BALTISTICA VIII PRIEDAS 2012 155–168

Inguna SKADIŅA, Madars VIRZA,
Lauma PRETKALNIŅA
LU Matemātikas un informātikas institūts

ANGĻU-LATVIEŠU STATISTISKĀS MAŠĪNTULKOŠANAS SISTĒMAS IZVEIDE: METODES, RESURSI UN PIRMIE REZULTĀTI

Ievads

Viens no pirmajiem uzdevumiem, kurš tika uzticēts datoriem, bija tekstu automatizēta tulkošana. Kopš pirmo tulkošanas sistēmu izveides pagājis vairāk nekā sešdesmit gadu. Šajā laikā tulkošanas sistēmu izveidē tikušas izmantotas dažādas stratēģijas. 20. gadsimta beigās galvenokārt tika pētītas un izstrādātas likumos balstītās tulkošanas sistēmas, izmantojot transfēra vai interlingvas stratēģijas (Hutchins 2007). Šādās sistēmās katras valodas specifika un zināšanas tulkošanai starp valodu pāriem tiek aprakstītas mašīnlasāmu likumu un zinību bāzu formā.

Jauns pavērsiens mašīntulkošanā iezīmējās 90. gadu sākumā, kad IBM radīja pirmo statistiskās mašīntulkošanas (SMT) sistēmu *Candide* (Berger 1994). Mūsdienās statistisko metožu lietojums automatizētās tulkošanas sistēmās kļuvis par dominējošo pētījumu virzienu pasaulē. Atšķirībā no likumos balstītām sistēmām, kur tulkošanas likumus un valodas specifiku apraksta cilvēks, SMT sistēmas tulkojumus „iemācās" no iepriekš pārtulkotiem tekstiem, kas dēvējami arī par paralēlo tekstu korpusiem.

Līdzīga ideja kā SMT sistēmās tiek izmantota arī tulkošanas atmiņās. Taču atšķirībā no tulkošanas atmiņas, kura var pārtulkot tikai tos teikumus vai teksta fragmentus, kas bijuši iepriekš pārtulkoti, statistiskā tulkošanas sistēma var pārtulkot teikumus, kuru vārdkopas vai vārdi ir bijuši paralēlo tekstu korpusā, no kura dators ir „iemācījies" tulkojumus.

Viens no būtiskiem priekšnoteikumiem SMT sistēmu izveidē ir daudzu miljonu vārdlietojumu lielu paralēlo tekstu korpusu pieejamība. Tāpēc sākotnēji SMT sistēmas tika izstrādātas valodām, kam ir uzkrāti lieli paralēlo tekstu korpusi. Latviešu valodai mašīntulkošanas pētījumu uzsākšanai nepieciešamais paralēlo tekstu apjoms elektroniskā formā ir pieejams kopš 2005. gada, kad

Eiropas Komisijas Apvienotais Zinātņu centrs (*Joint Research Centre*) publiskoja atlasītus ES tiesību aktus ES 20 valodās, tai skaitā latviešu valodā, tā saucamo JRC Acquis korpusu (Steinberger u. c. 2006). Tomēr šis korpuss ir salīdzinoši neliels ar citām valodām esošajiem paralēlajiem korpusiem un vēl aizvien paralēlo tekstu trūkums ir viens no būtiskākajiem ierobežojumiem vispārīgas kvalitatīvas SMT sistēmas izstrādei tulkošanai latviešu valodā.

Latvijā pētījumi statistiskajā mašīntulkošanā uzsākti 2005. gadā LZP projektā *Statistisko metožu izvērtējums angļu – latviešu tulkošanas sistēmā*. Kopš 2009. gada pētījumi tiek turpināti LZP projektā *Faktorēto metožu lietojums angļu-latviešu statistiskajā mašīntulkošanas sistēmā*. Jau vairākus gadus latviešu valoda ir ietverta uzņēmuma *Google* piedāvātajā SMT sistēmā *Google tulkotājs*[1]. Kopš 2009. gada SMT sistēmu izveide notiek arī sabiedrībā *Tilde*[2] (Skadiņš u. c. 2010), sadarbībā ar *Tildi* izstrādāta arī angļu-latviešu-angļu tulkošanas sistēma uzņēmuma *Microsoft* tulkotājā *Bing Translator*[3].

