Magyar nyelvű webes szövegek számítógépes feldolgozása

Varga Viktor[1], Wieszner Vilmos[1], Hangya Viktor[1],
Vincze Veronika[2], Farkas Richárd[1]

[1] Szegedi Tudományegyetem, Informatikai Tanszékcsoport
{viktor.varga.1991,vilmos.wieszner,hangyav}@gmail.com,
rfarkas@inf.u-szeged.hu

[2] MTA-SZTE Mesterséges Intelligencia Kutatócsoport
vinczev@inf.u-szeged.hu

Kivonat: Cikkünkben bemutatjuk a magyar nyelvű webes szövegek elemzésével kapcsolatos nehézségeket, elsősorban Facebook-bejegyzésekre és kommentekre támaszkodva, valamint tárgyaljuk ezeknek lehetséges javítási módjait. A webes szövegek elemzése a belőlük kinyerhető információ miatt fontos, azonban a szabályos szövegeken tanult elemzők nem képesek hatékonyan feldolgozni ezeket. A megoldást az eddigi angolra alkalmazott, illetve a magyar nyelv sajátosságaira finomhangolt módszerek hozhatják meg.

1 Bevezetés

Az emberek életének évről-évre egyre nagyobb részében van jelen az internet, főként a rajta átáramló kommunikáció (gondoljunk csak a Twitterre vagy a Facebookra). Nagy mennyiségű adat jön létre a felhasználók egymással való kommunikációja folytán, és ez sok számítógépes nyelvészeti alkalmazás számára hasznos lehet, például az információ- és véleménykinyerésnél. Az utóbbi időben ezért jelentős fontosságra tett szert a webes szövegek, főként az ún. közösségimédia-szövegek (felhasználók által írt szövegek: blogok, állapotjelentések, chatbeszélgetések, kommentek) feldolgozása.

A közösségimédia-szövegekkel (*social media texts*) és azok elemzésével foglalkozó kutatások ugyanakkor rávilágítottak, hogy nagy nehézséget okoz ezen szövegek ún. nem sztenderd nyelvhasználata, jelentősen lecsökkenti a meglévő, szabályos szövegen (mint amilyen a Szeged Korpusz [1] is) tanult elemzők hatékonyságát. Az ezzel kapcsolatos kutatások legnagyobb része angol nyelvre született ([2, 3, 4]) és ezeknek magyarra való alkalmazása – mint az a sztenderd szövegek elemzésénél is megállapítható – nem hozna tökéletes eredményt. A magyar és az angol nyelv közötti morfológiai és szintaktikai különbségek ugyanis más megközelítést, más típusú szabályok bevezetését követelik meg. Az alapvető lépések hasonlóak, normalizálni, standardszerűvé kell a szöveget, ennek kivitelezése több módon történhet.

Cikkünk célja, hogy összefoglaljuk a közösségimédia-szövegek elemzésével kapcsolatos (elsősorban a Facebook-kommentekből és -posztokból álló tesztkorpuszon végzett) eredményeket, főbb hibakategóriákat és lehetséges megoldási módjaikat.

