MASZEKER: projekt szemantikus keresőtechnológia kidolgozására

Szőts Miklós[1], Csirik János[2], Gergely Tamás[1], Karvalics László[3]

[1]Alkalmazott Logikai Laboratórium
1022 Budapest, Hankóczy J. u. 7
{szots, gergely}@all.hu
[2] Szegedi Tudományegyetem, Informatikai Tanszékcsoport,
Szeged, Árpád tér 2.
csirik@inf.u-szeged.hu
[3] Szegedi Tudományegyetem, Könyvtár- és Humán Információtudományi Tanszék,
Szeged, Egyetem u. 2.
zkl@hung.u-szeged.hu

Kivonat: Egy merész nyelvészeti projektről számolunk be, a MASZEKER szemantikus keresést megcélzó projektről, amelyen az Alkalmazott Logikai Laboratórium és a Szegedi Tudományegyetem közösen dolgozik. A cél olyan technológia kidolgozása, amely a jól formált keresőkifejezés jelentésreprezentációját illeszti a szövegekre olyan egyezést keresve, amely kifejezheti a keresőkifejezés jelentését. Két felhasználási területre, mégpedig a szabadalmi keresésre, valamint néprajzi keresésre prototípus rendszert kívánunk fejleszteni. A technológiát nyelvfüggetlennek tervezzük, természetesen egyes komponenseinek nyelvfüggőnek kell lenniük. Angol és magyar nyelvű változatot fogunk fejleszteni. Magát a keresést végző rendszert kiegészítik az archívumot feldolgozó modulok (tematikus klaszterezés, témafüggő szinonimagenerálás).

1 Bevezetés

Annak ellenére, hogy a Google látszólag „egyeduralkodóvá" vált a keresőrendszerek piacán (vagy tán épp ezért) folyamatosan „forró terület" a nagyobb tudású (vagy akár új elvű) keresők fejlesztése. Ezért az Alkalmazott Logikai Laboratórium és a Szegedi Tudományegyetem Informatikai Tanszékcsoportja, valamint Könyvtár- és Humán Információtudományi Tanszéke közös projektet (TECH_08_A2/2-2008-0092) indított az NKTH támogatásával.

A tervezett projekt célja egy olyan, új elveken alapuló integrált keresőrendszer, a MASZEKER kifejlesztése, amely adaptált (statisztikai és szimbolikus alapú) technológiák és újszerű megoldások kombinálásán keresztül a keresést végző felhasználó szemantikai kompetenciáját az eddigieknél nagyobb mértékben kiaknázva teszi lehetővé a természetes nyelvi dokumentumtárakban (szövegekben) történő valóban *tartalmi* keresést. Egyszerűen szólva: a felhasználó jól formált frázisokkal, mondatokkal specifikálhatja, milyen tartalmú dokumentumokat keres.

A projekt során kifejlesztett technológia magja *nyelvfüggetlen*, a rendszer prototípusát pedig magyar és angol nyelvű szabadalmi leírások, illetve néprajzi anyagok feldolgozására fejlesztjük ki.

2 State of art

A bevezetőben említett „forró terület" látképéből minket a szemantikai keresők érdekelnek. Természetesen – mint annyi szakszó az informatikában – a „szemantikai" is a lehető legkülönbözőbben értelmezhető. Sokan a szavak, szóösszetételek szintjén értelmezik: szavak közti jelentésösszefüggések feltárásával egészítik ki a kulcsszó szerinti keresést. Ilyen a már elterjedt látens szemantika algoritmusa[1] (l. [5]). Elterjedőben van a keresők valamilyen ontológiához, tezauruszhoz való kapcsolása, ilyen alapon működik a magyar fejlesztésű, de nemzetközi hírnevet szerző HealthMash kereső is (l. http://www.weblib.com/products/healthmash). A MEDLINE-on működő KLEIO kereső (ismertetőt találhatunk [2]-ben) szintén ontológiákhoz van kapcsolva, de a névelemfelismerés (NER) technikáját is használja. A keresőkifejezésben megengedi, hogy a kulcsszavakhoz a felhasználó megadja annak besorolását, pl. *PROTEIN*:cat. Már ezzel is jelentősen javítja a keresés recallját, amint az idézett példa is illusztrál. Mi azonban szemantikai keresés alatt olyan folyamatot értünk, amely összefüggő szövegrészek jelentése alapján ítél valamely dokumentumot relevánsnak.

