

# Delovni sklop 3: Semantične tehnologije

Simon Krek, Slavko Žitnik

[www.eu-skladi.si](http://www.eu-skladi.si)



# Kaj so semantične tehnologije?

“Semantične tehnologije uporabljajo metode **umetne intelligence**, da simulirajo razumevanje jezika in procesiranje informacij.”



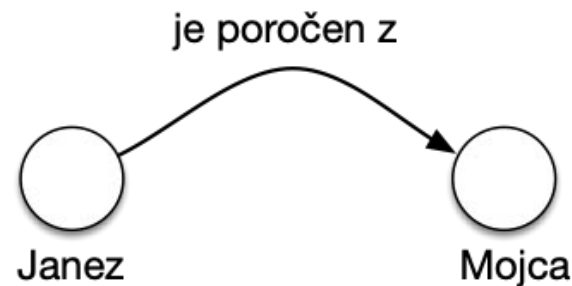
## Podatek

Originalno iz 1600 s pomenom  
“nekaj danega“



## Informacija

Originalno iz 1300 z nanašanjem na  
“akt informiranja“



# O Sklopu 3 – orodja

Nadgradnja “digitalne slovarske baze”

„Pregibalnik“

Prepoznavanje imenskih entitet (DEMO)

Ekstrakcija povezav (DEMO)

Odkrivanje koreferenčnost (DEMO)

Izdelava baze znanja (DEMO)

Orodje za razdvoumljanje (DEMO)

Prepoznavanje semantičnih premikov in izvajanje diahronih analiz (DEMO)

Avtomatsko povzemanje krajših in daljših besedil (DEMO)

Avtomatsko odgovarjanje na vprašanja (DEMO)

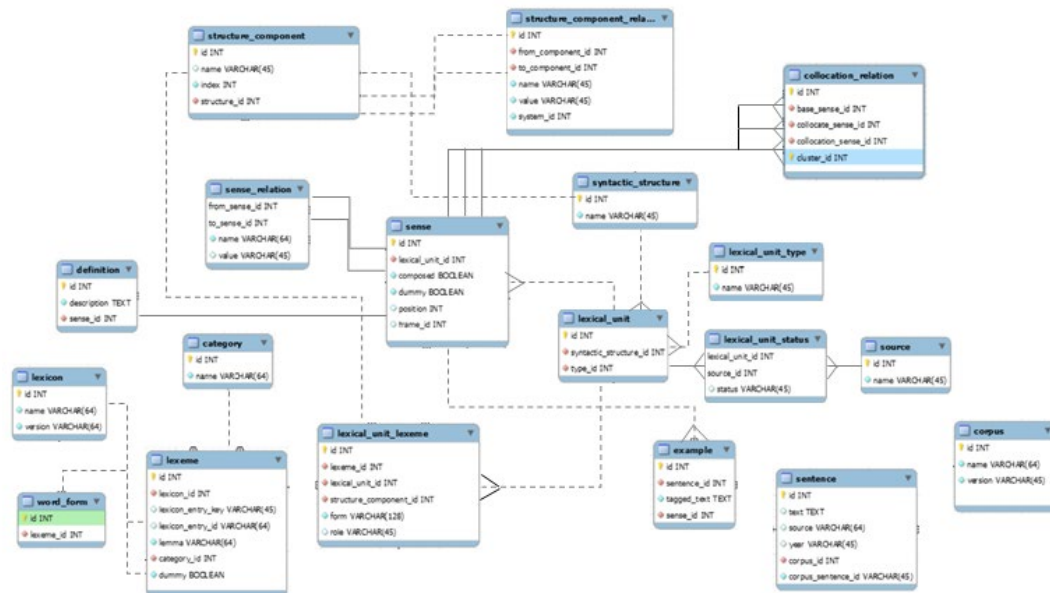
**Orodje za izračun kontekstnih vložitev**



# Nadgradnja "digitalne slovarske baze": opis podatkovnega modela in javni API



<https://wiki.cjvt.si/books/digital-dictionary-database>



## REST API

Public REST API for accessing the database.

