Frame-szemantikára alapozott információ-visszakereső rendszer

Szőts Miklós[1], Gyarmathy Zsófia[1], Simonyi András[1]

[1]Alkalmazott Logikai Laboratórium
1022 Budapest, Hankóczy j. u. 7
{szots,simonyi}@all.hu, gyzsof@gmail.com

Kivonat: Egy olyan információ-visszakereső rendszert mutatunk be, amely kontrollált természetes nyelven megadható keresőkifejezésekhez keres hasonló jelentésű szövegrészt tartalmazó természetes nyelvű dokumentumokat. A rendszer frame-szemantikai elemzéssel előállítja a keresőkifejezés szemantikus reprezentációját, és azokat a dokumentumokat adja vissza találatként, amelyekben található olyan szövegrész, amelyhez a reprezentáció illeszthető. Cikkünkben ismertetjük a rendszer működését és az általa használt szemantikus reprezentációk, illetve erőforrások felépítését – elsősorban a frame-szemantika alkalmazására koncentrálva. Röviden kitérünk a még hátralévő feladatokra és a lehetséges további kutatási irányokra is.

1 Bevezetés

Az Alkalmazott Logikai Laboratórium és a Szegedi Tudományegyetem Informatikai Tanszékcsoportja, valamint Könyvtártudományi Tanszéke közös projektet (TECH_08_A2/2-2008-0092) indított az NFÜ támogatásával szemantikus információ-visszakereső rendszer kifejlesztésére. A tervezett projekt célja egy olyan, új elveken alapuló integrált keresőrendszer, a MASZEKER kifejlesztése, amely 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. Terveinkről 2010-ben számoltunk be a VII. MSzNy konferencián [1], 2011-ben pedig az első prototípust mutattuk be [2]. A projekt 2012-ben lezárult. Eredményeinkről számol be ez az előadás.

A projekt során kifejlesztettük a szemantikai keresés technológiáját, és két rendszert: az egyiket angol nyelvű szabadalmi leírásokban, a másikat magyar nyelvű néprajzi anyagokban való keresésre. Itt a technológiát ismertetjük, a szabadalmi rendszert programbemutatón [3] ismerheti meg az érdeklődő. A magyar nyelven működő néprajzi keresőrendszert [4] ismerteti.

2 State of the art

Természetesen – mint annyi szakszó az informatikában – a „szemantikus információ-visszakereső rendszer" kifejezésben szereplő „szemantikus" is a lehető legkülönbözőbben értelmezhető. Sokan a szavak, szóösszetételek szintjén értelmezik: szavak közti jelentés-összefüggések feltárásával egészítik a ki a kulcsszó szerinti keresést. Ilyen a már elterjedt látens szemantika algoritmusa[1] (l. [5]) is. 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 [6]-ban) szintén ontológiákhoz van kapcsolva, de a névelem-felismerés (NER) technikáját is használja. A keresőkifejezésben megengedi, hogy a kulcsszavakhoz a felhasználó megadja annak besorolását, pl. *PROTEIN*:cat, amit a keresés pontosságának javítására használ. Mi azonban szemantikus keresés alatt olyan folyamatot értünk, amely összefüggő szövegrészek jelentése alapján ítél valamely dokumentumot relevánsnak.

A szemantikus keresők két nagy osztályba sorolhatóak (l. [7]): 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. Szintén gyakran használt osztályozási szempont az, hogy témafüggetlenek, vagy egy tématerületre specializáltak. Csak néhány keresőrendszert 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. [10]) egy átfogó, szintén statikus természetesnyelv-feldolgozó keretrendszer, amely egy témafüggetlen kereső motort is tartalmaz. Több, egy-egy területre vagy dokumentumhalmazra specializált alkalmazása van, pl. a Wikipediára, 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. [11]).

Az UpTake (l. [12]) 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 rendszer alapjául szolgáló ontológiát tanuló algoritmusokkal bővítik.

A GoWeb (l. [13]) az élettudományokra specializált kereső. Természetes nyelvű kifejezést fogad el bemenetként, s egy tradicionális, kulcsszó szerinti keresés eredmé-

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

nyeit veti alá szemantikus elemzésnek. Háttere a Gene és a MeSH ontológia. Az eredményhez ezeknek az ontológiáknak a releváns részleteit is megmutatja. E leírásból is kitűnik, hogy a GoWeb dinamikus kereső.

