Magyar nyelvű kérdő mondat elemző szoftver

Tikk Domonkos[1], Szidarovszky Ferenc P.[2], Kardkovács Zsolt Tivadar[1], Magyar Gábor[1]

[1] Budapesti Műszaki és Gazdaságtudományi Egyetem
Távközlési és Médiainformatikai Tanszék,
H-1117 Budapest, Magyar Tudósok krt. 2.
{tikk, kardkovacs, magyar}@tmit.bme.hu
[2] Szidarovszky Kft.
H-1392 Budapest, Pf. 283.
ferenc.szidarovszky@szidarovszky.com

Kivonat: Cikkünkben „A szavak hálójában" projekt keretében megvalósított kérdő mondatok elemzését végző programot ismertetjük. A projekt egyik célja egy olyan keresőszolgáltatás algoritmikus és architekturális feltételeinek megteremtése, amely lehetővé teszi, hogy természetes nyelvű magyar kérdésekkel internetes adatbázisok tartalmában az ún. mélyhálóban keressünk. Ezen cél elérésének egyik fontos része a kérdő mondatok szintaktikai és szemantikai feldolgozása. A bemutatott szoftver a nyelvtani szerkezetek felismerésén kívül minta- és szótáralapú névelem-detektáló feladatokat is elvégez, amelynek nagy jelentősége van a kérdéselemzés következő lépése, a kontextusfelismerés során.

1 Bevezetés

A projekt célkitűzése, hogy a felhasználók számára megkönnyítse az internetes adatbázisok tartalmában való keresést. Ezen adatbázisok tartalma, amelyet összességében *mélyhálónak* hívunk, sokszorosa az ún. felszíni világháló tartalmának, ráadásul a jellemzően relációs adatbázisokban tárolt információk pontosabbak, és hamarabb is frissülnek. A hagyományos keresőmotorok azonban nem tudják indexelni ezt az értékes információforrást, mivel a tartalmuk csak az adott internetes adatbázis felhasználói felületéről kezdeményezett keresések eredményeként, dinamikusan jelenik meg. A projektünk célja, hogy olyan keresőszolgáltatást nyújtson, amely a mélyhálós tartalomszolgáltatókkal együttműködve a szolgáltatott tartalmakat egy közös kereső platformon keresztül teszi elérhetővé és kereshetővé úgy, hogy a kereséseket *természetes magyar nyelvű mondatok* formájában lehessen megadni.

A projekt keretében megvalósuló prototípus-alkalmazás a jelenleg böngészővel közvetlenül el nem érhető, adatbázisban található tartalom egy részét kívánja elérhetővé tenni, amelyek a *könyv, film, labdarúgás* és *étterem* témakörébe esnek. Az alkalmazás csak olyan jellegű kérdésekre kísérel meg válaszolni, amelyre a válasz megtalálható a mélytartalmat szolgáltató partnerek adatbázisaiban. Ez természetesen

bizonyos megszorításokat jelent a kérdés típusára, jellegére és témájára vonatkozóan. Az alábbiakban vázlatosan bemutatjuk a szoftver működési elvét és funkcióit.

2 A program működése

A szoftver Windows platformra C++ nyelven íródott. A működéséhez különböző segédeszközöket (pl. morfológiai elemző), adattárakat használt fel (pl. névelem-tár).

A program bemenetként nem összetett, kérdőszóval kezdődő a magyar nyelvtan és helyesírás szabályainak megfelelő, tényszerű tartalomra vonatkozó kérdéseket elemez. Nem fogad el, illetve nem garantál jó elemzést eldöntendő, szubjektív, intencionális, kauzális kérdésekre. Az elemzés eredménye XML formátumú értelmezési *alternatívák* sorozata, valamint grafikusan megjelenített tokenizációs és elemzési fák. Ez utóbbiak láthatók cikkünk ábráin.

A felhasználónak lehetősége van bizonyos értelmezési opciók megadására, amelylyel elősegítheti a többértelmű szerkezetek egyértelműsítését. A felhasználó az alábbi értelmezési opciókkal segítheti a mondat helyes felismerését (ld. még ábrák).

