Az igekötők gépi annotálásának problémái

Elméleti háttér, problémafelvetés, célok, hipotézisek

Az automatikus szófaji egyértelműsítésben komoly hibaforrást jelentenek a homográf szavak. Kiugróan magas a tévesztések aránya az egyszótagú igekötők esetén: a *meg* például kötőszóként és igekötőként is nagyon gyakori és gyakran félreelemzett. A kutatásom olyan megoldást ad a problémára, amely – a munka jelen szakaszában – a rossz annotációk 66 százalékát képes kiszűrni.

Az igekötők három esetben kaphatnak hibás szófaji címkét:

(1) Homonim párjuk van a magyarban. Bizonyos igekötők (pl. *meg, ki*) egybeesnek egy élesen elkülönülő szófajú szóval, míg a legtöbb igekötő – nyelvészeti szempontból – hozzá közel álló szófajjal van átfedésben. Ilyenek az idiómarészből kialakult igekötők (pl. *tönkre*), amelyek sokszor esetragos főnévként annotáltak, valamint az olyanok, amelyek határozószóként, névutóként is funkcionálnak (pl. *együtt*). Számos ilyen szó igekötői státuszáról megoszlik a nyelvészek véleménye (ld. Komlósy 1992), a bizonytalanságot az annotáció következetlensége is tükrözi.

(2) Más nyelvek szavaival homográfak. Gyakori eset, hogy az angol *be* létige, a spanyol *el* és a francia *le* határozott névelők igekötőként jelennek meg a korpuszban.

(3) Elírás okozza a többértelműséget. Ez lehet felesleges szóköz (pl. *meg nézett*), ékezetek hiánya (pl. *fél → fel*) vagy internetes szövegekben gyakori rövidítések (pl. *megláttam → +láttam*).

A kutatásom célja az, hogy korpuszvezérelt módszert alkalmazva választ adjon két kérdésre: (1) Hogyan lehet automatikusan javítani az igekötők gépi annotációját? (2) Mennyire lehet hatékony a többértelműségek feloldása?

Abból a feltevésből indultam ki, hogy a kontextus mintázatai alapján egyértelműsíthető lehet az igekötő-gyanús szavak szófaja.

Módszertan

A Pázmány Korpuszról (Endrédy 2016) lekértem minden olyan mondatot, amely igekötőként annotált szavakat tartalmaz, ezután pedig az olyan mondatokat, amelyek tartalmazzák a vizsgált szóalakokat, viszont valamilyen más szófaji címkével. Az így kigyűjtött alkorpuszok elemzésével készítettem az annotációt korrigáló Python-scripteket.

A *meg*-hez Makrai (2007) már dolgozott ki mintaalapú szabályokat, amelyeket kisebb módosításokkal felhasználtam. Emellett figyelembe vettem a finit igék és elvált igekötőik várható távolságát, valamint lehetséges ige-igekötő kombinációkat is (Kalivoda 2016). Először nyelvészetileg is motivált szabályokat adtam meg, majd gyakorisági adatokat vizsgáltam. Az elkészült scripteket a Magyar Nemzeti Szövegtár 2-es verzióján (Oravecz et al. 2014) teszteltem.

Eredmények

Munkám eredménye egy scriptrendszer, amellyel az igekötők annotációját utólag lehet javítani. A kutatási kérdéseket a következőkkel válaszolom meg: (1) A kontextus mintázatai alapján javítható az igekötők annotációja. A keresett mintázatok aszerint változnak, hogy épp melyik igekötővel és melyik hibatípussal van dolgunk. (2) A hibás annotációt az esetek több mint két harmadában sikerült egyértelműen kiszűrni. Legnehezebbnek azok a szerkezetek bizonyultak, ahol a kérdéses szó (főként a *meg* és *ki*) olyan pozícióban áll, ahol igekötő is várható, és a tagmondat igéjével létező kombinációt alkothat.

Következtetések

A kontextus-alapú szűrés lényegesen javítja az eredeti annotáció minőségét az igekötők esetében. A módszer a jövőben más többértelműségek, így például a múlt idejű finit igék és a befejezett melléknévi igenevek közti különbségtétel alapjául is szolgálhat.

Hivatkozások

Endrédy István (2016): *Nyelvtechnológiai algoritmusok korpuszok automatikus építéséhez és pontosabb feldolgozásukhoz.* Doktori disszertáció. Budapest, Pázmány Péter Katolikus Egyetem, Információs Technológiai és Bionikai Kar.
http://users.itk.ppke.hu/~endis/phd/ei.phd.final.pdf

Kalivoda Ágnes (2016): *A magyar igei komplexumok vizsgálata.* Mesterszakos szakdolgozat. Budapest, Pázmány Péter Katolikus Egyetem, Bölcsészet- és Társadalomtudományi Kar.
http://nydi.btk.ppke.hu/sites/default/files/profil_files/Magyar_igei_komplexumok_MA_szakdolgozat_0.pdf

Komlósy András (1992): Régensek és vonzatok. In: Kiefer Ferenc (szerk.): *Strukturális magyar nyelvtan. 1. Mondattan.* Budapest, Akadémiai Kiadó. 299–527.
http://www.nytud.hu/publ/smny/mondattan.pdf
(Javított magyar nyelvű digitális kiadás, 2015, 274–486)

Makrai Márton (2007): *Többértelműségek magyar mondatok számítógépes elemzésében – a „meg" szó szófajának vizsgálata gyakoriságokkal.* Témalabor dolgozat.
http://math.bme.hu/~makraim/temalabor/20060702/makraim_temalabor_szoveg_2007.pdf

Oravecz Csaba – Váradi Tamás – Sass Bálint (2014): The Hungarian Gigaword Corpus. In: *Proceedings of LREC 2014.* Reykjavík, 2014. május 26-31. 1719–1723.
http://www.lrec-conf.org/proceedings/lrec2014/pdf/681_Paper.pdf