Dziļie neironu tīkli, vārdu reprezentāciju mācīšanās un teksta apstrāde

Artūrs Znotiņš, arturs.znotins@gmail.com

Darba vadītājs: Gunts Bārzdiņš, Dr.sc.comp.

LATVIJAS UNIVERSITĀTE
ANNO 1919

Mērķi

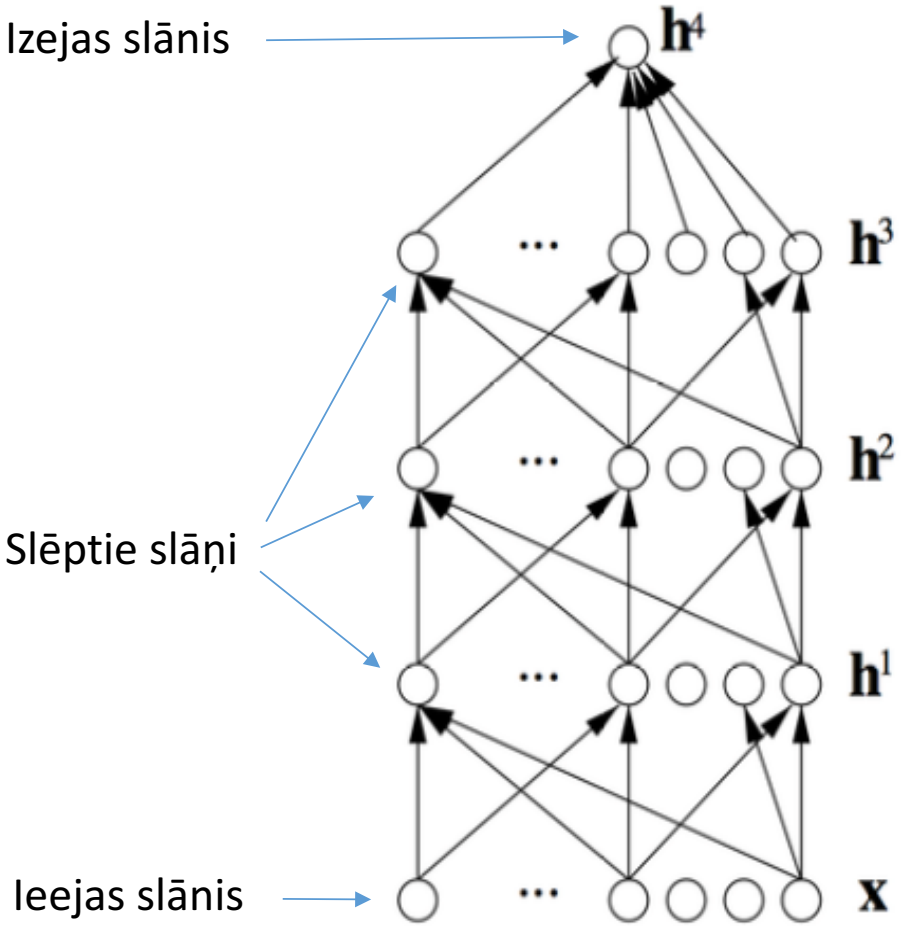
- Dziļās mašīnmācīšanās (*deep learning*) priekšrocības un ierobežojumi
- Teksta reprezentācijas un to pielietojumi dabīgās valodas apstrādē (NLP)
- Ierobežoti resursi (apmācības dati un to sagatavošana)

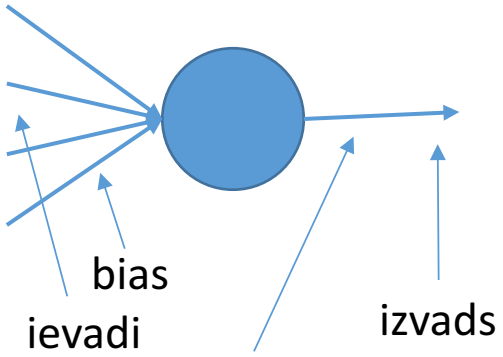
Saturs

- Neironu tīkli
- Ievads par vārdu vektoriem
- Latviešu valodas vārdu vektori un pielietojumi
- Secinājumi

Dziļie neironu tīkli

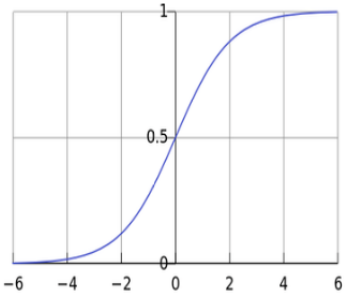
Dziļo neironu tīklu arhitektūra





$$h_{w,b}(x) = f(w^T x + b)$$

Nelinearitāte: $f(z) = \frac{1}{1 + e^{-z}}$



Dziļo neironu tīklu priekšrocības

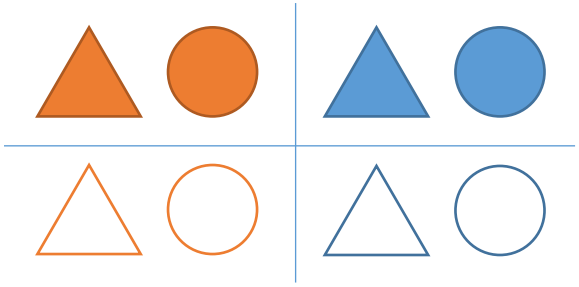
- Reprezentāciju mācīšanās (*end-to-end*)
- Vairāku līmeņu reprezentācijas
- Vajadzība pēc distributētām reprezentācijām (dimensiju lāsts)
- Neuzraudzīta pazīmju mācīšanās
- *Transfer-learning*

- Labāka ātrdarbība ar blīviem vektoriem/matricām uz vairākkodolu CPU, GPU

Vārdu reprezentācijas

Ievads: vārdu reprezentācijas

- Zema līmeņa reprezentācijas
  - Attēls –> pikseļu intensitāte
  - Audio –> spektrālā blīvuma koeficienti
  - Vārds –> ?
- Tradicionāli vārds tiek uzskatīts par diskrētu vērtību
  - Nav informācijas par vārdu attiecībām (ierobežoti ārēji resursi: sinonīmu kopas, taksonomijas)
  - Datu retuma problēma (*data sparsity*)



1-no-N vektors (*one-hot*):

Kaķis [0 0 0 0 0 0 0 0 0 1 0 0 0 0]

Suns [0 0 0 0 0 0 0 0 0 1 0 0 0 0]

Ievads: vārdu reprezentācijas

- **Līdzīgi vārdi tiek lietoti līdzīgos kontekstos**

  "*You shall know a word by the company it keeps*" (Firth, J. R. 1957)

  Vīrietis no rīta lasīja ***avīzi/laikrakstu****.*
  Viņš pērk šo ***avīzi/laikrakstu*** *katru dienu.*

- Varam reprezentēt vārdu ar tā kaimiņiem

Ievads: vārdu reprezentācijas

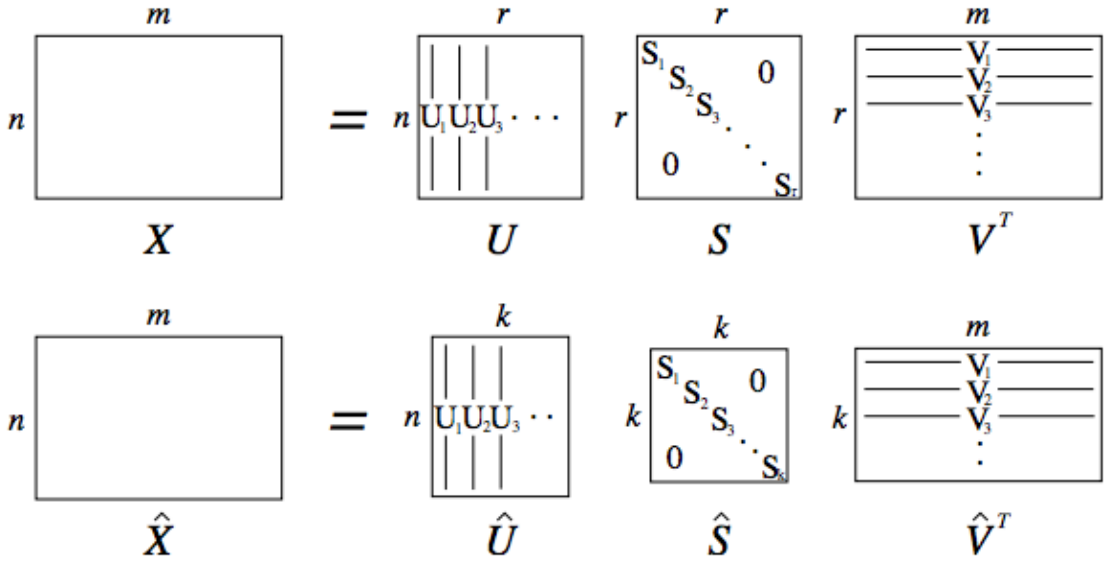
Blakus-sastopamību matrica

- I like deep learning.
- I like NLP.
- I enjoy flying.