A MEDIE (l. [6], [15]) 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. [6] beszámol további kutatási irányokról, amelyek hasonlóak a mieinkhez.

3. A technológia áttekintő ismertetése

A kifejlesztett technológia szerinti szemantikus információ-visszakeresés folyamata a következő lépésekből áll:

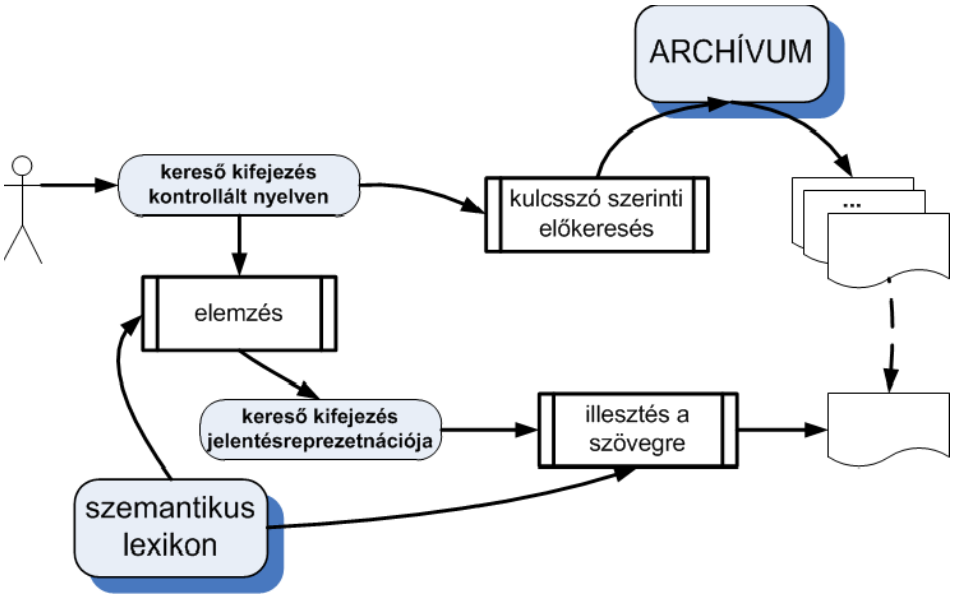
1. A felhasználó egy kontrollált nyelven megfogalmazott szöveget ad meg keresőkifejezésként.
2. Egy szintaktikai és kognitív szemantikai elemzési folyamat előállítja a keresőkifejezés jelentésreprezentációját. Az ehhez szükséges nyelvi és világtudást egy szemantikus lexikon írja le.
3. A keresőkifejezésben szereplő kifejezések és szinonimáik alapján egy kulcsszó alapú előkeresés kiválasztja a dokumentumokból azokat a szövegszegmenseket, amelyek a kulcsszavak előfordulása alapján találatok lehetnek.
4. Egy illesztési folyamat megy végbe, amely a keresőkifejezés jelentésreprezentációját ráilleszti azokra a szövegszegmensekre, amelyekben a szavak szerinti előkeresés találatai vannak, és az illesztés alapján elbírálásra kerül a keresőkifejezés és az adott szövegrészlet hasonlósága.
5. A felhasználó megkapja azon dokumentumrészletek listáját, amelyek a keresőkifejezés jelentéséhez leginkább hasonló jelentéssel bírhatnak. A lista a hasonlóság foka szerint rendezett.

A folyamat architektúráját az 1. ábra illusztrálja.

A fentiekből látható, hogy a technológia alapelve a következő:

A kontrollált nyelven megadott keresőkifejezésnek pontos jelentésreprezentációja generálható, a keresésnél a nyelvi és szemantikai elemzés bizonytalanságai heurisztikus szabályokkal oldatnak fel.

A következőkben a keresőkifejezés jelentésreprezentációjának előállítására koncentrálunk, e témában is a nyelvészeti kérdéseket hangsúlyozva.



1. ábra: A folyamat áttekintő architektúrája.