A szemantikus keresők két nagy osztályba sorolhatóak (l. [1]): lehetnek statikusak vagy dinamikusak. A statikus keresők előre elkészítik a keresett honlapok, dokumentumok szemantikus reprezentációját, és felindexelik azokat; míg a dinamikusak a keresőkifejezés jelentésreprezentációját a keresés alatt elemzett szövegrészekre illesztik. Másik általános osztályozási szempont az, hogy témafüggetlenek vagy egy területre specializáltak. Csak néhányat sorolunk itt fel, egy teljesebb áttekintés letölthető a www.maszeker.hu oldalról.

A HAKIA (l. [8]) általános célú, ontológiai szemantikára (l. [9]) alapozott, statikus keresőrendszer. Honlapok szövegei jelentésreprezentációjának alapján előre elkészíti a lehetséges kérdésekre adható válaszokat, amelyek közül az adekvátat a keresés közben csak ki kell választania. Inkább a tudáskinyerés területéhez tartozik, de a szemantikus keresés általában könnyen átfogalmazható tudáskinyerésre. A HAKIA egy erre a célra kifejlesztett, 8 500 fogalmat tartalmazó ontológiára támaszkodik. Ehhez csatlakozik egy kb. 100 000 szójelentést és több mint 1 000 000 szót tartalmazó szótár.

A Cognition (l. [3]) egy átfogó NLP framework, amely egy témafüggetlen keresőmotort is tartalmaz; szintén statikus rendszer. Több, egy-egy területre vagy dokumentumhalmazra specializált alkalmazása van, pl. a Wikipédiára, illetve a MEDLINE abstracts-ra is kifejlesztettek egy-egy speciális keresőt. Ontológiája 7 500 fogalmat tartalmaz, amelyekhez 536 000 szójelentés kapcsolódik.

A Powerset a Cognitionhoz hasonló rendszer. Sok információnk nincs róla, mivel a Microsoft megvette, és beépítette a fejlesztés alatt lévő keresőjébe (l. [10]).

[1] Részletes ismertetése letölthető a www.maszeker.hu honlapról.

Az UpTake (l. [14]) egy utazási információkat szolgáltató kereső, amely több mint 5 000 honlapot indexelt fel. Jellegzetessége, hogy a felhasználóval folytatott párbeszédet támogat, azaz az általánosabb kéréstől a specifikusabb felé mozoghat a felhasználó. Azt tervezik, hogy a rendszer alapjául szolgáló ontológiát tanulóalgoritmusokkal bővítik.

A GoWeb (l. [4]) az élettudományokra specializált kereső. Természetes nyelvű kifejezést fogad el inputként, s egy tradicionális, kulcsszó szerinti keresés eredményeit veti alá szemantikus elemzésnek. Háttere a Gene és a MeSH ontológia. Az eredményhez ezeknek az ontológiáknak releváns részleteit is megmutatja. E leírásból is kitűnik, hogy a GoWeb dinamikus kereső.

A MEDIE (l. [2], [7]) a már említett KLEIO-hoz hasonlóan a MEDLINE-on keres; azonban a KLEIO-hoz képest jelentős előlépés, hogy már szintaktikus és szemantikus elemzést alkalmaz az események kinyerésére. Egyelőre csak *alany-ige-tárgy* alakú kereső kifejezéseket kezel. [2] beszámol további kutatási irányokról, amelyek hasonlóak a mieinkhez.