### API design

Principles of the API design: All documented routes should be appended to <https://blisk.ijs.si/>...

### API routes

The API is being designed and developed, with priority on current needs. Specifications are avail...

### API implementation

The public API is being implemented using the Django REST Framework and APIViews in particular. I...

### API use cases

In addition to providing general public access to the database, the REST API can also be used to ...

# "Pregibalnik" (ORODJE)



Za podano besedo zgenerirati različne oblike besede (npr. sklon, oseba, število, spol, ...)

## **Generator:**

"eholokacija" -> "eholokacija", "eholokacije", "eholokaciji", ...

## **Naglaševalnik:**

"eholokácija", "eholokácije", "eholokáciji", ...

## **Fonetični pretvornik (IPA, SAMPA):**

"ɛxɔlɔ'ka:tsija", ..., 'ExOIO"ka:tsija', ...

Orodje je na voljo preko spletnega programskega vmesnika na <https://orodja.cjvt.si/pregibalnik>.

# Prepoznavanje imenskih entitet in koreferenčnosti

Audi je izdelovalec luksuznih avtomobilov.  
Podjetje je bilo ustanovljeno v Nemčiji.

Ustanovil ga je August Horch v letu 1910.  
Horch je pred tem imel že drugo podjetje s  
svojimi popularnimi modeli. V Audiju so  
začeli s štiri-cilindrskimi modeli. Do leta 1914  
so Horchovi avtomobili že dirkali in zmagovali.

August Horch je podjetje Audi zapustil leta  
1920 in prevzel mesto predstavnika za  
združenje motornih vozil Nemčije.

Audi je trenutno hčerinsko podjetje skupine  
Volkswagen in proizvaja kvalitetne avtomobile.



# Prepoznavanje imenskih entitet in koreferenčnosti

**Audi** je izdelovalec luksuznih avtomobilov.

**Podjetje** je bilo ustanovljeno v **Nemčiji**.

Ustanovil **ga** je **August Horch** v letu 1910.

**Horch** je pred tem imel že drugo podjetje s **svojimi** popularnimi modeli. V **Audiju** so začeli s štiri-cilindrskimi modeli. Do leta 1914 so **Horchovi** avtomobili že dirkali in zmagovali.

**August Horch** je **podjetje Audi** zapustil leta 1920 in prevzel mesto predstavnika za združenje motornih vozil **Nemčije**.

**Audi** je trenutno hčerinsko podjetje **skupine Volkswagen** in proizvaja kvalitetne avtomobile.



# Prepoznavanje imenskih entitet in koreferenčnosti



**Audi** je izdelovalec luksuznih avtomobilov.  
**Podjetje** je bilo ustanovljeno v **Nemčiji**.

Ustanovil **ga** je **August Horch** v letu 1910.  
**Horch** je pred tem imel že drugo podjetje s **svojimi** popularnimi modeli. V **Audiju** so začeli s štiri-cilindrskimi modeli. Do leta 1914 so **Horchovi** avtomobili že dirkali in zmagovali.