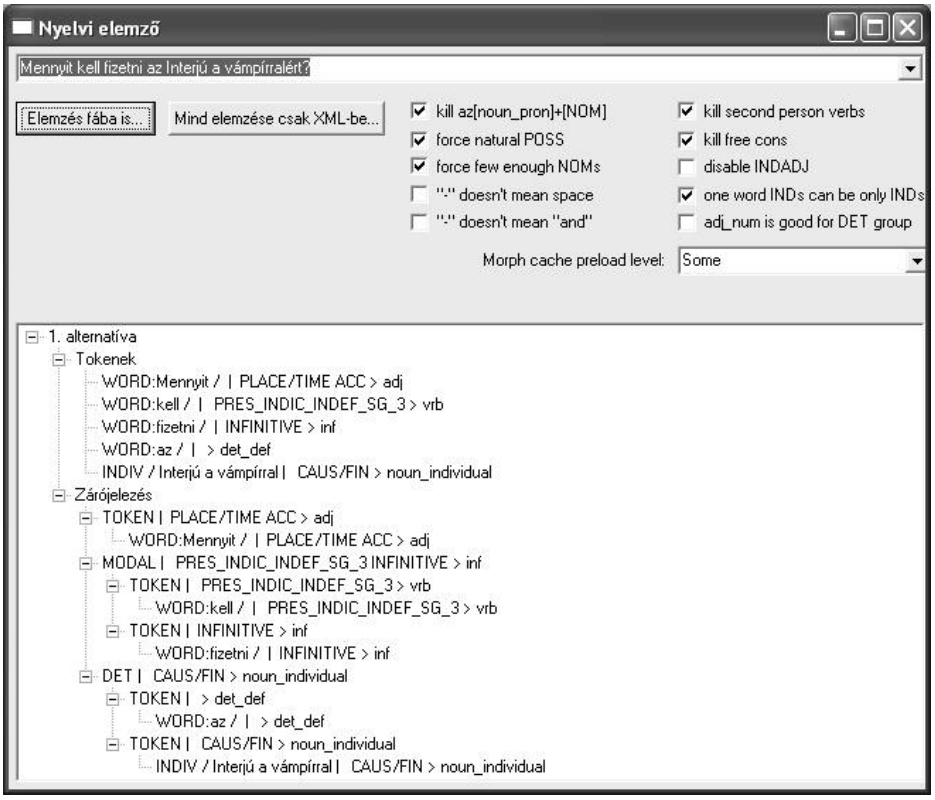
- *az* szó csak névelőt jelent (alapértelmezett); a morfológiai elemző által adott *főnév* elemzés általában hibás alternatívát generál.
- természetes birtokos sorrend preferálása (alapértelmezett); ha a birtokot közvetlenül megelőző token egyben lehetséges birtokos is, akkor más lehetséges birtokost nem keresünk.
- túl sok felső szintű alanyesetű főnév szűrése (alapértelmezett); ezzel az opcióval a 2-nél több valamilyen struktúrában nem szereplő alanyesetű főnevet tartalmazó mondatokat szűrjük ki.
- kötőjel értelmezése (2 opció); a kérdésben esetlegesen szereplő kötőjel karakterek értelmezésének megadása (szóköz; szóösszetétel; vagy „és”), amelyek hiányában az eredeti kérdésből több alternatívát készítünk.
- 2. személyű igéket tartalmazó alternatívák szűrése (alapértelmezett).
- szabad összekötőket tartalmazó alternatívák szűrése (alapértelmezett); a nem feldolgozott összekötők (és, vagy) tartalmazó mondatok eldobása.
- névelemek melléknévi szerepben (engedélyezett); *Shakespeare szonett* típusú kifejezések helyes felismerését lehetővé tevő opció.
- egy szavas névelemeket ne értelmezze csak névelemként (alapértelmezett); ha kikapcsoljuk, akkor a névelemként is szereplő közszavakat kétféleképpen tokenizálja.
- számnév elfogadása névelőként; az *egy* vagy más számnevek helyes értelmezése adható meg.

