JEZIKOVNI VIRI ZA STROJNO SIMULTANO PREVAJANJE GOVORA

mag. Darinka Verdonik

Povzetek

V članku razpravljamo o jezikovnih virih, ki jih je potrebno zagotoviti za razvoj sistemov strojnega simultanega prevajanja govora. Ugotavljamo, da se zadnji čas raziskovalci največ ukvarjajo s statističnimi pristopi k strojnemu simultanemu prevajanju govora, zato se osredotočamo na jezikovne vire, ki so potrebni za tak pristop. Ker so sistemi strojnega simultanega prevajanja govora sestavljeni iz modulov za razpoznavo govora, govorno orientirano prevajanje in sintezo govora, lahko v osnovi ločimo jezikovne vire, ki so potrebni za izboljšanje razpoznavanja in sinteze govora, ter vire, ki so potrebni za govorno orientirano prevajanje. Ugotavljamo, kateri takšni večji viri že obstajajo za tuje jezike in za slovenski jezik, njihove pomanjkljivosti, na koncu pa predstavimo jezikovne vire, ki se gradijo v projektu LC-STAR, in njihovo primerjavo z ostalimi podobnimi viri.

Ključne besede: jezikovne tehnologije, strojno simultano prevajanje govora, jezikovni viri, LC-STAR.

Language Resources for Speech-to-Speech Translation

Abstract

The article represents language resources needed for developing speech-to-speech translation systems. Lately most of researchers try to develop statistical approaches to machine translation, therefore we concentrate on language resources needed for statistical machine translation. Since speech-to-speech machine translation systems are composed from three modules: speech recognition, speech centered translation, speech synthesis, we can differ language resources needed to improve speech recognition and synthesis, and language resources needed for speech centered translation. Article gives overview of such most important language resources for foreign languages and for Slovenian language, shows their weaknesses and at the end represents language resources for Slovenian language, which are being built within the LC-STAR project, and their comparison to other similar resources.

1 UVOD

Ideja, kako prikladno bi bilo, če bi imeli prevajalni stroj, je zelo stara, prvi koraki v tej smeri pa so bili narejeni v letih po drugi svetovni vojni. Tako je takrat Američan Warren Weaver zapisal: "Pred seboj imam besedilo v ruščini, vendar se bom pretvarjal, da je v resnici zapisano v angleščini in zakodirano s čudnimi simboli. Vse, kar moram narediti, je razbiti kodo, da dobim informacijo, ki jo vsebuje besedilo." (Arnold et al., 1994) Takšen pogled na strojno prevajanje in razlike med jeziki je seveda zelo preprost, vendar je Weaver s tem spodbudil raziskave na tem področju in leta 1954 je bila demonstracija prototipa angleško-ruskega sistema strojnega prevajanja pisanega besedila (v nadaljevanju samo besedila v nasprotju s prevajanjem govora).

Začetnega optimizma glede strojnega prevajanja besedila je bilo konec s poročilom ALPAC-a (Automatic Language Processing Advisory Committee) leta 1966, ki je ugotavljalo, da to področje ni perspektivno in da zahteva preveč stroškov glede na končno doseženo kvaliteto produkta. Posledica je bila, da ameriška vlada ni bila več pripravljena financirati raziskav s tega področja, delo so nadaljevale le redke skupine zunaj ZDA.

V sedemdesetih so se vendarle zgodili nekateri pomembni premiki: zgrajena sta bila Systran za prevajanje besedil med ruščino in angleščino (za potrebe ameriškega letalstva) ter Meteo za prevajanje vremenskih napovedi. V Evropi so naredili angleško-francosko verzijo Systrana.

Pravo prebujenje tehnologije strojnega prevajanja besedila pa se je zgodilo v osemdesetih. Pomembnejši projekti na tem področju so bili evropski Eurotra, na Japonskem Mu, v ZDA pa t.i. Knowledge-Based Machine Translation. Nastajati so začeli tudi nekateri komercialni sistemi.

V poznih osemdesetih in začetku devetdesetih se je za področje strojnega prevajanja besedila začelo zanimati veliko podjetij, med pristopi se razvija statistično strojno prevajanje besedil. V tem času pa se začne tudi zanimanje za strojno simultano prevajanje govora, ki je veliko zahtevnejše in se, kot bomo videli v nadaljevanju, bistveno razlikuje od strojnega prevajanja besedil.

V poznih devetdesetih lahko opazujemo strojno prevajanje besedil na internetu, širjenje uporabe raznih elektronskih pripomočkov za prevajanje besedil (tudi pri nas, glej Hirci, 2003), med pristopi se začnejo razvijati na primerih temelječi sistemi strojnega prevajanja besedila (ang. example-based machine translation). V letu 2002 dobimo prvi večji strojni prevajalnik besedil tudi za slovenščino: podjetje Amebis predstavi slovensko-angleški prevajalni sistem Presis (Romih, Holozan 2002), poskuse statističnega strojnega prevajanja besedil prav tako iz slovenščine v angleščino delajo tudi na Fakulteti za računalništvo in informatiko v Ljubljani in Inštitutu Jožef Stefan (Vičič, Erjavec 2002).

V poznih devetdesetih pa se povečuje tudi zanimanje za strojno simultano prevajanje govora. Zasledimo nekaj večjih mednarodnih projektov strojnega simultanega prevajanja govora z različnimi pristopi: Janus, sistem z interlingvo (Waibel, Lavie, Levin, 1997), EuTrans, na primerih temelječ sistem (Internet), Nespole!, sistem z interlingvo (Metze et al., 2002), Verbmobile, uporablja različne metode, statistične in pravila (Ćavar, Menzel, 1998). V Sloveniji se v nekaterih znanstvenoraziskovalnih centrih ukvarjajo predvsem z razvojem razpoznave in sinteze slovenskega govora (sinteza: Vesnicer, Mihelič, Pavešić 2002; Šef, Gams, Škrjanc 2002; Rojc 2003, razpoznava: Rotovnik, Sepesy Maučec, Horvat, 2002; Kaiser et al., 2000; Sket, Imperl 2002; Pozne, Pavešić, Mihelič 2002), o razvoju govorno orientiranega strojnega prevajanja ne zasledimo nobene objave. Zaključen sistem strojnega simultanega prevajanja govora za slovenski jezik tako še ni bil predstavljen.