Raksta mērķis ir iepazīstināt ar pētījumiem, kas veikti Latvijas Universitāte Matemātikas un informātikas institūtā (LU MII), izstrādājot angļu-latviešu SMT sistēmas prototipu iepriekš minēto LZP projektu ietvaros, iegūtajiem rezultātiem, apzinātajām problēmām un nākotnes iecerēm. Rakstā analizētas latviešu valodai specifiskās problēmas un atrastie risinājumi statistiskajai mašīntulkošanai. Kaut arī rakstā sīkāk iztirzāti LU MII veiktie pētījumi, izstrādājot angļu-latviešu SMT sistēmas prototipu, pētījuma rezultāti attiecināmi arī uz citām SMT sistēmām, kas tulko latviešu valodā.

1. SMT sistēmu uzbūves pamatprincipi

Tradicionāli SMT sistēmas veido trīs komponenti: tulkošanas modelis, valodas modelis un dekoders. Tulkošanas modelis ir datora veidota vārdnīca, kuru dators „iemācās" no paralēlo tekstu korpusiem. Šādu vārdnīcu veido vārds vai vārdu virkne avotvalodā un tam atbilstošais vārds vai vārdu virkne mērķvalodā un tulkojuma ticamība (varbūtība). 1. attēlā parādīts vienkāršots tulkošanas modeļa fragments vārdam *legislative* angļu valodā. Kā redzams, par ticamākajiem tulkojumiem dators izvēlējies tulkojumus *tiesību aktu projektu* (varbūtība 0,25) un *tiesību akta* (varbūtība 0,222222), bet par vismazāk ticamu – vārdu *aktu* (varbūtība 0,00266667)

[1] http://translate.google.com
[2] http://translate.tilde.com
[3] http://www.microsofttranslator.com

1. attēls. **Vienkāršots SMT sistēmas tulkošanas modeļa fragments**

```
legislative ||| juridiskais ||| 0.0322581
legislative ||| juridiskas ||| 0.010101
legislative ||| juridiski ||| 0.0103093
legislative ||| aktu ||| 0.00266667
legislative ||| tiesību ||| 0.00725953
legislative ||| tiesību akta ||| 0.222222
legislative ||| tiesību akta projektu ||| 0.25
```

Valodas modelī tiek apkopota vārdu un vārdu virkņu biežuma statistika, kas iegūta no mērķvalodas tekstu korpusiem (skat. 2. attēlu). Tradicionāli valodas modeļi iekļauj vārdu virknes, kas sastāv no 1 līdz 3 vai līdz 5, reizēm līdz 7 vārdiem. Garākas vārdu virknes parasti valodas modelī neiekļauj, jo to biežums nav statistiski nozīmīgs.

2. attēls. **Valodas modeļa fragments**

```
0.000400858    visas procedūras
0.000248752    visi procedūras
0.000415881    visi procenti
0.000395173    visi projekti
0.000981681    ņemot vērā komisijai
0.0009327733   ņemot vērā komisijas
```

Gan tulkošanas modeli, gan valodas modeli dators automātiski izgūst no tekstiem. Tādējādi modeļos iekļautās vārdu virknes ne vienmēr atbilst vārdkopām vai tulkojošas vārdnīcas šķirklim. Piemēram, 2. attēlā minētā vārdu virkne *visi procedūras* neatbilst vārdkopai latviešu valodā, tomēr tekstā vārdi *visi* un *procedūras* var atrasties blakus, piemēram, *visi procedūras veicēji*, tāpēc dators šo vārdu virkni ir iekļāvis valodas modelī.

SMT sistēmā tulkošanas modeļa uzdevums ir ar statistisko līdzekļu palīdzību atspoguļot vārda, vārdu virknes tulkojuma iespējamību, savukārt valodas modeļa uzdevums ir atspoguļot vārdu kārtas ticamību un nodrošināt pareizā tulkojuma izvēli kontekstā. SMT sistēmas trešais komponents ir dekoders. Tā uzdevums ir atrast statistiski labāko tulkojumu starp visiem iespējamajiem tulkojumiem, izmantojot tulkošanas modelī un valodas modelī apkopoto informāciju. Tā kā tiek pieņemts, ka tulkošanas modelī lielāka varbūtība būs vārdu un vārdu virkņu pāriem, kas ir viens otra tulkojums, bet valodas modelī

lielāka varbūtība būs vārdu virknēm, kas valodā ir biežāk sastopamas (atbilst pareizai vārdu kārtai mērķvalodā), tad dekodera uzdevums ir starp visiem iespējamajiem datora tulkojumiem atrast visvarbūtiskāko, t. i., atrast teikumu, kam ir $argmax_e\ p(e|l)\ p(l)$, kur p(e|l) ir tulkošanas modeļa varbūtība un p(l) ir valodas modeļa varbūtība.

