Ékezetek automatikus helyreállítása magyar nyelvű szövegekben

Novák Attila[1,2], Siklósi Borbála[2]

[1] MTA-PPKE Magyar Nyelvtechnológiai Kutatócsoport ,
[2] Pázmány Péter Katolikus Egyetem, Információs Technológiai és Bionikai Kar
1083 Budapest, Práter utca 50/a
e-mail:{novak.attila, siklosi.borbala}@itk.ppke.hu

Kivonat Cikkünkben egy olyan rendszert mutatunk be, ami a statisztikai gépi fordítás módszereit használva megbízhatóan pótolja a hiányzó ékezeteket ékezetek nélkül írt magyar nyelvű szövegekben. Mivel magyar nyelv esetén elkerülhetetlen, hogy igen nagyméretű szöveges tanítókorpusz alkalmazása esetén is hiányozzanak bizonyos szóalakok a tanítóanyagból, morfológiai elemzőt integráltunk a rendszerbe, ami ékezetesített szóalakjelölteket generál ezekhez a szavakhoz. Az így létrejött rendszert kiértékelve a rendszer az esetek több mint 99%-ában helyes ékezetes alakot állított elő.

1. Bevezetés

Napjaink népszerű kommunikációs fórumai a közösségi oldalak. Az ezek használata során leírt szövegek létrehozása egyre inkább olyan mobil eszközökhöz kötődik, melyek szöveges beviteli felülete a magyar nyelv ékezetes karaktereinek elérésére nem nyújt kényelmes és gyors hozzáférést. Ezért egyre több olyan szöveg jön létre, ami nem tartalmaz ékezeteket. Az ebből adódó többértelműségek feloldása az emberek számára ritkán okoz problémát, ahhoz azonban, hogy ezek a szövegek a szokványos nyelvtechnológiai eszközökkel feldolgozhatóak, elemezhetőek legyenek, szükség van az ékezetek visszaállítására. Cikkünkben egy olyan magyar ékezetesítőrendszert mutatunk be, ami egy statisztikai gépi fordító (SMT) keretrendszer és egy morfológiai elemző kombinációjából áll.

Ugyan léteznek más ékezetesítőrendszerek magyar nyelvre, azonban azok vagy nem elérhetőek, vagy rosszabbul teljesítenek. Jellemző továbbá a probléma karakteralapú megközelítése, azonban az ilyen rendszereknél óhatatlanul megjelennek értelmetlen szóalakok. Ezzel szemben, a szótáralapú megoldások a szótárban nem szereplő szavak ékezetesítésére nem tudnak javaslatot tenni.

2. Kapcsolódó munkák

Több kutatás célozta már meg az ékezetek helyreállításának megoldását magyar nyelvű szövegek esetén. [11] és [3] gépi tanulási módszereket alkalmaztak, ahol a

beszúrandó ékezetek pozícióját az ékezet nélküli betű közvetlen környezete alapján határozzák meg. Ezzel a módszerrel 95%-os pontosságot értek el. A módszer előnye, hogy a tanítóanyagban nem szereplő, ismeretlen szavakat is kezelni tudja, hátránya viszont, hogy nem létező szóalakokat is generál. A feladat egy másik megközelítése szótár használatán alapul. Ezek a módszerek nagy szöveges korpuszból becslik meg a különböző ékezetes alakok disztribúcióját. [11] ezzel a módszerrel 98%-os pontosságról számol be. Ez a rendszer viszont nem tudja kezelni az ismeretlen szavakat. [4] egy többszintű nyelvfeldolgozó rendszert mutat be, amit egy text-to-speech alkalmazáshoz hoztak létre. Ennek keretein belül az ékezetek helyreállításához morfológiai és szintaktikai elemzést is végeznek, így az ékezetesítés pontossága erősen függ az elemzők teljesítményétől (95%-os pontosságot sikerült elérniük).

A Charlifter [8] egy nyelvfüggetlen ékezetesítőrendszer, ami lexikonalapú statisztikai módszereket alkalmaz, illetve egy bigram környezeti modellt és az ismeretlen szavak kezelésére egy karakteralapú statisztikai modellt is használ. A rendszert kipróbáltuk magyarra. Ennek teljesítményét alább a saját rendszerünkével összevetve részletezzük.

