Vrednotenje govornih vmesnikov z ogrodjem PARADISE

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 potencialno splošno metodologijo vrednotenja učinkovitosti sistemov za dialog, imenovano 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, zajema pa model učinkovitosti sistema, katerega osnovni cilj je maksimirati zadovoljstvo uporabnikov. Model učinkovitosti sistema, ki ga zajema ogrodje PARADISE, trdi, da lahko funkcijo učinkovitosti sistema določimo z multiplo linearno regresijo. Osredotočimo se na nekatere pomanjkljivosti, težave in nerešena vprašanja ogrodja PARADISE – opišemo vpliv normalizacije na natančnost napovedovanja zadovoljstva uporabnikov, navedemo regresijske predpostavke in poudarimo pomembnost dobre izbire regresijskih parametrov. Nakažemo tudi morebitne učinke razpoznavanja govora na rezultate vrednotenja in analiziramo vprašalnik, na podlagi katerega ogrodje PARADISE določa zadovoljstvo uporabnikov (tj. odvisno spremenljivko funkcije učinkovitosti). V dosedanjih raziskavah so bile nekatere izmed naštetih točk premalo poudarjene, zanemarjene ali celo napačno interpretirane.

Speech-interface evaluation using the PARADISE framework

We introduce the PARADISE (PARAdigm for DIalogue System Evaluation) framework, a potential general methodology for evaluating spoken-language dialogue systems. The PARADISE framework maintains that the system's primary objective is to maximize user satisfaction, and it derives a combined performance metric for a dialogue system as a weighted linear combination of task-success measures and dialogue costs. The PARADISE model of performance posits that a performance function can be derived by applying multivariate linear regression with user satisfaction as the dependent variable and task-success measures and dialogue costs as the independent variables. We focus on some PARADISE issues (with most of them arising from the application of multivariate linear regression) that have, up to now, not been sufficiently emphasized or have even been neglected by the dialogue-system community. These include considerations regarding the selection of appropriate regression parameters, normalization effects on the accuracy of the prediction, the influence of speech-recognition errors on the performance function, and the selection of an appropriate user-satisfaction measure.

1. Uvod

Avtomatizacija sporazumevanja z govorom je še vedno eden izmed največjih raziskovalnih izzivov. Razlogov je več:

- Govor je naraven – govoriti se naučimo, še preden znamo brati in pisati.
- Govor je učinkovit – večina ljudi je sposobnih govoriti petkrat hitreje kot tipkati in verjetno celo desetkrat hitreje kot pisati.
- Govor je fleksibilen – med sporazumevanjem z govorom se nam ni treba ničesar dotikati in ne opazovati.

Računalniški sistem, ki uporabniku omogoča, da z govorom dostopa do določenih aplikacij, imenujemo *sistem za dialog* ali *govorni vmesnik*.

Z razvojem sistemov za dialog se pojavljajo tudi potrebe po vrednotenju učinkovitosti in medsebojni primerjavi takih sistemov. Težava, ki se pojavi, je ta, da vrednotenja učinkovitosti sistema za dialog ni mogoče omejiti na primerjave z referenčnimi odgovori oz. referenčnimi poteki dialogov (Bates in Ayuso, 1991; Polifroni et al., 1992; Price et al., 1992). Množica sprejemljivih dialogov je namreč lahko zelo velika. Naslednja težava je veliko število potencialnih metrik dialoga. Sistem za dialog lahko npr. vrednotimo glede na njegovo sposobnost pomagati uporabnikom pri doseganju ciljev, glede na njegovo robustnost odkrivanja in premagovanja napak, ki se pojavljajo pri razpoznavanju oz. razumevanju govora, ali glede na skupno kakovost interakcije (Polifroni et al., 1992; Price et al., 1992; Danieli in Gerbino, 1995), ki jo omogoča.

Predlogi vrednotenja učinkovitosti sistemov za dialog, ki so se pojavili v zadnjih dveh desetletjih dvajsetega stoletja, se osredotočajo na razvoj različnih metrik dialoga. Predlagani so bili številni *objektivni parametri dialoga* (Price et al., 1992; Danieli in Gerbino, 1995) kot npr. število izjav, čas dialoga, povprečni čas odziva uporabnika, povprečni čas odziva sistema, delež izjav, sestavljenih iz več kot ene besede, ter povprečna dolžina izjav, sestavljenih iz več kot ene besede, ki jih lahko določimo brez mnenja človeka, in parametri, ki temeljijo na mnenju človeka, namreč *subjektivni parametri dialoga* (Shriberg et al., 1992; Danieli in Gerbino, 1995) kot npr. delež izjav, s katerimi sistem popravlja napake, delež kontekstno primernih izjav sistema, hevristično vrednotenje stopnje sodelovanja sistema na podlagi Griceovih maksim (Grice, 1975), delež pravilnih in delno pravilnih odgovorov, delež primernih in neprimernih izjav, s katerimi sistem usmerja uporabnika, ter zadovoljstvo uporabnika (Shriberg et al., 1992).

