DANIELIUS ALGIRDAS RALYS

MAŠININIS VERTIMAS LIETUVIŲ KALBAI

Santrauka

Vertimas iš kitos kalbos visuomet yra tam tikras intelektinis iššūkis. Gal čia gali padėti kompiuteris? 1949 m. Warrenas Weaveris pasiūlė panaudoti kompiuterius tekstams versti. Atsiranda terminas – *mašininis vertimas* (MV). Mašininis vertimas pirmaisiais dešimtmečiais buvo sparčiai plėtojamas, siekiant įgyti strateginį pranašumą šaltajame kare. Populiariausios verčiamos kalbos – rusų ir anglų. Vyrauja pažodinis vertimas, sukuriami dideli kompiuteriniai dvikalbiai žodynai, apimantys per 170 000 žodžių.

1950–1960 metais atsiranda mašininio (kompiuterinio) vertimo sistemų, kurias galima pavadinti taisyklinėmis (*rule-based*). Jos kuriamos laikantis požiūrio, kad kalbą galima aprašyti naudojant tam tikrų taisyklių (taip pat ir gramatinių) sistemą. Tai buvo optimistinis laikotarpis – tikėtasi per keletą metų sukurti tobulą mašininį vertimą. Tačiau kompiuteris sunkiai „supranta“ gramatiką. Verčiant fleksines kalbas reikia keliasdešimties tūkstančių taisyklių, kurios turi būti tarpusavyje suderintos. To niekas dar nėra tinkamai padaręs. Toliausiai pažengusios sistemos buvo SYSTRAN MV sistema, pradėta naudoti Europos Komisijoje ir rusiška PROMT vertimo sistema. Vėliau taisyklinio MV pažanga sulėtėjo. Europinis EUROTRA projektas (1982–1992 m.), kai kuriais skaičiavimais kainavęs per 50 000 000 ECU, baigėsi nesėkme – šimtai specialistų taip ir nesukūrė veikiančios MV sistemos. Taip prasidėjo rimta taisyklinio MV krizė. Vėliau daug metų buvo trypčiojama vietoje.

Kilo klausimas – jei negalime parašyti tiek daug taisyklių, tai gal galima versti be gramatikos? 1990 m. įvyksta naujas proveržis – *IBM Thomas J. Watson Research Center* tyrėjų grupė suformuluoja statistinio mašininio vertimo pagrindus. Vertimo procesas prilyginamas tam tikro pranešimo perdavimui triukšmingu kanalu. Dekoduojama remiantis Bajeso teorema. Vertimas remiasi tekstynais, ypač svarbūs yra dvikalbiai tekstynai. Statistinis MV buvo greitai tobulinamas. Europos Komisijos remto projekto *EuroMatrix* metu sukurtas universalus atvirojo kodo statistinio mašininio vertimo programinės įrangos paketas*Moses*, kurio pagrindu buvo sukurtos pramoninio lygio MV sistemos. Gauti geri rezultatai – pasirodo, galima versti neturint nei žodyno, nei jokio supratimo apie gramatiką! Šis metodas labai palengvino fleksinių kalbų vertimą.

Mašininio vertimo pasiekimai šiandien veiksmingai pritaikomi ir lietuvių kalbai. 2005–2007 m. Vytauto Didžiojo universitetas vykdė ES Struktūrinių fondų finansuojamą projektą

„Internetinė informacijos vertimo priemonė“. Rezultatas – vieša internetinė vertimo iš anglų į lietuvių k. paslauga (http://vertimas.vdu.lt/twsas/). Taisyklinio vertimo variklį pateikė rusų kompanija PROMT, o kiti kalbiniai ištekliai buvo paruošti Lietuvoje. Bendro pobūdžio tekstų vertimo kokybės įverčiai BLEU metrikoje (procentais) – apie 10. Praktikoje tai reiškia, kad adekvačiai suprantamas tik kas trečias sakinys. Ši vertimo priemonė dar turi nemažų galimybių pagerinti vertimo kokybę, pvz. plečiant frazių žodyną. Nuo 2008 m. rugsėjo 25 d. *Google Translate* palaiko ir lietuvių kalbą. 2014 m. atliktų testų duomenimis, bendro pobūdžio tekstų vertimo kokybės BLEU įverčiai buvo apie 17.

2012–2014 m. Vilniaus universitetas vykdė ES finansuojamą projektą „Anglų–lietuvių–anglų ir prancūzų–lietuvių–prancūzų kalbų mašininio vertimo, paremto statistiniais metodais, sistemos sukūrimas“. Rezultatas – vieša internetinė vertimo paslauga (https://www.versti.eu/). 2014 m. atliktų testų duomenimis, bendro pobūdžio tekstų vertimo kokybės BLEU įverčiai daugiau kaip dvigubai viršijo taisyklinio vertimo rezultatus ir buvo praktiškai tolygūs *Google Translate* vertimo sistemos rezultatams. Verčiant tam tikrų sričių dokumentus (pvz., teisės) BLEU įverčiai maždaug dvigubai didesni, nei verčiant bendrojo pobūdžio tekstus, ir gerokai viršijo *Google Translate* (2014 09 19) rezultatus. Visgi, tokie net geriausi mašininiai vertimai dažnai prašyte prašosi geresnio vertėjo žvilgsnio. Tad ar mašinos gali versti išties gerai?

Pastarieji keleri metai žada naujų proveržių naudojant neuroninį mašininį vertimą. Neuroniniai tinklai patys konstruoja transformavimo taisykles. Tikėtina, jog artimiausioje ateityje neuroninio MV sistemos vers geriau už vidutinišką vertėją. Vilniaus universitetas ruošiasi 2018 m. pradėti įgyvendinti naujos kartos neuroninį mašininį vertimą anglų, lietuvių, lenkų, prancūzų, rusų ir vokiečių kalboms.

Taigi, naujausi mašininio vertimo pasiekimai neaplenkia ir lietuvių kalbos.

ESMINIAI ŽODŽIAI: lietuvių kalba, istorija, kompiuterinė lingvistika, mašininis vertimas, neuroniniai tinklai, dirbtinis intelektas.

DANIELIUS ALGIRDAS RALYS
Vilniaus universiteto Taikomųjų mokslų institutas
M. K. Čiurlionio g. 29, 03100 Vilnius
danielius.ralys@gmail.com