Az Európai Médiafigyelő (EMM) magyar változata

Pajzs Júlia

MTA Nyelvtudományi Intézet Nyelvtechnológiai Kutatócsoport
1394 Budapest Pf. 360
pajzs.julia@nytud.mta.hu

Kivonat: A Közös Kutatóközpont – Europa (European Joint Research Centre) által fejlesztett európai médiafigyelő (http://emm.newsbrief.eu) világszerte több ezer hírportálról automatikusan gyűjti, és különféle kategóriákba sorolja a híreket, a nap 24 órájában, 10 percenként frissítve, nyelvtechnológia eszköztár használatával. Az MTA Nyelvtudományi Intézet Nyelvtechnológiai Kutatócsoportja együttműködési megállapodás keretében a szolgáltatás magyar nyelvű működését tette lehetővé. A magyar tulajdonneveknek az EMM rendszeren belüli felismerése és a toldalékolt változatok kezelése volt az elsődleges feladat. A nemzetközi jelentőségű híreket valamennyi feldolgozott nyelvi változatukban elérhetjük.

1 Bevezetés

Az Europe Media Monitor (EMM) teljesen automatikus médiafigyelő rendszer lehetővé teszi, hogy a felhasználók naprakészen tájékozódjanak az on-line média őket érdeklő tartalmairól. Több tucatnyi különböző nyelvből nyelvtechnológiai eszközök segítségével összegyűjti a híreket, és részleges elemzést, információkivonatolást hajt végre rajtuk. Mivel a rendszer megalkotói sok nyelv hatékony feldolgozását tűzték ki célul, nem kívántak az egyes nyelvek morfológiai, szintaktikai, szemantikai elemzésére az egyes nyelvekre kifejlesztett eszközöket alkalmazni. Némelyik korábban feldolgozott nyelv esetén járható út volt a várható toldalékolt alakok listaszerű feldolgozása [5], ez a magyar szövegekre nem lett volna reális célkitűzés [4]. A magyar modul illesztéséhez számos segédanyagot készítettünk és adtunk át, valamint az eredmény tesztelésében nyújtottunk segítséget. A cikkben az EMM rövid általános ismertetésén kívül az átadott anyagokat és az eredményeket ismertetem.

1.1 Információkinyerés az EMM-ben

A hírek klaszterekbe sorolása 10 percenként frissül. Ha egy hír jelentős részben azonos egy néhány órán belül korábban megfigyelt hírcsoport elemeivel, ennek a klaszternek tartalmához adódik. (Részletesebben kifejtve lásd [2, 3]).

Előre elkészített többnyelvű kategóriadefiníciókat tartalmazó állományok segítségével a klasztereket automatikusan témakörökbe sorolja a rendszer (természeti csapások, terrorizmus stb.).

Különféle segédállományok (ismert személynevek, titulusok, foglalkozások, népnevek stb. listák) felhasználásával igyekszik automatikusan felismerni a szövegben előforduló személyneveket. A napi hírösszesítőben (NewsExplorer) feltünteti a leggyakrabban szereplő személyneveket, ezek kapcsolatrendszerét más személyekkel, egyéb fontos neveket (pl. intézmények).

A hírek címében felismert földrajzi nevek alapján a világtérképen is elhelyezi a hírklasztereket.

A több nyelven megjelent azonos híreket mindegyik, az EMM által feldolgozott nyelven megtekinthetjük.

2 A magyar források

2.1 Javaslat a feldolgozandó hírportálok bővítésére

Már korábban is figyeltek néhány magyar nyelvű portált. Ezek bővítésére tettem javaslatot. Azt tartottam szem előtt, hogy a politikailag különböző térfélen állók képviselve legyenek. Fontos kiegészítés volt a határon túli magyar nyelvű portálok hozzáadása a rendszerhez, amely így jelenleg 66 magyar nyelvű portált kezel (http://emm.newsbrief.eu/NewsBrief/sourceslist/hu/list.html).

2.2 A személynevek felismerését segítő anyagok

- Az aktuálisan érvényes magyar keresztnév listák (3215 elem).
- Titulusok listája (úr, asszony, hölgy) (kb. 400 elem).
- Fontos beosztások listája (pl. miniszterelnök) (kb. 650 elem).
- Foglalkozásnevek, kategóriákba sorolva (pl. kutatóorvos → HEALTH, RESEARCH).
- Népnevek (francia, gall) (kb. 730 elem).

E listák segítségével igyekeznek felismerni a *Lech Kaczynski lengyel elnök, Németh Lászlóné miniszter asszony* jellegű szerkezeteket.

