Gépi tanulási módszerek ómagyar kori szövegek normalizálására

Oravecz Csaba, Sass Bálint, Simon Eszter

MTA Nyelvtudományi Intézet
e-mail:{oravecz,sass.balint,eszter}@nytud.hu

Kivonat A nyelvemlékek számítógéppel segített feldolgozása és elemzése számos problémát felvet, a nyelvtörténeti kérdésektől az egészen konkrét technológiai nehézségekig. A többféle, különböző nyelvtörténeti szakmai érvekkel alátámasztható lehetséges feldolgozási „forgatókönyv" egyik gyakori közös átalakító lépése a szokásos betűhű átírásban kiadott szövegek mai modern helyesírású változatának előállítása. Ez a szövegnormalizáló konverzió analóg több klasszikus nyelvfeldolgozási probléma során jelentkező feladattal, ezért az azokban sikerrel alkalmazott zajos csatorna modellt adaptáljuk és vizsgáljuk ennek eredményességét a transzliterációs feladatban.

Kulcsszavak: gépi tanulás, zajos csatorna modell, nyelvtörténet, normalizálás, transzliteráció

1. Bevezetés

A Nyelvtudományi Intézetben április óta folyik egy projekt, melynek a célja egy elektronikus nyelvtörténeti adatbázis létrehozása. Az adatbázis tartalmazza az összes ómagyar szövegemléket, a középmagyar korból pedig különféle szempontok szerinti arányosan válogatást úgy, hogy minden nyelvjárás, műfaj, regiszter súlyának megfelelően legyen képviselve benne. Ehhez első lépésben össze kell gyűjteni az összes elektronikus formában elérhető szöveget, majd egységes formátumra hozni őket. A szövegemlékek eredeti, betűhű változatukban és egy ún. *normalizált változatban* is elérhetők, kereshetők lesznek. Ez a normalizálási lépés a szövegfeldolgozási munkafolyamatnak az a lépése, amikor az eredeti betűhű szóalakokat mai magyar helyesírású szavakra alakítjuk át. A többféle, különböző nyelvtörténeti szakmai érvekkel alátámasztható lehetséges feldolgozási forgatókönyvek egyik gyakori közös átalakító lépése ez a fajta normalizálás (pl. (McEnery és Hardie, 2003)). A folyamat számítógépes modellezésének célja az, hogy választ kapjunk arra a nagyon fontos gyakorlati kérdésre, hogy a rendkívül időigényes manuális átírási munka kiváltható-e gépi eljárással, így a szükséges emberi erőforrás alkalmazása leszűkíthető-e a tanuló adatok előállításának feladatára. Mivel ez a szövegnormalizáló konverzió analóg több klasszikus nyelvfeldolgozási probléma során jelentkező feladattal, így feltétlen érdemesnek tűnik az

azokban sikerrel alkalmazott módszerek adaptálása és eredményességének vizsgálata.

A dolgozat központi kérdése annak meghatározása, hogy az átírási feladat miként illeszthető be meghatározott gépi tanulási modellekbe, és melyek azok a paraméterek, amelyek felhasználása ezekben a modellekben a feladat elfogadható pontosságú megoldását eredményezi. Ennek érdekében szükség van az adott modellben használt jegyeket tartalmazó specifikusan annotált tanító szövegekre, melyekből jelenleg korlátozott mennyiség áll a rendelkezésünkre — lévén a normalizálás nyelvtörténeti szakértelmet kívánó, időigényes munka. További nehézséget jelent, hogy az egyes nyelvemlékek írásmódja, a bennük előforduló speciális ómagyar karakterek halmaza is meglehetősen különbözik egymástól. A „könyvméretű magyar írásosságot" a latin nyelvű és vallásos tárgyú irodalom fordításának igénye hívta életre, de a latin ábécé magyarra alkalmazása számos problémát vetett fel. A legfőbb gond abból fakadt, hogy nyelvünk hangrendszerének több eleme a latinban ismeretlen, így ezek jelölésére új jeleket kellett bevezetni. A 14-16. században a helyesírás még egyáltalán nem volt egységesítve, sőt egy kódexet akár több kéz is jegyezhetett, ami további egyenetlenségeket okoz a szövegekben. Ezért nehéz egyértelmű konverziós szabályokat meghatározni, valamint emiatt kritikus kérdés az, hogy a tanult modellek milyen mértékben általánosíthatók az eltérő nyelvemlékekre. Mindezek miatt célszerű a problémát valamilyen valószínűségi alapú paradigma keretei között vizsgálni, egyik legkézenfekvőbb erre Shannon zajos csatorna modellje (Shannon, 1948).