counts	I	like	enjoy	deep	learning	NLP	flying	.
I	0	2	1	0	0	0	0	0
like	2	0	0	1	0	1	0	0
enjoy	1	0	0	0	0	0	1	0
deep	0	1	0	0	1	0	0	0
learning	0	0	0	1	0	0	0	1
NLP	0	1	0	0	0	0	0	1
flying	0	0	1	0	0	0	0	1
.	0	0	0	0	1	1	1	0

Ievads: vārdu reprezentācijas

- Risinājums: zemas dimensionalitātes blīvi vektori (*word vectors, embeddings*)
- Tipiski 25 – 1000 dimensijas
- *Singular Value Decomposition* (SVD), LDA, RI, …



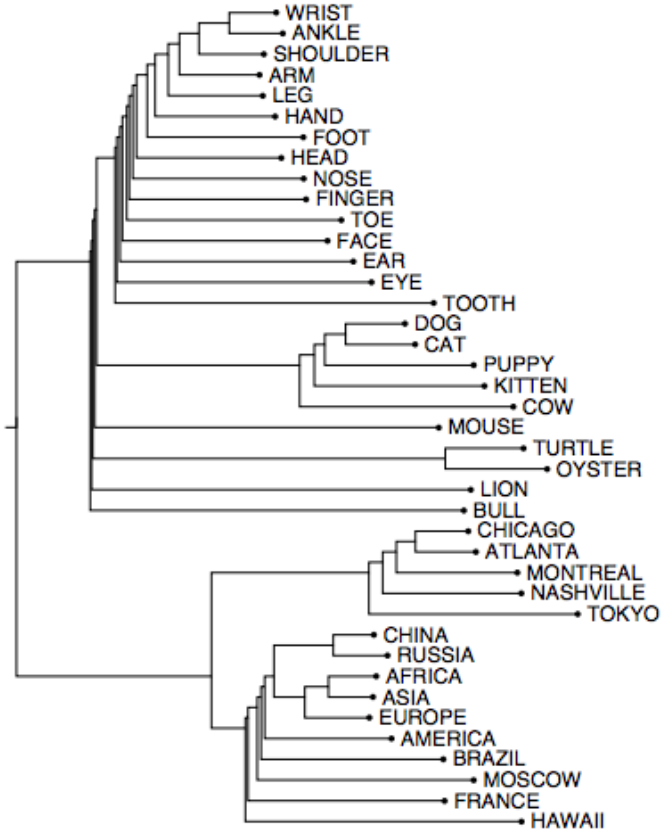
Kaķis =

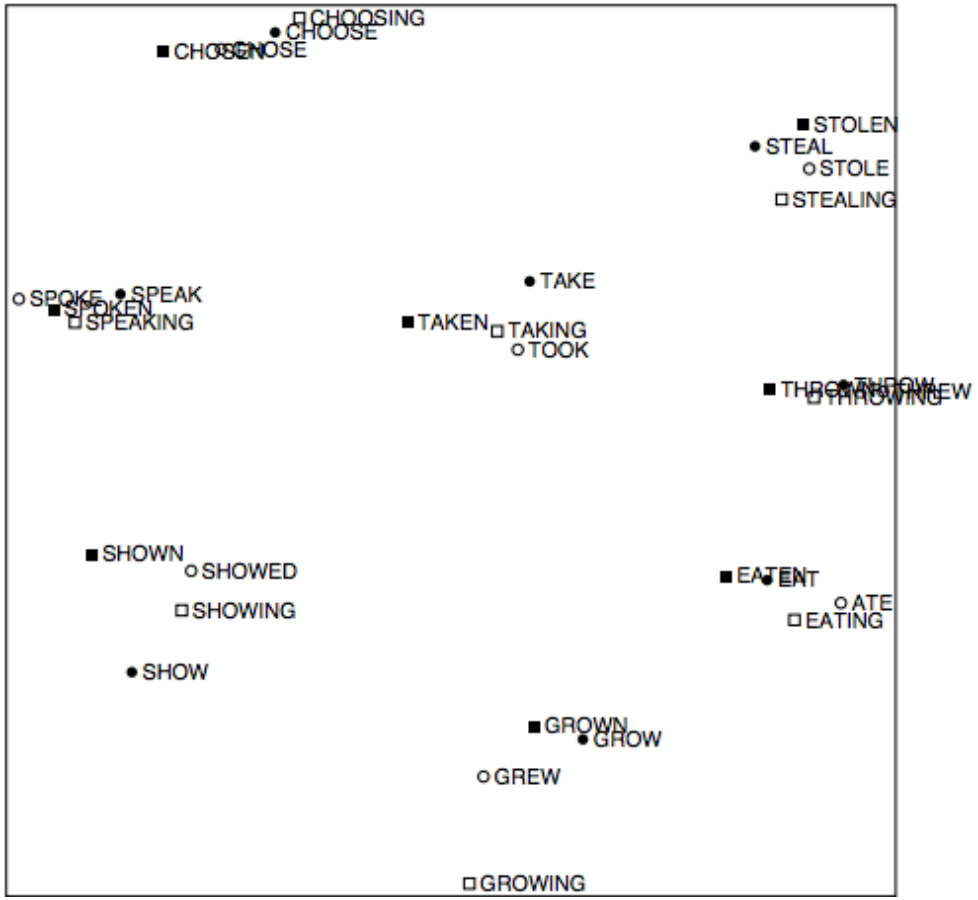
0.12
0.56
-0.34
0.03
0.23
-0.95

$\hat{X}$ - labākā k-ranka aproksimācija matricai X

$$PMI(w,w') = \log \frac{p(w,w')}{p(w)p(w')} \qquad TFIDF(w,z) = freq_{w,z} \cdot \log \frac{V}{n_z}$$

Ievads: vārdu reprezentācijas





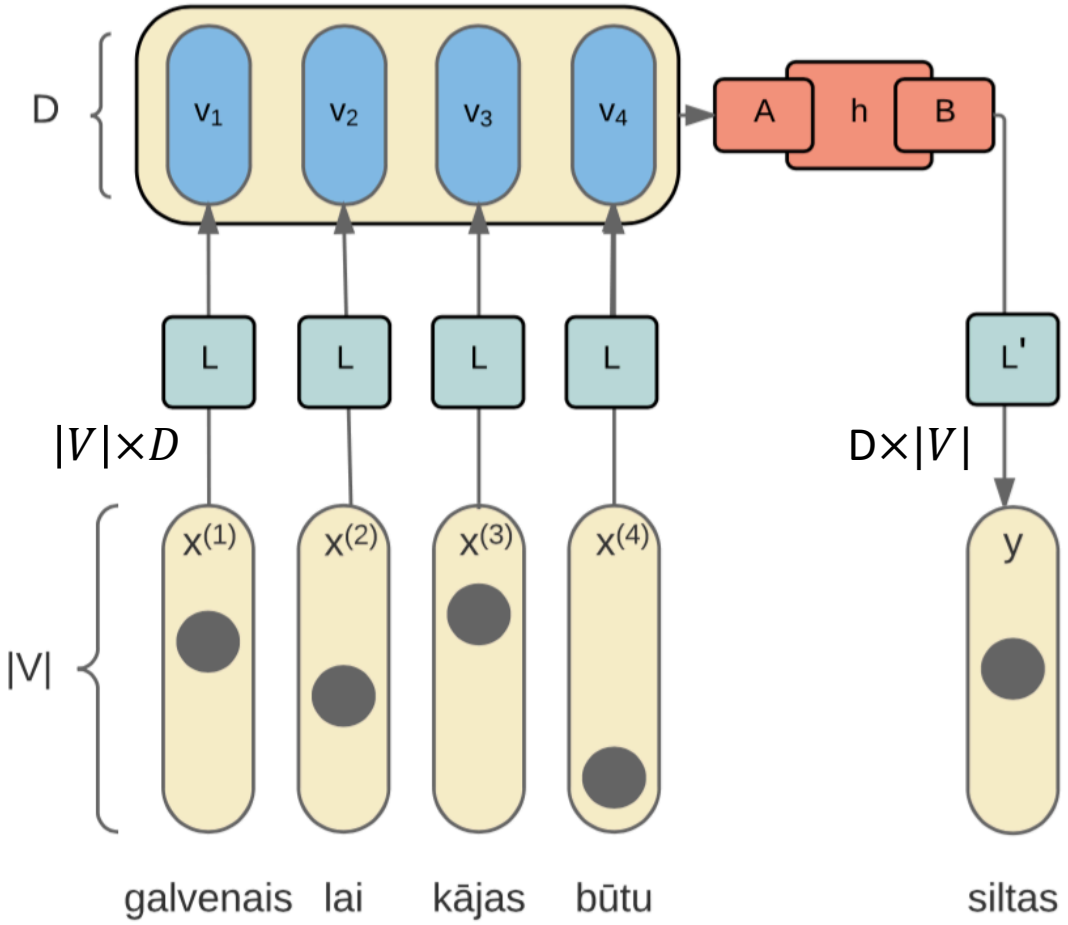
[Rohde et al., 2005, An Improved Model of Semantic Similarity Based on Lexical Co-Occurrence]

