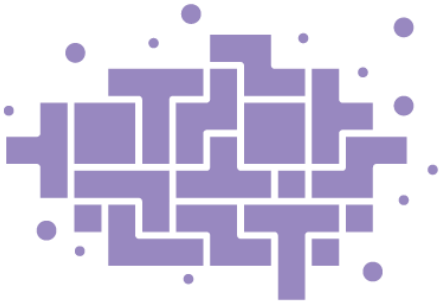
Latviešu valodas sapratne un tekstrade cilvēka-datora komunikācijas modelēšanā

Inguna Skadiņa

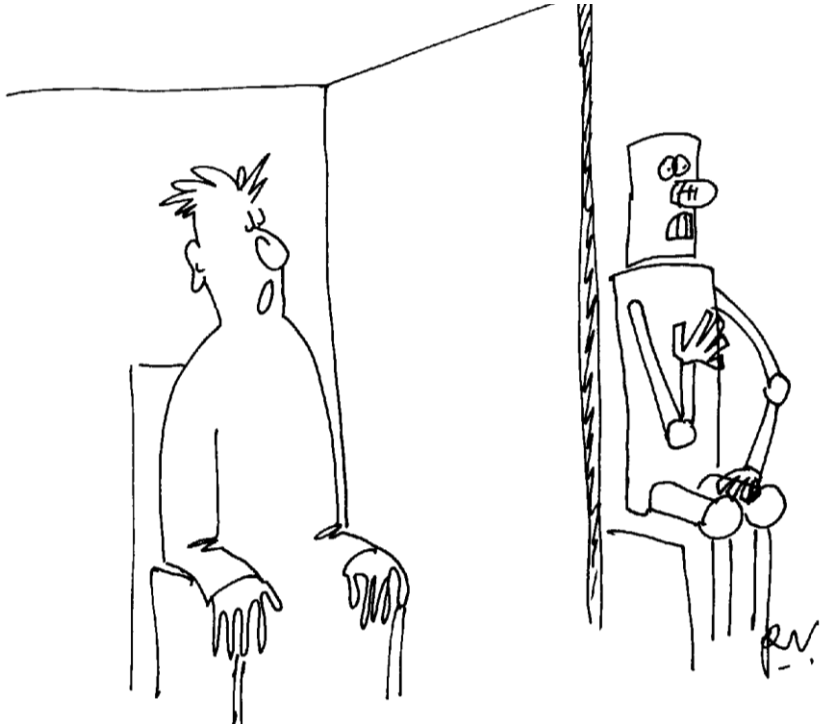


FLPP

FUNDAMENTĀLIE UN LIETIŠĶIE PĒTĪJUMU PROJEKTI

Mākslīgā intelekta laboratorija
LU MII

Prāts pret mašīnu



Tjūringa tests (1950)

Prāts pret mašīnu

B. Kristians. Prāts pret mašīnu, Rīgas Laiks
https://www.rigaslaiks.lv/zurnals/prats-pret-masinu-1162

Prāts pret mašīnu

Braitona, Anglija, 2009. gada septembris. Es pamostos viesnīcas numurā astoņtūkstoš kilometrus no savām mājām Sietlā. Pēc brokastīm izeju sāļajā āra gaisā un dodos pastaigāt gar jūras krastu. Es esmu zemē, kura izgudrojusi manu valodu, taču izrādās, ka man nav saprotama liela daļa ceļā ieraudzīto uzrakstu. *LET AGREED,* acīs krītošiem, lieliem burtiem vēsta viens no tiem. Man tas neko neizsaka.