Ustrezni jezikovni viri so osnova za izdelavo sistemov strojnega simultanega prevajanja govora, vendar je najprej potreben temeljit premislek o vrsti in velikosti teh virov, o podatkih, ki naj jih vsebujejo, o zagotavljanju njihove kvalitete, o njihovi usklajenosti s podobnimi viri za tuje jezike. Članek najprej opozarja na razliko med sistemi strojnega prevajanja besedila in sistemi strojnega simultanega prevajanja govora ter na kratko predstavlja pristope k strojnemu simultanemu prevajanju

govora; v osrednjem poglavju (3) so specificirani jezikovni viri, ki so potrebni za razvoj statističnih sistemov strojnega simultanega prevajanja govora, pri tem se osredotočamo zlasti na tiste vire, ki so v svetu in pri nas pomanjkljivo zastopani. V poglavju 4 predstavimo jezikovne vire za slovenski jezik, ki smo jih poleti 2003 začeli graditi na Fakulteti za elektrotehniko, računalništvo in informatiko v okviru projekta LC-STAR posebej za razvoj tehnologij strojnega simultanega prevajanja govora, in izpostavimo, v čem se razlikujejo od že obstoječih virov za slovenski jezik, namenjenih za sisteme procesiranja naravnega jezika (t.i. natural language processing). V poglavju 5 sledi zaključek.

2 SISTEMI STROJNEGA SIMULTANEGA PREVAJANJA GOVORA

2.1 Razlike med strojnim simultanim prevajanjem govora in strojnim prevajanjem besedila

Sistemi strojnega simultanega prevajanja govora se v marsikaterem pogledu razlikujejo od sistemov strojnega prevajanja pisanega besedila.

Prva pomembna razlika je, da mora sistem strojnega simultanega prevajanja govora najprej razpoznati govor (tj. znati prevesti zvok v takšno pisno obliko, kot to naredi človek), šele nato lahko sledi govorno orientirano prevajanje (ki pa je zaradi lastnosti govorjenega jezika, ki jih opisujemo v naslednjem odstavku, prav tako drugačna, zahtevnejša naloga kot strojno prevajanje besedila), potem pa je treba prevedeno besedilo ponovno pretvoriti v zvok. Sistemi strojnega simultanega prevajanja govora so torej sestavljeni iz treh osrednjih modulov: razpoznave, govorno orientiranega prevajanja, sinteze (Hoege, 2002). To pa pomeni veliko dodatno oviro za uspešnost strojnega simultanega prevajanja govora, saj je povsem natančna razpoznava zelo težavna, zaradi česar se lahko že na tej ravni vnesejo napake. Poleg tega so v besedilu z ločili podane nekatere informacije o skladnji in prozodiji, ki se pri razpoznavi govora izgubijo.

Razlike med govorno orientiranim strojnim prevajanjem in strojnim prevajanjem besedila pa so pogojene tudi z razlikami med pisnim in govorjenim jezikom. Primerjalne raziskave obeh (Wiebe idr. 1996) so pokazale, da v govoru ljudje posredujejo več informacij implicitno, kar se kaže v veliko večji pogostosti rabe zaimkov in nedokončanih stavkov. Prvo predstavlja problem v primerih, ko se v jezikih, med katerima prevajamo, zaimek ne ujema, npr. ang. "I saw a cat. ***It*** was crossing the street.", sln. "Videl sem mačko. Prečkal***a je*** cesto." (to je problem tudi pri strojnem prevajanju besedila). Poleg tega najdemo v govoru: napačne začetke, pomote pri pregibanju besed (npr. *vprašal sem vam namreč*), ponavljanje, izpuste, obotavljanje, mašila (npr. *eee, mhm*) (Wiebe et al., 1997; Kay, Gawron, Norvig, 1994). (Kay, Gawron, Norvig, 1994) navajajo, da je takih elementov pri pazljivem govoru povprečno 15 %, lahko pa tudi več kot polovica. Simultani prevajalci ne prevajajo vseh teh elementov dobesedno, ampak samo, če je to smiselno, in enako pričakujemo od strojnega simultanega prevajalnika govora. Zaradi vsega tega je naloga strojnega simultanega prevajanja govora bistveno zahtevnejša in tudi bistveno drugačna od naloge strojnega prevajanja besedila. Jezikovni viri, namenjeni za razvoj govorno orientiranega strojnega prevajanja, morajo zato izhajati iz govorjenega jezika in ne iz zapisanega besedila.

2.2 Osnovne strategije in pristopi

Pristopi k strojnemu simultanemu prevajanju govora so različni in zahtevajo tudi delno različne jezikovne vire za razvoj tehnologije. V tem poglavju na kratko predstavimo osnovne strategije in pristope.

Transferni sistemi (ang. trasfer systems): Čeprav so vsi prevajalniki na nek način transferni, se poimenovanje uporablja za jezikovno odvisne sisteme, pri katerih je rezultat analize abstraktna predstavitev (govorjenega) besedila v vhodnem jeziku, vnos za sintezo besedila pa je abstraktna predstavitev besedila v ciljnem jeziku. Naloga modula za transfer je, da abstraktno predstavitev besedila v enem jeziku prenese na abstraktno predstavitev besedila v drugem jeziku. Te predstavitve – abstraktna analiza, prenos, abstraktno generiranje – povezujejo različne module, zato jim pravimo tudi

vmesne predstavitve (interface representations). Nobena predstavitev pri transferni metodi ni jezikovno neodvisna.