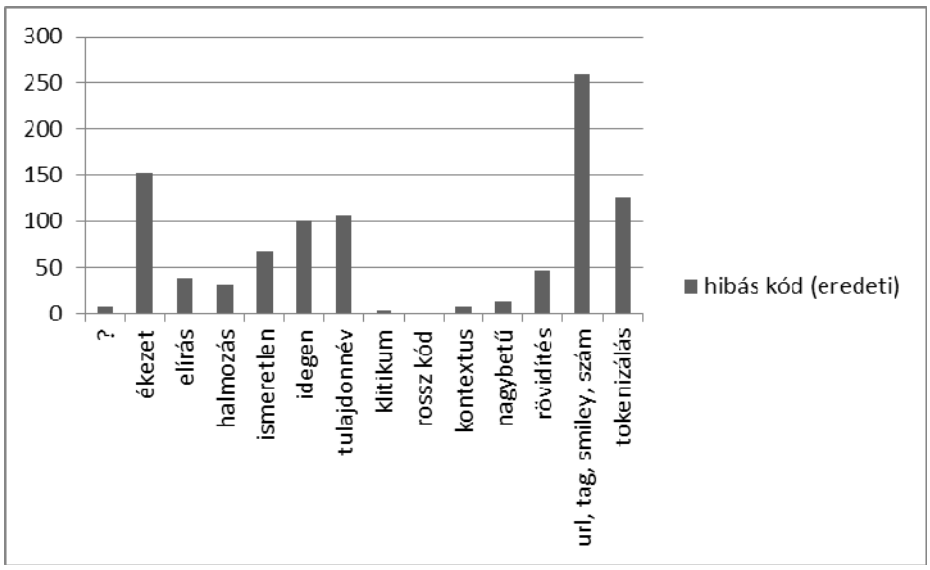
2 Problémák

A webes, azon belül a közösségimédia-szövegek nagy részének alapvető jellemzője, hogy írásbeli formájuk ellenére beszélt nyelvi sajátosságokat mutatnak. A szituációval ez könnyedén magyarázható: a szóbeli kommunikáció valósidejűségét (online) és multimodalitását egyszerre törekszik megtartani, így többek között az élmény (vagy vélemény) megosztásának gyorsasága és az érzelemkifejezés jelentős szerepet játszik a szövegekben, a hibák nagy része is ezeknek tudható be. A gyorsaságot ugyanis – a bevitelből adódóan – a gépelés gyorsításával lehet elősegíteni: többek között ékezetek mellőzésével (*ugyse* /úgyse/, *hat* /hát/, *lehet egy hulye kerdesem?)*, központozás és nagybetűk hanyagolásával, rövidítésekkel (*h*, *sztem*, *lécci*), egybeírással (*nemtom*, *énis*), valamint többnyire nem szándékoltan félregépeléssel (*mindegyekinek* /mindegyiknek/). A hétköznapi szóbeli kommunikációban elengedhetetlen érzelemkifejezés megnyilvánulhat a nagybetűhasználatban, a betű- és központozáshalmozásban (*jóóó*, *lehet ezekkel dolgozni???*), és az emotikonok használatában. Egyéb „zajok" a hezitáció explicitté tétele (*őőő*, *khm*), a nyelvi kreativitás termékeinek, illetve angol szavaknak és rövidítéseknek (*cool*, *wtf*, *pls*) a használata. Mindezek egyénenként és regiszterenként, illetve környezetenként változnak.

Az általános jellemzőkön kívül megállapítható, hogy a hibák szempontjából a közösségimédia-szöveg sem homogén kategória, az elemzők számára vannak könnyebben (blogok, Facebook-állapotjelentések) és nehezebben feldolgozható szövegek (kommentek, chat, mikroblogos bejegyzések). A blogok nagy részére jellemző a helyesírási szabályok lehetőség és képesség szerinti betartása, így ezekkel jobban boldogulnak, mint a beszélt nyelvre inkább hasonlító (akár több résztvevős) chatszövegnél, ahol a mondatra szegmentálás is problémát okoz az írásjelek és nagybetűk következetlen használata miatt.

Következő lépésben a tesztkorpuszt (150 Facebook státuszüzenet és 350 komment) a magyarlanc morfológiai és szintaktikai elemzővel [6] leelemeztük, majd kézzel részletes hibaellenőrzést végeztünk, ezután a hibákat a fentebb megállapított kategóriákba soroltuk. A különböző morfológiai hibakategóriák a nyers szövegben az 1. ábrán látható arányban fordultak elő. A számok a hibásan kódolt (X kódú, azaz le nem elemzett, illetve hibás szófaji kóddal ellátott) szóalakokat jelzik.