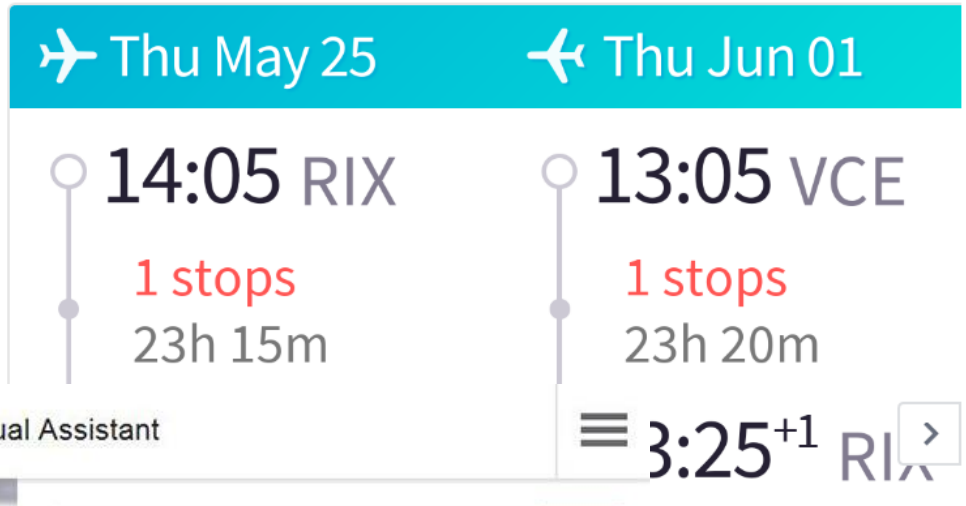
Es apstājos un brīdi truli veros jūrā, domās vēl un vēlreiz analizējot mīklaino paziņojumu. Parasti mani intrigē šādas lingvistiskas dīvainības un kultūru atšķirības; šodien tās drīzāk gan vieš manī bažas. Nepilnu divu stundu laikā man vajadzēs apsēsties pie datora un iesaistīties virknē piecminūšu virtuālo sarunu ar vairākiem man pilnīgi nepazīstamiem cilvēkiem. Mani neredzamie sarunu biedri būs psihologs, lingvists, datorzinātnieks un populāra tehnoloģijas jautājumiem veltīta britu televīzijas raidījuma vadītājs. Visi kopā šie cilvēki veidos tiesnešu grupu, kas vērtēs manu spēju tikt galā ar vienu no dīvainākajiem uzdevumiem, kādu savā mūžā esmu saņēmis.

Man vajadzēs viņus pārliecināt, ka es esmu cilvēks.

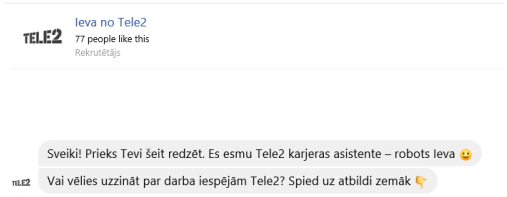
The Atlantic, 2011. gada marts

Virtuālie sarunu biedri ir realitāte

- Virtuāli sarunu biedri un «gudrās» iekārtas - Apple Siri, Microsoft Cortana, Google Assistant, Amazon Echo, Google Home,
- Virtuālie palīgi:
  - aviolīnijām
  - bankām
  - apdrošināšanai
  - utt.



