KUIDAS MASIN TÕLGIB

HEIKI-JAAN KAALEP, MARE KOIT

1. Sissejuhatus

Käesoleva aasta märtsis proovisime Google'i masintõlkeprogrammiga (*http://translate.google.ee/#*) tõlkida inglise keelde lauseid „Hea ilmaga sõidab paadiga sinna viie tunniga" ja „Paadiga sõidab hea ilmaga sinna viie tunniga". Vastuseks saime mõlemal juhul: „Good weather, a boat travels there for five hours." Tõlge pole perfektne, aga on täiesti arusaadav. Eesti keel on nende enam kui 50 keele hulgas (vt sama internetilehekülge), mis on arvuti abil tõlgitavad ja mille kõnelejaid-kirjutajaid ei saa oma salakeele ainuvaldajatena kasutada, nagu kasutati navaho indiaanlasi USA armees Teise maailmasõja ajal.

Teiselt poolt, kui tõlkisime sama programmiga eesti keelde lause „Blank Verse is any verse comprised of unrhymed lines all in the same meter, usually iambic pentameter", siis saime „Silosäe on iga salm koosneb riimitu read kõik samas meeter, tavaliselt neljajalalise jambi pentameter".

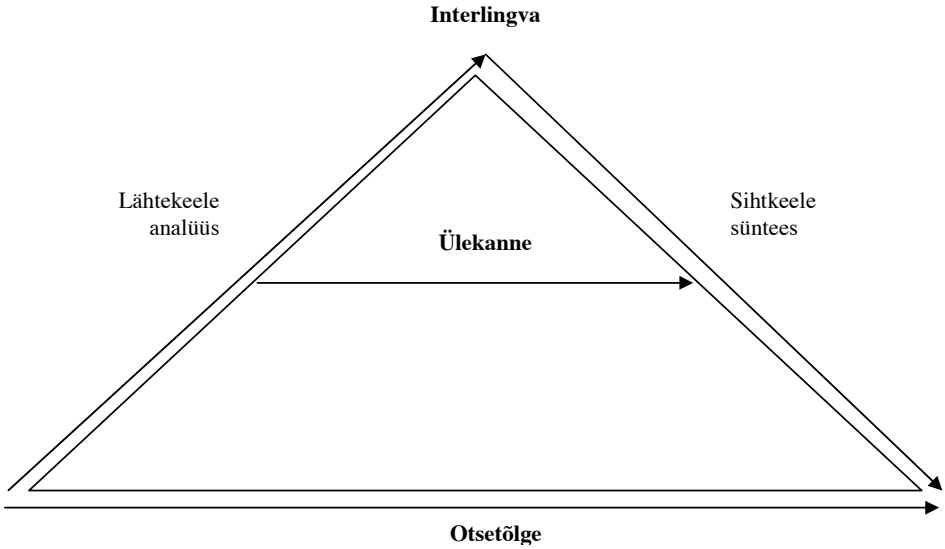
Milles on asi? Miks oli teisel juhul tõlge nii halb? Kust võeti väljend „neljajalalise jambi pentameter"? Võib tunduda veider, aga tegelikult ei tea seda ka tõlkesüsteemi tegijad ise. Alljärgnevalt püüame liigselt detailidesse minemata kirjeldada, kuidas masintõlkesüsteemid töötavad ja kuidas neid tehakse. Masinatega kipub ikka nii olema, et teadmine, kuidas asi töötab, võimaldab teda paremini kasutada (eriti sel juhul, kui ta hästi ei tahagi töötada).

2. Esialgne (nn klassikaline) lähenemisviis

Mõte kasutada tõlkimisel elektronarvutit on peaaegu niisama vana kui arvuti ise. Mõni masintõlke (MT) arengut soodustanud idee on aga pärit veel palju varasemast ajast. Näiteks soovitati juba XVII sajandil keelebarjääride ületamiseks kasutada keelest sõltumatut tähenduse esitust. 1933. aastal patenteeris vene teadlane Pjotr Smirnov-Trojanski mehaanilise tõlkimise protseduuri.

Tõsisemad MT-alased katsetused algasid koos esimeste elektronarvutite loomisega 1940. aastatel, olles niisiis – üllataval kombel – arvutusmasinate esimesi rakendusi.

Bernard Vauquois, kes töötas 1970. aastatel Grenoble'i Ülikoolis, mis oli juhtivaid reeglipõhise MT keskusi, võttis MT klassikalised meetodid kokku nn Vauquois' kolmnurgana (vt joonist 1), milles iga kõrgem aste nõuab lähtekeele detailsemat analüüsi ja vastavalt sihtkeele detailsemat sünteesi.



J o o n i s 1. Vauquois' kolmnurk: MT klassikalised meetodid.

2.1. Otsetõlge

Selle meetodi korral tõlgitakse lähtekeele tekst sõnahaaval, kasutades mahukat kakskeelset sõnastikku. Sisendit käsitletakse sõnavormide järjendina ja otse selle järjendiga sooritatakse mitmesuguseid operatsioone: asendatakse lähtekeele sõnad sihtkeele omadega, muudetakse nende järjekorda saamaks sihtkeele korrektsed laused jne.

Otsetõlge on võimalik, kui keeled on suhteliselt lähedase struktuuriga ja sarnase lauseehitusega. Esimesed MT-süsteemid, mis loodi 1950. aastatel, olid kõik otsetõlkesüsteemid. Tüüpiliselt luuakse selline süsteem ühe kindla keelepaari jaoks.

Ka Eestis hakati MT-ga tegelema 1950. aastate lõpus, mil Tartu Ülikooli matemaatikud eesotsas Ülo Kaasikuga proovisid tõlkida matemaatilisi tekste vene keelest eesti keelde. Koostati ka mõni programm (nt vene keele morfoloogiline analüüs) tolleaegsele arvutile Ural, mille töökiirus oli tänapäeva arvutitega võrreldes naeruväärne — 100 tehet sekundis — ega võimaldanud muidugi efektiivset tõlkimist. Hiljem see töö siiski soikus.

