DAIVA ŠVEIKAUSKIENĖ
Lietuvių kalbos institutas

Neuroniniai tinklai ir lietuvių kalbos gramatika

Šiek tiek istorijos

Iki XX a. vidurio visos gramatikos ir žodynai buvo spausdinami ant popieriaus ir skirti naudotis tik žmogui. Prieš atsirandant kompiuteriams, kalba apskritai buvo laikoma vien tik žmonėms būdingu reiškiniu. Kitų gyvūnų naudojamos signalų sistemos buvo per daug paprastos, kad galėtų būti pavadintos kalbomis, o mechaniniai ir elektriniai prietaisai tegalėjo išsaugoti ir perduoti kodų sekas, kurias šifruodavo ir suprasdavo tik žmogus. 1942 m. Harvardo universitete buvo sukurtas pirmasis pasaulyje kompiuteris *MARK I*, ir labai greitai paaiškėjo, kad kompiuteriai gali apdoroti ne tiktai skaičius, bet ir kitus simbolius, pavyzdžiui, raides, taigi, jie gali apdoroti ir kalbas. Popieriuje spausdintos gramatikos, kaip ir žodynai, persikėlė į kompiuterinę terpę ir atsirado naujas jų vartotojas – kompiuteris. Dirbtinio intelekto srityje kalbų apdorojimas tapo viena pagrindinių pritaikymo sričių. Nuo 1968 iki 1978 m. gramatinė analizė buvo pagrindinė dirbtinio intelekto tyrėjų darbo tema.

Pirmoji kompiuterių pritaikymo kalboms sritis, sulaukusi daug dėmesio, buvo tekstų vertimas iš vienos kalbos į kitą. 1947 m. Vorenas Vyveris (Warren Weaver) pirmasis iškėlė mintį apie kompiuterių pritaikymą vertimo darbams: „Įdomu, ar nebūtų įmanoma sukurti kompiuterio, kuris verstų?“ 1948 m. Alanas Tiuringas (Turing), dirbtinio intelekto pradininkas, vardydamas būdus, kuriais gali pasireikšti kompiuterių „protas“, trečią iš eilės pamini vertimą. Nors automatinio vertimo ištakų galima būtų įžvelgti ir daug anksčiau. Pirmą kartą mechaninio vertimo idėjos buvo užfiksuotos XVII amžiuje. 1629 m. Renė Dekartas (René Descartes) siūlė rašyti knygas, sudarytas iš šifrų, o žodynuose visų kalbų atitikmenims turėjo būti suteiktas tas pats kodo numeris. 1661 m. pasirodžiusiame Johano Joachimo Becherio žodyne dešimčiai tūkstančių lotyniškų žodžių buvo suteikti kodai. Įdomu tai, kad ši knyga, pavadinta „Apie mechaninį kalbų vertimą: bandymas programuoti 1661 metais“, buvo perspausdinta praėjus 300 metų nuo jos pasirodymo, t. y. 1962 m. Tačiau surasti ekvivalentus graikų, hebrajų, vokiečių, prancūzų, slavų ir arabų kalbomis, kaip buvo numatęs autorius, pasirodė ne taip paprasta. Vėliau lingvistai suprato, kad kalbų skirtumai yra tokie dideli, jog jų negali apimti vien tik žodynai, kad ir kaip „logiškai“ jie būtų sudaryti. Tai liudija 1903 m. spaudoje pasirodžiusi Vilhelmo Rygerio (Wilhelm Rieger) skaitmenimis koduota gramatika „Skaitmeninė gramatika, kuri kartu su žodynais leidžia mechaniškai versti iš vienos kalbos į visas kitas“. Joje skaičiais koduojamos ne tik morfologinės kategorijos (linksnis, giminė, skaičius, laikas, asmuo ir pan.), bet ir sintaksinės, pavyzdžiui, veiksmažodžių tranzityvumas.

Statistiniai metodai

Tikriausiai niekas neginčys teiginio, kad kompiuteriai geriau už žmogų atlieka aritmetinius veiksmus. Jie puikiai susidoroja su formaliai aprašytomis užduotimis ir niekada nepadaro klaidų. Bet, pabandžius kompiuteriui pateikti kalbų apdorojimo ar kitos intelektinės srities darbus, paaiškėjo, kad tókias užduotis labai sunku aprašyti formaliai. Taigi, kalbos, lengvai suprantamos žmonėms, pasirodė sunkiai įkandamos kompiuteriams. Todėl labai gerų rezultatų automatinio kalbų apdorojimo srityje iki šiol dar nepasiekta.