Esetünkben a normalizálás tulajdonképpen egybeesik azzal a fogalommal, amit a nyelvtörténészek értelmezésnek hívnak. Az értelmezés hagyományosan a régi nyelvi adatoknak mai magyar nyelvre való „fordítását" jelenti. A különböző helyesírási rendszerekben is ritka az egy hang–egy betű megfelelés (vagyis amikor egy hang jelölésére mindig ugyanaz a betű használatos, és az adott betűnek mindig egy hangértéke van), de egy alakulóban levő helyesírási rendszerben ilyenfajta következetesség még annyira sem várható el. Sőt inkább az a tipikus, hogy egy emléken belül is ingadozik egy-egy hang jelölésmódja (pl. ÓMS: Vylag uilaga [világ világa]), vagy kettős hangértéke van egy-egy betűnek (pl. MK: zerzete zerent [szerzete szerint]). Tovább bonyolítja a helyzetet, hogy néhány betű egyaránt utalhat magánhangzóra és mássalhangzóra is (pl. az u,v,w több évszázadon át jelölhette az u,ú,ü,ű,v hangok bármelyikét).

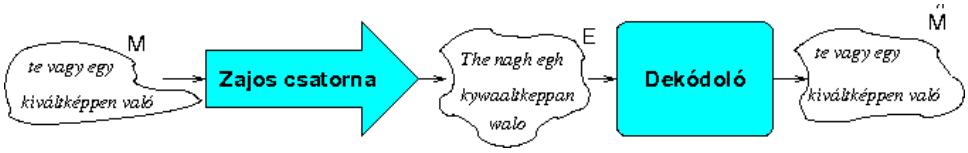
A dolgozat a következőképpen épül fel. A 2. rész rövid leírást ad az eddigi rokonítható kezdeményezésekről. A 3. rész az eljárás elméleti alapjait tárgyalja, míg a 4. részben a modell tanításának folyamatát mutatjuk be. Az 5. rész a modell alkalmazásáról és a lehetséges kiértékelési módszerről ad leírást. Rövid összefoglalás zárja a dolgozatot a 6. részben.

2. Kitekintés

A kitűzött feladat egyrészt lényegében tekinthető két reprezentáció közötti fordítási feladatnak, így közvetlenül rokonítható azokkal a megközelítésekkel, ahol a szövegnormalizáláshoz komplex gépi fordítási modelleket használnak (Raghunat-

han és Krawczyk, 2009; Kobus et al., 2008; Aw et al., 2006). További kapcsolódó problémakör a graféma-fonéma konverzió, ahol Lucassen és Mercer (1984) korai valószínűségi modelljére támaszkodik a legtöbb megoldási javaslat. Damper et al. (1999) tartalmaz részletes összehasonlítást, ahol kimutatja, hogy a gépi tanulási módszereket használó modellek jobb eredményeket adnak, mint a kézzel írt szabályokon alapulók. Számos analógiás továbbá rejtett Markov-modellen alapuló eljárást is eredményesen alkalmaztak (Bellegarda, 2005; Taylor, 2005). Az általunk használt módszer előzménye Kernighan et al. (1990) helyesírás-ellenőrzésre kidolgozott eljárása, illetve ennek továbbfejlesztett csatorna-modellt alkalmazó változatai (Brill és Moore, 2000; Toutanova és Moore, 2002).[1] A következő fejezet ezt modellt ismerteti részletesen. A fentiektől eltérő paradigmájú, szabály alapú megközelítésre példa Kiss et al. (2001).

3. Zajos csatorna alapú szövegnormalizáló modell



1. ábra. Szövegnormalizálás zajos csatorna modellben.

