Ismeretlen szavak helyes kezelése kötegelt helyesírás-ellenőrző programmal

Indig Balázs[1], Prószéky Gábor[1,2]

[1]Pázmány Péter Katolikus Egyetem, Információs Technológiai Kar,
MTA-PPKE Magyar Nyelvtechnológiai Kutatócsoport
1083 Budapest, Práter u. 50/a
indba@digitus.itk.ppke.hu, proszeky@itk.ppke.hu
[2]MorphoLogic, 1122 Budapest, Ráth György u. 36.
proszeky@morphologic.hu

Kivonat Napjainkban a közigazgatástól a könyvkiadásig jelentős szerepe van az összefüggő nagy terjedelmű szövegeknek. Ezek helyesírását meglehetősen nehéz és időigényes ellenőrizni, mert a szöveg vagy speciális tudást igényel egy adott szakterületről, vagy a nagy mennyiség miatt a javításra szánt idő válik jelentőssé. A gyakorlatban működő helyesírás-ellenőrző programok csak a szavak szintjére koncentrálnak, és legfeljebb egy-egy elgépelésre tudják figyelmeztetni a felhasználót, míg a létező, de a program számára ismeretlen, új szavakat, tulajdonneveket tévesen hibásnak jelölik. A cikkben bemutatásra kerülő módszer a nagyobb összefüggő szövegekben rejlő statisztikai sajátosságokra építve egy olyan helyesírás-ellenőrző módszert mutat be, amelynek célja a szövegösszefüggésekből kinyerhető ismeretlen, új, ám helyes szavak minél teljesebb felismerése, ragozási paradigmáik megsejtése, majd ezen szavak esetleges elgépeléseinek a javítása. A bemutatandó módszer lehetővé teszi hosszabb szövegek, például könyvek, intézményi dokumentumok egy lépésben történő gyors helyesírás-ellenőrzését.

1. Bevezetés

Az internet gyors átalakulásával és a számítógépek fejlődésével egyre szélesebb körben lehetővé válik, hogy mind nagyobb terjedelmű szövegeket hozzanak létre a felhasználók, és párhuzamosan elvárják, hogy a helyesírás-ellenőrző programok lépést tudjanak velük tartani. Ez nem kivitelezhető a hetvenes évek óta alig változó, szóról szóra haladó helyesírás-ellenőrző módszerekkel. Naponta új szavak, tulajdonnevek jelennek meg és keverednek a hagyományos szövegekkel, szófordulatokkal. Egyre több speciális területen rögzítik a szakszövegeket számítógépre, ahol egy általános helyesírás-ellenőrzőnek nincs lehetősége a szakterület speciális szavait ismerni, viszont az elgépelések esélye ugyanúgy fennáll.

Angol nyelven, ahol nincsenek túlsúlyban a ragozott szóalakok, a probléma kevésbé jelenik meg, viszont az erősen ragozó nyelvekben, mint a magyar, ez

határozottabban előkerül, ugyanis nemcsak az egyes új, helyesírás-ellenőrző környezetek által nem ismert szavakat „kellene" felismerni és javítani, hanem egyúttal ezek helyesen ragozott alakjait is. Bár az ismeretlen szavakról a gép jelenleg nem tudja eldönteni, hogy helyesek-e, egységesíteni tudja az írásmódjukat a statisztikailag releváns találatok alapján, illetve képes egy menetben csoportosítani és így egyszerre javítani vagy jóváhagyni több előforduló szóalakot a felhasználó kényelme érdekében. A módszer erősen támaszkodik arra, hogy egy szó jó alakja statisztikailag számottevően gyakoribb, mint az elgépelés. Természetesen ez a módszer a következetes helytelen írásmódot nem képes javítani.

Az alábbiakban ezen folyamat részleteit ismertetjük. Mi az általunk korábban kifejlesztett eszközöket használtuk, de a megoldás általánosabb, ezért a későbbiekben időnként tokenizálóként fogunk hivatkozni a PureTokenre [6], POS-taggerként fogunk hivatkozni a PurePOS-ra [3], és morfológiaként a Humorra [2].