2.2. Ülekanne

Selle strateegia järgi jaguneb tõlkimine kolme faasi: lähtekeele analüüs, ülekanne ja sihtkeele süntees ehk genereerimine. Üldiselt jaotatakse analüüs morfoloogiliseks, süntaktiliseks ja semantiliseks analüüsiks. Süntees (kui analüüsi pöördprotsess) jaguneb samal viisil semantiliseks, süntaktiliseks ja morfoloogiliseks sünteesiks.

Tuleb märkida, et MT jaoks võib analüüs erineda muudel eesmärkidel tehtavast analüüsist. Näiteks lause *Jüri vaatas pikksilmaga tüdrukut* süntaktilisel analüüsil ei ole oluline, missuguse sõna laiend on *pikksilmaga*.

Lähte- ja sihtkeelest sõltuvalt võib ülekanne toimuda kas pärast morfoloogilist, süntaktilist või semantilist analüüsi ja jätkuda vastavast sünteesi etapist, s.t osa analüüsi/sünteesi etappe võidakse ära jätta.

Olemasolevates reeglipõhistes MT-süsteemides toimub ülekanne enamasti pärast süntaktilist analüüsi. Näiteks ingliskeelse lause *It rains* eesti keelde tõlkimisel võiks leida lähtelause süntaktilise struktuuri ALUS ÖELDIS ning teisendada selle sihtlause omaks: ÖELDIS ALUS. Lisaks tuleb teha muidugi leksikaalne ülekanne, s.t asendada lähtekeele sõnad sihtkeele vastetega, saades tõlke: *Sajab vihma*.

Leksikaalse ülekande alus on otsing kakskeelsest sõnastikust. Tõlkeekvivalendiks võib olla nii sõna kui ka pikem fraas. Näiteks ingliskeelse sõna *misinterpret* vaste on eesti keeles 'valesti tõlgendama' ja saksakeelse fraasi *abgerahmte Milch* eesti vaste on 'lõss'. Üks tuntud MT-süsteem, mis ülekandemeetodit kasutab, on Kanada ilmateadete inglise-prantsuse tõlkesüsteem METEO.

2.3. Otsetõlke ja ülekande kombineerimine

Kuigi ülekandemeetod võimaldab nii lähte- kui ka sihtkeeles keerulisemaid nähtusi käsitleda kui otsetõlge, ilmneb, et lihtsad sõnajärje muutmise reeglid pole piisavad. Praktikas on vaja palju teisendusreegleid, mis kasutavad mõlema keele sõnades peituvaid teadmisi ning süntaktilisi ja semantilisi tunnuseid. Seetõttu kombineerivad kommertskasutuses olevad MT-süsteemid otsetõlget ja ülekannet, kasutades mahukaid kakskeelseid sõnastikke, aga ka morfoloogilist ja süntaktilist analüüsi. Näiteks SYSTRAN (SYStem TRANslation), mille arendamist alustati 1968. aastal Ameerika Ühendriikide õhujõudude tarvis vene keelest inglise keelde tõlkimiseks ja mis on üks vähestest praegu tööstuslikult kasutatavatest MT-süsteemidest, suudab tõlkida 35 keelepaari, täites kolme etappi: 1) pindanalüüs (morfoloogiline analüüs, sõnaliikide märgendamine, fraaside tuvastamine), 2) ülekanne (idioomide tõlge, sõnatähenduste ühestamine, eessõnade määramine verbide järgi), 3) süntees (sõnade tõlge kakskeelse sõnastiku abil, lause sõnajärje muutmine, morfoloogiline süntees). Seega teeb SYSTRAN analoogiliselt otsetõlkesüsteemidega suure osa oma tööst kakskeelse sõnastikuga, mis sisaldab nii leksikaalset, süntaktilist kui ka semantilist infot. Teiselt poolt kasutatakse lähtekeele süntaktilist ja pindsemantilist informatsiooni nagu ülekandesüsteemideski.

2.4. Vahekeele kaudu tõlkimine

Ülekandemeetodi puuduseks on, et iga erineva keelepaari jaoks on vaja erinevat ülekandereeglite hulka. Euroopa Liidu paljukeelses keskkonnas oleks eelistatavam teistsugune MT meetod, vahekeele kaudu tõlkimine. Vahekeele meetod käsitleb tõlkimist sisendi tähenduse mõistmisena ja selle tähenduse sihtkeeles väljendamisena. See eeldab, et on olemas keelest sõltumatu tähenduse esitamise keel ehk interlingva. Tõlkimise järjekord on järgmine: esmalt tehakse sisendlause täielik analüüs, saades tähenduse esitus vahekeeles, ning siis genereeritakse sellest lähtudes sihtkeele lause.

Sobiva vahekeele loomine on omaette probleem: kuidas valida vajalikud mõisted ja nendevahelised seosed. Näiteks inglise sõna *eat* 'sööma' saksa keel-

de tõlkimisel tuleb valida kahe vaste vahel: *essen* (inimese kohta) ja *fressen* (looma kohta). Seega peab vahekeel sisaldama mõlemad tähendused ja sisendlause analüüsil tuleb kindlaks määrata, kummaga on tegu.

Millist formalismi kasutada vahekeeles? On kasutatud näiteks predikaatarvutust, atomaarseid primitiive, sündmusepõhist esitust, kus sündmused on seostatud nende argumentidega väikese hulga temaatiliste rollide abil.

Vahekeele meetodi puhul peavad analüsaator ja süntesaator tegema rohkem tööd kui ülekandemeetodi rakendamisel, aga samal ajal võimaldab see vältida kahte keelt kõrvutavate teadmiste kasutamist ja piirduda iga üksiku keele analüüsi ja sünteesi moodulitega (millel on palju teisigi rakendusi peale MT). Vahekeele väljatöötamine nõuab MT-süsteemi loojalt muidugi ainevaldkonna semantika põhjalikku uurimist ja formaliseerimist. Mõne lihtsa valdkonna puhul (nt bussiinfo, hotelli reserveerimine, piletiost) võib semantika esitada andmebaasina. Andmebaas määrab võimalikud mõisted ja seosed ning MT-süsteemi looja ülesanne on üksnes otsustada, kuidas vastavad neile kahe käsitletava keele sõnad ja struktuurid.

