Rezultati vrednotenja dveh sistemov Čarovnik iz Oza

Melita Hajdinjak, France Mihelič

Laboratorij za umetno zaznavanje, sisteme in kibernetiko,
Fakulteta za elektrotehniko, Univerza v Ljubljani
Tržaška 25, 1000 Ljubljana
{melita.hajdinjak, france.mihelic}@fe.uni-lj.si

Povzetek

Opišemo postopek in rezultate vrednotenja učinkovitosti dveh sistemov Čarovnik iz Oza z ogrodjem PARADISE. Vpeljemo t. i. *parametre podatkovne zbirke*, ki odražajo velikost ter zgradbo podatkovne zbirke in se v literaturi o vrednotenju učinkovitosti sistemov za dialog ne pojavljajo. Izpeljemo funkciji učinkovitosti obeh sistemov Čarovnik iz Oza, ki nas vodita do spoznanja, da je predstavitev znanja oz. zgradba podatkovne zbirke sistema za dialog izjemnega pomena in da so parametri podatkovne zbirke pri vrednotenju sistemov za podajanje informacij nepogrešljivi.

Results from the Evaluation of two Wizard-of-Oz Systems

The results from the PARADISE evaluation of data from two Wizard-of-Oz experiments are given. The *database parameters* expressing the database size and the database structure, which have not so far been reported in the literature as costs for user satisfaction, are introduced. The performance functions for both Wizard-of-Oz systems lead to the conclusion that the system's knowledge representation is of great importance and that the database parameters are indispensable when evaluating the performance of information-providing dialogue systems.

1. Uvod

Z namenom omogočiti primerjavo različnih govornih vmesnikov, kjer nas zanima, v kolikšni meri posamezni dejavniki vplivajo na učinkovitost in kako strategija vodenja dialoga vpliva na zadovoljstvo uporabnikov, je bilo leta 1997 (Walker et al., 1997a) kot potencialna splošna metodologija vrednotenja učinkovitosti govornih vmesnikov predlagano ogrodje PARADISE (PARAdigm for DIalogue System Evaluation). Ogrodje PARADISE omogoča izpeljavo ocene učinkovitosti sistema kot uteženo linearno kombinacijo od domene odvisnih *parametrov uspešnosti naloge* in *cen dialoga* (tj. *parametri učinkovitosti dialoga* in *parametri kakovosti dialoga*), zajema pa model učinkovitosti sistema, ki za osnovni cilj postavlja maksimiranje zadovoljstva uporabnikov.

Model učinkovitosti sistema, ki ga zajema ogrodje PARADISE, trdi, da lahko funkcijo učinkovitosti sistema določimo z multiplo linearno regresijo (Seber, 1977) z zadovoljstvom uporabnikov kot odvisno spremenljivko ter parametri uspešnosti naloge in cen dialoga kot neodvisnimi spremenljivkami:

$$\text{Učinkovitost} = \alpha * \mathcal{N}(\kappa) - \sum_{i=1}^{n} w_i * \mathcal{N}(c_i).$$

Pri tem je α utež Kappa koeficienta κ, w_i so uteži cen dialoga c_i, $\mathcal{N}$ pa je funkcija normalizacije (Hajdinjak in Mihelič, 2006c). Z normalizacijo parametrov κ in c_i dosežemo relevantnost in primerljivost uteži preslikanih parametrov $\mathcal{N}(\kappa)$ in $\mathcal{N}(c_1), \ldots, \mathcal{N}(c_n)$.

Funkcija učinkovitosti tedaj omogoča napovedovanje zadovoljstva uporabnikov, vrednotenje učinkovitosti in izboljševanje sistema, primerjavo sistemov z istimi ali različnimi domenami, samodejno iskanje problematičnih dialogov in spreminjanje strategije vodenja dialoga že med interakcijo.

Ogrodje PARADISE smo uporabili pri vrednotenju učinkovitosti dveh nedograjenih sistemov za podajanje informacij o vremenu in vremenski napovedi (Žibert et al., 2004), s katerima smo izvajali eksperiment Čarovnik iz Oza (Hajdinjak in Mihelič, 2004). Eksperiment Čarovnik iz Oza je trenutno najboljša alternativa za zbiranje podatkov, ki izražajo jezik komunikacije človek–stroj. V teh eksperimentih so uporabniki prepričani, da se pogovarjajo s strojem – računalnikom, kar pa ni res. V resnici za računalnikom sedi človek (čarovnik), ki vsaj delno simulira delovanje sistema za dialog.