2. A módszer

Az összefüggő szövegeknek sajátossága, hogy a bennük előforduló szavak a Zipf-törvény szerinti eloszlással rendelkeznek [5]. Megfelelő méretű összefüggő szövegeket választva a statisztika mind jobban előtérbe tolódik, a nyelvspecifikus ismeretek mellé. Ahogy az Kornai és társai cikkében [7] is szerepel, az internetről is legyűjthetők ilyen szövegek, amelyekből statisztikai jellemzők kinyerhetőek későbbi felhasználásra.

2.1. A statisztikai jellemzők kinyerése és felhasználása

Ezen jellemzők kinyeréséhez a rendelkezésre álló nyelvtechnológiai eszközök mindegyikét végigfuttatjuk a szövegen, és a mondatokra és tokenekre bontott szöveg szavaihoz szófaji címkéket és szótöveket rendelünk, majd egy hagyományos helyesírás-ellenőrzővel megjelöljük azokat a szavakat, amelyek ismeretlenek. Az így létrejött annotált szövegben – immár csak az ismeretlen szavakat tekintve – statisztikai sajátosságokat keresünk, amelyek segítségünkre lehetnek a szavak osztályozásában, illetve ajánlatgenerálásban. Ilyen jellemzők például:

- az egyes szóalakok gyakoriságai
- az ismeretlen szavak (POS által meghatározott) szótöveinek gyakoriságai
- a fentiek kombinációja.

A szótövek szerint csoportosított szóalakokból a magyar nyelv ragozási jellemzőinek és ezek összefüggéseinek ismeretében – amit a morfológia tartalmaz a beépített szótárban szereplő szavak esetén – kellő számú és minőségű különböző ragozott alak megléte esetén megállapítható egy ragozási paradigma, amire vizsgálhatóak a kevésbé gyakori szóalakok, így eldöntve, hogy ragozásuk egységes-e vagy sem, ezzel felismerve a helytelenül ragozott, esetleg elgépelt szóalakokat. Az így szerzett információval lehet felismerni és javítani a csak különféle elgépelt formában előforduló változatokat is, melyeket a hagyományos helyesírás-ellenőrzők a többi helytelen szóval egyetemben egységesen hibásnak jelölnek, további

elemzés nélkül. Egy másik probléma az ismeretlen, de elgépelt szavakhoz megfelelő ajánlások generálása, amit a fenti módon gyűjtött információkkal orvosoltunk.

Az ismeretlen szavak osztályát tovább bontva egy-egy szóalakot, illetve szótövet a gyakorisága alapján tekinthetünk „biztosan jónak" vagy pedig „ritkának"[1]. A „biztosan jó" szóalakokból és a gyakori szótövekből végezzük el a csoportosítást és a ragozási paradigma meghatározását. Ezek a szóalakok és a belőlük nyert információk segítenek a ritka szóalakokhoz ajánlások generálásában[2].

A hagyományos helyesírás-ellenőrzők így átalakíthatóak, hogy a megadott szavak és szótövek alapján paradigmát építve újraellenőrizzék az ismeretlennek jelölt szavakat, és szükség szerint ajánlásokat generáljanak hozzájuk a meglévő belső működés felhasználásával. Ezzel megbízható módon és teljesen automatikusan lehet bővíteni a helyesírás-ellenőrző és a morfológia szótárát. Emellett a felhasználó visszajelzést tud küldeni a fejlesztőknek, vagy egy központi adatbázisban gyűjtheti a kollaboratív munka eredményeit egy helyesírás-ellenőrző esetleges doménspecifikus tudásának felépítéséhez.