2. Pētījumā izmantotā infrastruktūra

Rakstā aprakstīto angļu-latviešu SMT sistēmu prototipu izveidei izmantoti vairāki pasaulē plaši pazīstami rīki, kurus līdzīgu sistēmu izveidei izmanto gan universitātes un pētniecības institūti, gan arī uzņēmumi. Tulkošanas modeļa izveidei tika izmantots GIZA++ rīks (Och un Ney 2003), valodas modelis veidots ar SRLIM rīku (Stolcke 2002), bet tulkošanas procesa nodrošināšanai izmantots Moses dekoderis (Koehn u. c. 2007).

SMT sistēmu prototipu tulkošanas un valodas modeļa izveidei izmantots ES juridisko dokumentu daudzvalodu paralēlais korpuss JRC Acquis (versijas 2.2. un 3.0) un Acquis Communautaire daudzvalodu tulkošanas atmiņas (DGT TM)[4]. Šie teksti tika izvēlēti lielā apjoma dēļ (skat. 1. tabulu), kas ir svarīgs priekšnoteikums SMT sistēmas izveidei. Sagatavojot datus, izlases veidā no apmācības korpusa tika nodalīti teikumi testa korpusam.

SMT sistēmu novērtēšanai tika lietota BLEU metrika (Papineni u. c. 2002). BLEU metrika novērtē tulkojuma kvalitāti, salīdzinot datora tulkojumu ar cilvēku tulkojumu un aprēķinot tulkojumu līdzību robežās no 0 līdz 100. BLEU metrika bieži tiek kritizēta par ne vienmēr adekvāto vērtējumu, tomēr tā vēl aizvien ir visvairāk lietotā automātiskā novērtēšanas metrika SMT.

3. Frāzēs balstītā SMT

Pirmais angļu-latviešu SMT sistēmas prototips (Skadiņa, Brālītis 2008) tika izveidots, izmantojot frāzēs balstīto stratēģiju. Frāzēs balstītā stratēģija (Koehn u. c. 2003) ir viena no populārākajām SMT stratēģijām, kas kā vienīgo zināšanu avotu izmanto paralēlo tekstu korpusus tulkošanas modelim un mērķvalodas tekstu korpusus valodas modelim, kā tas aprakstīts 1. nodaļā.

3.1. Rezultātu kvantitatīvais raksturojums

Sistēmas apmācīšanai sākumā tika izmantota JRC Acquis korpusa versija 2.2, bet pēc JRC Acquis korpusa versijas 3.0 parādīšanās arī šis korpuss un DGT TM tulkošanas atmiņa. Sistēmu novērtējums BLEU punktos apkopots 1. tabulā.

[4] http://langtech.jrc.it/DGT-TM.html

1. tabula. **Angļu–latviešu frāzēs balstītā SMT prototipa novērtējums ar BLEU metriku**

Apmācībai izmantotais korpuss	Korpusa lielums (milj. vārdlietojumu)		BLEU punkti		
	angliski	latviski	JRC Acquis 2.2 testa korpuss	JRC Acquis 3.0 testa korpuss	Nejauši izvēlēts (JRC Acquis 3.0)
JRC Acquis 2.2	7, 51	5,66	**39,53**	27,06	37,79[6]
JRC Acquis 3.0	34, 59	27,59	41,51	**43,28**	39,25
DGT TM	2,19 (vienību)[5]	1,12 (vienību)[6]	48,42[7]	31,86	**43,73**

Lai arī BLEU punkti tikai daļēji atspoguļo sistēmu kvalitāti, iegūtie rezultāti ļauj izdarīt vairākus secinājumus. Pirmkārt, SMT sistēmu tulkojumu kvalitāte ir cieši saistīta ar paralēlā tekstu korpusa lielumu – jo lielāks ir paralēlo tekstu korpuss, jo sistēmā ir daudzveidīgāks vārdu krājums un iegūtie tulkošanas un valodas modeļi patiesāk atspoguļo reālo valodu, nodrošinot kvalitatīvāku tulkojumu. Taču iegūtais uzlabojums nav tik liels, kā būtu sagaidāms – kaut arī JRC Acquis versija 2.2 ir gandrīz 5 reizes mazāka par JRC 3.0 versiju, rezultāts, testējot uz JRC Acquis 2.2 testa korpusa, atšķiras tikai par nepilniem 2 BLEU punktiem.

Otrs secinājums attiecināms uz iegūto rezultātu ticamību. Lai kādai no sistēmām nebūtu priekšrocību, testa korpusam vienādā mērā jāatspoguļo dažādo sistēmu dati vai arī jāizvēlas līdzsvarots testa korpuss. JRC 2.2 testa korpusu veido katra 200. rindiņa no JRC 2.2. un JRC 3.0 korpusu kopējās daļas, bet JRC 3.0 testa korpusu veido nejauši izvēlēti teikumi no tās JRC 3.0 korpusa daļas, kas neietilpst JRC2.2. Tādējādi otrajā gadījumā sistēma, kas trenēta uz šī korpusa datiem, uzrāda vislabāko testa rezultātu – 43,28 BLEU punktus.

