Helyesírási hibák automatikus javítása orvosi szövegekben a szövegkörnyezet figyelembevételével

Siklósi Borbála[1], Novák Attila[1,2], Prószéky Gábor[1,2]

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

Kivonat: Cikkünkben egy korábban bemutatott orvosi helyesírás-javító rendszer lényegesen továbbfejlesztett változatát mutatjuk be, amely a korábbival ellentétben képes az egybeírások javítására, és a szövegkörnyezetet is figyelembe veszi ennek során, így alkalmas teljesen automatikus javításra is.

1 Bevezetés

A kórházakban keletkező szöveges dokumentumok olyan mennyiségű és minőségű gyakorlati tudást tartalmaznak, melyek feldolgozása és az eredmények felhasználása társadalmi szempontból hasznos, hozzájárulhat a ma sokszor hangsúlyozott életminőség javításához is. Mivel ezek a szövegek egyrészt mindenféle kontroll (pl. helyesírás-ellenőrző) alkalmazása nélkül készültek, másrészt az adott szövegtípusban nagyon magas arányban fordulnak elő a köznapi nyelvhasználattól idegen szóalakok: latin szavak, rengeteg rövidítés, gyógyszernevek, amelyeknek a helyesírására vonatkozó normákkal a szövegek íróinak nagy része nyilvánvalóan nincs tisztában, az ilyen szövegeknek a javítása nem könnyű feladat. A megvizsgált klinikai szövegekben jellemzően jelen vannak a hivatalos normától eltérő használatból fakadó, de következetesen elkövetett hibák, a véletlen melléütések, a következetlen szóhasználat, illetve az olyan többértelmű elírások, melyek helyességének megítélése még orvosi szakértelemmel sem egyértelmű (pl. elírt rövidítések). Emellett jellemző még az általános helyesírás-ellenőrzés során is felmerülő további probléma is: önmagukban helyes, de az adott környezetben téves szóalakok is előfordulnak. Mivel a szövegekben sok kifejezés gyakran egyáltalán nem fordul elő a helyesírási normának megfelelő formában, ezért úgy találtuk, hogy a gyakorlatban nem érdemes a szövegeket a helyesírási normáknak tökéletesen megfelelő formára hozni, ehelyett elegendő a szövegek egységesítése.

Cikkünkben bemutatjuk, hogy egy korábban létrehozott, helyesírási hibákat felismerő, s azok javítására automatikus javaslatokat generáló rendszer továbbfejlesztése során milyen stratégiákat alkalmaztunk a szövegkörnyezet figyelembevételére, illetve a generált javaslatok közül a megfelelő jelölt kiválasztására. Bemutatjuk, hogy az így létrehozott rendszer pontossága jelentősen javult, illetve az algoritmus kifinomultságának köszönhetően a paraméterek módosításával érzékenyebben hangolható. Ez a

megoldás könnyen kiterjeszthető azoknak a szavaknak a kezelésére is, amelyekben egynél több hiba van, mellyel további javulást érhetünk el. Így egy olyan rendszer kifejlesztéséről számolunk be, amely a jelentős minőségbeli javulás mellett használhatóságában is közelebb került egy teljesen automatikusan működő eszköz megalkotásához, mellyel az orvosi szövegek normalizálása megoldhatóvá válik a további felhasználási lépések előkészítéseként.

2 Helyesírási hibák

A klinikai dokumentumok jellegzetessége, hogy gyorsan, utólagos lektorálás, ellenőrzés, illetve automatikus segédeszközök (pl. helyesírás-ellenőrző) nélkül készülnek, ezért a leírás során keletkezett hibák száma igen nagy, valamint sokféle lehet. Így nem csupán a magyar nyelv nehézségeiből eredő problémák jelennek meg, hanem sok olyan hiba is felmerült a szövegekben, melyek a szakterület sajátosságaiból erednek. A legjellemzőbb hibák az alábbiak:

- elgépelés, félreütés, betűcserék,
- központozás hiányosságai (pl mondathatárok jelöletlensége) és rossz használata (pl. betűközök elhagyása az írásjelek körül, illetve a szavak között),
- nyelvtani hibák,
- mondattöredékek,
- a szakkifejezések latin és magyar helyesírással is, de gyakran a kettő valamilyen keverékeként fordulnak elő a szövegekben (pl. tensio/tenzio/tensió/ tenzió); külön nehézséget jelent, hogy bár ezeknek a szavaknak a helyesírása szabályozott, az orvosi szokások rendkívül változatosak, és időnként még a szakértőknek is problémát jelent az ilyen szavak helyességének megítélése,
- szakterületre jellemző és sokszor teljesen ad hoc rövidítések, amelyeknek nagy része nem felel meg a rövidítések írására vonatkozó helyesírási és központozási szabályoknak.

A fenti hibajelenségek mindegyikére jellemző továbbá, hogy orvosonként, vagy akár a szövegeket lejegyző asszisztensenként is változóak a jellemző hibák. Így elképzelhető olyan helyzet, hogy egy adott szót az egyik dokumentum esetén javítani kell annak hibás volta miatt, egy másik dokumentumban azonban ugyanaz a szóalak egy sajátos rövidítés, melynek értelmezése nem egyezik meg a csupán elírt szó javításával.

A feladat másik nehézségét az jelentette, hogy egyáltalán nem állt rendelkezésünkre nagyméretű helyesen írt klinikai korpusz, amely alapján elő tudtunk volna állítani a javításhoz használható hibamodelleket.

3 Statisztikai gépifordító-rendszer helyesírási hibák javítására

Célunk egy korábban [13]-ban már bemutatott, csak izolált szavak alapvető tulajdonságait figyelembe vevő és ezeket alkalmazó rendszer továbbfejlesztése volt. A fent leírt nehézségek miatt a rendszer elsősorban a szakterület nyelvére épített statisztiká-

kat vette legnagyobb súllyal figyelembe – természetesen a morfológia mellett –, ami biztosítja a speciális szövegek sajátosságainak megtartását az általános szövegekből átvett formák alkalmazása helyett. A korábbi eredmények során bemutattuk, hogy az így létrejött rendszer a szövegekben lévő hibákat felismeri, az azokhoz automatikusan generált és rangsorolt javaslatok között az első tíz az esetek 98%-ában tartalmazta a helyes alakot.

Mivel célunk a háttérben futó automatikus normalizálás, és nem az, hogy a javaslatokat egy felhasználónak mutassa meg a rendszer, aki aztán kiválasztja a helyes alakot, ezért fontos, hogy a rendszer képes legyen a generált javaslatok közül a valóban helyeset automatikusan kiválasztani. A legjobb javítás kiválasztásához kevésnek bizonyult a korábbi rendszerben alkalmazott, kizárólag morfológiára és különböző szóstatisztikákra épülő rangsorolás. Az automatikus javítás pontosságának növeléséhez szükséges az egyes szavakhoz tartozó szövegkörnyezet figyelembevétele is. E két követelmény alkalmazására a statisztikai gépi fordítás területén széles körben alkalmazott Moses keretrendszert használtuk. A fordítás során forrásnyelvnek az eredeti hibás szöveget tekintjük, míg a célnyelv ennek javított formája. Ebben az esetben a rendszer bemenete a hibás mondat: $H=h_1,h_2...h_n$, melynek megfelelő javított mondat a $J=j_1,j_2..j_k$ a várt kimenet. A helyesírás-javító rendszer zajos csatornamodellként tehát úgy fogalmazható meg, hogy az eredeti üzenet a helyes mondat, amely helyett a csatornán átért jel a zajos, azaz hibás mondat. Így a javítás az a $\hat{J}$ mondat lesz, melyre a

$$\hat{J} = argmax\ P(J|H) = argmax\ P(H|J)P(J) / P(H) \quad (1)$$

feltételes valószínűség a maximális. Mivel $P(H)$ értéke állandó, ezért a nevező elhagyható, így a számlálóban lévő szorzat a fordítási és nyelvmodellből számított statisztika alapján számítható.

