NAGY VIKTOR

HOGYAN LELJÜNK BARÁTOKAT A KORPUSZBAN?

Dolgozatom azokkal a korpuszlekérdezési problémákkal foglalkozik, amelyekben a keresési feltételek a korpuszban expliciten nem megjelenő tulajdonságokra hivatkoznak. Közülük is elsősorban homonim szópárok ritkábbik tagjának keresésére koncentrálok. Esettanulmányként a 'szerzetes' értelmű *barát* szó előfordulásait keresem a Magyar Nemzeti Szövegtárban.

1. A nyelvészetileg interpretált korpuszok

A nyelvészet számos területén fontos szerepük van a nagy méretű, nyelvészetileg interpretált elektronikus korpuszoknak. Az interpretálás (vagy annotálás) a szöveg kiegészítése morfológiai, szintaktikai, szemantikai vagy diskurzusbeli információkkal. Olyan információkkal, amelyek implicit módon megtalálhatóak a szövegben, de kiaknázásukhoz explicitté kell tenni őket. Ezt a kiegészítő információt, a korpuszannotációt egyrészt belekódolják a gépi szövegbe (általában az XML-jelölőnyelvet alkalmazzák erre), másrészt egy keresőfelületen elérhetővé, lekérdezhetővé teszik a használók számára.

A morfológiai interpretációra egy példa a szövegszók alapalakjának helyreállítása, a lemmatizálás. Egy lemmatizálatlan szövegben egy szótári szó összes előfordulását csak körülményesen tudjuk megkeresni. Az egyik megoldás lehet a csonkolt keresés: keressük a szótővel kezdődő szavakat. Ez egyrészt téves találatokat is eredményez (a *kátyú*-val kezdődő szavak között lesznek a *kátyúzás*, *kátyúmentesítés*, *kátyús*, *kátyúzik*, *kátyúsít*, *kátyúaszfaltozás*, *kátyúgyakorisági*, *kátyúmező* szavak alakjai a Magyar Nemzeti Szövegtárban), másrészt a több tőallomorffal rendelkező szavak nem kereshetők ilyen egyszerűen (lásd pl. *eszik*). A másik megoldás, ha az összes alakot felsoroljuk a lekérdezésben – az annotálás ezt a terhet hivatott levenni a kérdező válláról.

Az annotációnak megnő a jelentősége azokban a jelenségekben, amelyek formailag nem ragadhatóak meg, vagyis amelyeket kereséskor nem lehet megfogalmazni felszíni mintákkal. Már a szótári szó keresésekor is felléphet a homonímia problémája: a *török* a *török* és a *tör*, a *lappal* a *lap* és a *lapp* szótári szavak alakja is lehet.

A kézi annotálás, vagyis a szöveg emberi értelmezése és kódolása csak kis méretű korpuszok esetén lehetséges. A nagyobb, több tíz- vagy százmillió szövegszavas korpuszokon automatikus elemzőprogram állítja elő a korpuszannotációt. Emiatt az automatikus elemzők tudása határt szab az annotálási lehetőségeknek. A Magyar Nemzeti Szövegtár a strukturális információk (bekezdés, mondat, cím stb.) mellett a szövegszók lemmáját, szófaját és morfoszintaktikai kategóriáját nyújtja az annotációban. Az elemzést egy gépi tanulóalgoritmus végezte. Az elemzés nem képes szemantikai különbségeket értelmezni, ezért a szófajon belüli homonímiát nem tudja egyértelműsíteni. Ellenben a szó és a kontextus formai jegyei alapján alacsony hibaszázalékkal meg tudja határozni, hogy az adott szónak mi a szófaja (vagyis pl. a *tűz* igét és főnevet egyértelműsíteni tudja).

2. A gépi homonimaegyértelműsítés

Az MNSZ lemmaszófaj annotációja nem határozza meg egyértelműen, hogy mely szövegszó

mely lexémához tartozik. A magyar lexikográfiai gyakorlat az ugyanolyan szófajú homonimákat önálló szótári szavaknak tekinti. Ezek a szavak önállóan nem, csak homonim társukkal egyetemben kereshetők a MNSZ-ben. Ahhoz, hogy a kereséskor a homonimára is tudjunk hivatkozni, explicitté kell tenni ezt az információt az annotációban, ehhez pedig a homonim szavakat automatikus eljárással egyértelműsíteni kell.

