



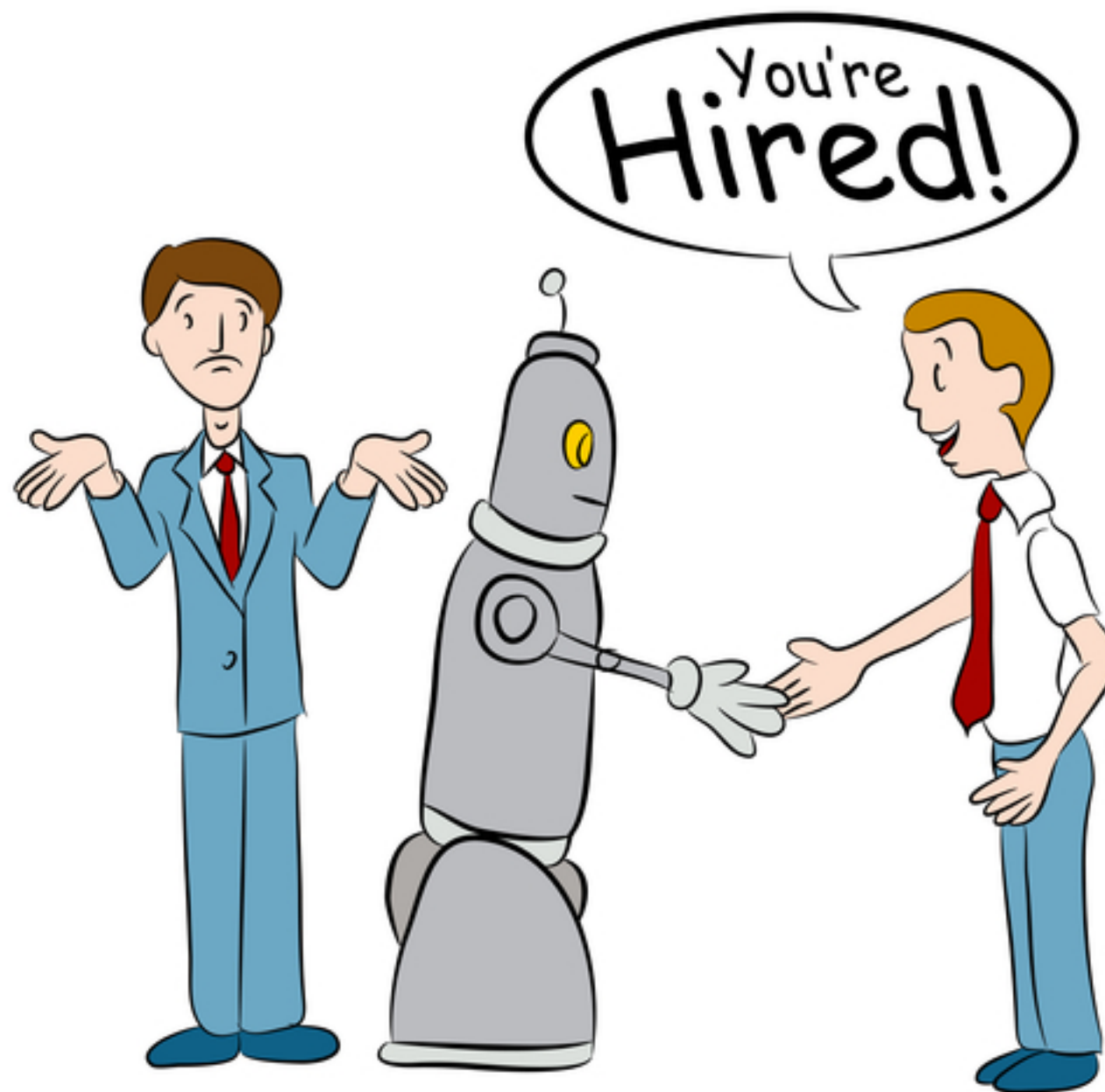
TARTU ÜLIKOOL

Masintõlge teoorias ja praktikas

Kairit Sirts

Tõlkimine ja tehnoloogia - 18.05.2017





Kava

- Kuidas töötab kaasaegne masintõlge?
- Hetkeolukord eesti-inglise-eesti masintõlkes
- Masintõlke potentsiaal tõlkeabina