Vorenas Vyveris, pirmasis iškėlęs mintį, kad kompiuteriai galėtų atlikti vertimus, 1949 m. siūlė automatinio vertimo problemas spręsti pasitelkiant statistiką. Tačiau anuomet mokslininkai greitai atsisakė tokių bandymų, matyt, dėl nedidelio tuometinių kompiuterių pajėgumo ir dėl nepakankamo kiekio tekstų, sukauptų elektronine forma. Po poros dešimtmečių, kai atsirado tekstynų ir taikant statistinius metodus buvo gauta gerų sakytinės kalbos atpažinimo rezultatų, vėl sugrįžtama prie tikimybių teorija paremto automatinio vertimo. Tam buvo itin palankios sąlygos Kanadoje, nes čia parlamento medžiaga saugoma ir anglų, ir prancūzų kalbomis (dvi valstybinės kalbos), todėl greičiausiai susikaupė pakankama lygiagrečių tekstų apimtis.

Taikant statistinius metodus, galima versti nesinaudojant nei gramatikos taisyklėmis, nei žodynu. Skirtingų kalbų žodžių atitikimas verčiant nustatomas iš dvikalbių lygiagrečiųjų tekstynų. Verčiant statistiniais metòdais, iš daugelio tekstynuose sukauptų vertimo variantų ieškoma labiausiai tikėtino. Tačiau labiausiai tikėtinas variantas ne visada būna tikslus, t. y. ne visada tas, kuris pavartotas nagrinėjamame tekste. Statistinio vertimo esmė ta, kad jis nebando generuoti vieno tikslaus vertimo, bet sudaro daug galimų vertimo variantų ir surikiuoja juos pagal tai, kiek tikėtina, kad kiekvienas jų yra tikslus.

Neuroniniai tinklai

Pastaruoju metu statistinį vertimą išstumia kita vertimo metodika – neuroniniai tinklai. Šio metòdo pagrindas – automatinis mokymasis. Naudojantis tradiciniu programavimu, kompiuteriui pateikiami pradiniai duomenys ir labai tiksliai nurodoma, kókias operacijas jis turi su jais atlikti, kad būtų gautas norimas rezultatas. Dirbant su neuroniniais tinklais nepasakoma, kaip turi būti sprendžiamas uždavinys; pateikiami pradiniai duomenys ir rezultatas, kuris turi būti gautas, o kompiuteris pats sudaro taisykles ir jas vėliau naudoja kitiems, naujiems, duomenims apdoroti. Kitais žodžiais tariant, kai taikomi tradicinio programavimo metòdai, geresnių rezultatų pasiekiama žmogui tobulinant programinės įrangos kodą, o, taikant automatinio mokymosi metòdą, programos parašomos taip, kad jos pačios pakoreguoja savo atliekamą darbą naudodamos vis daugiau gaunamų duomenų. Tačiau ir šiuo atveju ne viskas idealu – programoms ne visada pavyksta sėkmingai save patobulinti.

Kuriant neuroninius tinklus pagrindinis vienetas yra dirbtinis neuronas, t. y. matematinė funkcija, kuri yra suvokiama kaip biologinio neurono modelis. Dirbtiniai neuroniniai tinklai yra sudaryti iš tarpusavyje sujungtų dirbtinių neuronų. Jei neuroninis tinklas yra visai mažytis, jo veikimą suprasti galima, bet labai didelis, turintis šimtus sluoksnių ir tūkstančius neuronų sluoksnyje, tinklas žmogui tampa visai nesuprantamas. Neįmanoma tiesiog žvilgtelėti į giliojo neuroninio tinklo vidų ir pažiūrėti, kaip jis veikia. Pagal pateiktą programos kodą ir pàteiktus apmokymo duomenis (įvestį ir rezultatą) kompiuteris susikuria modelį – nustato neuroninio tinklo koeficientų réikšmes. Tų koeficientų paprastai būna labai daug, o jų prasmė neaiški – neaišku, kaip tam tikrą koeficientą ar koeficientus neuroniniame tinkle reikia pakeisti, kad jis veiktų geriau. Galima paminėti vieną sudėtingiausių neuroninių tinklų sistemų GPT-3, sukurtą 2020 m. Ji turi 69 sluoksnius, kurių kiekviename yra po kelias dešimtis tūkstančių neuronų, todėl susidaro apie 175 milijardus koeficientų. Šiai sistemai buvo naudojami superkompiuteriai, todėl apmokymui prireikė kelių mėnesių, su paprastais asmeniniais kompiuteriais tai būtų užtrukę daugiau nei 350 metų.