Z namenom omogočiti primerjavo sistemov z različnimi domenami, kjer je pomembno vedeti, v kolikšni meri posamezni parametri vplivajo na učinkovitost in kako strategija vodenja dialoga vpliva na zadovoljstvo uporabnikov, je bilo leta 1997 kot potencialna splošna metodologija vrednotenja učinkovitosti sistemov za dialog predlagano *ogrodje PARADISE* (PARAdigm for DIalogue System Evaluation) (Walker et al., 1997a).

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. Ogrodje PARADISE

Ogrodje PARADISE omogoča izpeljavo ocene učinkovitosti sistema kot uteženo linearno kombinacijo *parametrov uspešnosti naloge* in *cen dialoga*, zajema pa model učinkovitosti sistema, ki za osnovni cilj postavlja maksimirati zadovoljstvo uporabnikov, kar doseže z maksimiranjem parametrov uspešnosti naloge in minimiranjem cen dialoga.

Zadovoljstvo uporabnikov ponavadi merimo z vprašalniki, v katerih uporabniki podajo stopnjo strinjanja z izjavami o različnih vidikih svoje interakcije s sistemom za dialog. Avtorice ogrodja PARADISE (Walker et al., 1997a) v ta namen uporabljajo vprašalnik, podan v tabeli 1. Vprašanja (v naštetem vrstnem redu) sprašujejo po učinku modula za tvorjenje govora, učinku modula za razpoznavanje govora, težavnosti pridobivanja informacij, hitrosti interakcije, izkušenosti uporabnikov, ustreznosti odzivov sistema, pričakovanem obnašanju sistema in načrtovani rabi sistema v prihodnosti. Večino odgovorov podajo opisno kot *skoraj nikoli*, *redko*, *včasih*, *pogosto* in *skoraj vedno*, nekatere pa le z *da*, *ne* in *mogoče*. Te potem preslikajo v množico naravnih števil od 1 do 5, pri čemer 1 pomeni najmanjšo, 5 pa največjo stopnjo strinjanja. Parameter, ki ocenjuje zadovoljstvo uporabnikov, dobijo kot vsoto vseh ocen in ga poimenujejo **zadovoljstvo uporabnika** (US).

Cene dialoga, tj. parametre dialoga, katerih minimizacija ugodno vpliva na zadovoljstvo uporabnikov, razdelimo v dve kategoriji: *parametri učinkovitosti* dialoga in *parametri kakovosti* dialoga. Parametri učinkovitosti dialoga (npr. število izjav, ki jih uporabnik potrebuje, da uresniči svojo namero, ali čas dialoga) merijo, kako učinkovito sistem uporabniku pomaga pri doseganju njegove namere. Parametri kakovosti dialoga (npr. kolikokrat mora uporabnik ponoviti svojo izjavo, da ga sistem razume, ali kakšen je čas čakanja na odziv sistema) pa zajemajo ostale vidike, ki lahko na zadovoljstvo uporabnika prav tako močno vplivajo. Ker vnaprej ni jasno, katere cene dialoga bodo najmočneje vplivale na zadovoljstvo uporabnikov, je pomembno, da v empiričnih raziskavah uporabljamo širok spekter teh parametrov (Walker et al., 1998).

Tabela 1: Vprašalnik za ocenjevanje zadovoljstva uporabnikov, ki ga predlaga ogrodje PARADISE.

1.	Ali ste sistem brez težav razumeli?
2.	Ali vas je sistem razumel?
3.	Ali ste brez težav prišli do odgovorov na vaša vprašanja?
4.	Ali je bila hitrost interakcije s sistemom primerna?
5.	Ali ste na vsakem koraku dialoga vedeli, kaj morate povedati?
6.	Ali se je sistem na vaše izjave odzival hitro (brez pojasnilnih vprašanj)?
7.	Ali se je sistem obnašal tako, kot ste med dialogom od njega pričakovali?
8.	Glede na vašo trenutno izkušnjo s sistemom, ali mislite, da boste sistem še kdaj poklicali?