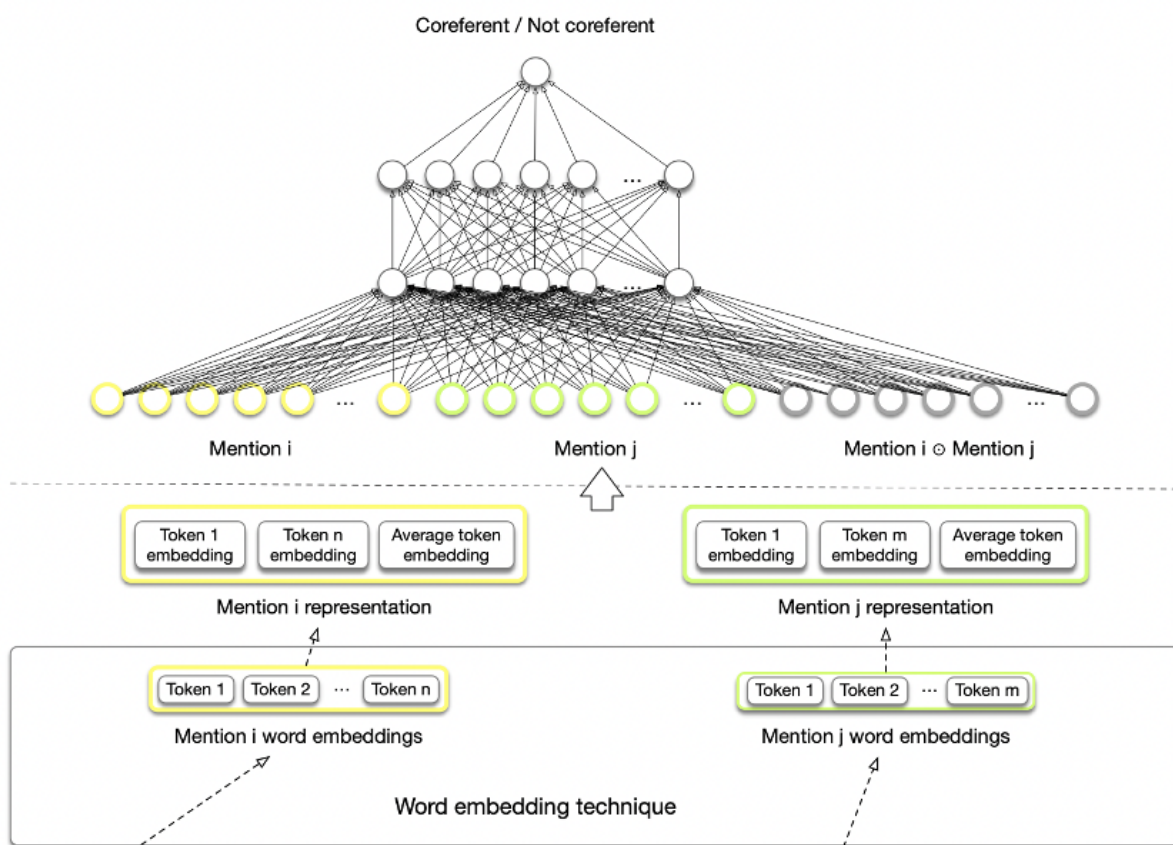
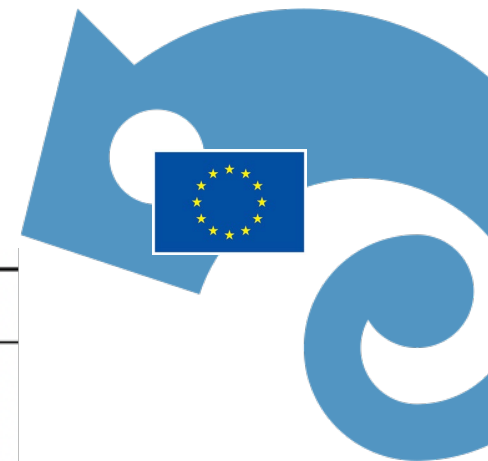
**August Horch** je **podjetje Audi** zapustil leta 1920 in prevzel mesto predstavnika za združenje motornih vozil **Nemčije**.

**Audi** je trenutno hčerinsko podjetje **skupine Volkswagen** in proizvaja kvalitetne avtomobile.





# Prepoznavanje imenskih entitet in koreferenčnosti (Orodje)



... Michel Barnier pointed that out as a difficult job. EC's Head of Task Force for Relations with the UK initiated ...

Statistic	coref149	SentiCoref 1.0
Documents	149	837
Tokens	26,960	433,139
Entities	1,277	14,572
Trivial	831	7,721
Mentions	2,329	42,738
Overlapping	196	4,212

	ACE 2004	SemEval2010	CoNLL-2012
	450	314	2,135
	191,387	102,952	1,468,889
	12,439	20,921	37,330
	-	-	-
	29,724	28,242	174,437
	-	-	-

Klemen M. and Žitnik S. (2021). **Neural coreference resolution for Slovene language**, in *Computer science and information systems*, vol. 19, num. 2, pp. 495-521.



