LIETUVIŲ – ANGLŲ KALBŲ VERTIMO SISTEMA

Dovilė Milisevičiūtė, Egidijus Paliulis
Šiaulių universitetas, Technologijos fakultetas

Įvadas

Automatinio (mašininio) vertimo sistemos pasaulyje tyrinėjamos beveik pusšimtį metų. Sudarytais algoritmais remiasi dauguma šiuo metu pasaulyje naudojamų vertimo sistemų. Lietuvoje darbas šioje srityje yra tik kelerių metų senumo. Todėl tyrimų ir mokslinių darbų skaičius yra nedidelis, bet turėtų augti.

Automatinio vertimo sistemos (AVS) naudojamos tada, kai reikalingas greitas, nors ne visada kokybiškas ir tikslus vertimas. Kadangi kompiuterizuotas vertimas nėra tobulas, tad diskutuojama, ar išvis įmanoma padaryti tokią sistemą, kuri galėtų galvoti kaip žmogus ir tekstą suprastų ta pačia prasme, kaip tai supranta žmogus, t. y. turėti realiojo pasaulio žinių (Hutchins, 2000). Tyrimai atliekami įvairaus lygio kompanijose ir tyrimų centruose visame pasaulyje. Automatinis vertimas siejasi ir su dirbtinio intelekto kūrimu.

AVS kūrimui reikalingas kelių sričių išmanymas – kompiuterinės ir matematinės lingvistikos, vertimo teorijos, informacinių technologijų, kalbos filosofijos (Rimkutė, Kovalevskaitė, 2007).

Internetu pasiekiamų AVS skaičius yra daugiau nei 30. Tačiau jų veikimo principai nėra visiškai skirtingi, turi panašumų. Pvz., *Babelfish*, *Apple*, *Worldlingo*, *Yahoo!*, *Google* vertimo portalai yra palaikomi *Systran* (Dugast, Senellart, Koehn, 2007) vertimo sistemos. Interaktyvi rusų-anglų programa *MagicGooddie* yra senos rusų kompanijos *Promt* produktas. Viena pirmųjų AVS dirbančių su lietuvišku tekstu yra VDU projektas (*http://vertimas.vdu.lt/tws/*), sukurtas tos pačios *Promt* kompanijos sistemos pagrindu, kuris anksčiau buvo skirtas rusų-anglų kalbų krypties teksto vertimui.

Tikslas – sukurti eksperimentinę lietuvių – anglų kalbų vertimo sistemą.

Uždaviniai:

- Išnagrinėti egzistuojančius vertimo algoritmus ir sistemas.
- Remiantis atlikta analize, sukurti lietuvių – anglų kalbų vertimo algoritmą, jį realizuoti ir atlikti jo patikrinimą.

Egzistuojantys algoritmai

Egzistuoja trys pagrindiniai automatinio (mašininio) vertimo metodai: *statistinis, taisyklėmis besiremiantis* (rule-based) ir *loginis* (arba tiesioginis). Praktiškiausia integruoti kelių tipų metodus į AVS. Taisyklėmis paremtoje AVS didžiausią vaidmenį atlieka žodynai ir taisyklių rinkiniai. Taip pat reikia išspręsti morfologinį, sintaksinį ir semantinį daugiaprasmiškumą. Statistiniai metodai nepasižymi išskirtinai geresne vertimo kokybe, atvirkščiai – nors taisyklėmis paremtos AVS sudarymas yra komplikuotas ir daug sunkesnis, tačiau gaunama aukštesne vertimo kokybė (Daudaravičius, 2004).

Automatinį vertimą realizuoja algoritmas, kuris sudaromas atsižvelgiant į kalbų lingvistinius požymius ir taisykles. Šio algoritmo sudėtingumas priklauso nuo kalbų, kurių gramatika yra realizuojama, panašumo. Žodynai ir jų sudarymas yra labai svarbus etapas AVS kūrimo eigoje. Nuo žodynų priklauso ir vertimo kokybė, ir algoritmo sudarymas. Sistemai, verčiančiai iš lietuvių kalbos, reikalingas žodynas su lietuvių kalbos žodžiais, jų gramatine (morfologine ir sintaksine) ir semantine informacija.

Iš esmės visi automatinio vertimo algoritmai (procesas) susideda iš tokių trijų pagrindinių dalių:

1) gramatinis pirmos kalbos (PK) nagrinėjimas, t. y. įvesto teksto struktūros analizė, paremta PK gramatika;
2) perdavimas / transformavimas/keitimas (transfer), t. y. PK teksto struktūros transformavimas į antros kalbos (AK) teksto struktūrą;
3) AK teksto formavimas (generavimas), t. y. perdarymas AK teksto struktūros į specialią žodžių seką (Ya S. Filiatov).