Uspešnost naloge, ki se lahko nanaša na celoten dialog ali del dialoga, ki predstavlja zaključeno celoto, pomeni stopnjo ujemanja med vsebino zahtev uporabnika in dojemanjem te s strani sistema za dialog. Ogrodje PARADISE uporablja en sam parameter uspešnosti naloge, namreč **Kappa koeficient** (Carletta, 1996). **Kappa koeficient** (κ) izračunamo z uporabo Cohenove metode (Di Eugenio in Glass, 2004) in kontingenčne tabele, ki podaja ujemanje med vsebino zahtev uporabnika in dojemanjem te s strani sistema.

2.1. Model učinkovitosti

Če želimo sistem za dialog vrednotiti z ogrodjem PARADISE, moramo podatke zbrati v eksperimentu, v katerem bodo uporabniki ocenili svoje zadovoljstvo. Ostale parametre modela (parametri uspešnosti naloge, cene dialoga) pa je treba določiti samodejno ali jih ročno označiti.

Model učinkovitosti sistema, ki ga zajema ogrodje PARADISE, trdi, da lahko funkcijo učinkovitosti sistema določimo z uporabo *multiple linearne regresije* (MLR) z zadovoljstvom uporabnikov kot neodvisno spremenljivko ter parametri uspešnosti naloge, parametri učinkovitosti dialoga in parametri kakovosti 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 edinega parametra uspešnosti naloge, namreč Kappa koeficienta κ, w_i so uteži cen dialoga c_i,

$\mathcal{N}$ pa je funkcija normalizacije:

$$\mathcal{N}(x) = \frac{x - \overline{x_0}}{\sigma_{x_0}}$$

Z $\overline{x_0}$ in σ_{x_0} smo označili srednjo vrednost in standardni odklon spremenljivke x_0 v učni množici, pridobljeni v ustreznem eksperimentu. Srednja vrednost s funkcijo normalizacije $\mathcal{N}$ preslikanih parametrov učne množice je 0, varianca in standardni odklon pa 1. Tako se znebimo težav, ki se pojavijo, če primerjamo vrednosti parametrov, ki se raztezajo na različnih intervalih in/ali so njihove vrednosti različno razpršene. Z normalizacijo parametrov κ in c_i dosežemo relevantnost in primerljivost uteži preslikanih parametrov $\mathcal{N}(\kappa)$ in $\mathcal{N}(c_1), \ldots, \mathcal{N}(c_n)$.

Rezultat multiple linearne regresije na učni množici parametrov, ki praviloma tvorijo predoločen sistem, je torej množica uteži, ki pomenijo sorazmeren prispevek teh parametrov k učinkovitosti sistema. Funkcija učinkovitosti, ki jo uvaja ogrodje PARADISE, zato omogoča:

- napovedovanje zadovoljstva uporabnikov,
- vrednotenje učinkovitosti sistema za dialog, tj. ugotavljanje vpliva posameznih parametrov na zadovoljstvo uporabnikov,
- izboljšanje sistema za dialog, tj. odpravljanje ali zmanjšanje vpliva parametrov, ki imajo najbolj negativne uteži in povečanje vpliva parametrov, ki imajo najbolj pozitivne uteži,
- primerjavo različnih sistemov za dialog, tj. primerjavo vplivov posameznih parametrov v pripadajočih funkcijah učinkovitosti, iz katerih lahko razberemo razlike med sistemi,
- samodejno iskanje problematičnih dialogov, tj. iskanje dialogov, katerih napovedano zadovoljstvo uporabnikov negativno izstopa, ter
- spreminjanje strategije vodenja dialoga med samo interakcijo, tj. spreminjanje načina sporazumevanja na osnovi napovedanega zadovoljstva uporabnika v že izvedenem delu interakcije.

V zadnjih letih je bilo opravljenih veliko študij učinkovitosti sistemov za dialog, ki so uporabljale ogrodje PARADISE (Walker et al., 1998; Kamm et al., 1998; Walker, 2000; Litman in Shimei, 2002; Larsen, 2003; Hajdinjak, 2006b). Ogrodje PARADISE je postala celo najbolj citirana metoda vrednotenja učinkovitosti sistemov za dialog.

3. Analiza ogrodja PARADISE

Osredotočili se bomo na nekatere pomanjkljivosti, težave in nerešena vprašanja ogrodja PARADISE (Hajdinjak in Mihelič, 2006a). Večina jih izvira ravno iz uporabe multiple linearne regresije.

