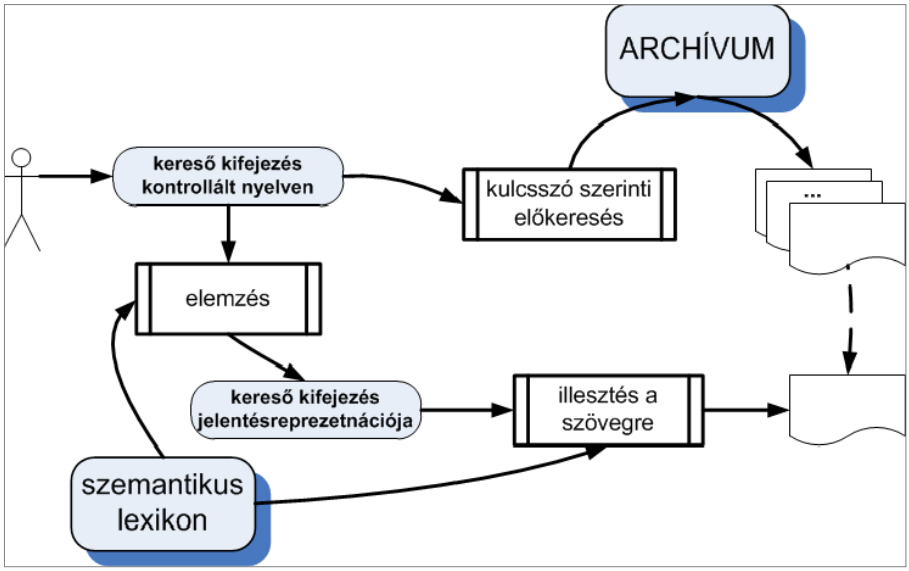
MASZEKER: szemantikus kereső program

Hussami Péter[1]

[1]Alkalmazott Logikai Laboratórium
1022 Budapest, Hankóczy j. u. 7
hussami@all.hu

Az Alkalmazott Logikai Laboratórium és a Szegedi Tudományegyetem Informatikai Tanszékcsoportja, valamint Könyvtár- és Humán Információtudományi Tanszéke egy, a Nemzeti Fejlesztési Ügynökség által támogatott közös projektet (TECH_08_A2/2-2008-0092) fejezett be 2012-ben. A projekt célja olyan, új elveken alapuló, integrált információ-visszakereső rendszer kifejlesztése volt, amely a keresést végző felhasználó szemantikai kompetenciáját az eddigieknél nagyobb mértékben kiaknázva teszi lehetővé a természetes nyelvi dokumentumtárakban (szövegekben) történő valóban *tartalmi* keresést. Egyszerűen szólva: a felhasználó jól formált frázisokkal, mondatokkal specifikálhatja, milyen tartalmú dokumentumokat keres. A rendszer áttekintő architektúrája az 1. ábrán látható.



1. ábra. A MASZEKER rendszer áttekintő architektúrája.

Az ábrának megfelelően a releváns dokumentumok keresése a következő lépésekből áll:

1.) a felhasználó egy kontrollált nyelven adja meg a keresőkifejezést,

2.) a rendszer szintaktikus és szemantikus elemzést végezve előállítja a keresőkifejezés jelentésreprezentációját,
3.) szavak szerinti kereséssel előszűri az archívumot,
4.) végül azokra a szövegszegmensekre, amelyekben a szavak szerinti keresés találatai vannak, illeszti a keresőkifejezés jelentésreprezentációját.

Az MSzNy VII konferencián tartott előadáson [1] ismertetésre kerültek a fenti elemek megvalósítására vonatkozó elméleti alapelvek, elsősorban a szemantikus reprezentáció felépítése mint sarokkő köré szervezve. Az MSzNy VIII konferencián tartott bemutatón [2] a rendszer első változatát mutattuk be, amely főnévi csoportokon mint keresőkifejezéseken működött. Idén a teljes, jól formált főnévi csoportokon és mondatokon működő rendszert kívánjuk bemutatni. Ugyanezen a konferencián egy előadás [3] számol be a rendszer alapjául szolgáló technológiáról.

A demóban az archívumot szabadalmi leírások főigénypontjaiból összeállított dokumentumgyűjtemény alkotja[1]. A keresőkifejezés több mondatból, ill. főnévi kifejezésből állhat, kontrollált angol nyelven megfogalmazva. A megszorítások az egyértelműséget biztosítják, a tipikusan nehezen egyértelműsíthető fordulatokat akartuk kizárni. Legfontosabb korlátozások (a teljes definíció [4]-ben hozzáférhető):

- csak kijelentő módú, jelen idejű mondatok használhatók,
- tiltott a mellérendelő mellékmondat (viszont a mondatok AND, OR kapcsolóval kapcsolhatóak, zárójelezhetőek),
- tiltott az alárendelő mellékmondatok bármiféle lerövidítése (pl. igeneves utómódosítók),
- az alárendelő mellékmondatnak a „which” vonatkozó névmással kell kezdődnie, és ennek a közvetlenül megelőző főnévi csoport fejére kell vonatkoznia,
- tiltottak az igeneves előmódosítók,
- felsorolás, koordináció csak főnévi csoportok közt megengedett, ezeket a felhasználónak jelölnie kell.

A felhasználói interfész segíti, és a morfoszintaktikai elemzés eredménye alapján ellenőrzi a szabályok betartását. Mivel a teljes szabályrendszer nem ellenőrizhető, a generált jelentésreprezentáció grafikusan bemutattatik – ha szükséges, a felhasználó módosíthatja a keresőkifejezést. Ez a megjelenítés segíti egyértelműsíteni az egyes szavak jelentésének megállapítását is, mivel ha több frame/synset van egy csomóponthoz rendelve, akkor a felhasználó választhatja a megfelelőt.

A rendszer a keresőkifejezéshez illő frázisokat keres az igénypontok szövegében, és az eredményt a grafikus interfészen megmutatja, kiemelve azokat a szavakat, amelyek olyan frázist alkotnak, melyet a keresőkifejezés egy szegmenséhez hasonlónak talált. Míg a keresőkifejezés feldolgozásánál maximálisan törekszünk a pontos jelentésreprezentációra, a keresés fázisában az aktuális szövegrészlet vizsgálatánál csak azt ellenőrizzük, hogy jelentheti-e a keresőkifejezés valamely frázisát.

[1] A projekt egyik kiemelt felhasználási területe a szabadalmi keresés, s a prototípust „gyógyhatású készítmények és kozmetikai szerek” témaköréből származó szabadalmakon mutatjuk be.

Köszönetnyilvánítás

A fejlesztés az NFÜ által finanszírozott, MASZEKER kódnevű, TECH_08_A2/2-2008-0092 számú projekt keretében valósult meg.

Hivatkozások

1. Szőts M., Csirik J., Gergely T., Karvalics L.: MASZEKER: projekt szemantikus kereső technológia kidolgozására In_ Tanács A., Vincze V. (szerk.): MSzNy 2010 – VII Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2010) 159–167
2. Hussami P.: MASZEKER: szemantikus kereső program In: Tanács A., Vincze V. (szerk.): MSzNy 2011 – VIII Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2011) 321–322
3. Szőts M., Simonyi A.: Frame-szemantikára alapozott információ-visszakereső rendszer In: Tanács A., Vincze V. (szerk.): IX. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2013) 275–285
4. A kontrollált nyelv definíciója http://www.maszeker.hu/?page=download