Skirtingose AVS jų vykdymo eiliškumas gali būti nevienodas. Pavyzdžiui, po įvesties teksto (pirmos kalbos teksto) morfologinio ir sintaksinio analizavimo, gali eiti arba sintaksės modifikavimo arba vertimo (perdavimo) etapas.

PROMT vertimo sistema

Promt sistemos vertimo algoritmai paremti ne nuosekliomis transformavimo procedūromis, bet hierarchiniu pagrindu, kuriame vertimo procesai suskaidyti į susisiekiančias transformavimo procedūras skirtinguose analizės etapuose.

Sistemą galima išskirstyti į tokius lygmenis:

1. *Leksikos vienetų lygmuo.* Leksikinis vienetas yra žodis arba kolokacija, priklausantis žemiausiam

lygiui. Žodis aprašytas kaip kamieno ir galūnės kompozicija. Iš vienos pusės, tai leidžia atpažinti PK žodį ir morfologiškai jį nagrinėti, iš kitos pusės, patogus kelias nustatyti vertimui pagal susijusius morfologinius duomenis (kamienas, kaitymo tipas ir galūnės adresas tokios kaitybos tipo galūnių masyve). Taigi, jei galima pritaikyti keitimo taisykles iš PK morfologinių duomenų į AK morfologinius duomenis, tada galima vykdyti transformavimo procedūras morfologiniame lygmenyje.

2. *Grupių lygmuo.* Jis atsako už sudėtingesnes struktūras: daiktavardžių grupes, būdvardžius, prieveiksmius ir sudėtines veiksmažodžių formas. Šio lygmens pagrindas yra formalios ryšių gramatikos, ir analizės metu tai leidžia jungti grupes į sintaksinius vienetus. Kiekvieną vienetą charakterizuoja susintezuoti struktūriniai duomenys ir pagrindinis junginio vienetas.
3. *Paprasto sakinio lygmuo.* Paprastas sakinys traktuojamas kaip struktūra, susidedanti iš sintaksinių vienetų. Jo analizė vykdoma pagal karkasines tarinio struktūras. Paprastuose sakiniuose, pagrindinis elementas yra veiksmažodis, o jo junglumas (valentingumas) nulemia aktyvaus karkaso užpildymą.
4. *Sudėtinio sakinio lygmuo.* Analizė reikalinga, kai reikia suderinti laikus ir teisingai išversti jungtukus.

Šie procesai tarpusavyje siejasi pagal tekstinio vieneto hierarchiškumą, keičiasi susintezuotais ir paveldėtais atributais. Toks algoritmo sudarymas leidžia panaudoti formaliuosius metodus algoritmams aprašyti skirtinguose lygiuose (Hutchins, Somers, 2004).

SYSTRAN vertimo sistema

Iš pradžių sistema suprojektuota tiktai vertimui iš rusų kalbos į anglų kalbą, dabar apima 80 kalbų porų, verčia iš 22 kalbų ir yra daugelio didžiųjų portalų variklis. *Systran* sistema tradiciškai traktuojama kaip taisyklėmis pagrįsta sistema. Projektas pradėtas kurti daugiau nei prieš 30 metų. Pirmieji vartotojai – USA vyriausybinės organizacijos, o po to ir Europos ekonomikos sąjunga, kuri *Systran* pertvarkė ir gausios dokumentacijos vertimas tapo įmanomas į daugelį jai priklausančių šalių kalbų.

Systran toliau tobulinamas, įtraukiant ir statistinius metodus. Čia aprašomi formalieji metodai ir vertimo algoritmas, nesusijęs su statistiniais metodais.

Systran sistemos dizainas yra agreguotas ir aukšto modalumo. Jame yra dviejų tipų programos – a) sisteminės, parašytos asemblerio kodu, nepriklausančios nuo kalbos; jos, pavyzdžiui, atsakingos už žodyno peržiūros paprogrames; ir b) vertimo programos, sudarytos iš daug atskirų modulių. Vertimo programos skirtos analizei ir generavimui. Analizės modulis PK yra pastovus nepriklausomai nuo AK, o generavimo moduliai yra pastovūs AK, nesvarbu, kokia bebūtų PK.

Pagrindinė sistemos dalis – didžiulis dvikalbis žodynas, talpinantis leksikos ekvivalentus, gramatikos ir semantikos informaciją, vartojamą analizei ir generacijai. Didelė šios informacijos dalis yra algoritmų formos, jie iškviečiami įvairiuose vertimo proceso etapuose. Pagrindiniai vertimo procesai valdomi sudėtingo dvikalbio žodyno.