Visbeidzot ne mazāk svarīga ir SMT sistēmas izveidē izmantoto datu kvalitāte. Abi JRC Acquis korpusi ir sastatīti, izmantojot automātiskās metodes, t. i., ietver kļūdainu sastatījumu. Savukārt DGT korpuss iegūts no tulkošanas atmiņām un tajā nav sastatījuma kļūdu. Kā redzams no 1. tabulas, tad gadījumos, kad testa korpuss ir vienlīdz atbilstošs visām sistēmām, visvairāk BLEU punktu iegūst sistēma, kas trenēta uz tulkošanas atmiņas datiem.

[5] Teikumi, teikuma daļas un rindkopas.

[6] Teikumi, teikuma daļas un rindkopas.

[7] Testa korpuss varētu saturēt apmācības korpusa datus.

3.2. Rezultātu kvalitatīvais raksturojums

Frāzēs balstītās SMT sistēmas ir uzrādījušas labus rezultātus analītiskām valodām, kurām nav tik bagātīga morfoloģija, bet ir uzkrāti lieli paralēlo tekstu korpusi. Arī izstrādātais angļu-latviešu sistēmas prototips uzrāda labus rezultātus atsevišķu teikumu tulkojumos. Piemēram, teikumu *I would be obliged if you could acknowledge receipt of this letter* prototips pārtulko kā *Es būtu pateicīgs saņemt jūsu apstiprinājumu par šīs vēstules saņemšanu.* Tomēr, lai arī sistēmu novērtējums BLEU punktos ir augsts, frāzēs balstītie modeļu tulkojumi fleksīvām valodām ne vienmēr ir apmierinoši. Lai noskaidrotu galvenās kļūdu grupas, tika veikta iegūto tulkojumu analīze.

Viena no visbiežāk sastopamajām kļūdām ir nepareiza locījuma izvēle vārdkopai. Piemēram, teikumu *the supervisory authorities concerned shall consult each other before approving such assignment*, SMT sistēma tulko kā *attiecīgo uzraudzības iestāžu apspriežas pirms apstiprina šādu norīkošanu*. Nepareiza locījuma izvēle vārdkopai *attiecīgo uzraudzības iestāžu* ir saistīta ar valodas modeļa ierobežojumiem. Tā kā valodas modelī izmantotas trigrammas (trīs vārdu savienojumi), tad šīs vārdkopas biežums apmācībā izmantotajos tekstos noteica ģenitīva formas izvēli nominatīva formas vietā.

Tāpat bieži sastopams ir nepareizs vārdkopu sakārtojums teikumā. Tas bieži izpaužas, tulkojot ģenitīva vārdkopas, kas angļu valodā ar prievārdu *of* tiek novietotas aiz galvenā vārda, bet latviešu valodā parasti – pirms galvenā vārda. Piemēram, teikuma *environmental risks of maritime traffic* datora tulkojums ir *vides risku, jūras satiksmes* (pareizi – *jūras satiksmes riski videi)*. Iemesls šai kļūdai tulkojumā, visticamāk, saistīts ar tulkošanas vārdnīcu, kurā atšķirības ģenitīva vārdkopas realizācijā dažādās valodās ir iekodētas konkrētiem piemēriem.

Visbeidzot SMT sistēmas slikti tulko garas vārdkopas, jo tulkojums tiek veidots no vārdiem un vārdu virknēm, tādējādi garas vārdkopas var tik pārdalītas. Piemēram, teikumu *do not have a direct, substantial and reasonably foreseeable impact on consumers in the Requesting Party's territory*, SMT sistēma tulko kā *nav tiešas, būtiska ietekme uz patērētājiem un pamatoti paredzamo pieprasījuma iesniedzējas puses teritorijā* (pareizi – *neatstāj tiešu, būtisku un visai prognozējamu iespaidu uz patērētājiem pieprasījuma iesniedzējas puses teritorijā*). Šīs kļūdas iemesls meklējams valodas modelī, kurā vārdu savienojums *būtiska ietekme* ir biežāk sastopams nekā *pamatoti paredzama ietekme.*

Iepriekš minētās kļūdas, kas īsos teikumos liekas maznozīmīgas, atstāj lielu iespaidu gadījumos, kad jātulko garāks teikums. Piemēram, teikumu *The*

ombudsman shall as soon as possible inform the person lodging the complaint of the action he has taken on it, SMT prototips tulko kā *Ombuds pēc iespējas drīz informē sūdzības iesniedzēju par to darbību ir veikusi persona*" (pareizi –*ombuds cik vien iespējams drīz informē sūdzības iesniedzēju par uzsāktajām darbībām sakarā ar sūdzību*), kuru lasītājs visdrīzāk uzskatīs par nesaprotamu.

