ügy közösségi médiában vagy mikroblogolás során történő felbukkanását szeretnénk vizsgálni.

II.2.3. példa

Érdekes alkalmazás az Európai Unió, az Európai Unió Belügyi Főigazgatósága közreműködésével kifejlesztett, az Európai Unió határvédelmi ügynökségéhez, a Frontexhez köthető nyílt forrású felderítő rendszer, amely a határvédelemhez köthetően képes azonosítani személyek és szervezetek neveit, helyeket, kapcso-

[6] A MUC-3 és MUC-4 adatbázisai ingyenesen hozzáférhetők: MUC Data Sets http://www-nlpir.nist.gov/related_projects/muc/muc_data/muc_data_index.html (Letöltés ideje: 2015. december 3.)

[7] Twitter NewsCloud alkalmazás https://sites.google.com/a/geox.hu/opinhu/elemzesi-eszkoezoek/twitter-newscloud-alkalmazas (Letöltés ideje: 2015. október 29.)

[8] Fehér Katalin: Microsoft – Steve Ballmer lemondása a Twitteren https://sites.google.com/a/geox.hu/opinhu/hirek-ujdonsagok/microsoft-steveballmerlemondasaatwitteren (Letöltés ideje: 2015. október 29.)

latfelvételi adatokat (pl. telefonszám, e-mail cím), hitelkártyaszámokat, továbbá adóazonosítókat. A rendszer az adatok összegyűjtésén és csoportosításán túl képes azok elemzésére, kapcsolatbányászatra is. A rendszer csak zárt körben elérhető, a végfelhasználók az Európai Unió egyes tagállamainak hatóságai, rendőrségei, határőrségei. A segítségével felhalmozott tudás a korábbi ismeretekkel kombinálva segíti elő a felderítést.[9]

Egy másik, a határőrizethez kapcsolódó alkalmazás a Frontex valós idejű híresemény-feldolgozó keretrendszere, amely képes nyolc nyelven, valós időben kigyűjteni az online hírekből és egyéb nyílt forrásokból (pl. mikroblogokból) a határőrizeti szempontból releváns strukturált információkat, szintén megkönnyítve a felderítést. Ilyen, lényeges információk lehetnek az illegális határátlépési ügyek és az ehhez kapcsolódó olyan bűncselekmények, mint az embercsempészet és emberkereskedelem, továbbá a válsághelyzetek (erőszakos események, tüntetések, fegyveres konfliktusok, humanitárius vagy természeti katasztrófák, illetve fertőző betegségek). A rendszer nem csupán az információk kinyerésére alkalmas, hanem annak egy térképen való elhelyezésére is képes. Segítségével nem csupán az illegális határátlépéshez kapcsolódó felderítés válik könnyebbé, hanem segítséget nyújthat a migrációs nyomás várható megnövekedésének előrejelzésére és az arra való felkészülésben is.[10]

Ellenőrző kérdések

- Mi a különbség a névelem és a tulajdonnév között?
- Mi a különbség a névelem-felismerés és az osztályozási probléma között?
- Melyek a névelem-felismerés előnyei a szózsák modellhez képest?
- Találjon ki egy feladatot, amelyet csupán a névelem-felismerés segítségével lehet megoldani, a szózsák modell segítségével nem!
- Milyen politikatudományi kérdés megoldására alkalmazná a névelem-felismerés módszerét?
- Tegyük fel, hogy az Országgyűlési Naplóból ki szeretnénk gyűjteni a személyneveket és a földrajzi neveket! Ez esetben milyen módszertani problémák merülhetnek fel a névelem-felismerés alkalmazása során?

[9] EMM Open Source Intelligence Suite http://btn.frontex.europa.eu/system/files/private/resources/tools/emm-osint-suite_productinfo.pdf (Letöltés ideje: 2015. október 30.)

[10] Frontex Real-time News Event Extraction http://btn.frontex.europa.eu/projects/internal/frontex-real-time-news-event-extraction (Letöltés ideje: 2015. december 3.)

Szószedet

Magyar	Angol
Címke	Label
Dokumentummegértési konferencia	Message Understanding Conference (MUC)
Felidézés, fedés, teljesség	Recall (score)
F-mérték	F1 score
Hitelesség	Accuracy
Információkinyerés	Information extraction
Kapcsolatbányászat	Relation extraction
Névelem-felismerés	Named entity recognition
Pontosság, megbízhatóság	Precision (score)
Szóalak	Token
Szövegelőkészítés, -előfeldolgozás	Preprocessing
Tématerület (szövegtartomány)	Domain
Természetesnyelv-feldolgozás	Natural language processing
Természetesnyelv-tanulási konferencia	Conference on Natural Language Learning (CoNLL)
Webes információkinyerő rendszer	Wrapper

Ajánlott irodalom

A névelem-felismeréssel kapcsolatos ismeretek mélyítéséhez, bővítéséhez az alábbi irodalmakat ajánljuk: Marrero et al. (2013); Russel – Norvig (2005); Tikk (2007); Tikk et al. (2005); Móra – Farkas (2010); Gosztolya – Tóth (2010).