Az 1. ábrán látható modellben az eredeti szöveget úgy tekintjük, mint a normalizált változat egy zajos kommunikációs csatornán átment „eltorzított" változatát. Jelölje M a modern helyesírású normalizált szövegváltozat pl. egy (rész)mondatnyi sztringjét, E pedig ennek eredeti betűhű átiratát. A dekódoló feladata annak az M karaktersorozatnak a megtalálása, melyre a $P(M|E)$ feltételes valószínűség maximális,

$$\hat{M} = \operatorname*{argmax}_M P(M|E) \tag{1}$$

illetve a szokványos átalakítással:

$$\hat{M} = \operatorname*{argmax}_M \frac{P(E|M)P(M)}{P(E)} = \operatorname*{argmax}_M P(E|M)P(M) \tag{2}$$

A feladat tehát egyrészt a $P(E|M)$ transzliterációs modell-eloszlás (csatornamodell) és a $P(M)$ normalizált szövegmodell-eloszlás (forrásmodell) meghatározása.

Forrásmodellként a normalizált szövegből készült karakter N-gram modelleket használhatunk, ahol vizsgálható a módszer pontossága N függvényében.

[1] Természetesen számos további gépi tanulási paradigma is alkalmazható a feladat megoldására, a döntési fáktól a log-lineáris osztályozókig.

Mivel a normalizált szöveg alapvetően mai magyar nyelvű anyag, a forrásmodell felépítésében nagy mennyiségű adat hozzáférhető és használható, így N a szómodelleknél megszokott 3-nál nagyobb is lehet. A transzliterációs modell paramétereinek meghatározására többféle lehetőség kínálkozik, melyeknek előfeltétele olyan tanító korpusz, amely $M_i^j \rightarrow E_k^l$ megfeleléseket tartalmaz.[2] Az 1-nél hosszabb sztringekre definiált megfeleltetésekkel a transzliterációs modell kontextuális információt is képes reprezentálni. A modell paramétereit a tanító korpuszból becsüljük, míg a lehetséges modern szövegváltozatok halmazát a megfeleltetésekből generáljuk. Az alkalmazott eljárás hasonló Brill és Moore (2000) gépelési hibákat javító módszeréhez, melynek alapján a transzliterációs modell formálisan az alábbi módon írható le.

Legyen $\mathrm{Part}(M)$ a modern nyelvváltozatú sztring minden lehetséges nemkeresztező partíciójának halmaza (hasonlóan $\mathrm{Part}(T)$ az eredeti alakra). Egy adott $R \in \mathrm{Part}(M)$ partícióra, ahol R $|R| = j$ darab szegmentumból áll, legyen R_i az i-edik szegmentum. Ekkor ($|T| = |R|$ esetén, ahol $T \in \mathrm{Part}(E)$)

$$P(E|M) = \sum_{R \in \mathrm{Part}(M)} P(R|M) \sum_{T \in \mathrm{Part}(E)} \prod_{i=1}^{j} P(T_i|R_i) \tag{3}$$

Egy meghatározott illesztés megfelel adott $M_i^j \rightarrow E_k^l$ megfeleltetések halmazának. Csupán a legjobb particionálást tekintve (3) az alábbira egyszerűsödik:

$$P(E|M) = \max_{R \in \mathrm{Part}(M), T \in \mathrm{Part}(E)} P(R|M) \prod_{i=1}^{j} P(T_i|R_i) \tag{4}$$

Brill és Moore (2000) modelljéhez hasonlóan $P(R|M)$ meghatározásával egyelőre mi sem foglalkozunk, vagyis ezt a tényezőt nem vesszük figyelembe (illetve a partíciók felett jobb híján jelenleg egyenletes eloszlást feltételezünk).

4. A modell tanítása

4.1. A transzliterációs modell tanító korpuszának előállítása

A tanító korpusz két ómagyar kori szövegemlék nyelvtörténészek által kézzel normalizált változatából állt elő. A Müncheni emlék (Haader, 2005) a 16. század elejéről származó, sajátos nyelvemlékünk. Sajátossága abban rejlik, hogy egyszerre tartalmaz egyházi és világi szövegeket, valamint latin és német nyelvű részleteket is (ezeket a normalizálás és a tanító korpusz építése során kihagytuk). A Szabács viadala (Imre, 1958) a 15. század második felében keletkezett, eredeti magyar nyelvű vers. A legrégebbi ránk maradt históriás ének, a Mátyás király egyik haditettét elbeszélő 150 sor egy hosszabb költeménye része lehetett. A két nyelvemlék tokenszáma (a nem magyar nyelvű részek elhagyásával) összesen 1525.