Az alternatívák elemzése lépések egymás utáni végrehajtásából áll. A lépések végrehajtása során keletkezhetnek szabálytalan alternatívák is, amelyeket a lépés végrehajtása után eldobunk.

2.1 Tokenizálás, névelemek keresése és felismerése

Az alternatívákat szavakra és vesszőkre tagoljuk. A szavak morfológiai elemzését a szoftverbe integrált Hunmorph[86] programmal végezzük. A szavak különböző morfológiai elemzései multiplikatíve alternatívákat generálnak. Például egy 5 szóból álló mondat esetén, ha minden szónak két alternatív morfológiai elemzése van, akkor 32 lehetséges mondatalternatívát generálunk.

A tokenizált alternatívákon először a szótár-alapú névelem felismerést végzünk (részleteket ld. [1]). Amennyiben valamely szó, vagy szósorozat névelemnek bizonyul, akkor a speciális „névelem" címkével látjuk el, amely függetlenül a benne szereplő szavak számától *egy* token lesz. Ha valamely token nem csak névelemként értelmezhető, akkor több alternatívát készítünk. Minta alapján az alábbi névelemeket ismerjük fel: postai címek; URL-k, e-mail címek; a pénzmennyiség; a dátumok/időpontok.



1. ábra: Példa többszavas névelem tokenizálására

Már a tokenizáció fázisában elvégzzük az alábbi szűréseket, amellyel csökkenthetjük a lehetséges alternatívák számát:

[86] http://mokk.bme.hu/eszkozok/hunmorph/

- összetett mondatok szűrése: ha több finite ige van a mondatban, az alternatívát eldobjuk. Ha nincs benne finite ige, akkor a „van” igét beszúrjuk.
- kérdőszó vizsgálat: ha az első szó nem megengedett kérdőszó, akkor az alternatívát eldobjuk

Az 1. ábrán egy többszavas névelem tokenizálását illusztráljuk.

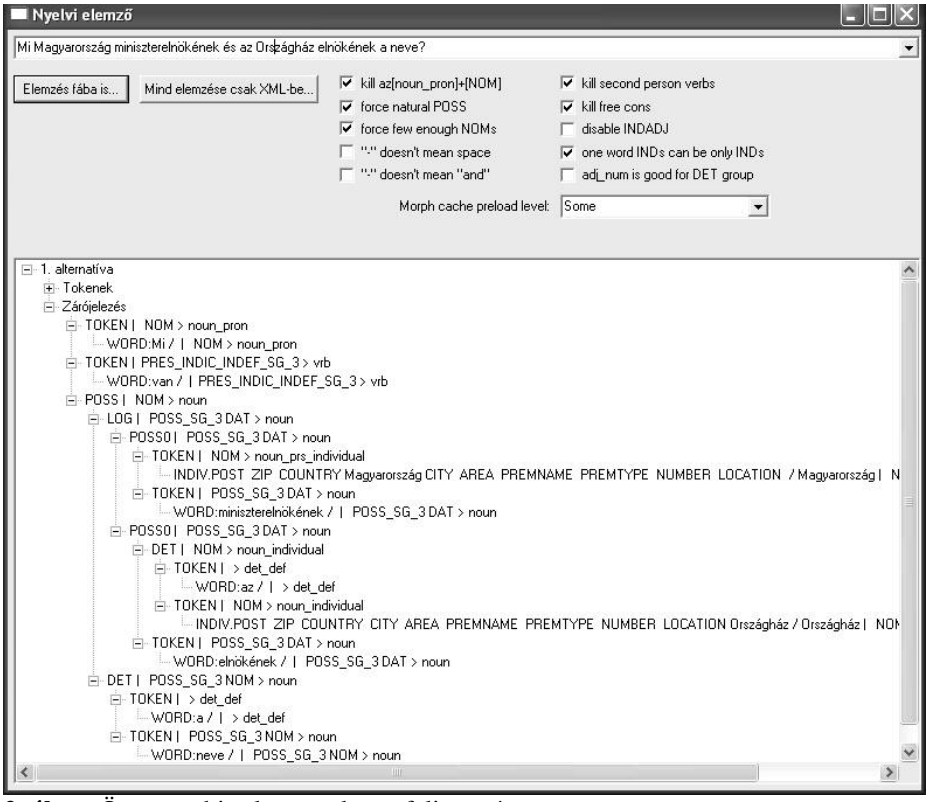
2.2 Zárójelezés

A zárójelezés célja, hogy az alternatívák tokenjeit egymásba ágyazott csoportokba ossza. A csoportok elemei szemantikai kapcsolatban lévő és így csoportban egységet képező (nem feltétlen szomszédos) tokenek.