Ezeket a modelleket a hagyományos statisztikai gépifordító-rendszerek esetén a forrás- és célnyelvű párhuzamos korpuszból számolt valószínűségek képezik. Ilyen korpusz azonban a mi esetünkben nem áll rendelkezésre, ezért a fordítási modellt a korábban létrehozott rendszer rangsorolásához használt számítási értékek valószínűségekké konvertálása képezi. A szövegkörnyezet figyelembevétele érdekében pedig a SRILM eszköz segítségével létrehozott nyelvmodell módosítja a "fordítás" során kapott eredményeket.

3.1 A fordítási modellek

A rendszerben három szó-, illetve hibatípus kezelésére három fordítási (hibajavítási) modellt alkalmazunk. Az egyik kifejezetten a rövidítések kezelésével, a másik a tévesen egybeírt szavak felbontásával, a harmadik pedig az egyéb elírásokkal foglalkozik. A modellek bemutatását ez utóbbival kezdjük.

3.1.1 Fordítási modell általános szavakra

A fordítási modellt, vagy más néven frázistáblát a korábban implementált javaslatgeneráló eredménye alapján építettük. Minden javítandó szóhoz (a potenciális rövidíté-

seket és stopszavakat külön kezeljük) az első 20-ként rangsorolt javaslatot vettük figyelembe. Ezek között a javaslatok között szerepelhet olyan eset is, amikor az eredeti alakba szóköz kerül, így az esetleg hibásan egybeírt szavak szétválasztásának lehetősége is belekerült a rendszerbe. Mivel korábban bemutattuk, hogy a javaslatok generálása során az esetek 98%-ában az első tíz javaslatban (amely még nem tartalmazott szóközbeszúrási lehetőséget) benne volt a helyes alak, ezért 20-nál több lehetőség figyelembevétele még a különírások figyelembevételével is csak fölösleges zajt generált volna. Az így megkapott javaslatok rangsorának kialakításához használt pontszámot (l. [13]) alakítottuk valószínűséggé, azaz az egy szóhoz tartozó lehetséges javítások valószínűségének összege 1. Ezzel a módszerrel helyettesítettük a párhuzamos korpuszból való tanítást.

1. táblázat: Részlet az általános szavakat tartalmazó frázistáblából.

hosszúságu ||| hosszúsági ||| 0.016497642667 ||| |||
hosszúságu ||| hosszúságú ||| 0.0156006851784 ||| |||
hosszúságu ||| hosszúsága ||| 0.013537671904 ||| |||
hosszúságu ||| hosszúságuk ||| 0.013537671904 ||| |||
hosszúságu ||| hosszúságul ||| 0.013537671904 ||| |||
hosszúságu ||| hosszúságé ||| 0.013537671904 ||| |||
hosszúságu ||| hosszúság ||| 0.013537671904 ||| |||

