Magyar szóalak- és morfológiaielemzés-adatbázis

Szidarovszky Ferenc P.[1], Tóth Gábor[1], Tikk Domonkos[2,3]

[1] F12 Kft., 1025 Budapest, Szépvölgyi út 191.
{ferenc.szidarovszky, gabor.toth}@f12.com
[2] Gravity Research&Development Kft., 1101 Budapest, Expo tér 5–7.
domi@gravityrd.com
[3] Budapesti Műszaki és Gazdaságtudományi Egyetem, Távközlési és Médiainformatikai Tsz, 1117 Budapest, Magyar Tudósok krt. 2.
tikk@tmit.bme.hu

Kivonat: Célunk egy olyan morfológiai elemző megoldás létrehozása, mely átlagos felhasználás mellett a szavak nagy arányát tudja elemezni, megengedve a helytelen szavak „közeli" értelmezését is. Ennek a megoldásnak műszakilag platformfüggetlennek és kevés szó elemzése esetén is hatékonynak kell lennie. Ennek érdekében egy olyan statikus MySQL adatbázist építünk, mely tartalmazza a szóalakokat és azok elemzését, így a szavak elemzése adatbázis-lekérdezéssel történhet. Kellő feltöltöttséggel ez az adatbázis megvalósíthatja célunkat.

1 Bevezetés

Az elmúlt években sikerrel és nagy megelégedésünkre használtuk az OcaMorph morfológiai elemzőprogramot [1]. Funkcionalitási szempontból magyar szavak morfológiai elemzésére a legjobb megoldások egyike. Technikai szempontból azonban vannak hátrányai:

- Csak külön folyamatként lehet elindítani, nehezen és/vagy nem hatékonyan integrálható más rendszerekbe.
- Magas a kezdeti inicializálás időigénye, gyakori, de kevés szót tartalmazó elemzési feladatokra nem hatékony. (Ilyen használat merül fel pl. ajánlórendszerek esetében.)

Célunk egy olyan morfológiai elemző megoldás létrehozása, mely a fenti technikai problémákat kiküszöböli. Ezt egy olyan statikus adatbázis létrehozásával igyekszünk elérni, mely tárolja a szóalakokat és azok morfológiai elemzéseit.

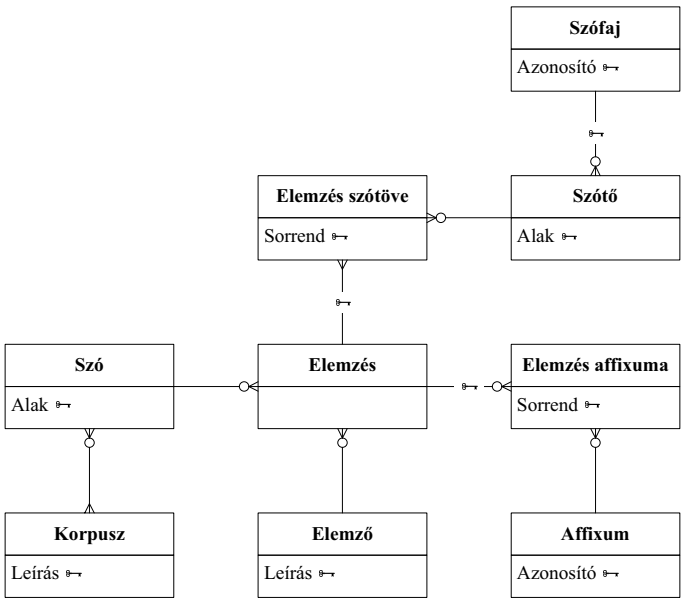
A megoldással kapcsolatos elvárásainkról fontos megjegyezni:

- A megoldástól nem várjuk, hogy teljes legyen, de törekvésünk, hogy átlagos felhasználás esetén a szóalakok minél nagyobb arányát tartalmazza.
- A megoldástól elvárjuk, hogy egy helyes szóalakra jó elemzéseket adjon, de helytelen szóalakok esetén csak annyit várunk el, hogy ha ad elemzést, akkor az alakhoz „közeli" elemzéseket adjon.
- A megoldástól nem várjuk, hogy tartalmazza az összetett szavakat. (Ezek elemzése jól visszavezethető több nem összetett szó elemzésére.)

