Új korpuszstatisztikai eszköztár kollokációkeresésre

Kis Balázs, Ugray Gábor

MorphoLogic
{kis,ugray}@morphologic.hu

Kivonat

A nyelvi erőforrások – korpuszok, lexikonok – előkészítése a számítógépes nyelvészet leginkább munkaigényes művelete, szakmai konferenciákon mégis viszonylag kevés előadás foglalkozik vele – talán mert tudományos szempontból itt lehet felmutatni a legkevesebb új eredményt. Ez az előadás is leginkább azt hangsúlyozza, hogyan lehet meglevő nyelvtechnológiai és egyéb számítógépes eszközök felhasználásával hatékonyabbá tenni a korpuszok előkészítését és feldolgozását. Az előadás olyan új korpuszelőkészítő és -statisztikai eszköztárt mutat be, amely általánosan használható kollokációkeresésre egynyelvű korpuszokban, és annotálatlan korpuszból kiindulva is ad értékes eredményt. Az eszköztár a teljes műveletsort felöleli, a korpusz előkészítésétől a statisztikai számítások kiértékeléséig.

1. A feladat

A kollokációkeresés során a leginkább munkaigényes feladat a nyelvészeti erőforrások – korpuszok – előkészítése és a kollokációjelöltek kivonása a szövegből. Ennek sokszor nem tulajdonítunk tudományos jelentőséget, mert feltételezzük, hogy kizárólag mechanikus munkáról van szó – ez azonban távolról sem igaz az általános esetben, amikor a kollokációkutatáshoz nem rendelkezünk kellőképpen előkészített korpusszal.

A kollokációkutatás alapvető fontosságú eleme a számítógépes nyelvészeti munkának, mert a legfőbb alkalmazások – a tartalomelemzés és a fordítástámogatás, illetve az ezek alapjául szolgáló szintaktikai elemzés – megkövetelik a többszavas lexémák, az idiómák, az igevonzatok és más rögzült frazémák megfelelő felismerését.

Ha a kész, nyelvileg alaposan annotált korpuszt mindenféle kollokációstatisztikai művelet előfeltételének tekintjük, akkor jelentős mértékben korlátozzuk a kutatásunkhoz felhasználható korpuszok körét. Ha azonban rendelkezünk olyan eszköztárral, amely hiányosan vagy egyáltalán nem annotált korpuszból is képes előállítani a kollokációjelöltek halmazát, már biztosak lehetünk abban, hogy rövid idő alatt előteremthetjük a statisztikailag is értékelhető nagyságú szövegbázist. (Előadásunkban nem szólunk a korpuszgyűjtés sajátos kérdéseiről.)

A fentiekből látható, hogy a szerzők véleménye szerint értékes kollokációkutatás csak az úgynevezett típusos kollokációk, illetve kollokációjelöltek feldolgozásával lehetséges. A típusos kollokáció mint lokális terminus technicus azt jelenti, hogy a

kollokációk (n-gráfok) komponensei szintaktikailag, morfoszintaktikailag determinálva vannak. Ez az angol vagy a holland nyelvben például az igék és az elöljárószói csoportok (*V + PP*) együttállásának vizsgálatát jelenti, míg ennek – legalábbis egy jelenleg is folyó nemzetközi projekt előfeltételezése szerint – a magyarban az igék és az esetragos, illetve névutós főnévi csoportok együttállása felel meg. (Villada-Bouma 2002) A típusos kollokációk hangsúlyozása azért fontos, mert számos esetben, így például terminológiakivonatoló rendszerekben megelégszenek a típus nélküli kollokációk gyűjtésével, vagyis kizárólag a felszínen előforduló lexémák együttállását vizsgálják, azok morfoszintaktikai jellemzőinek figyelmen kívül hagyásával. (Castellví et al. 2001)

2. A korpusz

Eszköztárunk részben kényszer hatására jött létre. Amikor egy statisztikai csomagot (lásd később!) saját munkakörnyezetünkhöz kellett adaptálnunk, szükségünk volt olyan magyar adatsorokra, amelyek segítségével ki lehetett próbálni. Az egyetlen korpusz, amely e munka közben a rendelkezésünkre állt, a nemrég befejezett SZAK-korpusz volt, amely akkori állapotában nem rendelkezett nyelvi annotációval (Kis-Kis 2003).