3.1.2 Fordítási modell rövidítésekre

A klinikai szövegekre jellemző, hogy az általános szövegeknél sokkal magasabb arányban tartalmaznak rövidítéseket. Ezek a fenti modellel két okból nehezen kezelhetőek. Egyrészt gyakran ugyanannak a szónak vagy kifejezésnek számos különböző rövidített alakja előfordul a szövegekben a dokumentum rögzítőjének egyedi szokásai miatt, vagy mert éppen úgy sikerült. Másrészt pedig ezekhez általában létezik több olyan generált javaslat, amelyek más, helyes szavakká alakítják ezeket, s mivel az ilyen valódi szavak előfordulása gyakoribb minden korpuszban, illetve a morfológia is megerősíti ezek helyességét, ezért könnyen átíródnának a szándékolt eredeti jelentéstől teljesen eltérő szóalakokká. Ezért automatikus módszerek alkalmazásával kigyűjtöttük az orvosi dokumentumokban előforduló rövidítéseket, melyek egy részét kézzel történő szűrés után a morfológiába is beépítettük. A fordítórendszer számára azonban ez az eljárás nem elégséges, hiszen egy-egy rövidítésnek számos alakja fordul elő a szövegekben (legjellemzőbb példa a rövidítések végén a pont elhagyása stb.). Ezért minden potenciális rövidítéshez kigyűjtöttük a dokumentumokból ezek variációit, gyakorisággal együtt, majd az így kapott gyakoriságokat szintén valószínűségekké alakítottuk. Így létrejött egy alternatív frázistábla a fordítórendszer számára. Mivel az ebben a modellben szereplő szóalakokhoz az előző pontban leírt módon nem generálunk javaslatokat, hogy ne javítsunk rövidítéseket egész más szavakká, ezért ezek csak ebben a második frázistáblában szerepelnek, az elsőben nem. A fordítórendszert úgy

alakítottuk ki, hogy az a fordítás során a bemenetre érkező szóalakhoz abból a táblából számít fordítási lehetőséget, amelyikben a szóalak megtalálható.

2. táblázat: Részlet a rövidítéseket tartalmazó frázistáblából.

conj. ||| conj. ||| 0.607803468208 ||| |||
conj ||| conj. ||| 0.869653179191 ||| |||
conj ||| conj ||| 0.130346820809 ||| |||
mko ||| mko ||| 0.489190805776 ||| |||
mko ||| mko. ||| 0.997094762027 ||| |||
mko. ||| mko. ||| 0.999316414595 ||| |||

3.1.3 A téves egybeírásokat kezelő modell

Mivel a gépi fordításra használt keretrendszert általában frázisalapú fordításra alkalmazzák a többnyelvű fordítás során, ezért általános jellemzője a hagyományos módon használt rendszerben lévő frázistáblának, hogy abban egy (vagy több) szóhoz tartozhat több szóból álló fordítás is. Így a mi esetünkben sem okozott problémát az, hogy nem csupán szóalapú megfeleltetéseink vannak, hanem egy szóból esetlegesen több szó is képződhet. Természetesen ezekhez is a javaslatgeneráló által kapott pontszámból számított valószínűség került a modellbe. Ezekben az esetekben a javaslatok pontszáma úgy adódik, hogy a szóköz beszúrásával kapott két lehetséges szóra számolja ki az értékeket, majd átlagolja, így kap a két szóból álló javaslat egy olyan pontszámot, amely nagyságrendileg illeszkedik az egy szóból álló javaslatok listájába.

3. táblázat: Részlet az általános szavakat tartalmazó frázistáblából.

soronkívül ||| soron kívül ||| 0.0207452583298 ||| |||
soronkívül ||| soronkívül ||| 0.0145949359186 ||| |||

3.2 A nyelvmodell

A nyelvmodell szerepe a javítás során az, hogy a fordítási modell alapján létrejött javított mondatokban szereplő szavak sorozatának előfordulási lehetőségét ellenőrizze, és a valós előfordulás felé súlyozza. Ennek a komponensnek a feladata, hogy a javítás során a szövegkörnyezetet is figyelembe vegye a rendszer. Ehhez az adott szövegtípusra jellemző helyes korpuszból kellene a kívánt hosszúságú szósorozatokat (szó n-eseket) tartalmazó statisztikát létrehozni. Mivel a rendelkezésünkre álló dokumentumoknak csak a tesztelésre használt része az, amely kézzel ellenőrzött módon helyesnek tekinthető, ezért nem volt lehetőségünk helyes korpuszból tanított nyelvmodellt létrehozni. Bár más témájú, illetve más stílusú szöveges korpuszok természetesen

léteznek, az ezekben található szó n-esek nem feltétlenül modellezik jól a klinikai szövegeket, ezért úgy döntöttünk, hogy nem használunk ilyen szövegeket. Azt láttuk azonban, hogy a klinikai dokumentumokban számos olyan szófordulat, szósorozat van, ami nagyon gyakori, összességében viszont kevés számú különböző szó n-es fordul elő, azaz a klinikai dokumentumok nyelvezete viszonylag korlátozottnak tekinthető ilyen szempontból.