Más nyelvekre is hasonló módszereket találunk. [10] átfogó elemzést mutat be francia és spanyol szövegek ékezetesítésére adott megoldásokról. Az esettanulmány a szövegkörnyezet jelentőségét hangsúlyozza, de mind a különböző szóalakok, mind az ékezetek száma jóval kevesebb ezekben a nyelvekben, mint a magyarban. [12] szintén francia nyelvre ad megoldást, azonban kifejezetten orvosi szakszövegekkel, szavakkal foglalkozik, aminek jellegzetessége az ismeretlen szavak magas aránya az általános nyelvhasználathoz képest. A módszer címkézési feladatként fogalmazza meg a problémát, amit transzducerekkel oldanak meg. A tesztek során 92%-os pontosságot értek el egy orvosi tezaurusz címszavain mérve, szövegkörnyezet nélkül.

A saját módszerünkhöz leginkább [6] módszere hasonlít. Ebben a kutatásban szintén gépifordító-rendszert alkalmaztak vietnami szövegek ékezetesítésére, 93%-os pontosságot érve el. Ez a rendszer azonban egy külső szótárt is használ, továbbá a vietnami[3] és a magyar nyelv sajátosságai közötti különbségek miatt az eredmények nem összemérhetőek.

3. Ékezetek helyreállítása

Az ékezetek helyreállításának problémáját fordítási feladatként fogalmaztuk meg, ahol a forrásnyelv az ékezet nélküli szöveg, a célnyelv pedig az ékezetes változat. Mivel ebben az esetben nagyon könnyű nagyméretű párhuzamos korpuszt létrehozni (hiszen egy egynyelvű korpuszból csak el kell távolítani az ékezeteket),

[3] A tonális vietnami nyelvben különböző mellékjeleket használnak egyes magánhangzófonémák megkülönböztetésére (négy különböző mellékjel) és a szóalakokban szereplő szótagok tónusának jelölésére (öt különböző mellékjel). A vietnamiban több az ékezet, mint a magyarban. Gyakran egy magánhangzót jelölő betűn két különböző mellékjel is megjelenik. Ugyanakkor a vietnami izoláló nyelv, ezért a produktív magyar morfológiából adódó rengeteg különböző szóalak a vietnamira nem jellemző.

magától értetődőnek tűnt a statisztikai gépi fordító (SMT) rendszer alkalmazása, melyben a fordítási modell az egyes frázisok lehetséges ékezetes változatainak eloszlását tartalmazza, a nyelvmodell pedig a szövegkörnyezetet képviselve az aktuális környezetben helyes alak kiválasztását biztosítja. A rendszer magjaként a Moses [2] keretrendszert használtuk a fordítási modell építéséhez és a dekódoláshoz, a nyelvmodellt pedig a SRILM [9] eszközzel hoztuk létre. A Moses használata során annak alapértelmezett konfigurációs beállításait használtuk a szóösszerendelő lépés kihagyásával, amire ebben a feladatban nem volt szükség, hiszen minden forrásoldali szó egyértelműen megfeleltethető a céloldali párjának. Ugyanez indokolta azt is, hogy a dekódolás során csak monoton fordítást engedélyeztünk, azaz a szórendnek változatlannak kellett maradnia.

3.1. Az alaprendszer

Az alaprendszerben csak a tanítóanyagból épített fordítási- és nyelvmodelleket használtuk. A dekóder bemenete a hiányzó ékezeteket tartalmazó magyar szöveg. A fordítási modell csupán unigramokat tartalmazott ebben a felállásban (magasabb rendű n-gramokkal is kísérleteztünk, ez azonban az eredményre nem volt hatással), a nyelvmodellben pedig legfeljebb 5 szó méretű frázisok szerepeltek. Így a fordítási modell meghatározta az egyes szavakhoz tartozó ékezetes alakok disztribúcióját, míg a nyelvmodell a szövegkörnyezetet képviselve választja ki a megfelelő alakot. Az így létrehozott baseline rendszer hiányossága azonban, hogy a tanítókorpuszban nem szereplő ismeretlen (OOV) szavakat egyáltalán nem kezeli.