A listákon a toldalékolható szavak végén szerepel a *„%”* karakter, jokerként (annak jelzésére, hogy a megadott szavakat egyéb karakterek követhetik). A változó tőalakokat is feltüntettem. Átadtam az egyszerű névszói toldalékok listáját is.

2.3 Idézetek felismerését segítő igék

Az alábbi igelistát adtam át, amelyek MNSZ-beli gyakoriságuk csökkenő sorrendjében szerepelnek: mond, jelent , elmond, ír, beszél, szól, közöl, kérdez, megállapít, jelez, kijelent, válaszol, hangsúlyoz, nyilatkozik, bejelent, megerősít, megjegyez, idéz, fogalmaz, beszámol, magyaráz, elárul, fenyeget, rámutat, tisztáz, felidéz, méltat, összegez, faggat, fenyegetőzik, nehezményez, deklarál, elbeszél, panaszol, tudakol.

Minden igét egyes és többes szám harmadik személyű, jelen és múlt idejű változatában adtam meg, az elváló és hátravetett igekötős igéknél ezeket a változatokat is feltüntettem.

2.4 Földrajzi nevek

A földrajzi neveket több nemzetközi adatbázist felhasználva dolgozták fel. A különböző listákból származó adatokat igyekeztek egyértelműsíteni [1]. A keletkezett adatbázis az egyes nevek földrajzi koordinátáit tartalmazza, valamint egy kódot, amely arra utal, mennyire nagy jelentőségű az adott helynév (ország, főváros, nagyváros stb.) Ezt az adatbázist kellett kiegészítenem részben magyar nevekkel, részben nemzetközi nevek magyar változataival (Wien→Bécs, Beijing→Peking). Valamennyi már korábban is meglévő nevet, szükség esetén ki kellett egészíteni olyan alakváltozattal, amely magyar toldalékok előtt állhat (Prága→Prágá%).

A feldolgozott magyar hírek címében augusztusban talált földrajzi nevek listáját ellenőriztem. A vizsgált nevek 40%-a toldalékolt formában fordult elő a szövegben. Az előfordult toldalékolt nevek 24%-a nemzetközi név volt, így beigazolódott, hogy nem csupán a toldalékolt magyar helynevek korrekt felismerése fontos. A jelenlegi megoldás elfogadható: minden legalább 5 karakter hosszú név végén ott van a „*%*" karakter, ami jelzi, hogy bármely karaktersorozat követheti a nevet. Ebből adódnak ugyan félreelemzések (pl. a *Gabonatermesztés*ről szóló cikket *Gabon* államnál helyezi el a térképen), a félreelemzések száma azonban a vizsgált egy hónapnyi mintában 4% alatt maradt. A félreértéseknek, félreelemzéseknek természetesen más forrása is van: nem egy keresztnév földrajzi név is egyúttal, de maguk a földrajzi nevek is sokszor utalnak különböző helyekre. A többértelműségeket és félreértéseket folyamatosan gyűjtjük, javítjuk.

2.5 Kategóriadefiníciók

A tematikus keresés lehetővé tételéhez többnyelvű kategóriadefiníciós állományokat használnak. Az egyes kategóriák definíciós állománya több részből áll össze az alábbiakban láthatunk erre példát. Az első részben azok a szavak szerepelnek, amelyek tipikusan előfordulhatnak az adott témájú hírekben. A szavak után látható szám az adott szó súlyára utal, minél nagyobb a szám, annál jellemzőbb a szó az adott kategóriára. A nagy negatív súllyal jelölt *film* és *könyv* szavak azt jelzik, hogy ha ezek a szavak is előfordulnak az adott hírben, ne tekintse a kategóriába tartozónak, hiszen akkor feltehetőleg egy ilyen témájú film vagy könyvismertetést tartalmaz a hír. Az állományok második részében szókombinációk megadására van lehetőség, bizonyos kombinációk ki is zárhatók (pl. ha a *bomba* szó közelében *foci, futball*, vagy *meccs* szerepel, ne sorolja a hírt a Terrorizmus kategóriába). A „*%"* karakter ebben a példában is a joker karaktert jelöli.