3.1. Vpliv normalizacije na natančnost napovedovanja zadovoljstva uporabnikov

Multipla linearna regresija temelji na metodi najmanjših kvadratov, tj. minimira vsoto kvadratov razlik med v eksperimentu pridobljenimi vrednostmi (tj. učne množice) in napovedanimi vrednostmi zadovoljstva uporabnikov. Za dano vrednost zadovoljstva uporabnika US torej velja

$$\mathcal{N}(US) = \widehat{\mathcal{N}(US)} + \epsilon,$$

kjer je $\mathcal{N}(US)$ normalizirana pridobljena vrednost zadovoljstva uporabnika, $\widehat{\mathcal{N}(US)}$ napovedana normalizirana vrednost zadovoljstva uporabnika, ϵ pa napaka napovedi. Ker je srednja vrednost napake ϵ enaka 0, sta srednji vrednosti odvisne spremenljivke in njene napovedi enaki. Nenormalizirano zadovoljstvo uporabnika US lahko tedaj ocenimo kot

$$US = \widehat{\mathcal{N}(US)}\sigma_{US_0} + \overline{US}_0 + \epsilon\sigma_{US_0} = \widehat{US} + \epsilon\sigma_{US_0},$$

kjer sta $\overline{US}_0$ in σ_{US_0} srednja vrednost in standardni odklon v eksperimentu pridobljenih vrednosti zadovoljstva uporabnikov. Vidimo, da se napaka ocene normaliziranega zadovoljstva uporabnika $\mathcal{N}(US)$ pri tem poveča za faktor σ_{US_0}.

Kako dobro $\widehat{US}$ napoveduje US, kaže razmerje absolutnih vrednosti njune razlike in pridobljene vrednosti zadovoljstva uporabnika US:

$$q(US, \widehat{US}) = \frac{|US - \widehat{US}|}{|US|}$$

Naslednje razmerje pa kaže, da ocena normalizirane vrednosti zadovoljstva uporabnika $\widehat{\mathcal{N}(US)}$ ni vedno tako dobra kot ocena nenormalizirane vrednosti $\widehat{US}$:

$$\frac{q(\mathcal{N}(US), \widehat{\mathcal{N}(US)})}{q(US, \widehat{US})} = \frac{\frac{|\mathcal{N}(US) - \widehat{\mathcal{N}(US)}|}{|\mathcal{N}(US)|}}{\frac{|US - \widehat{US}|}{|US|}} = \frac{|US|}{|US - \overline{US}_0|}$$

Za $US > \frac{\overline{US}_0}{2}$ namreč velja:

$$\frac{q(\mathcal{N}(US), \widehat{\mathcal{N}(US)})}{q(US, \widehat{US})} = \frac{|US|}{|US - \overline{US}_0|} > 1$$

Iz prikazanega sledi, da je napovedano normalizirano vrednost zadovoljstva uporabnika $\widehat{\mathcal{N}(US)}$ treba transformirati nazaj na začetni interval, saj je ocena nenormalizirane vrednosti zadovoljstva uporabnika $\widehat{US}$ v večini primerov veliko boljša. To naredimo s transformacijo

$$\widehat{US} = \widehat{\mathcal{N}(US)}\sigma_{US_0} + \overline{US}_0,$$

ki je inverzna normalizaciji.

Ne samo da ustrezna literatura (Walker et al., 1997a; Walker et al., 1998; Walker, 2000; Litman in Shimei, 2002; Larsen, 2003) vplivu normalizacije ne posveča pozornosti, ampak tudi ne omenja, da je treba vrednosti zadovoljstva uporabnikov, preden začnemo izpeljavo modela učinkovitosti, normalizirati, če želimo preprečiti prevelike napake ocen (Hajdinjak, 2006b).

Obstaja več načinov merjenja natančnosti MLR modelov. Najpogosteje se uporablja *koeficient (multiple) determinacije*,

$$R^2 = \frac{\sum_{i=1}^{m}(\widehat{\mathbf{X}}_i - \overline{\mathbf{X}})^2}{\sum_{i=1}^{m}(\mathbf{X}_i - \overline{\mathbf{X}})^2},$$

tj. razmerje pojasnjene variance in celotne variance $var(\mathbf{X})$, pri čemer smo z m označili število enačb učne množice. Celotna varianca je vsota pojasnjene variance in nepojasnjene variance:

$$var(\mathbf{X}) = \frac{1}{m}\sum_{i=1}^{m}(\widehat{\mathbf{X}}_i - \overline{\mathbf{X}})^2 + \frac{1}{m}\sum_{i=1}^{m}(\mathbf{X}_i - \widehat{\mathbf{X}}_i)^2$$

