

Nekatere težave pri strojnem prevajanju iz slovenščine v angleščino

■ IZVLEČEK

V članku je najprej na kratko opisana zgodovina strojnega prevajanja in osnovne metode, ki se zanj uporabljajo. Drugi del opisuje nekatere težave, ki so se pojavile pri razvoju slovensko-angleškega prevajalnika *Presis*, in nekatere možne rešitve zanje.

KLJUČNE BESEDE: *strojno prevajanje, zgodovina, metode, razvoj, prevajalnik*

■ ABSTRACT

A Few Difficulties with Machine Translating from Slovenian into English

The first part of the article briefly describes the basic methods and history of machine translation. The second part presents some problems encountered during the development of Slovenian to the English machine translator *Presis*, alongside with some possible solutions.

KEY WORDS: *machine translating, history, methods, development, machine translator*

ZGODOVINA STROJNEGA PREVAJANJA

Zametke zamisli o mehanizaciji prevajalskega procesa so poznali že v 17. stoletju. Dejanske možnosti za takšno prevajanje so nastale šele v prejšnjem stoletju. V petdesetih letih so se začele raziskave o uporabi računalnikov kot pripomočkov za prevajanje naravnih jezikov. V šestdesetih in sedemdesetih letih so prvi sistemi večinoma vsebovali obsežne dvojezične slovarje, v katerih so vnosi za besede iz izhodiščnega jezika imeli eno ali več ustreznih v ciljnem jeziku. Hkrati so vsebovali nekaj pravil za izbiro pravilne besede v končnem prevodu. Kmalu je postalo jasno, da so nekatera pravila za besedni red preveč zapletena in nepredvidljiva ter da je potrebna sistematičnejša metoda sintaktične analize.

Zaradi semantičnih pregrad, na katere so naleteli raziskovalci in za katere niso imeli rešitev, je kopnel tudi začetni zanos. Konec osemdesetih let 20. stoletja pa se je zgodil velik preobrat. Najprej je skupina družbe IBM objavila rezultate poskusov s sistemom (Candide), ki je temeljil na statistični metodi. Potem je neko japonsko podjetje začelo uporabljati metode, ki so temeljile na korpusih prevodnih primerov, torej je šlo za uporabo na primerih temelječega strojnega prevajanja.

V tem obdobju so se sistemi strojnega prevajanja razvili iz večinoma razi-skovalnih v bolj praktične aplikacije, aplikacije za poklicne prevajalce idr.

Te usmeritve so se nadaljevale še ob koncu devetdesetih let 20. stoletja. Uporaba prevajalnih sistemov in drugih prevajalskih orodij je naglo narasla, po-sebno v velikih podjetjih. Zelo hitro so te sisteme začeli uvajati pri lokalizaciji programske opreme. Močno sta narasli tudi prodaja prevajalnih sistemov za osebne računalnike in dostopnost prevajalnih sistemov na internetu (Altavista in drugi). V Evropi se dejavnosti, povezane s strojnim prevajanjem, izvajajo predvsem na tehle področjih: razvoj prevajalskih namizij, lokalizacija program-ske opreme, leksikološki viri in terminologija ter jezikovna tehnologija; to podpi-ra predvsem Evropska skupnost (Hutchins, 1996, str. 199).

Širitev uporabe sistemov strojnega prevajanja je v Evropi veliko počasnejša od pričakovane. Trgi so majhni in razdrobljeni, prevajalci pa sistemom niso na-klonjeni. Sisteme strojnega prevajanja uporabljajo zdaj le v večjih prevajalskih agencijah in večjih podjetjih. Manjša podjetja so bolj naklonjena prevajalskim namizjem in skupni rabi baz prevodov. Za cenejše različice sistemov strojnega prevajanja se zanimajo tisti, ki le občasno prevajajo. Vseeno so take različice najbolj prodajane v Severni Ameriki in na Japonskem.

METODE STROJNEGA PREVAJANJA

NEPOSREDNI PRISTOP (DIRECT APPROACH)

Neposredni pristop (Hutchins, 1992, str. 72) je značilen/značilna za prvo generacijo prevajalnih sistemov. Temelji na strategiji brez vmesnih stopenj v procesu prevajanja. Izvirno besedilo se pretvori neposredno v besedilo v cilj-nem jeziku. Prevajanje temelji na dobesednem prevajanju z nekaterimi prilago-ditvami besednega reda v ciljnem jeziku. Z jezikovnega vidika se je kmalu poka-zalo, da manjka pomemben element – analiza strukture izvirnega besedila, še zlasti analiza slovničnih razmerij.

