Ontológiaalapú szöveganntotáció a Sintagma projektben

Szekeres András Márk[1], Varga László Zsolt[1], Krauth Péter[2]

[1] MTA SZTAKI,
Budapest, 1111 Lágymányosi u. 10.
{szekeres, laszlo.varga}sztaki.hu
[2] IQSYS Informatikai ZRt.,
Budapest, 1134 Hun u. 2.
krauth.peter@kfki.com

Kivonat: A természetes nyelvű szövegek számítógépes feldolgozása során általános igény, hogy tovább lehessen lépni egyszerű karaktersorozatok felismerésének szintjén a szövegben. A Sintagma-projektben fejlesztett ontológiaalapú szöveganntotáció egyfelől a szöveget struktúráltan ragadja meg, vagyis a rendelkezésre álló karaktersorozatból szintaktikai elemzés segítségével levezetési fákat ill. más hasonló nyelvtani szerkezetet állít elő. Másfelől nemcsak a mondat szerkezetének, hanem a benne szereplő szavaknak is struktúráltabb reprezentálását használja, és ennek eredményeképp szavak mellett ontológiabeli fogalmakkal is képes *indexelni* a szöveget. A fejlesztés újszerű megközelítései közé tartozik a *többértelműségek* feloldására alkalmazott algoritmus és az egyedi példányokra, objektumokra utaló olyan *speciális kifejezések* (pl. rendszám, telefonszám, számlaszám, termékkód) felismerésének és értelmezésének módja, amelyek gyakran fordulnak elő szakszövegekben nehezítve ezzel a szövegek feldolgozását. Ez utóbbit a szöveganntotáció számára a Sintagma *szemantikus információintegráló rendszer* biztosítja, amely fogalmi és példányszintű háttérinformációkkal képes ellátni a szövegfeldolgozást, és növelni annak eredményességét.

1 Bevezetés

Az „ontológia” kifejezést különféle szűkebb-tágabb értelemben szokták használni. A továbbiakban alapvető tulajdonságának azt tekintjük, hogy öröklődést biztosító hierarchikus struktúrával rendelkezik.

A természetes nyelvű szövegekkel foglalkozó projektek különféle célokat tűznek ki: a keresés hatékonyabbá tételétől kezdve a szöveg tartalmának logikai rekonstrukciójáig terjedően meglehetősen széles a paletta. A több éve zajló kutatás azzal a problémával foglalkozik, hogy hogyan kapcsolható össze a szöveg az értelmezéséhez használt ontológiával, vagyis hogy az ontológia fogalmait hogyan lehet felismerni a szövegben. Az utóbbi két évben a kutatásnak a Sintagma-projekt keretein belüli fejlesztések adtak további perspektívát.

A jelen cikk a ragozott szavakból a szótövek előállítását megoldott problémának tekinti, az ezt követő lépésekre fókuszál. A kutatás tárgya ezért a többértelműségek és referenciák feloldása volt, vagyis az, hogy olyan esetekben is beazonosítható legyen egy adott fogalom, ha egy másik szóval hivatkoznak rá.

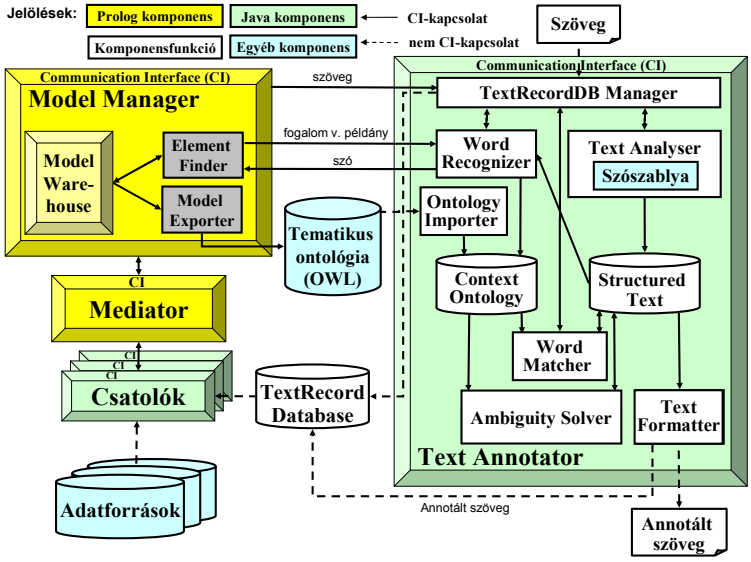

Fig. 1. A Text Annotator felépítése és kapcsolata a szemantikus integrációval.