Aiškinantis *Systran* vertimo sistemos principus, būtina susipažinti su sisteminiais žodynais, kadangi vertimo procese glaudžiai bendraujama su žodynų įrašais.

Leksikos duomenų bazės *Systran* yra suskaidytos į pagrindinių kamienų žodyną (Main Stem), dvikalbį žodyną (iš vieno žodžio įrašų) ir įvairius daugiažodžius kontekstinius žodynus. *Pagrindinių kamienų žodyne* kiekvienam žodžiui parengtas pilnas morfologinis, sintaksinis ir semantinis aprašas: gramatinė kategorija, valdymas, valentingumas, derinamumas, tranzityvumas, daiktavardžio tipas (gyvas, skaičiuotinis, abstraktus), semantinis žymuo (fizikinė savybė, maisto produktas, talpa, įrankis); ir pagrindinės formos vertimas į AK ekvivalentą, kartu su gramatikos informacija, kurios reikia generacijai. Skirtumas sudaromas tarp homografų (vienodai rašomų žodžių) su skirtingomis gramatikos kategorijomis. Pažymėtina, kad vertimas grindžiamas sintakse, tai yra pirma užsiimama sintaksės problemomis, ir tik tada pasitelkiama semantinė informacija likusioms problemoms spręsti. Kiekvienam įvestam žodžiui duodamas tiktai vienas jį atitinkantis ekvivalentas – „numatytasis“ („default“) vertimas, kuris paliekamas, jei nebuvo pakeistas kitų žodynų; pvz., angliškas žodis STATION turi numatytą vertimą prancūziškai POSTE.

Kontekstiniai žodynai:

- *Idiomų žodynas*, skirtas nekintantiems posakiams (pvz., „on the other hand“) tvarkyti.
- *Ribotos semantikos žodynas* apibrėžia sintaksinių ryšių galimybes tarp daiktavardžio frazių, kada junginiai identifikuojami kaip leksikos vienetai, užtikrinant pastovų vertimą (pvz., „mashine translation“ į prancūzų kalbą būtų verčiamas „traduction automatique“, o ne „traduction de mashine“. Taip pat prancūsų „pomme de terre“ būtų verčiamas vienu žodžiu „potato“).
- *Analitiniai žodynai talpina* sintaksių taisyklių išimtis, kurios taikomos individualiems žodžiams. Pvz., angl. NOR (nor could he see the difficulties).

- *Sąlyginės semantikos žodynas* sujungia sintaksinę ir semantinę informaciją galimiems AK atitikmenims atskirti. Pvz., numatytasis vertimas angl. GROW pranc. yra GRANDIR, bet su „animate“ (gyvas) papildymu jis jau tampa ELEVER ir su „plant“ (augalas) kaip objektas tai yra CULTIVER.

Systran, sudarant įrašų seką („baitų plotus“), naudoja linijinę duomenų struktūrą, po vieną kiekvienam žodžiui sakinyje. Kiekvienas baitų plotas sudarytas iš paties žodžio ir gramatinės informacijos ir vertimo atitikmenų, susijusių su žodžiu žodyno įrašuose. Kiekvienas baitas talpina tam tikro tipo informaciją. Pvz., baitas 1 – pirminę žodžio kategoriją (daiktavardis, veiksmažodis, būdvardis), baitas 2 – asmuo ir veiksmažodžio skaičius, baitas 3 – daiktavardžio linksnis ar veiksmažodžio rūšis laikas, nuosaka, baitas 4 – giminė ir skaičius, ir t. t. Baitų reikšmės varijuoja priklausomai nuo kitų baitų verčių (pvz., baitas 3, priklausomai nuo baito 1, yra daiktavardis ar veiksmažodis). Vertimo proceso specifika:

1. *Etapas „Prieš apdorojimą“*
   1.1. Programa įkrauna tekstą ir atpažįsta įvesties informacijos formatą.
   1.2. Identifikuojamos pastovios formos ir kai kurios gramatinės kategorijos (pvz., „in order to“ kaip prielinksnis) („*Idiomų žodyno*“ peržiūra).
   1.3. Žodžiai ieškomi *Pagrindinių kamienų žodyne* ir informacija kopijuojama į baitų plotus duomenų struktūroje.
   1.4. Atliekama žodžio morfologinė analizė, atskiriamas kamienas ir galūnė, jei įmanoma ir identifikuojama informacija apie juos.
   1.5. Remiantis *Limituotos semantikos žodynu*, identifikuojami sudėtiniai daiktavardžiai. Elementai, esantys šiame žodyne, visada traktuojami kaip sudėtiniai. Todėl iškyla problemų, kai šie žodžiai atsitiktinai eina paeiliui (kartu).
2. *Kiekvieno sakinio analizės etapas:*
   2.1. Homografai perprantami tikrinant šalia esančių žodžių gramatines kategorijas (pvz., angliško žodžio „states“ tikrinama, ar jis yra veiksmažodis, ar daiktavardis). Neišsiaiškinus tikros homografo reikšmės, paliekama pati tikėtiniausia.
   2.2. Sakiniai segmentuojami į pagrindinius ir šalutinius, atliekant skyrybos ženklų, jungtukų, santykinių įvardžių (kuris) paiešką t. t.
   2.3. Nustatomi „pirminiai sintaksiniai ryšiai“ tarp daiktavardžių ir būdvardžių, prielinksnių ir daiktavardžio frazių. Šiame etape taip pat nustatomas pagrindinis veiksmažodis, laikas, neiginys, laipsniai.
   2.4. Nustatomi susietieji žodžiai. Pavyzdžiai „Smog and pollution control *are* important factors“ ir „Smog and pollution control *is* under consideration“ rodo, kaip žodžiai „is“ ir „are“ padeda nustatyti žodžio „smog“ ir frazės „pollution control“ struktūrinius ryšius. Kitas pavyzdys: „zinc and aluminium components“ semantinis žymuo „chemical elements“ parodo, jog „zinc“ ir „aluminium“ yra vienarūšės sakinio dalys.
   2.5. Identifikuojami veiksnys ir tarinys. Anksčiau nustatyti asmeniniai veiksmažodžiai yra galimi tariniai, o daiktavardžiai (ar įvardžiai), dar anksčiau neidentifikuoti kaip „objektai“, yra galimi veiksniai.
   2.6. Identifikuojami ryšiai tarp tarinių ir argumentų. Jei reikia, tikrinamas analitinis žodynas.
3. *Transformavimo etapas:*
   3.1. Standartinės idiomos ir fiksuotos frazės surandamos analizės etape vartojant *Idiomų* ir *Ribotos semantikos* žodynus. Šiame etape pagal sąlygas, aprašytas *Sąlyginių semantikų žodyne*, transformuojamos leksikos. Pvz., homografas LEAD turi būti verčiamas kaip PLOMB, kai nustatoma, jog tai „cheminis elementas“.
   3.2. Prielinksnių vertimas, kuris dar neatliktas prieš tai buvusiuose etapuose.
   3.3. Struktūrinis transformavimas pagal žodyne nurodytus testus tam tikriems žodžiams ar sintaksės ir semantikos žodžių kategorijoms. Pvz., tinkamas vertimo parinkimas žodžiui angl. „as“ į prancūzų kalbą (comme, pendent que, a mesure que, puique).
4. *Sintezės etapas:*
   4.1. Numatytojo vertimo (default) reikšmės priskyrimas iš *Pagrindinio kamienų žodyno* tiems žodžiams, kurie liko neišversti.
   4.2. Morfologinė generacija, remiantis struktūrine informacija apie giminę, skaičių, laiką, t. t. iš ankstesniųjų etapų ir remiantis informacija (iš Main Stem žodyno) apie galūnes (kaitymą) ir priklausomybes.
   4.3. AK sakinio žodžių tvarkos generavimas. Pvz., perdarant žodžio tvarką į angliškąjį būdvardis-daiktavardis seką (Hutchins, Somers, 2004).

Eksperimentinė lietuvių – anglų vertimo sistema

Lietuvių – anglų vertimo sistemai sudaryti turime du pavyzdžius – VDU vertimas anglų – lietuvių ir „*Google Translator*“ lietuvių – anglų ir anglų – lietuvių. Šios sistemos veikia skirtingu principu: pirmoji yra „*tranfer-based*“ tipo sistema, o antroji – „*statistical-based*“ sistema. Eksperimentinė sistema at-

stovaus tiesioginio automatinio (direct) vertimo sistemos tipui, kuris yra vienas paprasčiausių metodų.