Egy másik alaprendszert is létrehoztunk az SMT rendszer hatásának vizsgálatára. Ebben a rendszerben minden ékezet nélküli szót mindig a leggyakoribb ékezetes alakjára cseréltük a szövegkörnyezet figyelembevétele nélkül. Ehhez a gyakorisági adatokat a tanítókorpuszból határoztuk meg.

3.2. Morfológiai elemző integrálása

A korpuszban nem szereplő ismeretlen szavak kezelésére a Humor morfológiai elemzőt [5,7] integráltuk a rendszerbe. Az eredeti elemzőnek egy olyan módosított változatát hoztuk létre és használtuk ebben a feladatban, ami az ékezet nélküli szóalakokat közvetlenül leképezi a lehetséges ékezetes változataikra. Továbbá, a morfémahatárokat is jelöli és meghatározza a morfoszintaktikai kategóriacímkét a kapott szóalakokhoz. A szegmentálásra vonatkozó jelölések (pl. szóösszetételi határok, képzők) és a kategóriacímkék az ékezetes alakok rangsorolásához használt pontszám számításakor szükségesek. Az ékezetes szóalakokat újraelemezzük, hogy a közvetlenül nem kinyerhető lemmákat is megkapjuk. A kísérleteinknél használt, 1 804 252 token méretű tesztanyagban a szavak kb. 1%-a nem volt benne a fordítási modellben a legnagyobb, 440 millió token méretű, tanítóanyag esetén sem. Az 1. táblázatban látható az ismeretlen szavak (OOV) aránya a fordítási modell létrehozásához használt különböző méretű tanítóanyagok esetén.

1. táblázat. Az ismeretlen (OOV) szavak aránya a különböző méretű tanítóanyagból épített fordítási modellek esetén.

tanítóanyag	**mondatok száma**	**millió szó**	**OOV a tesztanyagban**
100K	100 000	1,738	9,63%
1000K	1 000 000	18,078	3,44%
5000K	5 000 000	89,907	1,23%
10M	10 000 000	180,644	1,68%
ALL	24 048 302	437,559	0,81%

A tesztanyagban előforduló ismeretlen szavak esetén az elemző ékezetesített szóalakokat javasol. Ezeket a javasolt szóalakokat a Moses rendszer esetén azoknak a fordítandó szövegbe való beágyazásával továbbítani tudjuk a fordítórendszer felé. Ehhez azonban minden egyes javasolt szóalakhoz valószínűséget kell rendelni. Először egyenletes eloszlást feltételeztünk, így azonban a gyakori ékezetes alakok és a gyakorlatilag értelmetlen (bár nyelvtanilag helyes) alakok is azonos valószínűséggel szerepeltek. Hogy ezek a szóalakok ne jelenjenek meg az eredményben, a második változatban kifinomultabb algoritmust alkalmaztunk a valószínűségek becslésére.

Az ékezetesített javaslatokhoz egy-egy pontszámot rendelünk, amely alapján rangsorolhatók a kapott szóalakok. Mivel maga a szóalak nincs benne a korpuszban, ezért a pontszám a következő tényezők lineáris kombinációjaként jön létre: (1) lemmagyakoriság (LEM), (2) a szóalakban megjelenő ragsorozat gyakorisága (INF), (3) a szóalakban előforduló produktív összetételek (CMP), és (4) produktív képzők száma (DER). Az első két tényezőt a tanítókorpuszból számított statisztika alapján határoztuk meg. A modell használatakor ezek a tényezők pozitív súlyozást kaptak, előnyben részesítve ezzel a gyakori lemmákat, illetve gyakori toldalékkombinációkat. A második két tényező ezzel szemben negatív súlyozást kapott, csökkentve ezzel a többszörös összetételeket és képzőket tartalmazó jelöltek pontszámát. Az egyes ékezetes jelöltekhez rendelt pontszámot tehát az (1) egyenlet alkalmazásával határoztuk meg.

