Semantična analiza vremenskih napovedi

Melita Hajdinjak, France Mihelič

Fakulteta za elektrotehniko
Laboratorij za umetno zaznavanje, sisteme in kibernetiko
Tržaška 25, 1000 Ljubljana, Slovenija
melitah@luks.fe.uni-lj.si, mihelicf@luks.fe.uni-lj.si

Povzetek

Predstavljen je semantični analizator za tematsko omejeno področje *vremenske napovedi in obeti za Slovenijo*. S pomočjo tega analizatorja gradimo del podatkovne baze sistema za podajanje vremenskih napovedi. Del podatkov o vremenu, ki so podani v tekstovni obliki, analizator pomensko analizira in jih zapiše v računalniško berljivi strukturi.

1. Uvod

V Laboratoriju za umetno zaznavanje, sisteme in kibernetiko smo si zastavili nalogo izdelati sistem za dialog za podajanje vremenskih napovedi. Po zgledu podobnih sistemov, kot je *Jupiter* (Zue et al., 2000), so vir podatkov za podatkovno bazo, ki bo sestavni del sistema za podajanje vremenskih napovedi, internetne strani *Hidrometeorološkega zavoda Agencije RS za okolje*. Del podatkov o vremenu je na internetnih straneh podan v tekstovni obliki. Da bo sistem za podajanje informacij sposoben jedrnato odgovarjati na vprašanja o vremenu, mora to besedilo razumeti.

Kaj pomeni, da računalnik razume naravni jezik, kot je slovenščina, angleščina, nemščina,...? S pojmom *razumeti* ne mislimo, da bo računalnik razmišljal tako kot razmišlja človek. Mislimo le, da bo informacije pravilno uporabljal oziroma se pravilno odzival.

Tekstovno podani podatki pogosto vsebujejo veliko podrobnejše informacije kot pa uporabnika zanimajo. Da bo sistem za podajanje informacij sposoben ponuditi le zahtevano informacijo in uporabnika ne bo dolgočasil z dolgimi monologi, mora biti podatkovna baza sistema ustrezno zgrajena. Prav zaradi tega smo se odločili, da zgradimo semantični analizator, ki bo tekstovno podane informacije pomensko analiziral in jih zapisal v novi, računalniško berljivi strukturi.

Nekoliko drugačno strukturo podatkovne baze ima sistem *Jupiter*, katerega sestavni del je semantični analizator *Tina* (Seneff, 1992). *Tina* najprej vsaki povedi vremenske napovedi priredi ustrezne semantične kategorije ter zaporedni indeks. Podatkovno bazo sistema *Jupiter* sestavljajo semantičnim kategorijam prirejeni seznami indeksov tistih povedi, ki jim je analizator *Tina* priredil določeno semantično kategorijo. Na nivoju gradnje podatkovne baze se s tem semantična analiza zaključi. Šele, ko uporabnik sistemu *Jupiter* zastavi vprašanje, se spet vključi semantični analizator *Tina*, ki s pomočjo ustreznih povedi in njim prirejenih semantičnih kategorij sestavi odgovor na uporabnikovo vprašanje.

2. Semantična analiza

Osnova semantičnega analizatorja za tematsko omejeno področje *vremenska napoved in obeti za Slovenijo* je slovar, ki je rezultat 3-mesečnega snemanja podatkov na internetnih straneh.

2.1. Primer vremenske napovedi

Tematsko omejeno komunikacijsko področje vremenskih napovedi je zelo specifično. Podan je tipičen primer vremenske napovedi za Slovenijo.

Napoved za Slovenijo
Danes bo sončno, burja na Primorskem bo ponehala. Najvišje dnevne temperature bodo od 20 do 26, na Primorskem do 29 stopinj C. Tudi jutri bo sončno, popoldne bo predvsem na zahodu spremenljive oblačnosti več. Proti večeru ter ponoči lahko nastanejo posamezne nevihte. Najnižje jutranje temperature bodo od 8 do 15, najvišje dnevne od 22 do 29 stopinj C.