A szójelentés-egyértelműsítés a számítógépes nyelvészet részterülete. Célja automatikus módszerek kidolgozása annak megállapítására, hogy egy szó melyik lehetséges jelentésével fordul elő egy adott kontextusban. Egy szó előfordulásait egy ilyen eljárás vagy nyelvészek által előre felállított, vagy saját maga által felfedezett jelentésosztályokba sorolja. A mi esetünkben alkalmazandó feladat ennél valamivel egyszerűbb: egy szó jó néhány poliszém jelentése helyett két (ritka esetben több) homonima alkotja a jelentésosztályokat a nem egyértelmű esetekben.

A szójelentés-egyértelműsítésre alkalmazott, statisztikai módszereken alapuló (sztochasztikus) eljárások szemantikai felfogása szerint a szójelentést az határozza meg, hogy a szó milyen környezetben fordulhat elő. Egy szó különböző jelentései mögött egymástól elkülönülő kontextusok állnak. A sztochasztikus eljárások a jelentésosztályoknak megfelelő kontextusokat formai jegyekkel jellemzik, vagyis a szemantikai elemzést visszavezetik a formális, morfoszintaktikai elemzésre vagy pusztán a környezet szóalakjaira.

NAGY (2003) egy megoldást mutat be a homonimaegyértelműsítésre. A következő környezeti jegyeket alkalmazza:

- tartalmas lemma jelenléte a bővebb környezetben,
- szóalak jelenléte a szűkebb környezetben,
- inflexiós kategória jelenléte a szűkebb környezetben,
- szomszédos szóalakok,
- szomszédos szópárok,
- az egyértelműsítendő szó alakja.

A sztochasztikus eljárás minden homonim szóhoz egy kézzel egyértelműsített konkordancialistából nyerte ki a döntési modelljét, oly módon, hogy számba vette, hogy a különböző jegyek milyen gyakran járnak együtt a különböző jelentésekkel.

Az eljárás pontossága (70–90% a különböző szavakra) nem teszi lehetővé, hogy az általa adott elemzésekkel az MNSZ annotációját gazdagítsuk. Alkalmazható lenne konkordancia szűrésére, ha egy szónak csak az egyik homonimájára vagyunk kíváncsiak. Az így szűrt lista azonban további átvizsgálásra szorul a módszer viszonylagos pontatlansága miatt.

3. A ritka homonimák problémája

A gépi tanuláson alapuló eljárások sikeres működésének előfeltétele, hogy a program a tanulandó jelenségekből elegendő példával találkozzon. Ha a szemantikai egyértelműsítés esetén a tanulókorpuszban valamelyik jelentés kis számban vagy egyáltalán nem képviselteti magát, a sztochasztikus eljárás a kategorizálás során nem fogja kiosztani azt a jelentést. Ha a mintabeli jelentésarány megfelel az alapsokaság jelentésarányának, akkor az egyértelműsítő eljárás átfogó pontossága nem fog észrevehetően romlani. Ha viszont a célunk az, hogy az egyértelműsítéssel további példákat találjunk egy ritkább jelentésre, a módszer használhatatlan lesz.

A minta növelése annak érdekében, hogy a ritkább homonimára elegendő példát kapjunk, sok esetben reménytelen. Ha a homonim szó előfordulásainak csak egy százalékát teszi ki a ritkábbik homonima, akkor tízezres mintát kell átnézni ahhoz, hogy száz példát

találjunk rá.

Az alábbi táblázat olyan homonimákat mutat be az ÉKsz. szerint, amelynek tagjai közismertek, és legalább az egyik relatív (a többi taghoz képesti) gyakorisága 1% alatt van a vizsgált mintában. A minták az ÉKsz. gyakorisági adatainak meghatározásához az MNSZ-ből készültek, mindegyik homonimapárhoz kétszáz előfordulást tartalmaztak, háromtagú homonimára háromszázat stb.