Nepojasnjena varianca oz. srednja kvadratna napaka,

$$\overline{\epsilon^2} = \frac{1}{m}\sum_{i=1}^{m}(\mathbf{X}_i - \widehat{\mathbf{X}}_i)^2,$$

je ravno količina, ki jo multipla linearna regresija minimira. Koeficient determinacije zavzame vrednosti med 0 in 1. Vrednosti, ki so bližje 1, pomenijo večjo natančnost modela, tj. boljšo linearno zvezo med odvisno spremenljivko in neodvisnimi spremenljivkami. Če koeficient determinacije R^2 pomnožimo s faktorjem 100, rezultat imenujemo *odstotek pojasnjene variance*.

Izkaže se, da je v MLR modelu z normaliziranimi spremenljivkami koeficient determinacije R^2 enak varianci napovedanih vrednosti:

$$R^2 = \frac{\sum_{i=1}^{m}\widehat{\mathcal{N}(\mathbf{US}_i)}^2}{m} = var(\widehat{\mathcal{N}(\mathbf{US})})$$

Pri tem smo z $\mathbf{US}_i$ označili i-to komponento vektorja $\mathbf{US}$ pridobljenih vrednosti zadovoljstva uporabnikov, z $\widehat{\mathcal{N}(\mathbf{US}_i)}$ pa i-to komponento vektorja $\widehat{\mathcal{N}(\mathbf{US})}$ napovedanih normaliziranih vrednosti zadovoljstva uporabnikov. Zadnja enakost velja zato, ker je $\overline{\widehat{\mathcal{N}(\mathbf{US})}}$, torej srednja vrednost napovedanih normaliziranih vrednosti zadovoljstva uporabnikov, enaka $\overline{\mathcal{N}(\mathbf{US})} = 0$. Zanimiva posledica te ugotovitve je, da so uteži funkcije učinkovitosti po absolutni vrednosti navzgor omejene z 1. Za MLR model

$$\widehat{\mathbf{X}} = \sum_{i=1}^{n}\alpha_i\mathbf{X_i}$$

namreč velja naslednje:

$$var(\widehat{\mathbf{X}}) = \sum_{i=1}^{n}\alpha_i^2 + 2\sum_{i=1}^{n}\sum_{j=i+1}^{n}\alpha_i\alpha_j corr(\mathbf{X_i}, \mathbf{X_j}),$$

pri čemer je

$$corr(\mathbf{X_i}, \mathbf{X_j}) = \frac{\frac{1}{m}\sum_{k=1}^{m}(\mathbf{X_i}_k - \overline{\mathbf{X_i}})(\mathbf{X_j}_k - \overline{\mathbf{X_j}})}{\sigma_{\mathbf{X_i}}\sigma_{\mathbf{X_j}}}$$

korelacija oz. *korelacijski koeficient* spremenljivk $\mathbf{X_i}$ in $\mathbf{X_j}$. Ker ima dvojna vsota v $var(\widehat{\mathbf{X}})$ same nenegativne člene, sledi

$$1 \geq R^2 = var(\widehat{\mathcal{N}(\mathbf{US})}) \geq \alpha^2 + \sum_{i=1}^{n} w_i^2$$

in zato napovedan rezultat za uteži funkcije učinkovitosti:

$$|\alpha| \leq 1$$

$$|w_i| \leq 1 \text{ za } i = 1, \ldots, n$$

Velja, da je koren koeficienta determinacije R enak korelaciji spremenljivke $\mathbf{X}$ z njeno oceno $\widehat{\mathbf{X}}$ (Seber, 1977):

$$R = corr(\mathbf{X}, \widehat{\mathbf{X}})$$

Če upoštevamo običajno pojmovanje visoke koreliranosti, tj. korelacijski koeficient, ki je po absolutni vrednosti večji od 0.7, to pomeni, da lahko šele pri $R^2 \geq 0.5$ govorimo o zadovoljivi natančnosti MLR modela.