$$score = -\lambda_c \#CMP - \lambda_d \#DER + log_{10} LEM + \lambda_i log_{10} INF + MS \ , \quad (1)$$

ahol

$$MS = \begin{cases} |minscore| + 1 \text{ ha} & minscore \leq 0 \\ 0 & \text{egyébként} \end{cases} \quad (2)$$

Az MS komponens a pontszámok felskálázása miatt került bevezetésre, a kapott pontszámhoz $|minscore| + 1$-et adva hozzá, azaz az aktuális jelöltlistában szereplő legkisebb pontszámmal növelve az összes jelölt pontszámát, ezzel védve ki a negatív pontszámok megjelenését. A λ súlyokat a Moses `mert` optimalizáló programjával állítottuk be. Ehhez a korpusz egy előre elkülönített részén megvizsgáltuk a morfológiai elemző által elemzett OOV szavakban az összetételek, képzők és ragok eloszlását, majd a megfigyelt eloszlásnak megfelelően, de

véletlenszerűen kiválasztottunk 1000 szót. A `mert` optimalizálás célváltozója a rendszer pontossága volt erre az 1000 szóra, az így kapott λ súlyokat használtuk a modellben. Bár lineáris regressziót alkalmaztunk, melynek során bevett szokás egy torzító súly (bias weight) hozzáadása is, ezt nem tartottuk szükségesnek, mivel nem kellett a kapott becsléseket más forrásokból való becslésekkel szinkronba hozni. Végül normalizálással valószínűségi eloszlássá alakítottuk a kapott pontszámokat.

Bár a pontszámok megfelelő skálázásával a rangsorolt ékezetesített szóalak-jelöltek a fordítási modellben meglévő frázisokhoz hasonlóan mind felhasználhatóak lettek volna, a rendszerbe csak a legmagasabb pontszámot kapott jelöltet továbbítottuk. Ennek oka, hogy mivel a nyelvmodellben ezek a szóalakok nem szerepeltek (tehát a nyelvmodell szerint mindegyiknek azonos a valószínűsége), ezért mindenképp a legnagyobb valószínűségű szóalakot választja a rendszer az egyes ékezetlen szóalakokhoz generált jelöltek listájából.

4. Eredmények

A rendszer tanításához és teszteléséhez a magyar webkorpuszt használtuk [1]. Ebből 100 000 mondatot (1 804 252 token) tettünk félre teszteléshez, másik 100 000 mondatot optimalizáláshoz, a többit pedig több különböző felbontásban a tanításhoz használtuk. A legkisebb tanítóanyag 100 000 mondatból állt, a legnagyobb pedig több mint 24 millió mondatot tartalmazott. Az egyes tanítóanyagok méretét az 1. táblázatban foglaltuk össze. A kiértékelés során a tanítóanyag méretének növelése mellett vizsgáltuk a rendszer teljesítményét, illetve a morfológiai elemző integrálásának hatását.

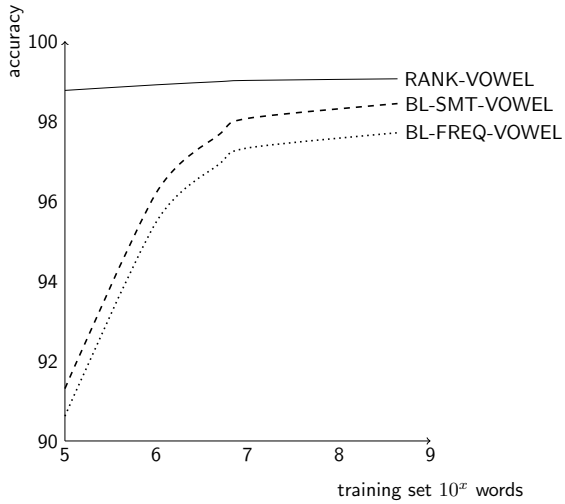
A teszthalmaz ékezetmentes, eredeti állapotában a tokenek 56,84%-a volt helyes szóalak, a magánhangzókat tartalmazó szavaknak pedig (ahol egyáltalán történhet ékezetesítés) 47,09%-a. Ebből az állapotból a morfológia nélküli SMT rendszer (BL-SMT) a legkisebb tanítóanyag esetén 91,31%-ra javította a magánhangzót tartalmazó szavak helyességének arányát, míg a teljes tanítóanyag felhasználása esetén ez az arány 98,44%-ra nőtt. A morfológiai elemző használata mellett (RANK) a pontosság 98,77%, illetve 99,06% volt a két tanítóanyag esetén. A minden szót mindig a leggyakoribb ékezetesített alakra cserélő alaprendszer pontossága (BL-FREQ) pedig 90,62%, illetve 97,71% volt. A 2. táblázatban láthatóak a részletes eredmények a legkisebb és a legnagyobb tanítóanyag használata mellett az összes szó (ALL) és a magánhangzókat tartalmazó szavak (VOWEL) esetén. Látható, hogy a rendszer pontossága a tanítóanyag növelésével csekély mértékben növekszik, viszont a fedés és a helyesség drasztikusan nő a morfológiai elemzőt nem használó rendszerek esetén.

