MISKOLCI EGYETEM

GÉPÉSZMÉRNÖKI ÉS INFORMATIKAI KAR

TÉMA- ÉS NYELVADAPTÁLHATÓ TERMÉSZETES NYELVI VEZÉRLŐ KERETRENDSZER

Ph.D. értekezés tézisei

KÉSZÍTETTE:
Barabás Péter
okleveles mérnök-informatikus

AKI DOKTORI FOKOZAT ELNYERÉSÉRE PÁLYÁZIK

HATVANY JÓZSEF INFORMATIKAI TUDOMÁNYOK DOKTORI ISKOLA
ALKALMAZOTT SZÁMÍTÁSTUDOMÁNY TÉMATERÜLET
ADAT- ÉS TUDÁSBÁZISOK, TUDÁSINTENZÍV RENDSZEREK TÉMACSOPORT

TÉMAVEZETŐ:
Dr. habil. Kovács László

TÁRSTÉMAVEZETŐ:
Prof. Dr. Juhász Imre

Miskolc, 2013.

Barabás Péter

Téma- és nyelvadaptálható természetes nyelvi vezérlő keretrendszer

Ph.D. értkezés tézisei

Miskolc, 2013.

VÉDÉSI BIZOTTSÁG

Elnök:

Prof. Dr. Szigeti Jenő, CSc	ME, egyetemi tanár

Titkár:

Dr. habil. Kovács Szilveszter, Ph.D.	ME, egyetemi docens

Tagok:

Dr. Johanyák Zsolt Csaba, Ph.D.	GAMF Kar, főiskolai tanár
Dr. Czap László, Ph.D.	ME, egyetemi docens
Dr. Kusper Gábor, Ph.D.	EKF, főiskolai docens

Opponensek:

Dr. Stanislav Ondáš, Ph.D.	TUKE, egyetemi adjunktus
Dr. Dudás László, CSc	ME, egyetemi docens

TARTALOMJEGYZÉK

1. BEVEZETÉS

A számítógépek megjelenése magával hozta bizonyos mesterséges nyelvek kialakulását, amelyek segítségével kommunikálni tudunk a gépi rendszerekkel. Ezen nyelvek tipikus példái a programozási nyelvek. Az emberek a számítógépek megjelenése óta szeretnék megvalósítani, hogy a gépekkel az egymás között megszokott, természetes nyelven „beszélgessenek". Ezen interakciónak a gyakorlati megvalósulását a természetes nyelvi felületek jelentik. A természetes nyelvi interfésszel (Natural Language Interface, NLI) rendelkező információs rendszerek gyökerei 1970-re nyúlnak vissza. Az úttörő LUNAR (Woods & Kaplan, 1977) projekt a holdkőzetek adatbázisában való lekérdezésekhez dolgozott ki természetes nyelvű interfész felületet. A RENDEZVOUS (Codd, 1974) rendszer volt az első általános célú adatbázis NLI modul. Az NLI modulok egyik alapfeladata a természetes nyelven beérkező parancsok átkonvertálása a feldolgozó modul saját parancsnyelvére. Ezen konverzió megvalósítása több lépcsőben történik, kezdve a természetes nyelvi mondat szintaktikai ellenőrzésével, elemzésével és folytatva a szemantikai analízissel, a tématerület ismert fogalmainak detektálásával.

Több ismert NLP keretrendszer is található a piacon, ezek közül a legismertebbek az Apache nyílt forráskódú OpenNLP (openNLP, 2010) és UIMA (Apache, 2013) rendszerei, illetve a Stanford Egyetem statisztikai alapú Stanford NLP (StanfordNLP, 2013) rendszere. Az előzőekben felsorolt rendszerekben egyaránt megtalálhatóak a legfontosabb szövegfeldolgozási modulok, mint a mondat- és szódetektáló, tulajdonnév azonosító, szótövező, mondatosztályozó, stb. A legtöbb létező keretrendszer alapvetően az angol nyelvet, illetve még esetleg néhány „könnyebben" feldolgozható nyelvet támogat. Esetenként előfordul, mint az UIMA esetén is, hogy a keretrendszer adaptálható különböző nyelvekre.

A magyar nyelv egy nehezen feldolgozható nyelv, az agglutináló tulajdonsága révén szavak és toldalékok nagyszámú kombinációjával rendelkezik, melyek között sok szabályszerűség és kivételes ragozás fedezhető fel. Magyar nyelvű kutatásokban (Zsibrita, Vincze, & Farkas, 2013), (Alexin, Csirik, Kocsor, Miháltz, & Szarvas, 2006) is találkozhatunk az előzőkben említett rendszerek használatával.

A kutatásom célja, hogy fentiekben említett NLP keretrendszerek egyes funkcióit magyar nyelvű szöveg feldolgozására implementáljam,

kiterjesszem a parancsnyelvi funkciókra történő konvertálással és azok implementációjának dinamikus meghívásával.

1.1. Irodalmi áttekintés

A természetes nyelvfeldolgozás kezdete az 50-es évek elejére tehetők, amikor is Alan Turing a „Computing Machinery and Intelligence" című tanulmányában (Turing, 1950) publikálta az intelligencia ismérveit. Ma Turing-tesztnek hívják a róla elnevezett eljárást, ahol egy emberi szakértő valósidejű írásbeli kommunikációt folytat egy emberrel és egy géppel. A számítógépprogram akkor mondható, hogy teljesíti a Turing-tesztet, ha a szakértő nem tudja eldönteni, hogy melyikük az ember és melyikük a gép.

A gépi fordítás volt az első természetes nyelvfeldolgozással kapcsolatos alkalmazás. Néhány kutatásról lehet olvasni az 50-es évek előttről (Hutchins J. , 2005). Ezek a rendszerek szótáralapú kereséseket használtak a szókereséshez, fordításhoz vagy a szóátrendezéshez és általában gyenge eredményt produkáltak.