Sistemi z interlingvo (ang. interlingua): Pri teh sistemih se vhodno besedilo, ki je bilo razpoznano iz govora, prepiše v interlingvo, ki zajema vse potrebne informacije. Glavna razlika med transfernimi sistemi in sistemi z interlingvo je, da se pri interlingvi besedilo v ciljnem jeziku sintetizira samo iz interlingve, brez da bi se gledalo nazaj vhodni jezik. Interlingva je torej predstavitev vhodnega besedila in hkrati osnova za tvorjenje besedila v ciljnem jeziku. Zares univerzalne interlingve zaenkrat še niso naredili, pač pa je vedno bolj ali manj omejena na jezike, ki so vključeni. Prednost pristopa z interlingvo je, da je za nov jezik treba dodati samo dva modula, dobimo pa več kombinacij (v vse vključene jezike in obratno), problem pa je seveda težavnost definiranja dobre interlingve, tudi za sorodne jezike.

Veliko transfernih sistemov in sistemov z interlingvo temelji na **jezikovnih pravilih** (ang. rule-based machine translation), ki predstavijo govorjeno besedilo na tako abstraktni ravni (oblikoslovno, skladenjsko, semantično), da je na podlagi tega mogoč prenos med različnimi jeziki. Stopnja abstrakcije je večja pri interlingvi.

Nekateri sistemi so pravila razširili zlasti na semantično in pragmatično vedenje o posameznem področju. Take sisteme imenujemo **na vedenju temelječi** (ang. knowledge-based machine translation).

V zadnjem času so predvsem zaradi večje uspešnosti (Ney, 2001) popularnejši empirični pristopi k strojnemu simultanemu prevajanju govora (podobni pristopi se uporabljajo tudi za strojno prevajanje besedila). Prvi taki so **na primerih temelječi** sistemi strojnega simultanega prevajanja govora (ang. example-based machine translation), pri katerih je treba zbrati dvojezični korpus govorjenega jezika z označenim ujemanjem na različnih ravneh (ujemanje stavkov, besednih zvez...), nato pa se poskuša najti najboljši ujemalni algoritem, ki bo za neznano besedilo, ki ga mora prevesti, v učnem korpusu skušal najti primer, ki bi bil najbližji temu, ki ga mora prevesti.

Statistični pristop k strojnemu prevajanju skuša na to področje prenesti postopke, ki so bili razviti za razpoznavo govora. Osnovna pojma pri tem sta jezikovni model in prevajalni model. Jezikovni model izračuna verjetnost zaporedja besed (v bistvu stavkov), prevajalni model pa izračuna verjetnost, da se bo ciljni stavek T, ki je prevod stavka S v vhodnem besedilu, pojavil v ciljnem besedilu. Obe verjetnosti skupaj izračunavata verjetnost vhodno-ciljnih parov povedi.

3 JEZIKOVNI VIRI ZA RAZVOJ SISTEMOV STROJNEGA SIMULTANEGA PREVAJANJA GOVORA

Jezikovni viri, potrebni za razvoj tehnologije strojnega simultanega prevajanja govora, so lahko nekoliko različni glede na vrsto govorno orientiranega prevajanja. Na pravilih temelječi pristopi zahtevajo veliko natančnih jezikovnih (oblikoslovnih, skladenjskih, semantičnih) informacij, ki morajo biti vključene v korpuse govorjenih besedil. Gradnja takih virov je počasna in draga, saj zahteva veliko ročnega dela. Ker v zadnjem času statistični pristopi k strojnemu simultanemu prevajanju govora dosegajo večjo uspešnost kot na pravilih temelječi (Ney, 2001), se gradijo predvsem viri za razvoj te vrste sistemov. Zato se tudi v tem članku osredotočamo na predstavitev jezikovnih virov, potrebnih za statistične pristope k strojnemu simultanemu prevajanju govora.

Kot je navedeno v 2.2, so sistemi strojnega simultanega prevajanja govora sestavljeni iz treh osnovnih modulov: razpoznave govora, govorno orientiranega strojnega prevajanja, sinteze govora iz prevedenega besedila. Jezikovni viri, potrebni za gradnjo teh modulov, se v grobem ločijo na dvoje: tiste, ki so namenjeni za razvoj razpoznave in sinteze govora, ter tiste, ki so namenjeni za razvoj govorno orientiranega prevajanja.

3.1 Jezikovni viri za razpoznavo in sintezo govora

Za obstoječo tehnologijo razpoznave (prikriti modeli Markova) in sinteze govora (konkatenativna sinteza), ki je v svetu najširše uporabljana, so potrebni predvsem veliki slovarji besed, ki pokrivajo številna različna področja rabe, in korpusi govorjenih besedil. Slovarji morajo vključevati oblikoslovne in glasoslovne podatke o besedah: besedno vrsto, spol, sklon, število ipd., t.i. lemo (osnovno obliko) ter fonetični prepis (vključno z označenim naglasnim mestom in zlogi). Korpusi govorjenih besedil so ključni predvsem za razpoznavo, za gradnjo jezikovnih modelov, medtem ko je za gradnjo akustičnih modelov potrebno imeti posnete baze izgovarjav s čim večjim in čim bolj razpršenim vzorcem govorcev.