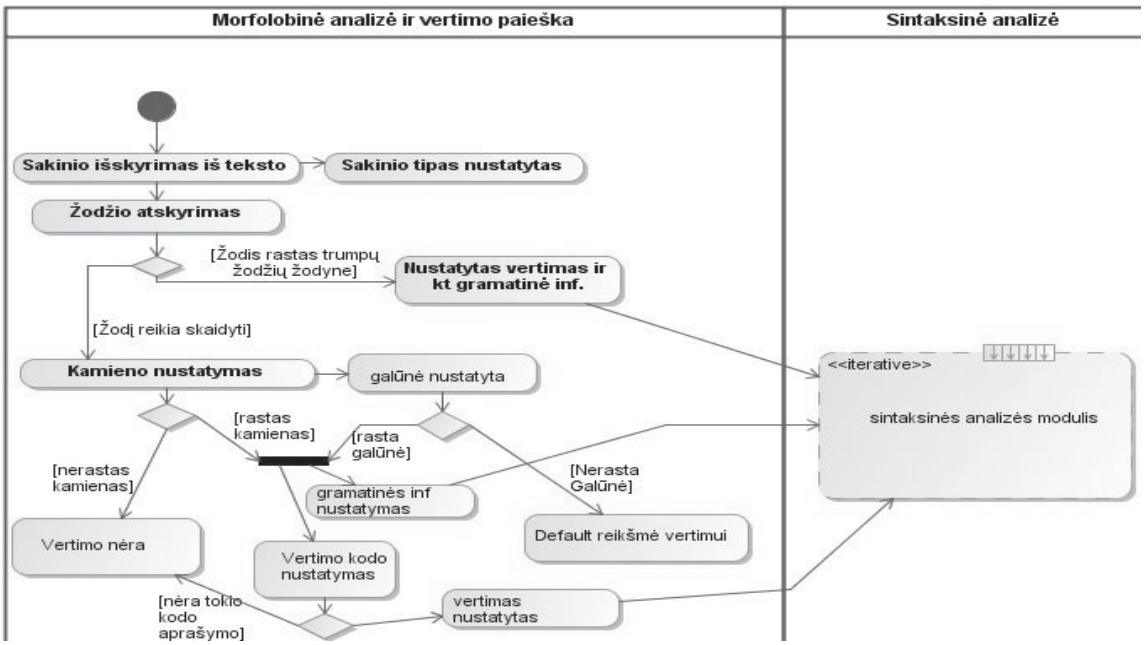
Sudarant naują algoritmą lietuvių – anglų porai, reikia atlikti detalią lietuvių kalbos analizę, susisteminti lietuvių kalbos gramatiką – sudaryti specialų sisteminį žodyną ir sukurti taisykles algoritmui, kuris atlieka paiešką duomenų bazėje bei suranda sintaksinius žodžių ryšius sakinyje.

Šiame darbe bandoma sukurti algoritmą lietuvių – anglų teksto vertimui, vartojant programines technologijas PHP (algoritmui aprašyti), duomenų bazę MySQL (lingvistei informacijai laikyti).

Kadangi lietuvių ir anglų kalbų gramatikos turi tas pačias kalbos ir sakinio dalis, laikyta, kad lietuviško sakinio žodis, einantis tam tikra sakinio dalimi, eina ta pačia sakinio dalimi ir anglų kalboje (dažniausiai taip būna ir tikrovėje). Todėl, konstruojant naują anglišką sakinį, reikia nustatyti lietuviško žodžio vaidmenį (kuria sakinio dalimi eina) lietuviškame sakinyje. Dažniausiai tokią informaciją teikia kalbos dalis (veiksmažodžiai beveik visada eis sakinyje tariniais) ir galūnės morfologinė informacija (linksnis dažniausiai apsprendžia sakinio dalį).

Žodynas ir leksikos duomenų bazė

Pagal lietuvių kalbos gramatikos taisykles žodyno bazėje pakanka saugoti žodžio kamieno požymius, o pagal linksniuotėms, giminei, asmenuotėms priimtas taisykles ir kitus vertimui naudingus atributus, atitinkamas žodžio formas galima konstruoti procedūriniu būdu, t. y. prijungiant atitinkamas žodžio dalis: galūnes, priesagas ir galūnes (Sipavičius, Tamulynas, 2004). Tad talpinome duomenų bazės lentelėse atskirai lietuviškų žodžių kamienus ir galūnes. Kamienai talpina tam tikrą gramatinę (pvz., kalbos dalį, linksniuotę / asmenuotę) ir leksinę (pvz., pagrindinę angliško atitikmens formą) informaciją. Pagrindinė angliškojo atitikmens forma saugoma prie kamieno, bet tikrasis vertimas priklauso ir nuo galūnės gramatinių kategorijų reikšmių.



1 pav. Algoritmo „Prieš apdorojimą“ ir morfologinės analizės veikimo schema

MySQL duomenų bazėje saugomi lietuviški kamienai, lietuviškos žodžių galūnės (atskirai pagal kalbos dalis), trumpieji arba dažnai vartojami lietuviški žodžiai ir angliškų atitikmenų lentelės. Visos lietuviškos žodžio leksemos turi atributus, saugančius individualią informaciją apie save, kuri reikalinga angliškajam atitikmeniui generuoti bei vėlesnei sintaksinei sakinio modifikacijai (į anglų kalbos taisyklėmis grįstą sakinį) vykdyti. Lietuviškieji kamienai ir galūnės turi unikalius kodus (ID), kurie reikalingi angliškam vertimui surasti vertimų lentelėje.