3. Miks on MT raskem, kui esialgu arvati?

Ammu on teada, et keeled erinevad oma ülesehituse poolest üksteisest süstemaatiliselt.

M o r f o l o o g i l i s e l t saab keeli iseloomustada kahe mõõtme abil. Esiteks, morfeemide arv sõna kohta: leidub isoleeritud keeli, kus iga sõna on üksainus morfeem (nt vietnami keeles), ja polüsünteetilisi keeli, kus üks sõna võib koosneda väga paljudest erinevatest morfeemidest, mis vastavad tervele lausele (nt eskimo keeles). Teine mõõde on segmenteeritavuse aste: on aglutinatiivseid keeli, kus morfeemidel on suhteliselt selged piirid (nt türgi keeles), ja flekteerivaid keeli, kus üks afiks võib esitada mitut grammatilist tähendust (nt vene keeles).

S ü n t a k t i l i s e l t erinevad keeled sõnajärje poolest, näiteks on olemas SVO-keeled (inglise), SOV-keeled (iiri, araabia), V2-keeled (rootsi, eesti) jne.

Iga tüpoloogiline mõõde võib põhjustada tõlkimisel probleeme, kui lähte- ja sihtkeel selle poolest erinevad. Näiteks SVO-keelest SOV-keelde tõlkimine nõuab suurt hulka struktuurilisi ümberkorraldusi, sest lähte- ja sihtkeeles on lauseliikmed erinevatel kohtadel.

Mõned lihtsamad struktuurierinevused on hästi teada nagu näiteks asjaolu, et prantsuse keeles järgneb enamik omadussõnu nimisõnadele, aga inglise keeles eelnevad omadussõnad reeglina nimisõnadele; jaapani ja ladina keeles on põhitegusõna lauses viimane, aga inglise ja saksa keeles asub põhitegusõna esimese nimisõnafraasi järel. Rohkem on siiski selliseid olukordi, kus keelte struktuuri võrdlemine ei ole nii lihtne.

Tegelik raskus seisneb MT jaoks aga selles, et keelte erinevused ei ole süstemaatilised. Sõnade puhul on ammu teada, et nende tähendusväljad eri keeltes ei kattu, seega tuleb õige vaste leidmiseks arvestada konteksti. Kuid kontekst ise koosneb sõnadest, millest igaühe puhul on samasugune mitmeti tõlgitavuse võimalus. Ka grammatilised tähendused, mida esindavad morfoloogilised kategooriad ja süntaktilised konstruktsioonid, nõuavad tõlkimisel konteksti arvestamist. Näiteks eesti kaasaütlevat võib kontekstist sõltuvalt tõlkida inglise keelde kas erinevate eessõnadega (nt *bussiga – by bus*, *habe-*

mega mees – man with a beard) või hoopis ilma nendeta (nt *sink munaga – ham and eggs*). Just see kontekstitingimuste ebasüstemaatilisus osutus klassikalisele MT paradigmale saatuslikuks: hoolimata suurest kulutatud inimtööst ei õnnestunud tõlkesüsteemide kvaliteeti enam parandada: praeguste teadmiste valguses võiks klassikalist MT paradigmat iseloomustada katsena süstematiseerida süstematiseeritamatut.

1966. aastal avaldati Ameerika Ühendriikides keele automaattöötluse komitee (Automatic Language Processing Advisory Committee, ALPAC) aruanne, mis oli MT võimalikkuse ja tasuvuse suhtes ülimalt kriitiline. See aruanne pärssis MT arengut paljudeks aastateks.

MT-alane töö jätkus siiski, aga piiratud ulatuses. 1968. aastal alustati Ameerika Ühendriikides MT-süsteemi SYSTRAN loomist, 1976. aastal Kanadas METEO väljatöötamist. Samal aastal hakkas Euroopa Liidu Komisjon kasutama SYSTRAN-i inglise-prantsuse versiooni.

4. Statistiline MT

4.1. Müraga kanali mudel

Klassikalisi MT meetodeid arendati põhiliselt MT algusaegadel, kuni 1980. aastate lõpuni. Viimase paarikümne aasta jooksul on MT-alases arendustöös domineerinud andmepõhised, statistilised meetodid. Ka Google'i tõlkeprogramm esindab statistilist MT-d. Statistiline MT vaatleb tõlkimist hoopis teistsuguse protsessina kui klassikaline MT.

1947. aastal esitas ameerika insener Warren Weaver idee rakendada MT-s dešifreerimise meetodeid, mida oli edukalt kasutatud äsja lõppenud maailmasõjas teadete edasiandmisel. Tema ettekujutuse järgi pidi ühest keelest teise tõlkimine olema nagu salakirja lahtimuukimine. Tõlkeprotsessi võis tema arvates käsitleda tagurpidi: algselt oli sõnum väljendatud ühes keeles, aga mingi protsess moonutas teda ja nüüd on ta kodeeritud nii, et me teda lugeda ei oska. Ehk teiste sõnadega: algne sõnum on meieni jõudnud läbi mingi mürarikka sidekanali, mis teda on moonutanud.

Tõlkimise ülesandeks on esialgse esitusviisi taastamine. Sellise lähenemisviisi juures on oluline ainult see, et me teame, milline on tõlke sihtkeel. Lähtekeeleoskus polegi rangelt võttes oluline. Tähelepanuväärne on ka ettekujutus tõlkimisest kui tõenäosuslikust protsessist, milles väljendub umbusk kõigi varasemate teooriate suhtes, mis tõlkeprotsessi kirjeldavad. Statistiliste meetodite rakendamine eeldab loomulikult, et oleks olemas materjal, millele tuginedes tõenäosuslikke järeldusi teha. Masintõlke puhul on selleks materjaliks tekstikorpus, milles on nii algkeelseid tekste kui ka nende tõlkeid.