4. táblázat: Általános magyar nyelvű és orvosi szövegekben előforduló különböző n-gramok száma (800000 mondatos korpuszban).

	Általános szöveg	Orvosi szöveg
1-gram	873951	275609
2-gram	4794135	1409290
3-gram	7886616	2440636

Ezért a hibás szót tartalmazó szósorozatok esetén ugyanezeknek a szósorozatoknak a helyes előfordulása gyakoribb, tehát a nyelvmodell-statisztikát a vártnál kisebb mértékben tekinthetjük torznak a hibás szavakat is tartalmazó korpuszból építve. Természetesen a kiértékelés során a mérésekhez használt teszthalmaz mondatait már a nyelvmodell építése előtt különválasztottuk a korpusztól, hiszen az ezekben megtalálható hibás szósorozatokra teljesen illeszkedő n-eseket találnánk a teljes mondatra, ami viszont már azzal járna, hogy a javítás helyett az eredeti szóalakok kapnának nagyobb súlyt.

A korpuszt alkotó dokumentumokat az előfeldolgozás során végzett tokenizálással egy időben a feltételezhető mondathatároknál mondatokra is bontottuk. Az így kapott mondatok átlagos hosszát (8,58 token/mondat) figyelembe vettük a nyelvmodell építése során, ezért a nyelvmodellt úgy hoztuk létre, hogy az abban szereplő szó n-esek maximális hossza három token. A rövid mondatok miatt nem várható el ennél hosszabb n-gramok esetén az illeszkedés. Ezt méréseink is megerősítették: nagyobb n esetén a végeredmény rosszabb lett.

Fontos megjegyezni még, hogy a nyelvmodell létrehozása előre megtörténik, ezért a dekódoláshoz szükséges idő az egyes mondatok esetén nem növeli számottevő mértékben a javításhoz szükséges időt.

3.3 Dekódolás

A fenti modellek alapján az (1) képlet alkalmazásával számított eredmény meghatározását a fordítórendszer magját képező dekódoló algoritmus végzi. Ehhez a Moses keretrendszert alkalmaztuk, amely a statisztikai gépi fordítás területén a legelterjedtebb eszköz. A dekódolás paramétereit a konfigurációs fájlban lehet beállítani. Így könnyen és gyorsan változtathatóak a rendszer paraméterei, ami igen rugalmassá teszi azt. A dekódolás során minden bemeneti mondathoz a fent részletezett modellek alapján mondatszintű fordítások (azaz a mi esetünkben javítások) jönnek létre, melyben a szószintű javítási lehetőségeket a frázistábla, a szövegkörnyezet figyelembevételét

pedig a nyelvmodell biztosítja. A dekódolás során a következő paramétereket használtuk:

- A frázistáblák súlyozása: mivel a frázistáblákban szereplő szavak halmazai diszjunktak, ezért ezek súlyának beállítása független egymástól. A szövegek javítása valójában inkább azok egységesítését jelenti, nem pedig a szigorú értelemben vett helyesírási normához való igazítását. Ezért a rövidítések sokféle megjelenési formája miatt ezeknél fontosabbnak láttuk az átírást egy meghatározott formára (amely általában a ponttal jelölt alak), így a rövidítésekhez tartozó fordítási modell nagyobb súlyt kapott.
- Nyelvmodell: trigram nyelvmodellt alkalmaztunk, azaz a nyelvmodellben szereplő szó n-esek hossza maximum 3. A dekódolás során a nyelvmodell a fordítási modellnél alacsonyabb súlyt kapott, hogy megakadályozzuk a hibás szövegekből készült nyelvmodellben előforduló hibás n-gramok előtérbe kerülését.
- Átrendezési korlát: a különböző nyelvek közötti fordítás során szükséges lehet a megfeleltetett szavak sorrendjének is a megváltoztatása, a helyesírás-javítás során azonban ezt nem engedhetjük meg, hiszen a módosítások csak szavakon belül, illetve szóközök beszúrásával történhetnek, a szavak sorrendjén a javítórendszer nem változtathat.
- Mondathossz-különbség: mivel a javított mondat hossza sem térhet el jelentősen az eredeti mondattól (ha minden szóba beszúrnánk egy szóközt, akkor érnénk el az elvi korlátot, de a valóságban ilyen nem fordulhat elő, mondatonként két egybeírási hiba volt a legtöbb, ami a teszthalmazban szerepelt), ezért nem szükséges a dekódolás során a mondathossz-eltérést külön súllyal büntetni.

4 Eredmények

A rendszer kiértékeléséhez szükséges volt az orvosi dokumentumokból létrehozott teszthalmaz kézzel való kijavítása, így létrejött egy 2000 mondatból álló, vegyes tartalmú (különböző klinikai osztályok anyagaiból származó) teszthalmaz. A nyelvmodell létrehozásához a fennmaradó 978000 mondatból álló korpuszt használtuk. Mindkét részhalmaz csak szabad szövegekből álló mondatokat tartalmaz, tehát az amúgy is szabványos BNO-kódokkal párosított betegségmegnevezéseket, kódokat, mérési és laboreredményeket nem tartalmazott a korpusz. Ennek ellenére számos olyan "mondat" került mind a tanítóanyagba, mind a teszthalmazba, amelyek nehezen értelmezhető, speciális tartalmú szavakat, rövidítéseket, gyakran rövidítéssorozatokat tartalmaztak. Ezek helyességének a megítélése külön feladat, amihez megfelelő szakterületi ismereteink hiánya miatt egy vagy több általunk helyesnek ítélt változatot fogalmaztunk meg elfogadható javításnak. Ráadásul az elvi helyesírási szabályoknak megfelelő formára való hozást el kellett vetnünk. Ennek egyik oka, hogy a gyakorlati alkalmazás során sok esetben a helyesírási szabályoknak ellentmondó írásváltozatok a korpuszban sokkal gyakoribbak voltak, mint a helyesírási normának megfelelő változat (amely sok esetben a korpuszban egyáltalán nem szerepelt). Úgy véljük, hogy a szövegekben szereplő fogalmak visszakereshetőségét az egységesítés abban az esetben is lehetővé

teszi, ha az alkalmazott egységes alak nem azonos valamely helyesírási norma által szentesített alakkal. Ezért a kézzel kijavított teszthalmaz létrehozásakor mindezen szempontokat figyelembe véve annak több változatát is elkészítettük, a lehetséges javításokkal.

A kiértékelés során ezen a teszthalmazon különböző metrikák szerint végeztünk méréseket. A gépi fordítás minőségére általánosan elterjedt mérőszám a Bleu érték meghatározása. Mivel a felfogásunk szerint a javítás folyamatát is egyfajta fordításként értelmezhetjük, ezért az egyik mérőszámként mi is ezt a metrikát használtuk. Ennek lényege, hogy a javítás eredményét a referenciafordításhoz hasonlítva a szavak sorrendjét is figyelembe vevő módosított pontosságértéket számol. Emellett a helyesírás-javítás hagyományos értelemben vett feladata során szokásos fedés, pontosság, F-mérték hármas mentén is vizsgáltuk a rendszer minőségét.

Mivel célunk a korábban létrehozott javító rendszer eredményeinek javítása volt, ezért azt tekintettük alaprendszernek. Ahogy az 5. táblázatból látszik, ez a korábbi rendszer megtalálta a legtöbb hibát (magas fedésérték), ám a szövegkörnyezetet figyelembe nem vevő pusztán rangsoroláson alapuló javítás pontossága rosszabb volt, ezért a kettő átlagából számított F-mérték is kisebb, csupán 72% volt.