Daži piemēri Latvijā



Rīki un platformas

Microsoft Bot Framework

ManyChat

Virtuālie sarunu biedri

Chit-Chat

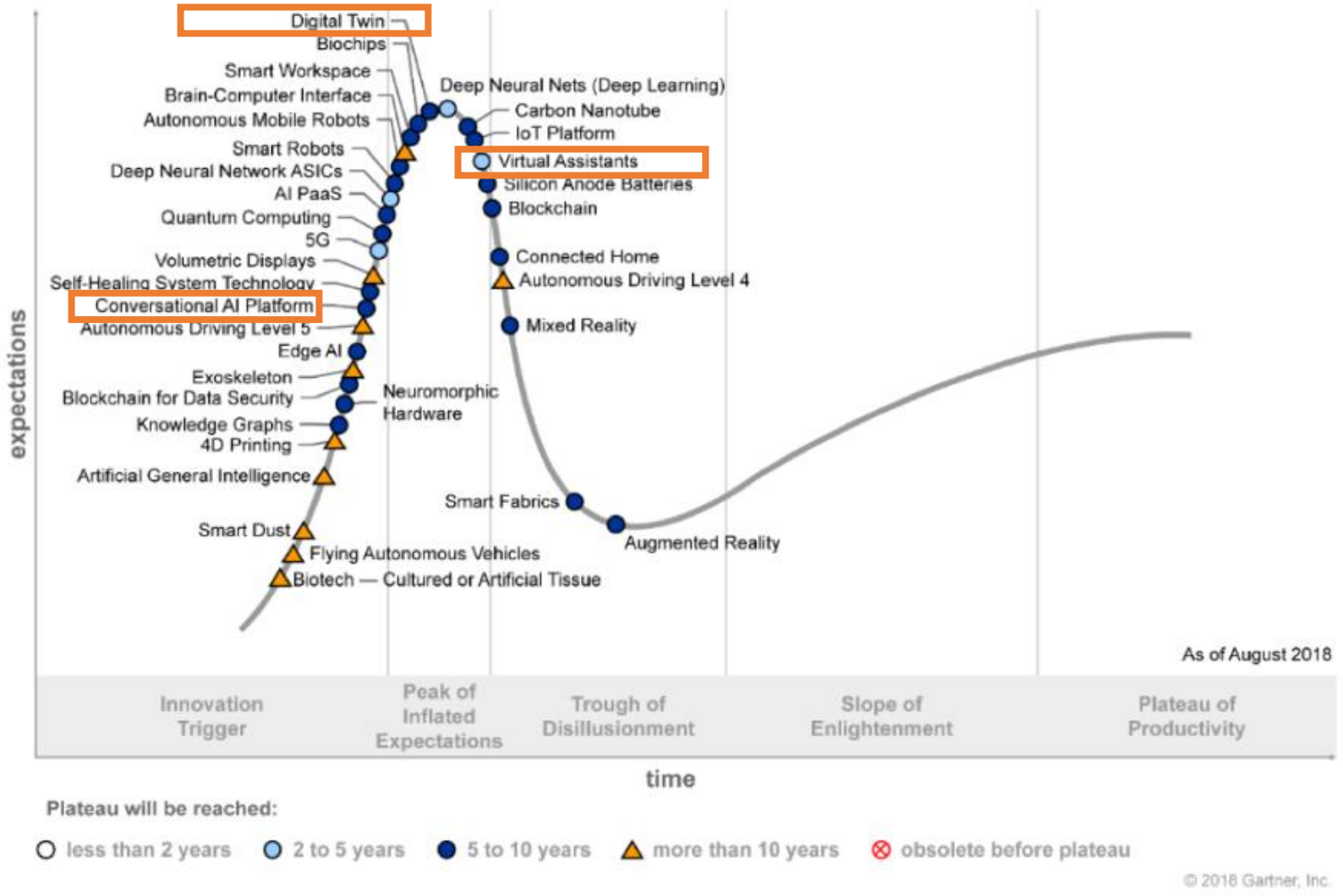
seq2seq models → Seq2seq with conversation contexts → Knowledge-grounded seq2seq models

Task-Oriented

Single-domain, system-initiative → Multi-domain, contextual, mixed-initiative → End-to-end learning, massively multi-domain