V skladu z našo trditvijo (Hajdinjak in Mihelič, 2006a), da je treba vplive samodejnega razpoznavanja govora iz sistema odstraniti, če želimo vrednotiti učinkovitost kakšnega drugega modula (v našem primeru modula za vodenje dialoga), je človek čarovnik v prvem sistemu simuliral razumevanje govora (razpoznavanje govora in razumevanje naravnega jezika) ter vodenje dialoga, v drugem sistemu pa le razumevanje govora.

Oba sistema sta se poleg načina vodenja dialoga razlikovala še v vrsti podatkovne zbirke – v prvem eksperimentu je sistem dostopal do relacijske zbirke vremenskih podatkov, v drugem pa do posebne sodelujoče podatkovne zbirke (Hajdinjak, 2006b). Ko bomo govorili o strukturi dialoga, bomo uporabljali pojma *konverzacijskih iger* in *konverzacijskih potez*. *Konverzacijske igre* povezujemo z željami oz. konverzacijskimi cilji, kot je na primer cilj pridobiti določeno informacijo, in so sestavljene iz zaporedja izjav, ki se začnejo s pobudo in končajo, ko je cilj igre dosežen ali igra prekinjena. Sestavne dele konverzacijskih

iger imenujemo *konverzacijske poteze*. To so izjave, deli izjav ali množice izjav, ki izražajo isto namero, kot je na primer potrditev ali preverjanje.

2. Izbira regresijskih parametrov

Tako kot avtorice ogrodja PARADISE (Walker et al., 1997a) smo izbrali en sam parameter uspešnosti naloge:

- **Kappa koeficient** (κ) meri uspešnost sistema pri reševanju nalog, ki mu jih naloži uporabnik. Napake, do katerih pride pri razumevanju govora in jih sistem v tekoči konverzacijski igri odpravi, ne znižajo vrednosti tega koeficienta. Ker je v naših eksperimentih razumevanje govora simuliral čarovnik, koeficient κ, izračunan iz podatkov prvega eksperimenta, kaže uspešnost oz. spretnost čarovnika in fleksibilnost grafičnega vmesnika, ki je čarovniku pomagal voditi dialog, pri reševanju navideznih nesporazumov med uporabnikom in čarovnikom. V drugem eksperimentu, ko je vodenje dialoga prevzel posebej za to nalogo zgrajen modul (Hajdinjak, 2006b), koeficient κ kaže uspešnost tega modula za vodenje dialoga pri reševanju navideznih nesporazumov med uporabnikom in čarovnikom, ki so nastali ali zaradi tipkarskih napak čarovnika ali zaradi neavtoriziranih posegov čarovnika v pomenske predstavitve uporabnikovih izjav.

Za parametre učinkovitosti dialoga smo izbrali:

- **Povprečni čas dialoga** (MET) meri povprečni čas trajanja informacijskih konverzacijskih iger, katerih namen je pridobiti določeno informacijo in jih uporabnik vpelje v času svoje interakcije s sistemom.

- **Povprečno število potez** (MUM) meri povprečno število konverzacijskih potez, ki jih uporabnik potrebuje za izvedbo ali prekinitev vpeljanih informacijskih iger.

Čeprav so cene dialoga definirane kot parametri, katerih minimiranje ugodno vpliva na zadovoljstvo uporabnikov, je včasih naravneje vzeti količine, katerih učinek je ravno obraten. Izbrali smo naslednje parametre kakovosti dialoga:

- **Izpolnitev naloge** (Comp) se nanaša na mnenje uporabnika o tem, ali je od sistema dobil odgovor na prvo vprašanje oz. prvo nalogo, ki smo mu jo v eksperimentu zastavili (Hajdinjak in Mihelič, 2004). Parameter Comp zavzame vrednost 0, če uporabnik meni, da ni dobil odgovora na svoje vprašanje, in vrednost 1 v nasprotnem primeru.

- **Število uporabnikovih iniciativ** (NUI) šteje začetne konverzacijske poteze, s katerimi uporabnik vpelje informacijske igre.

- **Povprečno število besed** (MWT) meri povprečno število besed, vsebovanih v konverzacijskih potezah uporabnika.

- **Povprečni čas odziva** (MRT) meri povprečni čas, ki ga sistem porabi, da se odzove. V prvem eksperimentu je bil ta čas povezan z izbiro odgovorov na grafičnem vmesniku, v drugem pa s tipkanjem pomenskih predstavitev uporabnikovih potez.

- **Število manjkajočih odzivov** (NMR) meri razliko med številom potez sistema in številom potez uporabnika. Ta parameter izraža tako število potez, ki sledijo, ko sistem v vnaprej določenem času ne zazna govora, kakor tudi nepripravljenost uporabnika, da bi sistem odzdravil.