3.2. Regresijske predpostavke

Uporaba multiple linearne regresije pri reševanju predoločenega linearnega sistema zahteva izpolnitev naslednjih pogojev (Johnson in Wichern, 2002):

1. LINEARNOST SPREMENLJIVK: Obstajati mora približno linearna zveza med odvisno spremenljivko $\mathbf{X}$ na eni strani in neodvisnimi spremenljivkami $\mathbf{X_1}, \ldots, \mathbf{X_n}$ na drugi strani, tj. pričakovana vrednost oz. matematično upanje odvisne spremenljivke mora biti linearna funkcija neodvisnih spremenljivk. Indikator linearnosti med odvisno spremenljivko in neodvisnimi spremenljivkami modela je velik koeficient determinacije R^2. 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 (Walker, 2000; Larsen, 2003), pogosto precej nižje (Walker et al., 1997b; Walker et al., 1998; Walker et al., 1999), le redko pa presežejo vrednost 0.6 (Litman in Shimei, 2002).

2. NEODVISNOST SPREMENLJIVK: Noben par neodvisnih spremenljivk $\mathbf{X_1}, \ldots, \mathbf{X_n}$ ne sme biti preveč koreliran, tj. korelacijski koeficienti $corr(\mathbf{X_i}, \mathbf{X_j})$ morajo biti po absolutni vrednosti manjši od 0.7. Če to ni tako, je dobljen model lahko zelo občutljiv na majhne merske napake ali spremembe vrednosti neodvisnih spremenljivk. Temu pojavu rečemo *multikolinearnost*. Odvečne neodvisne spremenljivke je zato treba odstraniti iz MLR modela. Zaradi težnje k čim večji natančnosti modela je smiselno odstraniti tiste spremenljivke, ki so z odvisno spremenljivko v nižji korelaciji.

Za napake dobljenega modela napovedovanja odvisne spremenljivke pa velja oz. mora veljati še naslednje:

3. NEPOŠEVNOST NAPAK: Srednja vrednost napake ϵ je enaka 0. To je posledica metode najmanjših kvadratov, na kateri temelji linearna regresija.

4. HOMOSKEDASTIČNOST NAPAK: Varianca napake ϵ mora biti po celotni učni množici enaka. V nasprotnem primeru je korelacija med odvisno spremenljivko in parametri modela lahko zavajajoče povprečje vzorcev višje in nižje korelacije.

5. NORMALNOST NAPAK: Napaka ϵ mora biti normalno porazdeljena slučajna spremenljivka.

Zanimivo vrsto vzorcev predstavljajo t. i. *osamelci*. Tako imenujemo meritve, ki se nenavadno razlikujejo od velike večine ostalih meritev in zato nepredvidljivo vplivajo na natančnost modela (Tabachnick in Fidell, 1996). Odstranitev osamelcev iz učne množice MLR modela je eden od običajnih regresijskih postopkov.

3.3. Pomembnost izbire regresijskih parametrov

Ko izbiramo podmnožico parametrov oz. neodvisnih spremenljivk MLR modela, se zastavi vprašanje, zakaj ne bi vzeli vseh parametrov, ki jih lahko pridobimo. To se zdi smiselno predvsem zato, ker koeficient determinacije R^2 s številom parametrov narašča. Izkaže pa se, da je uporaba vseh parametrov lahko neprimerna iz več razlogov:

- ⇝ Pridobiti celotno množico parametrov je včasih težko, časovno zahtevno in/ali samodejno nemogoče.
- ⇝ Če se omejimo na manjštevilnejšo množico parametrov, lahko to včasih bolj natančno določimo.
- ⇝ Varčnost je pomembna lastnost dobrih modelov – modeli z manj parametri omogočajo boljši vpogled v odnose med regresijskimi spremenljivkami.
- ⇝ Izračuni regresijskih koeficientov so v modelih z veliko spremenljivkami zaradi multikolinearnosti pogosto nestabilni.
- ⇝ Pokazati se da, da lahko neodvisne spremenljivke, ki so z odvisno spremenljivko v zelo nizki korelaciji (po absolutni vrednosti pod 0.1), povečajo srednjo kvadratno napako. Če take spremenljivke iz modela odstranimo, zmanjšamo napako napovedi.

  Za preizkus hipoteze o nekoreliranosti neodvisne spremenljivke $\mathbf{X_i}$ z odvisno spremenljivko $\mathbf{X}$ lahko uporabimo testno statistiko, ki temelji na Studentovi porazdelitvi.
- ⇝ Pokazati se tudi da, da lahko neodvisne spremenljivke, ki imajo v MLR modelu majhne neničelne (regresijske) koeficiente oz. uteži, povečajo srednjo kvadratno napako. Če takšne spremenljivke iz modela odstranimo, zmanjšamo napako napovedi.