3 A MASZEKER kereső felépítése

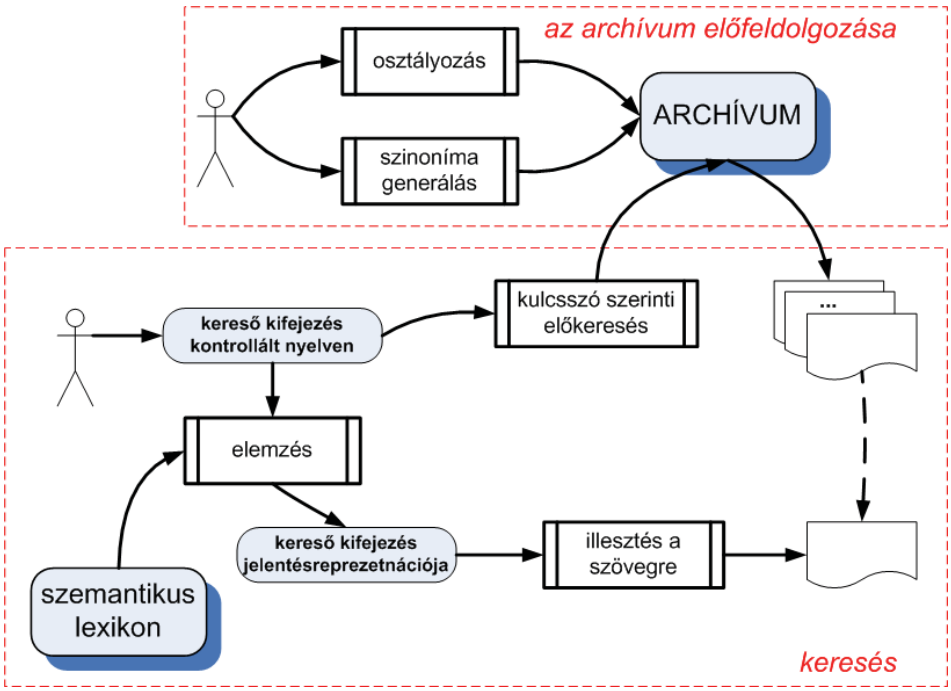
Ha a fent vázolt „tájképbe" illesztjük koncepciónkat, a következőképpen foglalhatjuk össze:

- általában nagyméretű ontológiákra épülnek a szemantikus keresők – mi egy kisméretű általános csúcsontológiát és ehhez csatlakozó, ugyancsak kisméretű tárgykörfüggő felső ontológiákat kívánunk használni;
- ennek megfelelően, – bár általános technológiai vázat építünk, – témakörökre kiélezett, tehát vertikális rendszereket kívánunk létrehozni;
- dinamikus keresőt tervezünk, bár bizonyos esetekben nem zárkózunk el az előzetes szemantikai feldolgozástól és felindexeléstől sem.

A rendszer áttekintő architektúrája az 1. ábrán látható.

Az ábrának megfelelően a releváns dokumentumok keresése a következő lépésekből áll:

1. a felhasználó egy kontrollált nyelven adja meg a keresőkifejezést,
2. a szintaktikus és szemantikus elemzés előállítja a keresőkifejezés jelentésreprezentációját,
3. a szavak szerinti keresés először az archívumot,
4. azokra a szövegszegmensekre, amelyekben a szavak szerinti keresés találatai vannak, illeszti a keresőkifejezés jelentésreprezentációját.



1. ábra. A MASZEKER rendszer áttekintő architektúrája.

3.1 Elemzés

A szintaktikus elemzésre egy robusztus algoritmust dolgoztunk ki, amely azokat a részeket, amelyekkel nem tud megbirkózni, átugorja. A keresőkifejezés megadására szolgáló kontrollált nyelvet azonban pontosan elemzi.

A szintaktikus elemzés két lépésben történik. Egy előfeldolgozás kijelöl bizonyos pontokat a szövegben, pl. a felsorolás elemeinek kezdetét. Ezután egy dependencia-nyelvtanon alapuló elemző fut végig a szövegen. A szabadalmi szövegekben sok kvantitatív jelző fordul elő, a legváltozatosabb formában (például: *aspirin crystals 20-60 mesh in size* vagy *about 3-10% by weight of a polymeric mixture*). Ezekre külön CFG nyelvtant dolgoztunk ki.

A szintaktikus és a szemantikus elemzés párhuzamosan történik. Ennek több oka van, a legfontosabb az, hogy a szemantikus elemzés a szintaktikus elemzés bizonytalanságait segít kezelni, azaz visszahat a szintaktikus elemzésre, sőt a POS-tagger ítéleteit is változtathatja. Ugyanis beleütköztünk olyan hibás szófaj-meghatározásba, amely eltorzítja a szintaktikus elemzést. Főleg az angol nyelvben sok az olyan szó, amely egyaránt szerepel igeként és főnévként, például az *extract* szó.

A jelentésreprezentáció kialakítását davidsoni alapokon [11] kezdtük el, azaz az igék és az eseményszerűségeket jelentő főnevek jelentését reifikáljuk: maga az ese-

mény egy token lesz, és a szereplőket kötik hozzá szereprelációk. Logikailag azt jelenti, hogy a többargumentumú relációkat áttranszformáljuk kétargumentumúakra.