- **Število neprimernih iniciativ** (NUR) in **delež neprimernih iniciativ** (URR) merita število oz. delež začetnih potez uporabnika, katerih vsebina ne ustreza domeni sistema.

- **Število neprimernih odzivov** (NIR) in **delež neprimernih odzivov** (IRR) merita število oz. delež kontekstno neprimernih potez sistema. Sem štejemo tudi poteze, s katerimi sistem uporabnika prosi, naj ponovi zadnjo izjavo.

- **Število napak** (Error) meri napake sistema, kamor štejemo prekinitve telefonske povezave, neustrezno oblikovane povedi in nasprotujoče si odgovore.

- **Število pomoči** (NHM) in **delež pomoči** (HMR) merita število oz. delež potez sistema, ki uporabniku pomagajo nadaljevati dialog.

- **Število preverjanj** (NCM) in **delež preverjanj** (CMR) merita število oz. delež potez, s katerimi sistem prosi za potrditev informacij, ki jih pridobi na osnovi zgodovine dialoga. V prvem eksperimentu čarovnik ni izvajal potez tega tipa. Čarovnik, ki je simuliral popolno razumevanje govora, je sicer na podlagi zgodovine dialoga sklepal o nenavedenih podatkih, za katere pa uporabnika ni prosil, da jih potrdi.

- **Število podanih informacij** (NGD) in **delež podanih informacij** (GDR) merita število oz. delež potez, s katerimi sistem uporabniku poda iskane informacije, ki jih najde v podatkovni zbirki.

- **Število relevantnih informacij** (NRD) in **delež relevantnih informacij** (RDR) merita število oz. delež potez sistema, ki uporabnika usmerjajo k izbiri relevantnih, dosegljivih podatkov.

- **Število nepodanih informacij** (NND) in **delež nepodanih informacij** (NDR) merita število oz. delež potez, s katerimi sistem uporabniku sporoča, da nima zahtevanega podatka in ga pri tem ne usmerja k izbiri relevantnih, dosegljivih podatkov. V prvem eksperimentu so to poteze, ki pravijo, da sistem zahtevane informacije trenutno nima ali je sploh ne ponuja. V drugem eksperimentu pusti sistem to vprašanje odprto.

- **Število prekinjenih zahtev** (NAR) in **delež prekinjenih zahtev** (ARR) merita število oz. delež informacijskih iger, ki jih uporabnik prekine še preden se končajo.

Tabela 1: Srednje vrednosti izbranih regresijskih parametrov v prvem (WOZ1) in drugem (WOZ2) eksperimentu Čarovnik iz Oza.

		WOZ1	WOZ2	p
uspešnost naloge	**Kappa koeficient** (κ)	0.94	0.98	
učinkovitost dialoga	**povprečni čas dialoga** (MET)	13.76 s	17.39 s	0.000
	povprečno število potez (MUM)	1.48 s	1.68 s	0.047
	izpolnitev naloge (Comp)	0.97	0.96	
	število uporabnikovih iniciativ (NUI)	6.49	7.51	0.005
	povprečno število besed (MWT)	9.32 s	7.56 s	0.000
	povprečni čas odziva (MRT)	5.13 s	6.38 s	0.000
	število manjkajočih odzivov (NMR)	0.60	0.75	
	število neprimernih iniciativ (NUR)	0.48	0.13	0.011
	delež neprimernih iniciativ (URR)	0.08	0.02	
	število neprimernih odzivov (NIR)	0.41	0.90	0.009
	delež neprimernih odzivov (IRR)	0.04	0.06	
kakovost dialoga	**število napak** (Error)	0.12	0.06	
	število pomoči (NHM)	0.32	0.40	
	delež pomoči (HMR)	0.03	0.03	
	število preverjanj (NCM)*	-	2.19	
	delež preverjanj (CMR)*	-	0.16	
	število podanih informacij (NGD)	4.07	4.35	
	delež podanih informacij (GDR)	0.67	0.58	
	število relevantnih informacij (NRD)	0.70	2.06	0.000
	delež relevantnih informacij (RDR)	0.10	0.28	0.005
	število nepodanih informacij (NND)	1.67	0.94	0.000
	delež nepodanih informacij (NDR)	0.22	0.12	
	število prekinjenih zahtev (NAR)	0.05	0.16	
	delež prekinjenih zahtev (ARR)	0.01	0.02	
	zadovoljstvo uporabnika (US)	34.08	31.96	0.015

Zgoraj uporabljene kratice za imena parametrov se nanašajo na angleške besedne zveze.