A „Georgetown-kísérlet" szerzői 1954 azt állították, miután implementáltak egy gépi fordítót, amely több mint 60 orosz mondatot automatikusan lefordított angolra, hogy a gépi fordítás problémáját fél évtizeden belül megoldják (Hutchins W. J., 2004).

1957-ben Chomsky „Syntactic Structures" (Chomsky, 1957) című munkájában a generatív nyelvtan ötletét publikálta. Ennek hatására jelentősen megerősödött a gépi fordítás területe. Időközben egyéb területek, mint a beszédfelismerés is kezdtek előtűnni.

A kutatók az 1960-as években egyre optimistábban tekintettek a természetes nyelvfeldolgozásra, miután az SHRDLU (Winograd, 1972) nevű rendszer (olyan természetes nyelvű rendszer, amely az építőkockák világában dolgozott) meglehetősen jó eredményeket ért el a maga korlátozott szókincsével. A következő NLP alkalmazás az ELIZA (Weizenbaum, 1966) nevet viselte, melyek Joseph Weizenbaum fejlesztett 1964 és 1966 között. ELIZA egy pszichológust szimulált, amelynek a gondolatokról és érzésekről szinte nulla információja volt, mégis érdekes ember-gép interakcióra volt képes.

Azon állítás, hogy a gépi fordítás 3-5 éven belül megoldódik, elbukott és az ALPAC beszámoló (J. R. Pierce, 1966) 1966-ban azt eredményezte, hogy a kutatások támogatását drasztikusan lecsökkentették, mivel az eredmények 10 évnyi kutató munka után is az elvárásokon alul teljesítettek.

A 70-es években „fogalmi ontológiákat” kezdett írni több programozó is. Az ontológiák a valós információkat képesek a számítógép számára „érthető” adatok formájába átalakítani, ahogy tették ezt a következő rendszerekben: MARGIE (Shank, 1975) SAM (Cullingford, 1981), PAM (Wilensky, 1978), TaleSpin (Meehan, 1976), QUALM (Lehnert, 1977), Politics (Carbonell, 1979), Plot Units (Lehnert 1981). Ebben az időben több chatrobot is napvilágot látott, mint a PARRY, Racter vagy a Jabberwacky. A már említett LUNAR rendszert is ekkoriban hozták létre. A 70-es évek vége felé a szemantikai irányok, a kommunikáció feldolgozásának céljai, tervei és a beszélgetés jelensége került előtérbe.

A 80-as évek végétől kezdve a növekvő számítási teljesítmény és az alacsonyabb költséges a statisztikai módszerek terjedését eredményezték.

1.1.1 NLP keretrendszerek és feladataik

Számos feladat van a természetes nyelvfeldolgozás területén, melyeket létező keretrendszerek már tartalmaznak. Némely megvalósítás valós alkalmazásként került a piacra, némelyek pedig mindössze nagyobb egységek építőelemeiként funkcionálnak. A legfontosabb természetes nyelvfeldolgozó feladatok a következők: gépi fordítás, morfológiai szegmentálás, tulajdonnév-azonosítás, természetes nyelv-generálás, szófajfeltárás, mondatelemzés, kérdésgenerálás, szövegszegmentálás, beszédfelismerés, szóegyértelműsítés, témadetektálás, nyelvdetektálás, stb.

A fent említett feladatok nagy részét a népszerű keretrendszerek, mint az NLTK (NLTK, 2012), Apache UIMA (Apache, 2013) vagy Stanford NLP (StanfordNLP, 2013) megvalósították.

Az NLTK Python alkalmazások fejlesztése során használható természetes nyelvű adatok feldolgozására. Több könnyen használható interfészt tartalmaz több, mint 50 korpusz és lexikális forráshoz, mint pl. a WordNet-hez. Az NLTK megvalósítja az osztályozás, tokenizálás, szótövezés, szófaj meghatározás, mondatelemzési funkciókat. Elsősorban az angol nyelvet támogatja, de megvan a lehetőség más nyelvekre történő adaptálására is.

A Stanford NLP keretrendszer a Stanford Natural Language Processing Group fejlesztette évekkel ezelőtt. Java-alapú statisztikai feldolgozó modulokat tartalmaz a főbb számítógépes nyelvészeti problémákra. A Stanford CoreNLP a keretrendszer egy több modult integráló része, amely angol nyelvre megvalósítja a tokenizálás, szófaj meghatározás, tulajdonnév-felismerés és a társhivatkozás meghatározás funkcióit. A CoreNLP mellett számos egyedi modullal is rendelkezik a Stanford NLP keretrendszer.

Az Apache UIMA a három ismertetett keretrendszer közül a legrobosztusabb, a komponensei mind Java, mind C++ nyelven elérhetőek. Az UIMA-ban annotáló komponenseket kell beállítanunk és egymás után fűznünk ahhoz, hogy a strukturálatlan információt fel tudjuk dolgozni. Saját annotálókat is írhatunk a keretrendszerhez vagy meglévőeket is terjeszthetünk ki egyaránt. Az UIMA feldolgozó motorja a következő funkciókat biztosítja: nyelv-beazonosítás, tokenizálás, szófajannotálás, felszínes mondatelemzés és tulajdonnév felismerés.

1.1.2 Kapcsolódó munkák

A magyar természetes nyelvfeldolgozás egyik legismertebb eredménye az iSpell szótár és a HunSpell (Németh, 2011) helyesírás és morfológiai elemző, amelyeket az OpenOffice programcsomag, a Thunderbird levelezőkliens, a Google Chrome, a Mozilla Firefox, az Internet Explorer, az Opera böngészők és még számos egyéb alkalmazás is sikeresen alkalmaz a pontos helyesírás elemzésre.