Ievads: vārdu reprezentācijas

Tiešā veidā iemācīties vārdu vektorus:

- Learning representations by back-propagating errors. (Rumelhart et al., 1986)
- A neural probabilistic language model (Bengio et al., 2003)
- NLP (almost) from Scratch (Collobert & Weston, 2008)
- **word2vec** (Mikolov et al. 2013)

FFNN valodas modelis



L, L' – ieejas un izejas vektori
D - vektoru dimensija
V – vārdnīca

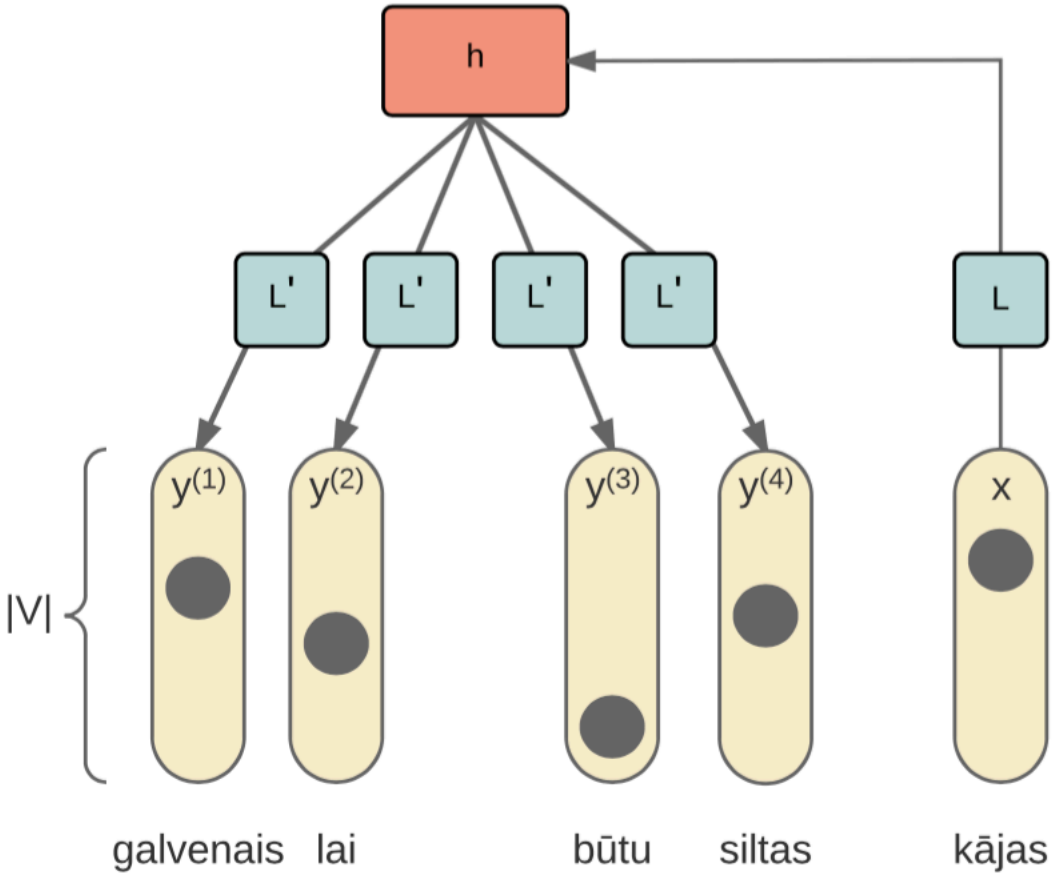
$$p(w_t|w_{t-1}, \dots, w_{t-K+1}) = \frac{\exp(y_{w_t})}{\sum_{w \in V} \exp(y_w)}$$

$$h = \sigma(Av_{t-K+1}^{t-1} + b_a)$$
$$y' = Bh + b_b$$

word2vec

- Aizmirstam par blakussastopamību skaitiem un mēģinām paredzēt konteksta vārdus
- Dažādas metodes un optimizācijas, kas ļauj iemācīties no ļoti liela tekstu korpusa (*unsupervised*)
- Saglabā dažādas attiecības starp vārdiem:
  - *v(krastmala) ≈ v(piekraste)*
  - *v(karaliene) ≈ v(karalis) - v(vīrietis) + v(sieviete)*
  - *v(grāmata) - v(grāmatas) ≈ v(ābols) - v(āboli)*

Skip-grammu modelis



$$v_x = x^T L$$

$$s^{(k)}\left(x, y^{(k)}\right) = u_{y^{(k)}}^T v_x, \mathrm{k} \in 1..K$$

$$\max p(y|x) = \max \log p(y|x)$$

$$= \max \log \frac{\exp\left(s(x,y)\right)}{\sum_{w \in W} \exp\left(s(x,w)\right)}$$

Mērķa funkcija:

$$J(L, L') = \log \frac{\exp\left(u_y^T v_x\right)}{\sum_{w \in W} \exp\left(u_w^T v_x\right)}$$

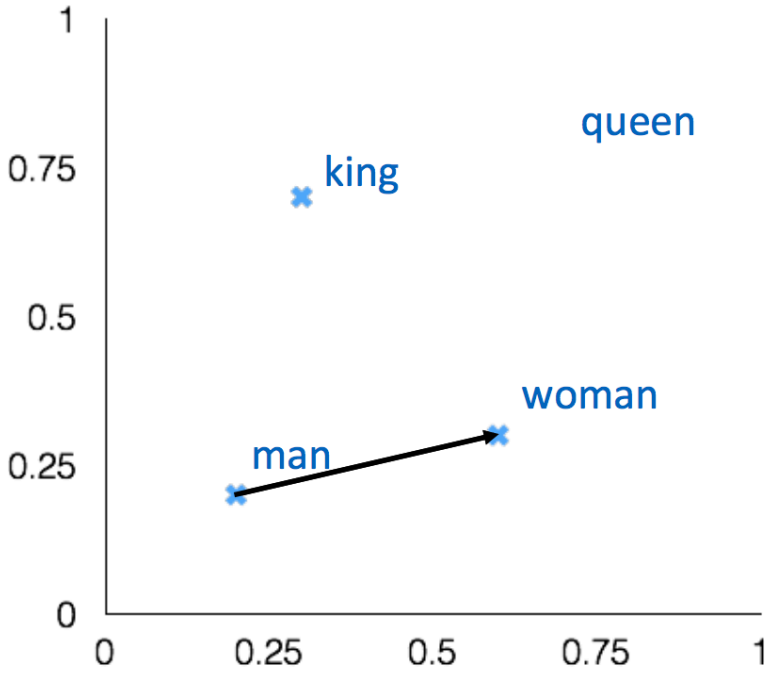
Vārdu analoģiju piemērs

a:b :: c:?

$$d = \arg\max_x \frac{(w_b - w_a + w_c)^T w_x}{||w_b - w_a + w_c||}$$

man:woman :: king:?

+	king	[0.30 0.70]
-	man	[0.20 0.20]
+	woman	[0.60 0.30]
	queen	[0.70 0.80]



Kā tas strādā?

X = King – man + woman

$$\arg\max_x \cos(x, k - m + w) = \arg\max_x \frac{x \cdot (k - m + w)}{||x||\,||k - m + w||}$$

1 konstante

$$= \arg\max_x (xk - xm + xw)$$

Ekvivalents ar:

$$\arg\max_x (\cos(x, k) - \cos(x, m) + \cos(x, w))$$

$xk - xm + xw$ => aditīva kombinācija (viens terms var dominēt)

$\frac{(xk)\cdot(xw)}{(xm)}$ => multiplikatīva kombinācija (daudz balansētāk)

$$\arg\max_x \frac{\cos(x, k)\cos(x, w)}{\cos(x, m) + \epsilon}$$

[Levy, Goldberg, 2014, Linguistic Regularities in sparse and explicit word representations]

Vārdu vektoru kompozīcija

Expression	*Nearest tokens*
Czech + currency	koruna, Czech crown, Polish zloty, CTK
Vietnam + capital	Hanoi, Ho Chi Minh City, Viet Nam, Vietnamese
German + airlines	airline Lufthansa, carrier Lufthansa, flag carrier Lufthansa
Russian + river	Moscow, Volga River, upriver, Russia
French + actress	Juliette Binoche, Vanessa Paradis, Charlotte Gainsbourg

Kā tas strādā?