Statistično strojno prevajanje (Statistical Machine Translation; SMT)

Statistično strojno prevajanje (Vičič, 2002, str. 13) je bilo do sedaj le redko uporabljano zaradi zahtevne matematične podlage, ki je potrebna za uporabo in razvoj statističnih metod strojnega prevajanja. Besedila morajo biti natančni pre-vodi iz izhodiščnega jezika v ciljni jezik. Ciljni in izhodiščni jezik se lahko pri uporabi zamenjata, ker so povedi v izhodiščnem jeziku natančni prevodi tistih v ciljnem jeziku. Tehnika statistične obdelave besedil je primerna za veliko količi-no besedil, ki ponujajo zadovoljivo število informacij o nekem jeziku oziroma

zadovoljivo število informacij za prevod med dvema jezikoma. Vzporedni korpusi se uporabljajo kot podlaga za učenje prevajanja. Za izboljšanje rezultatov prevajanja se uporablja različica trifaznega vzorca analiza-prenos-sinteza.

Sistem strojnega prevajanja, ki temelji na primerih (Example Based Machine Translation)

Ti sistemi temeljijo na dvojezičnem korpusu parov povedi. Z uporabo najboljšega algoritma želijo poiskati najboljši približek izhodiščne povedi. Tak sistem je še posebno privlačen za prevajanje zapletenih samostalniških povedi. Lahko se uporablja tudi za prevajanje povedi, ki so strukturno podobne že prej prevedenim povedim. Vseeno je treba to metodo razlikovati od sistemov, ki temeljijo na korpusih. Ti so bili namreč razviti za prevajanje besedil v nekem korpusu, pozneje pa se uporabljajo še za druga besedila. Metoda, ki temelji na primerih, pa ni nujno omejena na neki korpus ali podjezik (Hutchins, 1992, str. 125). Pravi sistem, ki temelji na primerih, ne uporablja slovničnih pravil, le primere povedi¹.

Interlingua

Interlingua (Hutchins, 1992, str. 73) je postopek, ki je prvi vključil analizo. Vmesna stopnja prevajalskega postopka je od jezikov neodvisni tako imenovani vmesni jezik (interlingua). Postopek je razdeljen na prevod iz izhodiščnega jezika v vmesni jezik in prevod iz vmesnega jezika v ciljni jezik. Izvirno besedilo je analizirano, prevod pa sestavljen na podlagi te analize, ki zagotovi vse potrebne podatke za sestavo prevoda. Ta metoda je najprivlačnejša za večjezične sisteme. Vsak modul analize je neodvisen od drugih modulov analize in tvorjenja. Ciljni jezik nima na rezultat analize nobenega vpliva. Cilj analize je medjezikovna/vmesna predstavitev. Prednost tega sistema je tudi preprosta možnost dodajanja novih jezikov. Za vsak jezik se dodata modula analize in tvorjenja. Težava tega sistema je opredelitev vmesnega jezika, tudi pri tesno povezanih jezikih (npr. romanski jeziki).

TRANSFERNO STROJNO PREVAJANJE (TRANSFER APPROACH)

Transferna metoda (Hutchins, 1992, str. 75) se uporablja v sistemih, ki imajo vgrajene dvojezične module med vmesnimi predstavitvami. Plod analize je abstraktna predstavitev izvirnega besedila. Naloga dvojezičnega transfernega

¹ <http://www.essex.ac.uk/linguistics/clmt/MTbook/HTML/book.html>

modula je preoblikovanje predstavitev izhodiščnega jezika v predstavitev ciljnega jezika. Transferna metoda ima tri stopnje, in sicer pretvorbo izhodiščnega besedila v abstraktno izhodiščno predstavitev, v kateri se reši večpomenskost ne glede na jezik, prenos predstave v abstraktno ciljno predstavitev in tvorjenje besedila v ciljnem jeziku.

Danes temelji večina sistemov na transfernem pristopu, pri čemer se v boljših sistemih vse bolj uveljavljajo empirični pristopi (statistično strojno prevajanje, na primerih temelječe strojno prevajanje).

Presis

Presis (PREvajalni SIStem) je strojni prevajalnik podjetja Amebis, ki za zdaj strojno prevaja iz slovenščine v angleščino in v nasprotni smeri. Program za prevajanje uporablja predvsem pravila. V prvi fazi analizira vhodno besedilo in ga pretvori v vmesni jezik, v drugi fazi pa ga iz vmesnega jezika prevede v izhodni jezik. Večina težav pri strojnem prevajanju nastane v prvi fazi (predvsem gre tu za težave, ki so kritične za razumevanje strojnega prevoda).