A Magyar Természetes Nyelvfeldolgozó Csoport (MTNC) szintén számos kutatási eredménnyel büszkélkedhet a természetes nyelvfeldolgozás területén. Fejlesztettek egy kiterjesztést a MALLET (McCallum, 2002) feltételes véletlen mező alapú tulajdonnév-felismerőhöz, amely egy paraméterezhető kivonatolót is tartalmaz az UIMA keretrendszerhez (Apache, 2013).

A magyarlanc (Zsibrita, Vincze, & Farkas, 2013) olyan magyar nyelvfeldolgozó eszközrendszer, mely a mondatszegmentálásra, tokenizálásra a MorphAdorner (MorphAdorner, 2009) rendszer adaptációját használja, a szófaj-meghatározója és szótövezője pedig a StanfordNLP (StanfordNLP, 2013) egy módosított változata. A stopszavak szűrésére és a mondatelemzésre a Bohnet elemző (Bohnet & Niver, 2012) magyar nyelvű adaptációját használja.

Az MTNC legjelentősebb eredményei a Szeged TreeBank (Csendes, Csirik, Gyimóthy, & Kocsor, 2005) és a Magyar Ontológia Tár (Magyar WordNet) (Csendes, Csirik, Gyimóthy, & Kocsor, 2005). A Szeged TreeBank olyan szótár, amely mondatok szintaktikai elemzéseit és annotációit tartalmazza. A Magyar Onológia Tár pedig egy természetes nyelvű fogalomgyűjtemény WordNet alapokon.

1.2. A kutatás célja

A kutatásom fő célja egy természetes nyelvi vezérlő keretrendszer-modell megalkotása, amely főként szabályalapú megközelítéseket alkalmaz

szemben a statisztikai módszerekkel. A célom olyan rendszer elkészítése, amely egyszerűen adaptálható különböző tématerülethez tartozó alkalmazások természetes nyelven történő vezérléséhez, valamint különböző nyelveken történő ember-gép kommunikáció megvalósításához. A keretrendszer megvalósításakor a már meglévő természetes nyelvfeldolgozó rendszerek moduljainak újrahasználhatóságát is biztosítani kell a tervezés során. A keretrendszer a következő követelményeket kell, hogy teljesítse:

1. **Téma-adaptálhatóság:** azon képesség, miszerint a keretrendszer a különböző tématerülethez tartozó fogalmakat, azok kapcsolatait a rendszer belső struktúrájának, folyamatainak változatlanul hagyása mellett megtanulja,
2. **Nyelv-adaptálhatóság:** azon képesség, miszerint a keretrendszer a különféle nyelveken írt bemeneti mondatokat a rendszer kizárólag nyelvfüggő moduljainak testre szabásával feldolgozza,
3. **Kiterjeszthetőség:** azon képesség, miszerint a hívandó alkalmazás-funkciók halmaza egyszerűen, a keretrendszer struktúrájának változatlanul hagyásával kibővíthető,
4. **Nyílt interfész:** azon képesség, miszerint a már meglévő természetes nyelvi rendszerek szövegfeldolgozó moduljainak újrahasznosíthatósága, beépítése a rendszer struktúrájának változatlanul hagyásával, jól definiált interfészhalmaz alkalmazásával megvalósítható.

A fenti követelmények teljesítése céljából a keretrendszer olyan szemantikai és alkalmazás-leíró modellt kell tartalmazzon, amely a következő követelményeket teljesíti:

- A szemantikai modell fő építőelemei a fogalmak és kapcsolataik.
- A központi fogalom a predikátum, azaz a mondat állítmánya.
- Előzetes tudással kell rendelkezni a mondatelemzés elvégzéséhez, amelyet a szemantikai modellel kell tudnunk reprezentálni.
- A szemantikai és alkalmazás-leíró modelleket egyaránt ki kell bővíteni a mondatbeli szerep információival a funkciótársítás megvalósíthatósága érdekében.
- A modellek nagyfokú rugalmasságot és kiterjeszthetőséget kell, hogy biztosítsanak.

Számos természetes nyelvfeldolgozó motor létezik, amelyek az alap szövegfeldolgozó feladatokra, mint a szegmentálás, helyesírás-elemzés, tulajdonnév-felismerés, morfológiai elemzés, mondatelemzés, jól működő megoldásokat biztosítanak. Az alkalmazásfejlesztő hatásköre, hogy ezen

módszereket a keretrendszer implementálása során felhasználja a meghatározott interfészek implementálása során.

A vizsgált rendszerek egyike sem teljesíti maradéktalanul a fenti követelményeket, ezért a disszertáció első feladata meghatározni a természetes nyelvi vezérlő keretrendszer-modell struktúráját a téma- és nyelvadaptálhatóság, illetve a meglévő feldolgozó modulok integrálhatóságának követelményeit figyelembe véve. Második feladata egyrészt egy szemantikai modell megalkotása, amely a tématerület „tudását” hivatott reprezentálni a mondatelemzés elvégzéséhez, másrészt a funkció-leíró modell definiálása a funkciótársításhoz, valamint az előbbi folyamatokat megvalósító algoritmusok létrehozása. Harmadik feladata a keretrendszerbeli optimálási pontok feltárása és a költségoptimalizálás, amelyek a természetes nyelvi vezérlés költségét minimalizálják. Végezetül pedig két mintaalkalmazást kell implementálni, amelyeken bemutatható az eredmények gyakorlati alkalmazhatósága.

2. ÚJ TUDOMÁNYOS EREDMÉNYEK

2.1. A természetes nyelvi vezérlő keretrendszer-modell architektúrájának kidolgozása

Kidolgoztam egy olyan általános természetes nyelvi vezérlő keretrendszer modellt [6][8], amely képes alkalmazások funkcióit természetes nyelvű parancsszöveggel elérni, emberközelibbé téve ezáltal az ember-gép kommunikációt. A kidolgozott modell kielégíti a következő követelményeket:

- téma-adaptálhatóság,
- nyelv-adaptálhatóság,
- kiterjeszthetőség,
- nyílt interfész,
- modularitás.

