Trumpa mašininio vertimo istorinė apžvalga

VU vykdomo projekto
„Anglų-lietuvių-anglų ir prancūzų-lietuvių-prancūzų kalbų mašininio vertimo, paremto statistiniais metodais, sistemos sukūrimas“ pristatymo dalis

Parengė projekto ekspertas
Danielius Algirdas Ralys
Vilnius, 2012 m. balandžio 18 d.

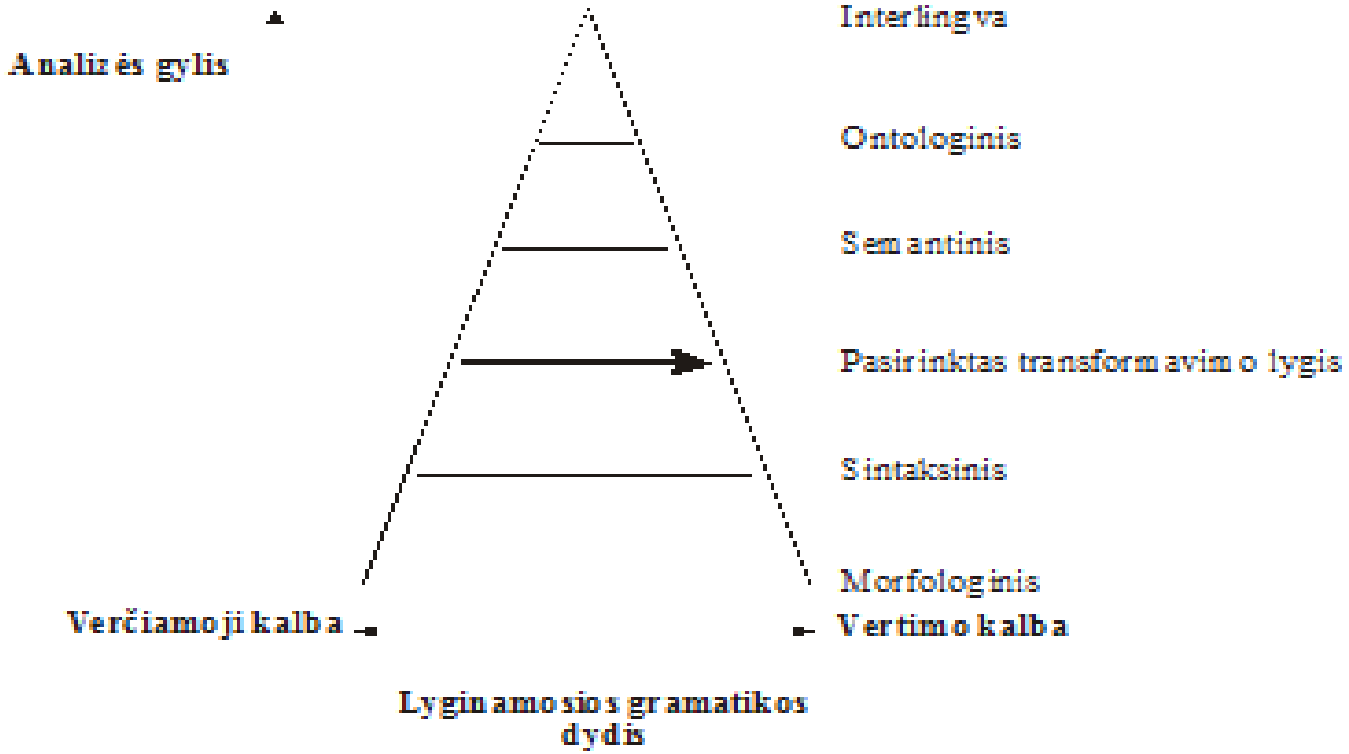
Mašinos gali versti

- 1947 m. Warren Weaver pasiūlė panaudoti kompiuterius tekstų vertimui. Atsiranda terminas – mašininis vertimas (MV).
- MV imamas sparčiai vystyti JAV ir TSRS, siekiant įgyti strateginį pranašumą šaltajame kare.
- Populiariausios verčiamos kalbos – rusų ir anglų.
- Vyrauja pažodinis vertimas, sukuriami dideli kompiuteriniai dvikalbiai žodynai, apimantys virš 200 000 žodžių.

Atsiranda taisyklinis mašininis vertimas

- 1950 – 1960 metais atsiranda mašininio (kompiuterinio) vertimo sistemos, kurias galima pavadinti taisyklinėmis (*rule-based*).
- Jos kuriamos laikantis požiūrio, jog kalbą galima aprašyti naudojant tam tikrų taisyklių (taip pat ir gramatinių) sistemą.
- Optimistinis laikotarpis – tikėtasi per keletą metų sukurti tobulą mašininį vertimą.

