Datorizēta semantiskā analīze

Strukturētas informācijas izguvei no teksta

Guntis Bārzdiņš, Didzis Goško,
Pēteris Paikens, Normunds Grūzītis

Latvijas Universitāte
ANNO 1919

Latvijas Universitātes
Matemātikas un informātikas institūts

Mākslīgā intelekta laboratorija (ailab.lv)

Teksta analīzes tehnoloģiju izmanto visā pasaulē

DATORZINĀTNE Latvijas Universitātes Matemātikas un informātikas institūta Mākslīgā intelekta laboratorijā radīta pilnīgi jauna mašīnmācīšanās metode un izveidots rīks valodas vienību nozīmes jeb semantiskā attēlojuma izguvei no angļu valodas teksta. Jaunā teksta semantiskās analīzes tehnoloģija ir ātra un precīza, un latviešu valodai pielāgoto versiju jau lieto praksē informācijas aģentūras LETA mediju monitoringa pakalpojumu nodrošināšanai. Patlaban tehnoloģiju ievieš tādos visā pasaulē pazīstamos mediju uzņēmumos kā *BBC* un "Deutsche Welle".

Pagājušajā gadā LU un LETA veidotā komanda triumfēja ASV notiekošajās sacensībās "SemEval-2016", līdz ar to Latvijā veidoto rīku var uzskatīt par pasaulē precīzāko semantiskā attēlojuma izguvei no angļu valodas teksta.

Latvijas komanda sacensībās "SemEval-2016" bija labākā starp 11 dalībniekiem.

Ilustrētā zinātne, Februāris 2017

SASNIEGUMI LIETIŠĶAJĀ ZINĀTNĒ

Izveidota precīza, ātra un praktiski izmantojama datorizētas semantiskās analīzes tehnoloģija. LZA korespondētājloceklis Guntis Bārzdiņš, *Mg.* Didzis Goško, *Mg.* Pēteris Paikens, *Dr.sc.comp.* Normunds Grūzītis Latvijas Universitātes Matemātikas un informātikas institūts.

LZA, Zinātnes Vēstnesis, 522(1), 2017

Uzvara SemEval-2016 (Task 8: AMR)

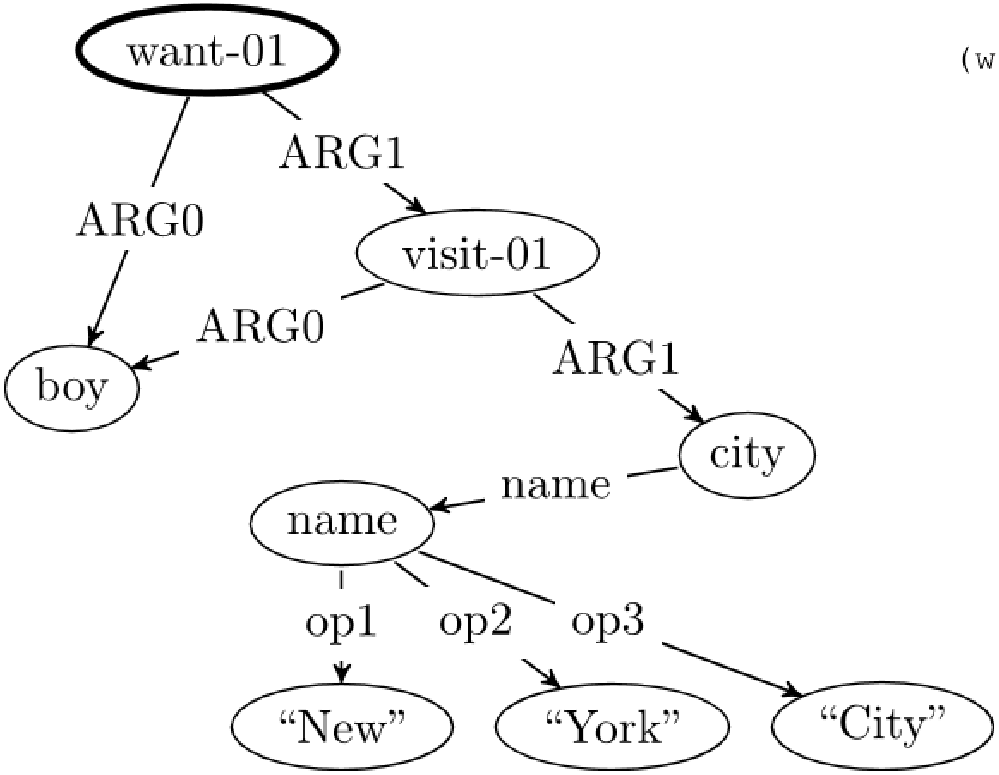
#	Nosaukums	OrganizAcija	Smatch
1	RIGA	University of Latvia, IMCS; LETA	0.6196
2	Brandeis/cx/RPI	Brandeis University; Boulder Learning Inc.; Rensselaer Polytechnic Institute	0.6195
3	ICL-HD	Ruprecht-Karls-Universitat Heidelberg	0.6005
4	UCL+Sheffield	University College London; University of Sheffield	0.5983
5	M2L	Kyoto University	0.5952
6	CMU	Carnegie Mellon University; University of Washington	0.5636
7	CU-NLP	OK Robot Go, Ltd.; University of Colorado	0.5566
8	UofR	University of Rochester	0.4985
9	MeaningFactory	University of Groningen	0.4702
10	CLIP@UMD	University of Maryland	0.4370
11	DynamicPower	National Institute for Japanese Language and Linguistics	0.3706