Večino večjih jezikovnih virov za tuje jezike je mogoče dobiti prek ELRE/ELDE (European Language Resources Association) (Internet #4) v Evropi in ameriškega LDC (Linguistic Data Consortium) (Internet #5). Pri tem nas zanima ponudba slovarjev, saj je teh manj. LDC ponuja od evropskih jezikov le vire za nemščino (Celex), španščino (Callhome – v okviru tega tudi viri za ameriško angleščino, kitajščino, arabščino in japonščino). ELRA ponuja velike slovarje za nemščino, francoščino in italijanščino. Poleg naštetih obstajajo slovarji, ki so bili narejeni vzporedno z velikimi govornimi bazami (SpeechDat, Callhome). Večina jih vsebuje nekaj tisoč do največ 50.000 vnosov. Od slovarjev lastnih imen velja omeniti slovarje Onomastica, ki vsebujejo ortografski in fonetični zapis lastnih imen. Narejeni so bili za večino evropskih jezikov, tudi za slovenščino. Dobiti jih je mogoče pri ELRI/ELDI.

Od jezikovnih virov za sisteme strojnega simultanega prevajanja govora za slovenščino smo že omenili slovar lastnih imen Onomastica, ki je dostopen prek ELRE/ELDE. Ta organizacija ponuja tudi govorno bazo posnetkov SpeechDat (II) za slovenski jezik, ki je prvi jezikovni vir za slovenski jezik, preverjen v mednarodnem centru za validacijo SPEX. Od ostalih govornih baz je prosto dostopen MobiLuz, govorna baza poizvedovanj o letalskih informacijah, ki vključuje tudi fonetični prepis in oblikoslovne oznake (Gros et al., 2000). V raziskovalne namene je mogoče dobiti tudi oblikoslovni slovar slovenskega jezika, ki je nastal v okviru projekta Multext-East (Internet #6), ta obsega 15.000 lem. Ostali jezikovni viri, ki se lahko uporabijo za razvoj sistemov strojnega simultanega prevajanja slovenskega govora, so bolj ali manj narejeni za lastno uporabo v organizacijah, ki so sodelovale pri njihovem nastajanju, so različno obsežni, vsebujejo različne podatke in so različno kvalitetni. Govorne baze za slovenski jezik so tako še Snabi (Kačič, 2002), Polidat (Zoegling Markuš, Kačič, Horvat, 2000), Luz (Gros, 1996), Gopolis (Dobrišek et al., 1998), govorna zbirka vremenskih napovedi (Žibret, Mihelič, 2000). Od slovarjev je za slovenski jezik narejen še en večji oblikoslovni slovar, SImlex (Verdonik et al., 2002) z 20.000 lemami, od večjih glasoslovnih slovarjev pa SIflex, ki vsebuje fonetični prepis vseh besed iz SImlexa (ibid.). O ostalih oblikoslovnih ali glasoslovnih slovarjih nismo zasledili objav, vendar lahko vsaj še za podjetje Amebis sklepamo, da tudi ima svojega (Internet #13). Večjega, referenčnega korpusa govorjenih besedil (eno-, dvo- ali večjezičnega) za slovenski jezik nimamo, lahko pa zasledimo objavo, ki napoveduje njegovo gradnjo (Stabej, Vitez, 2000).

3.2 Jezikovni viri za govorno orientirano prevajanje

Pri govorno orientiranem prevajanju je situacija manj jasna, saj na tem področju različni pristopi, opisani v 2.2, še tekmujejo med seboj. Kljub temu kažejo smernice k statističnim pristopom. Za te so osnovni vir dvo- ali večjezični poravnani korpusi govorjenih besedil, kar pomeni, da je treba govor najprej posneti, nato pa ga ortografsko prepisati, prevesti v izbran/e jezik/e ter poravnati. Poravnava je možna na različnih ravneh: osnova so vsekakor poravnane povedi, znotraj povedi pa so lahko dodani atributi, ki kažejo na ujemanje stavkov, besednih zvez ali celo besed. Bolj detajlna ko je poravnava, več ročnega dela zahteva, vendar posledično prinaša večjo uspešnost pri učenju statističnih modelov. Prav tako se lahko uspešnost modelov izboljša, če se korpus govorjenih besedil označi z osnovnimi oblikoslovnimi podatki: lemo in besedno vrsto. Pričakovali bi, da bi se uspešnost izboljšala tudi z dodajanjem skladenjskih informacij (t. i. parsed corpora), vendar so rezultati poletne delavnice 2003 na John Hopkins University v ZDA, kjer so delali eksperimente s statističnim prevajanjem s pomočjo skladenjsko označenih angleško-kitajskih korpusov (Internet #3), pokazali, da ni tako. Uspešnost

prevajanja se ni s pomočjo skladenjskih informacij prav nič povečala, seveda pa je med razlogi za to lahko tudi ta, da sta kitajski in angleški jezik veliko manj pregibna jezika, kot je npr. slovenščina, in bi za pregibne jezike bili rezultati povsem drugačni.

Glavna ovira pri gradnji korpusov govorjenih besedil je zamudna in draga produkcija, saj je zanje mogoče pridobivati besedila samo z ortografskim prepisom govorjenega jezika. Znana primera takšnega korpusa za angleški jezik sta Brownov korpus in korpus Wall Street Journala, ki sta bila uporabljena tudi za gradnjo nekaterih oblikoslovno in skladenjsko označenih korpusov govorjenih besedil (npr. Penn Treebank, Susanne), vendar sta enojezična, pri strojnem simultanem prevajanju govora pa želimo vsaj dvojezičen korpus govorjenih besedil. Takšen je angleško-francoski korpus Canadian Hansard, ki je veliko uporabljan v raziskovalnih sferah. Sicer pa se je večina takih korpusov snemala v manjšem obsegu in ad hoc, za potrebe posameznega sistema strojnega simultanega prevajanja govora. Tako so za:
– Verbmobil (Internet #12) v studiu posneli simulirane dialoge v nemščini, japonščini in angleščini; osebe so dobile različne naloge: dogovarjanje glede termina za sestanek, rezervacija poslovnega potovanja v Hannover v Nemčiji; spraševanje o znamenitostih ali kulturnih dogodkih; ok. 1000 dialogov;
– Nespole! (Burger et al., 2001) posneli dialoge v angleščini, nemščini, italijanščini, francoščini; osebe so se pogovarjale prek videokonferenčnega terminala; klicatelj je dobil nalogo, da sprašuje po turističnih informacijah o regiji Trentino v severni Italiji, na drugi strani je bil profesionalni delavec iz turizma, ki je dajal informacije; ok. 200 dialogov;
– Janus II (Internet #11) posneli bazo simuliranih spontanih dialogov, osebe so se morale dogovoriti za sestanek; 2000 dialogov v angleščini, nekaj manj v nemščini, španščini, korejščini in japonščini;
– EuTrans (Internet #1) posneli simulirane dialoge na temo potovanj v italijanščini, jih prepisali ter prevedli v španščino in angleščino.