4 A jelentésreprezentáció előállítása

4.1 Szemantika, jelentés, jelentésreprezentáció

Míg a nyelvészet különböző területeinek vizsgálata a nyelven belül marad, a szemantika – legalábbis ahogy mi értjük – kilép belőle: a nyelvi konstrukciók jelentései általában nyelven kívüli entitások. A számítógépes nyelvészet területén az alkalmazás célja határozza meg, mit értünk jelentés alatt. Sőt, a jelentés fogalma háttérbe húzódik, a célnak megfelelő jelentésreprezentáció lesz fontos. Az információ-visszakeresés esetén olyan jelentésreprezentációt keresünk, amely a keresőkifejezésre és az ahhoz illő nyelvi konstrukciókra ugyanaz lesz. Kissé absztraktabban: a jelentésreprezentáció reprezentálja azt a szituációt, amelyet a nyelvi konstrukció jelenthet. A szituációk központi szerepéből adódóan a jelentésreprezentáció legfontosabb feladata a predikatív szavak és argumentumaik által alkotott szintaktikai egységek jelentésének pontos ábrázolása.

A fentiek alapján érthető, hogy a jelentésreprezentációk kialakítását davidsoni alapokon [15] kezdtük el, azaz az igék és az eseményszerűségeket jelentő főnevek jelentését reifikáljuk: maga az esemény egy token lesz, amelyhez a szereplőket szerepreláció kötik hozzá. A problémát a tematikus szerepek kiválasztása és ehhez kapcsolódóan a szemantikus lexikon szerkezetének meghatározása jelentette.

4.2 Szemantikus lexikon

A projektvezető legszívesebben a keresési feladathoz illő saját rendszert alkotott volna, tematikus megoszlásban más-más szerepkészlettel – de az erőforrás megalkotásával, feltöltésével járó munkát a projekten belül nem vállalhattuk, ezért meglévő erőforrás után néztünk. A követelmények a következők voltak:

- tartalmazzon vonzatkereteket, mégpedig szemantikus információval (megfelelő szerepekkel), nemcsak az igéknél, hanem a predikatív szavaknál általában, lehetőleg rugalmasan, ne adott számú kötött szerep legyen kiosztva;
- a szinonimahalmazok ne legyenek olyan szűkek, mint a WordNet esetében; elsősorban a következőkre gondoltunk:
  - az igékkel együtt szerepeljenek a hasonló jelentésű, eseményszerűséget jelentő főnevek (pl. a *treat* és *treatment*),
  - azok a szavak, amelyek csak a nézőpontban különböznek (pl. *give* és *get*), szintén együtt szerepeljenek;
- szerepeljenek benne szelekciós megszorítások (minél több, annál jobb);
- a szinonimahalmazokat relációk kössék össze, elsősorban az öröklődési reláció.

Az erőforrások áttekintése (l. [16]) a következő hármat találta:

- PropBank (http://verbs.colorado.edu/propbank/framesets-english/),
- VerbNet (http://verbs.colorado.edu/~mpalmer/projects/verbnet.html),
- FrameNet (http://framenet.icsi.berkeley.edu/).

A PropBank Arg1, Arg2 … jellegű szerepeket használ, amelyek csak nagyjából felelnek meg jelentéshordozó szerepeknek (Actor, Instrument stb.). A VerbNet – ahogy neve is mutatja – csak igéket tartalmaz, és kötött tematikusszerep-készletet használ. Így végül a FrameNetre esett a választásunk, amelynek egy verziója már letölthető volt.

A FrameNet, ahogy a neve is mutatja, a frame szemantika alapján épül ([17], [18]). A frame szemantika szerint a szavak jelentését egy szemantikai frame-en, avagy szemantikai kereten keresztül lehet megragadni, amely legtöbbször egy esemény, illetve szituáció leírása, a benne lévő szereplőkkel együtt. Például a főzés fogalmához mint kerethez hozzátartozik szereplőként, avagy „frame element”-ként az étel, az a személy, aki a főzést végzi, az ételt tartalmazó edény, valamint az energiaforrás is. Ez a jelentésmegközelítés tehát bizonyos szintű világtudást is magában foglal. Egy-egy frame-et több szó is felidézhet (ezeket a már frame-hez rendelt szavakat hívják “lexical unit”-oknak, azaz lexikai egységeknek), amelyek nem feltétlenül szinonimái egymásnak. Például a gyógyítás frame-et a *rehabilitáció* és az *ápoló* szavak is előhívhatják.