2 Az adatbázis létrehozása

2.1 Adatstruktúra

Az adatbázis adatmodelljét az 1. ábra szemlélteti:



1. ábra. Az adatbázis adatstruktúrája

A **Szófaj** tábla tartalmazza a szófajok listáját (jelenleg 18 sor), kulcsa a szófaj azonosítója. Az **Affixum** tábla tartalmazza az affixum fajták listáját (jelenleg 137 sor), kulcsa az affixum azonosítója.

A **Korpusz** tábla tartalmazza a korpuszok listáját (jelenleg 3 sor), kulcsa a korpusz leírása. A **Szó** tábla tartalmazza az eddig talált elemezhető szóalakokat (jelenleg 2 300 717 sor), kulcsa az alak. A korpuszokat és a bennük megtalálható szavakat összekapcsoljuk.

A **Szótő** tábla tartalmazza az eddig talált szótövek listáját (jelenleg 199 822 sor), kulcsa a kapcsolódó szófaj és az alak párosa.

Az **Elemző** tábla tartalmazza a morfológiai elemzők listáját (jelenleg 1 sor), kulcsa az elemző leírása. Az **Elemzés** tábla tartalmazza a tárolt elemzések listáját (jelenleg 3 881 689 sor), kapcsolódik hozzá az elemző, és az elemzett szó.

Az **Elemzés szótöve** tábla (jelenleg 4 671 757 sor) tartalmazza a kapcsolódó elemzés által megadott szótöveket sorrendben. Az **Elemzés affixuma** tábla (jelenleg 9 543 740 sor) tartalmazza a kapcsolódó elemzés által megadott affixumokat sorrendben.

Mint látható, az adatmodellt felkészítettük a korpuszok szétválasztására és a jövőbeli esetlegesen előforduló többféle morfológiai elemző együttes kezelésére.

2.2 Feltöltés

Az adatbázis feltöltése az OcaMorph [1] felhasználásával történt úgy, hogy különböző korpuszok szavait leelemeztettük az OcaMorph-fal, és a kapott elemzéseket betöltöttük az adatbázisba.

Az alábbi korpuszok kerültek feldolgozásra:

- Web korpusz 2.0 [2, 3]
- Magyar wiki korpusz [4]
- Saját, 368 könyvből/regényből álló, az internetről letöltött korpuszunk.

3 Eredmények

3.1 Az adatbázis

Létrejött egy statikus (MySQL) adatbázis, mely:

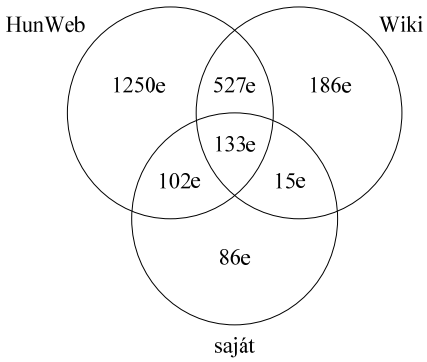
- platformfüggetlen;
- jól integrálható meglévő rendszerekbe;
- gyakran végrehajtásra kerülő, de kevés szó elemzését igénylő feladatokra is hatékony.

További előnye, hogy az elemzések egyszerre, hatékonyan állnak rendelkezésre, így alkalmassá váltak statisztikai elemzések elvégzésére, pl. szociolingvisztikai elemzésekhez.

3.2 Statisztikák

A fenti három korpusz feldolgozásával kb. 2,3 millió szóalak összesen kb. 3,8 millió elemzését tároltuk le. Ezek az elemzések közel 260 ezer szótőre hivatkoznak.

Az alábbi ábra szemlélteti a szóalakok korpuszokon belüli előfordulását:



2. ábra. Szóalakok korpuszokon belüli előfordulása.

Az egy szó alternatív elemzéseinek számának eloszlását az alábbi táblázat tartalmazza:

1. táblázat: Egy szóra eső alternatív elemzések számának eloszlása.

A szó alternatív elemzéseinek száma	Ilyen szavak száma
1	1 353 265
2	578 828
3	211 574
4	105 065
5	17 166
6	25 463
7	2 627
8	4 198
9	1 164
≥10	1 365

Az elemzésekben szereplő affixumok számának eloszlását az alábbi táblázat tartalmazza:

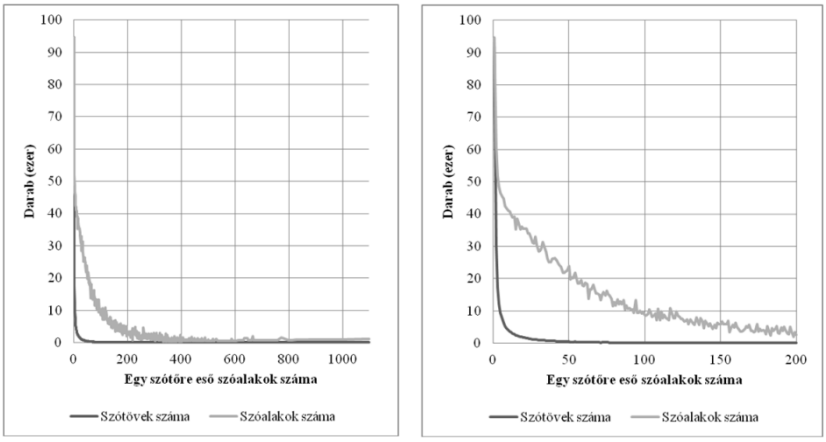
2. táblázat: Az elemzésekben szereplő affixumok számának eloszlása.

Elemzésben szereplő affixumok száma	Ilyen elemzések száma
1	1 106 984
2	798 212
3	896 217
4	468 085
5	238 277
6	119 013
7	30 890
8	15 925
9	2 183
10–12	1 034

Az elemzésekben közel 20 ezer különböző affixumsorozat szerepel.

A legtöbb különböző szóalak az *út* szótőhöz tartozott, összesen 1098. Az öt legtöbb különböző szóalakkal rendelkező szótő az *ad*, *gond*, *név*, *szó* és *út* voltak.

A 3. ábra mutatja, hogy hogyan alakul a szótövek, illetve szóalakok száma az egy szótőhöz talált különböző szóalakok számának függvényében:



3. ábra. Szótövek, illetve szóalakok száma az egy szótőhöz talált különböző szóalakok számának függvényében.

4 Jövőbeli tervek

4.1 További korpuszok bedolgozása

Tervezzük az adatbázis bővítését további korpuszok 1.2 pontban leírtak szerinti feldolgozásával.

Ennek első lépéseként learattuk az Országos Széchenyi Könyvtár online elérhető anyagait, ezek feldolgozásának előkészületei jelenleg folynak.

4.2 Szóalakok generálása

Vizsgáljuk egy ragozómotor kialakításának lehetőségét, mely egy szótőből és egy affixumsorozatból szóalakot képezne. Egy ilyen motorral korpusz nélkül lehetne célzottan bővíteni az adatbázist. A ragozómotor kialakítását segíti, hogy – amint a Bevezetőben is említettük – nem teljességre törekszünk, hanem a gyakorlati felhasználhatóság támogatására.

Az eddigi statisztikák alapján az adatbázis bővítése az eddig talált összes szótővel és alkalmazható affixumsorozattal jelentős, de megfelelő informatikai háttérrel kezelhető feladatnak tűnik.

4.3 Performancia mérése

Az Országos Széchenyi Könyvtár letöltött anyagainak bedolgozása után meg kívánjuk mérni az adatbázis teljességi mutatóit, továbbá működési sebességét. A jelenlegi mé-

retek mellett aggregációs segédtáblázat segítségével egy szálon kb. 9 ezer szó/másodperc sebességet tudtunk elérni.

5 Konklúzió

Az előzőekben ismertetett statikus MySQL adatbázisra épülő megoldás kellő feltöltöttség esetén megvalósítja a kitűzött célokat. Jó kilátások vannak arra, hogy nagy találati arányt adó adatbázist tudjunk építeni.

Bibliográfia

1. Trón, V., Németh, L., Halácsy, P., Kornai, A., Gyepesi, G., Varga, D.: Hunmorph: open source word analysis. In: Proceedings of the ACL 2005 Workshop on Software. (2005) 77–85
2. Halácsy P., Kornai A., Németh L., Rung A., Szakadát I., Trón V.: Creating open language resources for Hungarian. In: Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004) (2004)
3. Kornai, A., Halácsy, P., Nagy, V., Oravecz, Cs., Trón, V., Varga, D.:. Web-based frequency dictionaries for medium density languages. In: Proceedings of the 2nd International Workshop on Web as Corpus (ACL-06) (2006) 1–9
4. Héder, M., Farkas, M., Oláh, T., Solt, I.: Sztakipedia – Mashing Up Natural Language Processing, Recommender Systems and Search Engines to Support Wiki Article Editing. In: Proceedings of the AI Mashup Challenge 2011 at Extended Semantic Web Conference (ESWC). Iraklion, Greece (2011)