[2] $i < j$, $k < l$ karakterek közötti pozíciókat jelölő indexek, $j = i+1$, $l = k+1$ esetben karakter→karakter megfeleltetést kapunk.

A betűhű lejegyzés normalizálásánál két alapvető szempontot tartottunk szem előtt: az egységességet, és ugyanakkor az eredetihez való hűséget legalábbis a morfoszintaktikai reprezentáció szintjén. A normalizált alaknak alkalmasnak kell lennie arra, hogy automatikus morfológiai elemzést végezzünk rajta, ezért az erre a reprezentációs szintre való leképezésnél azokat a helyesírási és hangtani különbségeket neutralizáltuk, amelyek az egyébként azonos szóalakokat (ugyanazon lexikai szó ugyanazon morfoszintaktikai jegyekkel bíró előfordulásait) az eredeti szövegekben véletlenszerű módon megkülönbözteti. Hogy a normalizálást a lehető legegyszerűbb legyen megvalósítani, az automatikus elemzéshez használandó morfológiai elemző elkészítése minél kevesebb adaptációs munkát igényeljen, és minél kevesebb bizonytalansági tényező legyen a leképezés során, a normalizált alakok formáját úgy határoztuk meg, hogy azok a lehető legnagyobb mértékben kövessék a mai magyarban érvényes helyesírási konvenciókat.

A korpusz alapesetben mintegy 10000 $M_i^j \rightarrow E_k^l, j = i + 1,\ l = k + 1, j = l$ 1-1 megfeleltetést tartalmaz, továbbá nem egyenlő hosszú egymásnak megfelelő sztringek esetén olyan nem hosszúságtartó leképezéseket is, ahol a leképezés megfelelő oldalán üres szimbólum áll. A kiinduló leképezéseket kiterjesztjük olyan továbbiakkal, ahol a két oldalhoz konkatenáljuk adott N szomszédos leképezésből származó szimbólumokat. Körülbelül 7000 kiterjesztés adódik az eredeti megfeleltetésekhez. Az üres szimbólumot tartalmazó leképezések önmagukban nem, csak az összevont leképezésekben szerepelnek. Példaképpen legyen $N = 3, M = \mathrm{te}, E = \mathrm{the}$, ekkor az alábbi kiinduló leképezések kerülnek a tanítókorpuszba:

$$\mathrm{t} \rightarrow \mathrm{t}$$
$$\epsilon \rightarrow \mathrm{h}$$
$$\mathrm{e} \rightarrow \mathrm{e}$$

melyekből továbbá az alábbi helyettesítések generálódnak:

$$\mathrm{t} \rightarrow \mathrm{th}$$
$$\mathrm{e} \rightarrow \mathrm{he}$$
$$\mathrm{te} \rightarrow \mathrm{the}$$

A tanítókorpusz manuális előállítását gépi eszközökkel támogattuk. Automatikusan előállítottunk egy olyan változatot, ahol a régi szöveg karakterszinten közelítőleg párhuzamosítva volt a modern szöveggel. Ezt már csak javítani kellett kézzel, így nagy mértékben csökkent a manuális munkaigény. A Prószéky-kóddal kódolt régi szövegek esetében természetesen egy karakternek vettük a különféle Prószéky-kódokat (pl. 'y2', 's43'). A kimenet pontosságának javítása érdekében a következő heurisztikákat alkalmaztuk:

- ha a Prószéky-kód betűje egyezett a mai betűvel, elfogadtuk jó illeszkedésnek
- ha a jelen karakterpár nem egyezett, de a következő igen, akkor elfogadtuk ezt az eltérést az illeszkedésben
- ezt kiterjesztettük két egymás utáni nem egyező karakterpár esetére is

– ha a jelen karakterpár nem egyezett, de vagy a régi vagy a mai szövegben alkalmazott egy elcsúsztatással egyezést találtunk, akkor megfelelően beillesztettünk egy $\epsilon \rightarrow k$ vagy $k \rightarrow \epsilon$ illeszkedést, és csak az egyik szövegben léptünk tovább egy karakterrel.

