A HunNER Korpusz

Simon Eszter[1], Farkas Richárd[2], Halácsy Péter[1],
Sass Bálint[3], Szarvas György[2], Varga Dániel[1]

[1] BME MOKK
{daniel, halacsy}@mokk.bme.hu, esimon@cogsci.bme.hu
[2] Szegedi Tudományegyetem, Informatika Tanszékcsoport
{szarvas, farkas}@inf.u-szeged.hu
[3] MTA Nyelvtudományi Intézet
joker@nytud.hu

Kivonat: Cikkünkben egy folyamatban lévő projektet mutatunk be, melynek célja egy nagyméretű, manuálisan tulajdonnév-annotált korpusz létrehozása. A tervezett korpusz jól használható lesz gépi tanuláson alapuló tulajdonnév-címkézők tanítására és szabványos kiértékelésére, miközben elő- és utófeldolgozó eszközöktől független. A projektet a BME MOKK, a Nyelvtudományi Intézet és az SZTE közösen indította. A projekt fontos mellékterméke egy olyan klasszifikációs útmutató magyar nyelvre, amely időtálló, és a fenti intézmények közötti konszenzuson alapul. Az elkészült korpusz a konzorcium döntése alapján szabadon hozzáférhető lesz kutatási célokra.

1 Annotációs séma és útmutató

A projekt egyik fontos célja kialakítani egy egységes annotációs útmutatót. A konzorciumi tagok által eddig használt útmutatók között lényeges eltérések vannak. Ezek szabályait közös munkával konzisztens rendszerré ötvöztük.

Az annotálás során mindig szem előtt tartandó elveink a következők:

- Névnek nevezzük azt a kifejezést, ami unikusan, vagyis egyedi módon referál a világ valamely entitására. Tehát nem annotálunk olyan frázisokat, amelyek ugyan a világnak valamely egyedi részére utalnak, de nem teljesen egyértelmű módon.
- Nem annotálunk egymást átfedő vagy egymásba ágyazott neveket. Vagyis minden annotációnak be kell fejeződnie, mielőtt egy másik elkezdődik.
- Mivel a nevek nem kompozicionálisak, tehát jelöletük nem a részeik jelöletéből áll össze, ezért a neveket nem bonthatjuk részekre az annotáláskor. Például a *Kossuth Lajos utca* egy névként jelölendő, hiába van benne egy személynév. Mindig a leghosszabb nevet (a legkülsőbbet) jelöljük a jelölhetők közül.
- Az inflexiókról hagyományosan azt szoktuk gondolni, hogy nem vagy csak elhanyagolhatóan minimális mértékben változtatják meg a tulajdonnév "jelentését", vagyis ugyanarra utalnak, mint a toldalék nélküli alakok. Ezért ha az azonosított tulajdonnév ragozott formában szerepel a szövegben, a raggal

együtt, a teljes alakot annotáljuk. A képzők közül viszont csak néhányról gondoljuk ezt, ezért a képzett alakokat nem jelöljük, kivéve a földrajzi névből *-i/-beli* képzőkkel képzett melléknéveket.

A leginkább vitatott kérdéseink megegyeznek a tulajdonnév-klasszifikáció kapcsán nemzetközi szinten is felmerülő kérdésekkel. Ilyen például a *tag-for-meaning* elve, melyet követve a tulajdonneveket aktuális szövegbeli kontextusuk alapján osztályozzuk. Ezzel kapcsolatban problémák elsősorban a metonímiák esetében állnak elő, amikor valamely típusú entitással egy másik típusú entitásra utalunk; illetve az olyan neveknél, amelyek több dologra: épületre, emberi közösségre és intézményre is utalhatnak, mint a múzeumok, iskolák vagy színházak. A nevek metonimikus használatára a legjellemzőbb példát a szervezetre referáló helynevek, illetve a helyre referáló szervezetnevek adják. Az *A János kórházban sok a macska* példamondatban a *János kórház* egy intézménynek a neve ugyan, de ebben a kontextusban helyet jelöl. Ugyanígy: a *Washington Moszkvával tárgyal* mondatban mindkét helynév valamilyen szervezetre referál. A tag-for-meaning elvet követve a *János kórház*at helynévként, *Washington*t és *Moszkvá*t pedig szervezetnévként kellene annotálnunk. Egy másik lehetséges annotálási mód szerint egy típusú nevet kontextustól függetlenül, eredeti jelöletének megfelelően kell jelölni (*tag-for-tagging*). Ennek a konfliktusnak a kiküszöbölésére bevezettünk két új alkategóriát, melyekkel jelölni tudjuk a metonimikus használatot, és egyben lehetővé tesszük minden újrafelhasználó számára a neki tetsző elv követését.