A morfológiai elemző integrálásával azonban pótolni lehetett a kis tanítóanyagból hiányzó információt, amivel jelentősen megnőtt a fedés. Még a legnagyobb tanítóanyag esetén is 39,74%-os hibaarány-csökkentést eredményez az elemző használata, a hibás szavak arányát 1,56%ról 0,94%-ra csökkentve. A legkisebb tanítóanyag esetén a hibaarány-csökkenés 85,85%. Az elemzőt használó rendszer tehát a legkisebb tanítóanyag mellett is jobban teljesít, mint az elemző

nélküli SMT rendszer a legnagyobb tanítóanyagon tanítva. A 1. ábra az egyes rendszerek tanulási görbéjét mutatja, azaz a rendszer helyességét a tanítóanyag méretének függvényében.

2. táblázat. A pontosság alakulása az egyes rendszerparaméterek és a tanítóanyag mérete függvényében.

	100K			ALL		
rendszer	prec	rec	acc	prec	rec	acc
BL-FREQ-ALL	98,25	82,82	92,34	98,37	96,26	98,13
BL-FREQ-VOW	98,25	82,82	90,62	98,37	96,26	97,71
BL-SMT-ALL	99,03	83,88	92,91	99,09	97,36	98,72
RANK-ALL	98,81	98,08	98,99	99,01	98,56	99,23
RANK-VOW	98,82	98,08	**98,77**	99,02	98,56	**99,06**



1. ábra. A magánhangzót tartalmazó szavakon mért pontosság a tanítóanyag mérete függvényében az egyes rendszerek esetében.

Az eredményeinket összehasonlítottuk a Charlifter rendszerrel elért eredményekkel. Ennek teljesítménye 89,75% helyesség a leggyakoribb ékezetes alakok használata esetén, 90,00% a *lexicon-lookup+bigram* kontextuális modell esetén és 93,31% a *lookup+bigram context+character-n-gram* modell esetén. Az összehasonlításból látható, hogy az SMT modellben használt nyelvmodell jobban növeli a rendszer helyességét, mint a Charlifter által használt bigram kontextus modell,

a morfológiai elemzővel kiegészített SMT rendszer pedig szintén jobban teljesít, mint a karakteralapú n-gram modell.

5. Hibaelemzés

A tesztanyag egy 5000 mondatos részén részletes hibaelemzést is végeztünk. Ennek részletes eredményeit a 3. táblázatban foglaltuk össze.

A részletes elemzés során kiderült, hogy az eredeti és a rendszer által ékezetesített szöveg szavai közötti eltérés 14,7%-a nem valódi hiba. 3,55%-ban ekvivalens alakot kaptunk (pl. *lévő*∼*levő*), míg a többi a referenciában szerepelt hibásan, a rendszer által adott eredmény volt a helyes.

A referencia egy másik jelentős hányada (17,91%) szintén hibás volt, azonban ezekben az esetekben a hiba nem az ékezetek hiányából fakadt, ezért nem tudta a rendszerünk javítani. Ezek a hibák leggyakrabban hiányzó vagy felcserélt betűkből adódnak (10,81%), további 6,42% pedig valamilyen központozási hiba az eredeti referenciaszövegben.

