STROJNI PREVAJALNIKI IN ARHIVI

Zvone Štor *

UDK: 004.3:930.253

***Zvone Štor: Strojni prevajalniki in arhivi**. Tehnični in vsebinski problemi klasičnega in elektronskega arhiviranja. Zbornik referatov z dopolnilnega izobraževanja, Maribor 4/2005, št. 1, str. 184-188.*

Izvirnik v slovenščini, izvleček v slovenščini in angleščini, povzetek v angleščini.

Tehnologija strojnega prevajanja omogoča, da računalnik samodejno prevede določen tekst. S tem postanejo elektronski arhivi, članki in dokumenti dostopnejši, kar pa ne velja samo za tujce, ki jih zanima Slovenija, pač pa tudi za slovenske uporabnike, ki bi radi prebirali denimo kitajske časopise. Prve izkušnje z integracijo takšnega sistema v arhiv časopisnih člankov so pozitivne, vendar pa zahteva vzdrževanje in analiza delovanja sistema veliko pozornosti.

UDC: 004.3:930.253

***Zvone Štor: Machine Translations and Archives**. Technical and Field Related Problems of Traditional and Electronic Archiving. Conference Proceedings, Maribor 4/2005, No. 1, pp. 184-188.*

Original in Slovenian, abstract in Slovenian and English, Summary in English.

With the technology of machine translation texts are translated by computers. Moreover, the technology makes electronic archives, articles and documents much more accessible, not only for foreigners, who find Slovenia interesting, but also for those Slovene users, who would like to be able to read some Chinese newspapers, for example. In addition, some positive experiences have already been made in the field of integration of such systems into the newspaper archives. However, the process of maintaining and analyzing such systems requires attention to a high degree.

STROJNI PREVAJALNIKI IN ARHIVI

Ali lahko računalniki samodejno prevedejo slovenska besedila v angleščino? Morda v japonščino, kitajščino? Seveda lahko. Danes, ko razmišljamo o tehnologijah strojnega prevajanja, smo na prav takšnem začetku, kot sta bila decembra 1895 brata Lumiere na premieri prvega film z nekaj kratkimi in enostavnimi prizori. Sprva je šlo zgolj za zabavo, danes pa je to najpomembnejši ameriški izvozni artikel.

In zakaj ravno primerjava strojnega prevajalnika s filmsko industrijo? Razlog je preprost. Če v Sloveniji ne posnamemo nobenega filma, nas ni na kinematografskem zemljevidu. In analogija: če Slovenci sami ne bomo poskrbeli za postopke strojnega prevajanja, bo naš glas v prihodnje in še zlasti v svetu medmrežnega povezovanja, informacijske revolucije in navzkrižnega komuniciranja ostal neprepoznan, nerazumljiv in nezanimiv. Za ilustracijo velja izpostaviti dejstvo, da sta Poljska in Češka že leta 2000 začeli sistematično vlagali veliko sredstev v razvoj strojnega prevajalnika. Njihovi razlogi so bili preprosti: po vstopu v EU je skupnost pridobila kopico novih uradnih jezikov in prevesti je bilo potrebno gore pristopnih dokumentov. Samo v Sloveniji smo morali prevesti več kot 85.000 strani. Podatkov o tem, kakšne prihranke je prinesla tehnologija avtomatičnega računalniškega strojnega prevajanja, je težko oceniti. Pomen projektov se bo pokazal na dolgi rok.

* ***Zvone Štor, programer, organizator in vodja D. programov, Večer, 2000 Maribor, Slovenija.***

Že danes pa mora biti vsakomur jasno, da si brez tovrstnih rešitev ne moremo predstavljati informacijske družbe prihodnosti.

KAJ SPLOH JE STROJNI PREVAJALNIK?

Program, ki bi napisan tekst sproti prevedel v tuj jezik, ni več znanstvena fantastika. Strojni prevajalniki so poleg sintetizatorjev zvoka in tehnologij za prepoznavo govora trenutno najbolj vroča tema informacijskih raziskovalcev. Zlasti v svetu telekomunikacij niso redki primeri izgradnje sistemov, ko bi lahko na eni strani v slušalko govorili nemško, na drugi pa bi se slišal kitajski prevod. Dandanes so na voljo tako velike količine tekstov, baze znanja, pa tudi teoretičnih podlag ne manjka, da bi z dovolj truda, energije in natančnosti že lahko izdelali računalniški sistem, ki bi zadovoljivo prevajal besedila.