## Prepoznavanje imenskih entitet

Prepoznavanje imenskih entitet je naloga obdelave naravnega jezika, pri kateri se za vsako pojavnico (t.j. večinoma besedo) določi, kateri tip objekta predstavlja. Imenske entitete predstavljajo lastna imena oseb, organizacij, zemljepisna imena ter stvarna imena. Algoritmi poskušajo pri prepoznavanju imenskih entitet upoštevati pomen besede ali besedne zveze v kontekstu, v katerem se nahaja. Naloga poleg odkrivanja koreferenčnosti in ekstrakcije povezav predstavlja eno izmed osnovnih in ključnih nalog ekstrakcije informacij.

Spletna storitev je namenjena izključno demonstracijskim namenom in je omejena s številom zahtevkov na časovno enoto ter vnosom dolžine besedila. Za uporabo storitve v okviru vaših aplikacij, si prenesite rezultate projekta, ki so objavljeni v repozitoriju Clarin.si.

Janez Novak je šel v Volkswagnov salon. Tam je opazoval različna vozila, jih preskušal na relacijah od Maribora preko Ljubljane do Portoroža. Na koncu se je odločil za nakup novega Golfa 9. ✕

Vstavi vzorčno besedilo

189/500

Prepoznaj imenske entitete

Janez Novak je šel v Volkswagnov salon. Tam je opazoval različna vozila, jih preskušal na relacijah od Maribora preko Ljubljane do Portoroža. Na koncu se je odločil za nakup novega Golfa 9.



Vstavi vzorčno besedilo

189/500

Prepoznavaj imenske entitete

## Rezultat

Janez **B-PER** Novak **I-PER** je šel v Volkswagnov **B-ORG** salon **I-ORG** . Tam je opazoval različna vozila, jih preskušal na relacijah od Maribora **B-LOC** preko Ljubljane **B-LOC** do Portoroža **B-LOC** . Na koncu se je odločil za nakup novega Golfa **B-PRO** 9 **I-PRO** .

## Odkrivanje koreferenčnosti

Odkrivanje koreferenčnosti (oz. razreševanje koreferenc) pomeni gručenje omenitev oz. združevanje omenitev, ki se sklicujejo na isto entiteto. Omenitev predstavlja sklic na pomensko entiteto, ki je v besedilu omenjena kot imenska entiteta, z zaimkom ali pa je implicitno vsebovana v glagolski obliki. Rezultat naloge so tako gruče omenitev, kjer vsaka gruča predstavlja eno entiteto. Algoritmi za odkrivanje koreferenčnosti morajo delovati vsaj na nivoju celotnega dokumenta in poskušati razumeti diskurs na višjem semantičnem nivoju - raboslovju, da lahko dosegajo dobre rezultate. Poleg prepoznavanja imenskih entitet in ekstrakcije povezav je to bolj zahtevna naloga ekstrakcije informacij.

Spletna storitev je namenjena izključno demonstracijskim namenom in je omejena s številom zahtevkov na časovno enoto ter vnosom dolžine besedila. Za uporabo storitve v okviru vaših aplikacij, si prenesite rezultate projekta, ki so objavljeni v repozitoriju Clarin.si.

Janez Novak je šel v avtomobilski salon, kjer je preskušal novo vozilo. Le to mu je bilo zelo všeč, zato se je odločil za nakup novega Golfa 9.



Vstavi vzorčno besedilo

143/200

Prikaži tudi omenitve brez koreferenc

Minimalna stopnja zaupanja prepoznane koreferenčnosti:  0.25

Odkrij koreferenčnosti

Janez Novak je šel v avtomobilski salon, kjer je preskušal novo vozilo. Le to mu je bilo zelo všeč, zato se je odločil za nakup novega Golfa 9.

Vstavi vzorčno besedilo

143/200

Prikaži tudi omenitve brez koreferenc

Minimalna stopnja zaupanja prepoznane koreferenčnosti:  0.25

Odkrij koreferenčnosti

## Rezultat

Janez Novak PER je šel O v avtomobilski salon O, kjer je preskušal O novo vozilo O. Le to mu O je bilo O zelo všeč O, zato se O je odločil O za nakup O novega Golfa MISC 9.