Praeguste statistiliste MT-süsteemide ajalugu ulatub 1990. aastatesse, mil firma IBM seni kõnetuvastusega tegelnud rühm töötas välja katselise prantsuse-inglise MT-süsteemi Candide (Berger jt 1994), mis põhines müraga kanali mudelil. Prantsuse ja inglise keele valik on siin oluline, sest keelte tüpoloogilised ühis- ja erijooned (nt asjaolu, et noomenifraasis on sõnajärg keeliti erinev, kuid fraaside järjestus lauses on sageli sama) määrasid suures osas kogu tõlkesüsteemi ülesehituse.

IBM-i rühma üldist lähenemisviisi võib kirjeldada järgmiselt (vt Brown jt 1993: 264–265). Suvalist ingliskeelset sõnade jada saab tõlkida prantsuse

keelde mitmel viisil. Põhimõtteliselt võiks mis tahes prantsuskeelne sõnajada F olla inglise sõnajada tõlkeks E tingliku tõenäosusega P(F | E). Seda võib ette kujutada sel moel, et meil on hiigelsuur tabel, milles igale võimalikule inglise ja prantsuse lausepaarile on seatud vastavusse tõenäosust väljendav 0 ja 1 vahel olev murdarv. Valides tõenäosuste jaoks õige jaotusfunktsiooni, on võimalik saada kui tahes kõrgekvaliteedilisi tõlkeid. Muidugi on sellise kõikvõimalikke lauseid sisaldava tabeli tegemine võimatu. Kuid see on ainult praktiline, mitte põhimõtteline probleem. MT jaoks oluline küsimus ei ole seega filosoofiline, vaid praktiline: kuidas leida tõenäosuste jaotustele lähendeid, mis on piisavalt head, et anda kasutuskõlblikke tõlkeid?

Vaatame nüüd, kuidas rakendada müraga kanali mudelit statistilises MT-s. Oletame selguse mõttes, et meil on vaja tõlkida lause (s.t *m* sõnast koosnev jada) $F = f_1\ f_2\ \ldots\ f_m$ mingist võõrkeelest (olgu selleks prantsuse keel) tuttavasse keelde (näiteks inglise keelde). Parim tõlge (s.t *k* sõnast koosnev jada) $E^* = e_1\ e_2\ \ldots\ e_k$ on see, mis maksimeerib tingliku tõenäosuse P(E | F):

$$E^* = \underset{E}{\operatorname{argmax}}\ P(E \mid F).$$

Kasutades esiteks Bayesi teoreemi $P(A \mid B) = P(B \mid A) \times P(A) / P(B)$ ning teiseks asjaolu, et tõlkimist vajava lause tõenäosus P(F) tegelikult ei mõjuta seda, milline on parim tõlge, saame selle valemi ümber kirjutada järgmiselt:

$$\underset{E}{\operatorname{argmax}}\ P(E \mid F) = \underset{E}{\operatorname{argmax}}\ P(F \mid E) \times P(E) / P(F) \sim \underset{E}{\operatorname{argmax}}\ P(F \mid E) \times P(E).$$

Seega tuleb meil määrata kaks tõenäosust: P(F | E) ja P(E).

Intuitiivselt võiks öelda, et tuleb leida kompromiss, et tõlge oleks usaldusväärne lähtekeele mõttes (originaalilähedane) ja sorav sihtkeeles. Seega võib tõlkimise eesmärki kujutleda kui sellise väljundi E* produtseerimist, mis maksimeeriks mõlema nimetatud parameetri väärtused:

$$E^* = \underset{T}{\operatorname{argmax}}\ \mathit{originaalilähedus}(E,F) \times \mathit{soravus}(E),$$

kus F on lähtekeele lause ja E on sihtkeele lause.

Parim viis E* leida on panna lauseks kokku sõnast suuremate segmentide tõlkevasteid, s.t fraase. Fraasi all ei mõelda siin tingimata sisulist tervikut nagu keeleteaduses, vaid fraasiks sobib suvaline lauses järjestikku olevate sõnade jada. Fraaside kaupa tõlkides saab vähendada nii üksiksõnade mitmetähenduslikkusest tulenevat valesti tõlkimist kui ka grammatilist kohmakust. Kui proovida paigutada statistilist MT-d Vauquois' kolmnurgale, siis on tegemist kõige madalama taseme, s.t otsetõlkega.

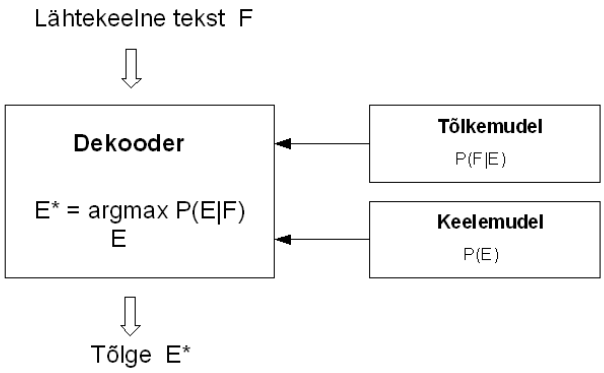
Müraga kanali mudeli rakendamiseks vajame kolme komponenti:

– tõlkemudelit, et arvutada P(F | E),

– keelemudelit, et arvutada P(E),

– dekoodrit, mis saab ette lähtelause F ja produtseerib kõige tõenäolisema tõlke E*.

Need kolm komponenti on ühendatud SMT-süsteemiks viisil, mida illustreerib joonis 2.



J o o n i s 2. SMT-süsteemi põhimõtteline skeem.