Az *Embercsempészet* témakör kategóriadefiníciós állománya

```
Alert definition
Alert ID: HumanTraffic

Description: Human Traffic
Patterns

Pattern Weight
emberkeresked% 20
ember%csempész% 20
illegális%bevánd% 20

film -999
könyv -999

A combination of at least one of
Proximity: 20
emberkeresked%
ember%csempész%
nő%keresk%
szex%rabszolg%
rabszolg%
kényszermunk%

and at least one of
szervez%+bűnöz%
prostitu%
áldozat%
gyermek%
csecsemő%
bevándorl%
```

3 Kiértékelés

3.1 Toldalékstatisztika

A különböző tulajdonnév listák kiértékelésének melléktermékeként toldalékstatisztika is készült. Ezek alapján megfontolásra érdemesnek tűnik, hogy csupán a leggyakoribb toldalékok felismerését célozzuk meg, néhány egyszerű reguláris kifejezéssel. Míg az összesített tulajdonnév listában a *t* tárgyrag különböző alakjai, a *bAn* és a *nAk* és a *vAl* fordultak elő leggyakrabban (az összes toldalékolt alak 90%-a), addig a földrajzi neveknél a bAn és az On toldalék alakjai szerepeltek nagyon gyakran (az összes toldalékolt alak 73%-a).

A részletes statisztikát az 1. táblázat tartalmazza.

1. táblázat: Toldalékstatisztika

Toldalék	Standard	Funkció	Fqs1	Fqs5	FqsTOP	Fqs700	Fq NewReC	FqGeo
AKAT	K+T	PL+ACC					1	
AS	S	N→A Der					1	
AT	T	ACC	1	5			2	4
BA	BA	ILL	1	5		3	13	25
BAN	BAN	INE	1	35	187	83	45	498
BE	BA	ILL					1	13
BEN	BAN	INE		10	69	16	10	88
BÓL	BÓL	ELA		5			4	41
BŐL	BÓL	ELA					3	5
CSAL	VAL	INS			17	1		
CSEL	VAL	INS		5				
DAL	VAL	INS		5				
DEL	VAL	INS					2	
É	É	POS				1	6	4
ÉK	ÉK	Der			19	31	5	
ÉKKAL	ÉK+VAL	Der+INS		5		5		
ÉKBÓL	ÉK+BÓL	Der+ELA				2		
ÉKNAK	ÉK+NAK	Der+DAT				2		
ÉKON	ÉK+ON	Der+SUP				2		
ÉKRÓL	ÉK+RÓL	Der+DEL				1		
ÉKTÓL	ÉK+TÓL	Der+ABL				3		
EN	ON	SUP		20		11	8	300
ÉRT	ÉRT	CAU					5	9
ET	T	ACC	1	15	34	6	19	12
ETT		SUP				1		
FÉLE	FÉLE	N→A Der				5		
FAL	VAL	INS						
GAL	VAL	INS		5			6	9
GEL	VAL	INS		5				
GYEL	VAL	INS			20		1	
HEZ	HOZ	ALL					1	
HOZ	HOZ	ALL		5		28	9	13

IG	IG	TER						1
JÁT	JA+T	PERS+ACC					1	
JE	JA	PERS					1	
JÉT	JA+T	PERS+ACC					3	
KAL	VAL	INS					1	
KEL	VAL	INS						1
KÉNT	KÉNT	FOR		5				
LAL	VAL	INS					1	
LEL	VAL	INS		10			3	
LYAL	VAL	INS				8	2	
LYEL	VAL	INS				4		
MAL	VAL	INS		5		2		3
MEL	VAL	INS					1	
N	ON	SUP	1	5		6	13	79
NAK	NAK	DAT	3	40	76	119	54	31
NEK	NAK	DAT	1	40	179	26	39	4
NAL	VAL	INS		20		18	8	52
NÁL	NÁL	ADE		5		4	9	
NEL	VAL	INS	1	5			3	2
NÉL	NÁL	ADE	1	10			1	18
ON	ON	SUP	1	15		6	16	189
ÖN	ON	SUP						6
OT	T	ACC	1	15	24	10	20	41
ÖT	T	ACC		5	57			
RA	RA	SUB	1	5		18	34	6
RE	RA	SUB		5		1	2	26
REL	VAL	INS				2	5	
RAL	VAL	INS		5		5		
RÓL	RÓL	DEL		10		7	21	9
RŐL	RÓL	DEL		10			2	11
SAL	VAL	INS				11	1	
SZAL	VAL	INS					3	
SZEL	VAL	INS					1	
T	T	ACC	9	110	197	239	130	33
TAL	VAL	INS	1				3	

TEL	VAL	INS			19	3	2	
TÓL	TÓL	ABL	1	25		35	24	22
TŐL	TÓL	ABL	1			3	14	16
UK	JUK	PERS				10		
VAL	VAL	INS	2	30	15	18	13	7
VEL	VAL	INS		10		1	4	
ZEL	VAL	INS					1	
Summa			**28**	**515**	**913**	**757**	**578**	**1578**

Fqs1 Az 1 gyakorisággal előforduló (Sample 1) tulajdonnév lista kézzel javított része alapján észlelt összegzett toldalék gyakoriság.