Izbrane parametre je treba določiti samodejno, če je to mogoče, v skrajnem primeru pa jih ročno označiti. Zavedati se namreč moramo, da neodvisne spremenljivke funkcije učinkovitosti, ki niso samodejno določljive, skrčijo uporabnost ogrodja PARADISE – samodejno iskanje problematičnih dialogov in spreminjanje strategije vodenja dialoga med interakcijo tedaj nista več mogoča.

V prvem eksperimentu Čarovnik iz Oza smo morali večino parametrov določiti ročno. Šele modul za vodenje dialoga (Hajdinjak, 2006b), vključen v drugi sistem Čarovnik iz Oza, ki je potek dialoga zelo dobro strukturiral, je omogočil samodejno določljivost velike večine izbranih parametrov. Še vedno je bilo samodejno nemogoče določiti naslednje parametre: **Kappa koeficient** (κ), **izpolnitev naloge** (Comp), **število neprimernih iniciativ** (NUR) in **število napak** (Error).

Zanimivo je, da se **število podanih informacij** (NGD) in **delež podanih informacij** (GDR), **število relevantnih informacij** (NRD) in **delež relevantnih informacij** (RDR) ter **število nepodanih informacij** (NND) in **delež nepodanih informacij** (NDR), ki jih imenujemo *parametri podatkovne zbirke*, v literaturi o vrednotenju učinkovitosti sistemov za dialog ne pojavljajo. Razlog je verjetno ta, da imajo razvijalci sistemov za dialog le redko na razpolago podatkovno zbirko, katere struktura bi bila tako zelo časovno odvisna in skopa, kot je naša. Omenjen tip parametrov pa vseeno ni ostal popolnoma neopažen. Walker, Litman, Kamm in Abella (Walker et al., 1998) razmišljajo, da bi velikost podatkovne zbirke lahko značilno vplivala na učinkovitost sistema za dialog.

Srednje vrednosti izbranih regresijskih parametrov v obeh eksperimentih Čarovnik iz Oza so podane v tabeli 1. Vrstice s parametri, katerih razlika srednjih vrednosti v obeh eksperimentih je statistično značilna (Studentov primerjalni test; $p < 0.05$), so potemnjene in navedena je pripadajoča p vrednost.

3. Izbira regresijskih parametrov

V obeh eksperimentih Čarovnik iz Oza so uporabniki ocenili svoje zadovoljstvo tako, da so podali stopnjo strinjanja z izjavami o obnašanju oz. učinkovitosti sistema (Hajdinjak in Mihelič, 2006c). Splošno **zadovoljstvo uporabnika** (US) smo dobili kot vsoto ocen, zbranih z vprašalnikom, ki ga predlaga ogrodje PARADISE (Hajdinjak in Mihelič, 2006c). Vrednosti parametra US zato ležijo med 8 in 40. Srednja vrednost US za prvi eksperiment je enaka 34.08 (s standardnim odklonom 5.07), za drugega pa 31.96 (s standardnim odklonom 4.99). Obe srednji vrednosti zadovoljstva uporabnikov se statistično značilno razlikujeta ($p < 0.015$). Glej tabelo 1.

Ker smo želeli poiskati razlike med obema različicama sistemov Čarovnik iz Oza, za odvisno spremenljivko MLR modela učinkovitosti nismo vzeli US, ampak le seštevek ocen, dodeljenih vprašanjem, ki se nanašajo na razlike med sistemoma. Menimo, da so vprašanja, ki te spremembe (tj. vodenje dialoga v povezavi s predstavitvijo znanja) najbolje merijo, naslednja:

2. *Ali vas je sistem razumel?* (ASR)
   Vprašanje naj bi merilo učinkovitost razumevanja govora. Ker pa je v naših eksperimentih čarovnik simuliral tako rekoč popolno razumevanje govora, to ni bilo tako. V drugem eksperimentu, ko čarovnik, v nasprotju s prvim eksperimentom, v pomenske predstavitve uporabnikovih potez ni dodajal podatkov, na katere se je dalo sklepati iz zgodovine dialoga, se to vprašanje nanaša predvsem na modul za vodenje dialoga oz. njegovo učinkovitost pri polnjenju predalčkov.

3. *Ali ste brez težav prišli do odgovorov na vaša vprašanja?* (TE)
   Vprašanje naj bi merilo težavnost pridobivanja informacij. Nedvomno se nanaša na uspešnost čarovnika pri uravnavanju dialoga oz. učinkovitost modula za vodenje dialoga. Pri tem ima pomembno vlogo tudi predstavitev znanja.

6. *Ali se je sistem na vaše izjave odzival hitro (brez pojasnilnih vprašanj)?* (SR)
   Vprašanje naj bi merilo ustreznost sistemovih odzivov. Uporabnike sprašuje po mnenju o strategiji vodenja dialoga, ki je bila v drugem eksperimentu del modula za vodenje dialoga.