Neuspešna analiza

Stavčni analizator še ne zna analizirati vseh slovenskih stavkov. V času pisanja članka tako na primer manjka možnost, ko je členek pred povedkom in ne le pred drugimi stavčnimi členi (*»Danes že delam.«*). Te pomanjkljivosti se postopoma odpravljajo z dodajanjem novih pravil v analizator in novih načinov zapisa vmesnega jezika.

Druga težava pri analizatorju so za zdaj večstavčne povedi, kjer so posamezni stavki razbiti z vrinjenimi stavki (*»Stavek, ki je tak, je problem.«*). Tako razbitih stavkov še ne zna združiti, zato analizira posamezne dele in to privede do napačnih analiz.

Naslednja možnost so nepopolni stavki (tipično pri odgovorih na vprašanja). Pri takih stavkih se lahko pojavijo težave, kakršna je ta, da sicer prehodni glagol nima ob sebi predmeta, to pa zmede analizator, ki predmet pričakuje.

Kadar analizatorju ne uspe analizirati vhodnega stavka, prevajalnik potem prevaja besedo za besedo. V splošnem so rezultati takega prevajanja zelo slabi, zato je čim boljši analizator kritičen za kakovost prevajanja.

Dvojno zanikanje

V slovenščini mora analizator upoštevati, da nekatere besede (npr. *nikoli, nič, nihče, noben*) predstavijo tudi povedek v nikalno obliko, v angleščini pa mora

biti v takem primeru povedek obvezno v trdilni obliki. Na srečo takih besed ni veliko in se da celoten seznam razmeroma preprosto vgraditi v analizator, ta pa potem to reši že pri pretvorbi v vmesni jezik.

Vezljivost glagolov

Pri glagolih potrebujemo za uspešno analizo še podatek, na kakšen način se glagol uporablja:

Strgal je hlače.

Hlače so se strgale.

Strgalo se mu je.

Presis za to uporablja (glagolske) predloge. Predloga vsebuje glagol in podatek, s katerimi predmeti ali predlogi se ta glagol kot povedek veže, ter morebitne omejitve pri uporabi predloge (nekateri predloge se lahko uporabljajo le v nekaterih časih, zahtevajo točno določene predmete in podobno).

Predloge so potem združene v pomene. Tukaj so dodane tudi slogovne oznake, da lahko Presis prevaja s predlogo s čim bolj podobno slogovno oznako (če ima vhodna predloga oznako pogovorno, bo kot prvi prevod izbrana izhodka predloga s tako oznako).

Del predlog so lahko tudi točno določene besede. Tako je mogoče v prevajalnik vnesti cele fraze. V Presisu je na primer vnesena predloga, kjer je z glagolom »reči« povezan samostalni »bob« v funkciji predmeta v dajalniku in tožilniku. V angleščini je vnesena podobna predloga z glagolom »call« in samostalnikom »a spade« v funkciji predmetov. Tako lahko prevajalnik brez težav prevede stavek »Jutri bo bobu rekel bob.« z ustrezno angleško frazo.

Pomeni glagolov

Včasih pa sama predloga ne zadošča za razločitev pomena:

Osvojil je prvo mesto.

Osvojil je zlato.

Osvojil je občinstvo.

Osvojil je dekle.

Take primere je mogoče reševati tako, da se pri pomenu, sestavljenem iz predlog, pove, kateri pomen za predmet oziroma za osebek tipično nastopa z njim.

Slaba stran te rešitve je to, da je treba pri vsakem glagolskem pomenu vpisati pogoste spremljajoče pomene. Vendar se da iskanje takih kombinacij delno

avtomatizirati s preiskovanjem korpusov, kjer se najde seznam dvoumnih kombinacij, ki jih je treba potem ročno obdelati. Pri preiskovanju korpusov si je mogoče kdaj pomagati tudi s tem, da so dvoumnosti v različnih jezikih različne. Tako je mogoče poiskati možne pare pomenov v dveh korpusih v različnih jezikih (pri čemer pri dovolj velikih korpusih sploh ni treba, da so v korpusih ista besedila) in narediti seznam parov pomenov, ki se pojavljajo v obeh korpusih. Taki pari so potem kandidati za ročno dodajanje na seznam povezanih pomenov.

Druga možnost je dodajanje povezav pomenov v trenutku, ko se naleti na neustrezne prevode, vendar je za to rešitev potrebnega preveč ročnega preverjanja prevodov, da bi lahko bila glavna metoda, je pa zelo pomembna kot dopolnitev.