Obeti
V nedeljo bo sprva še delno do spremenljivo oblačno, občasno so še možne manjše krajevne padavine, popoldne pa se bo postopno zjasnilo. V ponedeljek bo spet sončno in še topleje.

2.2. Slovar

Slovar vsebuje vse besede in to v vseh oblikah (spolih, sklonih in številih), v katerih se v bazi pojavljajo. Vseh vnosov v slovar je več kot 600. Tabela 1 vsebuje primere vnosov v slovar.

Besede z istim pomenom smo združili v isto *semantično kategorijo*. Besede, ki ne nosijo pomembne semantične informacije, pa smo že na nivoju slovarja izločili. Semantične kategorije smo naprej združevali še v *semantične tipe*.

Beseda	Kategorija	Tip
burjo	burja	veter
države	slovenija	pokrajina
jutri	jutri	rel_dan
kakšna	x	ignore
toplejše	topleje	temperatura

Tabela 1: Primeri vnosov v slovar.

Besede iz tematsko omejenega področja *vremenska napoved in obeti za Slovenijo* lahko predstavimo s 15 semantičnimi tipi. Te semantične tipe lahko razdelimo na *časovne* semantične tipe, na *krajevne* semantične, na semantične tipe s *ključno informacijo*, na *dopolnilne*

semantične tipe ter na semantične tipe *relativni izraz*, *ključna beseda* in *nepomembna informacija*. Razdelitev semantičnih tipov je prikazana v tabeli 2.

ČASOVNI SEMANTIČNI TIPI
relativni časovni izraz (rel_dan): *danes, jutri,...*
dan v tednu (dan_teden): *ponedeljek, torek,...*
del dneva (del_dneva): *podnevi, zjutraj,...*
KRAJEVNI SEMANTIČNI TIPI
pokrajina (pokrajina): *dolenjska, gorenjska,...*
lege (lege): *gore, morje,...*
stran neba (stran_neba): *sever, jug,...*
SEMANTIČNI TIPI S KLJUČNO INFORMACIJO
vreme (vreme): *sončno, nestalno,...*
veter (veter): *burja, severovzhodnik,...*
temperatura (temperatura): *vroče, topleje,...*
DOPOLNILNI SEMANTIČNI TIPI
lastnost (lastnost): *zmerno, pretežno,...*
stanje (stanje): *razjasnilo se bo, bodo ponehale,...*
število (stevilo): *0, 1, 2, ... 50.*
OSTALI SEMANTIČNI TIPI
relativni izraz (rel_izraz): *iznad, na, od, do,...*
ključna beseda (keyword): *stopinj, kmh, vreme,...*
nepomembna informacija (ignore): *bo, tako,...*

Tabela 2: Semantični tipi.

Semantični tip *ključna beseda* je pomemben predvsem zato, ker poleg semantičnih tipov s ključno informacijo najbolj zanesljivo določa pomen povedi. Semantični tip *nepomembna informacija* združuje kategorije, ki za razumevanje besedila niso ključnega pomena in jih lahko izločimo že na nivoju slovarja.

2.3. Semantični analizator

Na podlagi slovarja za tematsko omejenega področja *vremenske napovedi in obeti za Slovenijo* smo izdelali semantični analizator, ki vremensko napoved pomensko analizira. Vhod v analizator je seznam *besedilnih delov*, ki so v besedilu ločeni s pikami ali vejicami. Del tega vhodnega seznama je tako lahko po eni strani samo stavčni člen, po drugi strani pa tudi priredno zložena poved.

Naš cilj je posameznemu besedilnemu delu določiti pomen in pri vremenskih napovedih se izkaže, da lahko besedilne dele pomensko razdelimo na *vreme, veter* in *temperaturo*. V podatkovno bazo pa zapisujemo podatke o vremenu, vetru in temperaturi za določene kraje (pokrajine, deli Slovenije, lege) ob določenem času (dan, del dneva). Pomen posameznih besedilnih delov semantični analizator določi s pomočjo vnaprej definiranih predlog, to je točkovanih seznamov semantičnih kategorij in tipov. Podoben pristop je bil uporabljen pri pomenski analizi stavkov v sistemu za podajanje informacij o letalskih poletih (Pepelnjak, 1996; Pepelnjak et al., 1996). Vrednost posamezne predloge predstavlja seštevek točk, ki so pripisane posameznim semantičnim kategorijam in tipom besed, ki besedilni del sestavljajo. Zaradi pomena besedilnih delov, ki sestavljajo vremenske napovedi, smo definirali tri osnovne predloge: *VREME, VETER* in *TEMPERATURA*. Besedilne dele, ki ne nosijo pomembne informacije pa pripišemo dodatni četri predlogi *NEPOMEMBNA INFORMACIJA*.