7. *Ali se je sistem obnašal tako, kot ste med dialogom od njega pričakovali?* (EB)
   Vprašanje naj bi merilo ujemanje med pričakovanim in dejanskim obnašanjem sistema. Vsekakor je tesno povezano z načinom vodenja dialoga in predstavitvijo znanja, ki je predpogoj sodelujočega načina odgovarjanja.

Vsoto ocen, dodeljenih naštetim vprašanjem, smo imenovali **zadovoljstvo uporabnika z vodenjem dialoga in ravnijo sodelujočega odgovarjanja** (DM). Ta spremenljivka zavzame vrednosti med 4 in 20.

Tiste neodvisne spremenljivke, ki so bile z odvisno spremenljivko DM v zelo nizki korelaciji ($p > 0.05$), smo iz modela odstranili (Hajdinjak in Mihelič, 2006c). Z uporabo Studentovega testa z $n-2$ prostostnimi stopnjami, kjer je n velikost učne množice, tj. $n = 73$ v prvem eksperimentu in $n = 68$ v drugem eksperimentu, smo tako prišli do ugotovitve, da je v prvem eksperimentu z neodvisno spremenljivko DM značilno koreliralo 10 parametrov (in sicer MUM, Comp, NUI, NIR, IRR, NGD, GDR, NRD, NND in NDR), v drugem pa 8 (in sicer κ, MET, MUM, IRR, CMR, GDR, RDR in ARR). Iz teh množic smo odstranili še parametre, ki bi lahko povzročali multikolinearnost modelov.

4. Funkcije učinkovitosti

Po postopku vzvratne eliminacije (Seber, 1977) za delno F statistiko $F_{out} = 4$ pri p-vrednosti približno enaki 0.05 na celotni učni množici, pridobljeni v prvem eksperimentu Čarovnik iz Oza, z DM kot odvisno spremenljivko,

- ⇝ **povprečno število potez** (MUM),
- ⇝ **izpolnitev naloge** (Comp),
- ⇝ **število neprimernih odzivov** (NIR),
- ⇝ **število relevantnih informacij** (NRD) in
- ⇝ **število nepodanih informacij** (NND)

pa kot neodvisnimi spremenljivkami, smo identificirali in odstranili slabih 10% vzorcev osamelcev (Hajdinjak, 2006b). To so meritve, ki se nenavadno razlikujejo od velike večine ostalih meritev in zato nepredvidljivo vplivajo na natančnost modela (Tabachnick in Fidell, 1996).

Postopek vzvratne eliminacije smo ponovili na zmanjšani učni množici vzorcev. Tabela 2 podaja dobljene delne F statistike, pripadajoče koeficiente determinacije R^2 (Johnson in Wichern, 2002) ter parametre, ki jih v posameznih korakih iz modela učinkovitosti prvega sistema Čarovnik iz Oza odstranimo. Postopek vzvratne eliminacije ustavimo pred 4. korakom, ko delna F statistika preseže vrednost 4.

	F_i	R^2	odstranjen parameter
poln model	-	0.59	-
1. korak ($i = 1$)	0.00	0.59	NIR
2. korak ($i = 2$)	0.21	0.59	MUM
3. korak ($i = 3$)	3.32	0.57	NRD
4. korak ($i = 4$)	9.01	0.51	Comp

Tabela 2: Tabela vzvratne eliminacije za prvi sistem Čarovnik iz Oza in odvisno spremenljivko DM.

Iz začetnega MLR modela z vzvratno eliminacijo odstranimo tri parametre, in sicer NIR, MUM in NRD. Funkcija učinkovitosti za prvi sistem Čarovnik iz Oza in odvisno spremenljivko DM_1, ki se nanaša na podatke, pridobljene v prvem eksperimentu Čarovnik iz Oza, je zato taka:

$$\widehat{\mathcal{N}(\mathrm{DM}_1)} = 0.25 * \mathcal{N}(\mathrm{Comp}) - 0.65 * \mathcal{N}(\mathrm{NND}).$$

Dobljena funkcija učinkovitosti pojasnjuje 57% variance, tj. $R^2 = 0.57$. Najizrazitejši parameter, ki negativno vpliva na DM_1, je parameter podatkovne zbirke NND.

Po postopku vzvratne eliminacije za $F_{out} = 4$ pri p-vrednosti približno enaki 0.05 na celotni učni množici, pridobljeni v drugem eksperimentu Čarovnik iz Oza, z DM kot odvisno spremenljivko,

- ⇝ **Kappa koeficient** (κ),
- ⇝ **povprečni čas dialoga** (MET),
- ⇝ **delež preverjanj** (CMR),

$\rightsquigarrow$ **delež podanih informacij** (GDR) in

$\rightsquigarrow$ **delež prekinjenih zahtev** (ARR)

pa kot neodvisnimi spremenljivkami, smo identificirali in odstranili dobrih 7% vzorcev osamelcev.