Reprezentē teikumu/paragrāfu/dokumentu kā iekļauto vārdu vektoru vidējo svērto vektoru

=> visu vārdu pāru vidējā svērtā līdzība (efektīvā veidā)

$$docA = A + B + C$$
$$docB = X + Y + Z$$

$$\cos(\text{docA}, \text{docB}) = = \frac{docA \cdot docB}{||docA|| \cdot ||docB||} == \frac{(A + B + C) \cdot (X + Y + Z)}{||A + B + C|| \cdot ||X + Y + Z||}$$

$$(A + B + C) \cdot (X + Y + Z) = AX + AY + AZ + BX + BY + BZ + CX + CY + CZ$$

Latviešu valodas vārdu vektori un to pielietojumi

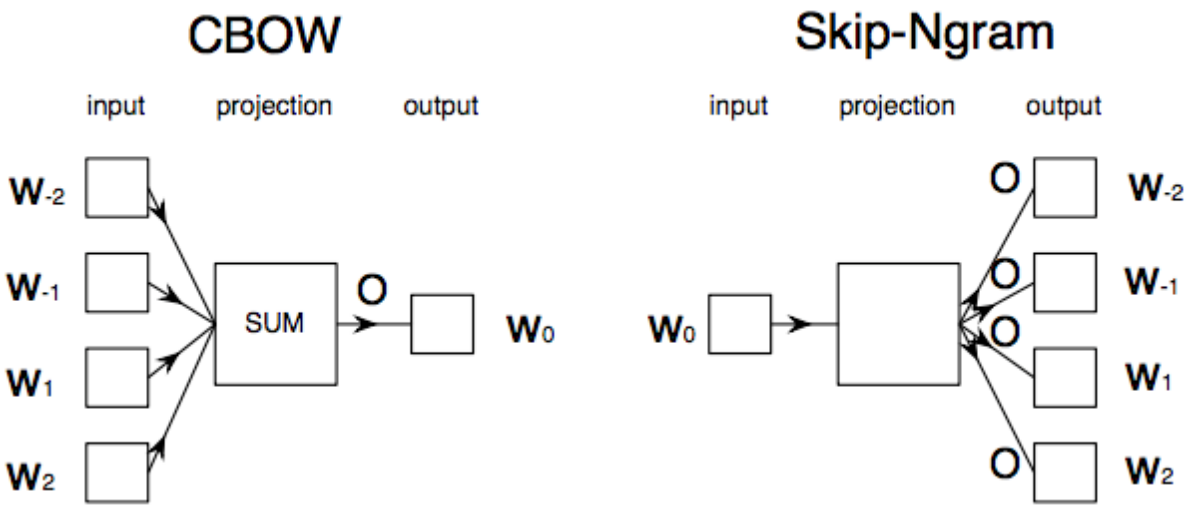
Latviešu valodas jēdzientelpa

- Latviešu valodas ziņu arhīvs
- Apstrādes soļi:
  - Noņemt HTML tagus, tabulas
  - Noņemt speciālos simbolus, punktuāciju
  - Aizvietot ciparus ar “0”
  - Sadalīšana teikumos, vārdos
  - Lemmatizēšana

Teikumi	66.8 M
Vārdi	1.7 B
Vārdnīcas izmērs:	
- >10 lietojumi	0.97 M
- lemmas	0.55 M

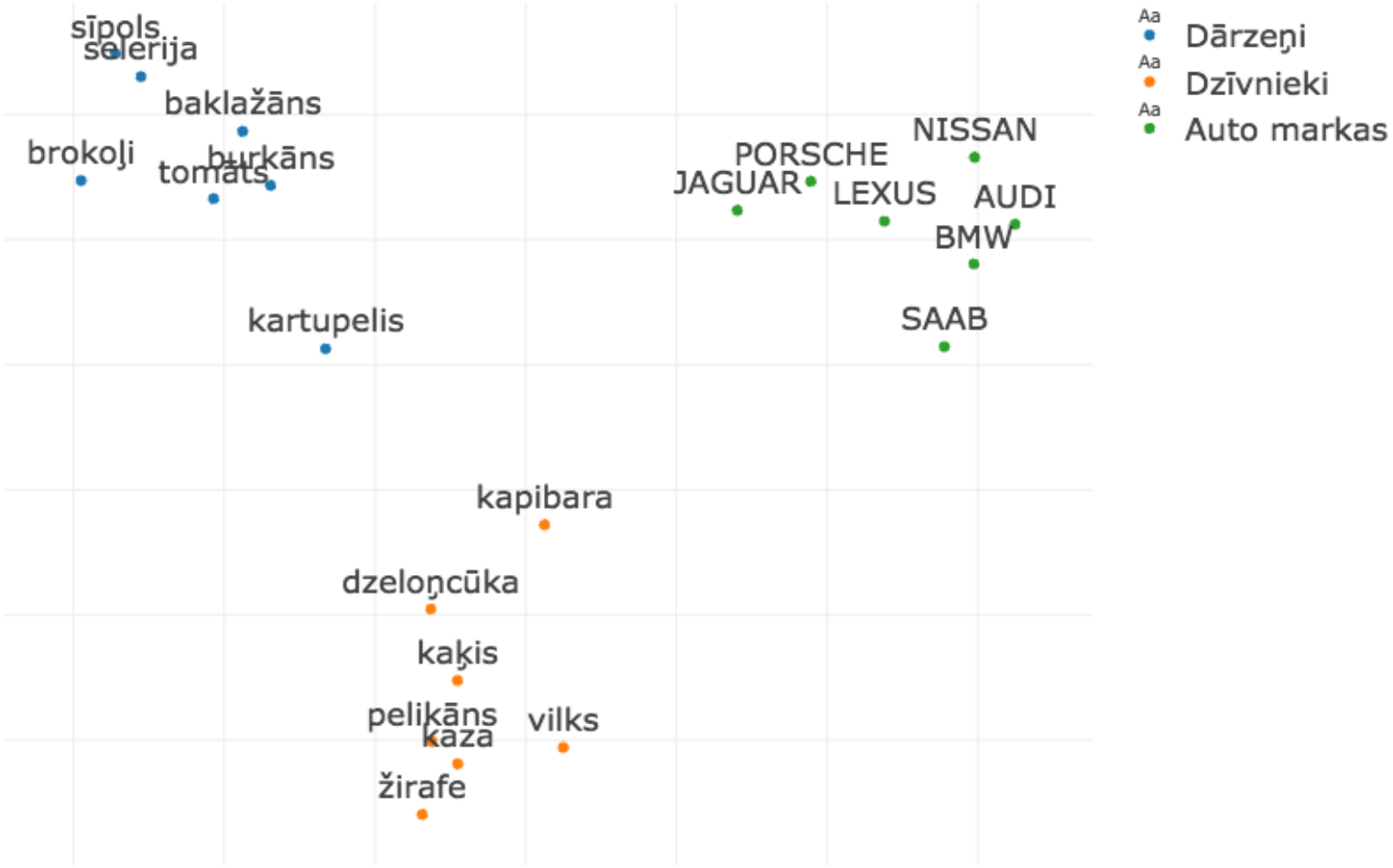
Dažādi modeļi

- Modeļi:
  - Skip-grams (SG), Continuous bag-of-words (CBOW)
  - Structured skip-grams (SSG), Continuous window (CWIN)
  - Character level word embeddings (CWE)
- Parametri: dimensionalitāte, konteksta izmērs
- Korpuss: apjoms, lemmatizēšana