V statistiki obstaja več načinov izbire 'dobre' podmnožice MLR parametrov, od katerih ima vsak svoje prednosti in slabosti. Najpogosteje se uporabljajo: *sprednja izbira*, *vzvratna eliminacija* in *postopna regresija* (Seber, 1977).

3.4. Merjenje zadovoljstva uporabnikov

Hone in Graham (Hone in Graham, 2000) sta opozorila na dejstvo, da vprašalnik (Tabela 1), s katerim avtorice ogrodja PARADISE merijo zadovoljstvo uporabnikov, ne temelji niti na teoriji niti na ustreznih empiričnih raziskavah in da je seštevanje ocen, ki naj bi merile popolnoma različne količine, sporno. Vsota naj bi bila smiselna le, če vsa vprašanja merijo isto količino.

Da bi bila vsota ali celo povprečje ocen, ki se nanašajo na učinkovitost katerega izmed modulov sistema za dialog, popolnoma nesmiselna, ni čisto res. Na izbran modul lahko gledamo kot na merjeno količino. Res je sicer, da lahko opazujemo različne vidike obnašanja tega modula, vendar nas ponavadi ne zanimajo le izolirane lastnosti, temveč tudi uspešnost modula kot celote. S tem v zvezi menimo, da tudi seštevanje ocen, dodeljenih vprašanjem za določanje zadovoljstva uporabnikov z različnimi vidiki delovanja sistema za dialog, ni popolnoma nesmiselno. Res je, da metoda ni dodelana, je pa lahko dober kazalec učinkovitosti sistema za dialog.

Ker za nobeno od obstoječih tehnik merjenja zadovoljstva uporabnikov sistemov za dialog ni dokazano, da izpolnjuje pogoje za veljaven psihometrični instrument, je treba vse sklepe, ki zajemajo zadovoljstvo uporabnikov, obravnavati zelo previdno. Žal je bil prvi resen poskus razvoja vprašalnika, ki bi zanesljivo, veljavno, objektivno in diskriminativno meril zadovoljstvo uporabnikov sistemov za dialog, (začasno) prekinjen (Hone in Graham, 2000).

Če se pojavi želja po vrednotenju katerega od modulov danega sistema za dialog (npr. modula za vodenje dialoga ali modula za razpoznavanje govora), pa je bolj smiselno sešteti ocene, dodeljene le tistim vprašanjem, ki se nanašajo na učinkovitost oz. obnašanje izbranega modula (Hajdinjak, 2006b).

3.5. Vplivi razpoznavanja govora na rezultate vrednotenja

Parameter, ki ima v funkciji učinkovitosti zaradi (po absolutni vrednosti) največje uteži najpogosteje najpomembnejšo vlogo, je parameter, ki meri učinkovitost modula za razpoznavanje govora (Walker et al., 1997b; Walker et al., 1998; Litman in Shimei, 2002; Larsen, 2003). To je, kakovost razpoznavanja govora ključno vpliva na zadovoljstvo uporabnikov – ob povečani učinkovitosti razpoznavanja govora se poveča tudi zadovoljstvo uporabnikov.

Kaj pa, če nas zanima npr. učinkovitost modula za vodenje dialoga ali razumevanje naravnega jezika? Na osnovi rezultatov, ki jih podaja literatura, smo prišli do sklepa, da bo vrednotenje učinkovitosti posameznih modulov zelo verjetno zanesljivejše in natančnejše, če odstranimo vpliv razpoznavanja govora, torej simuliramo tako rekoč popolno razpoznavanje. To lahko naredimo tako, da za pridobivanje regresijskih podatkov uporabimo eksperiment Čarovnik iz Oza (Hajdinjak in Mihelič, 2004), v katerem vlogo razpoznavalnika govora ali celo vlogo modulov za razumevanje govora prevzame človek. Ugotovili smo (Hajdinjak, 2006b), da pridejo v tem primeru v ospredje tudi tisti parametri modela učinkovitosti, ki jim zaradi izjemnega vpliva učinkovitosti razpoznavanja govora svoje vloge v preteklih študijah ni uspelo dokazati. Trdimo, da tako dobljene uteži funkcije učinkovitosti realneje izražajo vpliv parametrov na zadovoljstvo uporabnikov.

V skladu z našimi sklepi so tudi ugotovitve, do katerih so prišle Walker, Boland in Kamm (Walker et al., 1999). Ugotovile so, da se značilnosti in uteži parametrov modela učinkovitosti lahko spremenijo, če izboljšamo razpoznavanje govora.