4. Lingvistisko zināšanu iekļaušana SMT

Galvenie trūkumi frāzēs balstītajai SMT ir nepareiza locījuma formas izvēle vārdkopai un kļūdains vārdu sakārtojums teikumā. Lai to novērstu, jaunākās paaudzes SMT sistēmās papildus tiek integrētas lingvistiskās zināšanas, piemēram, zināšanas par morfoloģiju vai sintaksi.

Visbiežāk SMT sistēmās tiek integrētas zināšanas par morfoloģiju, veidojot faktorētus (factored) SMT sistēmu modeļus (Koehn, Hoang 2007). Faktorētos modeļos papildus norāda atbilsmes starp dažādām morfoloģiskajām pazīmēm, piemēram, var norādīt ne tikai atbilstību starp vārdformām abās valodās, bet arī vārdšķiru atbilsmi, gramatisko informāciju un citas morfoloģiju raksturojošas kategorijas. Angļu-latviešu faktorētajā SMT sistēmā uzdotas atbilsmes starp angļu valodas vārdformu un atbilstošo latviešu valodas vārda pamatformu un gramatisko informāciju. Šāds modelis izvēlēts tāpēc, ka angļu valodai nav bagāta morfoloģija. 3. attēlā redzamajā piemērā angļu vārdam *legislative* atbilst īpašības vārda *normatīvs* daudzskaitļa akuzatīva forma (acmpa1) un daudzskaitļa lokatīva forma (acmpl1), un vārdu virkne *tiesību aktos* (lietvārds *tiesība* daudzskaitļa ģenitīvā (ncfpg4), un lietvārds *akts* daudzskaitļa lokatīvā (ncmpl1)).

3. attēls. **Vienkāršots faktorētas SMT tulkošanas modeļa fragments**

```
legislative ||| normatīvs|acmpa1 ||| 0.181818 0.047619 0.0243902 0.030303 2.718
legislative ||| normatīvs|acmpl1 ||| 0.0357143 0.0081967 0.0121951 0.013468 2.718
legislative ||| tiesība|ncfpg4 akts|ncmpl1 ||| 0.0021322 0.00616645 0.0121951
                                                              0.00294754 2.718
```

Izveidoto SMT sistēmu vērtējums BLEU metrikā apkopots 2. tabulā. Kā redzams, tad, novērtējot SMT sistēmas ar BLEU metriku, ne vienmēr faktorētās sistēmas uzrāda labāku rezultātu kā frāzēs balstītās sistēmas, kas liek domāt par faktorēto modeļu nepiemērotību latviešu valodai.

2. tabula. **Faktorēto SMT sistēmu novērtējums ar BLEU metriku**

Apmācībai izmantotais korpuss	**BLEU punkti**					
	JRC Acquis 2.2 testa korpuss		JRC Acquis 3.0 testa korpuss		Nejauši izvēlēts (JRC Acquis 3.0)	
	Frāžu	Fakto-rētā	Frāžu	Fakto-rētā	Frāžu	Faktorētā
JRC Acquis 2.2	39,53	38,25	27,06	28,96	37,79[7]	35,84[7]
JRC Acquis 3.0	41,51	40,06	43,28	42,98	39,25	46,44
DGT TM	48,42[8]	46,21[7]	31,86	34,61	43,73	42.86

Kā jau iepriekš tika minēts, tad BLEU metrika nereti tiek kritizēta par neadekvātu vērtējumu, tāpēc papildus tulkojumu kvalitāti izvērtēja cilvēks, nosakot, kurš no sistēmu tulkojumiem, viņaprāt, ir labākais. Cilvēks novērtēja 200 teikumus, kas nejauši tika izvēlēti no testa korpusa. Cilvēka uzdevums bija no katra teikuma tulkojumiem izvēlēties vienu vai vairākus labākos teikuma tulkojumus. Sistēmu vērtējums apkopots 3. tabulā.[8]

3. tabula. **Cilvēka novērtējums tulkojumam**

SMT sistēmas veids	**Apmācībai izmantotais korpuss**	**Labu tulkojumu skaits**
Frāzēs balstītā	JRC Acquis 2.2	20
Faktorētā	JRC Acquis 2.2	42
Frāzēs balstītā	JRC Acquis 3.0	57
Faktorētā	JRC Acquis 3.0	74