Ezután az egyes helyettesítések valószínűsége a következőképpen számítható:

$$P(\alpha \rightarrow \beta) = \frac{C(\alpha \rightarrow \beta)}{C(\alpha)} \tag{5}$$

$C(\alpha \rightarrow \beta)$ a tanítókorpuszban látott $\alpha \rightarrow \beta$ helyettesítések, $C(\alpha)$ pedig az α sztring előfordulásainak száma.

4.2. A forrásmodell

A forrásmodell mintegy 10 millió szóból, 65 millió karakterből készült az MNSZ egyik alkorpuszából. Ilyen mennyiségben karakter alapú modelleknél különösebb jelentősége a szöveg regiszterének nincsen, ez a modell paramétereit lényegesen nem befolyásolja. Ugyancsak kevésbé sarkalatos kérdés ilyenkor az alkalmazott simító eljárás. A modell építésénél a CMU nyelvmodell készletet használtuk (Clarkson és Rosenfeld, 1997), és az alapbeállítású Good-Turing simítást alkalmaztuk (más eljárás kiválasztása nem változtatott az eredményen, így maradtunk az alapbeállításnál).

5. A modell alkalmazása

Adott E eredeti sztring esetén az $\mathrm{argmax}_M P(E|M)P(M)$ értéket kell kiszámítanunk. Ennek általunk alkalmazott (jelenleg teljesen nem optimalizált) módja a következő. Az eredeti szöveg minden partíciójából a transzliterációs modell helyettesítéseiből a lehetséges modern változatokat legeneráljuk, melyekhez a modell hozzárendeli a valószínűségüket is. Ennek alapján kapunk egy rangsort a kapott változatokra, amit aztán a nyelvmodell segítségével újrarendezünk, így alakul ki a az eljárás végleges kimenete.

5.1. Kiértékelés

A projekt kezdeti szakaszában egyelőre csak előzetes eredmények állnak rendelkezésre. Ennek illusztrációja a 2. ábrában látható. Az alkalmas kiértékelési módszer legjobb n-es listák vizsgálata, és ezekben a pontosság vizsgálata (a fedés ebben az esetben nem hordoz újabb információt). A módszer valós használhatósága abban mutatkozik meg, hogy a manuális annotáció redukálható a felkínált alakok közötti választásra, ami jelentősen felgyorsítja a szövegnormalizálás elkerülhetetlen kézi ellenőrzését. Kézenfekvő, hogy az alapmodell kiegészíthető az egyes tokenek feletti szóalapú n-gram nyelvmodellel, és a kimenet szűrhető illetve átrangsorolható morfológiai elemzés segítségével.

fwl (fül)=>		ygen (igen)=>	
-8,80780895229285	föl	-10,8729908279143	igén
-10,7227286786192	fel	-11,3178857141749	igen
-11,0558158154337	fül	-11,5989613202567	igény
-11,2756412387919	föl	-13,4229320257043	igyen
-12,4574295350367	fol	-14,3578433608162	igin
-12,790296695296	ful	-14,478835649955	igyén
-13,519092302452	fely		
honneg (honnét)=>		sabach (szabács)=>	
-19,1117218113907	honneg	-17,2582527599661	szabács
-19,5230300429664	honnég	-18,1187648297282	sabács
-20,8376176340216	honnét	-18,6771909747334	szabacs
-21,8538140705439	honyneg	-19,1848409742852	sábacs
-22,2098585020436	honynég	-19,5520665992527	szabach
-22,5639991398073	hónneg	-19,9685260661797	szabách

2. ábra. Legjobb n listák különböző bemenetekre.