A keretrendszer modelljét 4+1 modulra osztottam, amelyek mind egy-egy réteg szerepét töltik be: az *i.* modul bemenetként az (*i-1).* modul kimenetét kapja meg és dolgozza fel. Ezek alapján a következő modulokat definiáltam:

- 0. modul: bemeneti forrás modul,
- 1. modul: szövegfeldolgozó modul,
- 2. modul: morfológiai modul,
- 3. modul: szemantikai feldolgozó modul,
- 4. modul: funkciótársítási modul.

A keretrendszer megvalósítása az 1. modullal kezdődik annak ellenére, hogy a 0. modul is definiálásra került. Ez utóbbinak mindössze annyi a szerepe, hogy információt szolgáltat az analóg források (hang, kézírás, billentyűzet, …) típusáról, amely a szövegfeldolgozó algoritmusok hangolásánál használható fel.

Definiáltam az egyes modulok feladatait, szerepkörét, kapcsolódásait figyelembe véve, hogy létező természetes nyelvi keretrendszerek moduljai, kisebb-nagyobb átalakítással, adaptálással beépíthetőek legyenek a kidolgozott keretrendszerbe. Nagy hangsúlyt fektettem a téma- és nyelv-adaptálhatóság megvalósítására. A keretrendszer moduljait két csoportra osztottam: nyelvfüggő (1-2) és nyelvfüggetlen (3-4) modulokra és definiáltam egy módszert, illetve egy kódrendszert, amellyel ezen modulok kapcsolódása egyszerűen áthidalható.

Sikerült belátni, hogy a keretrendszer egyszerűen adaptálható egyrészt különböző tématerületekre a nyelvfüggetlen, szemantikai modulok tudásbázisának kiterjesztésével, cseréjével, másrészt a különböző nyelvekre a nyelvfüggő modulok testre szabásával és a kódrendszer új nyelvre történő bővítésével.

1. tézis: *[6][8]*
Megalkottam egy négy modulból álló természetes nyelvi vezérlő keretrendszer-modellt és annak formális információáramlási struktúráját, amely teljesíti a téma- és nyelv-adaptálhatóság, kiterjeszthetőség és nyílt interfész követelményeket, biztosítva ezzel a szoftverfejlesztésben a magas fokú újrahasznosíthatóságot az ember-gép kommunikáció területén.

2.2. Szemantikai modellek és algoritmusok kidolgozása

A keretrendszer modelljének harmadik és negyedik modulja végzi a szemantikai feldolgozást és funkciótársítást. A második feladat tehát az előbbi feladatokat kiszolgáló szemantikai és funkció-leíró modellek, valamint a folyamatokat megvalósító algoritmusok kidolgozása [2][3][6]. Egy adott tárgyterület vonatkozásában az ontológia a mesterséges intelligencián belül a jelenleg elfogadott meghatározás szerint a fogalomalkotás explicit specifikációja: egy tárgyterület fogalmainak és az azok között fennálló kapcsolatok formális specifikációja, amelyhez általában természetes nyelvű leírás is társul. (Gruber, 1995). A mondatelemzés alapfeltevése szerint nem vagy csak ritkán végezhető el tisztán szintaktikai elemzés útján, szükség van valamiféle előzetes tudásra is a szöveg témakörében, amely segítségével beazonosíthatjuk a fogalmakat és azok egymáshoz való kapcsolódásai. Az ECG modellt (Varga, 2011) kiterjesztettem három új csomóponttípussal és három új kapcsolattípussal. A szemantikai modellben a következő fogalomtípusokat határoztam meg:

- **predikátum fogalom:** olyan speciális fogalom, amely a mondatban az állítmány szerepét tölti be,
- **absztrakt fogalom:** olyan összefoglaló fogalom, amelynek további specializáltjai lehetnek,
- **egyedi fogalom:** olyan fogalom, amely egy absztrakt fogalom valamely konkrét előfordulását írja le.

A kidolgozott szemantikai modellben a kapcsolatok feladatait is kiterjesztettem: egyrészt hordozzák azon információt, hogy mely fogalmak kapcsolódnak egymáshoz, másrészt morfológiai és mondatbeli szerepük

információkat is tartalmazhatják, amelyek a mondatelemzést segítik. Ezek alapján a következő kapcsolattípusokat definiáltam:

- **ISA kapcsolat:** szülő-gyermek kapcsolat két absztrakt vagy absztrakt és egyedi fogalom között,
- **Predikátum kapcsolat:** predikátum-csomópont és kapcsolódó absztrakt vagy egyedi csomópont közötti kapcsolatot reprezentál. A kapcsolatnál megadhatók morfológiai és mondatbeli szerep információ is,
- **Attribútum kapcsolat:** absztrakt és/vagy egyedi fogalmak közötti kapcsolat és a jelző és jelzett viszonyt fejezi ki. Ennél a kapcsolatnál is adhatunk meg morfológiai és mondatbeli szerepinformációkat.

Ennek értelmében a következő szemantikai modellt definiáltam:

$$\mathcal{M} = (C, R), \tag{2.1}$$

ahol $C = C_P \cup C_A \cup C_I$ a fogalmak véges halmaza és $R \subseteq C \times C$, $R = R_{pred} \cup \mathrm{R_{isa}} \cup \mathrm{R_{attr}}$ a kapcsolatok halmaza.