Postopek vzvratne eliminacije smo ponovili na zmanjšani učni množici vzorcev. Tabela 3 podaja dobljene delne F statistike, pripadajoče koeficiente determinacije R^2 ter parametre, ki jih v posameznih korakih iz modela učinkovitosti drugega sistema Čarovnik iz Oza odstranimo. Postopek vzvratne eliminacije ustavimo pred $3.$ korakom, ko delna F statistika preseže vrednost 4.

	F_i	R^2	odstranjen parameter
poln model	-	0.48	-
1. korak ($i = 1$)	1.71	0.46	MET
2. korak ($i = 2$)	2.84	0.44	ARR
3. korak ($i = 3$)	12.59	0.32	κ

Tabela 3: Tabela vzvratne eliminacije za drugi sistem Čarovnik iz Oza in odvisno spremenljivko DM.

Iz začetnega MLR modela z vzvratno eliminacijo odstranimo dva parametra, in sicer MET in ARR. Funkcija učinkovitosti za drugi sistem Čarovnik iz Oza in odvisno spremenljivko DM_2, ki se nanaša na podatke, pridobljene v drugem eksperimentu Čarovnik iz Oza, je zato taka:

$$\widehat{\mathcal{N}(\text{DM}_2)} = 0.36{*}\mathcal{N}(\kappa) - 0.38{*}\mathcal{N}(\text{CMR}) + 0.40{*}\mathcal{N}(\text{GDR}).$$

Dobljena funkcija učinkovitosti pojasnjuje 44% variance, tj. $R^2 = 0.44$, in ima tri parametre – **Kappa koeficient** (κ) in **delež podanih informacij** (GDR) pozitivno vplivata na DM_2, **delež preverjanj** (CMR) pa negativno vpliva na DM_2.

Nobeden od parametrov, ki jih vsebuje $\widehat{\mathcal{N}(\text{DM}_1)}$, ni značilen za $\widehat{\mathcal{N}(\text{DM}_2)}$. Obratno pa ni res. Parameter podatkovne zbirke GDR, ki ima zelo velik pozitivni vpliv na DM_2, sicer ni vsebovan v $\widehat{\mathcal{N}(\text{DM}_1)}$, je pa visoko (negativno) koreliran s parametrom podatkovne zbirke NND, tj. najmočnejšim (negativnim) parametrom funkcije $\widehat{\mathcal{N}(\text{DM}_1)}$ (Hajdinjak, 2006b).

Analiza obeh funkcij učinkovitosti za DM omogoča vrednotenje učinkovitosti modula za vodenje dialoga, povezanega s sodelujočo podatkovno zbirko:

- Edini parameter, ki nastopa v funkciji učinkovitosti za DM_2 in je statistično značilen tudi za DM_1 ($p < 0.004$), je parameter podatkovne zbirke **delež podanih informacij** (GDR). V funkciji učinkovitosti za DM_1 namesto GDR sicer nastopa parameter podatkovne zbirke **število nepodanih informacij** (NND), ki je z njim visoko negativno koreliran in hkrati bolj značilen za DM_1 ($p < 0.0005$). Torej, parametri podatkovne zbirke predstavljajo edino podobnost med funkcijama učinkovitosti obeh sistemov Čarovnik iz Oza. Ta ugotovitev kaže na izjemno pomembnost predstavitve znanja oz. zgradbe podatkovne zbirke sistema za dialog. Pridemo do spoznanja, da so parametri podatkovne zbirke nepogrešljivi pri vrednotenju učinkovitosti sistemov za dialog, še posebej pa pri vrednotenju učinkovitosti sistemov za podajanje informacij.

- Medtem ko je parameter podatkovne zbirke **število nepodanih informacij** (NND) v prvem eksperimentu pomembno (negativno) vplival na zadovoljstvo uporabnikov, je njegov (negativni) vpliv v drugem eksperimentu izjemno splahnel. Vemo že (tabela 1), da se je srednja vrednost parametra **število relevantnih informacij** (NRD) v drugem eksperimentu značilno povečala, srednja vrednost NND pa zato značilno zmanjšala. Vse torej kaže na to, da zmanjšanje števila odzivov, s katerimi sistem uporabniku sporoča, da zahtevane informacije nima, hkrati pa mu ne ponudi nobenih dosegljivih, relevantnih informacij, negativno vpliva na zadovoljstvo uporabnika. Razvijalci sistemov za dialog morajo zato težiti k zmanjšanju števila takih odzivov oz. povečanju stopnje sodelujočega odgovarjanja. Sklepamo lahko tudi, da strategija usmerjanja uporabnika k izbiri dosegljivih, relevantnih podatkov, ki je implementirana v modulu za samodejno vodenje dialoga, na zadovoljstvo uporabnikov ne vpliva negativno.