Ostajajo zelo dvoumni primeri, kjer tudi tako ni mogoče ugotoviti pravega prevoda, saj je lahko več kombinacij poudarjenih:

Prevedel je program.

Ta stavek ima štiri osnovne pomene.

Nekdo je prevedel računalniški program v drug jezik.

Nekdo je prevedel program prireditve.

Nekdo je s prevajalnikom prevedel računalniški program v izvršilno obliko.

Računalniški program je nekaj prevedel.

Ugotavljanje pravega pomena je v tem primeru popolnoma odvisno od konteksta.

Več možnih besednih vrst za isto besedo

Pri nekaterih besedah v slovenščini se zgodi, da gre pri isti besedi lahko za več besednih vrst. Kadar sta obe možnosti sintaktično pravilni, se mora analizator odločiti, katera je verjetnejša.

Tipična primera za to sta predloga *od* in *brez*, kjer gre lahko tudi za rodilnik množine oziroma dvojine samostalnikov *oda* in *breza*.

Besedna zveza »*prtljaga brez nadzora*« lahko programu pomeni tudi »*prtljago od brez od nadzora*«.

Taki primeri se najpogosteje rešujejo tako, da se na primer zmanjša ocena analiz, ki vsebujejo samostalnik »*breza*«, s čimer dobi prednost predlog »*brez*«.

Časovna prislovna določila

V slovenščini je kar nekaj prislovnih določil časa, ki so po zgradbi sicer bolj podobna predmetom v četrtem sklonu. Najpogosteje so sestavljena iz pridevnika (oz. pridevniškega zaimka ali pa tudi števnik) in samostalnika, ki izraža čas.

V takih primerih jih je kdaj težko ločiti od osebk:

Vsak večer gleda televizijo.

Pokaže se, da je pri teh besedah smiselno kot prvo možnost upoštevati pri-slovno določilo časa, kar je v slovenščini večkrat pravilno. Ni pa mogoče tega vedno zanesljivo trditi:

Vsak večer je lep.

Ta stavek lahko pomeni, da so vsi večeri lepi, ali pa da je nekdo lep vsak večer.

Zaimki

Pri prevajanju zaimkov se pojavlja nekaj težav, ki jim Presis za zdaj še ni kos. *Osvojil je to.*

Za glagol »osvojiti« (pravzaprav natančneje za glagolsko predlogo »osvojiti PR4«) pozna Presis več pomenov, ki jih zna razlikovati glede na pomen predmeta, povezanega s tem povedkom. Vendar kadar je predmet osebni oziroma kazalni zaimek, teh kombinacij ni mogoče uporabiti, ne da bi se najprej ugotovilo, katero besedo zaimek nadomešča.

Druga težava je prevajanje samih zaimkov. Zaimke moškega oziroma ženskega spola, ki nadomeščajo predmete in ne živih bitij, je v angleščini namreč treba nadomestiti z zaimkom za srednji spol. Za to je potreben postopek za ugotavljanje, katero besedo nadomešča zaimek, poleg tega je za vse angleške besede treba imeti podatek o spolu.

Dvoumni stavki

»Dela brez besed,« se na prvi pogled zdi popolnoma enostaven stavek. Ampak računalnik najde zanj kar nekaj možnih razlag:

- nekdo dela brez besed,
- dela so brez besed,
- dela sta brez besed,
- dela so od brez od besed,
- pri čemer je povsod »beseda« lahko tudi čitalniška prireditel.

Stavki v naravnem jeziku so v splošnem veliko bolj dvoumni, kot opazimo na prvi pogled, ker pri branju samodejno izločamo manj verjetne možnosti; te pa delajo strojnemu prevajalniku težave, ker mu ne uspe, da bi se jih znebil.

Kontekst

Odprl je čebelnjak. Potem je prijel matico.

Prijel je ključ. Potem je prijel matico.

Ena od težav, ki jih bodo sčasoma morali rešiti strojni prevajalniki, če bodo hoteli biti res uporabni, je upoštevanje konteksta. Med možnimi rešitvami je

gradnja asociacijskih mrež pomenov (del te mreže so že povezave med pomeni predlog in samostalnikov). Beseda s posameznega področja bo morala povečati verjetnost pomenov s tega področja, seveda pa je treba upoštevati, da to ni vedno nujno pravilno:

Odprl je čebelnjak. Potem je privil matico.

V tem primeru je povezava med povedkom in predmetom pomembnejša od trenutnega konteksta.

Pri kontekstu se lahko zgodi tudi to, da so pomembni stavki za stavkom, ki se prevaja, ne le pred njim:

Ima mačka. Je črne barve.

Ima mačka. Včeraj je preveč pil.