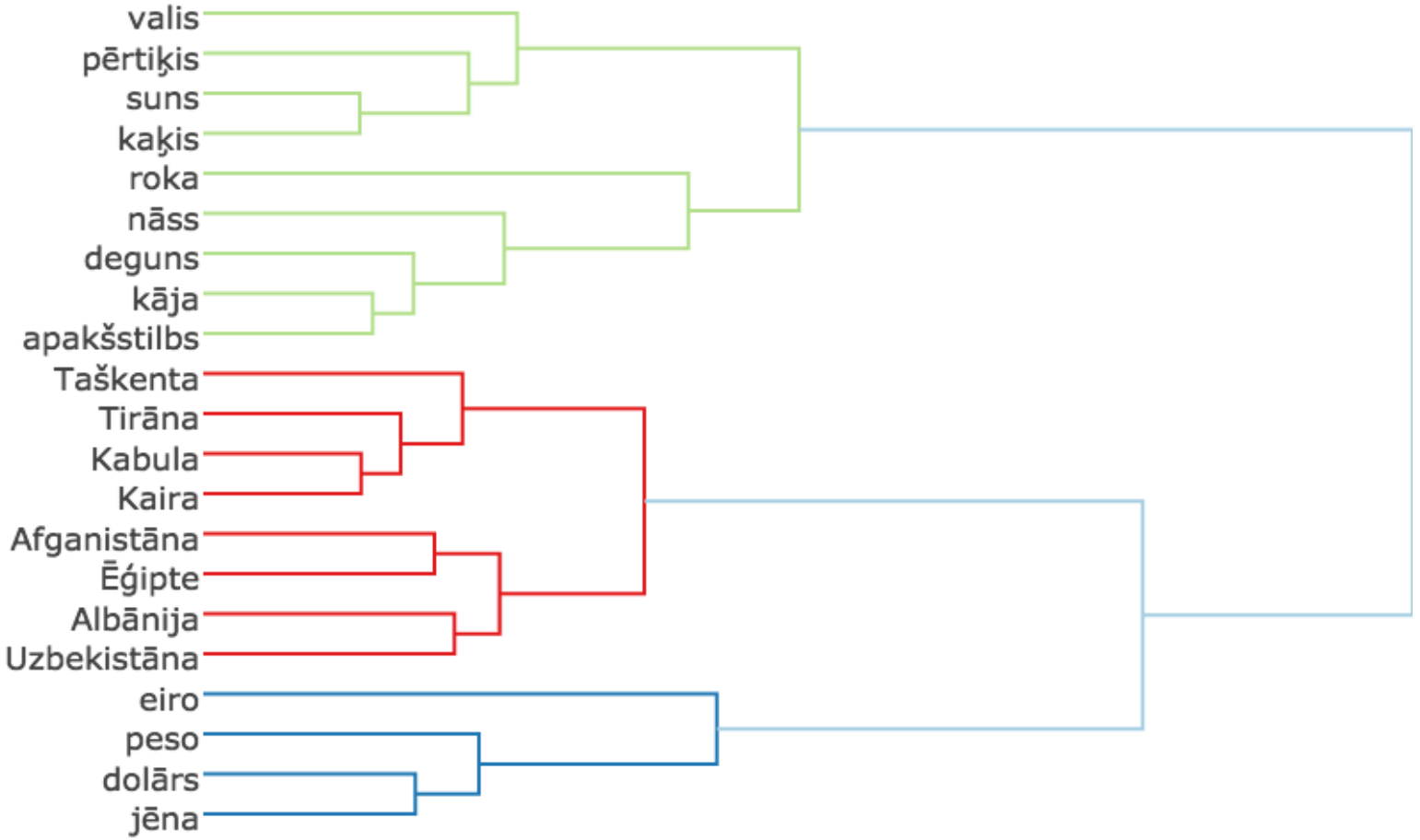
Piemērs

Līdzīgu vārdu klasterēšanās, t-SNE projekcija



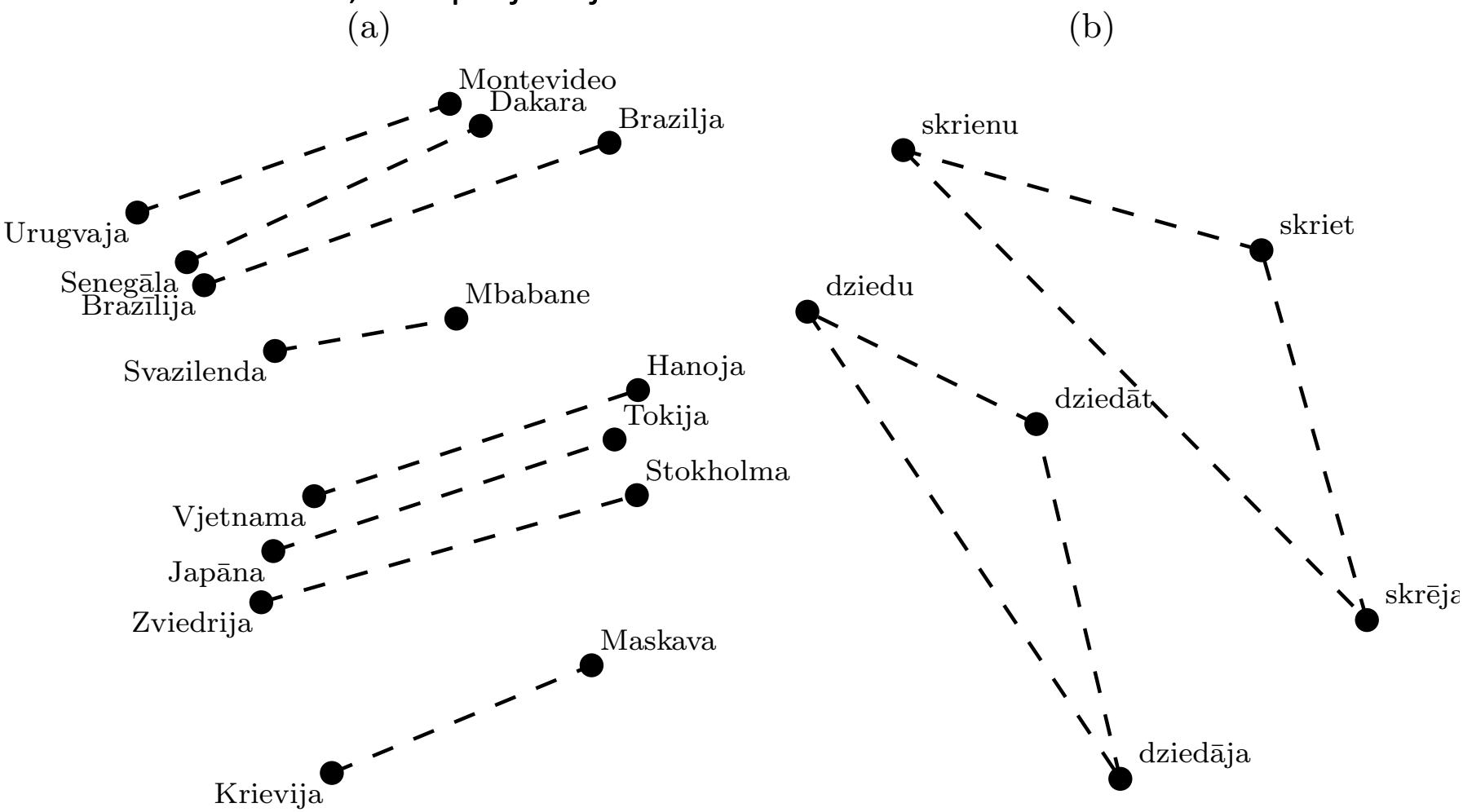
Piemērs

Vārdu vektoru kosinusu līdzības vairākos līmeņos, dendrogramma:



Piemērs

Vārdu attiecības, PCA projekcija



Īpašības

- Latviešu valoda – morfoloģiski bagāta valoda
- Blakus atrodošies vārdi:
  - Semantiska līdzība
  - Sintaktiska līdzība
  - Locījumi
  - Bieži kopā lietotas vārdu grupas
  - Pareizrakstības kļūda
- Kā to varam novērtēt?

$$argmax_{v \in V}(\cos(v, \text{'treniņš'}))$$

Vārdformu līmenis
kardiotreniņš (0.76)
aerobais (0.76)
treniņprocess (0.75)
treniņs (0.75)
iesildīšanās (0.74)
trenniņš (0.70)

Lemmu līmenis
iesildīšanās (0.79)
treniņprocess (0.79)
trenēties (0.77)
trenēšanās (0.77)
starts (0.77)
sparings (0.76)

Simbolu līmenis
treniņa (0.99)
treniņu (0.99)
treniņi (0.99)
treniņiem (0.99)
koptreniņš (0.99)
treniņam (0.99)

Novērtēšana un pielietojumi

- Morfoloģiskā tagošana (vārdšķira, vienkāršots morfoloģiskais tags)
- Nosaukto entītiju atpazīšana (PER, ORG, LOC)
- Sintaktiskā parsēšana (*UD*, *treebank*)

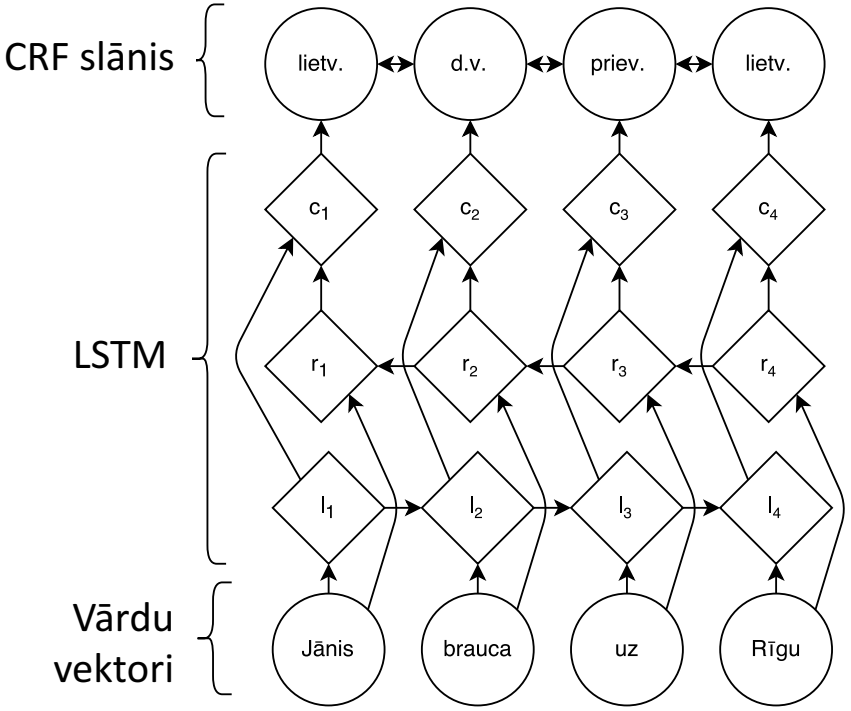
Izmantoto datu kopu statistika: vārdu skaits (teikumi)