Homonim szó	**Gyakori homonima (> 1%)**	**Ritka homonima (< 1%)**
aggaszt	1 'aggodalmat kelt'	2 'szikkaszt'
árt	1 'kárt okoz'	2 'beleavatkozik'
barát	1 'társ'	2 'szerzetes'
bocs	2 'bocsánat'	1 'kölyök'
bokszol	1 'öklöz'	2 'cipőt fényesít'
fogoly	1 'rab'	2 'madár'
fok	2 'fokozat, mérési egység' 3 'part kiszögellése'	1 'hegyes eszköz tompa fele'
lakik	1 'tartózkodik'	2 'étellel eltelik'
megér	2 'értéke van'	1 'életben van'
megvesz	1 'vásárol'	2 'kitör rajta a veszettség'
nyár	1 'évszak'	2 'nyárfa'
rét	1 'mező'	2 'hajtogatott rész'
század	2 'évszázad'	1 'századrész' 3 'katonai egység' 4 'rendkívül sok'
szó	1 'nyelvi egység'	2 'zenei hang'
szurkol	2 'drukkol'	1 'szurokkal ken'
ül	1 'alsó részén nyugszik'	2 'ünnepel'
vég	1 'végső hely, időpont'	2 'hosszegység'
vesz	1 'fog'	2 'pusztul'

4. Esettanulmány

Vegyük a *barát*[1] 'társ' és *barát*[2] 'szerzetes' homonimákat. A feladat az, hogy keressünk a második homonimához előfordulásokat az MNSZ-ben. A fenti táblázat szerint a *barát* lemma előfordulásaiból vett kétszázas mintában kevesebb mint kétszer, valójában egyszer sem képviseltette magát a ritkább homonima. Nagyobb, 1000 tagú mintában is csak párszor fordul elő. Maga a *barát* lemma összesen 26 543-szor fordul elő, fáradságos munka lenne további több ezer előfordulást egyenként átvizsgálni további 'szerzetes'-ek után kutatva. Ehelyett találjunk módot arra, hogy keressük meg a ritkább homonima előfordulásait a lehető legkisebb kézi munkával.

Eszünkbe juthat néhány szó, amely kollokációt alkothat a *barát*[2] szóval (*Julianus* és egyéb férfi keresztnevek, szerzetesrendnevek stb.). Ha ilyen kollokációkat keresünk, még ha sikerrel járunk is, akkor is egyoldalú találatokat kapunk. Ellenben ha nem a saját kútfőnkre, hanem a gépre bíznánk a kollokációk meghatározását, ezt az egyoldalúságot elkerülhetnénk.

Bár a ritkábbik homonimát nem tudjuk megtanítani a szemantikai egyértertelműsítést végző programnak, segítségünkre lehet a *barát*[2] közeli szinonimája, a *szerzetes*, illetve az ún. pszeudoszó eljárás.

A pszeudoszó eljárást a szemantikai egyértelműsítő programok tesztelésére használják. Több, alakilag különböző szót egyetlen pszeudoszóvá vonnak össze, és az eredeti szavakat a pszeudoszó különböző jelentéseinek tekintik. Például a *ló* és a *banán* szó alakjainak összevonásából képezhetünk egy **ló/banán* pszeudoszót, amelynek két jelentése van, a 'ló' és

a ’banán’. A különböző jelentéseket a pszeudoszó alakja egyértelműen jelzi, ezért nincs szükség a tanulókorpusz kézi összeállítására (SANDERSON 1994).

A feladat megoldásához abból a feltételezésből indulhatunk ki, hogy a *barát*[2] és a *szerzetes* a kontextusokban felcserélhető. Nyilvánvalóan ez nem teljesül mindig, hiszen *Julianus barát*-ot nem hívjuk **Julianus szerzetes*-nek, míg a *tibeti szerzetesek*-re nem alkalmazhatjuk a *tibeti barátok* kifejezést. Az első eset remélhetőleg nem okoz problémát, mert a *Julianus* utáni *barát* jelentésére a környezet más jegyei is utalhatnak. A második eset miatt semmiképp sem esik ki *barát*[2] előfordulás az algoritmus látóköréből, de könnyen előfordulhat, hogy a *tibeti barátaim* kifejezésben tévesen fedezi fel a *barát*[2] jelentést.