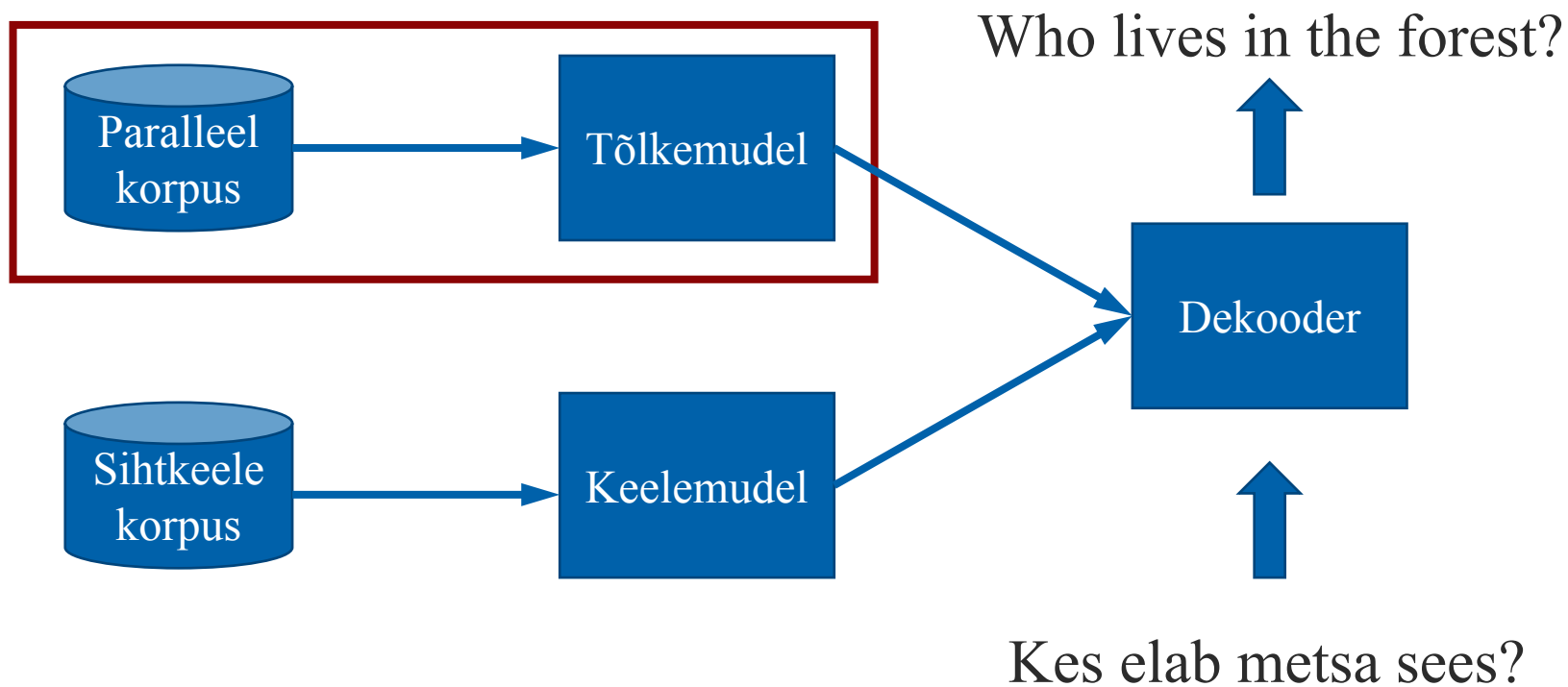
Masintõlke liigid

- Reeglipõhine masintõlge
- Statistiline masintõlge
- Närvivõrgupõhine masintõlge ehk neurotõlge

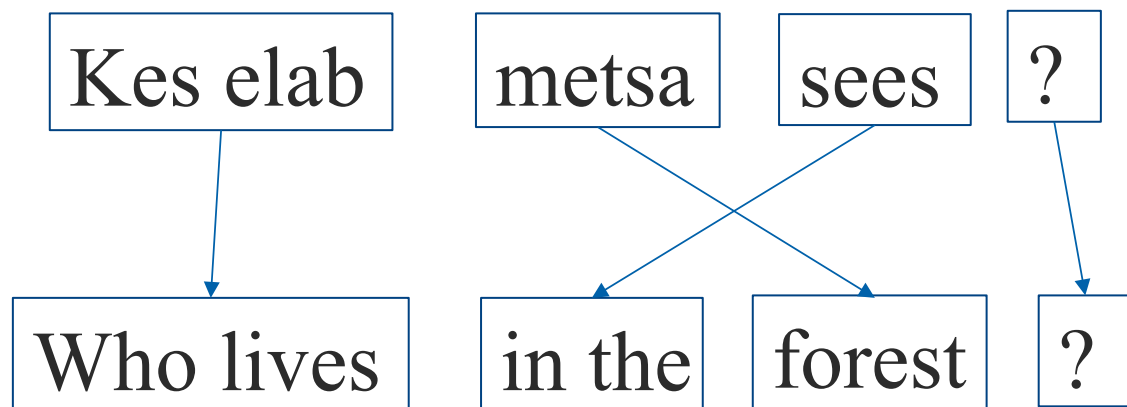
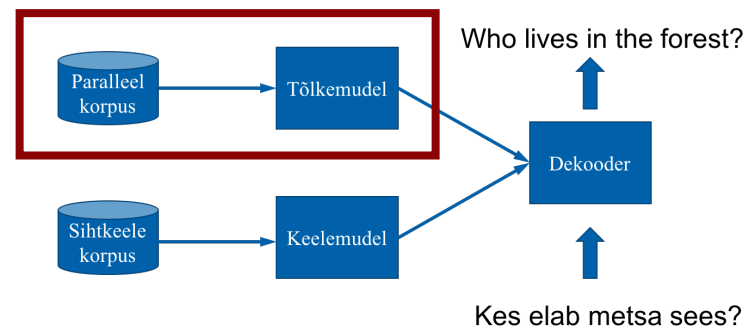
Masintõlke liigid

