A HunNER korpusz

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

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

Absztrakt

A tulajdonnév-felismerés (named entity recognition, NER) során egy bemeneti tokensorozatban kell tulajdonnevet alkotó intervallumokat kijelölnünk, ezeket véges sok kategóriába besorolva. Egy NER algoritmus kiértékelése manuálisan annotált korpusszal való összevetés útján történik, és szokásosan maga az algoritmus is egy ilyen korpuszból tanulja meg a paramétereit automatikus módon. Jelen poszterünkkel egy olyan 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 melléktermék 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. Bevezetés

A tulajdonnév-felismerés (*named entity recognition, NER*) a természetes nyelv feldolgozását célzó alkalmazások közül az egyik legnépszerűbb, mivel hatékonyan automatizálható, és eredménye hasznos bemenete különböző magasabb szintű információ-kivonatoló és információ-feldolgozó rendszereknek.

A NER során egy bemeneti tokensorozatban kell tulajdonnevet alkotó intervallumokat kijelölnünk, ezeket véges sok kategóriába besorolva. Egy NER algoritmus kiértékelése manuálisan annotált korpusszal való összevetés útján történik, és szokásosan maga az algoritmus is egy ilyen korpuszból tanulja meg a paramétereit automatikus módon.

Tulajdonnév-címkéző rendszerek fejlesztéséhez tehát elengedhetetlenül szükséges egy kellően nagy méretű, tematikusan heterogén, konzisztens annotálási szabályzaton alapuló, manuálisan feljelölt korpusz. Ennek létrehozására indult ez a projekt a három konzorciumi tag: a BME MOKK, a Szegedi Tudományegyetem Informatika Tanszékcsoportja és az MTA Nyelvtudományi Intézete részvételével.

2. 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 jelöletét, 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 mellékneveket.

Az annotációs séma kialakításánál 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).

2.1. Problémás esetek

A leginkább vitatott kérdéseink megegyeznek a tulajdonnév-klasszifikáció kapcsán nemzetközi szinten is felmerülő kérdésekkel, amilyen például a metonimikus névhasználat esete. Ilyen esetekben referenciaátvitel történik: a név nem az eredeti referensre utal, hanem egy másikra. Tipikus példáink az alábbiak:

Intézménynevek: eredeti referenciájuk az intézmény, de gyakran jelölnek helyet, illetve emberi közösséget.

Az [Eötvös József Gimnázium$_{INTÉZMÉNY}$] nem kap elegendő állami támogatást.
Nincs messze tőlünk az [Eötvös József Gimnázium$_{HELY}$].
Az [Eötvös József Gimnázium$_{KÖZÖSSÉG}$] idén Luxemburgba megy kirándulni.

Helynevek: tipikusan a politikai alapon definiált földrajzi egységek nevei tartoznak ide (országok, városok, megyék stb.), melyek jelölhetnek egy helyet, egy kormányzatot, egy közösséget vagy akár az adott terület iparát is.

[Franciaországot$_{HELY}$] kilenc ország határolja.
[Franciaország$_{KORMÁNYZAT}$] korlátozza a politikai menedékjogot.
[Franciaország$_{KÖZÖSSÉG}$] új elnököt választott.
[Franciaország$_{IPAR}$] bortermelése az idén visszaesett.

Helynévvel (jellemzően városnévvel) utalhatunk sportcsapatok nevére is:

[Manchester$_{HELY}$] ma London után Nagy Britannia második legnagyobb pénzügyi központja.
[München$_{HELY}$] fő közlekedésének jelentős részét bonyolítja az U-Bahn és az S-Bahn hálózat.
A [Manchester$_{SPORTCSAPAT}$] ma a [Münchennel$_{SPORTCSAPAT}$] játszik.

Ezek a metonímiák a fenti esetek mindegyikében egy-egy teljes szemantikai mezőre vonatkoznak, jellemzőek és megjósolhatóak, ezért valamilyen konzekvens jelölési módot kellett rájuk kitalálni. Két elv ismert és használt a NER területén az ilyen problémás esetek kezelésére.