A hibák kb. 2/3-a volt valódi hiba. Ezek 5,57%-ában a szótő ismeretlen volt a morfológiai elemző számára. Az esetek 3,55%-a olyan hiba volt, amikor a rendszer egy tulajdonnevet egy gyakoribb szóalakra alakított át: vagy egy másik tulajdonnévre, vagy csupán egy gyakori szóalakra. Hasonló hiba, amikor egy köznevet a rendszer egy gyakoribb tulajdonnévre alakít át (további 1,35%). Ezeknek a hibáknak egy részét kezelni lehetne, ha a rendszer figyelembe venné a kisbetű-nagybetű megkülönböztetést. Ez azonban más esetekben okozhatna hibát, a rendszer általános teljesítménye feltehetőleg romlana az adathiány miatt.

A hibák 2,20%-a a tanítóanyagban lévő hibákból fakadt. Mivel a magyarban gyakoriak a ritka szóalakok, ezért könnyen előfordulhat, hogy egy szó többször szerepel hibásan, mint helyesen (különösképpen igaz ez a korpuszban csupán egyszer szereplő szóalakokra). A vizsgált tesztanyagban előforduló hibák további 3,72%-a abból adódott, hogy a rendszer a szándékosan ékezet nélkül írt szóalakokat (fájlnevek, url-ek) is átalakította azok ékezetes alakjára, vagy valami más, értelmes szóalakra, vagy éppen ennek ellenkezője történt, a szövegben furcsa mód ékezetesen írt url-t nem ékezetesített (pl. *www.valamicég.hu*).

A hibák legnagyobb része (51,01%) olyan eset, amikor a rendszer nem tudta a környezet alapján sem eldönteni, hogy mi lenne a helyes szóalak. Ezeknek az eseteknek több, mint fele olyan hiba, amikor a rendszer felcserélte egy birtokos és a nem birtokos alakját egy adott főnévnek (pl.*gyereket*∼*gyerekét*, *gyereken*∼*gyerekén*, *gyereke*∼*gyereké*). További 26% a igék definit és indefinit alakjának hasonló tévesztéséből fakad (pl. *hajtottak*∼*hajtották*, *hajtanak*∼*hajtanák*, *hajtana*∼*hajtaná*).

3. táblázat. Hibaelemzés a rendszerkimenet és az eredeti szöveg eltéréseinek vizsgálatával egy 5000 mondatos tesztanyagon.

Hibatípus	**Arány**	**Példák**
A rendszer kimenete helyes	**14,70%**	
Ekvivalens alakok	3,55%	lévő→levő fele→felé áhá→aha periférikus→periferikus
Javított hibás név	1,01%	USÁ-ban→USA-ban Szóladon→Szóládon
Más javított hiba	10,14%	un.→ún. kollegánk→kollégánk lejto→lejtő lathato→látható
Valódi hibák	**67,40%**	
Hiányzik az elemzőből	5,57%	hemokromatózis-gén→hemokromatozis-gen
Helyes névből hibás kimenet	3,55%	MIG→míg Bösz→Bősz Ladd→Ládd Márton→Marton
Más helyes eredetiből hibás kimenet	2,20%	megőrzést→megorzést routeréhez→routerehez
Más helyes eredetiből a kontextusban hibás név	1,35%	logó→logo eperjeskein→eperjeskéin
Más helyes eredetiből a kontextusban hibás egyéb szó	51,01%	még→meg termek→termék gépét→gépet címét→címet vágyók→vagyok érméket→érmeket képé→képe
Az eredeti fájlnév vagy ékezetet tartalmazó URL	3,72%	latok→látok viz→víz szantok→szántók telepok→telepók www.valamicég.hu→www.valamiceg.hu
Nem javított hiba az eredetiben	**17,91%**	
Központozási hiba az eredetiben	6,42%	közalk.tan→kozalk.tan 1922.évi→1922.evi
Elválasztási hiba az eredetiben	0,68%	bemuta-tásra→bemuta-tasra
Egyéb hiba az eredetiben	10,81%	véri→veri ra→rá gonolkozásában→gonolkozasaban imátkoztok→imatkoztok hirújsághoz→hirujsasghoz változaban→változában környezetkíméli→kornyezetkimeli