4.2. Tõlkemudel

Tõlkemudeli ülesandeks on leida (meie näitel prantsuse keelest inglise keelde tõlkimisel) tõenäosus P(F | E) ehk see, millise tõenäosusega hoopis ingliskeelne lause E genereerib prantsuskeelse lause F. Tõlkesuuna ümberpööramine tuleneb müraga kanali mudelist.

P(F | E) leidmisel lähtutakse eeldusest, et seda tõenäosust saab tuletada lause osade, s.t fraaside tõlgete tõenäosuste kaudu. Sõnastikud üldjuhul ei sisalda fraase, mida SMT eelistab kasutada, ja isegi kui sisaldavad, siis konkreetse valdkonna puhul võivad nii sõnad kui ka fraasid olla kasutatud oma eripärasel moel. Seega tuleb algul fraasid ja nende tõlked paralleelkorpuse alusel leida ehk teiste sõnadega: konstrueerida fraasitabel, kus on iga fraasipaari juures kirjas ka sellise tõlke tõenäosus.

Fraasitabeli koostamine on tõlkemudeli treenimise kõige olulisem osa. Fraaside tõlkimise tõenäosused $\psi(\hat{f}, \hat{e})$ saaksime leida, kui meil oleks suur kakskeelne treeningkorpus, kus igale prantsuskeelsele lausele on vastavusse seatud tema tõlge – ingliskeelne lause – ja seejuures oleks täpselt teada, milline fraas prantsuskeelses lauses milliseks fraasiks ingliskeelses lauses tõlgitud. Sellise vastavuse leidmist nimetatakse f r a a s i d e j o o n d a m i s e k s. Tõenäosusi $\psi(\hat{f}, \hat{e})$ saaksime lähendada suhteliste sagedustega: lugedes kokku, mitu korda oli fraasi $\hat{f}$ tõlkevasteks fraas $\hat{e}$, ning jagades selle arvu fraasi $\hat{e}$ korpuses esinemiste arvuga. Tavaliselt paraku meie käsutuses sellist joondatud fraasidega treeningkorpust ei ole. Õnneks osutub, et kui meil on olemas lausete tasemel joondatud paralleelkorpus, siis saab fraaside joonduse tuletada ü k s i k s õ n a d e j o o n d u s e s t. Lausete automaatne joondamine on suhteliselt lihtne ülesanne, sest lausete järjekord on tõlkes sama mis originaalis. Sõnade joondamiseks leidub mitu mudelit, näiteks IBM mudelid 1 kuni 5 (esitatud 1993. a firma IBM masintõlkerühma seminaritöödes) ja Markovi peitmudel (*Hidden Markov Model*, HMM). Kõigi nende mudelite treenimiseks on olemas head algoritmid.

Fraaside joondamise saab sõnade joondusest tuletada järgmiselt. Algul treenime eraldi kahte sõnajoondajat: inglise-prantsuse ja prantsuse-inglise.

Tulemuseks on kaks joondust, kus üks lähtekeele sõna võib olla seotud *0…n* sihtkeele sõnaga. Et leida fraase, s.t juhtumeid, kus *1…m* lähtekeele sõna on seotud *1…n* sihtkeele sõnaga, kombineerime kahte saadud sõnajoondust omavahel. Et kõik joondatud fraasipaarid on kogutud treeningkorpusest, siis saame leida fraasi tõlkimise tõenäosusele kõige tõepärasema hinnangu, lähendades seda suhtelise sagedusega.

Lõpuks salvestame kõik fraasipaarid koos tõenäosustega suurde fraasitõlketabelisse. Seda tabelit kasutab dekooder, et leida tõlke tõenäosust.

Statistilise tõlkemudeli loomiseks on vaja suurt paralleelkorpust. Keelte arv, mille jaoks need on olemas, samuti paralleelkorpused ise kasvavad pidevalt, nii et asjakohane informatsioon vananeb kiiresti. Suurimad ja vähimate kasutuspiirangutega paralleelkorpused esindavad ametlikku, bürokraatlikku keelt, näiteks Kanada parlamendidebattide korpus Hansard prantsuse ja inglise keeles (300 miljonit sõna; *http://www.tsrali.com*); Euroopa Ühenduse õigustiku korpus JRC-Acquis, mis sisaldab omavahel joondatult 22 keelt (sh eesti), kusjuures igas keeles on vähemalt 2000 teksti ja üle 20 miljoni sõna (*http://langtech.jrc.it/JRC-Acquis.html*); Euroopa Meditsiiniakadeemia paralleelkorpus, mis sisaldab omavahel joondatult 22 keelt (sh eesti), kusjuures igas keeles on vähemalt 1900 teksti ja üle 10 miljoni sõna (*http://urd.let.rug.nl/tiedeman/OPUS/EMEA.php*); Euroopa Parlamendi debattide korpus Europarl, mis sisaldab 11 omavahel joondatud keelt, kusjuures igas keeles on ümmarguselt 45 miljonit sõna (*http://www.statmt.org/europarl/*). On ilmne, et vaja oleks ka muud tüüpi tekste, näiteks ilukirjandust, kuid on selgunud, et üks olulisemaid takistusi on mitmesugune õigusprobleemistik.

4.3. Keelemudel

Keelemudel peaks tagama tõlke soravuse, vähendama etteheiteid, et jah, tõlge justkui on, aga selles keeles nii ei öelda. Keelemudel eelistab tüüpilisi, s.t treeningkorpuses sagedamaid sõnu, fraase ja konstruktsioone ebatüüpilisematele, s.t harvematele.

Keelemudeli tegelikkuses realiseeritud lähendina kasutatakse statistilises MT-s tavaliselt N-gramm-mudelit, mille aluseks on idee, et lause iga järgmist sõna saab ennustada talle eelnevate N–1 sõna alusel (nn Markovi eeldus). Näiteks bigramm-mudelis, kus iga sõna sõltub üksnes ühestainsast, talle vahetult eelnevast sõnast, võib lause

$$E = e_1\ e_2\ \ldots\ e_k$$

tõenäosuse arvutada valemist

$$P(E) = P(e_1) \times P(e_2 \mid e_1) \times P(e_3 \mid e_2) \times \ldots \times P(e_k \mid e_{k-1}).$$

Seejuures saab vajalikud tõenäosused leida korpusest, lähendades neid suhteliste sagedustega (nn maksimaalse tõepära hinnang):

$$P(e_i \mid e_{i-1}) = Arv(e_{i-1},e_i) / Arv(e_{i-1}),$$

kus i = 2, …, k; $Arv(e_{i-1},e_i)$ on paaride (e_{i-1},e_i) arv korpuses ja $Arv(e_{i-1})$ on sõna e_{i-1} esinemiste arv korpuses.