Neuroniniais tinklais modeliuojama žmogaus galimybė apibendrinti, t. y. tai, ką mes išmokome, galime panaudoti ateityje, panašiose, bet ne tokiose pačiose situacijose, tarkim, žmogus neturi iš naujo mokytis vairuoti, jei sėda prie kito automobilio vairo ar važiuoja kitomis gatvėmis nei mokymosi metu.

Naudojant neuroninius tinklus įvairioms atpažinimo užduotims atlikti buvo pasiekta gana gerų rezultatų. Bene labiausiai jų gebėjimai pasireiškė atpažįstant vaizdus ir sakytinę kalbą. Vaizdams atpažinti skirtų neuroninių tinklų struktūra ir veikimas buvo įkvėpti smegenyse vykstančių vaizdo atpažinimo procèsų, tačiau techninis šio metòdo įgyvendinimas yra iš esmės kitoks. Kompiuteriai paveikslėlius mato kitaip nei žmogus. Kompiuterių pasaulis sudarytas vien iš skaičių, todėl visi vaizdai koduojami skaičių rinkiniais dvimatėje erdvėje – plokštumoje. 1 pav. parodyta, kaip paveikslėlį mato žmogus ir kaip jis vaizduojamas kompiuterio vidiniame formate.

08	02	22	97	38	15	00	40	00	75	04	05	07	78	52	12	50	77	91	08
49	49	99	40	17	81	18	57	60	87	17	40	98	43	69	48	04	56	62	00
81	49	31	73	55	79	14	29	93	71	40	67	53	88	30	03	49	13	36	65
52	70	93	23	04	60	11	42	69	24	68	56	01	32	56	71	37	02	36	91
22	31	16	71	51	67	63	89	42	92	36	54	22	40	40	28	66	33	13	80
24	47	32	60	99	03	45	02	44	75	33	53	78	36	84	20	35	17	12	50
32	98	81	28	64	23	67	10	26	38	40	67	59	54	70	66	18	38	64	70
67	26	20	68	02	62	12	20	93	63	94	39	63	08	40	91	66	49	94	21
24	55	58	08	66	73	99	26	97	17	78	78	96	83	14	88	34	89	63	72
21	36	23	09	75	00	76	44	20	45	35	14	00	61	33	97	34	31	33	95
78	17	53	28	22	75	31	67	18	94	03	80	04	62	16	14	09	53	36	92
16	39	05	42	96	35	31	47	55	58	88	24	00	17	54	24	36	29	85	57
86	56	00	48	35	71	89	07	05	44	44	37	44	60	21	58	51	54	17	58
19	80	81	68	05	94	47	69	28	73	92	13	86	32	17	77	04	89	33	40
04	52	08	83	97	35	99	16	07	97	57	32	16	26	26	79	33	27	98	66
88	36	68	87	57	62	20	72	03	46	33	67	46	33	12	32	63	93	53	69
04	42	16	73	38	25	39	11	24	94	72	18	08	46	29	32	40	62	76	36
20	69	36	41	72	30	23	88	34	62	99	69	82	67	59	85	74	04	36	16
20	73	35	29	78	31	90	01	74	31	49	71	48	86	81	16	23	57	05	54
01	70	54	71	83	51	54	69	16	92	33	48	61	43	52	01	89	19	67	48

1 pav. Vaizdas, matomas žmogaus akimis ir kompiuterio

Šie skaičiai, nors ir beprasmiški žmogui, tačiau kompiuteriui yra vieninteliai prieinami duomenys atpažįstant vaizdus. Naujausi neuroninių tinklų laimėjimai yra tokie, kad jų atpažinimas beveik prilygsta žmogaus galimybėms, tad natūraliai iškilo klausimas, kuo skiriasi žmogaus ir neuroninių tinklų matymas. Buvo atlikti du tyrimai, kuriuose analizuota, kuo skiriasi žmogaus ir neuroninių tinklų vaizdo atpažinimas. Pirmajame buvo parodyta, kad labai nedideli, žmogui nepastebimi vaizdo pakeitimai (trikdžiai) gali sukelti neuroninių tinklų klaidas (neatpažįstamas žmogui gerai žinomas objèktas). Ir tai nėra susiję su mokymo problemomis, nes tas pats trikdis sukelia klaidingą atpažinimą įvairiems tinklams, kurie buvo apmokomi naudojant ne tuos pačius duomenų rinkinius. 2 pav. pateiktas automobilių atpažinimo pavyzdys.