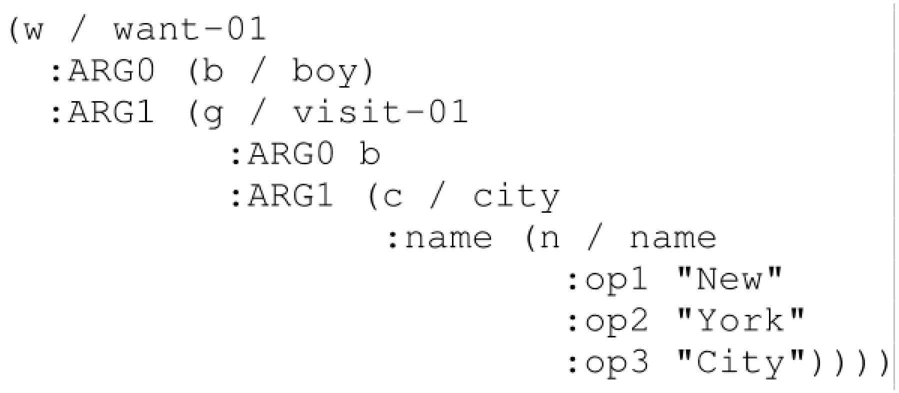
http://alt.qcri.org/semeval2016/task8/

http://www.aclweb.org/anthology/S/S16/S16-1166.pdf
http://www.aclweb.org/anthology/S/S16/S16-1176.pdf

http://amr.isi.edu/index.html

Abstract Meaning Representation (AMR)





```
(w / want-01
   :ARG0 (b / boy)
   :ARG1 (g / visit-01
            :ARG0 b
            :ARG1 (c / city
                     :name (n / name
                              :op1 "New"
                              :op2 "York"
                              :op3 "City"))))
```