6. Konklúzió

A cikkben egy magyar szövegek ékezetesítésére alkalmas rendszert mutattunk be. A statisztikai gépi fordítón alapuló alaprendszer fix méretű tanítókorpuszból épített fordítási-, és nyelvmodell használatával az esetek 98,44%-ában tudta helyesen ékezetesíteni az ékezet nélküli szóalakokat. Ez a rendszer azonban csak a tanítóanyagban szereplő szavak kezelésére képes. Ennek a problémának a megoldására morfológiai elemzőt integráltunk a rendszerbe, ami a tanítóanyagból hiányzó szavakhoz ékezetesített szóalakjelölteket generál. Ezzel a megoldással a helyesen ékezetesített szavak aránya 99,06%-ra nőtt. A rendszer további jellemzője, hogy a szövegkörnyezetet is figyelembe veszi nyelvmodell alkalmazásával, emellett nem generál értelmetlen szóalakokat, ami az ismeretlen szavak karakteralapú kezelésénél elkerülhetetlen lenne.

Hivatkozások

1. Halácsy, P., Kornai, A., Németh, L., Rung, A., Szakadát, I., Trón, V.: Creating Open Language Resources for Hungarian. In: LREC. European Language Resources Association (2004)
2. Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E.: Moses: Open Source Toolkit for Statistical Machine Translation. In: Proceedings of the ACL 2007 Demo and Poster Sessions. pp. 177–180. Association for Computational Linguistics, Prague (2007)
3. Mihalcea, R., Nastase, V.: Letter level learning for language independent diacritics restoration. In: Proceedings of the 6th Conference on Natural Language Learning - Volume 20. pp. 1–7. COLING-02, Association for Computational Linguistics, Stroudsburg, PA, USA (2002), http://dx.doi.org/10.3115/1118853.1118874
4. Németh, G., Zainkó, Cs., Fekete, L., Olaszy, G., Endrédi, G., Olaszi, P., Kiss, G., Kis, P.: The design, implementation, and operation of a Hungarian e-mail reader. International Journal of Speech Technology 3(3-4), 217–236 (2000)
5. Novák, A.: Milyen a jó Humor? In: I. Magyar Számítógépes Nyelvészeti Konferencia. pp. 138–144. SZTE, Szeged (2003)
6. Pham, L.N., Tran, V.H., Nguyen, V.V.: Vietnamese text accent restoration with statistical machine translation. In: Proceedings of the 27th Pacific Asia Conference on Language, Information, and Computation (PACLIC 27). pp. 423–429. Department of English, National Chengchi University (2013), http://aclweb.org/anthology/Y13-1044
7. Prószéky, G., Kis, B.: A unification-based approach to morpho-syntactic parsing of agglutinative and other (highly) inflectional languages. In: Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics. pp. 261–268. ACL '99, Association for Computational Linguistics, Stroudsburg, PA, USA (1999)
8. Scannell, K.P.: Statistical unicodification of african languages. Language Resources and Evaluation 45(3), 375–386 (2011)
9. Stolcke, A., Zheng, J., Wang, W., Abrash, V.: SRILM at sixteen: Update and outlook. In: Proc. IEEE Automatic Speech Recognition and Understanding Workshop. Waikoloa, Hawaii (Dec 2011)

10. Yarowsky, D.: A comparison of corpus-based techniques for restoring accents in Spanish and French text. In: Proceedings of the 2nd Annual Workshop on Very Large Text Corpora. pp. 19—-32. Las Cruces (1994)
11. Zainkó, Cs., Németh, G., Olaszy, G., Gordos, G.: Eljárás adott nyelven ékezetes betűk használata nélkül készített szövegek ékezetes betűinek visszaállítására (2000)
12. Zweigenbaum, P., Grabar, N.: Accenting unknown words in a specialized language. In: Johnson, S. (ed.) ACL Workshop on Natural Language Processing in the Biomedical Domain. pp. 21–28. ACL (2002)