2 pav. Automobilių atpažinimo klaidos

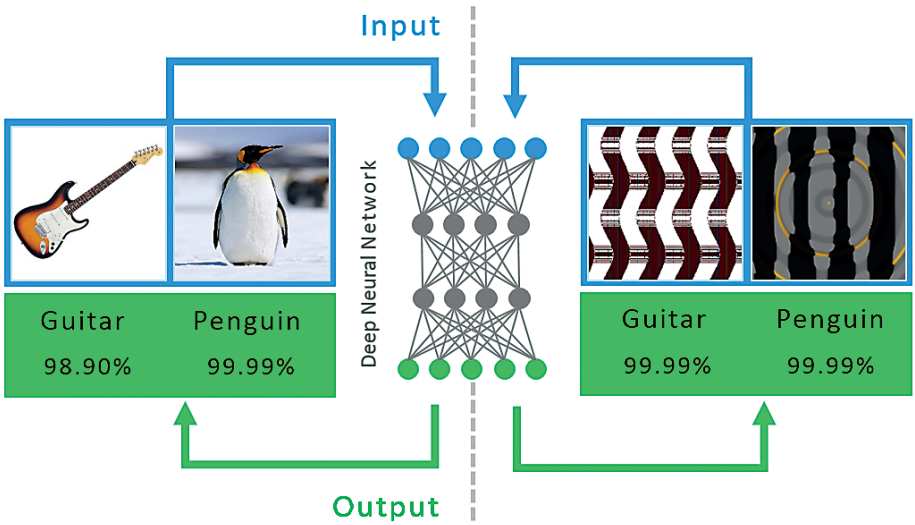
Kairėje pusėje esanti automobilio nuotrauka buvo atpažinta kaip automobilis, tačiau vidurinė nuotrauka nebuvo priskirta automobiliams. Dešinysis paveikslėlis rodo skirtumus tarp kairiajame ir viduriniame paveikslėlyje esančių automobilių, kurių žmogus pačiose nuotraukose pastebėti negali.

Atliekant antrąjį tyrimą bandyta parodyti klaidų atsiradimą kita kryptimi. Paaiškėjo, kad labai nesunkiai galima sukurti paveikslėlius, kurie žmogui yra visai nesuprantami, bet neuroninis tinklas su labai didele tikimybe juos laiko gerai pažįstamais daiktais. 3 pav. parodyta, kad neuroninis tinklas, gerai atpažįstantis gitarą ir pingviną, tiems patiems objèktams priskiria žmogui visai nepanašius į juos vaizdus.

Reikia pasakyti, kad neuroniniams tinklams bent jau kol kas geriau sekasi atpažinti objektus nuotraukose negu atlikti kalbų apdorojimo užduotis.

Automatinis vertimas

Viena naujausių neuroniniais tinklais besiremiančių automatinio vertimo sistemų yra *eTranslation*. Ji skirta teisės tekstams versti. Gana išsami šios sistemos veikimo analizė buvo pateikta 2019 m. vykusiame antrajame ELRC (Europos kalbų išteklių koordinavimo) seminare Lietuvoje. Be įvardytų pagrindinių šios vertimo sistemos pranašumų, kad *eTranslation* „dažnai pateikia gerus ar mažai taisytinus išversto teksto gabalus; beveik taisyklinga gramatika (tinkamos žodžių formos); atsižvelgia



3 pav. Klaidingas objektų priskyrimas pingvinų ir gitarų klasėms

į kontekstą“, buvo nurodyti ir kol kas dar pastebimi trūkumai: terminijos nenuoseklumas; iš piršto laužti (pačios sistemos susikurti) žodžiai, terminai, gramatinės formos, pavyzdžiui, *dish washing machines (indaplovės)* buvo išversta taip: *diškių skalbimo mašinos*. Pačiõs sistemos mokymasis buvo stebimas analizuojant gautų rezultatų pokyčius, kai versti buvo pateikiamas tas pats sakinys po tam tikro laiko (po kelių mėnesių). Antrajame etape gautas jau kitas vertimo variantas, tačiau ne geresnis už pirmąjį: *išk plovimo mašinos*.