- Reeglipõhine masintõlge
- Statistiline masintõlge
- Närvivõrgupõhine masintõlge ehk neurotõlge

Fraasipõhine statistiline masintõlge

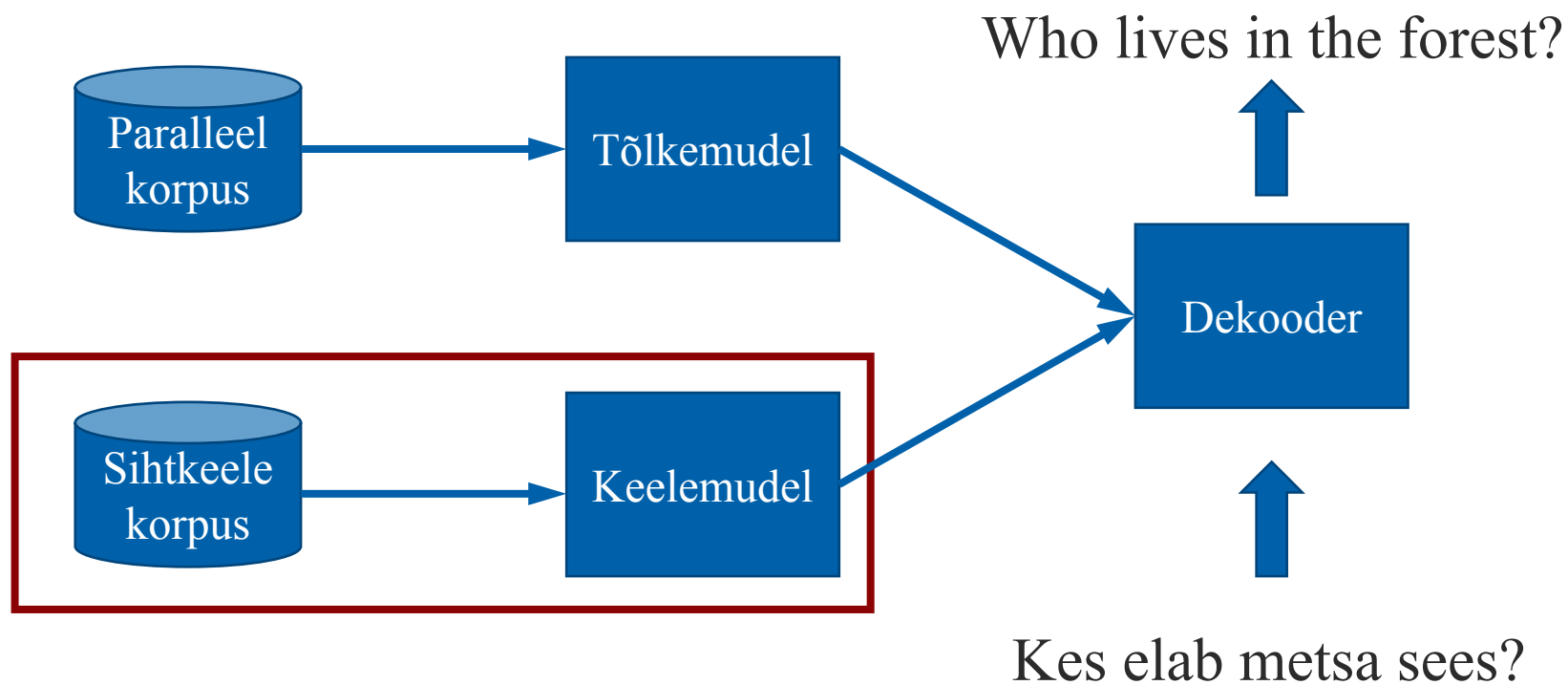


Tõlkemudel

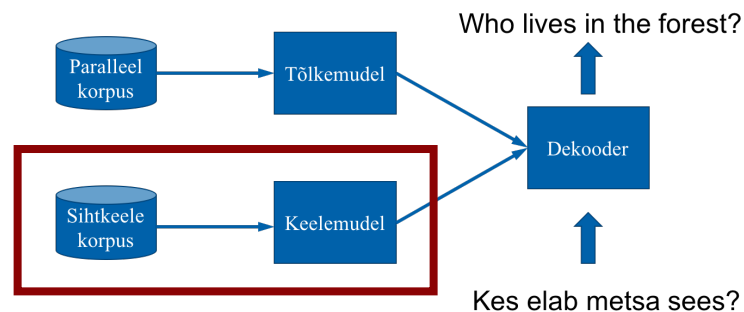


Lähtekeel	Sihtkeel	Tõenäosus
Kes elab	Who lives	0.95
metsa	forest	0.63
mets	forest	0.23
sees	in the	0.55
sees	in	0.12
sees	in a	0.05
sees	in an	0.07
...	...	...

