Budapesti Műszaki és Gazdaságtudományi Egyetem
Méréstechnika és Információs Rendszerek Tanszék

UIMA keretrendszerbeli nyelvi elemzők integrálása annotáló rendszerbe

Farkas Tamás (V9TI8U), 2. évfolyam (MSc), mérnökinformatikus
Konzulens: dr. Bartha Tamás, MIT
Külső konzulens: Tétényi István, SZTAKI ITAK
Szolgáltatásbiztos rendszertervezés szakirány
Önálló laboratórium 2 összefoglaló
2010/11. II. félév

Napjaink információs társadalmának egyik nagy vívmánya a kollaboratív tudásbázisok létrehozásainak lehetősége. A legismertebb ilyen szabadon elérhető és szerkeszthető tudásbázis a Wikipédia, 10 éves működése során 16 millió szócikkel rendelkezik. A tudásbázisok az emberek számára készülnek tehát a tárolt anyagok formázása az emberi fogyasztást támogatja. A gépeket csak mint eszközt használjuk, az anyagok gépi reprezentációjával nem foglalkozunk. Ha a tudást a gép számára is értelmezhetővé tennénk, nagyban segíthetnék a szerkesztők munkáját intelligens tanácsaikkal és javaslataikkal a gépek nemcsak a megjelenítés és tárolás funkcióját látnák el.

A Wikipédia szerkesztők munkájának segítéseként indult a Sztakipédia projekt az MTA SZTAKI ITAK osztályán belül. A projekt indításakor felmérésben megkérdeztük a Wikipédiát használókat milyen új funkciókat szeretnének a szerkesztőbe, és a következő eredmények születtek: a legtöbben képfeltöltési segítséget, belső link és kategória ajánlatokat, valamint helyesírás ellenőrzést és WYSIWYG szerkesztőt kértek. A Sztakipédia szerkesztő ezek közül a kategóriák, belső és külső linkek ajánlását, valamint egy vizuális editor elkészítését tűzte ki célul.

A megvalósított rendszer architektúrája 3 fő részből áll: webes frontend, háttértámogató rendszerek, és a háttérrendszereket integráló és vezérlő szerveroldali komponens. A webes szerkesztő TinyMCE alapú, kimenete HTML, továbbá képes a háttérrendszerekből érkező tartalomfüggő annotációk megjelenítésére. A háttérben három fő információforrás dolgozik: az UIMA (Unstructured Information Management Architecture) keretrendszerben integrált nyelvi feldolgozás, szövegelemzés, Yahoo! Term extractor and BOSS API a többszavas kifejezések és külső linkajánlatok keresésére, valamint DBPedia keresés a Wikipedia metaadatainak elérésére. Az aggregáló komponens kezeli a háttérrendszereket, gondoskodik a válaszaik egységes reprezentációjáról.

Önálló laboratóriumi munkám során az előző félévben elkészített UIMA komponensek felhasználásával, valamint új komponensek létrehozásával foglalkoztam. Célom a Wikipédián belüli linkek ajánlásának támogatása volt. Mivel az angol nyelvű Wikipédia már több mint 3 millió szócikket tartalmaz, szinte minden szóhoz tartozik egy cikk, ezért fontos, hogy a linkelésre felajánlott cikkek tényleg kapcsolódjanak a szerkesztett cikkhez. Meg kell keresni az adott szöveget reprezentáló szavakat, amik gyakran fordulnak elő, de a teljes korpuszon viszonylag ritkán. Erre a tf-idf (term frequency – inverse document frequency) alkalmas. A tf-idf meghatározásához adatbázist kellett építeni, ami az egyes szavakról tartalmazta, hogy hány dokumentumban fordultak elő a korpuszon. Korpuszként az angol nyelvű Wikipédia cikkek szolgáltak, feldolgozásukhoz az előző félévben elkészített elemzőket használtam. Végül egy memóriában eltárolható méretű hashtáblát kaptam, megfelelő szűrések után. Többszavas kifejezések keresésével is foglalkoztam, ehhez az egymás mellett gyakran előforduló szavakat kellett szintén egy adatbázisba rendezni, tárolni az előfordulás gyakoriságát, és rangját. Az adatbázisséma elkészítéséhez, és a feldolgozáshoz a Lipcsei Egyetemen fejlesztett Wortschatz egynyelvű korpuszfeldolgozó rendszert használtam. Végezetül elkészítettem mindkét modul UIMA annotáló komponensét. A kookkurencia elemző erősen épít a tf-idf elemző kimentére, ugyanis csak legmagasabb tf-idf értékű szavakkal, és azok szomszédjaival foglalkozik.

Az elkészült elemző komponensek némi teljesítményhangolás után integrálva lettek a Sztakipédia szerkesztőbe. A tf-idf értékek további felhasználásra kerültek a könyvajánlat kereső modulban is.

A Sztakipédia szerkesztőt sikeresen beneveztük az AI Mashup 2011 nevű versenyre, valamint készült egy cikk és egy előadás a 8th Extended Semantic Web Conference nevű konferenciára is.