- Ugotovili smo, da so bili uporabniki v prvem eksperimentu bolj dojemljivi za kvantitativne parametre (tj. NUR, NIR, NHM, NGD, NRD, NND, NAR), uporabniki v drugem eksperimentu pa za njim pripadajoče proporcionalne parametre (tj. URR, IRR, HMR, GDR, RDR, NDR, ARR). Funkcija učinovitosti za DM_1 vsebuje, poleg parametra Comp, še kvantitativni parameter **število nepodanih informacij** (NND). Funkcija učinkovitosti za DM_2 pa vsebuje, poleg parametra κ, še dva proporcionalna parametra, namreč **delež preverjanj** (CMR) in **delež podanih informacij** (GDR). Menimo, da je to posledica konsistentno povečanega ponujanja relevantnih informacij v drugem eksperimentu, ki je vodilo do več novih informacijskih iger in s tem do večje dojemljivosti uporabnikov za proporcionalne količine. Vsekakor so glede tega potrebne nadaljnje raziskave.

- Parametra **Kappa koeficient** (κ) in **izpolnitev naloge** (Comp) sta bila v naših eksperimentih nekorelirana. V prvem eksperimentu je na zadovoljstvo uporabnikov DM_1 močno (pozitivno) vplival Comp, κ ni imel statistično značilnega vpliva. V drugem eksperimentu je bilo ravno obratno – na zadovoljstvo uporabnikov DM_2 je močno (pozitivno) vplival κ, Comp pa ni imel statistično značilnega vpliva. Ugotovitev, do katere so prišle Walker, Litman, Kamm in Abella (Walker et al., 1998), da **izpolnitev naloge** (Comp) močneje vpliva na zadovoljstvo uporabnika kot **Kappa koeficient** (κ), torej ni vedno resnična. Le parameter Comp, katerega vrednost mora posredovati uporabnik, za vrednotenje učinkovitosti sistemov za dialog zato ni dovolj. Še vedno je dobro meriti tudi κ, ki pa ga na žalost prav tako ni mogoče določiti samodejno.

- Parameter, ki na zadovoljstvo uporabnikov DM_2

najmočneje negativno vpliva, je **delež preverjanj** (CMR). Sistem za dialog lahko torej izboljšamo, če zmanjšamo delež potez, ki preverjajo točnost podatkov, pridobljenih na osnovi zgodovine dialoga, ki jih uporabnik v svoji izjavi ne poda ali jih sistem ne razume. Vpliv parametra CMR v sistemih za dialog ni mogoče popolnoma odpraviti, zato ker je določeno število preverjanj nujno vsakič, ko imamo opravka s samodejnim razumevanjem govora. Napake, ki se pojavljajo pri samodejnem razumevanju govora, sistem namreč prisilijo, da svoje razumevanje uporabnikovih izjav preveri vsakič, ko o njihovi pravilnosti ni popolnoma prepričan. Če tega ne bi počel, bi nekontrolirano podajal napačne odgovore. To bi povečalo srednjo vrednost parametra **delež neprimernih odzivov** (IRR) in tako zelo verjetno vodilo do večjega nezadovoljstva s sistemom.

Funkciji učinkovitosti obeh sistemov Čarovnik iz Oza z **zadovoljstvom uporabnika** (US) kot odvisno spremenljivko sta se zelo razlikovali v natančnosti ($R^2 = 0.58$ proti $R^2 = 0.24$) (Hajdinjak, 2006b). Potem ko smo za odvisno spremenljivko vzeli **zadovoljstvo uporabnika z vodenjem dialoga in ravnijo sodelujočega odgovarjanja** (DM), nam je uspelo razliko v natančnosti izjemno zmanjšati ($R^2 = 0.57$ proti $R^2 = 0.44$). Upravičeno lahko torej trdimo, da se da DM veliko bolje modelirati kot US.

Povejmo še, da literatura o vrednotenju učinkovitosti sistemov za dialog z ogrodjem PARADISE v glavnem poroča o koeficientih determinacije R^2, ki so blizu mejne vrednosti 0.5, pogosto precej nižje (Walker et al., 1997b; Walker et al., 1998; Walker et al., 2001; Möller, 2005), le redko pa presežejo vrednost 0.6 (Litman in Shimei, 2002).