Az adatok azt mutatják, hogy az elemző a legtöbb hibát webcímek és egyéb kiszűrhető elemek miatt ejtette, a következő leggyakoribb a tokenizálással (szavak egybe- és különírása és egyéb szóközhiány), majd az ékezetekkel kapcsolatos hibák. Mint várható volt, az ismeretlen, de létező szavak (a diagramon *ismeretlen*, *idegen*, *tulajdonnév*, *rövidítések, kontextus* címszavak alatt) miatt történő hibák is jelentős számúak, valamint az elírás és a betűhalmozás is gyakori jelenség. A hibák természetesen halmozottan is előfordulhattak, az összetett hibákat a megfelelő hibakategóriákba külön-külön soroltuk be.



1. ábra: Morfológiai hibatípusok gyakorisága.

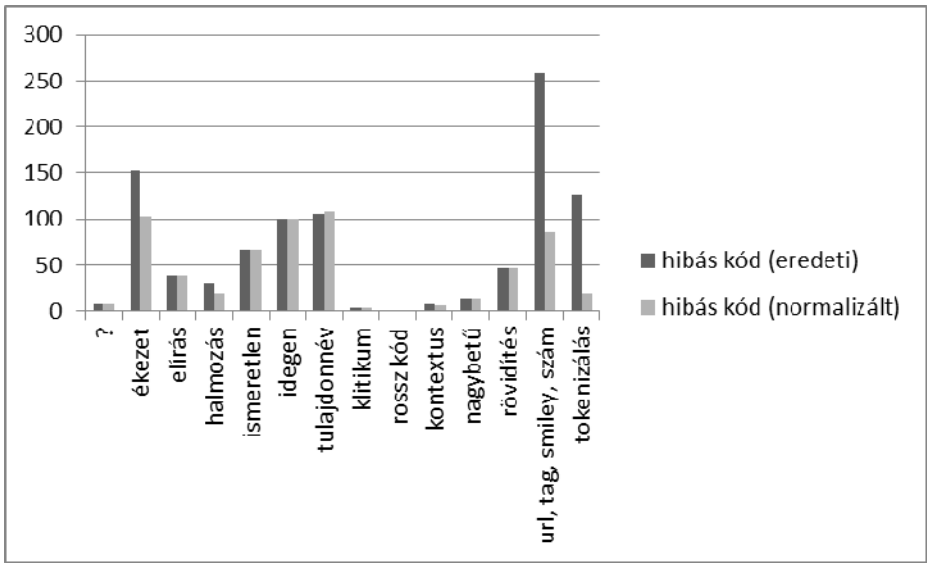
Látszik tehát, hogy a fentebb említett jelenségek a tokenizálásban és az automatikus morfológiai egyértelműsítésben problémát jelentenek, az elemző a számára ismeretlen szavakat nem tudja kiértékelni, vagy helytelen kódot ad. A kutatás egyelőre a morfológiára koncentrált, a NER tulajdonnév-felismerő [5] és a szintaktikai elemző eredményének kiértékelése folyamatban van. Annyi már látható, hogy a morfológiai hibák ezekre is hatással voltak: a helyes szintaktikai elemzéshez nélkülözhetetlen a pontos morfológiai egyértelműsítés, ami nem teljesül; a névelem-felismerő nem tudja kezelni a tiszta kisbetűvel írt neveket, a nagybetűvel írtakat – amelyeket nem látott a tanító adatbázison (pl. *Kedves Barátaim*) – pedig sokszor automatikusan névelemnek könyveli el.

3 Megoldások

A felmerült problémákat több oldalról is meg lehet közelíteni. Elméleti szempontból a hibák két csoportra oszthatók: amelyek benne vannak a tanulókorpuszban, de az elemző más alakban találkozik vele a szövegben; és amelyek semmilyen formában sincsenek a korpuszban. Az előbbire a forrásszöveg szabályalapú normalizálása (standard szöveghez hasonló formájúvá alakítása), utóbbiak nagy részére a szótár bővítése kínálhat megoldást.