Mašininio vertimo lygiai



Pav. 1 Taisyklinio vertimo lygiai.

Ar kompiuteris „supranta“ gramatiką?

Teksto struktūros nagrinėjimas didina automatinio vertimo tikslumą.

- Kaip turi būti skaidomas tekstas – į sakinius, frazes, žodžius, morfemas?
- Kokiame lygyje tekstas turi būti nagrinėjamas: morfologiniame, sintaksiniame, semantiniame?

Banguojančios viltys

- Vyravo optimistinės MV perspektyvos, tačiau pasiekti prasti praktiniai rezultatai.
- 1966 m. JAV įkurtas ALPAC (Automatic Language Processing Advisory Committee) komitetas nusprendžia, jog MV artimiausiu metu neturi perspektyvų.
- MV projektų finansavimas JAV nutraukiamas dvidešimčiai metų, jis sumenksta ir kitose šalyse.

Vis dėlto mašininis vertimas progresuoja!

Praktika parodė, jog ALPAC klydo.

SYSTRAN MV sistema pradedama naudoti Europos Komisijoje.

Įvairiose šalyse atsiranda veikiančios MV sistemos:

- ARIANE (Grenoble);
- SUSY (Saarbrücken);
- Mu (Kyoto);
- Interlingva metodo taikymas Nyderlanduose (Rosetta, DLT).

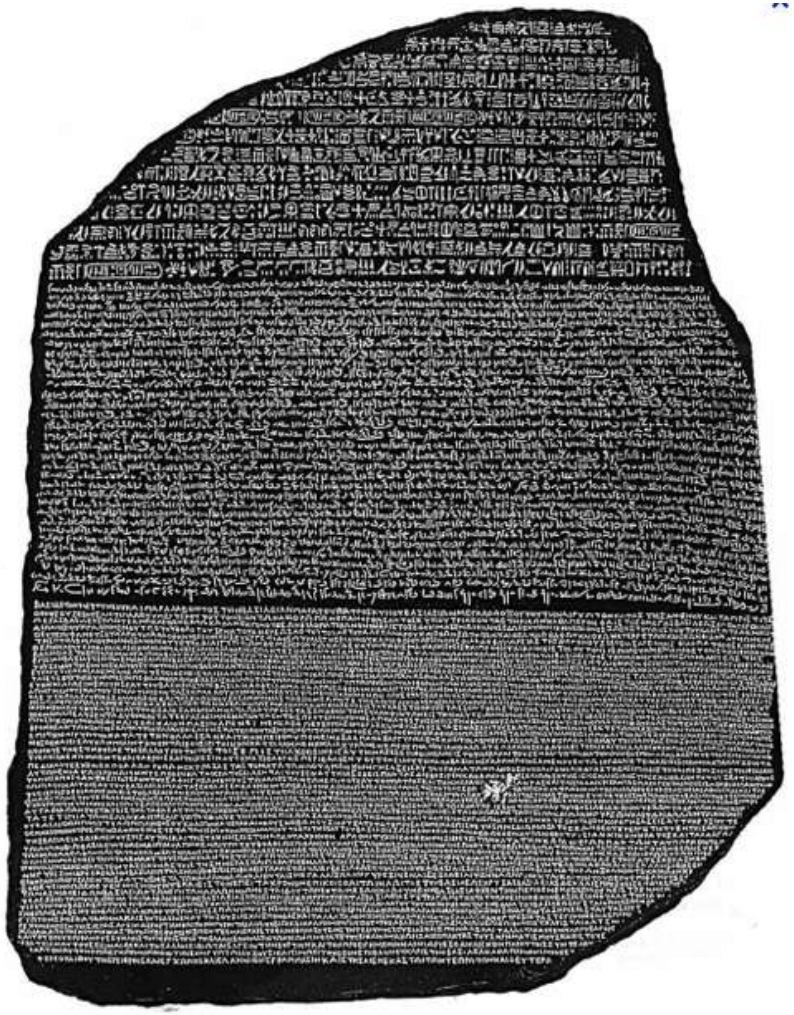
Taisyklinio MV pažanga lėtėja

- Europinis EUROTRA projektas (1982 – 1992 m. m.), kainavęs apie 50 000 000 ECU, baigiasi nesėkme – šimtai specialistų taip ir nesukūrė veikiančios MV sistemos.
- Tai – rimta taisyklinio MV krizė. Jau daug metų trypčiojama vietoje.
- Dar ir šiandien taisyklinio vertimo lyderiai – vis tas pats SYSTRAN bei kelių dešimtmečių senumo rusiška PROMT vertimo sistema.

Ar galima versti be gramatikos?