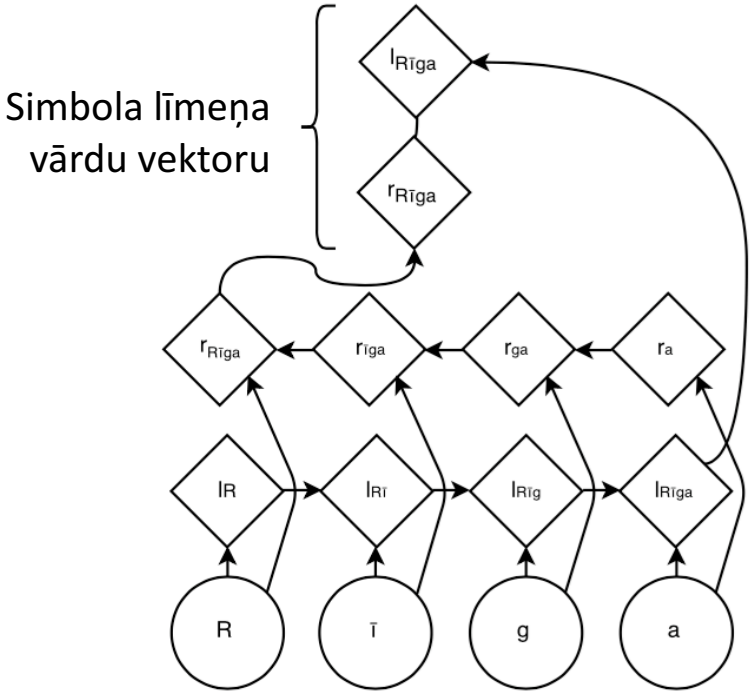
Datu kopa	train	dev	test
Morfoloģiski anotēts korpuss	88,600 (5,560)	11,500 (750)	8,000 (620)
Sintaktiski anotēts korpuss (*treebank*)	49,000 (3,900)	-	4,000 (220)
Universālās atkarības (*UD*)	12,600 (670)	3,500 (190)	4,000 (220)
Nosaukto entītiju anotēts korpuss	44,000 (2,400)	-	-

Izmantotā NN arhitektūra

Bi-directional LSTM-CRF, kas kombinē:

- iepriekš apmācītus vārdu vektorus;
- simbolu līmeņa vārdu reprezentācijas;
- *joint cross-label modeling*





[G. Lample et al., 2015]

Morfoloģiskā marķēšana (POS)

- Bāzlīnija: CMM morfoloģiskais marķētājs:
  - Leksikons
  - Morfoloģiskais analizators
- LSTM-CRF

Metode	POS (ACC, %)	Tags* (ACC, %)	Pilns tags (ACC, %)
CMM	97.7	-	93.6
LSTM-CRF	96.7	91.2	-
LSTM-CRF + WE	**98.3**	**94.6**	-

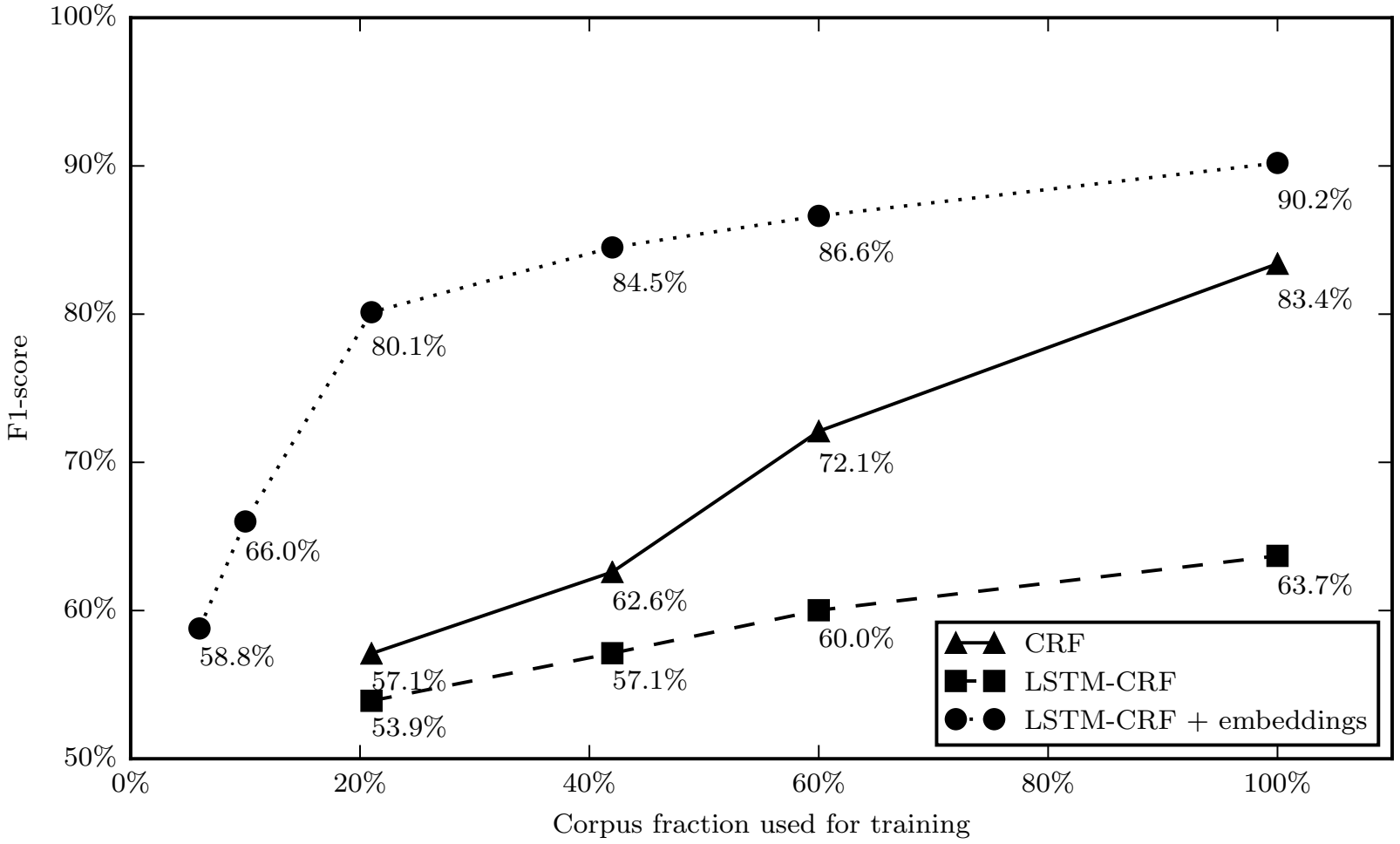
* Vienkāršots morfoloģiskais tags: vārdšķira, locījums, skaitlis, dzimums, persona, darbības vārda izteiksme (~200 unikāli tagi)

Nosaukumu atpazīšana (NER)

- Bāzlīnija: CRF nosaukumu atpazinējs
  - Vārdu saraksti (*gazetteers*)
  - Vārdu klāsteri
  - Morfoloģija
- LSTM-CRF