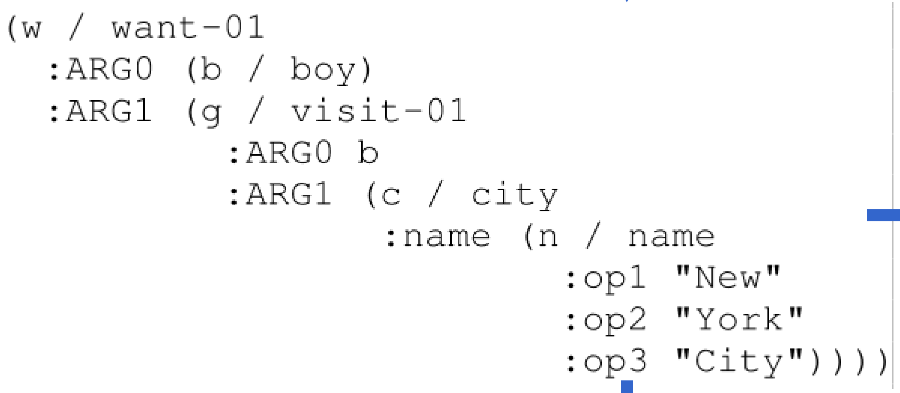
∃ w, b, g:
instance(w, want-01) ∧ instance(g, go-01) ∧
instance(b, boy) ∧ arg0(w, b) ∧
arg1(w, g) ∧ arg0(g, b)
arg1(g,c)...

“The boy wants to visit New York City.”

"NYC is what boys wanted to visit."

Semantiku saglabājoša tulkošana

“Zēns grib apmeklēt Ņujorku"



```
(w / want-01
   :ARG0 (b / boy)
   :ARG1 (g / visit-01
            :ARG0 b
            :ARG1 (c / city
                     :name (n / name
                              :op1 "New"
                              :op2 "York"
                              :op3 "City"))))
```

Strukturētas informācijas izguve no teksta (FreeBase, TAC-KBP, LETA)

“The boy wants to visit New York City.”

AMR tulkošanas / parafrāzes piemēri

Oriģināls	AMR (Grafs / FOL)	Parafrāze
Soldier injured during bomb defusion in Kathmandu after state of emergency expires.	(injure-01:ARG0(defuse-01:ARG1(bomb):location(city:wiki "Kathmandu":name(name:op1 "Kathmandu"))):ARG1(soldier):time(after:op1(expire-01:ARG1(state:mod(emergency)))))	Soldiers were defusing the bombs in Kathmandu were injured after expire on state of emergency
A Kathmandu police officer reports --	(report-01:ARG0(person:ARG0-of(have-org-role-91:ARG1(police:mod(city:wiki "Kathmandu":name(name:op1 "Kathmandu"))):ARG2(officer))))	Kathmandu police officers report
1 soldier of the Royal Nepal Army was seriously injured on 29 August 2002 when a bomb disposal team attempted to defuse the bomb left at an electricity pole in Okubahal near Sundhara in Lalitpur district in Kathmandu.	(injure-01:ARG1(soldier:quant 1:ARG2-of(have-org-role-91:ARG1(military:wiki "Nepalese_Army":name(name:op1 "Royal":op2 "Nepal":op3 "Army")))):time(date-entity:year 2002:month 8:day 29):ARG1-of(cause-01:ARG0(attempt-01:ARG0(team:mod(dispose-01:ARG1(bomb))):ARG1(defuse-01:ARG0 team:ARG1(bomb:ARG1-of(leave-12:ARG2(pole:mod(electricity):location(city:wiki - :name(name:op1 "Okubahal"):ARG1-of(near-02:ARG2(city:wiki "Sundhara":name(name:op1 "Sundhara"):location(district:wiki "Patan,_Nepal":name(name:op1 "Lalitpur"):location(city:wiki "Kathmandu":name(name:op1 "Kathmandu")))))))))))):ARG1-of(serious-02))	One soldier of the Royal Nepal Army has suffered 29 august 2002 because it is impossible to dispose the bomb team in an attempt to defuse the bomb that was left from the electricity pole in Sundhara Lalitpur district of Kathmandu city near Okubahal serious injury
Anti-government insurgents are believed to have planted the bomb.	(believe-01:ARG1(plant-01:ARG0(insurgent:ARG0-of(oppose-01:ARG1(government-organization:ARG0-of(govern-01)))):ARG1(bomb)))	Anti - government insurgents have planted the bomb , believed
The injured soldier has been admitted to the army hospital in Kathmandu.	(admit-02:ARG1(soldier:ARG1-of(injure-01)):ARG2(hospital:poss(army):location(city:wiki "Kathmandu":name(name:op1 "Kathmandu"))))	Injured soldiers admit to the army hospital in Kathmandu
There is no report of any other damage or casualty in the incident.	(report-01:polarity - :ARG1(or:op1(damage:mod(other:mod(any))):op2(casualty)):prep-in(incident))	No incident report that any of the other damage or casualties

SemEval-2017 Task 9

We peek inside your brain so you don't have to!