- 1990 m. įvyksta naujas proveržis - IBM tyrėjų grupė suformuluoja statistinio mašininio vertimo pagrindus (P. Brown et al.).
- Vertimo procesas prilyginamas tam tikro pranešimo perdavimui triukšmingu kanalu.
- Dekoduojama remiantis Bajeso teorema.
- Vertimas remiasi tekstynais, vertimui ypač svarbūs dvikalbiai tekstynai.
- Geri rezultatai – pasirodo, galima versti neturint nei žodyno, nei jokio supratimo apie gramatiką!

Lygiagretus tekstynas ant Rozetės akmens



Rozetės akmuo – pirmasis lygiagretus tekstynas, o taip pat ir statistinio vertimo objektas

Mašininio vertimo pritaikymas lietuvių kalbai

- 2005 - 2007 m. Vytauto Didžiojo universitetas vykdė Europos Sąjungos Struktūrinių fondų finansuojamą projektą „Internetinė informacijos vertimo priemonė“ . Rezultatas – vieša internetinė vertimo iš anglų į lietuvių k. paslauga. Vertimo variklį pateikė rusų kompanija PROMT. Nėra aišku, kiek laiko dar bus teikiama ši paslauga.
- http://vertimas.vdu.lt/twsas/
- Nuo 2008 m. rugsėjo 25 d. Google Translate palaiko ir lietuvių kalbą. Tačiau to neleidžiama naudoti komerciniams tikslams!
- Niūri realija – akademinė visuomenė neturi galimybių tobulinti šias sistemas.

Ar mašinos gali versti gerai?

- Žodžiai turi daug prasmių. Daugiaprasmiškumas buvo ir išlieka svarbiausia kompiuterinio mašininio vertimo problema.
- Sunki problema – kaip versti įvardžius (anaforos atpažinimas).
- *The soldiers killed ten women. They have been buried next day.* Kas buvo palaidoti, jie ar jos, kareiviai ar moterys?
- Sintaksinių struktūrų nustatymas šių vertimo problemų neišspręs.
- Ieškoma išsigelbėjimo semantikoje bei kuriant įvairias ontologijas.
- MV problemos stimuliuoja pažangą dirbtinio intelekto kūrimo srityje.
- Populiarėja mišrios (hibridinės) vertimo sistemos, apimančios tiek taisyklinį, tiek ir statistinį MV.
- Nuo 2010 m. SYSTRAN (Systran Server 7) inkorporavo ir statistinį vertimą į savo sistemą.
- Panašiu keliu eina ir PROMT.

Statistinio mašininio vertimo esmė ir pagrindiniai dalykai

Parengė projekto ekspertas Virginijus Dadurkevičius

2012 m. balandžio 18 d.

Vilnius

Statistinio mašininio vertimo prielaidos

Bendras kontekstas

- Norimus dalykus dažnai sužinom netiesiogiai, atlikdami rekonstrukciją:
  - reikia masės, bet matuojam svorį
  - reikia greičio, o matuojam laiką
  - temperatūrą nustatom, matuodami ilgį
  - domina žvaigždžių cheminė sudėtis, o matuojam spektrus
  - kūno erdvinę sandarą sužinom, analizuodami linijines rentgenogramas
  - gama šaltinių išsidėstymą danguje rekonstruojam iš laike moduliuoto signalo
  - ir t. t.
- Matavimai būna su paklaidomis, transformacijos – nevienareikšmiškos, rekonstrukcija – apsunkinta. Toli gražu ne atvirkštinės funkcijos suradimas $y = x^2$ atveju.
- Kuo vadovautis, atliekant rekonstrukciją?

Statistinio mašininio vertimo prielaidos

Bajeso metodas

- Atliekant atvirkštinio skaičiavimo tikimybinius uždavinius, jau nuo 1763 m. vadovaujamasi Bajeso teorema:

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

  - A ir B yra susiję įvykiai
  - $P(A)$ ir $P(B)$ – jų nepriklausomos tikimybės
  - $P(A|B)$ ir $P(B|A)$ – jų sąlyginės tikimybės
- Pritaikant mūsų aptariamiems atvejams, Bajeso teoremą galima perfrazuoti taip:
  - A – hipotezė (pvz., hipotetinė erdvinė kūno struktūra)
  - B – realiai gauti duomenys, matavimo rezultatai (pvz., linijinės kūno rentgenogramos)
  - $P(A)$ – apriorinė (išankstinė) hipotezės tikimybė (pvz., tuo mažesnė, kuo labiau nukrypstama nuo vidutinės kūno sandaros); $P(A)$ – konstanta, galima į ją neatsižvelgti
  - $P(B|A)$ – aposteriorinė (atsižvelgiant į įvykusį matavimo faktą) hipotezės tikimybė (pvz., kokia tikimybė, kad pasirinkus tokią tai hipotetinė erdvinę kūno struktūrą gali susigeneruoti realiai jau gauti duomenys)
  - Ta hipotezė, kuri maksimizuoja $P(A|B)$ yra pati tikimiausia
- Taikymo sudėtingumas
  - Galimų hipotezių gali būti be galo daug
  - Sunku įvertinti apriorinį žinojimą
  - $P(B|A)$ matematinis išreiškimas gali būti labai sudėtingas
  - Maksimumo paieška gali būti matematiškai ir praktiškai labai komplikuota