https://www.csie.ntu.edu.tw/~yvchen/doc/COLING18_Tutorial.pdf

Gartner Hype Cycle for Emerging Technologies, 2018



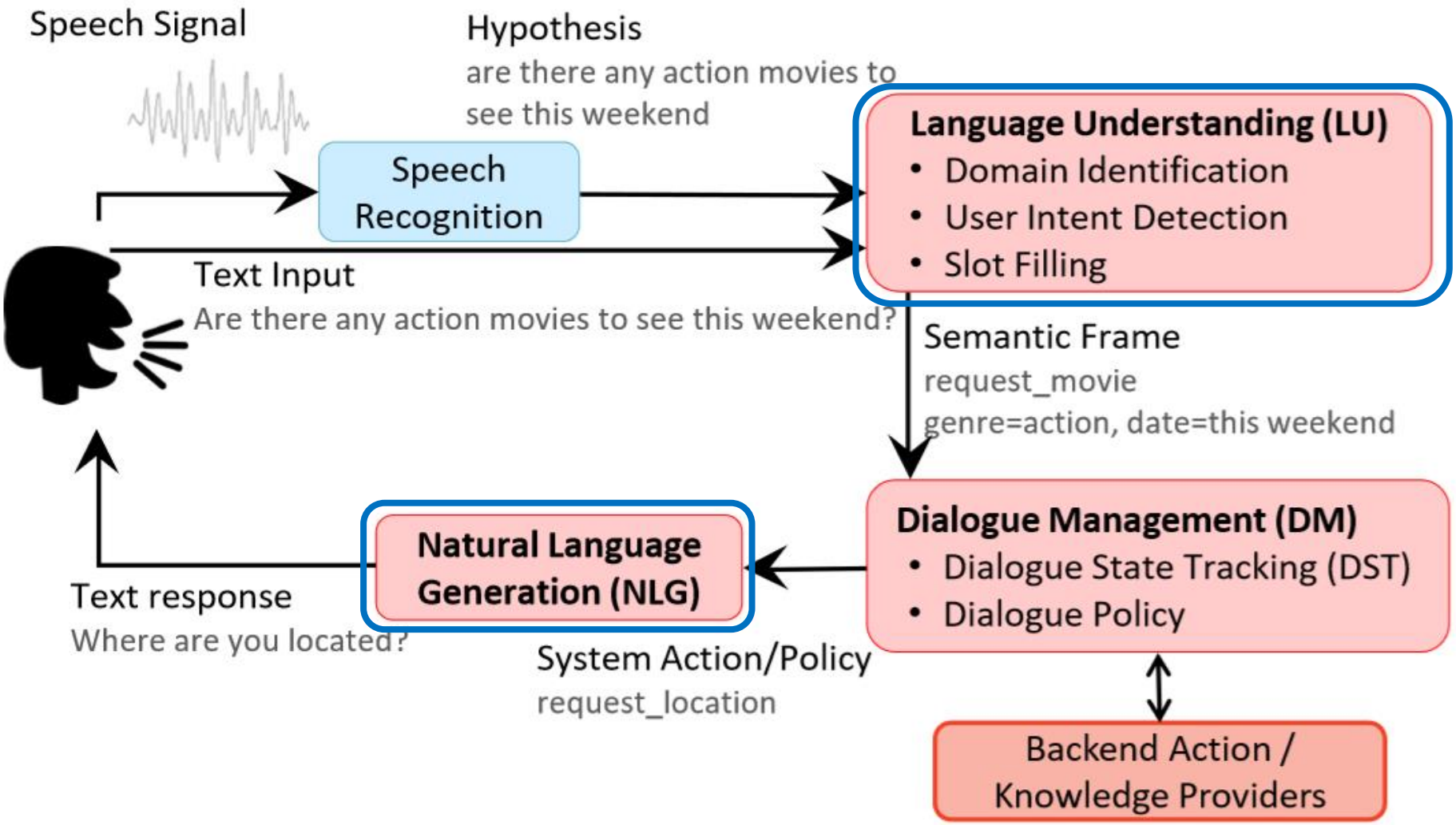
https://www.gartner.com/en/newsroom/press-releases/2018-08-20-gartner-identifies-five-emerging-technology-trends-that-will-blur-the-lines-between-human-and-machine

Bet

- Vai pašreizējie risinājumi sarunājas «mazajās» valodās?
- Vai dators saprot valodu?
- Kāds ir labākais risinājums, lai «saprastu» izteikumu un jēgpilni atbildētu?
  - Mašīnmācīšanās?
  - Zinību bāzes?
  - Hibrīdi risinājumi?