Iegūtie rezultāti ļauj secināt, ka faktorēto sistēmu tulkojums ir labāks nekā frāzēs balstīto sistēmu tulkojums: gan JRC Acquis 2.2 datos, gan JRC Acquis 3.0 datos apmācītās faktorētās sistēmas ir ieguvušas augstāku rezultātu nekā atbilstošās frāzēs balstītā sistēmas. No otras puses, sistēmas, kuru apmācībā izmantots lielāks datu apjoms (JRC Acquis 3.0) tulko labāk nekā sistēmas ar mazāku datu apjomu. Tādējādi varam secināt, ka visbūtiskāk tulkojuma kvalitāti ietekmē 1) datu daudzums (JRC Acquis 3.0 satur vairāk nekā 5 reizes vairāk vārdformu nekā JRC Acquis 2.2) un 2) sistēmā iekļautās lingvistiskās zināšanas (faktorēto sistēmu tulkojums tiek novērtēts kā piemērotākais).

[8] Testa korpuss varētu saturēt apmācības korpusa datus.

5. SMT sistēmu kļūdu analīze

Lai arī 3. tabulā apkopotie rezultāti parāda, ka lingvistiskajām zināšanām ir būtiska ietekme uz SMT sistēmu tulkojumu kvalitāti, tā nesniedz atbildi uz jautājumu, kādi trūkumi jānovērš pašreiz izstrādātajās SMT sistēmās. Tāpēc SMT sistēmu kļūdu analīzei bieži izmanto Vilar u. c. (2006) piedāvāto metodoloģiju. SMT sistēmās tiek analizētas piecas kļūdu grupas: izlaisti vārdi, nepareiza vārdu kārta, nepareizi pārtulkoti vārdi, nezināmi (nepārtulkoti) vārdi un pieturzīmju lietojums. Katrā kļūdu grupā izdalīti vairāki veidi.

4. tabulā apkopotas kļūdas 100 nejauši izvēlētiem teikumiem no testa korpusa. Izvēlētajos piemēros saglabāts SMT sistēmas tulkojums, kas var ietvert vairāk nekā vienu kļūdu. Visvairāk kļūdu ir saistītas ar nepareizas vārdformas izvēli, lieka vārda lietojumu un izlaistu vārdu. Tā kā juridiskos tekstos idiomātiskus izteicienus parasti nelieto, tad šo kļūdu skaits ir 0. Jānorāda, ka pieturzīmju lietojums un stila kļūdu izvērtēšana pašreiz mašīntulkošanas sistēmā ir salīdzinoši maznozīmīga. Analizējot tulkojumus latviešu valodā, ar daudznozīmību saistītās kļūdas ir grūti atšķiramas no kļūdām, kas saistītas ar nepareizu leksisko izvēli.

4. tabula. **SMT kļūdu analīze pēc Vilara klasifikācijas**

Kļūdu grupa	**Skaits**	**Piemērs**
Izlaisti vārdi	**65**	
Patstāvīgie vārdi	52	african swine fever must be considered an endemic disease in the province of nuoro, sardinia, italy. āfrikas cūku mēris jāuzskata par endēmisku slimību sardīnija, itāliju.
Palīgvārdi	12	should it not be possible for the member states concerned to reach agreement, the member states in question shall immediately inform the other member states and the commission, giving reasons for their decision. tas nav iespējams dalībvalsts var panākt vienošanos, konkrētās dalībvalstis nekavējoties par to informē citas dalībvalstis un komisiju, norādot sava lēmuma iemesliem.

Kļūdu grupa	Skaits	Piemērs
Kļūdaina vārdu kārta	**35**	
Vārdiem, kas ir blakus	11	member states shall ensure that the conditions laid down in paragraph 2 are met, when requesting the recognition of a protected zone as referred to in the first indent of the first subparagraph of article 2 (1) (h) of directive 77/93/ eec. dalībvalstis nodrošina, ka 2. punktā noteiktie nosacījumi ir ievēroti, ja tās pieprasa aizsargājamas zonas atzīšanu, kā minēts pirmās daļas pirmajā ievilkumā 2. panta 1. punkta h) direktīvas 77/93/ eek.
Vārdkopām, kas ir blakus	11	the first indent of point c (origin of produce) of title vi (provisions concerning marking) is replaced by the following: punkta pirmajā ievilkumā c (nosaukumu) vi sadaļas (noteikumi attiecībā uz marķēšanu) punktu aizstāj ar šādu punktu:
Vārdiem, kas nav blakus	6	conservation and protection of the environment on the basis of the principles of sustainable development saglabāšanu un vides aizsardzību, pamatojoties uz noturīgu attīstību
Vārdkopām, kas nav blakus	7	whereas for waste from the titanium dioxide industry it is advisable to lay down a special system which will ensure that human health and the environment are protected against the harmful effects caused by the uncontrolled discharge, dumping or tipping of such waste Tā kā ir ieteicams paredzēt tādu īpašu sistēmu titāna dioksīda ražošanas atkritumiem, kas nodrošina, ka cilvēku veselību un vidi no kaitīgas ietekmes, ko izraisa šādu atkritumu nekontrolēta novadīšana, nogremdēšana vai noglabāšana; aizsargā
Nepareizi izvēlēti vārdi	**128**	
Nepareiza leksiskā izvēle	25	the following characteristics relating to each employee in the sample ar šādiem parametriem, kas attiecas uz katru darba izlasē
Kļūda daudznozīmības novēršanā	10	members carry on a particular profession dalībnieki veikt īpašu profesiju