Az így kapott, osztályozott, javítási javaslatokkal ellátott szavak minden előfordulását a felhasználó könnyen, a teljes dokumentum átolvasása nélkül, mindössze a kritikus szövegkörnyezetre rápillantva, egy menetben kezelve képes javítani. A nyers szöveg mondatokra és tokenekre bontása közben ugyan elveszíti az eredeti formázást, de például dinamikus idővetemítéssel (DTW)[8] meghatározhatóak a szoros összefüggések (horgonyok) az eredeti szöveggel, arra az esetre, ha a javításokat nem szóalakonként egységesen, hanem a javítandó szavak környezetének függvényében kívánjuk elvégezni. Tipikusak az alábbi többértelműségek:

- *román*: a nemzetiség (román[MN][NOM]), a roma emberen (roma[FN][SUP])
- *rendben*: benne a rendben (rend[FN][INE]), rendben van (rendben[HA])
- *alma*: az állat alma (alom[FN][PSe3][NOM]), almafa (alma[FN][NOM])
- továbbá minden olyan toldaléksorra végződő alak, amelyek összetett szó utótagjaként is megjelenhet, például: *-ének: gyerekének, -ében: fejében, -ára: tanára, -inak: tanulóinak* [9]

2.2. A POS-tagger adaptálása a szöveghez a posteriori információkkal

A tokenizált szöveget a POS-taggernek átadva, az egyértelműen meghatározza a szavakhoz a lehetséges lemmákat a beépített morfológia segítségével.[3] Az ismert szavak esetén csak a néhány felkínált alternatíva közül kell választani a simított

[1] A gyakori, ugyanolyan módon történő elgépelést következetes hibának vesszük, és nem tudunk különbséget tenni következetes hibák szándékosságát illetően.

[2] Jelen mérésben csak egyszerű Damerau–Levenshtein távolságot [10] alkalmaztunk az ajánlások kereséséhez, de ez bővíthető több megszokott módszerrel is.

[3] Itt azt feltételeztük, hogy a helyesírás-ellenőrző nem szólista alapú, hanem morfológiát használ.

n-gram modell alapján. Ezzel szemben az ismeretlen szavak esetén a szótő és a szófaji címke meghatározása nem ilyen egyszerű: ekkor az ismeretlen szavakat egy ismeretlen szavakat elemezni képes modul, az ún. guesser megpróbálja megelemezni pusztán a beleépített nyelvi tudásra hagyatkozva. Az így kapott rengeteg elemzés közül kell kiválasztania a megfelelőt az egyértelműsítőnek, amely csak a lokális, n-gram modellt, illetve a mondatszintű beam search megoldást veszi figyelembe [3]. Más szóval: nem használja ki a nagy terjedelmű összefüggő szövegekben rejlő globális információkat. A POS-tagger hatékonyságának javítására olyan módszert dolgoztunk ki, amely a feldolgozott szöveg a posteriori információi alapján támogatja a feldolgozást: a szöveg feldolgozása közben a guesser által az egyes szavakhoz generált lehetséges lemmák közül a szóhoz tartozó címkének megfelelőkből mindig a globálisan leggyakoribbat választjuk. Ezzel előállítunk egy, a lemmák gyakorisága szerint rendezett listát, amelyből a megfelelően választott előfordulási küszöb fölötti, így gyakori szótöveket beadhatjuk a programnak listaként, hogy válassza ki azokat a lemma–címke párokat, amelyeknél a szótő szerepel a listán, ha van ilyen. Ezzel redukálja a lehetőségek számát, majd az így leszűkített halmazból kiválasztja a végleges verziót. Az eljárástól azt várjuk, hogy az egy szótőre visszavezetett ismeretlen szavak száma nő, ezzel pedig a helyes szótövek száma az ismeretlen szóalakok egészét tekintve arányosan javul.