Parsers and Resources from SemEval 2016

Below you can find links to the system description papers from the participants in SemEval 2016 Task 8. Nearly all task reports contain links to parser code bases. If you are new to AMR parsing or would like to try a new approach, these resources might be quite helpful. Note that they are provided 'as-is.' We have not verified that they will work for you, that the links work, that the descriptions in the papers are complete or accurate, nor that the developers will be able to provide support. Please respect the limitations of the licenseses that accompany this software, if provided.

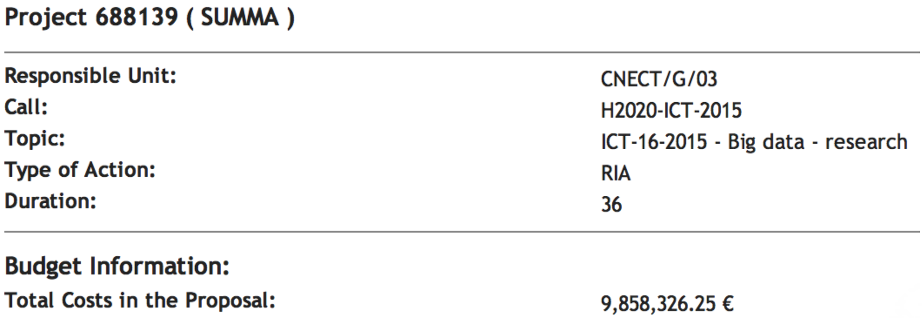
- Guntis Barzdins, Didzis Gosko (University of Latvia, IMCS and LETA)
- Alastair Butler (National Institute for Japanese Language and Linguistics)
- Yevgeniy Puzikov, Daisuke Kawahara, Sadao Kurohashi (Kyoto University)
- Lauritz Brandt, David Grimm, Mengfei Zhou, Yannick Versley (Ruprecht-Karls-Universitat)
- James Goodman[1], Andreas Vlachos[2], Jason Naradowsky[1] ([1]University College London and [2]University of Sheffield)
- Chuan Wang[1], Sameer Pradhan[2], Nianwen Xue[1], Xiaoman Pan[3], Heng Ji[3] ([1]Brandeis University, [2]cemantix.org, and [3]Rensselaer Polytechnic Institute)
- Johannes Bjerva, Johan Bos, Hessel Haagsma (University of Groningen)
- Xiaochang Peng, Daniel Gildea (University of Rochester)
- Sudha Rao, Yogarshi Vyas, Hal Daume III, Philip Resnik (University of Maryland)
- William R. Foland Jr.[1], James H. Martin[2] ([1]OK Robot Go, Ltd. and [2]University of Colorado)
- Jeffrey Flanigan[1], Chris Dyer[1], Noah A. Smith[2], Jaime Carbonell[1] ([1]Carnegie Mellon University and [2]University of Washington)

For more information on the 2016 task as a whole you can consult the task description paper

SUMMA

Scalable Understanding of Multilingual Media

- Horizon-2020
  BigData-Research projekts
- http://summa-project.eu
- Projekta ilgums: 36 mēneši
  (02/2016 – 01/2019)
- Budžets: **9,86M EUR**
  Latvijai: **1,16M EUR**



Project 688139 (SUMMA)

Responsible Unit:	CNECT/G/03
Call:	H2020-ICT-2015
Topic:	ICT-16-2015 - Big data - research
Type of Action:	RIA
Duration:	36

Budget Information:

Total Costs in the Proposal:	9,858,326.25 €

University of Edinburgh

Priberam Informatica S.A.

