Szemantikus annotációk létrehozása a weben nyelvtechnológiai eszközök támogatásával

Héder Mihály[1,2]

[1] MTA Számítástechnikai és Automatizálási Kutatóintézet
Internet Technológiák és Alkalmazások Központ, mihaly.heder@sztaki.hu
[2] Budapesti Műszaki és Gazdaságtudományi Egyetem
Filozófia és Tudománytörténet Tanszék

1. Absztrakt

A weben található hipertext minőségének egyik mutatója, hogy milyen mennyiségben található a szöveg mellett számítógép által is értelmezhető, azaz strukturált reprezentációja a szándékolt jelentésnek. Az így csatolt információkat szemantikus annotációknak is nevezhetjük, mivel úgy magyarázzák a számítógép számára az egyes szövegrészek értelmét, mint a széljegyzetek egy könyvben.

Az MTA Sztaki Internet Technológiák és Alkalmazások Központjában a szemantikus annotációkkal kapcsolatos kutatások és fejlesztések keretein belül létrehoztunk egy webes keretrendszert, amelynek segítségével az annotáció kérdései gyakorlati síkon is tárgyalhatókká váltak.

Az eszköz amellett, hogy megoldásokat kínál az annotációk granularitásával, szintaxisával és lekérdezhetőségével kapcsolatos néhány problémára, képes UIMA és egyéb interfésszel rendelkező nyelvfeldolgozó adapterek használatára is. Az angol nyelvű és nyelvfüggetlen adaptereken kívül az eszköz a Szegedi Tudományegyetem Nyelvtechnológiai csoportja által fejlesztett *magyarlanc* [1] UIMA-adaptereket is használja, a magyar nyelvű feldolgozás döntően ezen modulok segítségével történik.

Az annotáló szoftver leginkább végfelhasználóknak szóló alkalmazása egy Wikipédia-cikkszerkesztő. Ebben a konfigurációban a szoftver egy hivatkozásajánlásokat megfogalmazó névelem-felismerőt és a Hitec [2] keretrendszer magyar wikin betanított, webszervizen elérhető verzióját is használja. Ez az alkalmazás mutat rá a legmarkánsabban azokra a nehézségekre, amelyeket az annotációk helyes tárolása és karbantartása jelent egy olyan formátum (jelen esetben a wikitext) és létrehozási munkafolyamat esetében, amelyet ezen feladatokra nem készítettek fel.

Egy panaszlevél-kezelő alkalmazás is bemutatásra kerül. Ebben az alkalmazásban egyszerre próbáljuk segíteni a szemantikus keretekkel és szkriptekkel, illetve a szerkezeti egységekkel kapcsolatos kutatásainkat és a majdani felhasználót. Az Igazságügyi Minisztérium panaszleveleit tartalmazó korpusz, amellyel

ebben a projektben dolgozunk, megköveteli újfajta megközelítések és heurisztikák kikísérletezését.

Végezetül bemutatásra kerülnek azon kísérleteink, amelyek a nyelvfeldolgozással támogatott jogiszöveg-létrehozás és -annotálás gyakorlati kérdéseit vizsgálták.

Hivatkozások

1. Zsibrita, J., Nagy, I., Farkas, R.: Magyar nyelvi elemző modulok az UIMA keretrendszerhez. In: Magyar Számítógépes Nyelvészeti Konferencia. (2009)
2. : Hitec. (*categorizer.tmit.bme.hu/trac/wiki*)