3. Eredmények

A módszer hatékonyságát egy elméletileg csak helyes szavakat tartalmazó regényen (Orwell: 1984) és az internetről legyűjtött újságcikkekből és cikksorozatokból álló hasonló méretű korpuszon vizsgáltuk, a Szeged 2.0 korpuszt [4] használva nyelvi modellként. Az ellenőrzés során egy egyszerű heurisztikával szűrést végeztünk. Az eredetileg kapott adatokat az 1. táblázatban sz.e., a szűrés utániakat sz.u. jelzi. A szűréssel a statisztikából kivettük az egyértelműen önálló toldalékként azonosítható szavakat (pl. „-nak”) és az olyan szavakat, amelyek nem tartalmaztak legalább négy egymás melletti betűt (pl. „TU-154”, „MiG-24”). Ezáltal azt reméljük, hogy az „igazi” szavak és elgépeléseik jobban előtérbe kerülnek.

1. táblázat. A korpuszok adatai.

	1984		*Újságcikkek*	
	sz.e.	*sz.u.*	*sz.e.*	*sz.u.*
Tokenek:	99913	50586	74053	40716
Tokenek (egyedi):	20393	18211	20916	18465
Szegedben nem szereplő:	1149	1058	10001	8965
Szegedben nem szereplő (egyedi):	956	881	8321	7582
Humorban nem szereplő:	301	283	1431	1224
Humorban nem szereplő (egyedi):	181	168	1029	886
Humorban és Szegedben sem szereplő:	217	199	1362	1166
Humorban és Szegedben sem szereplő (egyedi):	129	116	992	859

2. táblázat. Példa a szavak gyakoriságára.

szó	**gyakoriság**	**szótő**
Obama	40	Obama
Obamaáról	1	Obamaá
Obamáék	1	Obamá
Obama-kormány	1	Obama-kormány
Obamának	3	Obam
Obamának	3	Obamá
Obamára	1	Obamá
Obamáról	3	Obam
Obamáról	3	Obamá
Obamát	5	Obam
Obamát	5	Obamát
Obamával	1	Obamával

A 2. táblázatban látható, hogy a globális információ nélküli program nem tudta megtalálni a kapcsolatot a különböző szóalakok között. Az elgépelés belesimul a helyes, ismeretlen alakokba. A szöveg méretétől függően érdemes beállítani a gyakorisági küszöböt, amitől egy szótő, illetve szóalak helyesnek számít. Mi a mérés során az alábbi paramétereket választottuk: szógyakoriság ≥ 2, tőgyakoriság ≥ 3 és Damerau–Levenshtein távolság $= 1$.

3. táblázat. Eredmények.

	1984	*Újságcikkek*
Szótőváltozás:	34	65
Szótőváltozás (egyedi):	19	48
Gyakori lemmák száma:	14	55
Gyakori szóalakok száma:	40	51
Paradigmák száma:	17	58
Ajánlások száma:	3	8

4. táblázat. Jó paradigmák.

1984	
szótő	
beszélír	
jó szóalakok	**ritka szóalakok**
beszélírba	beszélírja
beszélírral	beszélírtól
beszélír	
beszélírt	

Újságcikkek	
szótő	
Obama	
jó szóalakok	**ritka szóalakok**
Obamának	Obamáék
Obamáról	Obamára
Obamát	Obamával
Obama	

A ragozási paradigmák, amelyek a 4. táblázatban is láthatóak, akkor tekinthetőek jónak, ha megfelelő számú és minőségű olyan szóalakot találunk, amelyek alkalmasak az egyértelmű osztályozásra, így a bizonytalan, ritkább alakok ellenőrzésére. Rossz egy paradigma, ha a szótő sok ritka szóalak csoportosításaként, illetve ha túl kevés szóalak gyakori előfordulása miatt lett gyakori. Ez utóbbiak is természetes módon előfordulnak a szövegben. Az ajánlások a jónak tekintett szavak alapján történtek (5. táblázat).