6. Összefoglalás és további feladatok

A dolgozatban megmutattuk, hogy egyszerű sztochasztikus modellek miként alkalmazhatók két reprezentációs szint közötti fordítási feladatra. A további kutatásban számos újabb, a 2. részben említett gépi tanulási módszer alkalmazására van lehetőség (Chen, 2003; Marchand és Damper, 2000; Taylor, 2005), melyek kiértékelése megalapozottan kimutathatja, hogy a vizsgált modellek között melyik a leghatékonyabb, ezzel együtt pedig választ adhat arra a nagyon fontos gyakorlati kérdésre, hogy a manuális átírás hatékonyan kiváltható-e gépi eljárással, így a szükséges emberi erőforrás alkalmazása leszűkíthető-e a tanuló adatok előállításának feladatára illetve minimális kézi ellenőrzésre. Az itt használt megközelítés is számos részletében finomítható, így a szóhatárok kezelésére illetve a lehetséges partíciók feletti eloszlásra is kidolgozható modell, és természetesen a jelenlegi implementáció hatékonysága is nagy mértékben növelhető.

Irodalomjegyzék

Aw, AiTi, Zhang, Min, Xiao, Juan és Su, Jian. A phrase-based statistical model for SMS text normalization. In: *Proceedings of the COLING/ACL*, Sydney, Australia. Association for Computational Linguistics, 2006, 33–40.

Bellegarda, Jerome R. Unsupervised, language-independent grapheme-to-phoneme conversion by latent analogy. *Speech Communication*, 2005, 46(2): 140–152.

Brill, Eric és Moore, Robert C. An Improved Error Model for Noisy Channel Spelling Correction. In: *ACL-00*, Hong Kong. 2000, 286–293.

Chen, Stanley F. Conditional and Joint Models for Grapheme-to-Phoneme Conversion. In: *EUROSPEECH-03*, 2003.

Clarkson, P. R. és Rosenfeld, R. Statistical language modeling using the CMU-Cambridge toolkit. In: *EUROSPEECH-97*, 1. kötet, 1997, 2707–2710.

Damper, Robert I., Marchand, Yves, Adamson, M. J. és Gustafson, K. Evaluating the pronunciation component of text-to-speech systems for English: A performance comparison of different approaches. *Computer Speech and Language*, 1999, 13(2):155–176.

Haader, Lea. A Müncheni emlék. *Magyar Nyelv*, 2005, (101):161–178.

Imre, Samu. *A Szabács Viadala*. Aladémiai Kiadó, Budapest, 1958.

Kernighan, Mark D., Church, Kenneth W. és Gale, William A. A Spelling Correction Program Base on a Noisy Channel Model. In: *COLING-90*, II. kötet, Helsinki. 1990, 205–211.

Kiss, Gabriella, Kiss, Margit és Pajzs, Júlia. Normalisation of Hungarian Archaic Texts. In: *Proceedings of COMPLEX 2001*. University of Birmingham, 2001, 83–94.

Kobus, Catherine, Yvon, François és Damnati, Géraldine. Normalizing SMS: are two metaphors better than one? In: *Proceedings of the 22nd International Conference on Computational Linguistics*, 1. kötet, Manchester, United Kingdom. Association for Computational Linguistics, 2008, 441–448.

Lucassen, J. és Mercer, Robert L. An information theoretic approach to the automatic determination of phonemic baseforms. In: *ICASSP-84*, 9. kötet, 1984, 304–307.

Marchand, Yves és Damper, Robert I. A multi-strategy approach to improving pronunciation by analogy. *Computational Linguistics*, 2000, 26(2):195–219.

McEnery, Tony és Hardie, Andrew. Lancaster Newsbooks Corpus, 2003. http://www.lancs.ac.uk/fass/projects/newsbooks/default.htm.

Raghunathan, Karthik és Krawczyk, Stefan. Investigating SMS Text Normalization using Statistical Machine Translation. Stanford University, 2009.

Shannon, C. E. A Mathematical Theory of Communication. *Bell System Technical Journal*, 1948, 27(3):379–423.

Taylor, Paul. Hidden Markov Models for Grapheme to Phoneme Conversion. In: *INTERSPEECH-05*, Lisbon, Portugal. 2005, 1973–1976.

Toutanova, Kristina és Moore, Robert C. Pronunciation Modeling for Improved Spelling Correction. In: *ACL-02*, Philadelphia, PA. 2002, 144–151.