Za slovenščino obstajajo večji vzporedni korpusi zaenkrat samo za pisana besedila: korpus Elan (Internet #8), ki je enomilijonska zbirka večinoma pravnih in drugih besedil ter romana 1984 Georgea Orwella, 1,8-milijonski korpus prevodov zakonodaje EU (Internet #10) ter korpus Trans, enomilijonska zbirka besedil s področja medicine, strojništva, zakonodaje, geologije in turizma (Internet #9), ter slovensko-hrvaški korpus vremenskih napovedi (Žibert et al., 2000).

Pomemben člen jezikovnih virov za strojno simultano prevajanje govora so tudi dvo- ali večjezični slovarji. Pri tem je mogoče prilagoditi dvo- ali večjezične slovarje, ki že obstajajo v elektronski obliki (npr. Word Dictionary, Comlex, WordNet, za slovenščino elektronski slovarji DZS). Seveda pa je tudi tukaj važno izbrati besedje, ki ga je smiselno vključiti v tak slovar. Prej našteti sistemi strojnega simultanega prevajanja govora (Verbmobil, Nespole!, Janus II, EuTrans) so namreč vsi po vrsti osredotočeni ne samo na področje turizma, ampak nekateri tudi znotraj tega področja na zelo omejen govorni tip, saj lahko le tako dosegajo zadovoljive rezultate. Strojno simultano prevajanje govora, ki bi bilo jezikovnozvrstno neodvisno, ob obstoječi tehnologiji še ne dosega sprejemljive kvalitete.

4 GRADNJA JEZIKOVNIH VIROV V OKVIRU PROJEKTA LC-STAR

Namen mednarodnega projekta LC-STAR (Internet #2) je razviti jezikovne vire, s katerimi bi izboljšali uspešnost sistemov strojnega simultanega prevajanja govora. Pri tem cilju so se združila velika podjetja, ki se ukvarjajo (tudi) z razvojem jezikovnih tehnologij (Siemens AG, Nokia, IBM, NSC), in univerze (RWTH Aachen, UPC). Vključeni so večinoma večji svetovni jeziki: ameriška angleščina, nemščina, španščina, katalonščina, italijanščina, ruščina, grščina, klasična arabščina, finščina, turščina, kitajščina, hebrejščina. Kot zunanji partner sodeluje pri projektu tudi Fakulteta za elektrotehniko, računalništvo in informatiko z Univerze v Mariboru, ki v skladu s specifikacijami projekta LC-STAR pripravlja jezikovne vire za slovenski jezik. Narejeni jezikovni viri bodo na voljo prek ELRE/ELDE.

Ob obstoječih jezikovnih virih za strojno simultano prevajanje govora (glej poglavje 3) lahko ugotovimo pomanjkanje standardov, saj viri vsebujejo različne informacije, so različno kvalitetni,

različno obširni, prilagojeni posameznim jezikom, pogosto so tudi zakodirani različno. Tako je eden osrednjih ciljev projekta LC-STAR postaviti standarde, po katerih se bodo razvijali jezikovni viri za strojno simultano prevajanje govora, saj so primerljivi in usklajeni viri v različnih jezikih eden od temeljev za uspešen razvoj sistemov.

V nadaljevanju natančneje predstavljamo gradnjo jezikovnih virov v okviru projekta LC-STAR samo za slovenski jezik.

4.1 Slovar za razpoznavo in sintezo govora

Nabor besed za slovar za razpoznavo in sintezo govora se deli v osnovi na dva dela: občna in lastna imena. Seznam občnih imen je narejen glede na frekventnost iz 10-milijonskega korpusa, ki ga sestavljajo besedila iz šestih večjih področij, kot prikazuje tabela 1.

Področje	**Podpodročje**
šport/igre	šport
novice	lokalni in mednarodni dogodki komentarji
finance	gospodarstvo, domač in tuj trg
kultura/zabava	glasba, gledališče, razstave, ocene potovanje, turizem
potrošniške informacije	zdravje poljudna znanost tehnologija za potrošnike
osebne komunikacije	pisma uredništvom, elektronske komunikacije (vendar samo v knjižnem jeziku)

Tabela 1: Področja, po katerih so razdeljena besedila, iz katerih je bil narejen nabor občnih imen.

Tabela 2 prikazuje velikost korpusa za slovenski jezik.

Področja	**Surov korpus**		**Čist korpus**	
	Število besed	**Št. različnih besed**	**Število besed**	**Št. različnih besed**
šport/igre	2.445.831	135.243	1.888.753	42.029
novice	2.475.715	148.648	2.178.834	63.912
finance	3.812.582	155.551	3.411.268	62.837
kultura/zabava	3.126.704	229.674	2.716.028	82.305
potrošniške informacije	1.270.690	123.167	1.146.009	48.326
osebne komunikacije	1.052.911	97.484	950.179	40.500
skupaj	14.184.433	452.331	12.291.071	139.645

Tabela 2: Velikost surovega in čistega korpusa za slovenski jezik. Surov korpus pomeni zbirko besedil brez številk, ločil in drugih znakov, čist korpus pomeni zbirko besedil brez singletonov (besed, ki se pojavijo v besedilu samo enkrat), lastnih imen in okrajšav.