# Ekstrakcija povezav



Janez Drnovšek je bil predsednik Liberalne demokracije Slovenije.

# Ekstrakcija povezav



Janez Drnovšek je bil predsednik **Liberalne demokracije Slovenije**.

# Ekstrakcija povezav



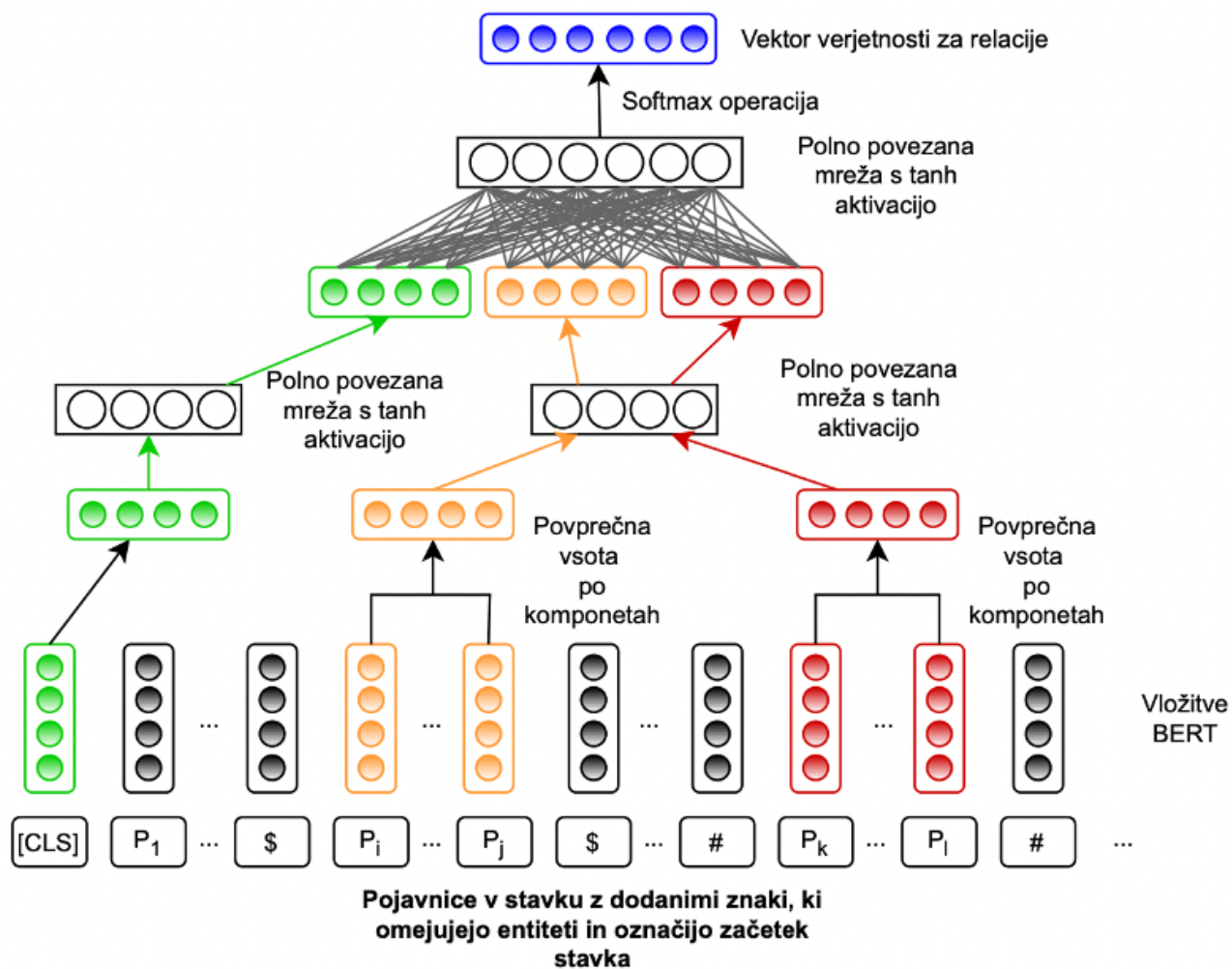
Janez Drnovšek je bil *predsednik* Liberalne demokracije Slovenije.