FLPP
FUNDAMENTĀLIE UN LIETIŠĶIE PĒTĪJUMU PROJEKTI

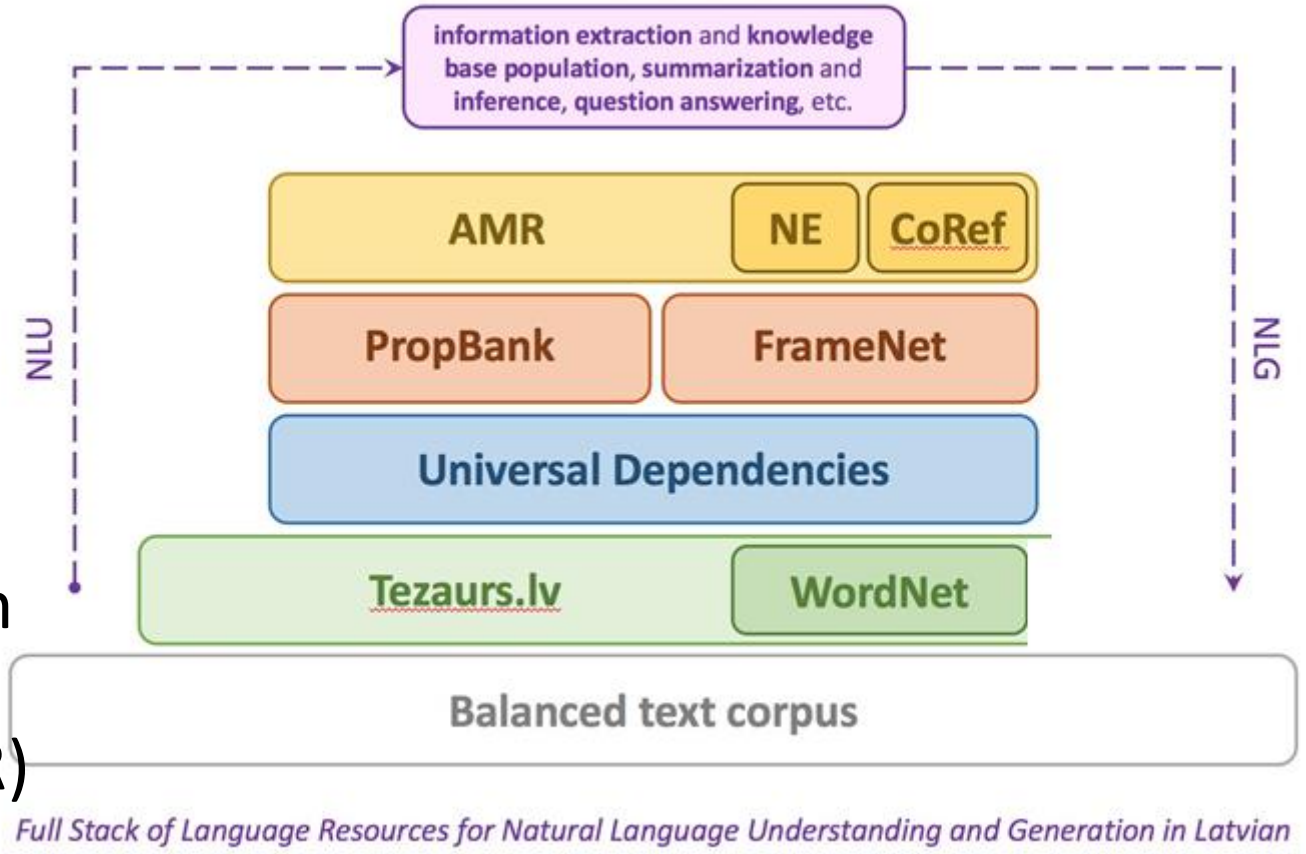
Dialogsistēmu arhitektūra



https://www.csie.ntu.edu.tw/~yvchen/doc/COLING18_Tutorial.pdf, Joung 2000

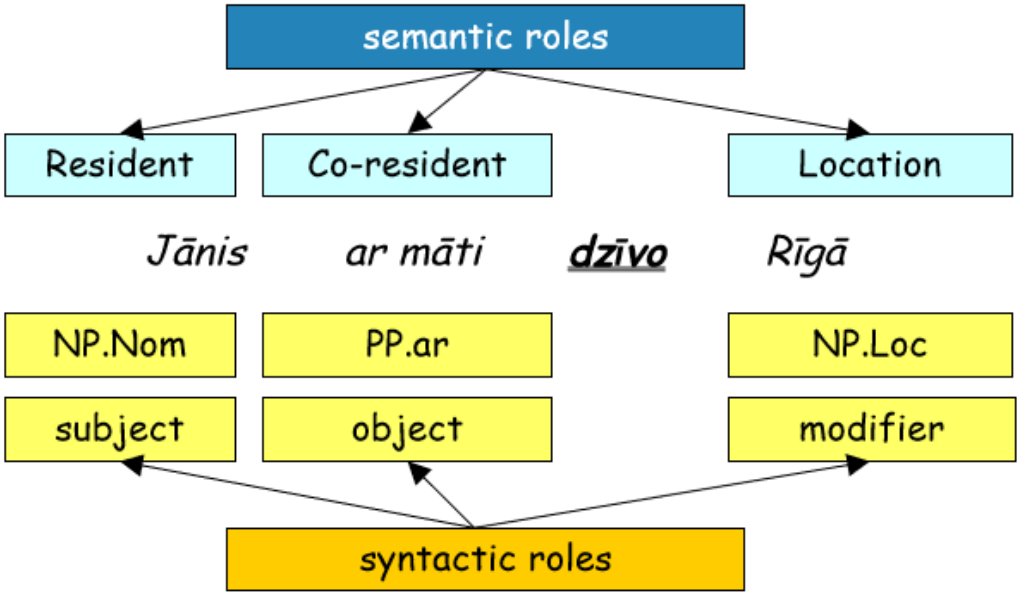
Mūsu piedāvājums: valodas sapratne

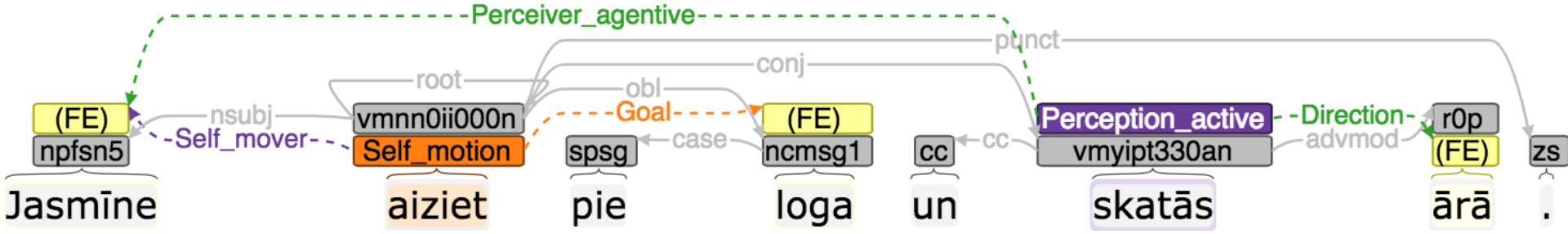
Radīt metodes un rīkus semantiskai parsēšanai, kas «mācās» no manuāli un automātiski marķētiem datiem (UD, Framenet, PropBank un AMR)



Mūsu piedāvājums: valodas sapratne

- Semantiskā parsēšana
- Ontoloģiju izveide jomai
- Nodomu noteikšana
- Slotu aizpildīšana





Iestrādes: NLP-pipe

Šajā mācību gadā Aizkraukles novada ģimnāzijas 8. klasē mācījās Marisa Butnere no Amerikas.

Go | tokenizer × | morpho × | parser × | ner × | NER | CONLL | JSON

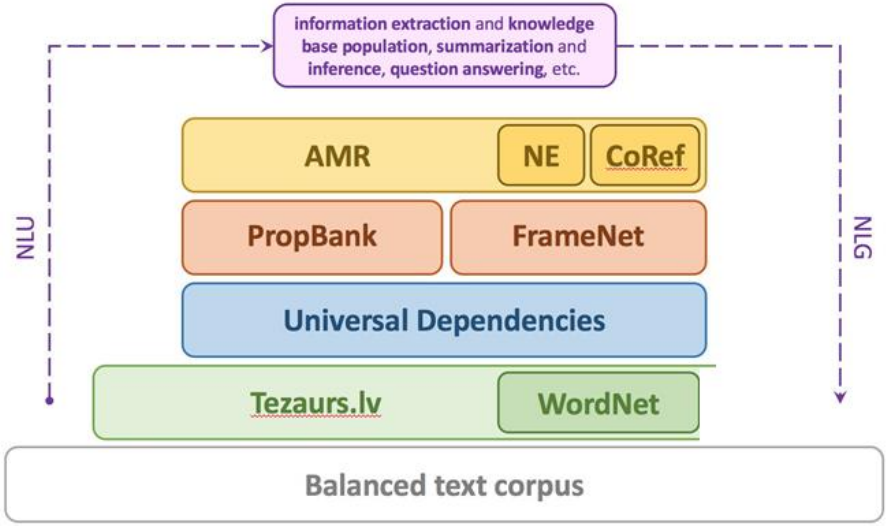
INDEX	FORM	LEMMA	UPOSTAG	XPOSTAG	FEATS	HEAD	DEPREL