A davidsoni közelítés több szempontból is kedvező. A szemantikus lexikon szempontjából célszerűbb az eseményjelentésű szavakból kiindulni, amikor a jelentéskapcsolatokat leírjuk. Rugalmasságot ad: bármikor újabb dependenssel lehet bővíteni a leírást, mivel nem kell a relációjelentések argumentumszámát meghatározni. Illeszkedik a dependenciaalapú szintaktikus elemzés eredményére – valójában az összefüggés fordított: a dependenciaalapon működő szintaktikus elemző algoritmust választottuk az eseményalapú szemantikus szerkezethez. Robusztus is: ha nem áll rendelkezésre elegendő információ, a részleges jelentésreprezentáció automatikusan előáll.

Látható, hogy a szereprelációk megfelelnek a tematikus szerepeknek [11]. A különbség annyi, hogy nem kívánunk általános tematikus szerepkészletet átvenni vagy alkotni, hanem témakörönként és kontextusokként definiálunk szereprelációkat (l. a szemantikus lexikonról szóló szekcióban az erről szóló részt). Néhány nyelvi jelenségre külön kidolgoztunk reprezentációs formalizmust, például a tagadásra, a „one of …” jellegű kifejezésekre, a tulajdonságok kifejezésére.

Az igénypont szakaszban a legnagyobb problémát a koordinációk, ill. a felsorolások detektálása jelenti, többször találkozunk egymásba ágyazott felsorolásokkal is. Jelenleg olyan algoritmuson dolgozunk, amelyek a koordinált frázisok hasonlósága alapján rendeli egymás mellé a megfelelő frázisokat. Nemcsak morfológiai, szintaktikai ismérveket veszünk figyelembe, hanem szemantikusakat is. Például tipikusak azok a felsorolások, amelyek valamely szabadalmazandó gyógyhatású készítmény összetételét adják meg, ilyenkor anyagmennyiségek vannak megadva.

A szintaktikus elemzés nemcsak párhuzamosan működik a szemantikussal, hanem párhuzamosan is fejlesztjük. Ezzel elkerüljük, hogy olyan problémába ütközzünk, mint amilyenről [2] beszámol, tudniillik, hogy a MEDIA esetében az elkészült HSPG nyelvtanhoz problémás hozzáilleszteni egy szereprelációkra alapozott jelentésreprezentációt.

3.2 Szemantikus lexikon

Ennek megfelelően a szemantikus lexikonunkban is a szintaktikus és szemantikus információk párhuzamosan lesznek elrendezve, például a vonzatkeretekkel együtt a megfelelő tematikus szerepek. A szemantikus lexikon kulcsfontosságú az elemzéshez. Mint írtuk, nem óriás ontológiát akarunk építeni vagy kölcsönözni. E helyett alkalmazunk egy általános csúcsontológiát (lényegében a DOLCE-ből [6] kölcsönözve), és ehhez kapcsolódnak témakörönként és kontextusokként szigetszerű ontológiák. Az ontológiák osztályai alatt szinonimahalmazok lesznek. Így egy háromrétegű lexikont kapunk, ahol a nyelvi elemek képezik a nagy tömegű információt, a felettük lévő ontológia pedig definiálja azokat az osztályokat, amelyekbe a szinonimahalmazok tartoznak, illetve meghatározza azokat a relációkat, amelyek szerepelhetnek a jelentésreprezentációban.

[2] beszámol arról, hogy a japán fejlesztésű MEDIE továbbfejlesztése is a szereprelációk bevonásával történik, azonban ők egy általános szerepreláció-készletet kívánnak alkalmazni. Mi célszerűbbnek találjuk több, de egyszerűbb szerepreláció-

készletet alkalmazni. Például a *kezel/treat* igéhez nemcsak más vonzatok társulnak, ha gyógyászati készítmények alkalmazásának témakörében használjuk (*treating a patient with a disease* vagy *treating a disease in a patient*[2]), vagy az előállításukban (*treating something with a material*), hanem más szereprelációk is. A *with* prepozíció az első esetben egy „kedvezőtlen állapot" szerepet játszó fogalmat kapcsol az eseményhez, a második esetben pedig „eszköz"-t. A példából az is látszik, hogy gyakorlati, alkalmazási szempontból szabadon eltérünk a nyelvészetben használt tematikus szerepektől, – ez is azt teszi lehetővé, hogy a szemantikus lexikon szerkezetét a második réteg kontextusok szerint is tagolja.

