Statisztikai lemmatizáló módszerek elemzése és továbbfejlesztése magyar nyelven

Berkecz Péter József
III. évf. programtervező informatikus BSc

Témavezetők:

Dr. Farkas Richárd és Szántó Zsolt

SZTE TTIK Számítógépes Algoritmusok és Mesterséges Intelligencia Tanszék

A lemmatizáló (szótári alakra, ragozatlan alakra hozó) a természetes-nyelvfeldolgozási rendszerek egyik fontos eleme, főleg a magyarban, mint agglutináló nyelvben. Például ennek segítségével kereshetünk hatékonyan az interneten, vagy érthetnek meg minket egyszerűbben a számítógépek, ami a mai világban nélkülözhetetlen. Könnyen tűnhet úgy, hogy egyszerűen leválaszthatnánk a toldalékokat, azonban ez nem ilyen egyszerű. Napjainkban a számítógépek felgyorsultak, így elérhetővé váltak a statisztikai alapú lemmatizálók. Korábban kézzel készített, nyelvészeti szabályok alapján határozták meg a lemmákat, azonban ez több okból is hátrányos, például nem tudja lekövetni a nyelv újulását, vagy a különböző doménbeli eltéréseket. Dolgozatomban ismertetem a lemmatizálást és annak különböző módjait. Megmutatom és összehasonlítom, hogy milyen polcról levehető eszközöket használhatunk erre a feladatra, illetve ezek milyen módszereket alkalmaznak. Továbbá bemutatom a jelenleg elérhető magyar nyelvű szövegfeldolgozásra szolgáló adatbázisokat, a szótövező algoritmusok kiértékelésének kihívásait.

Célom az, hogy a magyar nyelvre egy olyan state-of-the-art lemmatizáló álljon rendelkezésünkre, ami beleilleszkedik egy ipari felhasználásra alkalmas keretrendszerbe és meghaladja a többi rendszer pontosságát. Ehhez kiemelten ismertetek két, különböző típusú, de teljesen statisztikai alapú lemmatizálót, és ezekhez specifikusan javaslok néhány egyszerű heurisztikát, amivel tovább növelhetünk a pontosságukon. Elsőként a Lemmy lemmatizálót mutatom be, ami automatikusan készít szabályokat a látott példákról, majd statisztikai alapon alkalmazza ezeket. Továbbá az Edit Tree Lemmatizert, ami a tanulóadatokon úgynevezett szerkesztőfákat hoz létre és egy neurális háló segítségével kiválasztja, majd alkalmazza azokat.

Az eszközölt módosításokkal kicsivel több, mint 4,5%-ot javult összességében a lemmatizáló, ami ezen az architektúrán jelenleg state-of-the-art.