Za končni nabor besed smo za vsako področje določili 96-odstotno pokritost, tako je bilo najmanj besed (17.256) za področje športa in največ besed (40.837) za področje kulture in zabave. Skupaj obsega nabor 65.096 besed, kar je 98,11-odstotna pokritost vhodnega korpusa.

Ker s korpusom ne zajamemo vseh besed, ki so potrebne za govorno vodene aplikacije, je bil dodatno narejen referenčni seznam 5.700 besed ali besednih zvez, ki se pogosto pojavljajo v teh aplikacijah. Seznam je bil izvorno v ameriški angleščini, nato pa preveden v posamezen jezik. Večbesedni prevodi so bili razcepljeni na posamezne besede, tako smo za slovenščino dobili dodaten seznam 6040 besed,

od tega ok. 2.500 takšnih, ki se niso pojavile v korpusu. Posebej je bil narejen tudi seznam t.i. končnih naborov (closed sets), tj. seznam besed tistih besednih vrst, ki jih je v posameznem jeziku končno število. Za slovenščino so to: predlogi, vezniki, zaimki, členki zanikanja in pritrjevanja, pomožni glagol in modalni glagoli. Zaimki in glagoli so bili zbrani v vseh oblikah. Tako smo dobili nov nabor nekaj več kot 1000 besed, od tega jih je bila nekaj manj kot polovica že zajeta s prvim naborom. Skupaj vsebuje seznam občnih imen okoli 68.000 besed.

Ločeno so bila zbrana lastna imena, razdeljena po področjih, kot prikazuje tabela 3.

Področje	**Število besed**	**%**	**Podpodročje**
osebna imena	22.469	49,2	
zemljepisna imena	11.828	25,9	imena slovenskih krajev in mest
			ostala zemljepisna imena (reke gore, pokrajine...)
			svetovne prestolnice
			velika svetovna mesta
			pomembne nacionalne in kulturne znamenitosti
			slovenska imena ulic
			države
stvarna imena	11.357	24,9	organizacije velika mednarodna podjetja
			blagovne znamke
skupaj št. vnosov	45.654	100	
skupaj št. razl. vnosov	45.027		

Tabela 3: Prikaz razporeditve lastnih imen po posameznih področjih za slovenski jezik v številu vnosov in v odstotkih.

V tabeli 3 vidimo, da je skoraj polovica osebnih lastnih imen. S pomočjo raznih atlasov in telefonskega imenika smo namreč zajeli večino slovenskih zemljepisnih in stvarnih imen ter morali dodati še nekaj najbolj znanih mednarodnih imen podjetij in krajev, da nismo presegli zgornje dovoljene meje, največ 50 % imen iz enega področja.

Celoten nabor ok. 113.000 besed bo oblikoslovno označen z oznakami, ki so enotne za vse jezike, vključene v projekt. Ker so vključeni jeziki izredno raznoliki (kitajski, arabski, hebrejski, finski, ruski, nemški, španski, slovenski...), se je bilo treba pri tem osredotočiti samo na tiste oblikoslovne lastnosti, ki so za tehnologije strojnega simultanega prevajanja govora pomembne. Posledično je nabor oblikoslovnih oznak, ki sledijo oznaki o besedni vrsti (te so samostalnik, pridevnik, številnik, zaimek, glagol, pomožni glagol, prislov, predlog, veznik, členek, medmet), nekoliko ožji kot pri slovarjih Multext-East in SImlex, vendar so zajete vse pomembne: spol, število, sklon, oseba ipd. Oblikoslovni oznaki v slovarju sledita podatek o lemi in fonetični prepis besede z abecedo SAMPA (Zemljak et al., 2002).

4.2 Slovar in korpus za govorno orientirano strojno prevajanje

Kot ugotavljamo v 3.2, so najredkejši viri za strojno simultano prevajanje govora ustrezni dvo- ali večjezični korpusi govorjenega jezika, saj je gradnja teh izredno dolgotrajna in draga. V okviru projekta LC-STAR se gradijo korpusi govorjenih dialogov v osnovi za tri jezike: španskega, katalonskega in angleškega. Tema dialogov je posredovanje turističnih informacij in je razdeljena na štiri večja področja: komunikacija v hotelu med receptorjem in gosti, komunikacija v turistični agenciji, komunikacija v turistični pisarni, komunikacija na letališčih in železniških postajah med informatorji ali prodajalci kart ter potniki. Dialogi so posneti v studiu in v celoti simulirani. Korpus govorjenih besedil bo obsegal 500.000 besed v posameznem jeziku. Posnetki so bili narejeni v španskem jeziku, ortografsko prepisani ter prevedeni v katalonščino in angleščino. Ker je v projekt vključena tudi slovenščina, obstaja možnost, da se korpus prevede tudi v naš jezik.

V okviru projekta pa bo narejen tudi manjši večjezični slovar na podlagi referenčnega nabora v angleškem jeziku. Slovar bo zajemal 10.000 fraz, ki so v komunikaciji v turizmu najpogosteje uporabljane. Preveden bo v 9 v projekt LC-STAR vključenih jezikov, tudi slovenščino, vsaka beseda bo označena z lemo in besedno vrsto.

