Kratek oris računalniškega prevajanja

Jaro Lajovic

V članku obravnava avtor računalniško prevajanje, in sicer s kratkim orisom sistemov za tako prevajanje. Te loči v dve skupini: konvencionalne (klasične) oziroma na pravila oprte (rule-based) in na analogne sisteme.

The article gives a short outline of existing computer translation systems. It distinguishes between conventional (classic) or rule-based and analogous types.

Računalniško (strojno, avtomatično) prevajanje (RP) je računalniška obdelava besedil, ki naj izvirno besedilo čim ustrezneje prevede v ciljni jezik. Sistemi za RP razčlenijo izvirnik in oblikujejo besedilo v ciljnem jeziku, vendar poimenovanje »prevajanje« zavaja, saj postopek ne vključuje razumevanja predloge.

V novejšem času je nastalo tudi več sistemov za t. i. računalniško podprto prevajanje (npr. translation memories – prevodni pomnilniki), ki sicer ne spadajo v ožji okvir RP, vendar se deloma približujejo nekaterim njegovim različicam.

Možnosti

Zamisel o uporabi elektronskih računalnikov za prevajanje se je porodila že kmalu po izdelavi prvega. Pričakovanja so bila velika: računalniško prevajanje naj bi bilo povsem avtomatično in visokokakovostno[1]. Nestrokovnjaki ob omembi RP še danes pričakujejo prav takšne sisteme, kdor pa nekoliko pozna prevajalsko delo, je ob tej predpostavki do možnosti RP skrajno zadržan[2]. Vendar so takšna pričakovanja nestvarna, kar so ugotovili že sredi 50. let. Še več: tedaj so se po začetnem obdobju pokazale številne težave, zaradi katerih so se zanimanje in sredstva za razvoj RP v 60. letih močno zmanjšali. Toda precej težav je bilo z leti sprejemljivo rešenih (kot bomo videli pri orisu sistemov), kar je spremenilo položaj računalniškega prevajanja.

Spremenilo pa ga je tudi drugačno razumevanje vloge prevoda. Čeprav se zdijo prevodi strokovnih besedil v nekaterih pogledih enotna skupina, jih po namembnosti lahko razdelimo na tiste za asimiliranje in tiste za diseminiranje informacij[3]. Slednji se približno ujemajo s tradicionalnim razumevanjem strokovnega

prevoda, pri katerem sta bistvena vsebinska enakovrednost izvirniku in jezikovna kakovost. Namenoma poudarjamo strokovni prevod; RP je povsem neprimerno v leposlovju, kjer je enakovrednost učinka vsaj tako pomembna (in včasih še pomembnejša) kot enakovrednost vsebine[4].

V sodobnem svetu se zaradi obsežnih komunikacij in izdatnega pretoka informacij povečuje potreba po prevajanju, še zlasti za asimiliranje informacij. Bližnji zgled je večjezična Evropska skupnost (kjer pa npr. administracija njenih direktoratov tem potrebam že zadošča z uporabo računalniškega prevajanja)[5,6]. Podobne potrebe obstajajo vsepovsod za prečesavanje tehničnih poročil, osnutkov, patentov in podobnih zapisov, v zadnjih letih vse bolj tudi sporočil in informacij v svetovnem računalniškem omrežju. V teh primerih je temeljna zahteva hitrost, ne jezikovna dovršenost. Večini teh potreb prevajalci ne morejo zadostiti, zato sistemi za RP niso njihov tekmec. Rekli bi lahko, da gre pogosto za dokumente oz. vire, ki bi brez RP nikdar ne bili prevedeni.

Računalniško prevajanje je kljub svojim otroškim boleznim preživelo prav zaradi potrebe po asimilaciji informacij: v času hladne vojne so na obeh straneh železne zavese budno spremljali nasprotnikov (zlasti tehnični) razvoj. To je bila osnovna pobuda projektov, iz katerih so kasneje nastali nekateri najvidnejši komercialni sistemi (npr. Systran, Logos), in sicer v času, ko je bil odnos do računalniškega prevajanja precej negativen.