Eksperimentinės vertimo sistemos algoritmas

1. „Prieš apdorojimas“ (kiekvienam sakiniui):
   1.1. Pirminis teksto apdorojimas (sakinio atskyrimas iš teksto ir kartu jo tipo nustatymas pagal sakinio pabaigos ženklą).
   1.2. Žodžių atskyrimas pagal tarpo žymę ir kitus skyrybos ženklus (kablelį, brūkšnį ir t. t.).
2. Morfologinė analizė (kiekvienam sakinio žodžiui):
   2.1. Žodžių paieška trumpųjų žodžių lentelėje arba žodžio kamieno ir galūnės nustatymas ir jų gramatinės informacijos surinkimas iš ka-

mienų ir galūnių lentelių MySQL duomenų bazėje.

2.3. Žodžio angliško vertimo kodo generavimas. Vertimų lentelėje saugomas raktinis vertimo identifikavimo kodas, kuris generuojamas iš individualaus kamieno ID ir galūnės ID. Vienas žodis gali turėti vieną, bet gali turėti ir daugiau sugeneruotų vertimo kodų rinkinių. Tarkim, kamieno „lauž“ raktinis laukas ID = „5“ (kai kalbos_d = „daiktavardis“) ir ID = „10“ (kai kalbos_d = „veiksmažodis“). Galūnė „-o“ turi net 6 skirtingus ID, kadangi kiekvienas atskiras ID talpina skirtingą galūnės gramatinę informaciją. Algoritmas sugeneruoja vertimo kodų rinkinius kiekvienam kamieno ID su kiekvienu galūnės ID, todėl gaunami 2 x 6 = 12 ID kodų rinkinių. Paieška vertimų lentelėje vykdoma pagal sugeneruotą kodų rinkinį. Jei toks kodų rinkinys aprašytas vertimų lentelėje, grąžinamas žodžio vertimas. Jei bus aprašyti visi galimi kodų rinkiniai, tai šituo atveju grąžins 12 žodžio vertimų variantų. Tačiau realybėje vertimų lentelėje yra aprašyti tik 3 kodų rinkiniai, kurie grąžina tokius žodžio vertimus: 1) *of the fireplace*; 2) *break*; 3) *breaks*. Reikšmės pagal nutylėjimą (default) įvedimas: pagrindinėje duomenų bazės lentelėje *LT_KAMIENAI* šalia kamieno įvedama numatomoji kamieno reikšmė anglų kalboje. Administratorius turi parinkti kamienui tokią angliską žodžio formą, kurią jis turi omenyje, jog lietuviškas kamienas turi būti verčiamas. Pvz., kamieno „sūn-“ numatoma reikšmė bus „son“, kadangi administratorius turi omenyje lietuviškąjį „sūnus“. Vertimo proceso metu neradus žodžio „sūnus“ kurios nors formos (neaprašytos vertimų lentelėje), sistema vertimu laikys šią numatytąją reikšmę. Tokiu atveju nenustatomos gramatinės žodžio kategorijų reikšmės. Pateikiamas preliminarus vieno žodžio vertimas, taigi, nenustatoma, kuria sakinio dalimi jis eina sakinyje, todėl sintaksinė sakinio modifikacija negalima (paliekamas pažodinis vertimas).

3. Sintaksinės sakinio struktūros keitimas:

3.1. Šiame žingsnyje gali prireikti vartotojo įsikišimo, kai žodžiui rastas daugiau nei vienas vertimas. Vartotojas nurodo reikiamą vertimo variantą ir sistema tęsia darbą.

3.2. Surandamos galimai derinamosios sakinio dalys (būdvardis ir daiktavardis), prielinksniniai junginiai (prielinksnis ir daiktavardžio linksnis) per vieną žingsnį nuo analizuojamojo žodžio (tikrinamas po jo einantis žodis; tęsiama realizacija).

3.3. Naujas sakinys sudaromas iš gautų vertimų, sudėliojant juos tokia žodžių tvarka, kokios pagal griežtas anglų kalbos žodžių tvarkos sakinyje taisykles reikalauja tam tikras sakinio tipas.