Metode	NER (F1, %)
CRF	83.4
LSTM-CRF	63.7
LSTM-CRF + WE	**90.5**

NER apmācības līkne



Sintaktiskā parsēšana

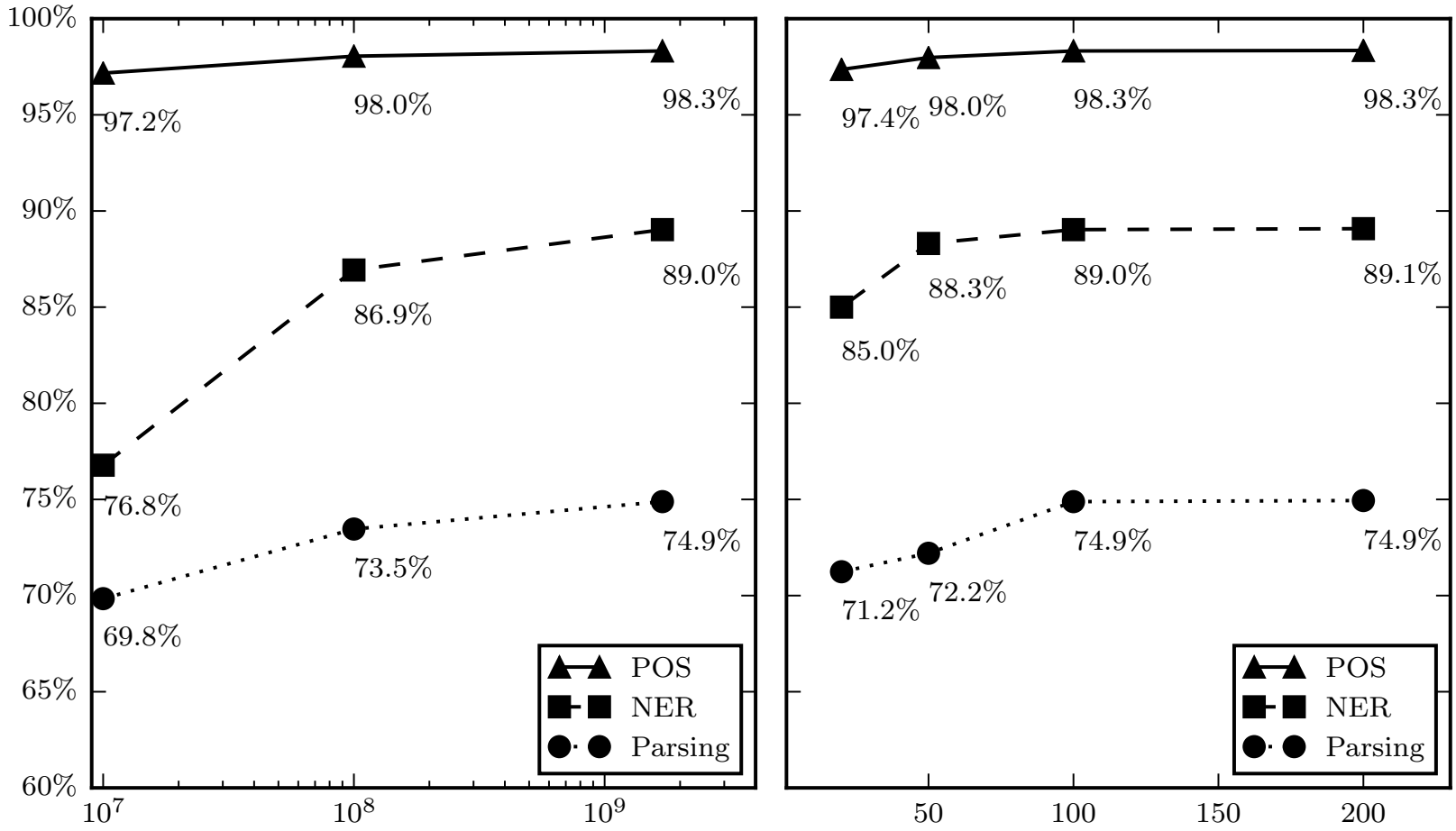
- Bāzlīnija: *Maltparser* rīkkopa, hibrīda atkarībās balstīta gramatika
- LSTM balstīts atkarību modelis
  [M. Ballesteros et al., 2015, Improved Transition-Based Parsing by Modeling Characters instead of Words with LSTMs]

Metode	Treebank (UAS, %)	UD Treebank (UAS, %)
Maltparser	**76.8**	-
LSTM + WE	**76.8**	75.1
LSTM + WE + morpho-tag	-	**75.4**

Rezultāti

Modelis	POS (ACC, %)	NER (F1, %)	UD (UAS, %)
ssg_200	**98.3**	89.1	**74.9**
ssg_100	**98.3**	89.0	**74.9**
sg_100	97.5	86.0	72.6
cbow_100	97.5	85.9	72.6
cwin_100	**98.3**	88.9	74.8
cwe_100	96.9	74.6	68.2
sg_100_lem	97.3	84.5	71.2
avg(ssg_100, sg_100_lem)	**98.2**	86.5	72.8
concat(ssg_100, sg_100_lem)	**98.3**	**90.2**	**75.1**

Svarīgākie parametri



(a) Corpus size (word count)

(b) Dimension size

Secinājumi

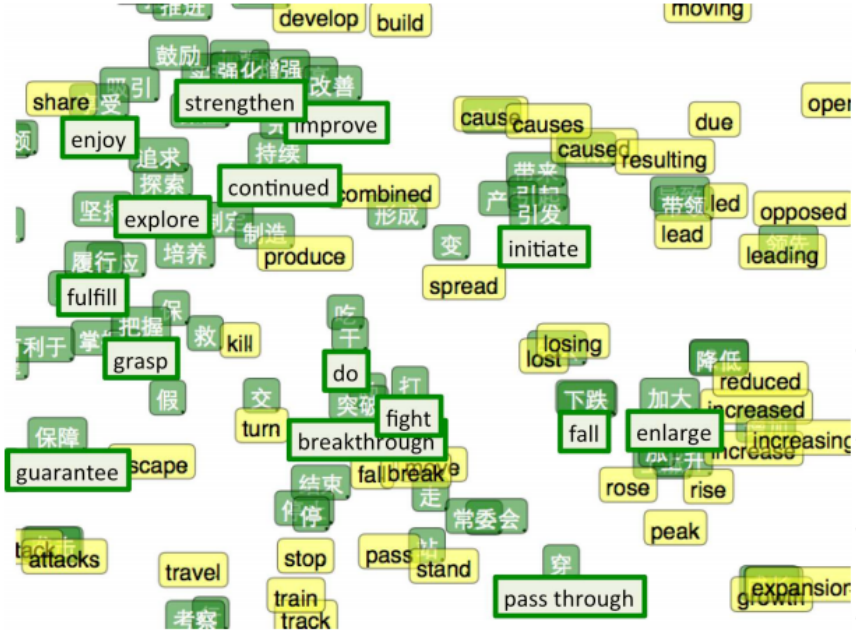
- *State-of-the-art* rezultāti trijos latviešu valodas NLP uzdevumos
  - Neizmanto ārējus resursus
  - Neizmanto manuāli veidotas/atlasītas pazīmes
- Lemmatizēšana var palīdzēt uzlabot rezultātus

Tālāk

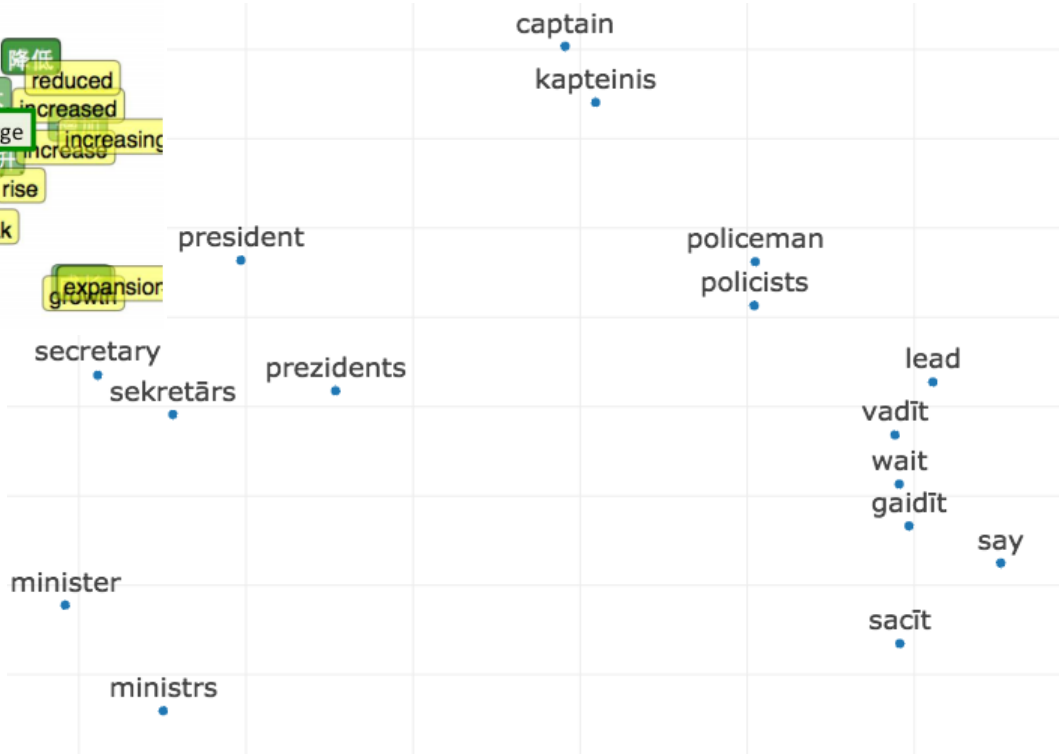
- Dokumentu vektori no vārdu vektoriem
- Vārdu reprezentāciju uzlabošana flektīvām valodām
- Ārpus vārdnīcas vārdi
- Dažādu neironu tīklu arhitektūru salīdzināšana
- Specifisku nosaukumu un terminu atpazīšana runā

- Vairākvalodu vektortelpas
- Vārda nozīmju nodalīšana, inducēšana
- Multimodālas vektortelpas