S pomočjo predlog lahko besedilne dele dodatno analiziramo. V bazi vremenskih napovedi besedilni deli vsebujejo največ dva različna pomena. Ravno take besedilne dele, ki vsebujejo dva različna pomena, želimo razdeliti. Za predstavljen semantični analizator je namreč pomembno, da analizira besedilne dele, ki vsebujejo samo informacije o vremenu, samo informacije o vetru ali pa samo informacije o temperaturi. Edini priredni veznik, ki se je pojavil v bazi vseh povedi in pred katerim v povedi ni vejice, je veznik *in*. Tako priredno zloženo poved lahko testiramo glede na pomena, ki ju določata oba priredno zložena stavka. Če ugotovimo, da oba vsebujeta pomembno informacijo in imata različna pomena glede na predloge, ju lahko ločimo.

Semantični analizator vremenskih napovedi izvede analizo v šestih korakih :

1. korak : PRIREJANJE SEMANTIČNIH KATEGORIJ IN TIPOV. Vsakemu besedilnemu delu iz seznama na vhodu priredi ustrezne semantične kategorije in semantične tipe. To naredi na osnovi slovarja, v katerem sta dodatna informacija k vsaki besedi tudi njena semantična kategorija in semantični tip. Vzemimo na primer stavek *Jutri bo na Primorskem sončno in toplo.* Analizator priredi temu stavku naslednje semantične kategorije : [jutri], [primorska], [sončno], [toplo].

2. korak : OBDELAVA SEMANTIČNIH KATEGORIJ IN TIPOV. V drugem koraku semantični analizator nekatera zaporedja semantičnih kategorij in tipov združi, nekatera zaporedja pa preoblikuje. Ta dodatna obdelava je potrebna predvsem zaradi *relativnih izrazov*. Besedne oblike iz semantične kategorije *relativni izraz* imajo v kontekstu namreč lahko različne pomene. Obdelava kategorij in tipov pa je potrebna tudi v nekaterih primerih, ko v besedilnem delu *relativni izraz* ne nastopa. V tabeli 3 so primeri takšne obdelave semantičnih kategorij in tipov.

PREDLOGA VREME:		
tip	vreme	20 točk
	stanje	15 točk
	lastnost	10 točk
kategorija	vreme	15 točk
	prevladovati	10 točk
	zajeti	10 točk
	razširiti	10 točk
	sijati	5 točk
	količina	5 točk
	padati	5 točk
	širiti	5 točk

Slika 1: Primer predloge *VREME*.

3. korak : UJEMANJE BESEDILNIH DELOV S PREDLOGAMI. V tretjem koraku semantični analizator vsakemu vhodnemu besedilnemu delu priredi ustrezno predlogo in mu s tem določi tudi pomen. Vnaprej so definirane tri

Besedna zveza	Začetni okvirji [kategorija / tip]	Obdelani okvirji [kategorija / tip]
v večjem delu Slovenije	[večji/rel_izraz] [slovenija/pokrajina]	[slovenija/pokrajina]
večji del dneva	[večji/rel_izraz] [podnevi/del_dneva]	[podnevi/del_dneva]
do jutra	[do/rel_izraz] [zjutraj/del_dneva]	[ponoči/del_dneva]
v noči na torek	[ponoči/del_dneva] [na/rel_izraz] [torek/dan_teden]	[ponedeljek/dan_teden] [ponoči/del_dneva]
v prvi polovici dneva	[prvi/rel_izraz] [podnevi/del_dneva]	[dopoldan/del_dneva]
v drugem delu noči	[drugi/rel_izraz] [ponoči/del_dneva]	[ponoči/del_dneva]
v severnih krajih	[sever/stran_neba] [slovenija/pokrajina]	[sever/stran_neba]
na jugu Štajerske	[na/rel_izraz] [jug/stran_neba] [štajerska/pokrajina]	[štajerska/pokrajina]
vzhodni veter	[vzhod/stran_neba] [veter/keyword]	[vzhodni veter/veter]