A FrameNet egy, a frame szemantikán alapuló lexikális adatbázis, amely a predikatív szavak leírását célozza meg. Az egyes predikatív szavakhoz megadja, hogy mely frame-eket hívják elő (az ekként jelentés-egyértelműsített szavak a “lexical unit”-ok), hogy milyen vonzatkeretekben (valence pattern, azaz valenciamintázat, amely nem csupán vonzatokat, hanem szabad határozókat is tartalmazhat) fordulhatnak elő, valamint hogy milyen relációban állnak velük az egyes vonzatkeretekben előforduló bővítményeik. Ez utóbbi relációt a “frame element” (frame-elem) fogalma fedi, amely egy adott frame-hez tartozó megfelelő szereplőt jelöl. A hagyományos tematikus szerepekkel ellentétben a frame-elemek frame-specifikusak (például a gyó-

gyítás frame-hez tartozik a Gyógyító frame element), azonban a frame-ekhez hasonlóan fennállhatnak közöttük öröklődési és egyéb kapcsolatok.

A FrameNet bővítése lexical unitokkal és valence patternekkel egy korpusz kézi frame szemantikai annotálásán keresztül, majd az annotált mondatok gépi feldolgozásával és adatbázissá alakításával történik. Ennek megfelelően, lévén korpuszalapú, időnként nyelvészetileg érthetetlen redundanciák vagy épp hiányok tapasztalhatók, és bizonyos mértékig (a gyakorisági hatást leszámítva) esetleges, hogy egy szó mely jelentése, illetve valenciamintázata kerül be az adatbázisba.

A FrameNet teljesíti legfontosabb követelményeinket, bár nem mindegyiket kielégítően:

- Az úgynevezett valence pattern-ök (VP-k, valencia mintázatok) jól leírják a régens és dependensei közti viszonyt. Nincs megkülönböztetve a vonzat és a szabad határozó. Nincsenek általános frame-elemek, első pillanatban úgy érezzük, hogy minden frame-hez frame-elemek egy külön rendszere tartozik. Azonban kimutatták [19], hogy az ugyanolyan névvel ellátott frame-elemek sok esetben azonosnak vagy közel azonosnak tekinthetőek.
- A frame szemantika szerint egy szó akkor tartozik egy frame-hez, ha a szóról az adott frame-re asszociálunk. Egy frame-hez tartozásnak nem feltétele a szófaj egyezése, így igék, főnevek, melléknevek, sőt, prepozíciók kerülhetnek egy frame-hez. Látható, hogy az egy frame-hez tartozó lexikális egységek kielégítik a keresés igényeit. Egyedül az az eltérés, hogy a nézőpontjukban különböző szavak külön frame-ekben vannak, viszont ezeket egy közös frame-hez köti a Perspective_on reláció. Mi összevonjuk ezeket a frame-eket egyetlen frame-be.
- Szemantikus megszorítások vannak a FrameNet-ben, de teljesen használhatatlanok. Az úgynevezett ontológiai szemantikus típusok meghatároznak egy kezdetleges fogalmi hierarchiát, ezek megadásával definiálják a szemantikus megszorításokat. Azonban az ontológiai szemantikus típusok semmiképp nincsenek összekapcsolva a frame-ekkel, ezért önmagukban használhatatlanok.
- A frame-eket relációk kötik össze, amelyek szerint öröklődnek a frame elemek is. Legfontosabbak az Inheritance, a Subframe és a Using relációk. Azonban a relációk értelmezése nem tiszta, használatuk nem következetes. Számunkra a leginkább fájó az volt, hogy az öröklődés nem követ valamilyen ontológiai elvet, hanem csak az egymásnak megfelelő frame elemekre figyel. Például: van négy „Cause_change of ….” kezdetű névvel ellátott frame, de egy sem öröklődik a „Cause_change” frame-ből[2].

Vannak a FrameNet-nek árnyoldalai is. Számunkra érthetetlen, hogy a VP-knél nem jelzik, hogy az ige aktív vagy passzív alakjához tartozik-e – ezt az annotált mondatokból kellett kiszűrnünk. Találtunk hibás besorolást, következetlen döntéseket.