A mondatelemzés célja a morfológiailag elemzett mondatok mondatelemzési fává alakítása, amelyet formálisan a következőképpen adtam meg:

$$f_{analysis} : W \rightarrow S_{tree}, \tag{2.2}$$

$$S_{tree} = (N, \rightarrow_s), \tag{2.3}$$

$$N = (C', \kappa), C' \in C, \kappa \in \mathrm{K}, \tag{2.4}$$

$$\rightarrow_s \subseteq N \times N, \tag{2.5}$$

ahol

- S_{tree} : a mondatelemzési fa,
- N_s : a fa csomópontjai,
- $\rightarrow_s$: a csomópontok közötti élek,
- C' : egy fogalom,
- κ : egy mondatbeli szerep.

A természetes nyelvű vezérléssel történő kiterjesztéshez az alkalmazás funkciót le kell tudnunk írni olyan formában, hogy a mondatelemzési fából az erre a leírásra történő transzformáció megvalósítható legyen. Egy funkciót alapesetben a neve és paraméterlistája határoz meg. Ez azonban még kevés ahhoz, hogy tudjuk, hogy a mondatelemzés során feltárt fogalmak mely paraméterekhez kapcsolhatók. Definiáltam egy alkalmazás-

leíró modellt, amelybe a következő kiterjesztések kerültek az alapleíráshoz képest:

- minden funkcióhoz tartozik egy predikátumhalmaz,
- minden paraméterhez tartozik egy fogalomtípus,
- minden paraméterhez tartozik egy vagy több mondatbeli szerep,
- egy paraméterhez tartozhatnak olyan függő fogalmak, mondatbeli szerepek, amelyek a mondatelemzési fában az aktuális fogalom szülőjeként szerepelnek. Ezzel pontosítani tudjuk a társítást a nem egyértelmű helyzetekben.

A modellek definiálása után kidogoztam két algoritmust: az egyik elvégzi a mondatelemzést, a másik elvégzi a funkciótársítást a mondatelemzési fa és a funkcióleíró között.

2. tézis: *[2][3][6][8]*
Kidolgoztam egy új szemantikai reprezentációs modellt, amely kiterjeszti az ECG modellt (Varga, 2011) *három új csomóponttípussal és három új, morfológiai és mondatbeli szerepinformációkkal ellátott kapcsolattípussal, biztosítva ezáltal a hatékonyabb tudástranszformációt a bemenő mondatok és az alkalmazás funkcióhívás között. A kifejlesztett funkció leíró modell újdonságát a mondatbeli szerepek bevonása és a funkciótársításban történő alkalmazása adja, amelyet egy olyan algoritmus végez, ami a mondatelemzési fát funkcióhívássá alakítja.*

2.3. A természetes nyelvi vezérlő keretrendszer optimalizálása

A keretrendszer alkalmazhatóságának vizsgálata kulcsfontosságú kérdése volt a kutatásomnak [2][3][9][15]. Ipari környezetben akkor lehet egy efféle rendszert alkalmazni, ha a kiterjesztendő alkalmazás vezérlési ideje nem változik az eddigiekhez képest, azaz az új módszer szinte valósidejű működést tud biztosítani. Egy gombra történő kattintással pontosan meg tudtuk mondani a rendszernek, hogy mit csináljon, jelen esetben a keretrendszer feladata kitalálni a cselekvést, pl. a gombra kattintás akcióját. Ebből kifolyólag egy másik fontos szempont volt a működésbeli pontosság biztosítása. Ezek optimalizálási feladatok, mely során megalkottam a rendszer többtényezős

$$c = \frac{(w_t \cdot c_t + w_r \cdot c_r)}{c_a} \rightarrow min., \tag{2.6}$$

költségfüggvényét ahol

- c : teljes költség [-],
- c_t : futási időköltség [ms],
- w_t : futási időköltség súlytényezője [1/ms],
- c_r : erőforrás költség (memória/háttértár) [B],
- w_r : erőforrás költség súlytényezője [1/B],
- c_a : működési pontosság [-].

Ahhoz, hogy a fenti költséget minimalizálni tudjam, meghatároztam azokat a folyamatokat a természetes nyelvi feldolgozásban, amelyeket optimalizálás alá kell vetni a fenti paraméterek szempontjából. Ezek a következők:

- szófajmeghatározás,
- mondatelemzés,
- funkciótársítás.

A szófajmeghatározás és mondatelemzés során összehasonlítottam a Markov, a feltételes valószínűségi mező (LCRF) és a gráfalapú megoldásokat. A nyelv-adaptálhatóság mellett az is fontos szempont volt a kutatásomban, hogy magyar nyelvre is hatékonyan működhessen a rendszer. A magyar nyelv feldolgozását az agglutináló tulajdonsága mellett a szabad szórendje is nehezíti. Arra a következtetésre jutottam, hogy effajta nyelvi sajátosságok miatt olyan megoldást kell alkalmaznom, amelyeket a specialitások nem, vagy csak kis mértékben befolyásolják. Ezeket figyelembe véve megállapítottam, hogy a Markov és LCRF megoldások kötött szórendű nyelvek esetén használhatóak, a magyarhoz hasonló nyelvek esetén a preferált megoldás egy többszintű

$$G = \{l, K, G_N, G_D, G_P\} \tag{2.7}$$

gráfmodell, ahol l a gráf azonosítója, K a gyökércsomópont, G_N a csomópontok halmaza, G_D a gyökércsomópontból kiinduló élek halmaza és G_P a megelőzési relációk halmaza.

A mondatelemezés és funkciótársítás során az előzetes tudást reprezentáló fastruktúrát illesztettem a morfológiailag elemzett fastruktúrában ábrázolható mondathoz, illetve a mondatelemzési fát a funkcióleíró fához. Összevetettem több fa struktúra (Kovács, Adatbázisok tervezésének és kezelésének módszertana, 2004), (Guttman, 1984) és (Kovács, Rule approximation in metric spaces, 2010) és illesztő algoritmust is, mely eredményeképpen a prefix-fa alkalmazása mellett döntöttem.