4. Sklep

Podrobno smo preučili ogrodje PARADISE, ki velja za potencialno splošno metodologijo vrednotenja učinkovitosti sistemov za dialog. Opozorili smo na nekatere pomanjkljivosti in omejitve te metode ter predlagali morebitne rešitve (Hajdinjak in Mihelič, 2006a).

Prvič, opozorili smo na dejstvo, da je treba, če se želimo izogniti prevelikim napakam ocen, normalizirati tudi odvisno spremenljivko funkcije učinkovitosti, ki izraža zadovoljstvo uporabnikov, ter napovedano normalizirano vrednost zadovoljstva uporabnika transformirati nazaj na začetni interval.

Drugič, poudarili smo, da vprašalnik, s katerim avtorice ogrodja PARADISE merijo zadovoljstvo uporabnikov, ne temelji niti na teoriji niti na ustreznih empiričnih raziskavah, in zato ne more šteti za veljaven psihometrični instrument.

Tretjič, prišli smo do sklepa, da bo vrednotenje učinkovitosti posameznih modulov zelo verjetno zanesljivejše in natančnejše, če odstranimo vpliv razpoznavanja govora, torej simuliramo tako rekoč popolno razpoznavanje. Omenili smo, da v tem primeru pridejo v ospredje tudi tisti parametri modela učinkovitosti, ki jim zaradi izjemnega vpliva učinkovitosti razpoznavanja govora svoje vloge v preteklih študijah ni uspelo dokazati, in trdili, da tako dobljene uteži funkcije učinkovitosti realneje izražajo vpliv parametrov na zadovoljstvo uporabnikov.

5. Literatura

M. Bates in D. Ayuso. 1991. A proposal for incremental dialogue evaluation. V: *Proceedings of DARPA Speech and Natural Language Workshop*, str. 319–322. Pacific Grove, ZDA.

J. C. Carletta. 1996. Assessing the reliability of subjective codings. *Computational Linguistics*, 22(2):249–254.

M. Danieli in E. Gerbino. 1995. Metrics for evaluating dialogue strategies in a spoken language system. V: *Proceedings of the 1995 AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation*, str. 34–39. Stanford, ZDA.

B. Di Eugenio in M. Glass. 2004. The kappa statistic: A second look. *Computational Linguistics*, 30(1):95–101.

H. Grice. 1975. *Logic and Conversation (Syntax and Semantics, Speech Acts, Vol. 3)*. Academic Press, New York.

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.

K. S. Hone in R. Graham. 2000. Towards a tool for the subjective assessment of speech system interfaces (sassi). *Natural Language Engineering*, 6(3/4):287–303.

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

C. A. Kamm, D. J. Litman, in M A. Walker. 1998. From novice to expert: The effect of tutorials on user expertise with spoken dialogue systems. V: *Proceedings of the 5th International Conference on Spoken Language Processing*, str. 1211–1214. Rundle Mall, Avstralija.

L. B. Larsen. 2003. Issues in the evaluation of spoken dialogue systems using objective and subjective measures. V: *Proceedings of the Automatic Speech Recognition and Understanding Workshop*, str. 209–214. St. Thomas, ZDA.

D. J. Litman in P. Shimei. 2002. Designing and evaluating an adaptive spoken dialogue system. *User Modeling and User-Adapted Interaction*, 12.

J. Polifroni, L. Hirschman, S. Seneff, in V. Zue. 1992. Experiments in evaluating interactive spoken language systems. V: *Proceedings of DARPA Speech and Natural Language Workshop*, str. 28–33. Harriman, ZDA.

P. Price, L. Hirschman, E. Shriberg, in E. Wade. 1992. Subject-based evaluation measures for interactive spoken language systems. V: *Proceedings of the DARPA Speech and Natural Language Workshop*, str. 34–39. Harriman, ZDA.

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

E. Shriberg, E. Wade, in P. Price. 1992. Human-machine problem solving using spoken language systems (sls): Factors affecting performance and user satisfaction. V: *Proceedings of the DARPA Speech and Natural Language Workshop*, str. 49–54. Harriman, ZDA.

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

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, J. Boland, in C. Kamm. 1999. The utility of elapsed time as a usability metric for spoken dialogue systems. V: *Proceedings of the Automatic Speech Recognition and Understanding Workshop*, str. 317–320. Keystone, ZDA.

M. A. Walker. 2000. An application of reinforcement learning to dialogue strategy selection in a spoken dialogue system for email. *Journal of Artificial Intelligence Research*, 12:387–416.