Vrste sistemov

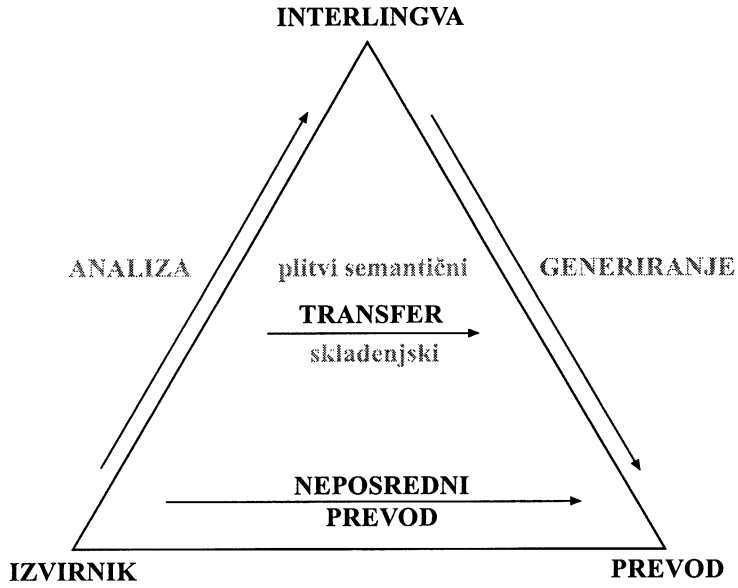
Sistemi za RP temeljijo na različnih pristopih. Ločimo dve skupini. Prva so konvencionalni (klasični) sistemi, zgrajeni na bolj ali manj poglobljenih jezikoslovnih temeljih; imenujemo jih tudi na pravila oprti (*rule-based*). Druga skupina so t. i. analogni sistemi, ki temeljijo na uporabi velikih (ponavadi dvojezičnih) podatkovnih zbirk in matematičnih metod.

Vsi so sestavljeni iz dveh delov: programa in skladišča znanja; pri klasičnih sistemih je slednji slovar (leksikon), pri analognih podatkovna zbirka. Zlasti na področju na pravila oprtih je približno takšna tudi delitev dela: za programski del skrbijo računalničarji, za leksikonski jezikoslovci.

Konvencionalni sistemi

Njihovo razvrstitev in odnos prikazuje t. i. grenoblski trikotnik. Više ko je sistem, globlja je njegova jezikovna analiza (slika).

Zgodnji (neposredni, direktni) sistemi so sloneli na zelo preprostih jezikoslovnih pravilih. Izvirnik so analizirali plitvo in lokalno ter nadomeščali besede. Prevodi so bili zelo dobesedni, sistemi niso mogli razrešiti homografov, izbira sklopov oz. izrazov v ciljnem jeziku je bila slaba.



Sledil je drugi rod, t. i. transferski sistemi. Za zapis lingvističnih pravil uporabljajo različne formalne slovnice (s čimer se približujejo računalniško primerni obliki), kodiranje leksikona je zahtevnejše itd. Besedilo obdelajo v treh fazah: analitični, transferski in generativni. Analitični modul (razpoznavalnik, parser) pripravi skladenjsko drevo izvirnega stavka. Transferski ga preoblikuje skladno s slovnico ciljnega jezika; v leksikonu izbere besede v ciljnem jeziku. Generativni oblikuje prevedeni stavek in zagotovi njegovo slovnično usklajenost. Moduli so jezikovno vezani: analitični na izvirni jezik, generativni na ciljnega, transferski na oba. Zato v večjezičnem sistemu število modulov hitro narašča (transferskih skoraj s kvadratom števila jezikov)[7].

Analiza je lahko zgolj skladenjska, a za boljšo kakovost je potrebna vsaj še plitva semantična, ki jo danes uporabljajo vsi večji transferski sistemi. Seveda postane priprava leksikona s tem zahtevnejša, saj mora poleg slovničnih in skladenjskih informacij zajeti še semantične, katerih formalni zapis je težaven.

Z vzpenjanjem po grenoblskem trikotniku se razdalja med analitično in generativno stranico (ki predstavlja transferski del) zmanjšuje. Interlingvalni sistemi imajo le dva modula: analitičnega in generativnega. Prvi pripravi jezikovno neodvisno predstavitev izvirnika v zapisu, imenovanem interlingva, drugi iz nje oblikuje prevod. Modula sta jezikovno neodvisna, kar je velika prednost. Vendar jezikovno neodvisna interlingva ne sme predstavljati besed, ampak koncepte, kar je v praksi težavno[8].

V nasprotju s transferskimi sistemi dodatek novega jezika v interlingvalni sistem zahteva le dva nova modula: enega za analizo, drugega za generiranje. Zaradi

jezikovne nevtralnosti analizatorju ni treba »vedeti« za ciljni jezik in generatorju ne za izvirnega.

Analogni sistemi

Na pravila oprti sistemi so uveljavljeni, vendar niso brez pomanjkljivosti. Zahtevajo npr. izredno velik začetni vložek dela (priprava programa, leksikonov), uporabljena pravila so praviloma pragmatična, povečevanje sistemov (npr. za komercialno uporabo) je zelo težavno, zahtevajo veliko računalniško moč, prevodi pa so – brez dodatnega prilagajanja potrebam specifičnega uporabnika – ponavadi še vedno preveč dobesedni. Če niso namenjeni zgolj okvirnemu informiranju, brez izjeme potrebujejo dodatno urejanje (*post-editing*).

