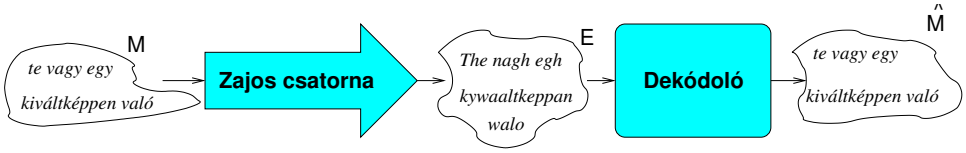
Gépi tanulási módszerek ómagyar kori szövegek normalizálására

Oravecz Csaba[1]

MTA Nyelvtudományi Intézet

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 fonetikus átírásban kiadott szövegek mai modern helyesírású változatának előállítása. 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.

Az előadás 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 eredményezik. Ennek érdekében szükség van az adott modellben használt jegyeket tartalmazó specifikusan annotált tanító korpuszokra. További nehézséget jelent, hogy az egyes nyelvemlékek nagyon különböző fonetikusan átírt reprezentációval rendelkeznek, ezért kritikus kérdés az, hogy a tanult modellek milyen mértékben általánosíthatók az eltérő nyelvemlékekre. Egy-egy nyelvemléken belül sem egységes a reprezentáció, és természetesen elírások is nehezítik illetve lehetetlenné teszik egyértelmű konverziós szabályok meghatározását. Mindezek miatt célszerű a problémát valamilyen valószínűségi alapú paradigma keretei között vizsgálni, egyik legkézenfekvőbb erre Shannon zajos csatorna modellje (Shannon, 1948).



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 fonetikus átiratát. A dekódoló feladata annak az M karaktersorozatnak a megtalálása, melyre a $P(M|E)$ feltételes valószínűség maximális,

$$\hat{M} = \underset{M}{\operatorname{argmax}}\, P(M|E)$$

[1]Sass Bálinttal és Simon Eszterrel közös kutatás.

illetve a szokványos átalakítással:

$$\hat{M} = \underset{M}{\operatorname{argmax}} \frac{P(E|M)P(M)}{P(E)} = \underset{M}{\operatorname{argmax}}\, P(E|M)P(M)$$

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 használunk, ahol vizsgáljuk a módszer pontosságát N függvényében. 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.

A modell alkalmazásakor tehát adott E eredeti sztring esetén az $\operatorname{argmax}_M P(E|M)P(M)$ értéket kell kiszámítanunk. Ennek egyik lehetséges módja, hogy 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 az eljárás kimenete. A keresési tér bejárásakor természetesen a szokásos optimalizáló eljárások alkalmazhatók.

Az elsődleges kimenetként kapott eredmények[3] további tanulómodellek kombinációjával jelentősen javíthatók. Az eddig alkalmazott döntési fa és maximum entrópia tanulók segítségével a kombinált osztályozó pontossága legjobb 3-as lista esetén megközelíti a 90%-ot, mely elegendően jó eredmény ahhoz, hogy a rendszer a gyakorlatban is használható legyen. A részletes kiértékelés megmutatja, hogy milyen mennyiségű tanuló adat előállítására van szükség az elfogadható eredmény érdekében, illetve választ kapunk az általánosíthatóság kérdésére is.

Hivatkozások

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

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

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

[3] Legjobb n-es lista módszerrel kiértékelve 75%-os pontosság pl. $n = 5$ esetén.