Testavimas

Gautų vertimų lietuvių – anglų kryptimi pavyzdžiai ir palyginimas su statistinio tipo algoritmo pagrindu veikiančia kol kas vienintele automatinio vertimo sistema internete *Google Translator*, verčiančia ta pačia lietuvių – anglų kryptimi:

1 lentelė. ***Eksperimentinio vertiklio ir* Google Translator *teksto vertimų palyginimas***

Originalus tekstas	**Eksperimentinis algoritmas (tiesioginio vertimo metodas)**	***Google Translator* (statistinis vertimo metodas)**
Automatinio vertimo sistemomis naudojamasi tada, kai reikalingas greitas vertimas	*Automatic translation system* naudojamasi then , when *necessary fast translation*	Automatic translation systems for use when a translation is needed fast
Žmonės turi vienintelę priemonę apsiginti – tai valstybė	The people has / have vienintelę device apsiginti – that (is) the commonwealth	People have a single measure to protect themselves – to the country
Mėgsta kepti dešreles ant laužo	Likes to bake the sausages on break / breaks / (of the) fireplace	Love fried sausages on a campfire

Pateiktoje lentelėje pasviręs tekstas reiškia, kad vertimas preliminarus (naudojama “default” reikšmė pagal kamieno informaciją). Pasvirusiais brūkšniais skiriami daugiaprasmiško žodžio vertimo variantai.

Algoritmo privalumai:

- Randa visus įmanomus žodžio vertimo variantus;
- Interaktyvumas – minimalus vartotojo įsikišimas į vertimo procesą, pasirenkant vertimo variantą iš kelių galimų, kai tokių atsiranda.

Algoritmo trūkumai:

- Artikeliai dedami neatsižvelgiant į tikslią jiems skirtą vietą (pvz., (of the) man (of the) head).
- Dar nebaigtas realizuoti sintaksės modifikavimo etapas, tad kol kas atliekamas tik pažodinis teksto vertimas. Kai AK yra anglų kalba, būti-

na realizuoti šį etapą, kadangi dažniausiai nuo žodžių tvarkos angliškame sakinyje priklauso jo prasmė.
- Aprašytos ne visos galimos lietuviškų žodžių darybos galūnės (nėra išvestinių veiksmažodžio formų priesaginių galūnių, t. y. dalyvių, pusdalyvių, padalyvių, ir skaitvardžių).

Išvados

1. Lietuvių – anglų algoritmui sukurti pirma reikia gerai apgalvoti žodyno (leksinės bazės) sudėtį ir apimtis. Tam reikia įsigilinti į lietuvių kalbos gramatikos ir leksikos ypatumus.
2. Kalbos vertimo algoritmas tuo lengvesnis, kuo aprašomoji algoritmo kalbų pora yra panašesnė. Lietuvių ir anglų kalbos turi panašumų (tos pačios kalbos ir sakinio dalys, sakinių tipai), tačiau nemažiau ir skirtumų. Angliški žodžiai beveik nekaitomi, o lietuviškųjų kaitymas yra ypač platus. Anglų kalboje vartojami artikeliai.
3. Automatinio vertimo algoritmo kūrimas su pasirinktomis programinėmis technologijomis pasiteisino. MySQL duomenų bazė gali būti ir toliau pildoma naujais įrašais, o į PHP kodą galima įterpti papildomas funkcijas ir išplėsti tikrinimo, sąlygų ir kitas taisykles.
4. Sudarytas algoritmas nepretenduoja į aukštą vertimo sistemos lygį. Jis reprezentuoja paprasčiausio galimo sudaryti lietuvių – anglų kalbų porai algoritmą ir susidaryti aiškesnį vaizdą apie tokios sistemos kūrimo sudėtingumą. Yra daug būdų sistemai toliau tobulinti, pavyzdžiui, semantinės informacijos tikrinimo taisyklių suprogramavimas.
5. Eksperimentinio algoritmo vertimo kokybė tuo geresnė, kuo daugiau žodžių yra žodyne ir kuo daugiau vertimų aprašyta vertimų lentelėje.
6. Dėl rusų ir lietuvių kalbų lingvistinių panašumų jau sukurtus modelius, skirtus rusų – anglų kalbų porai, įmanoma pertvarkyti ir pritaikyti lietuvių – anglų kalbų porai. Būtina sumodeliuoti lietuvių kalbą, tai bus analogiškas procesas rusų kalbos modeliavimui.
7. Kai kuriuos *Promt* ir *Systran* architektūros ir vertimo principus galima įžvelgti sudarytame lietuvių – anglų kalbai algoritme, kurie būdingi visoms automatinio vertimo sistemoms etapas prieš apdorojimą, morfologinė ir sintaksinė analizė, žodžių junginių sakinyje paieška, naujo išvesties teksto generavimas.
8. Sudarytas eksperimentinis algoritmas teisingai išverčia tekstą tada, kai įvesties tekstas yra taisyklingas, o daugiaprasmiškumą išsprendžia žmogaus įsikišimas. Tokią sistemą galima naudoti tik eksperimentiškai. Norint kad ji atitiktų aukštesnius vertimo kriterijus, reikalinga praplėsti žodyno informaciją apie leksinius vienetus ir įvesti semantinės žodžių informacijos įvertinimo / tikrinimo taisykles.