Keelemudel on ükskeelne ja nii on suuremahuliste treeningandmete saamine suhteliselt lihtne. Google'i uurijad (Brants jt 2007) on näidanud, et sihtkeele treeningkorpuse ja keelemudeli fraasi pikkuse N suurendamine toob kaasa tõlkekvaliteedi paranemise, kusjuures paranemine jätkub ka väga suurte korpuste (1 800 000 000 000 sõna) ja N väärtuste juures.

4.4. Dekooder

Statistilise MT kolmas komponent dekooder püüab leida prantsuskeelsele lähtelausele F vastavat parimat ingliskeelset tõlget E*, s.t maksimeerida tõlkemudelist ja keelemudelist pärinevate tõenäosuste korrutist.

On ülimalt tõenäoline, et tõlkimist vajavat lauset pole tervikuna varem kohatud, mistõttu tuleb sisendlause jagada fraasideks, mille jaoks on treenimisetapil olnud piisavalt materjali, et teha tõlkimisviisi kohta usaldusväärseid otsuseid. Fraasideks saab jagada mitmel eri moel ja igal võimalikul fraasil on mitu võimalikku tõlget, nii et tõlkehüpoteeside arv kasvab eksponentsiaalselt koos lausepikkusega.

Dekodeerimist võib käsitleda tehisintellektist tuntud otsinguprobleemina: otsing kulgeb olekute ruumis, kus olekuteks on lähtelause F tõlkehüpoteesid. Otsing lähtub algolekust (kus lause F ühtegi sõna ei ole veel sihtkeelde tõlgitud) ja suundub läbi osaliste tõlgete lõppolekusse (selleks on lause F täielik tõlge). MT dekoodrid kasutavad tüüpiliselt otsimisrinde meetodit: igas jooksvas olekus hinnatakse järgmisi võimalikke olekuid ja valitakse mõningad paremad, millest otsingut jätkata. Parima valikul võetakse ühelt poolt arvesse siiani leitud osalise tõlke tõenäosust ja teiselt poolt veel tõlkimata sõnade tõlkimise raskust.

4.5. Faktorid

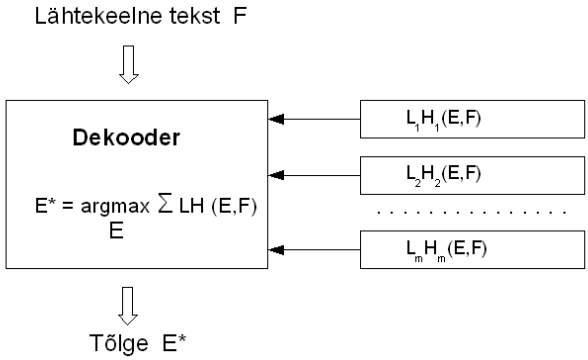
Eespool jätsime täpsustamata, kas ja kuidas kasutatakse SMT-s teadmist, et loomuliku keele väljenditel on oma struktuur – sõnad koosnevad morfeemidest, laused lauseliikmetest jne – ning et lingvistilist sisendit võib mitmel tasemel analüüsida, näiteks leida sõna lemma, kääne, sõnaliik, semantiline klass vms.

Olgu meil programm, mis suudab lähteteksti sõnajada teisendada mingiks lingvistiliste kategooriate jadaks, näiteks lause *Karu sööb mett* jadaks „karu sööma mesi" või „nimisõna tegusõna nimisõna" või „alus öeldis sihitis". Iga selline jada esindab esialgse sõnajada suhtes teatavat üldistust. Tõlkides tundmatut lauset, oleks otstarbekas selletaolisi üldisi keelestruktuurialaseid teadmisi ära kasutada. Võib ette kujutada, et üldistatud kategooriad annavad oma osa sorava tõlke leidmisesse, aidates ületada andmete hõredusest tingitud raskusi. SMT-süsteemide arendajad käsitlevad lingvistilisi kategooriaid faktoritena, mille arvestamine võib (ehkki ei pruugi) anda paremaid tõlkeid. Eri faktoritele (nt sõnavormile, lemmale, sõnaliigile) antakse mingid kaalud, mis määravad, kui palju dekooder arvestab ennustatavate tõlgetega, kui tõlkehüpoteese omavahel võrreldakse. Praktika on siiski näidanud, et kaugelt kõige olulisemateks faktoriteks on sellised, mis opereerivad sõnavormidega, mitte lingvistiliste kategooriatega.

Franz Josef Ochi ja Hermann Ney (2002) järgi esindab iga faktor $H_i(E,F)$ mingit tõenäosusfunktsiooni, mis annab oma osa (mida väljendab kaal L_i) tõenäosuse $P(E \mid F)$ kujunemisse: $P(E \mid F) = \sum L_i H_i(E,F)$.

Näiteks ülalkirjeldatud müraga kanali meetodit võib käsitleda faktoriseeritud mudeli erijuhuna, kus faktoreid on kaks ja korrutamine on asendatud logaritmide liitmisega: $H_1(E,F)=\log P(E)$, $H_2(E,F)=\log P(F \mid E)$, $L_1=L_2=1$.

Faktoriseeritud SMT põhimõtteline skeem on joonisel 3.



J o o n i s 3. Faktoriseeritud SMT-süsteemi põhimõtteline skeem.