Tabela 3: Primeri obdelave semantičnih kategorij in tipov.

predloge : *VREME, VETER* in *TEMPERATURA*. Točke kategorij in tipov, ki predlogo določajo, smo določili eksperimentalno. Dodatna četrta predloga *NEPOMEMBNA INFORMACIJA* omogoča, da besedilne dele, ki ne nosijo za podatkovno bazo pomembne informacije, izpustimo. Analizator besedilni del pripiše predlogi *nepomembna informacija*, če največje število točk, ki ga besedilni del dobi glede na tri osnovne predloge, ne doseže vnaprej določenega praga. Ta prag smo določili na osnovi baze vremenskih napovedi, ki jih črpamo z internetnih strani. Primer predloge je na sliki 1.

4. korak : DODATNO RAZCEPLJANJE BESEDILNIH DELOV. V četrtem koraku poteka analiza besedilnih delov, ki vsebujejo veznik *in*. Tak besedilni del analizator razdeli na dva dela, razpolovi ga na mestu veznika *in*. Sedaj obema deloma določi pomen. Če se ujemata z različnima osnovnima predlogama, poteka nadaljna analiza na vsakem delu posebej. Primeri besedilnih delov, ki jih analizator na tej stopnji razcepi, so naslednji:

Jutri bo sončno in toplo.
V soboto in nedeljo bo sončno in poletno vroče.
Nadaljevalo se bo sončno in vroče poletno vreme.

Analizator pa ne razcepi naslednjih besedilnih delov:

Danes in jutri bo pretežno jasno.
Sredi dneva in popoldne bodo pogoste krajevne plohe.
V sredo bo povsod delno jasno in suho.

5. korak : GENERIRANJE IZHODNIH PREDLOG. V petem koraku semantični analizator generira izhodne predloge, to je sezname časovnih in krajevnih ter za pomen besedilnega dela ključnih informacij. Ključna informacija v izhodni predlogi besedilnega dela, ki ga analizator pripiše predlogi *vreme*, je na primer *vreme*. Izhodne predloge so tri : *VREME, VETER* in *TEMPERATURA*. V tabeli 4 sta podana dva primera izhodnih predlog. Iz drugega primera izhodne predloge vidimo, da nekateri besedilni deli ne nosijo vse informacije, ki jo potrebujemo za zapis v podatkovno bazo. V primeru manjka krajevna informacija, časovna pa ni popolna. Prav zaradi tega je potreben še en korak do končne izhodne predloge.

6. korak : DOPOLNJEVANJE IZHODNIH PREDLOG. V zadnjem, šestem koraku, analizator nepopolne izhodne predloge dopolni. Manjkajoče informacije poišče v prejšnjih predlogah. Prvi besedilni del v seznamu vhodnih besedilnih delov ima vnaprej določeno časovno in krajevno informacijo, namreč *Slovenija - danes podnevi*, če seveda sam besedilni del ne vsebuje drugačnih krajevnih ali časovnih informacij. Vsak naslednji besedilni del, katerega začetna predloga ni popolna, deduje podatke od predloge prejšnjega besedilnega dela. Nedoločene krajevne in časovne informacije, kot so *drugod*, *nato*, *tam*, pa analizator v tem koraku s pomočjo prejšnjih predlog natanko določi. Natanko določi pomeni, da na primer besedo *drugod* nadomesti z ustreznimi *pokrajinami* ali ustreznimi *legami* ali ustreznimi *stranmi neba*, odvisno pač od krajevne informacije v prejšnji predlogi. Besedo *nato* pa nadomesti z naslednjo časovno enoto, to je naslednjim dnem ali naslednjim delom dneva.