Literatūra

1. Daudaravičius V., 2006, Pradžia į begalybę. Mašininis vertimas ir lietuvių kalba. *Darbai ir dienos: Pažangos šuoliai.* Nr. 45. P. 7–18. <http://donelaitis.vdu.lt/publikacijos/dd45_vidas.pdf>.
2. Dugast L., Senellart J., Koehn P., 2007, Statistical Post-Editing on SYSTRAN's Rule-Based Translation System: *Proceedings of the Secont Workshop on Statistical Machine Translation*. P. 220–223. Praha: Association of Computational Linguistics.
3. Hutchins W. J., 2000, Mashine Translation: *Encyclopedia of literary translation into English*. P. 884–885. London: Fitzroy Dearborn Publishers.
4. Hutchins W. J., Somers H. L., 2004, *An introduction to machine translation.* Chapter 10, Cambridge: University Press.
5. Ya S. Fitialov, Automatic translation. <http://eom.springer.de/A/a014110.htm>.
6. Rimkutė E., Kovalevskaitė J., 2007, Mašininis vertimas – greitoji pagalba globalėjančiam pasauliui, *Gimtoji kalba.* Nr. 9. P. 3–11. <http://www.apiekalba.lt/index.php?option=com_content&task=view&id=41>.
7. Sipavičius A., Tamulynas B., 2004, Valdomojo kompiuterinio vertimo technologinių galimybių tyrimas. <http://oras.if.ktu.lt/moduliai/p175m007/grafika/public_html/2004_konferencija/Straipsniai_2004/>.
8. Sokolova S., 2005, How the computer translates. <http://www.promt.com/company/technology/overview/>.

LITHUANIAN-ENGLISH LANGUAGE TRANSLATION SYSTEM

Dovilė Milisevičiūtė, Egidijus Paliulis

Summary

The article briefly describes the existing Systran and Promt machine translation systems and principles of their operation. Choice of such systems was due to the fact that many of the automated translation systems on the Internet and commercial translation programs operate under these two systems, therefore, we believed that their algorithms are among

those of the most advanced ones. It is difficult to say which type of automatic translation method both systems represent. Systran previously was considered to be a representative of the direct method, and now it is a representative of the rules-based approach. Implementation of new technologies is now changing this approach again. Promt is considered to be a transfer type of machine translation system; however, it incorporates new technologies as well. The article deals with the abstract principles of these systems.

The presentation of the experimental system of Lithuanian English language pair translation. Algorithm development was on the impact on existing systems models. When developing algorithms of automated language translation systems, at first language grammar needs to be analyzed in detail and turned into a programming algorithm. A large dictionary creation and filling, which affects the quality of translation, requires a lot of work and linguistic information. The development of the direct type and the type of rules-based translation systems is to program the close cooperation between algorithm programs and dictionaries.

Keywords: language translation, translation algorithm.

LIETUVIŲ – ANGLŲ KALBŲ VERTIMO SISTEMA

Dovilė Milisevičiūtė, Egidijus Paliulis

Santrauka

Straipsnyje trumpai apibūdinamos esamos Systran ir Promt mašininio vertimo sistemos bei jų veikimo principai. Tokios sistemos pasirinktos todėl, kad daugelio automatizuotų vertimo sistemų internete bei komercinių teksto vertimo programų veikimas pagrįstas šiomis dviem sistemomis, todėl manoma, kad jų algoritmai yra vieni pažangiausių. Sunku pasakyti, kurį automatinio vertimo metodo tipą atstovauja abi sistemos. Systran anksčiau laikyta tiesioginio metodo atstove, o dabar ji yra taisyklėmis grįsto požiūrio atstovė. Naujų technologijų diegimas dabar vėl keičia šį požiūrį. Promt laikoma perkėlimo tipo mašininio vertimo sistema, tačiau joje taip pat yra naujų technologijų. Straipsnyje nagrinėjami abstraktūs šių sistemų principai.

Bandomosios lietuvių – anglų kalbų poros vertimo sistemos pristatymas. Algoritmų kūrimui įtaką daro esamų sistemų modeliai. Kuriant automatizuotų kalbos vertimo sistemų algoritmus, pirmiausia reikia išsamiai išanalizuoti kalbos gramatiką ir paversti ją programavimo algoritmu. Daug darbo ir kalbinės informacijos reikalauja didelio žodyno kūrimas ir pildymas, kuris daro įtaką vertimo kokybei. Tiesioginio ir taisyklėmis pagrįsto vertimo sistemų plėtojimo esmė – užprogramuoti glaudžią sąveiką tarp algoritmų programų ir žodynų.

Prasminiai žodžiai: kalbos vertimas, vertimo algoritmas.

Įteikta 2009-05-04