Kļūdu grupa	Skaits	Piemērs
Nepareiza forma	51	catches are taken during the course of scientific investigations nozvejas laikā tiek veikti zinātniskos pētījumus
Lieki vārdi	34	means an industrial plant which is operational on the date of notification of this directive. ir rūpniecības uzņēmums, kas darbojas šīs direktīvas paziņošanas dienas šo direktīvu.
Stils	8	annex I shall be supplemented by the addition of the following I pielikumu papildina, papildinot to ar šādu:
Idiomas	0	
Nezināmi vārdi	**4**	the community financial contribution shall be granted per calender year kopienas finansiālais ieguldījums ir piešķīrusi calender tonnu gadā
Pieturzīmes	**20**	moreover, where appropriate, the rules governing alterations to the memorandum and articles of association shall apply. Turklāt, ja vajadzīgs, noteikumus, kas reglamentē izmaiņas memoranda un statūtu_piemēro.

6. Jaunākie virzieni SMT sistēmās

Lai arī SMT sistēmas salīdzinoši labi tulko tekstus, kas ir līdzīgi tiem tekstiem, uz kuriem sistēma apmācīta, tās uzrāda sliktus rezultātus darbā ar atšķirīgiem tekstiem. Rakstā apskatīto sistēmu labākais rezultāts BLEU punktos ir 46,44, tulkojot ES juridiskos dokumentus, bet, ja sistēma tiek darbināta uz līdzsvarota tekstu korpusa (Skadiņš u. c. 2010), kurā ietverti dažādi teksti, tad sistēmas novērtējums ir tikai 13,42 BLEU punkti. Tādējādi viena no aktuālām pētniecības problēmām ir jaunu metožu izveide, kas ļautu paplašināt SMT sistēmu lietojamību. Lai to paveiktu, tiek strādāts gan pie inovatīvām metodēm jaunu piemērotu lingvistisko resursu atrašanai tīmeklī, gan pie metodēm, kas ļauj izgūt tulkošanai nepieciešamo informāciju no cita veida resursiem, piemēram, salīdzināmiem korpusiem, terminoloģiskām vārdnīcām.

Līdztekus jaunu lingvistisko resursu izveidei ne mazāk būtisks jautājums ir par SMT sistēmu papildināšu ar lingvistiskajām zināšanām. Tāpēc pēdējā

laikā tiek pētītas iespējas SMT sistēmās iekļaut sintaktiskās zināšanas. Diemžēl līdz šim radītās metodes vairāk piemērotas valodām, kurām izveidoti lieli sintaktiski anotēti valodas korpusi. Latviešu valodai tāds ir tikai tapšanas stadijā.

Līdztekus SMT sistēmu papildināšanai ar zināšanām par valodu arvien biežāk izskan doma par hibrīdo MT sistēmu attīstīšanu, tajās kombinējot zināšanas no likumos balstītājām sistēmām un no statistiskajām sistēmām (Federmann 2010). Arī tulkošanai latviešu valodā šis varētu izrādīties perspektīvākais virziens, jo nav nepieciešams veidot milzīgus lingvistiski anotētu tekstu korpusus, kas prasa lielus resursu ieguldījumus.

7. Secinājumi un nākotnes ieceres

Tulkošanas sistēmas gūst arvien lielāku popularitāti valodas nezinātāju starpā. Pašlaik esošās SMT sistēmas uzrāda labus rezultātus, tulkojot tekstus, kas līdzīgi tiem tekstiem, kuri izmantoti sistēmas apmācīšanai. Tāpēc šādas sistēmas var izmantot arī cilvēki ar labām valodas zināšanām, piemēram, ar lokalizāciju saistītu uzdevumu veikšanā.