Létre kell hozni tehát a **barát/szerzetes* pszeudoszót, és ennek előfordulásaiból véletlenszerű mintát kell összeállítani. Mivel a *barát*[2] előfordulása elenyésző *barát*[1]-éhez képest, nyugodtan kijelenthetjük, hogy a pszeudoszó alakjai jelzik az aktuális jelentést. Ebből a mintából megtanítjuk az egyértelműsítő eljárást a két jelentés megkülönböztetésére, majd végrehajtjuk az egyértelműsítést immár a *barát* összes előfordulására.

A **barát/szerzetes* tanulókorpusz 2000 véletlenszerűen kiválasztott előfordulást tartalmazott, mindkét „jelentésére” 1000-1000 példát. A tanulókorpuszból Naiv Bayes-eljárással nyertem ki a sztochasztikus modellt, az előző fejezetben tárgyalt környezeti jegyek figyelembevételével. Természetesen a pszeudoszó alakjával nem számolt környezeti jegyként a tanulóalgoritmus, ugyanis ha így tenne, akkor minden *barát*-nak az első homonimát tulajdonítaná, pusztán a szó alakja miatt. Helyette a pszeudoszó inflexiós kategóriáját vette figyelembe, hiszen pl. a *barátom* alak nagyobb valószínűséggel tartozik az első homonimához, mint a másodikhoz.

A 26543 *barát* lemmából 2938-at jelölt meg ’szerzetes’ jelentésűként az algoritmus. Ebből egy ötszázas véletlen minta alapján megbecsültem, hogy a besorolásnak csak 9%-a volt helyes, 91%-ban téves. Azonban ha a ’szerzetes’ előfordulásokat a Bayes-módszer által adott pontszámok alapján csökkenő sorrendbe állítjuk, a tévedések nagy része a lista aljára rendeződik, és a lista első száz elemében már 50%-os a pontosság.

A célunk az volt, hogy a *barát*[2] szó előfordulásait megkeressük minél kevesebb kézi munkával. A fenti eljárás a szóba jöhető 26543 előfordulást leszűkítette 2938-ra, amely kb. 260 *barát*[2] előfordulást tartalmaz. A szűkebb halmaz átvizsgálása még így is jelentős kézimunkát igényel, de a *barát*[2]-k sűrűségét sikerült megnövelni. Ha megelégszünk 50 előfordulással, akkor elegendő csak a 100 legmagasabb pontszámú előfordulást átvizsgálni. Az eljárás azonban nem ad teljes eredményt, nincs kizárva, hogy a 23605 kiszűrt *barát* is tartalmaz *barát*[2] előfordulást.

5. Összegzés

A ritka homonimák megtalálása amolyan „tű a szénakazalban” típusú probléma. Egy kiválasztott homonima esetén sikerült megkönnyíteni valamelyest a keresést. Hogy ez a megoldás alkalmazható-e a többi, hasonló problémára, attól függ, hogy létezik-e a ritkább homonimának közeli szinonimája vagy több olyan szinonimája, amelyek együttesen lefedik a keresett szó jelentésmezőjét.

Irodalom

[ÉKsz.] l. Pusztai (főszerk.) 2003

Gale, William–Church, Kenneth–Yarowsky, David 1992. A Method for Disambiguating Word Senses in a Large Corpus. Computers and the Humanities 26: 415–439.

Nagy Viktor 2003. Korpuszegyértelműsítés – a morfoszintaxison túl. In: Alexin Zoltán–

Csendes Dóra (szerk.): Magyar számítógépes nyelvészeti konferencia. SZTE, Szeged. 1–7.
Oravecz, Csaba–Dienes, Péter 2002. Efficient Stochastic Part-of-Speech Tagging for Hungarian. In: Proceedings of the Third International Conference on Language Resources and Evaluation, LREC2002. Las Palmas. 710–717.
Pusztai Ferenc (főszerk.) 2003. Magyar értelmező kéziszótár. Akadémiai Kiadó. Budapest.
Sanderson, Mark 1994. Word sense disambiguation and information retrieval. In: Proceedings ACM Special Interest Group on Information retrieval. 142–151.
Váradi Tamás 2002. The Hungarian National Corpus. In: Proceedings of the Second International Conference on Language Resources and Evaluation, LREC2002. Las Palmas. 385–389.