Prvi poskusi mehanskega prevajanja segajo v 40. leta prejšnjega stoletja, javnosti pa je bila s tem seznanjena šele leta 1954, ko so v ZDA izdelali sistem, ki je znal v angleščino prevesti 49 ruskih stavkov, pri tem pa so uporabili 250 besed in 6 slovničnih pravil. To je povzročilo takšno navdušenje, da so s tovrstnimi raziskavami pričeli tudi drugod po svetu (tudi v Jugoslaviji na Beograjski univerzi). Hladen tuš je sledil leta 1966, ko je ALPAC (Automatic Language Processing Advisory Committee) izdal poročilo, v katerem so zapisali, da so »tovrstne raziskave nesmiselne in da vsi dosežki niso kaj dosti vredni«. Področje strojnega prevajanja je za več kot 10 let zamrlo in praktično izginilo iz besednjaka vseh resnih akademskih in strokovnih krogov.

Pomembnejših dosežkov ni bilo vse do leta 1995, ko so v Singapurju razvili lokalni sistem za prevajanje iz angleščine v kitajščino, malajščino, japonščino in korejščino, zasnovan pa je bil kot pripomoček vladnim prevajalcem. Sistem, ki še danes deluje, omogoča prevajanje ogromnih količin dokumentov za naročnike z vsega sveta, ponuja pa tudi lokalizacijo podjetjem, ki razvijajo programsko opremo za kitajsko govoreči del tržišča.

Danes imamo po internetu kopico (tudi brezplačnih) servisov, ki pa podpirajo prevode zgolj za najbolj zastopane jezike: angleščino, nemščino, francoščino, španščino, japonščino, kitajščino in tu in tam še italijanščino.

KAKŠNO JE STANJE V SLOVENIJI?

V javnosti je najbolj odmevala informacija iz leta 2002, ko je podjetje Amebis iz Kamnika napovedalo predstavitev strojnega prevajalnika iz slovenščine v angleščino. V akademskih sferah je od tedaj nastalo veliko magistrskih in doktorskih nalog, ki sistematično in analitično razdeljujejo matematične postopke prevajanja. A slovenščina je slovanski jezik, ki je zelo pregiben in s skoraj prostim besednim redom. Večino funkcij, ki jih v slovenščini izražamo s končnicami besed (pregibanje), v angleščini izražamo z besednim redom in dodatnimi funkcijskimi besedami. Veliko težav denimo povzročajo samostalniki, s katerimi lahko tvorimo edninsko, dvojinsko ter množinsko obliko v šestih sklonih. Večina pridevnikov lahko tvori 3 spole, vsa 3 števila, 6 sklonov, 3 osnovne ravni stopnjevanja. In za nameček je slovenščina še jezik z mnogimi izpuščanji. Osebni zaimki (jaz, on, oni) imajo ponavadi ničto obliko in so izpuščeni. V našem jeziku ni določnih in nedoločnih členov. Dokaz kompleksnosti jezika pa je tudi velikost korpusa besed, ki je za 12 odstotkov manjši od angleškega.

Čeprav so rezultati slovenskih strojnih prevajalnikov včasih na zelo abstraktnem nivoju (beri: pogosto so zaradi nesmiselnosti zelo zabavni), pa pri kamniškem Amebisu optimistično verjamejo v tovrstne rešitve. Začeli so z 10.000 vnesenimi predlogami, danes jih imajo 20.000, za zadovoljivo delovanje prevajalnika pa naj bi po njihovih ocenah bilo potrebnih okoli 100.000 vnesenih predlog. Danes omenjeni manjko v naboru prevodov rešujejo enostavno: z ugibanjem.

STROJNI PREVAJALNIK IN ARHIVI

Odgovor na vprašanje, zakaj vključevati sisteme strojnega prevajanja v arhive, je preprost. Slovenski članki, prevedeni v angleščino (pa četudi gre za slab prevod) so veliko več vredni za tujce, kakor originalno slovensko besedilo. Pomembno je, da bralec uporabnik iz prevodov razbere bistvo članka, po potrebi pa si bo tako in tako uredil popolni prevod. Doslej so bili arhivi pač omejeni na jezikovno področje, ki so ga obvladovali. Izkušnje s prvim strojnim prevajalnikom, integriranim v arhiv člankov pri časopisni hiši Večer, so zelo pozitivne, kar priznavajo tudi raziskovalni novinarji iz tujine. Med rednimi uporabniki so namreč sodelavci New York Timesa, BBC-ja, pa tudi Vatikan je med pogostimi uporabniki, zabeleženi so celo uporabniki iz Kitajske, Avstralije, Rusije ...