Fraasipõhine statistiline masintõlge



Keelemudel

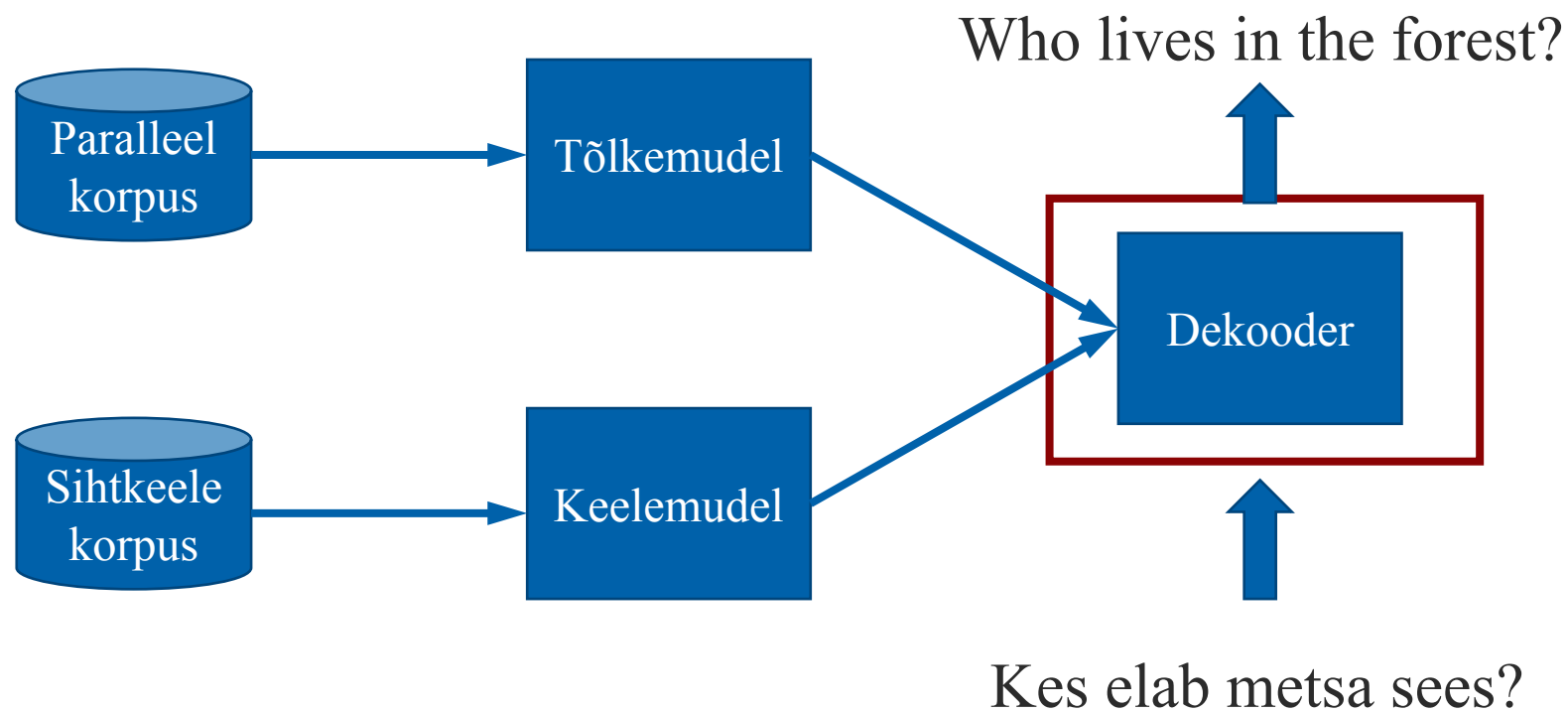


TARTU ÜLIKOOL

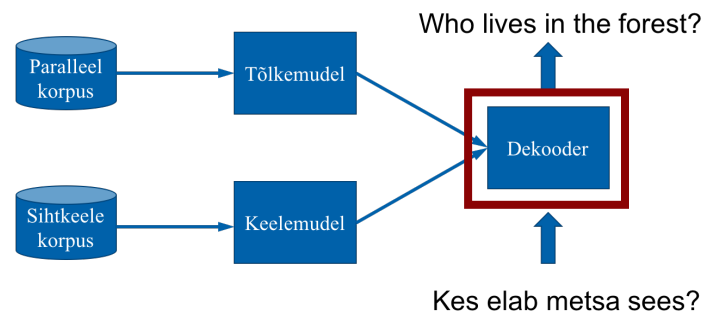
Who lives in the forest ?
 Who works in the forest ?
 Who lives in the house ?

Kontekst	Sõna	Tõenäosus
Who	lives	0.66
Who	works	0.33
lives	in	1.00
works	in	1.00
in	the	1.00
the	forest	0.66
