Szövegfeldolgozás ontológiák segítségével – fogalmak azonosítása

Szekeres András Márk

Ontológia fogalmainak azonosítása

- Minden ontológia alapú szövegfeldolgozás egyik kulcslépése.
  - Keresés „szemantikusabbá” tétele
  - Ontológia (vagy Topicmap) alapú keresés, kategorizálás
  - Szövegből logikai állítások kinyerése
- Ragozott szóalakok szótövezését megoldottnak tekintjük. (A projektben a szószablyát alkalmazzuk).

Referenciális többértleműség

- Továbbra is probléma, hogy az ontológia fogalma más szóként szerepel a szövegben.
  - Utalószavak. Szintaktikai elemző feladata, hogy ha fel nem is oldja (van amikor emberi olvasó se képes erre), de legalább jelölteket adjon. A „referenciális többértelműség” az irodalomban csak erre vonatkozik, én kiterjesztve használom.
  - Szinonímák. Szinoníma szótárakkal felismerhetőek.
  - Van egy harmadik eset is, ennek egy példája „Feltettem a rízst főni. Leültem TVzni, és nem vettem észre, hogy odaégett az étel”. Ezzel fogunk foglalkozni, nevezzük Asszociatív referenciának.

Asszociatív referencia

- Ez az eset vizsgálata szinte teljesen elhanyagolt, pedig rendkívül gyakori, az emberi beszéd alapvető jellemzőjének tűnik.
- Egy (újságcikkek kis corpusára kiterjedő) vizsgálatom alapján a mondatok 90%-ban van ilyen fajta referencia. Orvosi zárójelentések (úgyszintén kis corpusára kiterjedő) vizsgálatom alapján a szinoníma vagy asszociatív referencia a mondatok 40%-ban van.
- Leggyakrabban a referált fogalmat egy általánosabb kategória helyettesít. A projekt során ezzel a jelenséggel foglalkozunk, de a későbbiekben ki tervezzük terjeszteni az algoritmust más esetekre is.

Eljárás

- Egy, az ontológiában megtalált szóhoz érve generáljuk az alatta levő fogalmakat, ez a jelöltek halmaza.
- A jelölteket a szövegkörnyezet (más, már beazonosított fogalmak) alapján súlyozzuk (relációk mentén milyen távol vannak egymástól).
- Példa: „A hallócsont ép, nyálkahártya egészséges”. Itt a nyálkahártya valójában a „dobüregi nyálkahártyára” vonatkozik, amely a kontextusból ki is derül (a hallócsont located-in relációban van a dobüreggel, a dobüregi nyálkahártya is).

MTIs alkalmazás

- Az MTI híradatbázisan több demo alkalmazást tervezünk: a hírek kategorizálását és egy szemantikus(abb) keresést.
- „Meghalt Zámbó Jimmy kedd délelőtt a Honvéd Kórházban - közölte Katona István főorvos a helyszínen tartózkodó újságírókkal. A népszerű előadóművész reggel hat óra tájban saját fegyverével lőtte fejbe magát, a mentők életveszélyes állapotban, koponyasérüléssel szállították a Honvéd Kórházba. Az énekes - az eddigi ismeretek szerint – fegyverviselési engedéllyel rendelkezett.”
- A fenti szövegben egyszer szerepel „Zámbó Jimmy” és „Katona István”. Hagyományos kereső szerint a cikk ugyanannyira szól mindkettőjükről. Azonban az algoritmusunkat használva kiderül, hogy Zámbó Jimmy háromszor is szerepel, asszociatív referenciával. Ezeknek a referenciáknak a felismerése nagy mértékben növeli a keresés és/vagy kategorizálás hatékonyságát.

Ontológia készítés

- A megközelítés egyben ontológiák ellenőrzésére is alkalmas.
- Az ontoclean-nel megjelent az első komoly, elméleti alapokon nyugvó módszertan (amely azonban még mindig csak Arisztotelészig jutott, az informatika számára releváns episztemológia, ontológia és nyelv-filozófia több ezer éves irodalmában.)
- A nyelvfeldolgozás szempontjából az ontológiáknak a nyelvet kell leírniuk. Ez és ehhez hasonló alkalmazások visszajelzést adnak arra, hogy mennyire jól sikerült ez. Kísérleti ellenőrzésee az ontológiák helyességének.