Ker se tehnologija ubada s porodnimi težavami in so tudi omenjene storitve v začetni fazi razvoja, je uporabnikov računalniško prevedenega arhiva občutno manj. Konec koncev si tovrstne rešitve šele utirajo pot. Vzorec je premajhen, da bi lahko iz njega sklepali o pravilih, so pa zanimive vsebine, ki jih tujci iščejo pri nas. Na prvem mestu je seveda Tito in povojna zgodovina. Leta takoj po drugi svetovni vojni so zlasti zanimiva za Italijane in Avstrijce (Večerov povojni arhiv je bil podlaga trem seminarskim nalogam dijakov iz obmejnih srednjih šol). Uporabniki iz Kitajske (morda sploh ne gre za novinarje, raziskovalce ali naključne uporabnike interneta) so iskalnik uporabili 12-krat, največ pozornosti pa so namenili tibetantskemu duhovnemu vodji Dalaj Lami, zanimal pa jih ni samo visok obisk leta 2002, pač pa tudi poročilo o podelitvi nobelove nagrade leta 1989. Na drugem mestu »kitajskih interesov« se je pojavilo ime Bainqen Erdeni. Statistika vpogledov in seznam odprtih člankov kaže na to, da je nekdo zbiral podatke o dinamičnem duhovniku in politiku, ki je umrl v 51. letu starosti (uradni viri navajajo, da je šlo za srčno kap), bil pa je najvišji predstavnik struje, ki zagovarja avtonomijo Tibeta znotraj kitajskega ozemlja. Danes je le redkim znano, da sta znotraj budizma že od 14. stoletja naprej dve struji.

Med dostopi z ameriškimi domenami je bilo največ zanimanja za Irak in seveda za Georga Busha. Povečan obisk je bilo zaznati zlasti v času predvolilnih kampanj, jeseni 2004. Tina Maze je med posamezniki očitno najbolj znana Slovenka v ZDA. Preseneča zanimanje Indijcev za slovensko-hrvaški spor, zlasti o nuklearki, Nizozemce zanima »monopolni« položaj Mobitela. Slovenijo očitno zelo dobro poznajo v Švici in Veliki Britaniji, vendar zgolj zaradi športa. Ob vsakem pomembnejšem nogometnem srečanju obišče Večerov strojno preveden iskalnik veliko (očitno športnih) novinarjev iz Velike Britanije, ki brskajo za izjavami slovenskih nogometašev in trenerjev.

Interes tujcev za Slovenijo vendarle ni majhen, je pa v veliki meri odvisen od intenzitete dogodkov. Časopisni arhiv, ki ga ponuja Večer, so očitno doslej odkrili zgolj novinarji, do akademskih sfer in strokovnih krogov pa informacija o strojno prevedenem arhivu časopisnih člankov še ni prišla. Tudi na tovrstne storitve se je treba privaditi.

IN KAKO V PRAKSI DELUJE STROJNI PREVAJALNIK?

Podatkovna baza člankov je še naprej v izvornem jeziku. Ključne besede, ki jih vpiše uporabnik, se najprej iz angleščine prevedejo v slovenščino. Sistem nato izvede vse iskalne procedure, ki so enake kot pri iskanju v slovenskem jeziku. V tretjem koraku pa prevajalnik namesto rezultatov v slovenščini izpiše prevedene rezultate v angleškem jeziku. Tak način je vsekakor počasnejši, kot če bi za tujejezične uporabnike že vnaprej v celoti prevedli bazo člankov, vendar pa omogoča, da je kvaliteta prevodov z vsako novo različico programa boljša. Dokler intenzivnost uporabe ne preseže 1000 iskanj na dan, zmorejo sodobni internetni strežniki brez težav opraviti poleg drugega dela še vse procese prevajanja.

PRIHODNOST PREVAJALNIKOV

V Sloveniji lahko storitve strojnega prevajanja preštejemo na prste ene roke, večina pa jih nima praktične vrednosti. Online prevajalniki zmorejo prevesti vrstico ali dve slovenskega ali angleškega besedila, vendar zgolj kot prikaz delovanja sistema. Zanimivo je dejstvo, da ko se matematično razdela in dovolj natančno opiše izvorni jezik (v našem primeru je to slovenščina), je potem razmeroma enostavno dodajati nove jezike. V pripravi je že prevajalnik za nemščino in v roku 10 let bodo ti sistemi zagotovo tako dodelani, da bomo enakovredno brskali po kitajskih, japonskih, španskih in seveda angleških arhivih. Želimo si, da bi bilo tudi obratno, saj bi s tem pomembno prispevali k boljši prepoznavnosti naše države.