5. táblázat. Ajánlások

Újságcikkek	
hibás szóalak	**ajánlás**
BruxInfo	Bruxinfo
Gingrics	Gingrich
Mtelekom	MTelekom
Obamaáról	Obamáról
Osama	Obama
Sandber	Sandberg
stent	sztent
Unicredit	UniCredit

1984	
hibás szóalak	**ajánlás**
aszondom	Aszondom
beszélírja	beszélírba
jógondoló	jógondol

Látszik, hogy érdemes egy már meglévő helyesírás-ellenőrző program motorját használni, mert különben a primitív algoritmusnak köszönhetően olyan hamis ajánlások is születhetnek, amelyek elkerülhetők lennének.

A vizsgált korpuszokon a hagyományos helyesírás-ellenőrző programok által helytelenül hibásnak jelzett szavak aránya csökkent, és néhány esetben sikerült a hibásan gépelt ismeretlen szavakat helyesre javítani, minimális zajarány mellett.

4. További kutatási irányok

A módszer jelen pillanatban önmagában még nem alkalmas automatikus helyesírás-ellenőrzésre, de a kutatásnak ez a kezdeti fázisa azt mutatja, hogy az új módszer használatával a teljes ellenőrzési folyamat a szöveg méretének növelésével egyszerűbbé és gyorsabbá válik.

Az újfajta helyesírási hibák ember által felügyelt javításával pedig már most is kielégítő eredményt kapunk, a felhasználó pedig az összefüggő szövegek javítását gyorsabban, kényelmesebben és pontosabban tudja végezni. További kutatásainkban a módszer alábbi alkalmazási lehetőségeit vizsgáljuk:

- a helyesírás-ellenőrző tudásának bővítése hatékonyan;
- egy erre a célra hasznos elgépelésszótár automatikus építése;
- felhasználók közötti kollaboráció a helyesírás-ellenőrzésben és javításban megosztott lexikonnal;
- mindezek által gyors doménadaptáció elérése.

A felsorolt folyamatok jelenleg meglehetősen emberigényesek, de a javasolt módszer az egységnyi idő alatt feldolgozható szöveg mennyiségét egyértelműen növeli.

Köszönetnyilvánítás

Köszönjük a TÁMOP-4.2.1.B – 11/2/KMR-2011–0002 projekt részleges támogatását.

Hivatkozások

1. Miháczi A., Németh L., Rácz M.: Magyar szövegek természetes nyelvi feldolgozása. In: Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2003). SZTE, Szeged (2003) 38–43
2. Prószéky, G., Novák, A.: Computational Morphologies for Small Uralic Languages. In: Inquiries into Words, Constraints and Contexts. Stanford, California (2005) 150–157
3. Novák A., Orosz Gy., Indig B.: Javában taggelünk. In: Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2011). SZTE, Szeged (2011) 336–340
4. Csendes D., Hatvani Cs., Alexin Z., Csirik J., Gyimóthy T., Prószéky G., Váradi T.: Kézzel annotált magyar nyelvi korpusz: a Szeged Korpusz. Magyar szövegek természetes nyelvi feldolgozása. In: Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2003). SZTE, Szeged (2003) 238–247
5. Zipf, G.: Selective Studies and the Principle of Relative Frequency in Language. Cambridge, Mass (1932)
6. Indig B.: PureToken: egy új tokenizáló eszköz. In: Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2013). SZTE, Szeged (2013) 305–309

7. Kornai, A., Halácsy, P., Nagy, V., Oravecz, Cs., Trón, V., Varga, D.: Web-based frequency dictionaries for medium density languages. In: Proceedings of the 2nd International Workshop on Web as Corpus (WAC '06). Association for Computational Linguistics, Stroudsburg, PA, USA (2006) 1–8
8. Bellman, R., Kalaba, R.: On adaptive control processes. IRE Transactions on Automatic Control, Vol. 4, No. 2 (1959) 1–9
9. Novák A., M. Pintér T.: Milyen a még jobb Humor. In: IV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2006). SZTE, Szeged (2006) 60–69
10. Damerau, F. J.: A technique for computer detection and correction of spelling errors. Commun. ACM, Vol. 7, No. 3 (1964) 171–176