A szinonima fogalmát tágabban értelmezzük, mint szokásos: nem a kifejezések felcserélhetősége az ismérv, hanem az, hogy azonos szituációt/objektumot írnak-e le. Például a *kap* és *ad* szinonim lesz, a vonzatkeret különbözőségét a szereprelációk egyenlítik ki. Ebből következően a szavakat a vonzataikkal együtt kell szerepeltetni; a párhuzamos szintaktikai elemzés miatt a vonzathoz a nekik aktuálisan megfelelő szereprelációkat is hozzá kell rendelni. Sőt, amikor a vonzatok csak bizonyos osztályból kerülhetnek ki, ezeket is.

Nemcsak a szinonim kifejezések lesznek illeszthetőek, hanem azok is, amelyek valamilyen módon implikálják a jelentésreprezentációban szereplőt. Ilyen implikációs viszony a *fajtája* reláció (például az *ékszer* szóhoz illeszthető a *gyűrű*), de nem csak ez. Ilyen a *szükségszerűen következik* reláció is – például, ha a kereső kifejezésben az *érintkezik* ige szerepel, az *irritál* illeszthető hozzá. Természetesen tagadás esetén a szükségszerű következményen alapuló implikációs viszonyok megfordulnak. Tehát a szinonimahalmazok mind a fajtája, mind a szükségszerű következmény relációk szerint rendezve vannak.

3.3 Keresés

A kulcsszó szerinti keresés eredményéül kapott dokumentumokon folyik a szemantikus keresés. Kijelöltetnek azok a szövegszakaszok, amelyekben kulcsszavak szerepelnek, és ezekre kísérli meg rendszerünk a keresőkifejezés jelentésreprezentációjának illesztését.

A keresőkifejezés jelentésreprezentációjának illesztése elvileg háromféle módon hajtható végre:

- generálható a kijelölt szövegszegmens jelentésreprezentációja, és hasonlóságot keresünk a keresőkifejezés jelentésreprezentációjával;
- a szövegszegmenst csak szintaktikusan elemezzük, és a szemantikus lexikon segítségével az algoritmus azt állapítja meg, hogy a szöveg kifejezései és a köztük lévő szemantikus reláció illenek-e a jelentésreprezentációra;
- a szövegszegmens elemzését a keresőkifejezés jelentésreprezentációja vezérli egy rekurzív algoritmussal.

[2] Tisztán pragmatikus okokból a fenti frázisokban a *with* és *in* prepozíciókkal jelzett vonzatokat az igéhez kötjük, nem a főnevekhez.

Az első megoldás nyilvánvalóan pazarló. A harmadik változatot választjuk, bár lehetséges, hogy a szabadalmak igénypont szakasz közti keresés esetén a második változatot célszerű használni.

A találatokat relevancia-sorrendbe rendezzük pontosságuk szerint. Négy nagy osztályt szándékozunk megkülönböztetni:

- teljes találat,
- részleges találat,
- csak kulcsszó szerinti találat,
- ellentmondásos.

3.4 Az archívum feldolgozása

Mint az 1. ábra mutatja, a tulajdonképpeni keresési feladatot – annak megkönnyebbítése érdekében – kiegészítettük az archívum feldolgozásával. Ez két tevékenységet takar: a dokumentumok tematikus klaszterezését és osztályozását és a szakterületekre jellemző szinonimaosztályok generálását.

Több klaszterezési algoritmust kipróbáltunk. Választásunk a Cluto g1p módszerre esett, amely kísérleteinkben meglehetősen pontosnak bizonyult. A kapott eredmények: precision 89,4%, recall 99,1%, f-measure 94%.

A szinonimagenerálás során a mondatokból kiválogatott minták összehasonlítása alapján (kölcsönösinformáció-nyereség) keresünk "szemantikusan" hasonló főneveket. Igaznak bizonyult az a feltevés, hogy sokszor nem szinonim szavakat talál meg az algoritmus, hanem antonímákat, illetve olyan klasztereket, amelyekben hasonló szerepű fogalmak vannak (pl. egyesülés, bomlás, vegyülés, feloldódás). Az azonban a mi esetünkben nem baj, ha a szokottnál lazább szinonimafogalommal dolgozunk. A kísérletezés még kezdeti fázisban van, később dől el, hogyan vezérelhetjük a tanulást, illetve milyen mértékben van szükség emberi kontrollra.