Tomēr, tulkojot patvaļīgu tekstu, SMT sistēmas tulkojumu nevar pielīdzināt cilvēka tulkojumam. Rakstā aprakstītajiem prototipiem, tulkojot tekstu latviešu valodā, bieži vien tekstā sastopamas nepareizi lietotas vārdformas, tulkojumā mēdz tikt izlaisti vārdi un teikumā ne vienmēr vārdi pareizi sakārtoti vārdkopās, kā arī atsevišķi vārdi mēdz vispār nebūt pārtulkoti. Tādējādi par galvenajiem turpmākā darba virzieniem var tikt minēta efektīva metožu, kas ļautu integrēt lingvistiskās zināšanas, izveide SMT sistēmām un hibrīdo MT sistēmu izveide.

DEVELOPMENT OF ENGLISH-LATVIAN STATISTICAL MACHINE TRANSLATION SYSTEM: METHODS, RESOURCES AND FIRST RESULTS

Summary

This paper presents research and development of English-Latvian Statistical Machine Translation (SMT) prototypes for legal domain. Several methods have been investigated, i.e., phrase-based models and factored models. Translation quality has been evaluated using automated metrics (BLEU score) and human evaluation. In automatic evaluation the best score (46.44 BLEU points) was assigned to factored model trained on JRC Ac-

quis corpus (version 3.0) which was also evaluated as the best from the human viewpoint. In addition, error analysis of SMT output was performed. This analysis showed that although the output of the best prototype demonstrated a reasonable quality, it had several frequent common errors, i.e., incorrect form, missing words and wrong word order. For the future, work on tree-based SMT and hybrid systems is proposed.

LITERATŪRA

Berger Adam L., Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, John R. Giuett, John D. Lafferty, Robert L. Mercer, Harry Printz, Luboi Urei 1994, The Candide system for machine translation, in *Proceedings of the ARPA Conference on Human Language Technology*, 157–162.

Federmann Christian, Andreas Eisele, Hans Uszkoreit, Yu Chen, Sabine Hunsicker, Jia Xu 2010, Further experiments with shallow hybrid MT systems, in *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, 77–81.

Hutchins John 2007, Machine translation: a concise history, in Chan Sin Wai. (ed.), *Computer aided translation: Theory and practice*, Chinese University of Hong Kong.

Koehn Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Corbett Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, Evan Herbst 2007, Moses: Open source toolkit for statistical machine translation, in *ACL 2007,* 177–180.

Koehn Philipp, Franz Josef Och, Daniel Marcu 2003, Statistical phrase based translation, in *Proceedings of the Joint Conference on Human Language, Technologies and the Annual Meeting of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL)*, Edmonton, Canada, 48–54.

Koehn Philipp, Hieu Hoang 2007, Factored translation models, in *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Prague, Czech Republic, June 2007.

Och Franz Josef, Hermann Ney 2003, A systematic comparison of various statistical alignment models, *Computational Linguistics* 29(1), 19–51.

Papineni Kishore, Salim Roukos, Ward Tood, Wei-Jing Zhu 2002, BLEU: a method for automatic evaluation of machine translation, in *Proceedings of the 40th Annual Meeting of the Association of Computational Linguistics (ACL)*, Pennsylvania, 311–318.

Skadiņa Inguna, Edgars Brālītis 2008, Experimental statistical machine translation system for Latvian, in *Proceedings of the 3rd Baltic Conference on HLT*, Vilnius, 281–286.

Skadiņš Raivis, Kārlis Goba, Valters Šics 2010, Improving SMT for Baltic languages with factored models, in *Proceedings of the Fourth International Conference Baltic HLT 2010*, IOS Press (= *Frontiers in Artificial Intelligence and Applications* 219), 125–132.

Steinberger Ralf, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiş 2006, The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages, in *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006),* 2142–2147.

Stolcke Andreas 2002, SRILM – an extensible language modelling toolkit, in *ICSLP-2002*, 901–904.

Vilar David, Jia Xu, Luis Fernando D'Haro, Hermann Ney 2006, Error analysis of machine translation output, in *Proceedings of the 5th Internation Conference on Language Resources and Evaluation (LREC'06)*, 697–702.

Inguna SKADIŅA, Madars VIRZA, Lauma PRETKALNIŅA
Mākslīgā intelekta laboratorija
LU Matemātikas un informātikas institūts
Raiņa bulvāris 29
LV-1459, Rīga
Latvia
[*inguna.skadina@lumii.lv*]
[*madars.virza@lumii.lv*]
[*lauma.pretkalnina@lumii.lv*]