5. táblázat: Az automatikus helyesírás-javító rendszerek minősége automatikus metrikák alapján.

	Pontosság	Fedés	F-mérték	Bleu
Alaprendszer	0,70	0,75	0,725	-
SMT	0,8814	0,8857	0,8826	0,8085

A statisztikai fordítórendszer alkalmazásával jelentősen javultak a mért eredmények, a legjobb paraméterbeállítás során 88%-os F-mértéket értünk el. Több olyan konfigurációval is elvégeztük a javítást, melynek eredményei valamivel rosszabb értéket produkáltak, de bizonyos jelenségek kezelésére mégis alkalmasabbak voltak.

6. táblázat: Eredetileg hibás mondatok és a hozzájuk tartozó automatikus javítás az alaprendszerrel, illetve a statisztikai módszer használatával (melynek eredménye megegyezik a referencia javítással).

Eredeti mondat:	*csppent előírés szerint ,*
Alaprendszer javítása:	*cseppent előír és szerint ,*
SMT javítás:	*cseppent előírás szerint ,*
Eredeti mondat:	*th : mko tovább 1 x duotrav 3 ü-1 rec , íb : 2 x azoipt 3 ü-1 rec*
Alaprendszer javítása:	*th : mko tovább 1 x duotrav 3 ü-1 sec , kb : 2 x azoipt 3 ü-1 sec*
SMT javítás:	*th. : mko tovább 1 x duotrav 3 ü-1 rec , kb : 2 x azopt 3 ü-1 rec*
Eredeti mondat:	*/alsó m?fogsor .*
Alaprendszer javítása:	*/alsó műfogsor .*
SMT javítás:	*alsó műfogsor .*
Eredeti mondat:	*vértelt nyállkahártyák , kp erezett conjuctiva , fehér sclera .*
Alaprendszer javítása:	*vértelt nyálkahártyák , kp erezett conjunctiva , fehér sclera .*
SMT javítás:	*vértelt nyálkahártyák , kp. erezett conjunctiva , fehér sclera .*

5 Nehéz esetek

Az automatikus javítás során több olyan jelenség van, amelyeket a javítónak nem sikerül kezelnie. Néhány példa:

- Vannak esetek, amikor a javító egy helyes szót, egy másik helyes szóra ír át, illetve egy elírt szót nem a megfelelő, ámde helyes szóalakra javít, melyek az adott mondatban is helytállóak. Ilyen esetek a nyelvmodellben való előfordulások miatt kerülnek előtérbe, a korábban említett nyelvmodellel kapcsolatos problémák miatt - különösen a rövid mondatoknál. További probléma, hogy két szó megváltoztatása esetén nem kap elég hangsúlyt az ezért járó büntetés, aminek erősítése viszont a ténylegesen szükséges javítások esélyét csökkentené.

  eredeti mondat: homályos látást panaszol .
  javított mondat: homályos látás panaszok .

  eredeti mondat: panasz nem volt .
  javított mondat: panasza nem volt .

  (A második példában a mondat átírása helyes változatot eredményez, de a referencia és a valós környezet nem tartaná szükségesnek az átírást.)

- Az egynél több hibát tartalmazó szavak javítása egyelőre nem lehetséges:
  eredeti mondat: gyógyógyszerei : ld lázlap
  javított mondat: gyógyógyszerei : ld lázlap

6 Összefoglalás

Cikkünkben bemutattuk, hogy a nagyon zajos, magyar nyelvű, de speciális nyelvezetű és rövidítésekkel teletűzdelt orvosi szövegekben lévő helyesírási hibák automatikus javítása a szövegkörnyezet figyelembevételével elég nagy pontossággal elvégezhető. A helyesírási normáknak megfelelő szaknyelvi korpusz hiányában a szövegek egységesítésének egyik lehetséges útja egyelőre az lehet, ha a normát nem úgy értelmezzük, hogy csak a szabványos helyesírásnak megfelelő alakokat fogadjuk el, hanem a korpuszban túlnyomó többségben szereplő alakokat is normalizálási célpontnak tekintjük. Így létrejöhet a szövegek egységes reprezentációja, ami a további feldolgozás szempontjából alapvető fontosságú.