# Ekstrakcija povezav (Orodje)



## Modeli

BERT

RECON

LSTM arhitektura po meri

# Ekstrakcija povezav

Avtomatsko izdelan korpus  
za ekstrakcijo povezav

**29 tipov povezav**

Wikidata značka	ime relacije	celoten korpus	učna množica	validacijska množica	testna množica
P0	prazna	0	30721	4455	8901
P131	se nahaja v	31745	28644	4110	8217
P150	administrativne podenote	22061	15400	2165	4496
P31	primerek od	18463	13064	1819	3580
P106	poklic	16605	11543	1670	3392
P527	vsebuje	13521	9460	1410	2651
P17	država, kateri objekt pripada	72806	9299	1333	2701
P156	naslednik	10841	7733	1041	2067
P155	predhodnik	10709	7647	1030	2032
P361	je del	9930	6988	1005	1937
P19	kraj rojstva	9812	6094	880	1799
P3450	sezona športne prireditve	5820	4105	603	1112
P20	kraj smrti	7041	3650	491	989
P138	poimenovano po	4459	3083	440	936
P641	šport povezan s tem	4147	2910	410	827
P172	etnična skupina	4015	2837	366	812
P27	država državljanstva	8861	2830	389	762
P276	kraj, lokacija	3641	2575	361	705
P3373	sorojenec	3139	2206	322	611
P607	bitka povezana s tem	3013	2093	301	619
P1001	spada pod upravo	2667	1884	258	525
P279	podpomenka od	12621	1818	271	500
P50	avtor	1344	1444	193	431
P140	vera	2052	1421	219	414
P136	žanr	1673	1191	160	322
P40	otrok	2455	1047	133	238
P39	uradni položaj	1294	910	129	255
P463	član organizacije	1231	902	113	216
P22	oče	1989	837	102	203
P25	mati	451	220	28	38

## Ekstrakcija povezav

Ekstrakcija povezav je naloga obdelave naravnega jezika, ki poskuša med besedami in/ali besednimi zvezami odkriti pomenske povezave. V splošnem je lahko naloga definirana le kot iskanje povezave brez specifične identifikacije tipa povezave. Prav tako je lahko omejena na iskanje povezav med dokumenti ali znotraj posameznega stavka. V okviru projekta smo polavtomatsko zgradili korpus in naučili model, ki je sposoben prepoznavati 29 tipov povezav med omenitvami znotraj posameznega stavka. Omenitev lahko predstavlja imensko entiteto ali sklic na imensko entiteto (npr. zaimsek). Trenutni model je prilagojen na besedila, ki so dostopna na Wikipediji, zato deluje slabše na splošnih besedilih. Poleg prepoznavanja imenskih entitet in odkrivanja koreferenčnosti ekstrakcija povezav predstavlja eno izmed osnovnih in ključnih nalog ekstrakcije informacij.

Spletna storitev je namenjena izključno demonstracijskim namenom in je omejena s številom zahtevkov na časovno enoto ter vnosom dolžine besedila. Za uporabo storitve v okviru vaših aplikacij, si prenesite rezultate projekta, ki so objavljeni v repozitoriju Clarin.si.

Janez, predsednik stranke Slovenski ljubitelji, je rojen na Malem Lipoglavu.



Vstavi vzorčno besedilo

76/200

Kot omenitve upoštevaj le imenske entitete

Minimalna stopnja zaupanja prepoznane povezave:  0.25

Prepoznavaj povezave

Janez, predsednik stranke Slovenski ljubitelji, je rojen na Malem Lipoglavu.