4 A felhasználási területekről

A projekt két felhasználási területet vállalt fel: a szabadalmi keresést és a néprajzi információkeresést. Többé-kevésbé vakon választottuk ezt a két területet, azaz nem jól átgondolt szakmai érvek döntöttek. Azonban sikerült két olyan területet találni, amelyek a lehető legnagyobb mértékben különböznek egymástól[3]. Míg a szabadalmi keresés nagy múlttal, általánosan használt keresőrendszerrel, technológiával rendelkezik, tematikailag nagyon részletesen osztályozottak a dokumentumok, addig a néprajzi területen alapvető eszközök hiányoznak – elsősorban Magyarországon. Míg a szabadalmak legfontosabb része, az igénypont szakasz, félformális szövegnek tekinthető, a néprajzi gyűjtések feldolgozásához a szöveg normalizálásával kell kezdeni (l. [12]). Ugyanakkor a néprajz és a számítástudomány közös területe lett a narrációk kutatása, azaz a néprajzi szövegekre alkalmazható formális rendszerek kutatása. A

[3] Mind a szabadalmi keresésre, mind a néprajzi témájúra vonatkozó helyzetfelmérés, ill. követelményfeltárás letölthető a www.maszeker.hu honlapról.

célok is különbözőek: a szabadalmi kutatásban a szabadalmi bejelentéshez hasonló tartalmú dokumentumot kell keresni[4], a néprajzban motívumok, típusok szerint kell keresni. Igaz, ez utóbbiak definiálása is kutatási feladat.

A fenti különbségekből adottan az általunk fejlesztett technológia különböző módon lesz hasznosítva e két területen.

- A szabadalmi keresés területén az Európai Szabadalmi Hivatal (EPO) által rendelkezésre bocsátott speciális kereső programot (EPOQUENet) használnak. Ez természetesen kulcsszavak szerint keres. A szemantikus keresést végző modul az EPOQUNet találataiból alkotott archívumon fog működni. Képes lesz az igénypont szakaszt teljesen feldolgozni. Arra nem vállalkozunk, hogy következtetésekkel megállapítsuk a talált dokumentum viszonyát a benyújtotthoz[5], – ez mérhetetlen nagyságú és komplexitású világtudást kívánna meg. Azonban súlyt fektetünk arra, hogy a szabadalmak szövegét, ill. találatainkat strukturáltan jelenítsük meg, hogy a keresőt segítse annak áttekintésében.
- A néprajznál viszont maga a korpusz összeállítása is feladat, jelenleg magyar nyelvű hiedelem-, táltosszöveg és mesegyűjteményünk van, amely nyelvészeti feldolgozása megtörtént (l. [12] [13]). A néprajzos kutatóknak már az is nagy eredménynek számít, hogy kollokációkereső programot tudnak futtatni az anyagon (motívumkeresés). Most úgy látjuk, hogy a néprajzi keresésnél a legfontosabb annak feldolgozása lesz, hogy az egyes motívumok milyen hierarchiát alkotnak (pl. a *segítő* lehet *segítő állat*, vagy még specifikusabban *segítő kutya*), és az, hogy milyen megfogalmazásokból lehet következtetni ezek előfordulására. Például a *varázstárgyat ad* jelentésű frázisok alanya *segítő*.

A szemantikus keresés technológiájának kidolgozásánál a szabadalmi keresésre koncentrálunk, a néprajzi keresésnél a már kifejlesztett technológiát alkalmazzuk. Viszont a tematikus osztályozó modult a néprajzi anyagokon teszteljük, és szerepe a néprajzi információkeresésben lesz.

5 A projekt állása

Ebben az évben egy 0. prototípus kerül megvalósításra, amely a fontos funkciókat végrehajtja, de még az algoritmusok finomhangolása nem történik meg – azaz számos ritkábban előforduló nyelvi fordulattal nem fog megbirkózni. Hasonlóképpen a szemantikus lexikon sem a végleges szerkezetben fog rendelkezésre állni, s csak korlátozott tartalommal.