Tegelikult võimaldab faktoriseeritud SMT ignoreerida müraga kanali teooriaga seotud meetodi nõuet, et tõlkides keelest F keelde E, peab tõlkemudel olema tagurpidine, s.t $P(F \mid E)$: kui iga faktorit pidada lihtsalt üheks liidetavaks, mis annab oma osa summaarsesse tõlke tõenäosusse, siis võib keelemudelina kasutada ikkagi tõenäosust $P(E \mid F)$, mis tundub ka intuitiivselt loomulik. Ka Google'i masintõlge on faktoriseeritud SMT (Hoang jt 2009).

4.6. Kolmanda keele kaudu tõlkimine

Mida teha sel juhul, kui tõlkimist vajavat sõna või fraasi ei õnnestu leida paralleeltekstide põhjal koostatud fraasitabelist? Üks ilmne võimalus on proovida tõlkida kolmanda keele vahendusel: näiteks eesti-läti korpusest võib tõlgitav fraas puududa, aga eesti-inglise ja inglise-läti korpuse kaudu võiks tõlge olla tuletatav. Eesti keele puhul on kolmandaks keeleks sageli soome keel. See seletab ka esmapilgul arusaamatut olukorda, et inglise keelest eesti keelde tõlkimisel pakkus Google, et *blank verse* on *silosäe*: inglise-eesti paralleelkorpusest puudus lause inglise fraasiga *blank verse* ja seega ei olnud võimalik midagi selle vasteks pakkuda. Inglise-soome paralleelkorpuses on aga vastav fraas olemas, ning oletades, et eesti ja soome keel on sarnased (üldiselt see ju nii ongi), pakkuski Google'i tõlkesüsteem blankvärsi vasteks soome sõna *silosäe*.

Kolmanda keele kasutamine võib aidata tõlke kvaliteeti parandada ka üldjuhul, mitte ainult puuduva fraasi tõlkimisel. Meenutame, et tõlkevaste leidmisel kombineeritakse tõlkemudelit, mis fraase tõlgib, ja keelemudelit, mis fraasid omavahel lauseks kombineerib. Kolmanda keele kaudu tõlkimine puudutab ainult tõlkemudelit, keelemudel sõltub ainult sihtkeelest. Võib juhtuda, et kolmanda keele kaudu tõlkides leitakse selline parafraas, mis sobitub sihtkeele lausesse paremini.

5. MT-süsteemide hindamine

Lähenemisviis, mis oletab, et tõlkimisprotsess on parimale võimalikule tõlkele võimalikult lähedase versiooni leidmine, vajab ka viisi, kuidas tulemust hinnata. Tõlkekvaliteedi parim hindaja on muidugi inimene. Tõlkeid hinnatakse kahe mõõtme, originaalilähedus ja soravuse järgi, mis vastavad tõlkesüsteemi kahele komponendile, s.t tõlkemudelile ja keelemudelile. Soravust hinnates otsustatakse, kui intelligentne, kui selge, kui loetav ja kui loomulik on MT väljund. Sageli kasutatakse hindamisel skaalasid (1–5).

Inimhindajate töö on aga kallis ja aeganõudev. Seetõttu on kasutusele võetud automaatseid hindamismeetodeid, näiteks BLEU (Papineni jt 2002), MaxSim (Chan, Ng 2008), RTE (Pado jt 2009), ULC (Giménez, Màrquez 2008). Kõigil juhtudel on aluseks idee, et hea MT väljund peaks olema väga sarnane inimtõlkega. Eeldatakse, et on olemas teatav kogus testlauseid ja inimeste tehtud tõlkeid. Testlausete tõlkimine on küll ajamahukas, aga loodetavasti saab tõlkeid testimisel korduvalt kasutada.

Hindamismeetodid erinevad üksteisest selle poolest, mille alusel nad tõlgete läheduse üle otsustavad. Näiteks BLEU hindab MT väljundit inimtõlgetega kokkulangevate n-grammide arvu kaalutud keskmise alusel. BLEU kasutab uni-, bi-, tri- ja sageli ka kvadrigramme. Peale selle liidab ta veel juurde karistuse liiga lühikeste tõlkekandidaatide eest. MaxSim arvestab peale n-grammide täpse kokkulangevuse positiivsena ka seda, kui kokku langeb ainult lemma, ainult sõnaliik või lemmaga sünonüümne sõna. RTE oletab, et kui MT väljundlausest on võimalik järeldada, et ka vastav inimese tõlgitud lause on tõene, või inimese tõlgitud lausest järeldub masina tõlgitud lause tõesus, siis on masina tõlge hea. RTE teisendab kummagi lause lihtsaks süntaktiliseks sõltuvuspuuks, leiab (leksikaalsete tunnuste järgi) teineteisele vastavad osad ja võrdleb neis osades kümnete lingvistiliste tunnuste väärtusi, näiteks süntaktilise sõltuvuse liiki (subjekti, objekti), modaalsust, eituse olemasolu, sünonüümiaseost. ULC on paljude muude automaatsete mõõdikute väärtuste aritmeetiline keskmine. Prantsuse-inglise ja hispaania-inglise katsetatud MT-süsteemide puhul olid kõigi mõõdikute ja inimeste hinnangud kooskõlas, ungari-inglise puhul aga hindas mõni mõõdik, näiteks BLEU, tõlkeid hoopis vastupidiselt. Kui tõlkesuund oli inglise keelest mõnda teise keelde, siis oli mõõdikute ja inimeste hinnangute kooskõla üldse palju halvem (Callison-Burch jt 2009).

Automaatsed hindamismeetodid sobivadki eelkõige selleks, et hinnata ühe MT-süsteemi järgnevaid versioone või sarnase ülesehitusega MT-süsteeme, kusjuures nii keelesuund kui ka kasutatav tekstikorpus peavad olema samad.