VREME		VETER	
čas:	*sobota ponoči, nedelja zjutraj*	čas:	*popoldne*
kraj:	*slovenija*	kraj:	
vreme:	*dež*	veter:	*severni veter*

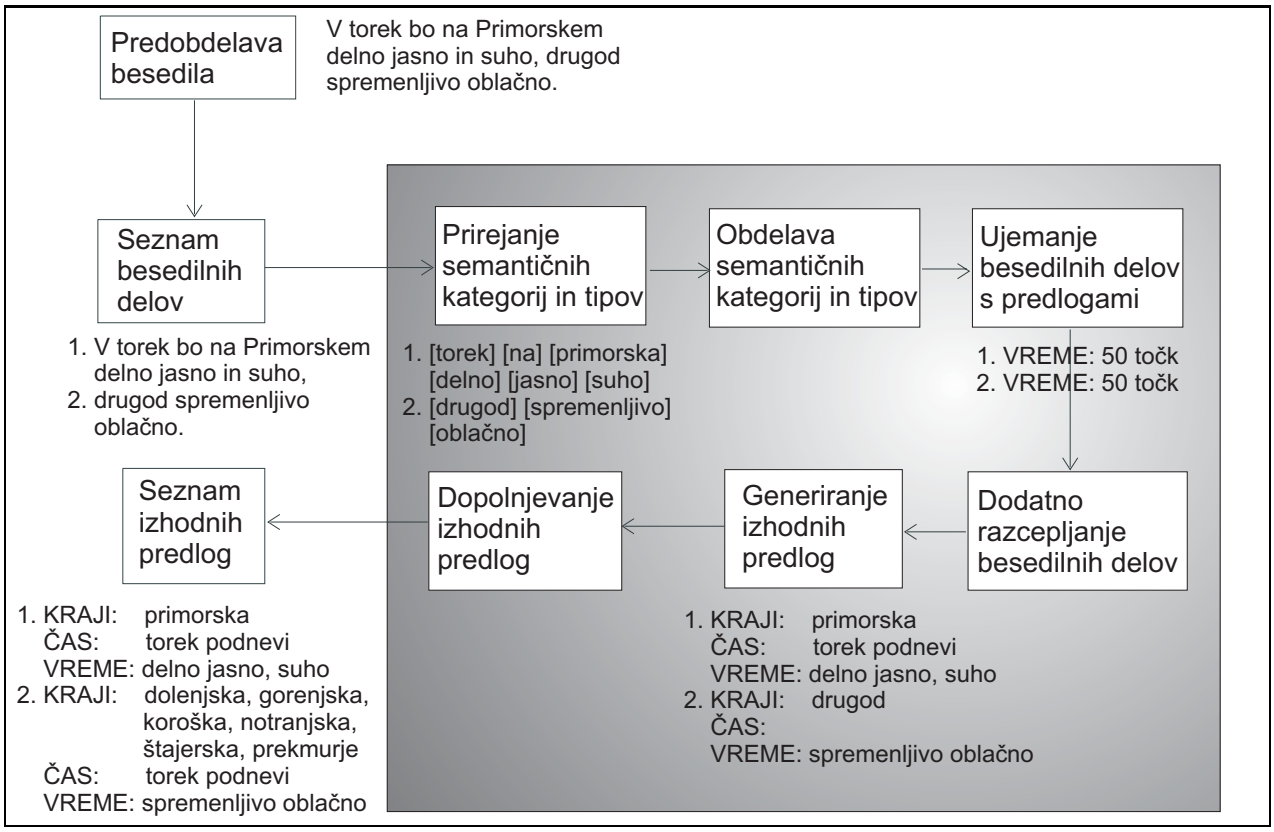
Tabela 4: Primer nedopolnjenih izhodnih predlog.

Šele sedaj je besedilni del z izhodno predlogo tako opisan, da ga lahko zapišemo v podatkovno bazo. Čas je natanko določen z dnevom in delom dneva, kraj pa je natanko določen s pokrajino, lego ali stranjo neba. Besedilni deli s pomenom *vreme* vsebujejo še opis vremena, besedilni deli s pomenom *veter* še opis vetra, besedilni deli s pomenom *temperatura* pa še opis temperature.