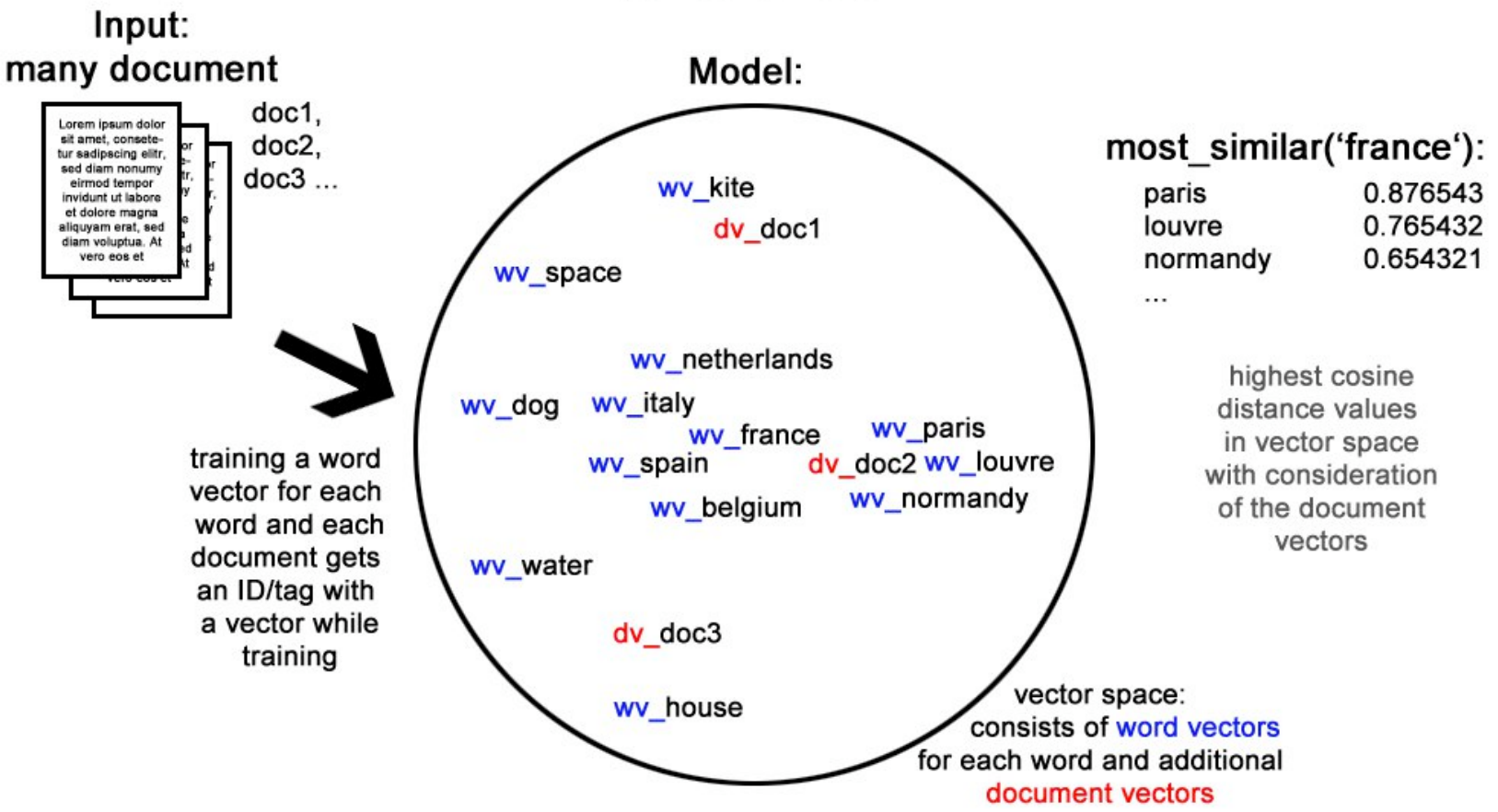
Bilinguāli vārdu vektori



[Socher et al., 2013]



doc2vec



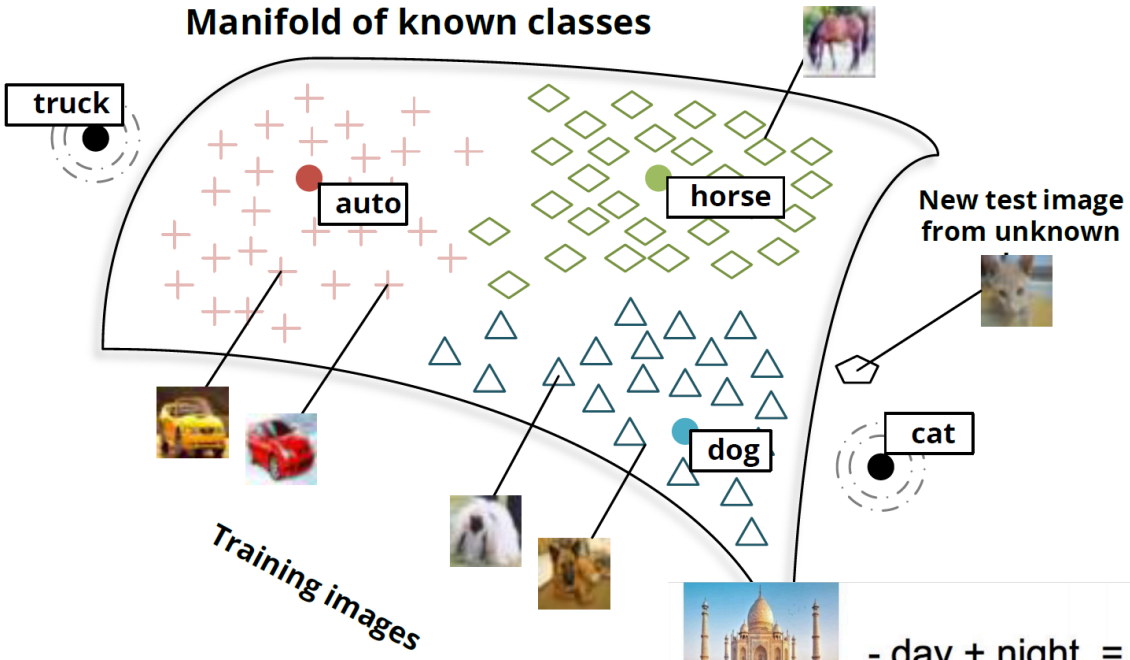
http://gensim.narkive.com/RavqZorK/gensim-4914-graphic-representations-of-word2vec-and-doc2vec:i.5.2.full

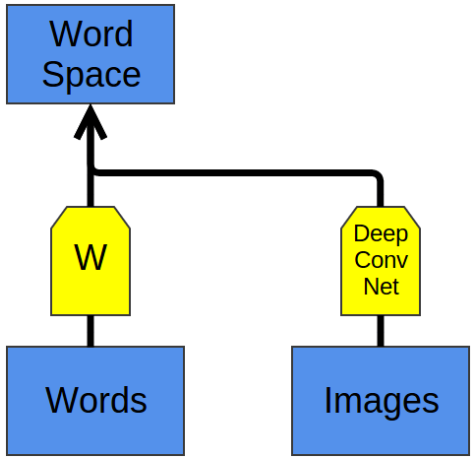
Skip-thought vektori

Query and nearest sentence
he ran his hand inside his coat , double-checking that the unopened letter was still there . he slipped his hand between his coat and his shirt , where the folded copies lay in a brown envelope .
im sure youll have a glamorous evening , she said , giving an exaggerated wink . im really glad you came to the party tonight , he said , turning to her .
although she could tell he had n't been too invested in any of their other chitchat , he seemed genuinely curious about this . although he had n't been following her career with a microscope , he 'd definitely taken notice of her appearances .
an annoying buzz started to ring in my ears , becoming louder and louder as my vision began to swim . a weighty pressure landed on my lungs and my vision blurred at the edges , threatening my consciousness altogether .
if he had a weapon , he could maybe take out their last imp , and then beat up errol and vanessa . if he could ram them from behind , send them sailing over the far side of the levee , he had a chance of stopping them .
then , with a stroke of luck , they saw the pair head together towards the portaloos . then , from out back of the house , they heard a horse scream probably in answer to a pair of sharp spurs digging deep into its
" i 'll take care of it , " goodman said , taking the phonebook . " i 'll do that , " julia said , coming in .
he finished rolling up scrolls and , placing them to one side , began the more urgent task of finding ale and tankards . he righted the table , set the candle on a piece of broken plate , and reached for his flint , steel , and tinder .

http://gabgoh.github.io/ThoughtVectors/

Multimodālas vektortelpas





[Socher *et al.*, 2013]

Publikācijas

- Miranda, S., Znotins, A., Cohen, S.B., Barzdins, G. (2017). Online Monolingual and Crosslingual Clustering of News Articles. In EMNLP.
- Znotiņš, A. (2016). Word Embeddings for Latvian Natural Language Processing Tools. In *Human Language Technologies: The Baltic Perspective*, IOS Press. Web of Science.
- Znotiņš, A., Polis, K., and Darģis R. (2015). Media monitoring system for Latvian radio and TV broadcasts. In *Proceedings of the 16th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 732–733. Web of Science, Scopus.
- Darģis, R. and Znotiņš, A. (2014). Baseline for Keyword Spotting in Latvian Broadcast Speech. In *Human Language Technologies: The Baltic Perspective*, IOS Press, 75–82. Web of Science.
- Znotiņš, A. and Paikens, P. (2014). Coreference Resolution for Latvian. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*, 3209–3213. Web of Science.
- Pretkalniņa, L., Znotiņš, A., Rituma, L., Goško, D. (2014). Dependency parsing representation effects on the accuracy of semantic applications — an example of an inflective language. In Proc*eedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*, 4074–4081. Web of Science.

Paldies par uzmanību!

Q & A

arturs.znotins@gmail.com