V takih primerih si bo prevajalni program moral zapomniti prejšnje dvoumnosti in se po potrebi vračati. Druga možnost je dvojna analiza – v prvem prehodu prevajalnik poišče najtipičnejše poudarke konteksta in jih potem uporabi v drugem prehodu analize.

Ni enakovrednega prevoda

Prva težava so besede, za katere še ni prevoda in jih je treba prevajati opisno. Tipična težava so različne vrste jedi – angleščina pač nima ustreznega izraza za povitico. Te vrste težave so še pogostejše pri prevajanju iz angleščine v slovenščino, kjer slovenskemu tehničnemu izrazoslovju pogosto ne uspe sproti slediti novostim. Določanje prevodov za nove besede je zelo zapleteno delo, ki mu strojni prevajalniki zlepa ne bodo kos.

Včasih je težava tudi to, da se stvari navadno v enem jeziku izrazijo drugače kot v drugem.

Primer »*Prisilil me je, da sem to prebral.*« bi se v angleščino najbolje prevedel s »*He made me read this.*«, kar bi bilo bolj dobesečno »*Prisilil me je prebrati to.*«. Dobeseden prevod »*He made me that I read this.*« v angleščini zveni nenavadno.

Napake v vhodnem besedilu

Strojni prevajalniki so izjemno občutljivi za napake v vhodnem besedilu. Ne znajo uganiti, katera beseda bi morala biti namesto zatipkane, kje manjka vejica ... V takih primerih analiza stavka navadno ne uspe (ali pa uspe, vendar je pomen nesmiseln: »*To nebo poceni.*«, kjer nekaj poceni nebo). Posledica tega so popolnoma napačni prevodi.

Možna rešitev za to je, da slovar pozna pogosto zatipkane besede (»življenski«) in so te potem pri analizi in prevajanju popolnoma enakovredne praviim (»življenjski«). Nekatere besede same po sebi sicer niso napačne, vendar

so napačne v nekaterih pomenih (»nebo« kot zanikana oblika glagola biti za prihodnost). Prav tako mora analizator tolerirati pogoste slovnične napake (»na velikemu vrtu«, »Nisem prebral knjigo.«) in narediti analizo, ki je enakovredna analizi pravilnega stavka.

Prevajalnik Presis prepozna večino tipov slovničnih napak, ki jih najde slovnični pregledovalnik BesAna, s čimer je omogočeno pravilno prevajanje stavkov, v katerih so take napake.

ZAKLJUČEK

Na poti do zanesljivega strojnega prevajanja nas čaka še zelo veliko dela. Nekatere težave so tehnično razmeroma preprosto rešljive (npr. vezave glagolov in kombinacije pomenov glagolov in samostalnikov), vendar je veliko dela z vnosom podatkov. Pri drugih (zaimki) bo več dela pri analizatorju, vendar ne bo nujno dolgotrajno vnašanje v slovar. Pri tretjih (nove besede, širši kontekst) računalniki še dolgo ne bodo nadomestili človeka.

Kljub vsem tem težavam pa strojni prevajalnik vedno uspešneje prevaja enostavna besedila. Pogosto sicer še zgreši pravi pomen, vendar potem ponudi pravi pomen med drugimi možnostmi. In včasih nam šele računalnik s svojimi »nenavadnimi« prevodi pomaga, da se zavemo, kako dvomni so pravzaprav kdaj stavki.

Viri

- Arnold, D., in sod. 2001. *Machine Translation: An Introductory Guide*. London: Blackwells-NCC. Dostopno na medmrežju.
<http://www.essex.ac.uk/linguistics/clmt/MTbook/HTML/book.html>
 Prevezeto 20. 4. 2005
- Hutchins, W. John, in Somers, Harold L. 1992. *An introduction to machine translation*. London: Academic Press.
- Hutchins, W. John. 1996. »The state of Machine translation in Europe«. V *Expanding MT Horizons: proceedings of the Second Conference of the Association for the Machine Translation in the Americas*, 2–5 October 1996. Montreal, Quebec, Canada. 198–205.
- Hutchins, W. John. Ni podatka. »The History of MT in a nutshell«. Neobjavljen članek. Dostopno na medmrežju.
<http://ourworld.compuserve.com/homepages/WJHutchins/Nutshell.htm> Prevezeto 15. 1. 2005
- Vičič, J. 2002. »Avtomatsko prevajanje iz slovenskega v angleški jezik na osnovi statističnega strojnega prevajanja«. Magistrska naloga. Ljubljana: Univerza v Ljubljani, Fakulteta za računalništvo in informatiko.