A SZAK-korpusz informatikai szakszövegeket tartalmazó párhuzamos korpusz, amely komponensenként kb. 1,2 millió szövegszóra rúg. Ennél lényegesen nagyobb magyar nyelvi korpuszok is rendelkezésre állnak, azonban ez a terjedelem szaknyelvi korpusz esetén elfogadható, eléri a statisztikai számításokhoz szükséges kritikus tömeget.

Az előadásban ismertetett eszköztár a SZAK-korpusz magyar nyelvi komponensét használja fel, azonban az eszköztár első néhány tagját a teljes korpusz formátumának egységesítésére és nyelvi annotálására is felhasználtuk.

3. Az eszköztár

Az eszköztár a korpusz előkészítésének minden műveletét felöleli, a formátum egységesítésétől a végső adatsor összeállításáig. Az alapvető cél az NSP statisztikai csomag (Pedersen-Banerjee 2003) kiszolgálása volt, így az adatsor a statisztikai csomag által megkövetelt formátumot állítja elő.

Az eszköztár által megvalósított protokol főbb elemei a következők:

1. A korpuszbeli szövegek formátumának egységesítése;
2. A szövegek részleges nyelvi elemzése, a típusos kollokációjelöltek kivonatolása;
3. A kollokációjelöltek által reprezentált események heurisztikus utószűrése
4. Az eseménytípusok megszámlálása
5. A statisztikai csomag által megkívánt formátum előállítása

3.1. A korpusz előkészítése

A korpusz előkészítésének legfontosabb lépése a konzisztens – s ilyenformán jól feldolgozható – formátum előállítása. Erre az eszköztár olyan XML-struktúrát alkalmaz, amelyben egyaránt ábrázolhatók a nyelvi annotációval ellátott, illetve az annotációval nem rendelkező szövegek. Ezt a formátumot az eszköztár robusztus módon, közismert állományformátumokból is automatikusan és hibátlanul elő tudja állítani.

Olyan XML-formátumot választottunk, amely az eredeti dokumentum formázásából csak a legfontosabb elemeket őrzi meg, így például azonosítja a címsorokat, és megtartja a címfokozatokkal kapcsolatos információkat. Nem őrzi meg a táblázatokat és képeket, és általában feltételezi, hogy folytonos szövegről van szó. Ezt a formátumot a 2.-ben említett konkrét korpusz előkészítésekor dolgoztuk ki (Kis-Kis 2003).

3.2. Az adatsor kivonatolása

Ha kollokációkeresésről beszélünk, a második fontos lépés a kollokációjelöltek kivonása a korpuszból; ezeket a jelölteket kell később – például statisztikával – értékelni abból a szempontból, hogy valóban kollokációt alkotnak-e.

A kollokációkivonatolás az itt ismertetett eszköztárban a nyelvtechnológia legteljesebb körű kihasználásával történik. A kivonatolás során párokat (bigráfokat), illetve hármasokat (trigráfokat) lehet kiemelni a szövegből; a bigráfok és a trigráfok elemeit morfoszintaktikai, szintaktikai szempontok alapján már a kivonatolási szabályok segítségével szűrni lehet. Ezt az eszköztár úgy éri el, hogy szintaktikai elemző programot alkalmaz (a szerzők HumorESK, illetve Moose nevű eszközét, rendre magyar, illetve angol nyelvű korpuszokhoz), s a kivonatolás során az elemző programok által adott eredményeket (egyes részfák gyökereit) is felhasználja.