Az alábbi csoportokat különböztetjük meg:

- igekötős szerkezetek; az igekötő [prv] címkével ellátott szavakhoz keresünk finite, nem utólagosan beillesztett igét. Elváló igekötők és az ige összekapcsolására alkalmas. Az ige nélküli igekötőt tartalmazó mondatalternatívákat eldobjuk.
- főnévi igeneves szerkezetek (ld. 1. ábra); [inf] címkéjű szavakhoz keresünk olyan segédigét vagy egyéb szót, amellyel a modalitást jellemző csoportot tudunk alkotni.
- főnévi csoport (jelzős szerkezetek), ahol a jelző melléknév vagy szótári névelem;
  - *-(j)ú,-(j)ű* ragú melléknévi szerkezetek felismerése; ekkor olyan szegmenseket keresünk, amelyek {[noun]+NOM}{[adj]_jÚ}{[noun]} alakúak, pl. *Mátrix című filmet.*
  - egyéb melléknévi szerkezetek felismerése; a zárójelezés feltétele, hogy valamely főnév [noun] előtti szó melléknév [adj], szótári névelem (ld. még opciók), vagy már korábban alkotott jelzős csoport legyen. Az eljárást rekurzívan végezzük.
- névelős szerkezetek; főnév, vagy főnévi csoport előtti névelőt [det] a csoporthoz kapcsoljuk.
- névutós szerkezetek; névutóból [post] és előtte álló főnévi csoportból képezzük.
- logikai összekapcsolások (és, vagy); a logikai kapcsoló két oldalán azonos morfológiai jellemzőkkel ellátott tokenekből (csoportokat) logikai csoportot képezünk. Azokat a felsorolásokat, ahol csak az utolsó két tag között szerepel kötőszó, a többi tag között pedig vessző, többszintű logikai csoportként ismerjük fel, ha a fenti feltételek fennállnak.
- birtokos szerkezetek; legfeljebb 3 elemű birtokos szerkezeteket ismerünk fel a morfológiai elemző által megadott birtokosra és birtokra jellemző toldalékok alapján (ld. 2. ábra.). A keresést a lehetséges birtokkal kezdjük, és ahhoz illesztjük a birtokosok láncát.
- értelmező jelzős szerkezetek; feltétele, hogy vessző előtt és után ugyanolyan morfológiai jellemzőjű csoport álljon. Ekkor két alternatívát gyárt, az egyikben csak az értelmezett kifejezés szerepel (vessző előtti rész), a másikban pedig csak az értelmezés (vessző utáni rész).

Ha valamely csoportképzés nem egyértelmű, akkor több alternatívát készítünk. A csoportok előfordulásainak keresését célszerűen kialakított, rögzített sorrendben végezzük. Vannak olyan csoportok, melyek keresése a rögzített sorrendben többször szerepel.



2. ábra: Összetett birtokos szerkezet felismerése

A program utolsó fázisa a szűrés, amikor a valamely szempontból hibás alternatívák eldobásra kerülnek. Az eldobás indokát az XML kimenetben feltüntetjük, ahol a teljes elemzési folyamatot is végig lehet követni.

3 Köszönetnyilvánítás

A cikk a Nemzeti Kutatási és Fejlesztési Pályázatok NKFP-0019/2002 jelű projektjének támogatásával készült.

Irodalomjegyzék

[1] D. Tikk et al: Ismert névelemek felismerése és morfológiai annotálása szabad szövegben, In: III. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, (2005), ebben a kötetben.