Fqs5 Az 5 gyakorisággal előforduló (Sample 5) tulajdonnév lista kézzel javított része alapján észlelt összegzett toldalék gyakoriság.

FqsTop A leggyakoribb (legalább 15) gyakorisággal előforduló (Sample Top) tulajdonnév lista kézzel javított része alapján észlelt összegzett toldalék gyakoriság.

Fqs700 A teljes felismert tulajdonnév lista „magyar" elemeiből 700 tétel (Sample 700) kézzel javított része alapján észlelt összegzett toldalék gyakoriság.

Fq NewRec A teljes felismert tulajdonnév lista „magyar" elemeiből 700 tétel (Sample New Rec) kézzel javított része alapján észlelt összegzett toldalék gyakoriság.

FqGeo Az augusztusban felismert földrajzi név lista (Geo Names) kézi javítása alapján észlelt összegzett toldalék gyakoriság.

3.2 Toldalékolt tulajdonnevek részaránya

Mivel a magyar morfológiai eljárások alkalmazását az EMM fejlesztői lehetőleg el szeretnék kerülni, igen fontos kérdés volt, hogy a felismert tulajdonnevek hány százaléka volt toldalékolva, azaz valójában mennyi az információveszteség abból adódóan, hogy *Bajnai Gordon, Bajnai Gordonékról, Bajnai Gordonékat* 3 különböző tétel a felismert nevek adatbázisában. A rendszer egyéves működése után készültek a teljes felismert tulajdonnév lista különböző részeiből azok a kézzel ellenőrzött listák, amelyeknek összesített és toldalékolt type/token arányát mutatja be a 2. táblázat.

2. táblázat: Type/Token arány

	Type	Token	Toldalé-kolt Type	Toldalé-kolt Token	Told Type/Type	Told Token/Token
Teljes névlista	22.464	102.041				
Sample1	100	100	28	28	0,280	0,280
Sample5	771	3. 855	103	515	0,133	0,133
SampleTop	1. 057	48 .938	37	913	0,035	0,0186

Sample 700	700	9.056	197	760	0,281	0,083
SampleNewRec	1.167	2.136	187	330	0,160	0,154
GeoNames	788	4.056	350	1581	0,444	0,389

3.3 Tulajdonnevek és idézetek felismerése

Azonos napon megjelent 11 hírből gyűjtöttem ki a bennük előfordult 100 személynevet, a rendszer csak a felét ismerte fel személynévként. Ez azzal magyarázható, hogy csak akkor tekinti az egymást követő nagybetűs szavakat személynévnek, ha a) már a rendszer által ismert személynév b) a nevet követően a személynév felismeréshez megadott különféle listák elemei közül legalább egyre illeszkednek a nagybetűs szavakat követő szavak. Így természetesen felismeri *Angela Merkel*t (15 előfordulás) és *Orbán Viktor*t (3 előfordulás), de nem ismeri fel *Csík János zenekarvezető*t, mivel ez utóbbi nem szerepelt a foglalkozásnevek listáján, ugyanezért *Varga Gábor hatóanyag-szakértő* is ismeretlen marad az EMM számára.

Az idézetek felismerésének aránya sajnos még ennél is rosszabb. Ez részben a személynév felismerés hiányosságából adódik. További probléma, hogy csak akkor tekint idézetnek egy szövegrészletet, ha idézőjelben van, és ugyanabban a mondatban szerepel a felsorolt igék (*mond, jelent* stb.) valamelyike és egy felismert személynév. Emellett túl szigorú volt az ige és a felismert személynév együttes előfordulásnak szabálya is (csak egy-egy előre megadott listán felsorolt szót engedtek meg közöttük: pl.: *tegnap, korábban, délelőtt* stb.), ezt a szabályt időközben javították (legfeljebb 3 bármilyen szó lehet az ige és a személynév között).