A kivonatolás során fontos szempont volt, hogy a későbbi mérési eredmények kompatibilisak legyenek annak a kutatócsoportnak az eredményeivel, ahonnan az NSP-csomagot közvetlenül átvettük. Így a keresés elsősorban a holland *V+PP* kollokációknak leginkább megfelelő magyar *V+NP+case* minták kigyűjtésére, vagyis az igék és az esetragos, illetve névutós főnévi csoportok együttállásának vizsgálatára irányult.

A megfelelő kollokációjelölteket úgynevezett metaszabályok segítségével lehet meghatározni. A mondatelemző programhoz olyan utószűrő modult illesztettünk, amely e metaszabályokat értelmezve a mondatelemző által létrehozott egyes részfák relatív gyökérszimbólumait, illetve a szimbólumok egyes jegyeit szűri ki, és ezeket mint komponenseket használja fel bi-, illetve trigráfok létrehozására. Példa metaszabályra:

```
VX!(lex),NP-FULL!(lex,case):4
```

A fenti szabály olyan trigráfok létrehozását írja elő, amelyek komponensei:

1. egy VX (ige) típusú szimbólum lemmája (*lex* jegye)
2. egy NP-FULL (főnévi csoport) típusú szimbólum lemmája (*lex* jegye) és
3. ugyanazon NP-FULL szimbólum esetragja (*case* jegye)

A trigráfot a rendszer akkor tekinti érvényesnek, ha a VX és az NP-FULL szimbólumok egy 4 terminális pozíciót fedő ablakon belül fordulnak elő együtt. Ez az ablakméret vitatható, mivel így a kivonatolási folyamat figyelmen kívül hagyja az igevonzatban szereplő összetettebb főnévi csoportokat. Példa a kivonatolás eredményére:

küld üzenet ACC 'üzenetet küld'

3.3. A kivonatolt adatsor utószűrése

A kivonatolás természetesen zajosabb, ha a kivonatolást annotálatlan korpuszon végezzük; sokat segít, ha a korpuszt előbb lemmatizáljuk, illetve egyértelműsített szófaji/morfoszintaktikai annotációval (POS-tagging) látjuk el. Az eszköztárhoz az előadás írása idején illesztjük a szófaji annotáló modult.

A jelöltek között megjelenő zajt azonban jelentősen csökkenthetjük egy utószűrő modul segítségével, amely szintén része az eszköztárnak. Ha a kivonatolás során a program egyes jelölteket tévesen ismer fel, s a tévedések egy része szabályokkal leírható, ezek még a statisztikakészítés előtt eltávolíthatók a jelöltek listájából (az eseménylistából).

A heurisztikus utószűrő is metaszabályokat alkalmaz, méghozzá ugyanazon morfológiai elemző felhasználásával. Ez azt jelenti, hogy kiszűrünk olyan többértelműségeket, ahol a kivonatoló választott egy adott trigráfot, amelynek egy komponense másképp is értelmezhető, s nyelvtudásunk alapján épp a másik értelmezés a gyakoribb. E metaszabályok alkalmazása vitatható, hiszen ez a legközvetlenebb beavatkozás a mérési eredményekbe; megfelelő egyértelműsítő (*POS-tagger*) modul alkalmazása esetén szükségtelenné válhat. Addig is alapszabálynak tekintjük, hogy a metaszabályokkal csak a nyilvánvaló zajt szabad szűrni.

3.4. Az események megszámlálása és a statisztikai csomag által megkívánt adatformátum előállítása

Az NSP statisztikai csomag már rendelkezésre álló gyakorisági adatokra alkalmaz különböző statisztikai függvényeket. Ezért a korpusz előkészítéséhez az egyes események gyakoriságának megszámlálása is hozzátartozik. Eseménynek egy kivonatolt n-gráfot (bigráfot, trigráfot) tekintünk.