#text=Šajā mācību gadā Aizkraukles novada ģimnāzijas 8. klasē mācījās Marisa Butnere no Amerikas .							
1	Šajā	šis	DET	pd3msln	Skaitlis=Vienska	3	det
2	mācību	mācība	NOUN	ncfpg4	Skaitlis=Daudzsk	3	nmod
3	gadā	gads	NOUN	ncmsl1	Skaitlis=Vienska	9	obl
4	Aizkraukles	Aizkraukle	PROPN	npfsg5	Skaitlis=Vienska	5	nmod
5	novada	novads	NOUN	ncmsg1	Skaitlis=Vienska	6	nmod
6	ģimnāzijas	ģimnāzija	NOUN	ncfsg4	Skaitlis=Vienska	8	nmod
7	8.	8.	ADJ	xo	Reziduāļa_tips=I	8	amod
8	klasē	klase	NOUN	ncfsl5	Skaitlis=Vienska	9	obl
9	mācījās	mācīties	VERB	vmyis_330an	Laiks=Pagātnelk	0	root
10	Marisa	Marisa	PROPN	npfsn_	Skaitlis=Vienska	9	nsubj
11	Butnere	Butnere	PROPN	ncfsn5	Skaitlis=Vienska	10	flat:name
12	no	no	ADP	spsg	Skaitlis=Vienska	13	case
13	Amerikas	Amerika	PROPN	npfsg4	Skaitlis=Vienska	10	nmod
14	.	.	PUNCT	zs	Galotnes_nr=209	9	punct