A kialakításnál figyelembe vettük azt a szempontot is, hogy az általunk használt annotációs séma kompatibilis legyen nemzetközileg elfogadott tulajdonnév-klasszifikáló sémákkal. Ezek közül a számunkra legfontosabbak a Szeged NER korpusz [1] építéséhez már adaptált CoNLL [2,3], valamint a Linguistic Data Consortium által alkalmazott [4] sémák. Ezek alapján a korpuszunkban jelölendő típusok:

a személynevek (PERSON),
az embereknek valamely szervezetnél betöltött szerepét jelölő frázisok (ROLE),
a cím- és rangjelölő szavak (RANK),
a szervezetnevek (ORGANIZATION),
a helynevek (LOCATION),
a szervezetre referáló helynevek (ORG:LOC),
a helyre referáló szervezetnevek (LOC:ORG),
a márkanevek (BRAND/PRODUCT),
a műcímek (TITLE) és
egyéb tulajdonnevek, vagyis amelyek nem tartoznak a fenti kategóriák egyikébe sem, de tulajdonnevek (MISC).

2 Külső annotáció

A korpusz minden feldolgozási lépésében (tokenizálás, mondatra bontás, tulajdonnév-címkézés) külső (*standoff*) annotációt használunk. Ennek lényege, hogy az eredeti dokumentumokat sima szövegként rögzítjük, és az annotációkat nem beágyazott markupként, hanem egy külső fájlban jelöljük úgy, hogy megadjuk, hogy az eredeti szöveg melyik karaktertartományára vonatkozik a címkézés, és hogy milyen címkét

kap a szövegrészlet. A külső annotáció előnye, hogy az annotálást teljesen különválasztja a használt feldolgozó eszközöktől, és minden formai információ elérhető a feldolgozottság minden fázisában.

Páros számú annotátorral dolgozunk, és az annotátorok közötti egyetértést mérjük:

$$\frac{2 * |\text{ugyanúgy jelölt entitások}|}{|\text{A annotátor által jelölt entitások}| + |\text{B annotátor által jelölt entitások}|}$$

Az annotátorok mindig csak a rögzített útmutató alapján dolgozhatnak, amit a menet közben felmerülő problémás kérdések megvitatásával folyamatosan fejlesztünk, egész addig, amíg az egyetértés 95% feletti nem lesz.

3 A korpusz forrásai

A korpusz méretének lehetővé kell tennie, hogy az azon tanított statisztikus tulajdonnév-felismerő modellek általános szövegen is megállják a helyüket, és specifikus szövegen is jól tudjanak működni. Számításaink szerint legkevesebb félmillió szövegszó címkézése teszi lehetővé, hogy a korpuszban megfelelő méretű részkorpuszok legyenek.

A korpusz elsődleges forrása magyar nyelvű valódi hírek teljes szövege. A korpusz téma szerinti eloszlása a következő: gazdaság (100 ezer szövegszó), sport, belföldi politika, nemzetközi politika, törvények/rendeletek, tudomány/technika, fórum/blog, szoftverkézikönyvek, filmszövegek/szépirodalom (50-50 ezer). A gépi fordítási alkalmazásokra tekintettel a szövegeket úgy választjuk ki, hogy minden műfajban legalább 1/5 rész angolból magyarra fordított szöveg legyen.

4 Jogok

Alapvető cél, hogy a létrejött korpuszt bárki teljesen szabadon használhassa, és kiegészíthesse további standoff annotációkkal.

Bibliográfia

1. György Szarvas, Richárd Farkas, László Felföldi, András Kocsor, János Csirik: A highly accurate Named Entity corpus for Hungarian. Proceedings of International Conference on Language Resources and Evaluation 2006.
2. Nancy Chinchor, Erica Brown, Lisa Ferro, Patty Robinson: Named Entity Recognition Task Definition. MITRE and SAIC. 1999.

3. Erik F. Tjong Kim Sang, Fien De Meulder, Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In: *Proceedings of CoNLL-2003*, Edmonton, Canada, 2003, pp. 142-147.
4. http://projects.ldc.upenn.edu/LCTL/Specifications/SimpleNamedEntityGuidelinesV6.5.pdf