3. Primer analize

Poglejmo, kako semantični analizator vremenskih napovedi analizira poved *V torek bo na Primorskem delno jasno in suho, drugod spremenljivo oblačno.* Potek analize je prikazan na sliki 2.

Po predobdelavi dane povedi dobi semantični analizator na vhodu seznam dveh besedilnih delov. Analizator pripiše obema besedilnima deloma pomen *vreme*. Prvi besedilni del (*V torek bo na Primorskem delno jasno in suho*) analizator dodatno ne razcepi, čeprav vsebuje veznik *in*. To pa zato, ker obema deloma pripiše enak pomen, namreč pomen *vreme*. Ta besedilni del vsebuje vse potrebne podatke za napolnitev predloge (časovno in krajevno informacijo ter informacijo o vremenu), zato pred-



Slika 2: Model semantičnega analizatorja.

loge ni potrebno dodatno dopolnjevati. V drugem besedilnem delu (*drugod spremenljivo oblačno*) manjka časovna informacija, krajevna informacija pa je nedoločena. Zaradi tega je dopolnjevanje predloge tega besedilnega dela potrebno. Časovno informacijo analizator prepiše iz predloge prejšnjega besedilnega dela, krajevno pa določi na podlagi krajevne informacije v prejšnji predlogi. Določiti mora kategorijo *drugod*, zato najprej ugotovi, da je krajevna informacija v prejšnji predlogi podana s semantičnim tipom *pokrajina*. Zaradi tega v krajevno informacijo predloge zapiše vse pokrajine, ki jih prejšnja predloga ne vsebuje.

4. Zaključek

Predstavljen semantični analizator vremenskih napovedi je trenutno še v fazi testiranja in izboljševanja. Za uspešno analizo vremenskih napovedi skozi vse leto je pomembno, da bo baza *Vremenske napovedi in obeti za Slovenijo*, na podlagi katere gradimo semantični analizator, pokrivala vse letne čase. Nadaljnje zbiranje vremenskih napovedi z internetnih strani je zaradi tega neizogibno. Naslednji korak bo rezultate analizatorja zapisati v podatkovno bazo, ki bo vir informacij sistema za podajanje vremenskih napovedi. Ko bo podatkovna baza zgrajena, se bomo lotili semantične analize uporabnikovih izjav, začeli pa bomo tudi z zbiranjem podatkov za postavitev strategije vodenja dialoga. Načrtujemo, da bo to zbiranje potekalo s pomočjo *Wizard-of-Oz* tehnike (Dahlbäck et al., 1993) oziroma eksperimenta, ki ga imenujemo *Čarovnik iz Oza*. V teh eksperimentih človek misli, da se pogovarja s strojem, v resnici pa za računalnikom sedi človek - *čarovnik* ali *wizard*, katerega odgovori so posredovani s sintetiziranim govorom. Tako lahko že pred konstrukcijo sistema simuliramo dialog med človekom in računalnikom. To tehniko pa lahko uporabimo tudi za izboljševanje in ocenjevanje sistema za dialog.

5. Viri

Dahlbäck, N. & Jönsson, A. & Ahrenberg, L. (1993) Wizard of Oz studies: why and how. *Proceedings of the international workshop on Intelligent user interfaces*, USA, 193–200.

Pepelnjak, K. (1996). *Pomenska analiza stavkov v sistemu za razumevanje tekočega govora.* magistrsko delo, Fakulteta za elektrotehniko, Univerza v Ljubljani.

Pepelnjak, K. & Mihelič, F. & Pavešić, N. (1996) Semantic decomposition of sentences in the system supporting flight services. *CIT. J. Comput. Inf. Technol.*, 4 (1), 17–24.

Seneff, S. (1992). TINA: A Natural Language System for Spoken Language Applications. *Computational Linguistics*, 18 (1), 61–86.

Zue, V. & Seneff, S. & Glass, J. & Polifroni, J. & Pao, C. & Hazen, T.J. & Hetherington, L. (2000). JUPITER: A Telephone-Based Conversational Interface for Weather Information. *IEEE Transactions on Speech and Audio Processing*, 8 (1), 85–96.