Természetesen a feladatot még nem oldottuk meg tökéletesen, bemutattuk azokat a hibajelenségeket, amelyek kezelése még előttünk áll. További terveink között szerepel, hogy a rendszer részévé tegyük a PurePos szófaji egyértelműsítőt is, amely olyan kiegészítő információkat biztosítana, melyekre támaszkodva tovább javíthatnánk a rendszer teljesítményét a jelenleg még kétséges esetekben. Ezen túl további javulást remélünk a nyelvmodell építésére használt korpusznak az automatikus javítások utáni

iteratív újraépítésétől, ami a jelenleg igen zajos nyelvmodell helyességén javítana, így annak torzító hatását kiküszöbölné.

Köszönetnyilvánítás

Ez a munka részben a TÁMOP 4.2.1.B – 11/2/KMR-2011–0002 pályázat támogatásával készült.

Hivatkozások

1. Dustin, B.: Language Models for Spelling Correction CSE 256 (2004)
2. Brill, E., Moore, R.C.: An improved error model for noisy channel spelling correction. In: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics (2000) 286–293
3. Contractor, D., Faruquie, T.A., Subramaniam, L.V.: Unsupervised cleansing of noisy text. In: Proceedings of the 23rd International Conference on Computational Linguistics (2010) 189–196
4. Heinze, D.T., Morsch, M.L., Holbrook, J.: Mining Free-Text Medical Records. A-Life Medical, Incorporated (2001) 254–258
5. 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 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions. Association for Computational Linguistics, Prague, Czech Republic (2007) 177–180
6. Mykowiecka, A., Marciniak, M.: Domain-driven automatic spelling correction for mammography reports. In: Intelligent Information Processing and Web Mining Proceedings of the International IIS: IIPWM'06. Advances in Soft Computing, Heidelberg (2006)
7. Ehsan, N., Faili, H.: Grammatical and Context-sensitive Error Correction Using a Statistical Machine Translation Framework. In: Software Practice and Experience (2011)
8. Novák A.: Milyen a jó Humor? In: Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2003). Szegedi Tudományegyetem (2003) 138–145
9. Orosz, Gy., Novák, A.: PurePos — an open source morphological disambiguator. In: Proceedings of the 9th International Workshop on Natural Language Processing and Cognitive Science (2012) 53–63
10. Patrick, J., Sabbagh, M., Jain, S., Zheng, H.: Spelling Correction in Clinical Notes with Emphasis on First Suggestion Accuracy. In: 2nd Workshop on Building and Evaluating Resources for Biomedical Text Mining (2010) 2–8
11. Pirinen, T.A., Lindén, K.: Finite-State Spell-Checking with Weighted Language and Error Models – Building and Evaluating Spell-Checkers with Wikipedia as Corpus. In: SaLTMiL Workshop on Creation and Use of Basic Lexical Resources for Less-Resourced Languages, LREC (2010) 13–18
12. Prószéky, G., Novák, A.: Computational Morphologies for Small Uralic Languages. In: Inquiries into Words, Constraints and Contexts (2005) 150–157

13. Siklósi, B., Orosz, Gy., Novák A., Prószéky G.: Automatic structuring and correction suggestion system for Hungarian clinical records. In: Proceedings of the 8th SaLTMiL Workshop on Creation and use of basic lexical resources for less-resourced languages (2012) 29–34
14. Stevenson, M., Guo, Y., Amri, A., Gaizauskas, R.: Disambiguation of biomedical abbreviations. In: Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing (2009) 71