Vertimas, kaip statistinis procesas

Pagrindinės idėjos

- 1990 m. *IBM Thomas J. Watson Research Center* padaryta prielaida:

- Viskas vyksta statistiškai! Todėl galioja Bajeso formulė, ir:

$$\tilde{L} = \arg\max_{L} P(A \mid L) * P(L)$$

A – angliškas sakinys, kurį reikia išversti

L – hipotetinis lietuviškas sakinys

$\tilde{L}$ – tinkamiausias lietuviškas vertimas

$P(A|L)$ – tikimybė, kad hipotetinis lietuviškas sakinys gali būti išverstas į duotą anglišką sakinį (statistinis vertimo modelis)

$P(L)$ – hipotetinio lietuviško sakinio tikimybė (statistinis kalbos modelis)

- Problemos:
  - šimtai tūkstančių galimų žodžių kiekvienoje kalboje
  - reikia milžiniškų skaičiaus jau išverstų sakinių vertimo modeliui sudaryti
  - kiekvienoje kalboje yra savi žodžių tvarkos dėsniai
  - fleksuotos kalbos (lietuvių kalboje gali būti iki 1,5 mlrd. teoriškai galimų žodžių formų!)

Vertimas, kaip statistinis procesas

Dabartinės galimybės

- Sukurti metodai operuoti ne tik žodžiais, bet ir sustabarėjusiomis frazėmis.
- Prieš vertimą žodžiai gali būti lemuojami ir anotuojami morfologinėmis žymelėmis (“*factors*”), pvz., “*žvejams*” keičiamas į “*žvejas*” ir pažymima, kad originali forma buvo daiktavardžio daugiskaitos naudininkas. Atskirai “verčiant” lemas ir jų žymeles išvengiama milijardinių formų gausos ir sumažėja reikalavimai tekstynų dydžiams. Paskutinėje vertimo stadijoje lemos (bet jau kitoje kalboje) vėl sujungiamos su žymelėmis ir atstatoma jų normali forma.
- Žymelėse gali būti nurodoma ne tik morfologinė, bet ir sintaksinė-semantinė informacija.
- Europos Komisijos remto projekto *EuroMatrix* metu sukurtas universalus atviro kodo statistinio mašininio vertimo programinės įrangos paketas MOSES.

Statistinis vertimas VU projekte

Užduotys ir kokybė

- Vienkalbiai tekstynai:
  - lietuvių kalba – apie 850 mln. žodžių
  - anglų kalba – apie 1 mlrd. žodžių
  - prancūzų kalba – apie 1 mlrd. žodžių
- Dvikalbiai tekstynai:
  - anglų-lietuvių kalbų – apie 8 mln. sakinių
  - prancūzų-lietuvių kalbų – apie 7 mln. sakinių
- Kompiuterinės lingvistikos instrumentai lietuvių, anglų ir prancūzų kalboms
  - morfologiniai analizatoriai, vienareikšmintojai ir sintezatoriai, tikslumas – apie 95 proc.
  - sintaksiniai-semantiniai analizatoriai
- Vertimo tikslumas, įgyvendinus projektą:
  - bendrieji tekstai – virš 37 proc. (naudojant BLEU metriką)
  - teisinės ir IT sričių tekstai – virš 50 proc.

Projekto tęstinumas

- 5 metus bus užtikrintas vertimo paslaugos teikimas
- 2 metus bus analizuojami vartotojų atsiliepimai ir pagal tai tobulinama paslauga
- 1 metus bus viešinama sukurta paslauga
- Kasmet bus apdorojama apie 200 000 užklausų.

Mokslinis darbas

- Puiki galimybė projekto pagrindu sukurti (suburti?) VU modernios kompiuterinės lingvistikos laboratoriją
- Sukaupti tekstynai – nejkainojama ir absoliučiai būtina dirva moderniems lingvistiniams tyrinėjimams
- Galimybė tyrinėti ir tobulinti projekte naudojamus lingvistinius instrumentus (pvz., įvairių kalbų morfologinius analizatorius-sintezatorius)
- Moksliniai straipsniai, daktaro darbai

Pedagoginė veikla

- Galimybė studentams filologams ir informatikams susipažinti su modernia veikiančia mašininio vertimo sistema
- Kursiniai ir magistro darbai
- Savanoriai galės prisidėti kaupiant ir apdorojant papildomus tekstynus
- Puikus būdas reklamuoti VU kaip šiuolaikišką, įdomų ir patrauklų universitetą tarp būsimųjų studentų