Nyelvészeti problémák a névfelismerés területén

Az *információkinyerés* (information extraction) a számítógépes nyelvészet egyik fontos és mostanában elég felkapott alterülete. Célja, hogy a számítógép által olvasható, ámde strukturálatlan szövegből gépi eszközökkel, automatikusan információt nyerjünk ki. Egy információkinyerő rendszer feladata, hogy automatikusan adatbázisba rendezze ezeket az adatokat, amely így már használható az adatok analizálására, összegzést kaphatunk belőle természetes nyelvi jelenségekről, vagy bármilyen online eszköz bemenetéül szolgálhat. A feladatok igen tág köre tartozik ez alá, de a lényeg, hogy hatalmas mennyiségű szöveg átnyálazása helyett csak a számunkra fontos, specifikált információt kapjuk meg, ami manapság az interneten található beláthatatlan mennyiségű információt tekintve nem tűnik rossz ötletnek.

Az információkinyerés egyik alfeladata a *névfelismerés*. Az angol terminus *(named entity recognition)* nehezen lefordítható, a magyar szakterminológiában is inkább az angol kifejezést használjuk. Lényegében a szövegben található olyan elemek megkereséséről van szó, amelyek a világ valamely entitására egyedi módon: névvel, becenévvel, mozaikszóval vagy rövidítéssel referálnak. A névfelismerés két fő lépésből áll: először lokalizálni kell a szövegben a nevet, aztán besorolni egy előre definiált névosztályba. Tipikus ilyen kategóriák: a személy-, a földrajzi, az intézménynév, a dátumok és egyéb időre referáló kifejezések, valamint a különböző mennyiségeket jelölő elemek.

Egy szöveg nyelvi elemzése általában azzal kezdődik, hogy a szöveg szavait főnévként, igeként stb. azonosítjuk szótárak segítségével. Viszont a legtöbb szöveg tartalmaz neveket, amelyeket nem tud értelmes nyelvi egységként azonosítani a rendszer. Így tehát a névfelismerés nélkülözhetetlen lépése bármilyen szöveg nyelvi elemzésének, az eseménykivonatolásnak (event extraction) és a gépi fordításnak.

Előadásomban a névfelismerés során jelentkező nyelvészeti problémákat szeretném felvázolni, illetve megpróbálok rájuk egzakt válaszokat találni. Egy jellemző kérdése a névfelismerésnek, hogy a metonimikus kifejezéseket a szövegben felvett aktuális jelentésük vagy általában vett referensük alapján jelöljünk. Erre szolgáltatnak tipikus példát az olyan nevek, amelyek jelentenek egy épületet, egy intézményt és egy közösséget is, mint például az iskolák és a múzeumok. Ha a *tag-for-meaning* elvét követjük, vagyis minden nevet az aktuális kontextusának megfelelően annotálunk, akkor ugyanazt a tulajdonnevet különböző szöveghelyeken különböző névosztályokba soroljuk. Hasonlóan problémás a többi metonimikus kifejezés is, például amikor személynév áll egy szervezet neve vagy egy mű címe helyett.