Šajā mācību gadā Aizkraukles novada ģimnāzijas organization 8. klasē mācījās Marisa Butnere person no Amerikas GPE .

http://nlp.ailab.lv/

Mūsu piedāvājums: tekstrade

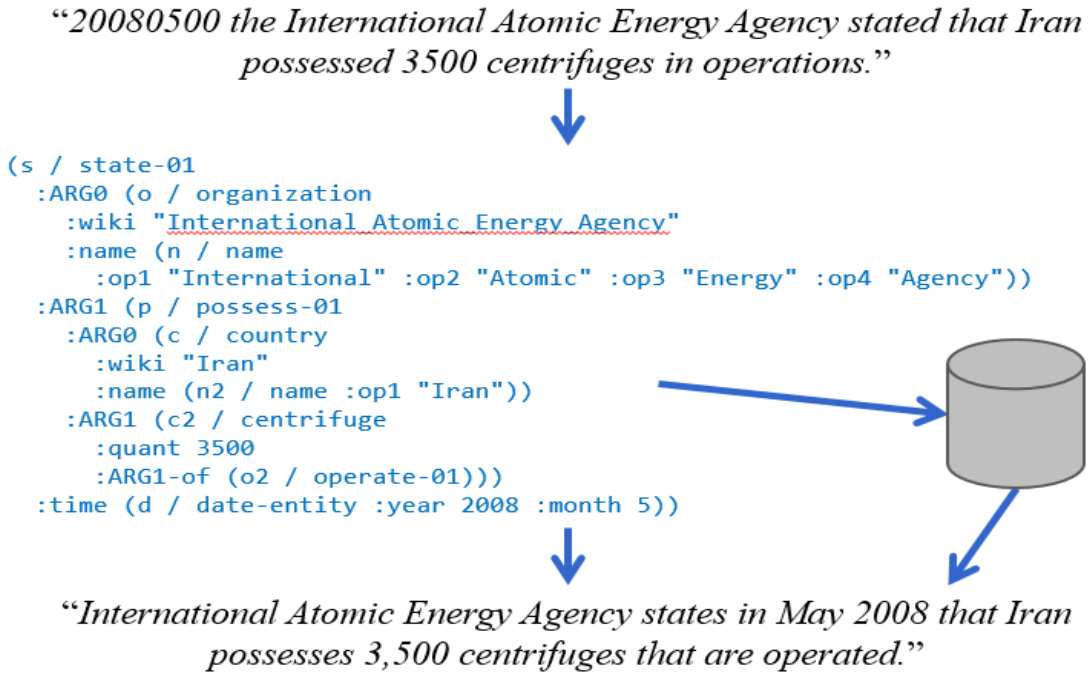
Plūstoša un jēgpilna tekstrade no mašīnlasāmas nozīmes reprezentācijas, apvienojot datos balstītu un gramatikā balstītu pieeju



Full Stack of Language Resources for Natural Language Understanding and Generation in Latvian

Mūsu piedāvājums: tekstrade

- Abstraktās sintakses izveide GF
- Konkrētās sintakses izveide jomas lietojumam latviešu un angļu valodai