the	house	0.33
forest	?	1.00
house	?	1.00

Fraasipõhine statistiline masintõlge



Dekooder



Kes elab metsa sees?

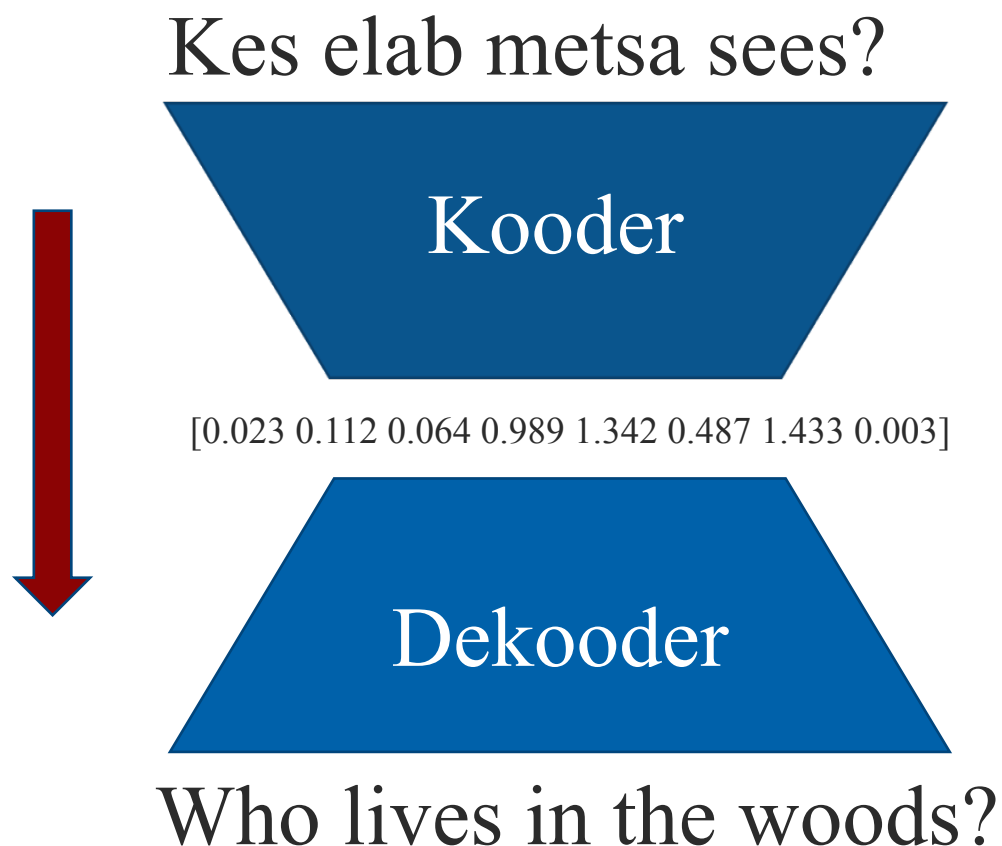


Kandidaat	Tõenäosus
Who lives in the forest?	0.77
Who lives in the woods?	0.74
Who lives in a forest?	0.25
Who lives in the wood?	0.05
Who lives in a house?	0.0004

Who lives in the forest?