Az eszköztárnak ezért része egy robusztus gyakoriságszámláló program is, amelynek legfőbb előnye a méretezhetősége: milliós, milliárdos eseményhalmazból is rövid idő alatt előállítja a gyakorisági listát. Vizsgálataink során a fent említetthez hasonló kivonatolást végeztünk a British National Corpus (BNC) anyagának egyharmadán, körülbelül százmillió szövegszón: a kapott eseményhalmaz gyakorisági listájának előállítása egy átlagosnak tekinthető PC-n, hibakereső üzemmódban kb. 40 másodpercig tartott. Sem a korpuszt, sem az eseményhalmazt nem osztottuk részekre, az átalakított, nyelvileg annotált rész-BNC terjedelme 4 gigabájtra rúgott!

Az NSP-csomagban implementált statisztikai függvények paraméterként nemcsak a teljes n-gráfok gyakoriságszámait követelik meg: az n-gráfok komponenseit is meg kell számolni. Így trigráfok esetén a teljes trigráf gyakorisága mellett fel kell tüntetni az 1. és 2., az 1. és 3., illetve a 2. és 3. komponens együttes előfordulásának gyakori-

ságát is, valamint az egyes komponensek önálló előfordulásainak számát is (amikor nem vizsgáljuk, hogy az illető komponens része-e kollokációnak). Ezeket a fent említett gyakoriságszámláló programmal egyszerűen meg lehetett számlálni, de az eszköztárt ki kellett egészíteni egy olyan programmal is, amely összefésüli a különböző számolási menetek során kapott adatsorokat is. Egy trigráf esetén az NSP-csomag által megkívánt bemenet a következő:

```
hoz<>lét<>SUB<>722 1044 1108 22506
```

Ebben még csak az egyes komponensek gyakorisága szerepel. A kételemű részhalmazok gyakoriságát egy, az NSP-csomag részét képező programmal adhatjuk hozzá az adatsorhoz.

Mind a gyakoriságszámláló, mind az adatokat összefésülő programnak kellően robusztusnak kell lennie, mert az adatsor több millió eseményt is tartalmazhat. Ezekben a programokban a Naszódi Mátyás (MorphoLogic) által kifejlesztett, és Kis Balázs által adaptált GammaTrie nyelviadat-indexelő modult használjuk, amely még a MorphoLogic műhelyében létrehozott programok között is kitűnik gyorsaságával.

3.5. Az események statisztikai értékelése, a jelöltek sorbaállítása

A statisztikai feldolgozáshoz eszköztárunk az NSP statisztikai csomagot (Pedersen 2003) használja, amelynek legfőbb előnye, hogy tetszőleges statisztikai függvény illeszthető hozzá. Az első kísérlet során a jól ismert *log-likelihood*, illetve a Villada (2002) által adaptált és az NSP-hez illesztett *salience*-függvényt használtuk; a holland elöljárószói kollokációk keresésekor (Villada 2002) szerint e két függvény bizonyult a legeredményesebbnek, bár terminológiakivonatolási kísérletekben a kölcsönösinformáció-számítás is jól használható (Jacquemin 2001).

A statisztikai függvények olyan rendezett eseménylistát adnak, ahol a lista élén a legrelevánsabbnak tekintett események, a végén pedig a legkevésbé releváns események találhatók. Ez nem feltétlenül van összhangban az események relatív gyakoriságával; a statisztikai vizsgálatok legnagyobb kihívása, hogy az alacsony gyakoriságú, de releváns eseményeket is észre kell venni, és a statisztikailag értékelt eseménylisták elejére kell rendezni.