Erinevate masintõlkesüsteemide võistlused – milline neist tõlgib mingi etteantud valdkonna tekste kõige paremini – aitavad kiirendada paremate süsteemide loomist. Eri uurimisgrupid ja firmad üle maailma töötavad paralleelselt sarnaste probleemide kallal ja proovivad erinevaid lähenemisi, mille edukust võistlused siis kontrollivad. Uurimistöö paralleelsus pole seega ressursside raiskamine, sest võimaldab hoida kokku aega. Viimased võistlused näitasid, et statistilised MT-süsteemid, sh Google'i oma, kuuluvad oma kvaliteedilt maailma tippude hulka (Callison-Birch jt 2009). Samast uurimusest ilmneb, et prantsuse-inglise uudisteksti tõlke puhul arvasid hindajad ainult

iga teise lause puhul, et Google'i tõlge on vastuvõetav, ning ka muude keelte (hispaania, saksa, ungari, tšehhi) ja süsteemide puhul ei olnud vastuvõetavate lausete hulk suurem.

6. Lõpetuseks

SMT-süsteemi, mille üheks keeleks on eesti, võib teha kes tahes, ka see, kes ei oska sõnagi eesti keelt, kui tal on kasutada eesti keele korpus. Kuid praegu kasutatav SMT lähenemisviis sobib teatavat tüüpi keeltele paremini kui teistele. Fraasipõhine SMT on suures osas saksa keele selliste eripärasuste nagu pikkade liitsõnade, artiklite ja mitmesõnaliste verbide arvessevõtmise tulemus. Varasem, IBM-i prantsuse-inglise sõnapõhine lähenemisviis oli inglise-saksa keelepaari jaoks ilmselt liiga puudulik. Praeguse SMT jaoks on rasked rikkaliku morfoloogia ja vaba sõnajärjega keeled, näiteks soome-ugri ja slaavi keeled. Lihtne loogika ütleb, et tuleks leida viis, kuidas paralleelkorpusest tuletada sõnade segmenteerimisviis (morfoloogia) ning sõnadevahelised sõltuvused (süntaks), nii et need oleks just masintõlke jaoks optimaalsed. Kuid mõlemad on väga keerukad probleemid, mille jaoks häid algoritme praegu pole. Seetõttu on raske ennustada masintõlkesüsteemide võimete paranemise kiirust, ehkki mingi paranemine toimub kindlasti, kui ka praegusi meetodeid kasutatakse järjest suuremate tekstikorpuste toel.

Artikli valmimist on toetanud Euroopa Regionaalarengute Fond Eesti Arvutiteaduse Tippkeskuse kaudu ning Haridus- ja Teadusministeerium (sihtfinantseeritav teema sF0180078s08 „Loomulike keelte arvutitöötluse formalismide ja efektiivsete algoritmide väljatöötamine ning eesti keelele rakendamine" ja riiklik programm „Eesti keele keeletehnoloogiline tugi").

Kirjandus

Berger, Adam, Brown, Peter, Della Pietra, Stephen, Della Pietra, Vincent, Gillett, John, Lafferty, John, Mercer, Robert, Printz, Harry, Ureš, Luboš 1994. The Candide System for Machine Translation. – HLT '94. Proceedings of the Workshop on Human Language Technology. Plainsboro, NJ: Morgan Kaufman, lk 157–162.

Brants, Thorsten, Popat, Ashok, Xu, Peng, Och, Franz Josef, Dean, Jeffrey 2007. Large Language Models in Machine Translation. – Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL). Stroudsburg, PA: ACL, lk 858–867.

Brown, Peter, Della Pietra, Stephen, Della Pietra, Vincent, Mercer, Robert 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. – Computational Linguistics, kd 19, nr 2, lk 263–311.

Callison-Burch, Chris, Koehn, Philipp, Monz, Christof, Schroeder, Josh 2009. Findings of the 2009 Workshop on Statistical Machine Translation. – StatMT '09. Proceedings of the Fourth Workshop on Statistical Machine Translation. Stroudsburg, PA: ACL, lk 1–28.

C h a n, Yee Seng, N g, Hwee To 2008. MAXSIM. A Maximum Similarity Metric for Machine Translation Evaluation. – Proceedings of ACL-08: HLT. Stroudsburg, PA: ACL, lk 55–62.

G i m é n e z, Jesús, M à r q u e z, Lluís 2008. A Smorgasbord of Features for Automatic MT Evaluation. – Proceedings of the Third Workshop on Statistical Machine Translation. Stroudsburg, PA: ACL, lk 195–198.

H o a n g, Hieu, K o e h n, Philipp, L o p e z, Adam 2009. A Unified Framework for Phrase-Based, Hierarchical, and Syntax-Based Statistical Machine Translation. – Proceedings of the International Workshop on Spoken Language Translation. Tokyo, lk 152–159.

O c h, Franz Josef, N e y, Hermann 2002. Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. – Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Philadelphia, PA: ACL, lk 295–302.

P a d o, Sebastian, Ga l l e y, Michel, J u r a f s k y, Dan, M a n n i n g, Christopher 2009. Machine Translation Evaluation with Textual Entailment Features. – StatMT '09. Proceedings of the Fourth Workshop on Statistical Machine Translation. Stroudsburg, PA: ACL, lk 37–41.

P a p i n e n i, Kishore, R o u k o s, Salim, W a r d, Todd, Z h u, Wei-Jing 2002. BLEU. A Method for Automatic Evaluation of Machine Translation. – Proceedings of the 40th Annual Meeting on Association For Computational Linguistics. Philadelphia, PA: ACL, lk 311–318.

How does a Machine Translate?

Keywords: rule-based machine translation, statistical machine translation

The article gives an informal overview of machine translation. It describes both the classical rule-based paradigm and the statistical phrase-based paradigm. The aim is to present some assumptions about what is the adequate model of translation as a process, and what are the important typological aspects of language in this context.

Heiki-Jaan Kaalep (*b. 1962*), *PhD, University of Tartu, Institute of Computer Science, senior researcher, heiki-jaan.kaalep@ut.ee*

Mare Koit (*b. 1945*), *PhD, University of Tartu, Institute of Computer Science, professor of language technology, mare.koit@ut.ee*