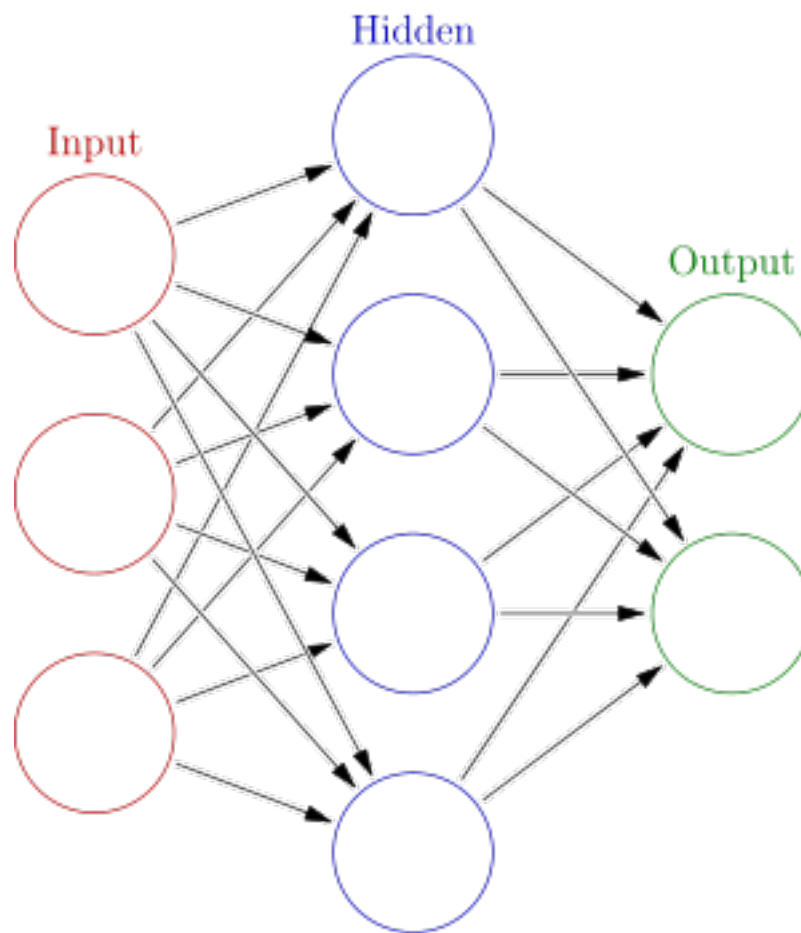


Närvivõrgupõhine masintõlge

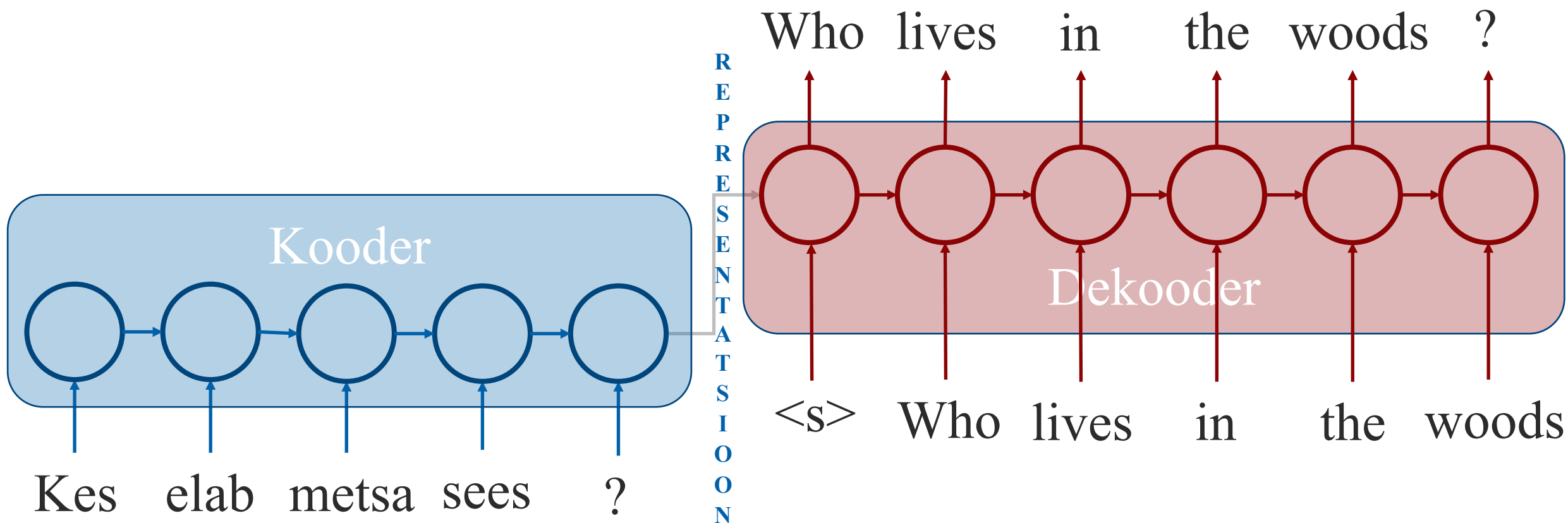


Lause mudelisisene
esitus ehk representatsioon

Tehisnärvivõrk



Kooder-dekooder arhitektuur



Kuidas õpetada närvivõrku tõlkima?