Az első kísérleti eredmények kiértékelésekor – az igék és az esetragos főnévi csoportok kollokációinak vizsgálatának esetében – azt láttuk, hogy mindkét felhasznált statisztikai függvény az eseménylista elejére rendezte az olyan 'ál-igevonzatokat', amelyek valójában olyan elváló igekötős igék voltak, ahol az igekötő helyén esetragos főnév jelent meg ('vesz észre'). Ez két alapvető tényt bizonyít:

1. A statisztikai csomag alkalmas a magyar nyelvi környezetben, megfelelően előkészített korpuszokon való használatra.
2. A korpuszelőkészítő eszköztár alkalmas arra, hogy statisztikai feldolgozásra akár annotálatlan korpuszokat is előkészítsen, azonban megfelelő eredményt csak akkor várhatunk, ha a részleges elemzéshez használatos magyar nyelvtant megfelelően átalakítjuk, és a többértelműségeket visszaadó morfológiai elemző program helyett egyértelműsítő szófaji címkéző programot (POS-tagger) alkalmazunk.

4. Összefoglalás

Az előadásban bemutattunk egy új, nyelvtechnológiát is felhasználó korpuszstatisztikai eszköztárat, amelyet összekapcsoltunk egy, a nemzetközi irodalomból ismert és különböző kutatóműhelyekben széles körűen használatos statisztikai csomaggal Az eszköztár kollokációkeresésre (bigráfok, trigráfok statisztikai értékelésére) alkalmas. Konkrét kollokációkeresési példán bizonyítottuk az eszköztár által alkalmazott megközelítés helyességét és a nemzetközi irodalomból átvett statisztikai csomag magyar nyelvi korpuszokra való alkalmazhatóságát.

Az itt leírt kutatás jelentős eredményt hozott a MorphoLogic számítógépes nyelvészeti kutatócsoportja számára is, ugyanis itt eddig nem állt rendelkezésre átfogó statisztikai rendszer, így az alkalmazott nyelvtechnológiai megoldásokat meglehetősen nehéz volt korpuszbeli példákkal igazolni. Azonban az itt ismertetett eszköztár és az NSP statisztikai csomag segítségével a kutatócsoport immár egészséges módon össze tudja kapcsolni a számítógépes nyelvészetben alkalmazott szabályalapú és statisztikai módszereket.

Köszönetnyilvánítás

A szerzők köszönetet mondanak a groningeni egyetem (Rijksuniversiteit Groningen) kutatócsoportjának (Begoña Villada Moirón, Gosse Bouma, Bíró Tamás, John Nerbonne) a statisztikai eszköztár adaptálásában nyújtott segítségükért. Köszönet illeti Naszódi Mátyást a GammaTrie nyelviadat-indexelő modulért, Pál Miklóst, Tihanyi Lászlót és Novák Attilát a Humor morfológiai elemző, illetve HelyesLem lemmatizáló modulért.

Irodalomjegyzék

CASTELLVÍ, M. Teresa Cabré – BAGOT, Rosa Estopà – PALATRESI, Jordi Vivaldi: Automatic Term Detection: A Review of Current Systems. In: *Bourigault, Didier – Jacquemin, Christian – L'Homme, Marie-Claude (eds.): Recent Advances in Computational Terminology.* John Benjamins, Amsterdam-Philadelphia, 2001. pp. 53–88.

JACQUEMIN, Christian (2001): Spotting and Discovering Terms through Natural Language Processing. The MIT Press, Cambridge, MA, USA–London.

KIS Balázs (1997): Mi van a szavakon túl? Nyelvtani szerkezetek felismerése számítógéppel. *Előadás a VII. Országos Alkalmazott Nyelvészeti Konferencián.* Külkereskedelmi Főiskola, Budapest, 1997

KIS, Ádám–KIS, Balázs (2003): A Prescriptive Corpus-based Technical Dictionary. Development of a multi-purpose technical dictionary. In: *Proceedings of COMPLEX 2003*, Budapest.

PEDERSEN–BANERJEE (2003): The Design, Implementation and Use of the Ngram Statistics Package. In: *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics* (Mexico City).

VILLADA, Begona–BOUMA, Gosse (2002): A corpus-based approach to the acquisition of collocational prepositional phrases. In: *Proceedings of EURALEX 2002*, Copenhagen, Denmark.