Zato so skušali najti načine, pri katerih ne bi bilo treba eksplicitno formulirati lingvističnih pravil. Tako so se razvili na znanje (*knowledge-based*) in na vzorce oprti (*example-based*) ter statistični načini.

Na znanje oprti sistemi so sorodni interlingvalnim. Izhodišče zanje je, da je za prevajanje potrebno razumevanje, to pa je povezano z znanjem o stvarnem svetu[9]. Ustrezna računalniška »zbirka znanja« bi morala (v povezavi s programskim modulom, praviloma z uporabo umetne inteligence) zagotoviti RP, podobno človeškemu. Gradnja takšnih sistemov je težavna in omejeni ostajajo na ozka področja. So pa v rabi: primer je sistem KANT, razvit na univerzi Carnegie Mellon[10].

Na vzorce oprte sisteme sta omogočila razvoj računalniške tehnologije (velike pomnilniške zmogljivosti, hitra obdelava) in dosegljivost velikih dvojezičnih zbirk izvirnikov in njihovih človeških prevodov. Izhodišče je, da je pri prevajanju zelo pomemben priklic analognih besednih sklopov, ki jih je mogoče najti v dvojezični zbirki. Zbirke morajo biti za takšno uporabo uravnane: parno razvrščene na določeni (npr. stavčni) ravni. Uravnava je ključni del njihove priprave.

Sistem v zbirki poišče sklope, čim podobnejše (ne le enake) izvirnim. Z rekombiniranjem (in morebitnimi dodatnimi postopki) na generativni strani oblikuje prevedeni stavek. Lingvistično znanje je v sistemu implicitno, v nasprotju z eksplicitnim pri konvencionalnih[11]. Na vzorce oprte sisteme lahko povežemo s konvencionalnimi, kar se zdi obetavna smer nadaljnjega razvoja RP.

Najizrazitejši odmik od lingvističnih pristopov pomenijo statistični načini. V nasprotju s starejšo uporabo statističnih metod (npr. za odkrivanje določenih pravil za RP) sta njihov cilj analiza in generiranje zgolj s statističnimi postopki. Zahtevajo veliko dvojezično zbirko besedil in obsegajo tri stopnje: uravnavo zbirke; statistično izbiro ekvivalentnih besed (oz. sklopov) v izvirnem in ciljnem jeziku – t. i. prevodni model; in statistično izbiro ustrezne strukture za oblikovanje stavka v ciljnem jeziku (verjetnosti n-gramov) – t. i. jezikovni model[12].

Kljub omejitvam, zaradi katerih se niso uveljavili kot samostojni temelj sistemov, utegnejo imeti statistični načini pomembno vlogo v hibridnih sistemih za RP.

Računalniško podprto prevajanje

Omenili smo, da programi za računalniško podprto prevajanje (RPP) ne sodijo med sisteme za RP. Vendar jih zaradi sodobnega razvoja omenjamo. Po eni strani so zmogljivi prevodni pomnilniki sorodni na vzorce oprtim sistemom, po drugi pa jih povezujejo s sistemi za RP (npr. Eurolang Optimizer z Logosom)[13]. Zanimive so zemljepisne razlike: medtem ko sta ameriško in pacifiško območje zelo odprta za računalniško prevajanje (kar kaže tudi število uporabnikov), so Evropejcem ljubši programi za RPP[14].

Zmogljive programe za RPP lahko uporabnik (kot dodatno izbiro) vključi v svoj urejevalnik besedil in jih hkrati poveže z enim od sistemov za RP. Osnovna in dodatna orodja omogočajo uravnavo starih prevodov z izvirniki, izrabo dobljene dvojezične zbirke in njeno dopolnjevanje z novimi prevodi, ustvarjanje terminološke zbirke (slovarja) in njeno aktivno vključitev v RPP, ohranjanje oblikovne podobe izvirnika ter ponujajo nekatere dodatne pripomočke (npr. pripravo dvojezičnih konkordanc, analizo deleža ponavljanj)[15].

Uporaba

Uporaba računalniškega prevajanja se povečuje. Dejanski obseg je težko ugotoviti, vendar tudi bolj zadržane ocene govorijo o več kot milijonu tipkanih strani na leto[16]. Obsežen je seznam velikih podjetij in organizacij, ki uporabljajo sisteme za RP, uveljavljajo pa se tudi RPP.