University College London

Idiap Research Institute

Latvian News Agency

British Broadcasting Corporation

Deutsche Welle

Qatar Computing Research Institute

full stack

Daudzslāņu valodas resursu kopa teksta semantiskai analīzei un sintēzei latviešu valodā

NACIONĀLAIS ATTĪSTĪBAS PLĀNS 2020

EIROPAS SAVIENĪBA
Eiropas Reģionālās attīstības fonds

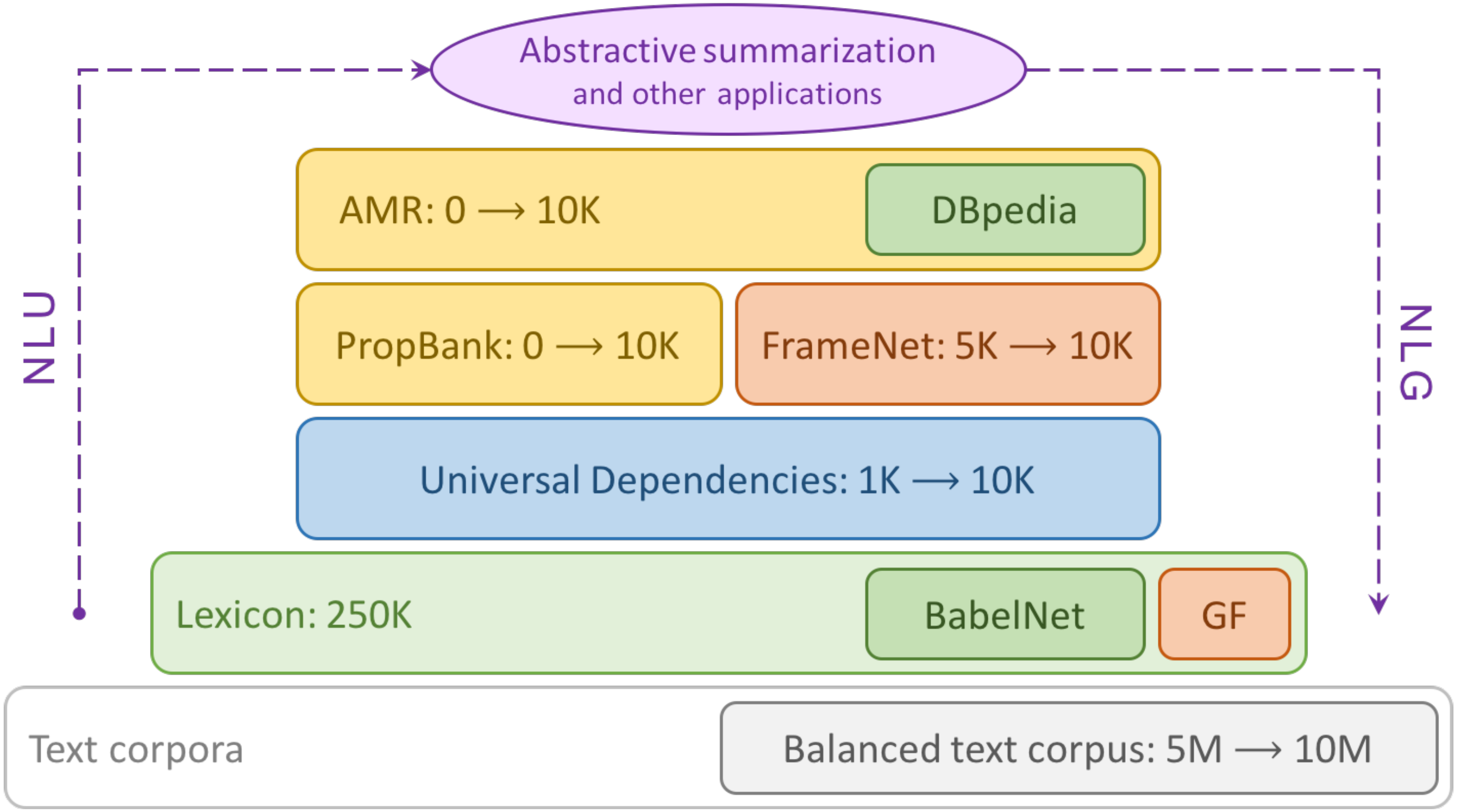
projekts

IEGULDĪJUMS TAVĀ NĀKOTNĒ

Latvijas Universitātes
Matemātikas un informātikas institūta
Mākslīgā intelekta laboratoriju (AI-Lab)

LETA
nacionālā informācijas aģentūra

full stack



NLU

Abstractive summarization and other applications

NLG

AMR: 0 ⟶ 10K

DBpedia

PropBank: 0 ⟶ 10K

FrameNet: 5K ⟶ 10K

Universal Dependencies: 1K ⟶ 10K

Lexicon: 250K

BabelNet

GF

Text corpora

Balanced text corpus: 5M ⟶ 10M

Priekšvēsture

- IT Kompetences centra pētījumi
  - Pētījums par teksta automātiskas analīzes iespējām jauna informācijas arhīva produkta izstrādē (2013)
  - Runas korpusa izveide, principi, metodes, realizācija (2013)
  - Runas atpazīšanas iespēju izpēte audiomateriāla automātiskai transkribēšanai mediju monitoringā (2014)
  - Pētījums par runas atpazīšanas sistēmas pielāgošanu zemas kvalitātes audiofailu apstrādei (2015)
- Pētījums par publicistikā pieminēto entītiju savstarpējo saišu identificēšanu, tām atbilstošo grafu strukturēšanu un datu bāzu vaicājumu attēlošanu grafu veidā (2014–2015)
- Valsts pētījumu programma SOPHIS (2014-2017) http://www.edi.lv/lv/projekti/vpp-projekti/vpp-sophis-projekts-nr2-/
- Valsts pētījumu programma NexIT (2014-2017) http://www.lumii.lv/resource/show/836

Publikācijas