A 0. prototípus kifejlesztése nemcsak azt a célt szolgálja, hogy az algoritmusainkat teszteljük és finomítsuk, hanem azt is, hogy a jövendő felhasználókkal – jelen esetben a szabadalmi hivatal munkatársaival és a néprajzi korpuszokat feldolgozó munkatársakkal egyeztessük a keresés működését, a keresőkifejezés megadási módjait és az eredmény bemutatását. Ugyanis nemcsak magát a keresés technológiáját dolgozzuk ki, hanem olyan felhasználói interfész felületeket, amelyek a szemantikus kereséshez

[4] Nagyon elnagyolt leírás, vannak különböző, de lényegileg ehhez hasonló keresési feladatok is.

[5] A szabadalmak elbírálásánál ennek több fokozatát definiálták.

illenek. Különösen izgalmas probléma megmutatni az egyes találatoknál azt, hogyan illik a dokumentum szövege a találatra. Erre a funkcióra a szöveg grafikus megjelenítését tervezzük.

A jövő évre tervezünk egy fejlettebb prototípus változatot, amely már teljes fegyvertárral mutatja be a kifejlesztésre kerülő technológiát.

Bibliográfia

1. Abolhassani, H., Esmaili K. S.: A categorization scheme for semantic web search engines. In: 4th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA-06) (2006)
2. Ananiadou, S., Thompson, P., Nawaz, R.: Improving Search through Event-based Biomedical Text Mining. In: Darányi, S., Lendvai, P. (szerk.): Proceedings of the First International AMICUS Workshop on Automated Motif Discovery in Cultural Heritage and Scientific Comminication Texts (2010) 42–54
3. Dahlgren, K.: Technical overview of Cognition's semantic NLP (as applied to search). Technical report, Cognition Technologies, Inc. (2007) http://www.cognition.com/pdfs/Cognition_Semantic_NLP_for_Search_Overview.pdf
4. Dietze, H., Schroeder, M.: GoWeb: A semantic search engine for the life science web. In: Burger, A., Paschke, A., Romano, A., Splendiani, A. (szerk.): Proceedings of the Intl. Workshop Semantic Web Applications and Tools for the Life Sciences SWAT4LS. Edinburgh (2008)
5. Landauer, T.K., McNamara, D.S., Dennis, S., Kintsch, W.(szerk.): Handbook of Latent Semantic Analysis. University of Colorado Institute of Cognitive Science Series, Psychology Press (2007)
6. Masolo, C., Borgo, S., Gangemi, A., Guarino, N., Oltramari, A.: WonderWeb Deliverable D18: Ontology Library (2001)
7. Miyao, Y., Ohta, T., Masuda, K., Tsuruoka Y., Yosida K., Ninomiya T., Tsujii J.: Semantic Retrieval for the Accurate Identification of Relational Concepts in Massive Textbases. In: Annual Meeting - Association for Computational Linguistics (2006) 1017–1024
8. Nirenburg, S.: Homer, the author of the Iliad and the computational linguistic turn. In: Words and Intelligence II. Springer (2007)
9. Nirenburg, S., Raskin, V.: Ontological Semantics. The MIT Press (2004)
10. Montalbano, E.: Microsoft testing Kumo search engine internally. NetworkWorld, March 3, 2009. WWW document. http://www.networkworld.com/news/2009/030309-microsoft-testing-kumo-search-engine.html (accessed March 27, 2009)
11. Parsons, T.: Events in the Semantics of English: A Study in Subatomic Semantics. MIT Press, Cambridge (1990)
12. Szauter D., Vincze V., Almási A., Alexin Z., Kiss M.: Morfoszintaktikailag annotált néprajzi korpusz. In: Tanács, A., Szauter, D., Vincze, V. (szerk.): VI. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2009)
13. Szőts, M., Darányi, S., Alexin, Z., Vincze, V., Almási, A.: Semantic Processing of a Hungarian Ethnographic Corpus. In: Darányi, S., Lendvai, P. (szerk.): Proceedings of the First International AMICUS Workshop on Automated Motif Discovery in Cultural Heritage and Scientific Comminication Texts (2010) 112–115
14. UpTake under the hood—the Interview. Alt-SearchEngines, May 14, 2008. WWW document. http://www.altsearchengines.com/2008/05/14/uptake-under-thehood-exclusive-interview/ (accessed March 27, 2008)