5. Sklep

Ogrodje PARADISE smo uporabili pri vrednotenju učinkovitosti dveh nedograjenih sistemov za podajanje informacij o vremenu in vremenski napovedi, s katerima smo izvajali eksperiment Čarovnik iz Oza. Za namene vrednotenja smo izbrali in določili 25 regresijskih parametrov. Pri vrednotenju učinkovitosti sistemov za podajanje informacij smo predlagali še neuveljavljene parametre podatkovne zbirke, ki izražajo velikost in sestavo podatkovne zbirke. V raziskave smo vključili kvantitativne in proporcionalne parametre podatkovne zbirke. Ugotovili smo, da so bili uporabniki prvega sistema bolj dojemljivi za kvantitativne parametre, v drugem pa za proporcionalne parametre.

Ker smo želeli poiskati razlike med dvema sistemoma Čarovnik iz Oza, ki sta se razlikovala le v načinu vodenja dialoga in predstavitvi znanja, smo mero zadovoljstva uporabnikov definirali kot vsoto ocen, ki se nanašajo na vpeljane spremembe. Po vzvratni eliminaciji smo dobili funkciji učinkovitosti, ki ne vsebujeta nobenega skupnega parametra. Edini parameter, ki nastopa v funkciji učinkovitosti drugega sistema in je bil statistično značilen tudi v prvem eksperimentu, je eden od parametrov podatkovne zbirke. Prišli smo do spoznanja, da so parametri podatkovne zbirke edina podobnost med funkcijama učinkovitosti obeh sistemov Čarovnik iz Oza in da ima predstavitev znanja v sistemih za podajanje informacij velik pomen.

6. Literatura

M. Hajdinjak in F. Mihelič. 2004. Conducting the wizard-of-oz experiment. *Informatica*, 28(4):425–430.

M. Hajdinjak in F. Mihelič. 2006a. The paradise evaluation framework: Issues and findings. *Computational Linguistics*, 32.

M. Hajdinjak. 2006b. *Predstavitev znanja in vrednotenje učinkovitosti sodelujočih samodejnih sistemov za dialog, Doktorska disertacija*. Fakulteta za elektrotehniko, Univerza v Ljubljani, Ljubljana.

M. Hajdinjak in F. Mihelič. 2006c. Vrednotenje govornih vmesnikov z ogrodjem paradise. V: *Zbornik IS-LTC 2006 9. mednarodne multikonference Informacijska družba IS'2006*. Ljubljana, Slovenija.

R. A. Johnson in D. W. Wichern. 2002. *Applied multivariate statistical analysis*. Prentice-Hall, Upper Saddle River (NJ).

D. J. Litman in P. Shimei. 2002. Designing and evaluating an adaptive spoken dialogue system. *User Modeling and User-Adapted Interaction*, 12(2-3):111–137.

S. Möller. 2005. Evaluating telephone-based interactive systems. V: *Proceedings of the COST278 Final Workshop and ISCA Tutorial and Research Workshop (ITRW) on Applied Spoken Language Interaction in Distributed Environments*. Aalborg, Danska.

G. A. F. Seber. 1977. *Linear Regression Analysis*. John Wiley & Sons, New York.

B. G. Tabachnick in L. S. Fidell. 1996. *Using Multivariate Statistics, Third Edition*. Harper Collins, New York.

J. Žibert, S. Martinčić-Ipšić, M. Hajdinjak, I. Ipšić, in F. Mihelič. 2004. Development of a bilingual spoken dialog system for weather information retrieval. V: *Proceedings of the 8th European Conference on Speech Communication and Technology*, str. 1917–1920. Ženeva, Švica.

M. A. Walker, D. Litman, C. A. Kamm, in A. Abella. 1997a. Paradise: A framework for evaluating spoken dialogue agents. V: *Proceedings of the 35th Annual Meeting of the Association of Computational Linguistics*, str. 271–280. Madrid, Španija.

M. A. Walker, D. Hindle, J. Fromer, G. Di Fabbrizio, in C. Mestel. 1997b. Evaluating competing agent strategies for a voice email agent. V: *Proceedings of the 5th European Conference on Speech Communication and Technology*, str. 2219–2222. Rodos, Grčija.

M. A. Walker, D. J. Litman, C. A. Kamm, in A. Abella. 1998. Evaluating spoken dialogue agents with paradise: Two case studies. *Computer Speech and Language*, 12(3):317–347.

M. A. Walker, R. Passonneau, in J. E. Boland. 2001. Quantitative and qualitative evaluation of darpa communicator spoken dialogue systems. V: *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, str. 515–522. Toulouse, Francija.