Magyar Webkorpusz II.

Halácsy Péter[1], Kornai András[1], Németh Péter[2], Varga Dániel[1]

[1] Budapesti Műszaki Egyetem, Média Oktató és Kutató Központ,
{hp, kornai, daniel}@mokk.bme.hu
[2] Kitchen Budapest, spacecadet@kitchenbudapest.hu

1. Bevezetés

Az alábbiakban bemutatjuk a *Magyar Webkorpusz* készülő új, második kiadását. A első Magyar Webkorpusz [2] elkészülte óta feldolgozási láncunk minden fontos elemét továbbfejlesztettük, és a láncba integráltuk morfológiai egyértelműsítő rendszerünket. A második kiadás létjogosultságát technológiai fejlesztéseinken kívül természetesen az is indokolja, hogy 2003 ősze, az első Webkorpusz anyagának begyűjtése óta a magyar Web mérete lényeges mértékben nőtt, és a zsánerek arányai is jelentősen megváltoztak. Rövid absztraktunkban csak a legfontosabb új fejlesztéseket emeljük ki.

2. Crawling, Tokenizálás, Nyelvazonosítás

A nyers weboldalak begyűjtéséhez ezúttal a WIRE crawlert [1] alkalmaztuk. A WIRE nagy teljesítményű, erősen párhuzamosítható működésű webcrawler. Elemi duplikátum-szűréssel rendelkezik, amit saját, nyelvfeldolgozásra hangolt duplikátumszűrőnkkel egészítettünk ki.

A Szószablya projekthez kifejlesztett tokenizáló és mondatra szegmentáló rendszerünket véges állapotú technológiára alapozva újraimplementáltuk, lényegesen felgyorsítva ezzel.

A .hu domainből kinyert dokumentumok nem elhanyagolható százaléka nem magyar nyelvű. A dokumentumok nyelvének azonosításához szintén egy véges állapotú technológián alapuló rendszert építettünk. Ennek teljesítménye is lényegesen nagyobb, mint az első korpusz építésekor alkalmazott megoldásé.

3. Morfológiai egyértelműsítés

Morfológiai egyértelműsítőnk [3] a morphdb.hu magyar morfológiai erőforrást [4] felhasználva dolgozik. Rendszerünk sebessége megfelelő ahhoz, hogy a teljes Webkorpuszt feldolgozásnak vethessük alá. Hangsúlyozzuk, hogy az egyértelműsítés nem csupán szófaji azonosítást jelent: minden tokenhez részletes morfológiai információt rendelünk, képzést és produktív szóösszetételek felbontását

is beleértve. A rendszer elfogadható pontossággal oldja meg a sem lexikonja, sem tanítókorpusza által nem ismert szavak elemzésének (guessing) nehéz feladatát is.

4. A morfológiai erőforrás

morphdb.hu morfológiai erőforrásunk fejlesztésében az eredeti Webkorpusz felbecsülhetetlen segítséget nyújtott. Elsősorban természetesen oly módon, hogy lehetőséget adott gyakori, de a morfológiai erőforrás által mégsem ismert szóalakok megtalálására. Megjegyezzük, hogy a Webkorpusz leggyakoribb szóalakjai között is előfordul idegen (főként angol) nyelvű szó, elgépelés, helyesírási hiba és ékezethiány. Az angol nyelvű szavak automatikus kiszűrésére lehetőséget adott az angol nyelvű morphdb.en erőforrásunk használata. Az ismeretlen, de nem angol nyelvű szavak leggyakoribbjainak átvizsgálását, és az indokolt esetekben erőforrásba felvételét elvégeztük. Ezen a ponton elmondható, hogy az első Webkorpusz 60,000 leggyakoribb szóalakját a morphdb.hu helyesen elemzi. A morphdb.hu erőforrás fedésének növelése visszahat az új Webkorpusz minőségére, amennyiben javítja a morfológiai egyértelműsítés és a nyelvfelismerés pontosságát.

5. Webes keresőfelület korpusznyelvészeknek

A korpuszból épített, szógyakoriság-listát nyers és morfológiailag egyértelműsített változatában publikussá tesszük. De ezen túl építettünk a gyakoriságlistához egy olyan web-alapú kereső-felületet, amely minden a korpusznyelvészek által hagyományosan alkalmazott keresési feltételt és rendezési elvet támogat.

6. Tervbe vett fejlesztések

Technológiáink jelentős része nyelvfüggetlen. Ezért természetesen adódik az a célkitűzés, hogy a webkorpusz-építést a környező országok miénkhez hasonló méretű webes jelenléttel bíró nyelveire is elvégezzük. Ez a munka az absztrakt írásának pillanatában cseh nyelvre zajlik, de reményeink szerint több más nyelvre is elvégezzük majd.

Terveink között szerepel továbbá a szógyakoriság-lista interaktív keresőfelületének kiterjesztése olyan módon, hogy szövegkörnyezet-információt is indexeljen és kereshetővé tegyen.

Hivatkozások

1. Carlos Castillo and Ricardo Baeza-Yates. Wire: an open-source web information retrieval environment. In *Workshop on Open Source Web Information Retrieval (OSWIR)*, 2005.

2. Péter Halácsy, András Kornai, László Németh, András Rung, István Szakadát, and Viktor Trón. Creating open language resources for Hungarian. In *Proceedings of Language Resources and Evaluation Conference (LREC04)*. European Language Resources Association, 2004.
3. Péter Halácsy, András Kornai, Csaba Oravecz, Viktor Trón, and Dániel Varga. Using a morphological analyzer in high precision POS tagging of Hungarian. In *Proceedings of LREC 2006*, pages 2245–2248, 2006.
4. Viktor Trón, Péter Halácsy, Péter Rebrus, András Rung, Péter Vajda, and Eszter Simon. Morphdb.hu: Hungarian lexical database and morphological grammar. In *Proceedings of LREC 2006*, pages 1670–1673, 2006.