Gramatinė analizė

Neuroninių tinklų pagrindu veikiančios morfologinės ir sintaksinės analizės sistemos taip pat kartais pateikia klaidingų rezultatų. Kodėl taip atsitinka, galima paaiškinti pavyzdžiu, kaip nustatomi lietuviškų raidžių diakritiniai ženklai. Jei tekstas parašytas lotynų abėcėlės raidėmis ir lietuvių kalbos žodžiuose gali būti keli diakritinių ženklų variantai (tarkim, du – „sunelis susirgo“ gali būti „sūnelis susirgo“ ir „šunelis susirgo“, arba net trys – „rastas“ gali būti „raštas“, „rąstas“ ir „rastas“) visada bus imamas tik tas variantas, kuris apmokymo duomenyse pasitaikė daugiau kartų, t. y. jis yra labiau tikėtinas. Tačiau taip pasirinktas variantas gali būti ne tas, kuris iš tikrųjų pavartotas tekste.

Tinklalapis *Parts-of-speech.Info* kiekvienam vartotojo įvestam žodžiui nurodo kalbos dalį. Programinės įrangos pagrindą sudaro Stanfordo universiteto morfologinis analizatorius *Stanford University Part-Of-Speech-Tagger*. Jo veikimas remiasi neuroniniais tinklais, o šalia teksto analizės pateikiama pastaba, kad 100 proc. tikslumas gali niekada ir nebūti pasiektas. Vokiečių kalbos apmokymo duomenis šiam analizatoriui

pateikė Zarlando universitetas. Jie apima 50 morfologinių žymų. Tačiau paaiškinama: „Automatinio mokymosi metòdas veikia taip, kad, šiurkščiai tariant, programinė įranga šeriama tekstais, kuriuose žmogaus yra nurodyta, kuriai kalbos daliai koks žodis priklauso. Tad programinė įranga iš gautų duomenų turi pati susikurti taisykles (pavyzdžiui, kad po artikelio su labai didele tikimybe eina būdvardis arba daiktavardis) ir privalo nurodyti kalbos dalį net ir tiems žodžiams, kurių ji niekada nėra „mačiusi“. Kompiuteriui tai yra labai sunki užduotis, ir todėl nustatant kalbos dalį klystama.“

Morfologinę bei sintaksinę lietuvių kalbos sakinių analizę gali atlikti sistema *UDPipe*, veikianti neuroninių tinklų pagrindu. Kol lietuvių kalbos sakinio struktūra panaši į anglų kalbos, t. y. jame pavartota tipiška anglų kalbos žodžių tvarka (veiksnys – tarinys – netiesioginis papildinys – tiesioginis papildinys, tarkim, *I give her an apple*), morfologinis analizatorius dirba gerai. Sakinys *Mama padovanojo man mažą šuniuką* išnagrinėjamas be klaidų. Tačiau jei sakinyje yra lietuvių kalbai būdinga laisva, neangliška žodžių tvarka, padaroma nemažai klaidų. Atrodytų, toks nesudėtingas sakinys *Lauke sninga* išanalizuojamas neteisingai: žodžiui *sninga* kaip pradinė forma nustatoma: būdvardžio vardininkas *sningas*, o kaip tekste pavartota forma nurodoma jo bevardė giminė. Matyt, pagal analogiją su *lauke gera* (šiuo atveju tai išties būdvardžio bevardė giminė, o pradinė forma gaunama atmetus galūnę *-a* ir pridėjus vardininko galūnę *-as*: *geras*) buvo gauta ir pradinė forma *sningas*. Panašu, kad apmokymo duomenyse nebuvo žodžio *snigti* todėl kompiuteris, naudojosi tuo, ką turi, kas jam buvo pateikta, ir darė savas išvadas iš turimų duomenų. Panaši situacija buvo ir analizuojant sakinį *Mokytojas įėjo ir vaikai atsistojo*. Kaip žodžio *įėjo* pagrindinė forma (bendratis) pateikiama *įėti*. Tikėtina, kad *UDPipe* apmokymo duomenyse nebuvo žodžio *įeiti* bei jo formų. Analizatorius pagal susikùrtas pradinės formos nustatymo taisykles sugeneravo šio žodžio bendratį remdamasis analogija su kitais veiksmažodžiais, kaip *kalbėjo – kalbėti*, *stovėjo – stovėti*. Sakinyje *Liūdną jis mums pranešė žinią: šuo nebegrįžo į namus* žodžiui *šuo* nurodyta, kad tai įvardžio įnagininkas (panašiai kaip *tuo*), o vardininko forma nurodyta *šas* (pagal analogiją su *tas*). Taigi, apmokymo duomenyse, iš visko sprendžiant, nebuvo ir žodžio *šuo*.