4.3 Zagotavljanje kvalitete jezikovnih virov LC-STAR za slovenski jezik

Vse slovarje, ki bodo narejeni v okviru projekta LC-STAR, bodo preverili neodvisni strokovnjaki v validacijskem centru ELRE, SPEX-u (Speech Processing EXpertise centre). Validacija bo delno avtomatska in delno ročna. SPEX je pripravil orodja, s katerimi se izvede formalno preverjanje leksikonov, CST (Center for Sprogteknologi) iz Danske pa bo opravil ročno pregledovanje oblikoslovnih oznak in fonetičnega prepisa ter t.i. končnih naborov. S tem v projektu zagotavljamo kakovost narejenih jezikovnih virov.

Pomembni novosti v projektu LC-STAR pripravljenih slovarjev za slovenski jezik sta izredno premišljen in skrbno pripravljen nabor besed glede na izbrana področja (izbrana so bila glede na interese komercialnih partnerjev) ter usklajenost kode s številnimi izvorno povsem različnimi svetovnimi jeziki. Prav tako bodo to prvi pisni jezikovni viri za slovenski jezik, ki bodo preverjeni v mednarodnem validacijskem centru.

Dvo- ali večjezičnega elektronskega slovarja fraz s področja turizma, ki bi bil prilagojen za uporabo v jezikovnih tehnologijah, za slovenščino še nimamo, ta vir bo prvi tak. Pomembna pridobitev za slovenski jezik pa bi bil tudi poravnan večjezični korpus govorjenih besedil za področje posredovanja turističnih informacij, saj tudi tega za slovenščino še nimamo.

5 ZAKLJUČEK

V članku razpravljamo o jezikovnih virih, ki jih je potrebno zagotoviti za razvoj sistemov strojnega simultanega prevajanja govora. Za slovenščino v nekaterih centrih, kjer se ukvarjajo z jezikovnimi tehnologijami, že razvijajo posamezne module, ki so lahko del sistemov strojnega simultanega prevajanja govora (razpoznava govora, sinteza govora), vendar še nikjer ne obstaja zaključen sistem. Preden ga zgradimo, je treba zagotoviti ustrezne jezikovne vire.

V članku ugotavljamo, da je zadnji čas najboljše rezultate dalo statistično strojno simultano prevajanje govora, zato se osredotočamo na jezikovne vire, ki so potrebni za tak pristop. V osrednjem delu članka ugotavljamo, da lahko jezikovne vire za sisteme strojnega simultanega prevajanja govora ločimo na vire, potrebne za sintezo in razpoznavo govora, ter vire, potrebne za govorno orientirano prevajanje. Pri prvem so zlasti pomanjkljivo zastopani ustrezni slovarji: problem je predvsem pomanjkanje standardov, saj viri vsebujejo različne informacije, so različno kvalitetni, različno obširni, prilagojeni posameznim jezikom, pogosto so tudi zakodirani različno. Za govorno orientirano prevajanje je največji problem zagotoviti dovolj velike dvo- ali večjezične poravnane korpuse govorjenih besedil, saj je gradnja teh izredno dolgotrajna in draga.

Naštete pomanjkljivosti skušajo odpraviti v projektu LC-STAR, v katerega je med svetovne jezike vključena tudi slovenščina, jezikovne vire zanjo pripravljamo na Fakulteti za elektrotehniko, računalništvo in informatiko v Mariboru. Zgrajen bo oblikoslovni slovar 68.000 občnih besed in 45.000 lastnih imen, ki bo vseboval informacije o lemi, besedni vrsti in ostalih oblikoslovnih lastnostih ter fonetični prepis v abecedi Sampa. Za govorno orientirano prevajanje bo zgrajen večjezični slovar 10.000 fraz s področja turizma, ki bo prav tako oblikoslovno označen in fonetično prepisan, obstaja pa tudi možnost za pridobitev večjezičnega korpusa govorjenih besedil (500.000 besed, vključeni jeziki so španski, katalonski in angleški). Pomen pridobljenih virov za slovenski jezik je za slovarje zlasti premišljen nabor besed in usklajenost s slovarji drugih, svetovnih jezikov, večjezični slovar fraz in večjezični korpus za področje posredovanja turističnih informacij pa bi bila sploh prva takšna vira za slovenski jezik.