Első lépésben a mondatra és tagmondatokra szegmentálást segítő, csere alapú szabályokkal (emotikonok és hiperhivatkozások egységes kezelése, szóköz és központozás helyzetének rögzítése) javítottuk a tokenizálás eredményeit. A legnagyobb problémát egyértelműen az ékezetek használata jelenti, a többi szabály elsődlegesen erre a problémakörre irányul. Az idegen ékezetek magyarra cserélése mellett toldalékokra

vonatkozó, nyelvészeti jellegű cseréket állítottunk fel (*-ság, -szerű, -ő* stb), illetve gyakori szótövek ékezetesítése (*és, csinál, tehát*, stb.). A másik normalizálási kísérlet a betűhalmozásokra irányult, ugyanis a magyarban kettőnél több azonos betű nem fordulhat elő egymást követően. A szabályok alkalmazása utáni elemzési eredmények a 2. ábrán találhatók.



2. ábra: Morfológiai hibatípusok gyakorisága a normalizálási lépések után.

Mint látható az ábrán, a kiszűrhető elemek (webcím, emotikon stb.) okozta kódolási hibák nagy része az egységes kezelés segítségével eltűnt, mint ahogy a tokenizálással kapcsolatos hibák is. A toldalék- és tőalapú ékezetesítés nem hozott akkora eredményt, azonban egy helyesírás-elemző ezzel együtt várhatóan jobb eredményt fog mutatni, mint ahogy a betűhalmozási problémák esetén is.

A szótár bővítése főként az emotikonokra, magyar és angol rövidítésekre és gyakori szavakra nyújthat megoldást, ez a munkafázis jelenleg is folyamatban van.

4 Összegzés

A közösségimédia-szövegekből kinyerhető információ egyre nagyobb jelentőségű lesz, ezek elemzése azonban – zajosságuk miatt – nem egyszerű, a standard szövegen tanult elemzők nagy hibaszázalékkal futnak le. Kutatásunk a közösségimédia-szövegekkel kapcsolatos elemzési problémák feltérképezését tűzte ki célul, számba vettük a morfológiai hibalehetőségeket és lehetséges megoldási módjukat. A kutatás jelenlegi eredményei már megkönnyíthetik egy helyesírás-elemző munkáját, ami a szöveg standardizálásának szempontjából jelentős eredményt hozhat.

Köszönetnyilvánítás

A kutatás a futurICT.hu nevű, TÁMOP-4.2.2.C-11/1/KONV-2012-0013 azonosítószámú projekt keretében az Európai Unió és az Európai Szociális Alap társfinanszírozása mellett valósult meg.

Hivatkozások

1. Csendes, D., Csirik, J., Gyimóthy, T., Kocsor, A.: The Szeged Treebank. In: Proceedings of the Eighth International Conference on Text, Speech and Dialogue (TSD 2005). Karlovy Vary, Czech Republic 12-16 September, and LNAI series Vol. 3658 (2005) 123–131
2. Khan, M., Dickinson, M.: Does Size Matter? Text and Grammar Revision for Parsing Social Media Data. In: Proceedings of the Workshop on Language Analysis in Social Media (2013) 1–10
3. Liu, Fei, Weng, Fuliang, Jiang, Xiao: A Broad-Coverage Normalization System for Social Media Language. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (2012) 1035–1044
4. Mott, Justin, Bies, Ann, Laury, John, Warner, Colin: Bracketing Webtext: An Addendum to Penn Treebank II. Guidelines. URL (2013. 11. 25.) = http://catalog.ldc.upenn.edu/docs/LDC2012T13/WebtextTBAnnotationGuidelines.pdf
5. Szarvas, Gy., Farkas, R., Kocsor, A.: A Multilingual Named Entity Recognition System Using Boosting and C4.5 Decision Tree Learning Algorithms. In: Discovery Science (2006) 267–278
6. Zsibrita, J., Vincze, V., Farkas, R.: magyarlanc: A Toolkit for Morphological and Dependency Parsing of Hungarian. In: Proceedings of RANLP-2013. Hissar, Bulgaria (2013) 763–771