Mudeli parameetrid

$\begin{bmatrix} -0.07 & 0.29 & -3.11 \\ 0.54 & -0.06 & 0.93 \\ -0.10 & 0.10 & 0.70 \\ -2.27 & -1.09 & 0.98 \\ 0.94 & 0.76 & -0.34 \\ 1.34 & -0.46 & 0.63 \end{bmatrix}$
Kooder

$\begin{bmatrix} -0.51 & -2.78 & -0.64 \\ -0.73 & -0.07 & -0.72 \\ -1.85 & 0.27 & -0.30 \\ -0.26 & 0.26 & -0.30 \\ 0.44 & 0.31 & -0.37 \\ 1.97 & -1.36 & -0.54 \end{bmatrix}$
Dekooder

Kes elab metsa sees?



, , and is . . .



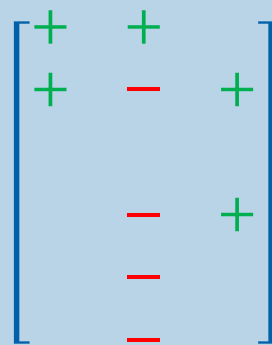
Kuidas õpetada närvivõrku tõlkima?

Who lives in the woods?

—

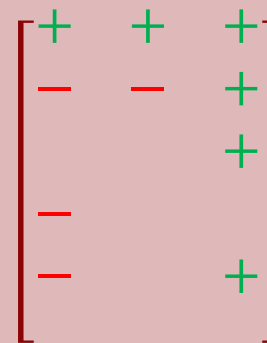
, , and is . . .

=



A 4x3 matrix with blue brackets. The first column has two green '+' signs. The second column has one green '+' and three red '-' signs. The third column has two green '+' signs.

Kooder



A 4x3 matrix with red brackets. The first column has one green '+' and three red '-' signs. The second column has two red '-' signs. The third column has three green '+' signs.

Dekooder

Kuidas õpetada närvivõrku tõlkima?

Mudeli parameetrid

Kes elab metsa sees?

Sisend

-0.06	0.30	-3.11
0.55	-0.07	0.94
-0.10	0.10	0.70
-2.27	-1.11	0.99
0.94	0.75	-0.34
1.34	-0.49	0.63

Kooder

-0.50	-2.77	-0.63
-0.75	-0.08	-0.71
-1.85	0.27	-0.29
-0.27	0.26	-0.30
0.43	0.31	-0.35
1.97	-1.36	-0.54

Dekooder

, , who is . . in the

Väljund

Kuidas õpetada närvivõrku tõlkima?

Mudeli parameetrid

Kes elab metsa sees?



Kooder		
-0.04	0.33	-3.12
0.61	-0.01	0.93
-0.02	0.05	0.66
-2.25	-1.10	1.01
0.88	0.72	-0.38
1.34	-0.45	0.67

Dekooder		
-0.50	-2.77	-0.63
-0.75	-0.08	-0.71
-1.85	0.27	-0.29
-0.27	0.26	-0.30
0.43	0.31	-0.35
1.97	-1.36	-0.54

Who lives in the woods?



Kokkuvõtteks

- Masintõlkemudeleid treenitakse paralleelkorpuste peal
- Mudel püüab õppida treeninglauseid võimalikult hästi tõlkima
- Treeningtekstide maht peab olema piisavalt suur – muidu mudel ei “õpi”, vaid “tuubib” treeninglausepaarid lihtsalt pähe
- Mudelite (eriti närvivõrgumudelite) treenimine on arvutuslikult ressursimahukas – see võib võtta terve nädala või isegi rohkem

Kava

- Kuidas töötab kaasaegne masintõlge?
- Hetkeolukord eesti-inglise-eesti masintõlkes
- Masintõlke potentsiaal tõlkeabina

Masintõlkesüsteemid

Ettevõtte/asutus	Aadress	Tehnoloogia
Tartu Ülikool	neurotolge.ee	närvivõrk
Tilde	translate.tilde.com	statistiline närvivõrk
Google	translate.google.com	statistiline
Microsoft	www.bing.com/translator	statistiline
Yandex	translate.yandex.com	statistiline
Baidu	translate.baidu.com	tõenäoliselt statistiline

Statistiline vs neurotõlge

- Kas neurotõlge on parem kui statistiline?
- Kui on, siis mille poolest?
- Kui on, siis kas alati?