Literatura

Arnold, D., Balkan, L., Meijer, S., Humphreys, R. L., Sadler, L. (1994). Machine translation: An introductory guide. NCC Blackwell Ltd., Oxford.
Bub, T., Schwinn, J. (1996). Verbmobil: The evolution of a complex large speech-to-speech translation system. V: Proceedings ICSLP. 2371.
Burger, S., B. Laurent, P. Colleti, F. Metze, C. Morell (2001). The Nespole! VoIP Dialogue Database. V: Proceedings Eurospeech, Aalborg.
Ćavar, D., Menzel, W. (1998). Verbmobil: A speech-to-speech translation system. V: IS'1998: Jezikovne tehnologije za slovenski jezik. 25.
Dobrišek, S., Gros, J., Ipšić, I., Pepelnjak, K., Mihelič, F., Pavešić, N. (1998). Gopolis: slovenska podatkovna zbirka govorjenih poizvedovanj. V: Informacijska družba IS'1998: Jezikovne tehnologije za slovenski jezik. 105.
Erjavec, T. (2002). Compiling and using the IJS-ELAN parallel corpus. V: Informatica, 3. 299.
Gros, J., I. Ipšić, F. Mihelič, N. Pavešić (1996). Segmentation and labelling of Slovenian diphone inventories. COLING096. 298.
Gros, J., F. Mihelič, S. Dobrišek, T. Erjavec, M. Žganec (2000). A phonetically and prosodically annotated Slovene speech corpus. V: . IS'2000: Jezikovne tehnologije. 27.
Hirci, N. (2003). Prevajanje danes in jutri: delo s sodobnimi prevajalskimi orodji in viri. V: Jezik in slovstvo 3–4. 89.
Hoege, H. (2002): Project Proposal TC-STAR - Make Speech to Speech Translation Real. V: Third International Conference on Language Resources and Evaluation. 136.
Imperl, B., Sket, G. (2002). M-Vstopnica – uporaba avtomatskega razpoznavanja govora v praksi. V: Informacijska družba IS'2002: Jezikovne tehnologije. 116.
Internet #1 (12. 2. 2004): http://www.hltcentral.org/usr_docs/project-source/eutrans/AR-99/index.htm.
Internet #2 (12. 2. 2004): LC-STAR homepage. http://www.lc-star.com.
Internet #3 (12. 2. 2004): http://www.clsp.jhu.edu/ws2003/groups/translate/.
Internet #4 (12. 2. 2004): http://www.elra.info/
Internet #5 (12. 2. 2004): http://www.ldc.upenn.edu/
Internet #6 (12. 2. 2004): Multext-East. http://nl.ijs.si/ME/
Internet #7 (12. 2. 2004): http://www.fida.net
Internet #8 (12. 2. 2004): http://nl.ijs.si/elan/
Internet #9 (12. 2. 2004): http://www-ai.ijs.si/~spela/trans-index.html
Internet #10 (12. 2. 2004): www.sigov.si/evrokorpus
Internet #11 (12. 2. 2004): http://www.c-star.org/main/english/cstar2/tech/janus.html
Internet #12 (12. 2. 2004): citeseer.nj.nec.com/kurematsu00verbmobil.html
Internet #13 (12. 2. 2004): http://www.amebis.si/sklanjanje/
Kačič, Z. (2002). Pomen združevanja raziskovalnih potencialov pri preseganju jezikovnih pregrad v okviru jezikovnih tehnologij naslednjih generacij. V: Informacijska družba IS'2002: Jezikovne tehnologije. 111.
Kaiser, J., Sepesy Maučec, M., Kačič, Z., Horvat, B. (2000). Razpoznavanje tekočega slovenskega govora z velikim slovarjem. IS'2000: Jezikovne tehnologije. 39.
Kay, M., J. M. Gawron, P. Norwig (1994). Verbmobil: A translation for face-to-face dialog.CSLI, Stanford.
Metze, F., C. Langley, A. Lavie, J. McDonough, H. Soltau, A. Waibel, S. Burger, K. Laskowski, L. Levin, T. Schultz, F. Pianesi, R. Cattoni, G. Lazzari, N. Mana, E. Pianta, L. Besacier, H. Blanchon, D. Vaufreydaz, and L. Taddei (2002). The NESPOLE! speech-to-speech translation system. V: Proceedings of the Second International Conference on Human Language Technology Conference.
Ney, H. (2001). The statistical approach to spoken language translation. citeseer.nj.nec.com/article/ney01statistical.html.
Pozne ml., A., N. Pavešić, F. Mihelič (2002). Samodejno razpoznavanje govora: od unimodalnih k bimodalnim sistemom. V: Jezikovne tehnologije za slovenski jezik. 94.
Požgaj Hadži, V., Tadić, M. (2000). Slovensko-hrvatski paralelni korpus. IS'2000: Jezikovne tehnologije. 70.

ROJC, M. (2003). Časovo in pomnilniško optimalna struktura večjezičnega in poliglotskega sintetizatorja govora - arhitektura s končnimi stroji. Doktorska disertacija. Maribor: FERI.
Romih, M., Holozan, P. (2002). Slovensko-angleški prevajalni sistem. V: Jezikovne tehnologije za slovenski jezik. 167.
Rotovnik, T., Sepesy Maučec, M., Horvat, B. (2002). Uporaba algoritma ROVER pri razpoznavanju slovenskega govora. V: Informacijska družba IS'2002: Jezikovne tehnologije. 58.
Stabej, M., Vitez, P. (2000). KGB (korpus govorjenih besedil) v slovenščini. IS'2000: Jezikovne tehnologije. 79.
Šef, T., Gams, M., Škrjanc, M. (2002). Naglaševanje nepoznanih besed pri sintezi slovenskega govora. V: Informacijska družba IS'2002: Jezikovne tehnologije. 149.
Verdonik, D., Rojc., M., Kačič, Z. (2002). Zasnova in izgradnja oblikoslovnega in glasoslovnega slovarja za slovenski knjižni jezik. V: Informacijska družba IS'2002: Jezikovne tehnologije. J. 44.
Vesnicer, B., Mihelič, F., Pavešić, N. (2002). Sinteza govora z uporabo prikritih Markovovih modelov. V: Informacijska družba IS'2002: Jezikovne tehnologije. 28.
Vičič, J., Erjavec, T. (2002). Vsak začetek je težak: avtomatsko učenje prevajanja slovenščine v angleščino. V: Informacijska družba IS'2002: Jezikovne tehnologije. 20.
Waibel, A., Lavie, A., Levin, L. (1997). Janus: a system for translation of conversational speech. V: Kunstliche Intelligenz, 4, 51.
Wiebe, J., D. Farwell, D. Villa, T. O'Hara, K. McKeever, T. Sandgren, K. Payne (1997). Artwork: discourse processing in machine translation of dialog. Tech. rep. MCCS-96-294, Computing Research Laboratory, New Mexico State University.
Zemljak, M., Kačič, Z., Dobrišek, S., Gros, J., Weiss, P. (2002). Računalniški simbolni fonetični zapis slovenskega govora. V: Slavistična revija, 2, 159.
Zoegling Markuš, A., Kačič, Z., Horvat, B. (2000). Razvoj slovenske baze izgovarjav "POLIDAT". IS'2000: Jezikovne tehnologije. 95.
Žibert, J., Mihelič, F. (2000). Govorna zbirka vremenskih napovedi. IS'2000: Jezikovne tehnologije. 108.