Az irodalom (l. pl. [20]) arról tanúskodik, hogy a szóegyértelműsítés hibái különösen nagy kárt okozhatnak, mert a frame-specifikus szereprelációk miatt nagyobbat tévedhet a rendszer, ha rossz frame-hez köti az adott kifejezést, mint ha általánosabb szereprelációkat használnánk. Ezt azzal védjük ki, hogy a keresőkifejezés jelentésreprezentációjában az összes olyan frame-t felvesszük, amelyben a lexikális egységnek

2 Nem volt időnk alaposan végig vizsgálni ezeket, így nem tudjuk, hogy az öröklődés elve okozta ezt, vagy csak rosszul van alkalmazva.

van megfelelő VP-je, és a felhasználó egyértelműsíthet. A keresés során pedig a keresőkifejezés szabja meg a választott frame-et: ha egy szó tartozhat egy, a keresőkifejezés jelentésreprezentációjában szereplő frame-hez, odatartozónak vesszük.

A FrameNet másik hátránya az, hogy viszonylag kevéssé feltöltött. A 2011-es állapotban 960 frame és 11 600 lexikai egység volt, azonban ez utóbbiakból csak 6 800 teljesen annotált. Érthetően elsősorban a predikatív szavakat veszik fel, bár vannak kivételek. Ezért a nem predikatív szavakra nem megfelelő forrás, és mondanunk sem kell, hogy nem a tudományos jellegű szövegek szókincsét nyújtja.

Ezen okok miatt úgy határoztunk, hogy a szemantikus lexikon három rétegből álljon:

- a predikatív (vonzatkerettel rendelkező) szavak,
- a nem predikatív főnevek, melléknevek, határozók,
- tematikus csoportosításban az egyes szakterületek terminológiája.

A FrameNet csak az első réteg alapja lett, a másodiknak a WordNetet, a harmadiknak egy orvosbiológiai erőforrás, a MeSH[3] megfelelő szegmenseit vettük. A letöltött FrameNet-verzió csak alapja az első rétegnek, számos új lexikai egységet vettünk fel, és frame-ekkel, relációkkal, szemantikus megszorításokkal is gazdagítjuk. Bizonyos FrameNet-beli szelekciós megszorításokat helyettesíteni tudtunk WordNetre hivatkozókkal, felhasználva az ontológiai szemantikus típusoknak megfelelő WordNet szinonimahalmazokat.

4. 3 Jelentésreprezentáció

A jelentésreprezentációk ciklusmentes, címkézett irányított gráfok. A gráfban a predikatív szavaknak olyan csomópontok felelnek meg, amelyek az adott szónak megfelelő frame-mel vannak címkézve, míg a belőlük kiinduló élek a bővítményekre jellemző szintaktikus viszonynak megfelelő frame-elemmel címkézettek. Az élek a bővítmények reprezentációiba futnak. Tehát mi a frame elemeket frame-ek közt értelmezett relációknak tekintjük.

A jelentésreprezentáció előállításához természetesen morfoszintaktikai elemzés szükséges. Azonban csak olyan mértékben van szükségünk a szöveg szintaktikai szerkezetére, hogy a jelentésreprezentációt generálni lehessen. A jelentésreprezentáció sokszor igen kissé hasonlít a szintaktikus gráfhoz, ahogy a következő példán is láthatjuk. Tekintsük a következő két frázist: *a tablet containing aspirin* és *a tablet contains aspirin*. A jelentésreprezentációjuknak meg kell egyezni, mindkét esetben a *contain* ige a fej, míg az első frázisnál a *contain* a *tablet* bővítménye. Ebben az esetben a szintaktikus elemzésnek azt is be kell jelölnie, hogy a *tablet* a *contain* alanya. A szintaktikus elemzésről [21]-ben olvashatunk.

A jelentésreprezentáció előállításához a szavakhoz tartozó frame-eket és a szintaktikus viszonyokhoz tartozó frame-elemeket kell a szemantikus lexikonból kinyerni. Problémát az alternatívák nyilvántartása jelent; nemcsak az, hogy egy csomópont vagy él több címkét kaphat, hanem az is, hogy ezek összefüggenek. Egy szülő csomópont

3 Mivel a prototípus a gyógyszerek és kozmetikai szerek témaköréhez adaptált.

frame címkéjétől függ a belőle induló él frame-elem címkéje, és a gyermek csomópont címkéje függhet a hozzá vezető él címkéjétől (a szemantikus megszorítások miatt). Sőt, ugyanaz a csomópont lehet több szülő gyermeke is. A legcélszerűbb az lett volna, ha a szintaktikus elemzés és a jelentésreprezentáció generálása egyszerre történik, azonban történeti okokból nem így lett.