A funkciótársítás hatékony megoldására egy pontrendszeren alapuló megoldást dolgoztam ki, mely során minden funkcióleírás pontot kap, ha:

- a mondat predikátuma releváns a funkcióra nézve *(x)*,
- kötelező paraméterhez rendelhető a mondatelemzés valamely csomópontja,

- opcionális paraméterhez rendelhető a mondatelemzés valamely csomópontja.

Ezek alapján definiáltam a

$$\Phi_i = \alpha x_i + \beta y_i + \delta z_i \tag{2.8}$$

súlyozott jóság függvényt, amelyet ha az F funkcióhalmaz összes elemére alkalmazunk, akkor a maximális értéket eredményező funkcióleírót kapjuk győztesként.

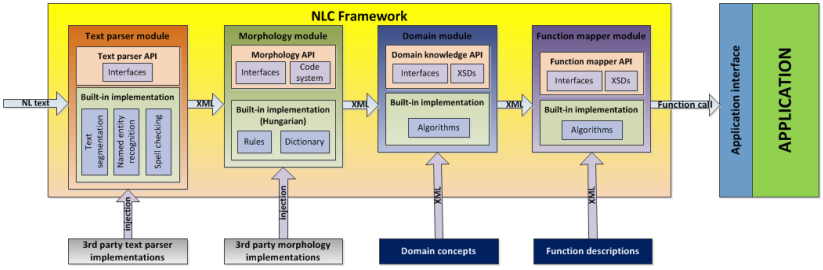
$$w = \arg\max_{i} \phi_i \,, i = 1 \ldots |F|. \tag{2.9}$$

3. tézis: *[2][3][9][15]*

Meghatároztam a természetes nyelvi vezérlő keretrendszer-modell költség szempontjából kulcsfontosságú komponenseit és felállítottam egy többtényezős költségmodellt, amely a futási időt, a pontosságot és az erőforrás-felhasználást veszi figyelembe. Elvégeztem az optimalizálást a szófaj-meghatározó, a mondatelemző és funkció-társító modulokon, amely eredményeként a javasolt gráfalapú algoritmusok a keretrendszer hatékony működését biztosítják nagyméretű tudáshalmaz esetén is.

2.4. Mintaalkalmazások kidolgozása, az eredmények gyakorlati igazolása

Az elméleti eredmények alkalmazhatóságának bemutatására elkészült a Java nyelven implementált vezérlő keretrendszer *[8]* és két mintaalkalmazás: robotvezérlésre és navigációra. A vezérlő keretrendszer működési modellje a következő ábrán látható:



2.1. ábra: *A természetes nyelvi vezérlő keretrendszer működési modellje*

A robotvezérlő alkalmazás *[7]* lehetővé teszi egy humanoid robot vezérlését magyar nyelvű természetes parancsszöveggel. Az alkalmazás kiterjeszthető beszédvezérléssel is a megfelelő beszédfelismerő modul illesztésével. A robot egyszerű utasításokat tud értelmezni és végrehajtani, mint pl. „állj

fel", „menj előre 30 centit", „nézz balra", stb. Jelenleg 19 különböző funkciót lehet természetes nyelven elérni. A funkciók köre egyszerűen kibővíthető, mindössze a funkcióleírásokat és az esetleges új fogalmakat kell „megtanítani" a rendszernek.

A navigációs alkalmazás jó példa arra, hogy pl. egy GPS eszközt miként lehetne természetes nyelvű utasításokkal használni. Az alkalmazás a Google Maps szolgáltatásait használja az eredmények előállítása céljából. A program a következő funkciókat tartalmazza: távolság és útvonal kiszámítása két város vagy cím között, eredmények megmutatása útvonaltervként vagy egyszerű válaszként a térképen, a nevezetes helyekre való rákeresés a kapcsolódó cselekvések alapján, megtalált nevezetes helyek adatainak elérése (weboldalának mutatása, telefonon történő felhívása, stb.).

Az alkalmazások igazolják a keretrendszerrel szemben támasztott követelmények teljesülését.

4. tézis: *[7][8]*

Megalkottam egy szöveg-funkció konverziós API könyvtárat, mely demonstrálja a javasolt természetes nyelvi vezérlő keretrendszer hatékonyságát. A könyvtár funkcionalitása két eltérő tématerülethez (robotvezérlés, navigáció) tartozó ipari megoldásban került kipróbálásra, mely során az eredmények azt mutatják, hogy a kifejlesztett rendszerek teljesítik az ipari elvárásokat a futási idő és a pontosság tekintetében egyaránt.

3. TOVÁBBI KUTATÁSI FELADATOK

A kidolgozott keretrendszert célszerű lenne valódi, nagy tudásbázissal tesztelni, illetve esetlegesen hangolni, pl. egy konkrét GPS eszköz teljes tudását felépítve. Jövőbeli tervek között szerepel annak vizsgálata is, hogy például a WordNet-tel összekapcsolható-e a szemantikai modul.

A kutatás további jövőbeli célja, hogy az egyszerű bővített mondatok mellett összetett mondatokat is tudjon feldolgozni a keretrendszer. Erre a szövegfeldolgozó és morfológiai modul fel van készítve, azonban a mondatelemző és a funkciótársító modulok algoritmusai átdolgozandóak, finomítandóak.

A nyelv-adaptálhatóság feltételeit a keretrendszer megteremti, a jövőbeli tervek között szerepel a különböző típusú nyelvekre (flektáló, izoláló, agglutináló) történő adaptáció, mint angol, szlovák, német, spanyol.

Napjainkban rohamosan terjednek a felhő alapú szolgáltatások. További vizsgálatok tárgyát képezi, hogy a jelenleg megvalósított keretrendszert hogyan lehetne felhőszolgáltatásként üzemeltetni.