A pomemben vidik strojnih prevajalnikov je vendarle dejstvo, da ti programi zmorejo že danes prevesti kitajske tekste v angleščino, kar je veliko bliže slovenskim uporabnikom, kot pa za večino nečitljivi članki v azijskih pismenkah. Na ta način je mogoče enostavneje obvladovati globalni medijski prostor.

Interes kapitala več kot očitno narekuje tempo in smernice razvoja, kar za našo majhnost pomeni, da bi morala vodilno vlogo na tem področju prevzeti država. Razen velikih količin filigranskega prevajalskega dela tovrstni projekti ne zahtevajo veliko vlaganj. V danem trenutku je pomembno, da Slovenci dobimo enoten korpus besed, poenoteno zbirko prevodov in določimo standardno strukturo, kot jo ima denimo nekomercialna Menola (www.menola.org - slovenski internetni online prevajalnik), ki sloni na korpusu IJS-ELAN. Šele takrat, ko bodo izpolnjeni ti pogoji, bo mogoče razmišljati o boljših matematičnih modelih prevajanja.

Tovrstne rešitve pa niso uporabne samo na nivoju dokumentov ali člankov, pač pa jih je mogoče integrirati v glasovne informacijske postaje, tudi v telefonske centrale, z njimi je mogoče avtomatično generirati podnapise na filmih, televizijskih kanalih, poenostavljeno je prevajanje prospektov, reklamnega materiala, knjižic z navodili za uporabo ... Pomembno in veliko področje je tudi usklajevanje zakonodaje z EU, kandidiranje na javnih razpisih in še kopica drugih možnosti.

Ključ do uspeha so torej velike elektronske baze tekstov, kajti le tako se bomo dokopali do najbolj zastopanih besednih zvez, ki tvorijo jedro jezika in temelj prevajalskih mehanizmov. Prevedeni arhivi so danes že realnost.

VIRI

- *Statistika dostopa do internetnega iskalnika www.Vecer.com/arhiv.*
- *Jernej Vičič: Menola, magistrska naloga: Avtomatsko prevajanje iz slovenščine v angleščino na osnovi statističnega strojnega prevajanja, 2002.*
- *Programski paket Presis, Amebis.*
- *Gradivo časopisne dokumentacije Večera.*

SUMMARY

MACHINE TRANSLATIONS AND ARCHIVES

When in May 2004 many new members joined the European Union, many countries decided to support and encourage the development of the technology used in the so called machine or automatic translation. In Slovenia more than 85.000 pages of extremely difficult texts have already been translated. Since the Slovene language is one of the official languages of the EU, even more translations will be required in the future. In Slovenia there is the famous machine translator called Presis, a product manufactured by the company Amebis and the academic system Menola. In both cases this is only the beginning of developments in the field of machine translation. At present a computer can translate approximately 20.000 pages, but in the case of a high-quality translation about 100.000 pages must be translated at least. However, some positive experiences have been made with the integration of translation systems into electronic archives of newspaper articles. But of utmost importance is the analysis of accesses and information searches. With key words, written by foreign users, and readers we may see which topics, contents and services are found interesting by foreigners. According to these research data the historical facts and reports of Slovenian media about a certain event are said to be the most interesting. Older issues of newspapers are of greater importance for users from abroad than for Slovenian ones. According to the statistics of the newspaper publisher Večer more than 85 per cent of the searches have been done in the issues older than 10 years and only 9 percent of the Slovene users were interested in facts about Slovene history. The key to the use of machine translators are large electronic text bases, since this is the only way possible to produce high-quality translations, which represent the essence of the language and lay the groundwork for the translation mechanisms. Translated archives have already become reality.

Zvone Štor *se z raziskovanjem novih medijev ukvarja že več kot 10 let, pri časopisni hiši Večer iz Maribora pa skrbi za spletne strani in vodi projekte digitalizacije vsebin. V zadnjih letih so njegovo področje dela elektronski arhivi in elektronske knjige, veliko časa pa posveča integraciji novih tehnologij v obstoječe procese (jezikovne tehnologije, prepoznava in sinteza govora, lokacijske informacije, strojni prevajalnik ...). Redno objavlja članke v različnih slovenskih medijih, kot predavatelj pa je sodeloval na več konferencah in okroglih mizah. Je tudi ustanovni član IAB Slovenija, član projekta Maribor e-mesto in pobudnik iniciative za pokritost mesta z brezžičnim internetom.*