Egy másik ontológiákhoz kapcsolódó fejlesztésre nagy méretű példány-adatbázisok kapcsán merült fel az igény. Például közéleti szereplők különféle információkkal ellátott nyilvántartásánál nem praktikus egy OWL-ben leírt ontológiába több ezer, esetleg tízezer példányt felvenni: a felismerő algoritmust jelentősen lelassítaná. Ezért az ontológiába a példányok és metainformációik „on-demand” (szükség szerint) dinamikusan töltődnek be a háttérként szolgáló, a Sintagma által integrált adatforrásokból.

Fontos felismerés volt az is, hogy a háttérinformációk adatelemzésével (adatprofilozás) olyan metainformációk (pl. adatmaszkok) származtathatók, amelyek segítik egyes speciális formátumú kifejezések felismerését mint pl. egy rendszám (ABC-123) alapján annak a kikövetkeztetését, hogy valójában valamilyen járműről van szó. Ha van hozzáférés a vonatkozó adatbázishoz, akkor a szóbanforgó jármű (ti. az „ABC-123” rendszámú) egyéb tulajdonságait (pl. tulajdonos, típus) is származtatni lehet.

2 Többértelműségek feloldása

Többértelműségek feloldása alatt sokkal tágabb problémakör értendő, mint ami általában szokás. Nem csupán a többértelmű szavakkal és nyelvtani utalószavak okozta

többértelműségekkel kell foglalkozni, hanem általában azzal a jelenséggel, amikor egy szó egy másik helyett szerepelhet, és így el kell dönteni, hogy melyiket jelenti.

2.1 Többértelműségi problémák

Egy tipikus esete ennek a jelenségnek a következő „Feltettem a rizst főni. Azután leültem a TV elé, és nem vettem észre, hogy odaégett az étel". Itt az „étel" szó a „rizs"-re utal, az emberi olvasó természetesen ezt rögtön megérti. A számítógépes feldolgozás során viszont minden ma létező megoldás, amikor megtalálja az „étel" szót a szövegben, majd az ontológiában/szótárban, akkor itt megáll, és fel sem merül, hogy kiderítse, vajon esetleg hivatkozik-e egy már korábban a szövegben megjelenő dologra?

Ez a jelenség igen gyakori a szövegekben, különösen újságcikkek esetén, ahol általában a szóismétlést súlyos stilisztikai hibának tekintik. Az újságcikkek esetében a mondatok többségében szerepel legalább egy ilyen fajta referencia. De formális szövegekben is előfordul, például orvosi zárójelentésekben: pl. egy fülvizsgálat után az orvos egyszerűen „nyálkahártyát" ír a „dobüregi nyálkahártya" helyett.

2.2 A megoldás fő elve

A probléma feloldása két lépésből áll, először generálni kell a jelentésjelölteket, másodszor pedig dönteni kell közöttük.

A jelöltgenerálást úgy történik, hogy az ontológiában megtalált fogalom leszármazottait tekinti jelölteknek. Ugyanis leggyakrabban a szóismétlés elkerülése érdekében használt szavak az eredeti szó *általánosabb kategóriái* (mint például a „rizs" helyett használt „étel" kifejezés). Informális szövegekben előfordulnak *tulajdonságokon alapuló referenciák* is: például egy jelenet szereplői közül úgy hivatkoznak az egyikre, mint „a kövérebbikre". Jelenleg ezzel az esettel az algoritmus nem foglalkozik, de éppen a már említett háttérinformációk kezelése teremti meg a lehetőséget arra, hogy a példányokat tulajdonságokkal is azonosítani lehessen.

A leszármazottak generálása mellett „hagyományosabb" módon is vesz fel jelölteket, szinonímalisták alapján. Így a többértelmű szavak problémáját is ugyanaz az algoritmus oldja meg: például a „körte" szó feldolgozása során a körte mint „gyümölcs" és mint „villanykörte" is a jelöltek közé kerül.

A jelöltek közül a kontextus alapján az ontológia segítségével történik a választás. A szövegkörnyezetben már beazonosított fogalmak és az adott jelölt között az ontológiában a relációk mentén mért *távolságon* alapul a választás. Ez egyrészt lehet saját maga korábbi előfordulása (mint például a „rizs-étel" példánál), de akár akkor is működik, mikor maga a fogalom nem is szerepelt. Például a „szögek beverésére alkalmas szerszám" mondatrész esetében, ha az ontológiába felvettünk olyan relációt, hogy „bekalapál: eszköze kalapács, tárgya: szög", akkor a „szerszám" fogalom leszármazottai közül a „kalapács" van legközelebb a kontextushoz (mivel a „bekalapál" reláció összeköti a „szög"-gel), és az algoritmus azt az eredményt hozza ki, hogy a „szerszám" szó itt speficikusan a „kalapács" fogalomra utal.