Ha a *tag-for-meaning* elvét alkalmazzuk, akkor a kontextusnak megfelelően, az éppen aktuális referens címkéjét kapja a név. A *tag-for-tagging* elve alapján viszont egy név kontextustól függetlenül mindig ugyanazt a címkét kapja, vagyis a kiinduló referensét.

A HunNER korpuszban a kettő közötti kompromisszumos megoldást alkalmazzuk, vagyis jelöljük a referenciaátvitelt helyről intézményre és intézményről helyre.

Az [Európai Uniónak$_{LOC:ORG}$] 20 országgal van közös szárazföldi határa.
[Washington$_{ORG:LOC}$] [Moszkvával$_{ORG:LOC}$] tárgyal.
A [Fehér Ház$_{ORG:LOC}$] semmi információt nem ad ki.

3. Az annotáció menete

A korpusz szövegein a következő feldolgozási lépéseket végezzük el:

mondatra bontás: a hírek felbontása mondatokra;

tokenizálás: az egységként kezelt szövegelemek azonosítása (tipikus tokenhatároló elemek pl. a szóköz, kötőjel);

tulajdonnév-címkézés: a szöveg tulajdonneveinek azonosítása, osztályozása.

3.1. Minőségbiztosítás

Az annotátorok csak a rögzített útmutató alapján dolgozhatnak, amit a felmerülő problémás kérdések megvitatásával folyamatosan fejlesztünk.

Az útmutató pontosítása addig folytatódik, míg a használatával az annotátorok egyetértési rátája 95% feletti lesz. Páros számú annotátorral dolgozunk, és az egyetértést a következő képlet szerint mérjük:

$$\frac{2*|\text{egyformán annotált NE-k}|}{|\text{az A annotátor által jelölt NE-k}|+|\text{a B annotátor által jelölt NE-k}|}$$

3.2. Külső annotáció

- Az eredeti dokumentumokat egyszerű szövegként rögzítjük.
- Az annotációkat egy külső fájlban tároljuk úgy, hogy megadjuk, hogy az eredeti szöveg melyik karaktertartományára vonatkozik a címkézés.

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

4. A korpusz forrásai

A korpusz elsődleges forrása magyar nyelvű valódi hírek teljes szövege. A korpusz méretének lehetővé kell tennie, hogy:

- az azon tanított statisztikai tulajdonnév-felismerő modellek általános szövegen is megállják a helyüket, illetve hogy
- specifikus szövegen is kiemelkedően tudjanak működni.

Ezt legkevesebb félmillió szövegszó címkézése teszi lehetővé.
A korpusz téma szerinti eloszlását a 1 táblázat mutatja.

Téma	Szövegszó
gazdaság	100.000
sport	50.000
belföldi politika	50.000
nemzetközi politika	50.000
törvények / rendeletek	50.000
tudomány / technika	50.000
fórum / blog	50.000
szoftverkézikönyvek	50.000
filmszövegek / szépirodalom	50.000

1. táblázat. Az egyes részkorpuszok és méretük

5. Jogok

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

Hivatkozások

[1] Szarvas, G., Farkas, R., Felföldi, L., Kocsor, A., Csirik, J.: A highly accurate named entity corpus for hungarian. In: Proceedings of International Conference on Language Resources and Evaluation (LREC2006). (2006)

[2] Chinchor, N., Robinson, P.: MUC-7 named entity task definition version 3.5. In: Proceedings of the 7th Message Understandig Conference (MUC-7). (1998) Available from http://www-nlpir.nist.gov/related_projects/muc/proceedings/ne_task.html.

[3] Tjong Kim Sang, E.F., De Meulder, F.: Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In: Proceedings of CoNLL-2003, Edmonton, Canada. (2003) 142–147

[4] Linguistic Data Consortium – LCTL Team: Simple Named Entity Guidelines For Less Commonly Taught Languages Version 6.5. (2006) http://projects.ldc.upenn.edu/LCTL/Specifications/SimpleNamedEntityGuidelinesV6.5.pdf.