5. Keresés

A keresési folyamat az előkeresés által kiválasztott szövegszegmensek és a keresőkifejezés jelentésreprezentációja közti hasonlóság megállapításából áll. A feldolgozás szövegszegmensenként zajlik. Első lépésként meg kell vizsgálni, hogy a feldolgozandó szövegszegmens mely szavai tartozhatnak olyan frame-hez, illetve szinonimahalmazhoz, amely szerepel a keresőkifejezés szemantikai reprezentációjában. A második lépés annak meghatározása, hogy a keresőkifejezés elemeinek a szövegszegmensben megfelelő szavak lehetnek-e ugyanabban a szemantikai viszonyban, mint amelyben a keresőkifejezés szemantikai reprezentációjában nekik megfelelő elemek állnak. Ez a lépés három módon valósítható meg:

- Az adott szövegszegmens teljes szintaktikai elemzése alapján készül el a szegmens teljes jelentésreprezentációja, és ez a teljes reprezentáció kerül összehasonlításra a keresőkifejezés reprezentációjával.
- A szegmens teljes szintaktikai elemzése elkészül, és a keresés ezen a teljes szintaktikus gráfon működik, de csak azon szegmensbeli kifejezések szemantikai értéke kerül előállításra és vizsgáltatik meg, amelyek kapcsolatban állnak a keresőkifejezés reprezentációjában előforduló elemnek megfelelő frázissal.
- Teljes egészében a keresőkifejezés által vezérelt keresés megy végbe, tehát a szegmensnek a keresőkifejezés elemeinek megfelelő kifejezéseiből kiindulva részleges és párhuzamos szintaktikai és szemantikai elemzés történik, amely a grammatikai viszonyok megállapításával egy időben osztja ki a szemantikai szerepeket.

Az 1. megoldást valósítottuk meg. A megközelítés hátránya az, hogy nincsen megbízható módszer a nem kontrollált nyelven megfogalmazott szövegek pontos jelentésreprezentációjának előállítására. Viszont nincs is szükségünk a pontos jelentésreprezentációra, alkalmazhatjuk a „*jóindulatú olvasat*” elvét. Ez azt jelenti, hogy ha egy szó tartozhat olyan frame-be, synset-be, amely a keresőkifejezés jelentésreprezentációjában szerepel, vizsgálat nélkül elfogadja ezt az eljárás; hasonlóképpen, ha a bővítményekhez tartozó lehetséges frame-elemek közt van a keresőkifejezés jelentésreprezentációjában szereplő, az eljárás azt veszi figyelembe. Figyelni kell arra, hogy ugyanaz a csomópont a keresőkifejezés jelentésreprezentációjában előforduló több frame-hez/synset-hez is illik, - ekkor a hozzá tartozó frázis reprezentációját meg kell sokszorozni, mintha többször fordulna elő.

6. További teendők, kutatási irányok

A projekt sikeres volt: hatékony információ-visszakereső technológiát sikerült kidolgoznunk. A sikert egyrészt az biztosította, hogy csak a keresőkifejezés elemzésének kell pontosnak és egyértelműnek lennie, másrészt a frame-szemantikán alapuló jelentésreprezentáció. Természetesen a sikeresen megvalósított információ-visszakereső rendszer nem jelenti azt, hogy minden kutató-fejlesztő tevékenységet lezárhatunk. A találati halmaz pontossága megfelelőnek tűnik, viszont a fedést növelnünk kell. Vannak olyan fejlesztési feladatok, amelyek elméletileg tisztázottak, specifikálva is vannak (pl. a raising és control igék kezelése), ezekre itt nem térünk ki.

Van azonban két nyelvészetileg meg nem oldott probléma, amellyel szembetalálkoztunk:

- A felsorolások és koordinációk jellemzőek az igénypontok szövegére. A keresőkifejezés szerkesztésénél a felhasználónak meghatározott módon jeleznie kell a felsorolásokat, koordinációkat, a dokumentumok elemzésénél közelítő eljárást alkalmazunk.
- Az összetett szavak problémája sokkal fájóbb. Bonyolult kémiai kifejezéseket találtunk, sokat mi sem tudtunk értelmezni. A probléma a kifejezések zárójelezése. Az irodalomban sem találtunk nagy mennyiségű kézi annotálás nélkül implementálható megbízható megoldást problémánkra. [22] szerint a balra való zárójelezés a legjobb default, alapértelmezett eljárás – bár mi számos ellenpéldát láttunk.