4. SUMMARY

The main goal of our research is to develop a natural language controlling framework which uses mostly rule-based approaches. The aim is to make such a NLC system which can easily be adapted for several domains and for different languages as well. The implementation of the framework should consider the integration of existing solutions, working NLP engine modules into the framework components. The capabilities of the framework are fixed in advance which are:

1. **Domain-adaptivity:** the ability to easily learn concepts and relations of different domains without modifying the inner structure and workflows of the framework
2. **Language-adaptivity:** the ability to parse natural language sentences in different languages with only teaching language-dependent parts of framework
3. **Extendibility:** the ability to extend the set of functions which intended to be called by natural language commands
4. **Open interface:** the ability to reuse existing components of NLP engines and to implement and refine any part of the framework for own needs.

The new scientific results which are achieved during the completion of the project are summarized as follows.

Thesis 1: *[6][8]*

A novel structure of natural language controlling framework model has been developed fulfilling requirements specified in recommendations. The architecture of the framework is based on the developed formal information flow model. The framework contains four modules in a linear structure. The proposed architecture provides domain-adaptivity, language-adaptivity and high extensibility with an open interface. These properties provide high level reusability of the framework in software development in the field of human-machine interfaces.

Thesis 2: *[2][3][6][8]*

A novel semantic model is developed which satisfies the requirements of the knowledge representation format in our proposed natural language controlling system. The model is an extension of ECG model proposed by (Varga, 2011). *The implemented extensions (three new types of nodes, three novel edge types completed with morphology and POS labels) ensure a more efficient knowledge transformation between input sentences of a*

language and the function call generator module. The novelty of the developed function signature model is the inclusion of POS information for function mapping. A deterministic algorithm has been implemented to convert the sentence analysis tree into API function calls.

Thesis 3: *[2][3][9][15]*

I have determined the key cost components in the natural language controlling framework. The proposed multi-criteria cost model includes execution time, accuracy and resource consumption factors. An algorithm optimization was performed for the POS tagging, sentence analysis and function mapping modules. The proposed graph-based algorithm provides an efficient execution for the implementation of the framework in large-scale domains.

Thesis 4: *[7][8]*

A prototype Text-to-Function API library was developed to demonstrate the efficiency of the proposed natural language controlling framework. The functionality of the library was tested in different industrial solutions on two different domains (humanoid robot controlling, Google Maps navigation). The experiments show that the implemented systems meet the industrial requirements concerning the accuracy and the response time.

SAJÁT PUBLIKÁCIÓK AZ ÉRTEKEZÉS TÉMAKÖRÉBEN

Külföldön megjelent idegen nyelvű könyvfejezet

[1] **Kovács, L., Barabás, P., Répási, T.:** *Ontology-Based Semantic Models for Databases*, Handbook of Research on Innovations in Database Technologies and Applications: Current and Future Trends, IGI Global Publisher, Hersey (USA) 2009, ISBN 978-1-60566-242-8, Pp. 443-451

Nemzetközi folyóiratban megjelent, lektorált idegen nyelvű publikációk

[2] **Barabás, P., Kovács, L.:** *Optimization tasks in Conversion of Natural Language Text into Function Calls*, Topics in Intelligent Engineering and Informatics, ISSN: 2193-9411, Springer, 2013

[3] **Barabás, P., Kovács, L.:** *Efficient Encoding of Inflection Rules in NLP Sytems*, Scientific Bulletin, vol. 9 (XXVI), no. 2, 2012, ISSN 2285-438X, Pp. 11-16

Hazai folyóiratban megjelent, lektorált idegen nyelvű publikációk

[4] **Barabás, P., Kovács, L.:** *Estimation of Misclassification Error using Bayesian Classifier*, Publication of University of Miskolc, Production Systems and Information Engineering , Vol. 5, 2009, ISSN 1785-1270, Pp. 41-50.

[5] **Barabás, P., Kovács, L.:** *Efficient Classification of String Transformations using Markov Model*, GAMF Közlemények, XXI. évf., 2008, ISSN-1587-4400, Pp. 145-151

Nemzetközi folyóiratban megjelent, lektorált magyar nyelvű publikációk

[6] **Barabás Péter:** *Parancskinyerés magyar nyelvű szövegből*, A Gépipari Tudományos Egyesület Műszaki Folyóirata, LXIII. Évfolyam, 2012, pp. 71-74.

Nemzetközi konferencia kiadványban megjelent, lektorált, idegen nyelvű publikációk

[7] **Barabás, P., Kovács, L., Vircikova, M.:** *Robot Controlling in Natural Language*, The 3rd IEEE International Conference on Cognitive Infocommunications (CogInfoCom2012), Kosice, Slovakia, December 2-5, 2012, Pp. 181-186

[8] **Barabás, P., Kovács, L.:** *Requirement Analysis of Internal Modules of Natural Language Processing Engines*, 10th International Symposium on Application Machine Intelligence and Informatics, Her'lany (Slovakia) 2012, ISBN 978-1-4577-0196-2, pp. 41-46

[9] **Kovács, L., Barabás, P.:** *Experiences of building of context-free grammar tree*, 9th International Symposium on Application Machine Intelligence and Informatics, Smolenice (Slovakia) 2011, ISBN 978-1-4244-7429-5, pp. 67-71

[10] **Barabás, P., Kovács, L.:** *Implementation of Sentence Parser for Hungarian Language in Natural Language Processing*, 8th International Symposium on Applied Machine Intelligence and Informatics, Herl'any, Slovakia, 1/2010, ISBN 978-1-4244-6422-7, pp. 59-63

[11] **Kovács, L., Baksa-Varga, E., Répási, T., Barabás, P.:** *Clustering Based on Context Similarity*, Complexity and Intelligence of the Artificial and Natural Complex Systems, IEEE computer Society, ISBN 139780769536217, 2009., pp. 157-166

[12] **Kovács, L., Barabás, P.:** *Generalization Analysis of the CL and MM-based Classifications*, 6th International Symposium on Applied Machine Intelligence and Informatics, Herl'any (Slovakia) 2008, ISBN 978-1-4244-2105-3, pp. 39-43.