Ma jätsin auto seisma, tulin välja ja võtsin päikesepriid eest.

TÜ NMT	I stopped the car, came out and took the sunglasses off.
TILDE NMT	I left my car, i came out, and took off my sunglasses.
TILDE SMT	I left to pull over, i come out and took off your sunglasses.
Google	I left the car stopped, got out, and took the sunglasses.

The new protocol now put to us by the Commission represents very rigorous cuts compared with the old one.

TÜ NMT	Uus protokoll, mille komisjon andis, esindab väga rangeid kärpeid võrreldes vanaonuga.
TILDE NMT	Komisjon esitab nüüd uue protokoll, mis esindab väga rangeid kärpeid võrreldes vana protokolliga.
TILDE SMT	Uus protokoll, mille komisjon meile nüüd esitab, sisaldab vanaga võrreldes väga karme piiranguid.
Google	Uus protokoll nüüd panna meile komisjon esindab väga range kärpeid võrreldes vanaga.

Kummaline on juulikuus mõeldagi, et aasta kõige valgemaal ajal on pimenemine juba alanud.

TÜ NMT	It's a little weird. it's just a little bit of a hair
TILDE NMT	It is a strange thing to imagine that the worst of the year in the mother of the year has already started.
TILDE SMT	It is strange to think that in July at the time of the most valgemaal pimenemine has already begun.
Google	Strangely, in July to think that by the time the lightest darkness has already begun.

BlackBerry AtHoc provides government agencies with the most secure cloud-based crisis communication solution available.

TÜ NMT	BlackBerry AtHoc võimaldab valitsusasutusi kõige turvalisemates pils- ja pilvelahendustes
TILDE NMT	BlackBerry AtHoc pakub valitsusasutusi kõige turvalisemate pilvepõhiste kriisilahenduste lahendustega.
TILDE SMT	BlackBerry AtHoc pakub valitsusasutustele kõige turvaline pilvepõhine kriisikommunikatsiooni lahendus saadaval.
Google	BlackBerry AtHoc pakub valitsusasutustele kõige turvalisem pilvepõhise kriisikommunikatsiooni lahendus olemas.

Statistilise ja neurotõlke sarnasused ja erinevused

- Neurotõlge genereerib grammatilisemaid lauseid kui statistiline masintõlge
- Neurotõlke ja statistilise masintõlke vead on erinevad:
 - Neurotõlge võib genereerida täiesti teise tähendusega “tõlkeid”, mis on samas grammatiliselt õiged
 - Statistiline masintõlge kasutab õigeid sõnu, aga paneb neid valesti kokku
- Mõlemad süsteemid oskavad ainult seda, mida nad on õppinud:
 - Europarlamendi kõnedel treenitud süsteem ei oska tõlkida ilukirjandust
 - Üldiselt jäävad masintõlke süsteemid hätta keerulisemate ja pikemate lausetega
- Neurotõlge on ca 2 aastaga jõudnud sinna, kuhu statistilisel masintõlkel jõudmiseks kulus 25 aastat

Kava

- Kuidas töötab kaasaegne masintõlge?
- Hetkeolukord eesti-inglise-eesti masintõlkes
- Masintõlke potentsiaal tõlkeabina

Masintõlke potentsiaal tõlkeabina

- Masintõlge ei suuda asendada inimest
 - Nüansi, konteksti ja emotsiooni tabamine
- Masintõlkesüsteemid võivad tõlketööd kiirendada ja lihtsustada
 - Madala kunstiväärtusega valdkonnad – tehnilised tekstid, tootekirjeldused vmt
 - Valdkonnale adapteeritud tõlkesüsteem
- Masintõlge on võrreldes inimtõlketööga odavam
 - Lihtsad toote- ja reklaamtekstid

Lõpetuseks

- Närvivõrgupõhised masintõlkemeetodid arenevad väga kiiresti
- Lähiajal masintõlge inimest ei asenda
- Masintõlge aitab kaasa keelte säilitamisele
- Tööjaotus inimese ja masina vahel
 - Inimene tõlgib keerulisemaid ja huvitavamaid materjale
 - Masintõlge aitab spetsiifiliste valdkondade tekste tõlkida
 - Masintõlget kasutatakse lihtsate tekstide tõlkimiseks