Magyar nyelvi elemző modulok az UIMA keretrendszerhez

Zsibrita János[1], Nagy István[1], Farkas Richárd[2]

1 Szegedi Tudományegyetem, Informatikai Tanszékcsoport
6720, Szeged, Árpád tér 2.
{zsibrita, nistvan}@inf.u-szeged.hu
[2] MTA-SzTE Mesterséges Intelligencia Kutatócsoport
6720 Szeged, Tisza Lajos krt. 103. III. lépcsőház
rfarkas@inf.u-szeged.hu

1 Az UIMA keretrendszer

Az UIMA (Unstructured Information Management Application) keretrendszer [1] célja olyan szoftverrendszerek fejlesztésének támogatása, amelyek nagy mennyiségű strukturálatlan adat elemzését célozzák meg. Az Apache UIMA[1] az UIMA specifikáció nyílt forráskódú implementációja, amely kifejezetten szöveges dokumentumok feldolgozását támogatja.

Az UIMA keretrendszer platformfüggetlen, törekszik az elemzés során minél inkább szabványos megoldások használatára. Fő célja, hogy az egyes elemző modulok könnyen beilleszthetőek legyenek elemzési láncokba (letöltöm és már használom is) és hogy a felhasználó számára megkönnyítse a leginkább megfelelő komponens kiválasztását (azonos feladatot ellátó komponensek gyorsan cserélhetőek).

A keretrendszer lehetőséget ad egy komplex probléma kisebb részproblémákra történő szétbontására, mint például: mondatra bontás, tokenizálás, tulajdonnévfelismerés. Minden feldolgozási egység egy meghatározott interfészt implementál (Java vagy C++ nyelven), a keretrendszer felügyeli az elemzési lánc összeállítását és futtatást, gondoskodik az egységek közötti adatáramlásról, performanciamérésről stb. A programozónak csak az adott modul megírására kell fókuszálnia, minden egyebet a keretrendszer hajt végre.

2 Magyar nyelvi elemző modulok

A Szegedi Tudományegyetem Informatikai Tanszékcsoportjánál elkészítettünk egy magyar nyelvi elemző láncot JAVA programozási nyelven. A munka elsősorban meglévő JAVA nyelvű modulok magyar nyelvre adaptálásából és létező magyar nyelvi modulok „JAVA-sításából" állt. A JAVA nyelvű modulok egyrészről könnyedén beilleszthetőek az utóbbi években népszerűvé vált UIMA keretrendszer alá, másrészről könnyen építhetőek be webes alkalmazásokba (például Google Web Toolkit).

[1] http://incubator.apache.org/uima

Az elemzési folyamat első lépése a szöveg mondatokra bontása, ehhez a Northwestern University nyelvi csomagjának (MorphAdorner) [2] *SentenceSplitter*ét használtuk, kiegészítve a beépített szótárat azon speciális magyar rövidítésekkel, amelyek után bár a szövegben . áll, mégsem mondatvégek. Ilyen például a *zrt.*, a *szül.* vagy a hónapnevek rövidítései. Második lépésben a mondatokon belüli tokenek azonosítása történik, szintén a MorphAdorner-ben található *Tokenizer* segítségével.

Az így kapott tokenek morfológiai elemzése a magyar nyelvre készült, szintén szabad forrású, Hunspell [3] rendszer JAVA-sított verziójával történik. A lehetséges morfológiai kódok halmazából a szövegben betöltött szerep (szófaji kódok és szótövek) kiválasztásához a Stanford Maximum Entrópia POS taggert [4] tanítottuk a Szeged Korpuszon.

Ezeken felül UIMA modulként is használható a magyar nyelvű újsághíreken tanult tulajdonnév-felismerő algoritmusunk. Ez saját paraméterezhető jellemzőkészletet és a MALLET Conditional Random Fields implementációt[2] használja.

Az így megalkotott elemzési lánc segítségével lehetővé vált magyar nyelvű szövegek standard elemzési eszközökkel történő feldolgozása, illetve egyszerűbbé válik egy feladatot megoldó különböző algoritmusok cseréje és tesztelése.

Köszönetnyilvánítás

A kutatást – részben – a TEXTREND projekt (Jedlik Ányos program) keretében az NKTH támogatta.

Hivatkozások

1. Gotz, T., Suhre, O.: Design and implementation of the UIMA Common Analysis System, IBM Systems Journal (2004)
2. Kumar, A.: MONK Project: Architecture Overview. Technical Report of the Northwestern University (2009)
3. Németh, L., Halácsy, P., Kornai, A., Trón, V: Nyílt forráskódú morfológiai elemző. Magyar Számítógépes Nyelvészeti Konferencia (2004)
4. Toutanova, K., Klein, D., Manning, C., Singer, Y. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In: Proceedings of HLT-NAACL (2003) 252-259

[2] http://mallet.cs.umass.edu/