Nemzetközi konferencia kiadványban megjelent, nem lektorált, idegen nyelvű publikációk

[13] **Barabás, P., Kovács, L.:** *Usability of Summation Hack in Bayesian Classification*, 9th International Symposium of Hungarian Researchers on Computational Intelligence and Informatics, Budapest 2008

[14] **Barabás, P.:** *Rule Learning with MM-based Classification*, MicroCAD 2008 International Scientific Conference, Miskolc, 03/2008

[15] **Kovács, L., Barabás, P.:** *Cost Analysis of Classification using CL and MM*, 8th International Symposium of Hungarian Researchers on Computational Intelligence and Informatics, Budapest, 2007., pp. 227-237.

[16] **Kovács, L., Barabás, P.:** *Statistical Methods for Morhological Parsers*, 7th International Symposium of Hungarian Researchers on Computational Intelligence and Informatics, Budapest, 2006, pp. 523-531.

[17] **Barabás, P.:** *Automated* Type Checking in VFP, MicroCAD 2005 International Scientific Conference, Miskolc, pp. 229-234

Helyi konferencia kiadványban megjelent, nem lektorált, idegen nyelvű publikációk

[18] **Barabás, P.:** *Cost Analysis of Classification using (H)MM*, Forum of Ph.D. Students, Miskolc, 11/2007

[19] **Barabás, P.:** *Automations in Grammar Induction*, Forum of PhD Students, Miskolc, 11/2006

HIVATKOZÁSOK

Alexin, Z., Csirik, J., Kocsor, A., Miháltz, M., & Szarvas, G. (2006). Construction of the Hungarian EuroWordNet Ontology and its Application to Information Extraction. In *Proceedings of the Third International Global WordNet Conference (GWC-06)* (pp. 291-292). Jeju Island, Korea.

Apache, U. (2013, May 29). Retrieved from http://uima.apache.org

Bohnet, B., & Niver, J. (2012). A Transition-Based System for Joint Part-of-Speech Tagging and Labeled Non-Projective Dependency Parsing. EMNLP-CoNLL.

Chomsky, N. (1957). *Syntactic Structures. The Hague: Mouton & Co.*

Codd, E. (1974). Seven steps to rendezvous with the casual user. In *IFIP Working Conference Data Base Management* (pp. 179-200).

Cullingford, R. (1981). *SAM.*

Csendes, D., Csirik, J., Gyimóthy, T., & Kocsor, A. (2005). The Szeged Treebank. In V. e. Matouš ek, *Proceedings of the 8th International Conference on Text, Speech and Dialogue (TSD 2005)* (pp. 123-131). Karlovy Vary, Czech Republic: Springer LNAI 3658.

Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing. *International journal of human computer studies , 43,* 907-928.

Guttman, A. (1984). R-trees: a dynamic index structure for spatial searching. ACM.

Hutchins, J. (2005). *The history of machine translation in a nutshell.*

Hutchins, W. J. (2004). *The Georgetown-IBM experiment demonstrated in January 1954.* Springer Berlin Heidelberg.

J. R. Pierce, J. B. (1966). *Language and Machines — Computers in Translation and Linguistics. ALPAC report.* Washington, DC: National Academy of Sciences, National Research Council.

Kovács, L. (2004). *Adatbázisok tervezésének és kezelésének módszertana.* Budapest: ComputerBooks.

Kovács, L. (2010). Rule approximation in metric spaces. In *Applied Machine Intelligence and Informatics (SAMI), 2010 IEEE 8th International Symposium* (pp. 49-52). IEEE.

McCallum, A. K. (2002). *MALLET: A Machine Learning for Language Toolkit.* Retrieved from http://mallet.cs.umass.edu

MorphAdorner. (2009). Retrieved from http://morphadorner.northwestern.edu

Németh, L. (2011). *Hunspell: open source spell checking, stemming, morphlogical analysis and generation under GPL, LGPL or MPL licenses*, 1.3.2. Forrás: http://hunspell.sourceforge.net

NLTK. (2012). Retrieved from http://nltk.org

openNLP, A. (2010). *openNLP*. Retrieved from http://opennlp.apache.org.

Shank, R. C. (1975). *Conceptual Information Processing.* North Holland, Amsterdam.

StanfordNLP. (2013). *Standford NLP*. Retrieved from http://www-nlp.stanford.edu

Turing, A. (1950). *Computing machinery and intelligence.*

Varga, E. B. (2011). Ontology-based Semantic Annotation and Knowledge Representation in a Grammar Induction System. *Ph.D. Dissertation* . Miskolc, Hungary.

Weiner, P. (1973). Linear pattern matching algorithms. In *Switching and Automata Theory* (pp. 1-11). SWAT'08. IEEE Conference Record of 14th Annual Symposium: IEEE.

Weizenbaum, J. (1966). *ELIZA-A Computer Program for the study of Natural Language Communication between man and machines.*

Winograd, T. (1972). *Understanding Natural Language.* New York: Academic Press.

Woods, W., & Kaplan, R. (1977). Lunar rocks in natural English: Explorations in natural language question answering, Linguistic Structures Processing. In *Fundamental Studies in Computer Science 5* (pp. 521-569).

Zsibrita, J., Vincze, V., & Farkas, R. (2013). magyarlanc 2.0: szintaktikai elemzés és felgyorsított szófaji egyértelműsítés. In V. V. Tanács Attila, *IX. Magyar Számítógépes Nyelvészeti Konferencia.* (pp. 368-374). Szeged, Szegedi Tudományegyetem.