Főnévi csoportok azonosítása szabályalapú és hibrid módszerekkel

Recski Gábor

MTA SZTAKI
Nyelvtechnológiai Kutatócsoport
e-mail: recski@sztaki.hu

Az előadásban először egy a magyar főnévi csoportok azonosítására épített mondattani elemzőt (parser-t) mutatunk be, melynek pontossága megközelíti a hasonló célú, gépi tanuláson alapuló rendszerünk (`hunchunk`, Recski et al. 2009) eredményeit, majd megmutatjuk, hogy a két eszköz egyesítésével a feladat nagyobb pontossággal végezhető, mint az önálló statisztikai alapú rendszerrel.

Az NLTK nyelvtechnológiai programcsomag (Bird 2009) segítségével implementáltuk Kornai (1985) környezetfüggetlen nyelvtanát. A nyelvtan szabályai nem csupán a szavak szófajára, hanem azok számos morfológiai jegyére is képesek hivatkozni – a KR-formalizmus által kódolt jegyek pedig egyértelműen lefordíthatóak a nyelvtan által hivatkozott jegyekre. Az így létrejött parser azonosít minden olyan szósorozatot egy mondatban, melyet az NP-nyelvtan elfogad – ezek alapján a mondat diszjunkt főnévi csoportokra osztását néhány egyszerűbb heurisztikus lépés sorozatával végezzük.

A parsert a Szeged Treebank (Csendes et al. 2005)-ből készült NP-korpuszon értékeltük ki. Módosítások nélkül az eredeti nyelvtan használatával 81.76%-os F-pontszámot érünk el, mely még jócskán elmarad a `hunchunk` rendszer 94.75%-os eredményétől. Ezt követően a szabályrendszert számos különböző módon bővítjük: egyrészt néhány egyszerű szabállyal nyelvtant írunk a melléknévi, határozói és számnévi csoportok azonosítására, majd kezeljük a magyar NP-k több olyan csoportját, melyet az eredeti nyelvtan nem ismert fel (ezekre a bővítésekre az előadásban több példát is mutatunk). A végső rendszer 89.36%-os F-pontszámot ér el a Szeged Korpuszon. Ez továbbra is elmarad a gépi tanuló rendszer teljesítményétől, de lehetővé teszi, hogy megpróbálkozzunk hibrid rendszer kifejlesztésével.

A főnévi csoportok szabályalapú azonosítását a teljes Szeged Korpuszon elvégeztük, majd a `hunchunk` rendszert ezen a korpuszon tanítottuk úgy, hogy a parser kimenetének megfelelő chunk-címkéket (B-NP, I-NP) felvettük a tanulóalgoritmusnak átadott jegyek közé. Az így megalkotott vegyes rendszer 95.48%-os F-pontszámot ér el, a parser által szolgáltatott jegyek tehát kb. 15 százalékkal csökkentették a statisztikai alapú rendszer hibáinak számát. A parser-jegyek a `hunchunk` teljesítményét annak másik feladatán, a maximális NP-k azonosításán is csaknem fél százalékponttal javították. A két rendszer egyesítésére – és mindkettőnél magasabb pontosságú hibrid rendszer létrehozására további lehetőségek is nyílnak, az előadásban ezek közül is említünk néhányat.

Hivatkozások

1. S. Bird, E. Klein, and E. Loper. *Natural language processing with Python.* O'Reilly Media, 2009.
2. D. Csendes, J. Csirik, T. Gyimóthy, and A. Kocsor. The Szeged Treebank. In *Lecture Notes in Computer Science: Text, Speech and Dialogue*, pages 123–131, 2005.
3. A. Kornai. The internal structure of Noun Phrases. *Approaches to Hungarian*, 1:79–92, 1985.
4. G. Recski, D. Varga, A. Zséder, and A. Kornai. Fonevi csoportok azonosítása magyar-angol párhuzamos korpuszban [Identifying noun phrases in a parallel corpus of English and Hungarian]. *VI. Magyar Számitógépes Nyelvészeti Konferencia [6th Hungarian Conference on Computational Linguistics]*, 2009.