200 hír kézi ellenőrzésnél azt tapasztaltam, hogy az általam észlelt 120 idézetből mindössze 9-et azonosított a rendszer! 36 esetben adódott a hiba a személynév felismerés hiányából. 6 esetben az igét nem ismerte fel (például mert az elváló igekötő nem közvetlenül az ige után volt), egyes esetekben felmerült az igelista kibővítésének igénye is: *tette hozzá, zárta, véli*. Összességében úgy tűnik, a fontos közszereplők egymondatos idézeteinek van esélye a korrekt felismerésre. Bár ezek a részeredmények meglehetősen gyengének tűnnek, az EMM fő célkitűzése lényegében teljesülni látszik: a hírekben rendszeresen, gyakran szereplő fontos személyek (főként vezető politikusok) nézeteinek és kapcsolatrendszerüknek számontartása.

3.6 Témakörök szerinti keresés

A 2.5-ben definiált embercsempészet témakör keresésének eredménye

Két szabadkai embercsempészt fogtak el a rendőrök Mórahalomnál

16 felnőtt és gyermek határsértő akart két személyautóba bezsúfolódni, amikor tetten érték őket a

Szegedi Határrendészeti Kirendeltség járőrei Mórahalom külterületén

Segítséget kap Bulgária a menekültügy megoldásához

Technikai és pénzügyi segítséget nyújt az ENSZ Menekültügyi Főbiztossága és az Európai Unió is....

Hihetetlen: harminc éve rabszolgasorban tartott nőt szabadítottak ki

Három, harminc éve rabszolgasorban tartott nőt szabadított ki a brit rendőrség Londonban. Ketten külföldiek, a harmadik a fogságban születethetett. A rabszolgatartókat letartóztatta a brit rendőrség....

"Jól felszerelt" embercsempészeket buktattak le a magyar határrendészek

16 helyszínen 3 embercsempészt és összesen 91 határsértőt fogtak el a Szegedi Határrendészeti Kirendeltség járőrei a polgárőrökkel együttműködve Csongrád megye déli részén 24 óra alatt....

Éjjellátót is vitt magával a fülön csípett embercsempész

Éjjellátóval és mobiltelefonokkal szerelkezett fel egy Ásotthalomnál elfogott embercsempész....

4 Összegzés

Az EMM magyar modulja működőképes. A személynevek esetében a toldalékolt alakok aránya viszonylag alacsony (általában 10% alatt, a leggyakoribb nevek esetén mindössze 1,8%) ezért a morfológiai elemzés hiánya nem okoz jelentős problémát. A már ismert neveket biztonságosan felismeri a rendszer, az addig ismeretleneket csak akkor, ha kellőképpen fontos pozíciójuk vagy foglalkozásuk van és ez közvetlenül a név után, ugyanabban a mondatban expliciten megjelenik. A földrajzi nevek felismerésének aránya megfelelő (96%). A kiemelt témakörökre a tematikus hírkeresés használható. Jól működő nemzetközi hírfigyelő rendszerbe sikerült beillesztenünk a magyar modult.

Hivatkozások

1. Pouliquen, B.; Kimler, M. Steinberger, R., Ignat, C., Oellinger, T., Blackler, K. Fluart, F., Zaghouani, W., Widiger, A Forslund, Clive: Best Geocoding Multilingual Texts: Recognition, Disambiguation and Visualisation LREC (2006)
2. Steinberger, R., Pouliquen B., van der Goot E.: An Introduction to the Europe Media Monitor Family of Applications. In: Fredric Gey, Noriko Kando & Jussi Karlgren (eds.): Information Access in a Multilingual World - Proceedings of the SIGIR 2009 Workshop (SIGIR-CLIR'2009), Boston, USA. (2009) 1–8
3. Steinberger, R.: A survey of methods to ease the development of highly multilingual Text Mining applications. Language Resources and Evaluation Journal, Springer, Volume 46, Issue 2, (2012). 155–176
4. Steinberger, R., Maud, E., Pajzs, J., Mohamed, E., Steinberger, J. Turchi, M.: Multilingual media monitoring and text analysis - Challenges for highly inflected languages. In: Habernal, I., Matoušek, V. (eds). Text, Speech and Dialogue. 16th International Conference, TSD 2013, Pilsen, Czech Republic, September 2013, Proceedings. Springer Lecture Notes in Artificial Intelligence LNAI 8082 (2013) 22–33
5. Steinberger, R., Ombuya S., Kabadjov M., Pouliquen, B., Della Rocca, L., Belyaeva, J., De Paola. M., van der Goot, E.: Expanding a multilingual media monitoring and information extraction tool to a new language: Swahili. Language Resources and Evaluation Journal (DOI 10.1007/s10579-011-9165-9), Volume 45, Issue 3 (2011) 311–330