1. Guntis Bārzdiņš, Didzis Goško. RIGA at SemEval-2016 Task 8: Impact of Smatch Extensions and Character-Level Neural Translation on AMR Parsing Accuracy. Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval), Association for Computational Linguistics, 2016, pp. 1143-1147 [http://aclweb.org/anthology/S16-1176]
2. Barzdins G., Renals S., Gosko D., Character-Level Neural Translation for Multilingual Media Monitoring in the SUMMA Project. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), pages 1789–1793, Portorož, Slovenia, 2016. URL http://www.lrec-conf.org/proceedings/lrec2016/pdf/4_Paper.pdf
3. Guntis Bārzdiņš, Pēteris Paikens and Didzis Goško. RIGA: from FrameNet to Semantic Frames with C6.0 Rules. Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval), Association for Computational Linguistics, 2015, pp. 960-964 [http://aclweb.org/anthology/S15-2160]
4. Normunds Grūzītis and Dana Dannélls. A Multilingual FrameNet-based Grammar and Lexicon for Controlled Natural Language. Journal of Language Resources and Evaluation, Springer (in press; SNIP 2014: 2.335) [http://arxiv.org/abs/1511.03924]
5. Normunds Grūzītis and Guntis Bārzdiņš. The role of CNL and AMR in scalable abstractive summarization for multilingual media monitoring. Controlled Natural Language, LNCS 9767, Springer, 2016, pp. 127-130 [http://arxiv.org/abs/1606.05994]
6. Barzdins, G., Gosko, D., Rituma, L., Paikens, P. (2014). Using C5.0 and Exhaustive Search for Boosting Frame-Semantic Parsing Accuracy. In Proceedings of the 9th Language Resources and Evaluation Conference (LREC). Reykjavik. http://www.lrec-conf.org/proceedings/lrec2014/summaries/515.html