Zadnja leta so na področju RP prinesla občutne spremembe. Še pred petimi leti so programi zanj tekli samo na velikih računalnikih ali delovnih postajah, uporabljali so jih pretežno za prevajanje tehničnih priročnikov, namenjenih objavi (t. j. za diseminiranje), seveda z nujnim dodatnim urejanjem. Že tedaj je veliko prednost pomenilo ohranjanje oblike izvirnika. Sistemi namreč v predpripravi ločijo ukazna zaporedja (npr. za okvire, polkrepki tisk in podobno) in prevedejo »golo« besedilo. Tik pred koncem v prevod vstavijo ukazna zaporedja in tako zagotovijo, da je prevod oblikovno enak izvirniku.

To in druge prednosti velikih sistemov imajo zdaj tudi sistemi za RP na osebnih računalnikih, ki so se izredno razmahnili v zadnjih treh letih. Med njimi so tudi znana imena, kot sta Systran ali sistema PAHO (Panameriške zdravstvene organizacije). Računalniško prevajanje vse bolj prodira na svetovno računalniško omrežje. Mrežje CompuServe uporabnikom npr. ponuja avtomatično prevajanje elektronske pošte z Intergraphovim Transcendom[17]. Nemara najzanimivejša smer na tem področju pa se odpira z možnostjo neposrednega prevajanja spletnih strani (*World Wide Weba*)[18].

Takšna je skica današnjega stanja. In jutrišnje? Predvidevanje je na tako dinamičnem področju nehvaležna naloga. A najbrž lahko pričakujemo po eni strani integriranje različnih pristopov v hibridne sisteme (kakršen je npr. Pangloss, povezava

transferskega, na znanje oprtega in na vzorce oprtega modela) po drugi strani pa nadaljnje povezovanje računalniškega in računalniško podprtega prevajanja.

Literatura

[1]HUTCHINS, W. J., SOMERS, H. L, An Introduction to Machine Translation, London, Academic Press, 1992, 6.

[2]KOLLER, W., Einführung in die Übersetzungswissenschaft, Heidelberg, Quelle & Meyer, 1992, 75–79.

[3]VASCONCELLOS, M., Machine translation, Byte 1993, 1, 162.

[4]KOLLER, W., Einführung in die Übersetzungswissenschaft, Heidelberg, Quelle & Meyer, 1992, 52, 152–154.

[5]HEARN, P., Machine Translation, Current Applications in Europe, MT News International 1996, 14, 6.

[6]PETRITS, A., The Commission's Operational Machine Translation System, Multilingual Action Plan, CEC – DG XIII Translation Service, Luxembourg, July 1995.

[7]HUTCHINS, W. J., SOMERS, H. L., An Introduction to Machine Translation. London, Academic Press, 1992, 69–76.

[8]LAJOVIC, J., InterLan's EUROTRA, MT News International 1996, 15, 6.

[9]LENAT, D. B., GUHA, R. V., Building Large Knowledge-Based Systems, Reading, Addison-Wesley, 1990, 20.

[10]MITAMURA, T., NYBERG, E., CARBONELL, J. KANT, Knowledge-Based, Accurate Natural Language Translation, v: Technology Partnerships for Crossing the Language Barrier, 1st AMTA Conference Proceedings. Columbia, AMTA, 1994, 232.

[11]ARNOLD, D., et al., Machine Translation, an Introductory Guide, Oxford, NCC Blackwell, 1994, Chapter 10.

[12]BROWN, P. F. et al., The Mathematics of Statistical Machine Translation, Computational Linguistics 1993, 19 (2), 263.

[13]ANON, New Versions of Eurolang Optimizer, MT News International 1995, 11, 3, 7, 14. Brace C, Vasconcellos M, Chris Miller L. MT Users and Usage, Europe and the Americas, MT News International 1995, 12, 14.

[14]NIXON, S., Transit for Windows, Language International 1994, 6 (4), 3–5.

[15]VASCONCELLOS, M., The Present State of Machine Translation Usage Technology, V: MT Summit IV, Proceedings, Kobe, IAMT, 1993, 35–45.

[16]KINGSCOTT, G., MT on-line from CompuServe, Language International 1996, 8 (1), 18–20.

[17]MACKLOVITCH, E., et al., MT Online, vč: Expanding MT Horizons. 2nd AMTA Conference Proceedings, Montreal, AMTA, 1996, 220–223.

An Outline of Computer Translation

Soon after the appearance of the first computers, there occurred the idea of their use in translation. It was accompanied by great expectations. The efforts for its implementation revealed a huge number of problems, but with time also various ways of solving them were found. Today there exist a number of computer translation systems and they produce over a million pages of translated text a year. They are of two types: rule-based and analogous.

There are also software programmes for computer-assisted translation, which do not belong to the domain of computer translation in the strict sense of the word, but which in many ways resemble the analogous systems. Future development may bring the two even closer together or connect them, as well as produce various new solutions adapted to the specific needs of different groups of potential users.