A kognitív nyelvészethez tartozó frame szemantika alkalmazása nem volt céltudatos, hanem a FrameNet választása szinte észrevétlenül sodort minket bele. Természetesen a kognitív nyelvészet nagyon sok sajátossága a feladatunkhoz mellékes volt. Azonban most látjuk, hogy a frame szemantika beágyazási mechanizmusa alkalmas természetesnyelv-feldolgozási feladatok megoldáshoz. Az egyes szavakhoz tartozó frame-ek beágyazása elvezethet közös doménhez, így alkalmas egy-egy szövegszegmens témájának meghatározásához. Ez segítené a szóegyértelműsítést is.

Köszönetnyilvánítás

A kutatás az NFÜ által finanszírozott, MASZEKER kódnevű, TECH_08_A2/2-2008-0092 számú projekt keretében valósult meg.

Szeretnénk köszönetet mondani a projekt összes résztvevőjének is – nélkülük nem számolhatnánk be eredményeinkről.

Hivatkozások

1. Szőts M., Csirik J., Gergely T., Karvalics L.: MASZEKER: projekt szemantikus keresőtechnológia kidolgozására. In: Tanács A., Vincze V. (szerk.): VII. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2010) 159–167

2. Hussami P.: MASZEKER: szemantikus kereső program. In: Tanács A., Vincze V. (szerk.): MSzNy 2011 – VIII Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2011) 321–322
3. Hussami P.: MASZEKER: szemantikus kereső program. In: Tanács A., Vincze V. (szerk.): IX. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2013) 302–304
4. Zsibrita J., Vincze V.: Magyar nyelvű néprajzi keresőrendszer. In: Tanács A., Vincze V. (szerk.): IX. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2013) 361–367
5. Landauer, T.K., McNamara, D.S., Dennis, S., Kintsch, W. (eds.): Handbook of Latent Semantic Analysis (Universita of Colorado Institute of Cognitive Science Series). Psychology Press (2007)
6. Ananiadou, S., Thompson, P., Nawaz, R.: Improving Search through Event-based Biomedicasl Text Mining. In: Darányi, S., Lendvai, P. (eds.): Proceedings of the First International AMICUS Workshop on Automated Motif Discovery in Cultural Heritage and Scientific Comminication Texts (2010) 42–54
7. Abolhassani, H., K. S. Esmaili: A categorization scheme for semantic web search engines. In: 4th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA-06) (2006)
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. 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.
11. 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 (Letöltve: 2009. március 27.).
12. UpTake under the hood: the Interview. Alt-SearchEngines, 2008. május 14. WWW document. http://www.altsearchengines.com
13. Dietze, H., Schroeder. M.: GoWeb: A semantic search engine for the life science web. In: Burger, A., Paschke, A., Romano, A., Splendiani, A. (eds.): Proceedings of the Intl. Workshop Semantic Web Applications and Tools for the Life Sciences SWAT4LS. Edinburgh (2008)
14. 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, Vol. 2 (2006) 1017–1024
15. Parsons, T.: Events in the Semantics of English: A Study in Subatomic Semantics, MIT Press, Cambridge, MA (1990)
16. A szemantikus nyelvészeti erőforrások áttekintése. http://www.maszeker.hu/?page=download
17. Fillmore, C. J.: Frame Semantics And The Nature Of Language. In: Annals of the New York Academy of Sciences, Vol. 280, No.1 (1976) 20–32
18. Baker, C. F., Fillmore, C. J., Lowe, J. B.: The Berkeley FrameNet Project. In: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (ACL '98), Vol. 1. Association for Computational Linguistics, Stroudsburg, PA, USA (1998) 86–90
19. Matsubayashi, Y., Okazaki, N., Tsujii, J.: A comparative study on generalization of semantic roles in FrameNet. In: Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP (2009) 19–27

20. Shen, D., Lapata, M.: Using semantic roles to improve question answering. In: Proceedings of EMNLP-CoNLL (2007) 12–21
21. Kiss, M., Nagy, Á., Vincze, V., Almási, A., Alexin, Z., Csirik, J.: A Manually Annotated Corpus of Pharmaceutical Patents. In: Proceedings of TSD 2012 (2012) 135–142
22. Nakov, P., Hearst, M.: Search Engine Statistics Beyond the n-gram: Application to Noun Compound Bracketing. In: CoNLL-2005. Ann Arbor, MI (2005)