The Second Conversation Intelligence Challenge (ConvAI2)

http://convai.io/

Human Evaluation Leaderboard

Rank	Creator	Rating	Persona detect
1	Lost in Conversation [code]	3.11	0.9
2	(Hugging Face)	2.68	0.98
3	Little Baby(AI小奶娃)	2.44	0.79
4	Mohd Shadab Alam	2.33	0.93
5	Happy Minions	1.92	0.46
6	ADAPT Centre	1.6	0.93
KV Profile Memory	ParlAI team	2.44	0.76
Human	MTurk	3.48	0.96

Automatic Evaluation Leaderboard (hidden test set)

Rank	Creator	PPL	Hits@1	F1
1	(Hugging Face)	16.28	80.7	19.5
2	ADAPT Centre	31.4	-	18.39
3	Happy Minions	29.01	-	16.01
4	High Five	-	65.9	-
5	Mohd Shadab Alam	29.94	13.8	16.91

DSTC7

Dialog System Technology Challenges

Honolulu, Hawaii, USA, January 27, 2019

DSTC8

SUBMITTED TRACK PROPOSALS

- End-to-end Task Completion Dialog Challenge - Microsoft Research & Tsinghua University
- Multi-Domain Human-to-Human Dialog Understanding, a machine reading approach - Naver Labs Europe & Adobe Research
- Dialog State Tracking for Conversational Image Editing - Adobe Research
- NOESIS II: Predicting Responses, Identifying Success, and Managing Complexity in Task-Oriented DialogueIBM & University of Michigan
- Audio Visual Scene-Aware Dialog Track - MERL & CMU
- Scalable Schema-Guided Dialogue State Tracking - Google

SQuAD2.0

Stanford Question Answering Dataset (SQuAD) is a reading comprehension dataset, consisting of questions posed by crowdworkers on a set of Wikipedia articles, where the answer to every question is a segment of text, or *span*, from the corresponding reading passage, or the question might be unanswerable.

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Jan 15, 2019	BERT + MMFT + ADA (ensemble) *Microsoft Research Asia*	**85.082**	**87.615**
2 Jan 10, 2019	BERT + Synthetic Self-Training (ensemble) *Google AI Language* https://github.com/google-research/bert	84.292	86.967
3 Dec 13, 2018	BERT finetune baseline (ensemble) *Anonymous*	83.536	86.096
4 Dec 16, 2018	Lunet + Verifier + BERT (ensemble) *Layer 6 AI NLP Team*	83.469	86.043
4 Dec 21, 2018	PAML+BERT (ensemble model) *PINGAN GammaLab*	83.457	86.122
5 Jan 10, 2019	BERT + Synthetic Self-Training (single model) *Google AI Language* https://github.com/google-research/bert	82.972	85.810

https://rajpurkar.github.io/SQuAD-explorer/

ELMO: Deep contextualized word representations

ELMo is a deep contextualized word representation that models both

(1) complex characteristics of word use (e.g., syntax and semantics)

and

(2) how these uses vary across linguistic contexts (i.e., to model polysemy).

ELMo representations are *Contextual*: The representation for

Task	Previous SOTA		Our baseline	ELMo + Baseline	Increase (Absolute/Relative)
SQuAD	SAN	84.4	81.1	85.8	4.7 / 24.9%

BERT: Bidirectional Encoder Representations from Transformers

SQuAD1.1 Leaderboard

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar et al. '16)	82.304	91.221
1 Oct 05, 2018	BERT (ensemble) *Google AI Language* https://arxiv.org/abs/1810.04805	**87.433**	**93.160**
2 Sep 09, 2018	nlnet (ensemble) *Microsoft Research Asia*	85.356	91.202
3 Jul 11, 2018	QANet (ensemble) *Google Brain & CMU*	84.454	90.490

https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html

“Messenger is growing as a platform and chatbots provide a unique way to connect,” said Krista Neher, bestselling author and CEO of Boot Camp Digital. “What I expect in 2019 is that they hype of chatbots will die down and businesses will be more strategic in how they actually use them. Marketers will shift from ‘we need a chatbot’ to ‘how do we use this strategically as a communications tool?’”

Search Engine Journal

169 tūkst. people like this

Mediju/Ziņu uzņēmums

FLPP

FUNDAMENTĀLIE UN LIETIŠĶIE PĒTĪJUMU PROJEKTI

Paldies par uzmanību!