Tad čia pritartina lietuvių kalbos kompiuterizavimo srityje dirbusio Vido Daudaravičiaus teiginiui: „Naivu manyti, kad metòdai, kurie sėkmingai taikomi anglų kalbai, tinka ir kitoms kalboms.“

Pokalbių robotas

Visai neseniai pasirodė dirbtinio intelekto prodùktas – pokalbių robotas *ChatGPT*, naudojantis neuroninius tinklus. Paprašius jo išanalizuoti jau minėtą sakinį *Liūdną jis mums pranešė žinią: šuo nebegrįžo į namus*, rezultatai buvo dar blogesni nei *UDPipe*. Apie žodį *jis ChatGPT* pateikė tokią informaciją: daiktavardis, vardininkas, vienaskaita, vyraslystė. Apie žodį *mums* pateikta informacija buvo: įvardis, kilmininkas, dau-

giskaita, bendratis. Sintaksinė šio sakinio analizė buvo tokia: *Liūdną jis mums pranešė žinią* – šis sakinys yra pagrindinis sakinys, kuriame yra veiksmažodžio *pranešė* dalyvis *liūdną*, su daiktažodžiu *žinią*, priešdalyvio *jis* ir įvardžio *mums*; *šuo nebegrįžo į namus* – tai priklausomas sakinys, kuriame yra veiksmažodžio *nebegrįžo* dalyvis *šuo* su priešdalyvio *-ne*, vietininko *į* ir vietininko *namus*. Taigi, visą sakinį galime suprasti kaip pranešimą apie tai, kad liūdnas asmuo (jis) mums pranešė žinią, jog šuo nebegrįžo į namus.

Matome, kad gramatinėje analizėje atsiranda naujų terminų: kas yra „vyraslystė“, „daiktažodis“, „priešdalyvis“, tektų dar pasiaiškinti. Nauja yra ir tai, kad *jis* yra daiktavardis; veiksmažodžio *nebegrįžo* dalyvis yra *šuo*; žodis *mums* yra bendratis ir kt.

Apibendrinant galima pasakyti, kad neuroniniais tinklais besiremiantys dirbtinio intelekto prodùktai kol kas negali atlikti rimtų lietuvių kalbos gramatinės analizės uždavinių, reikalaujančių didelio tikslumo. Prieš imantis tokių uždavinių, turi būti atlikta nuoseklių mokslinių parengiamųjų tyrimų.

Literatūra

Daudaravičius V. 2012: Teksto skaidymas pastoviųjų junginių segmentais: daktaro disertacijos santrauka. Kaunas: Vytauto Didžiojo universitetas. Prieiga internete: *https://www.vdu.lt/cris/bitstream/20.500.12259/124748/1/vidas_daudaravicius_dd.pdf*

Hutchins J., Sommers H. 1992: An Introduction to Machine Translation. London: Academic Press.

Šveikauskienė D. 2022: Lietuvių kalbos gramatikos kompiuterizavimas. Mokslo studija. Vilnius: Lietuvių kalbos institutas. Prieiga internete: *http://lki.lt/wp-content/uploads/2023/01/Daiva-Sveikauskiene-Mokslo-studija-2022-pataisytas-leidimas.pdf*

Žaliauskas N. 2017: Kalbėtojo atpažinimas naudojantis dirbtiniais neuroniniais tinklais: baigiamasis bakalauro darbas. Vilniaus universitetas. Prieiga internete: *http://talpykla.elaba.lt/elaba-fedora/objects/elaba:23159558/datastreams/MAIN/content*