Vstavi vzorčno besedilo

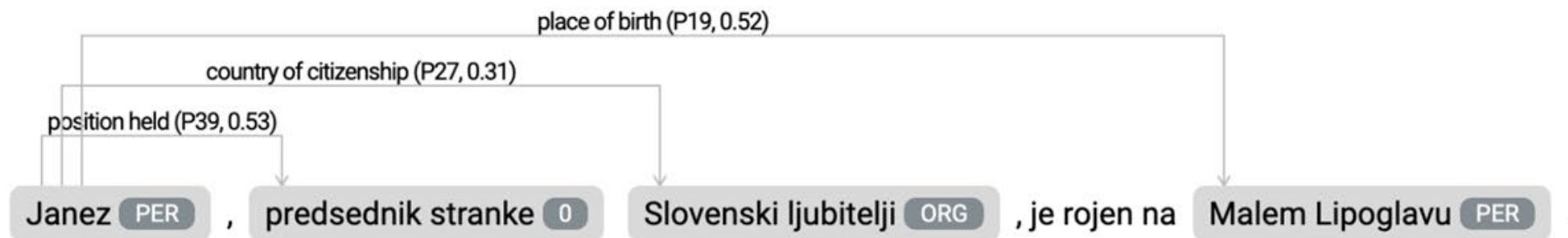
76/200

Kot omenitve upoštevaj le imenske entitete

Minimalna stopnja zaupanja prepoznane povezave:  0.25

Prepoznaj povezave

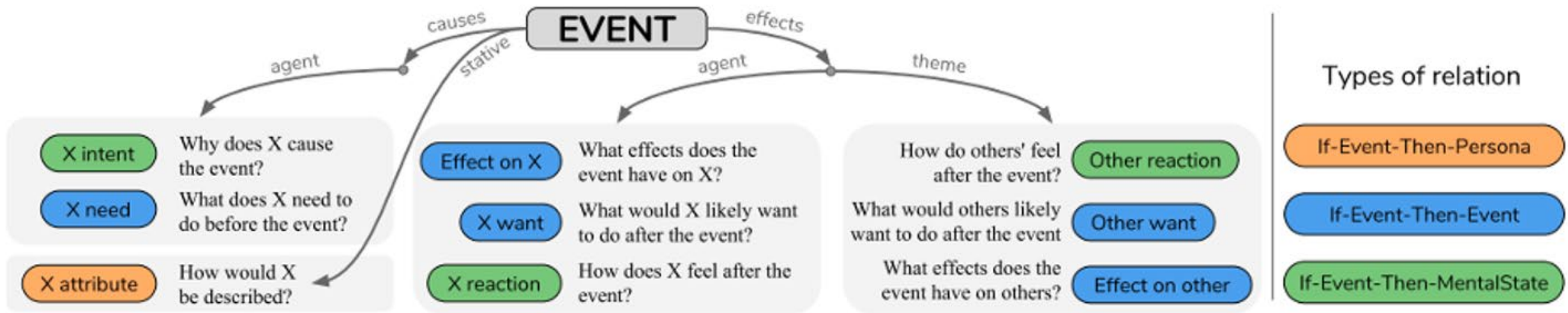
## Rezultat



# Baza znanja

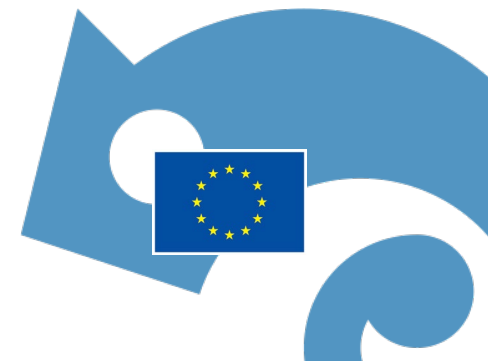


Allen Institute for AI (2020): *Commonsense Inferences about People and Events*



# Baza znanja (Orodje za prevajanje)

Ročno prevedenih cca. 10% množice (10.000 primerov).



OK HEAD: PersonX acts quickly (izvirnik)  
OsebaX deluje hitro (prevod)

RELATION: xWant

## Vnos prevodov

Izvirnik: PersonX acts quickly OsebaX deluje hitro

Izvirnik: xWant xWant (posledično, OsebaX želi)

Izvirnik: to escape from him pobegniti pred neki popravek

## Sinonimi

Izberite besedo z miškinim kazalcem, da sprožite iskanje sinonimov

X

Naziv

Razlaga

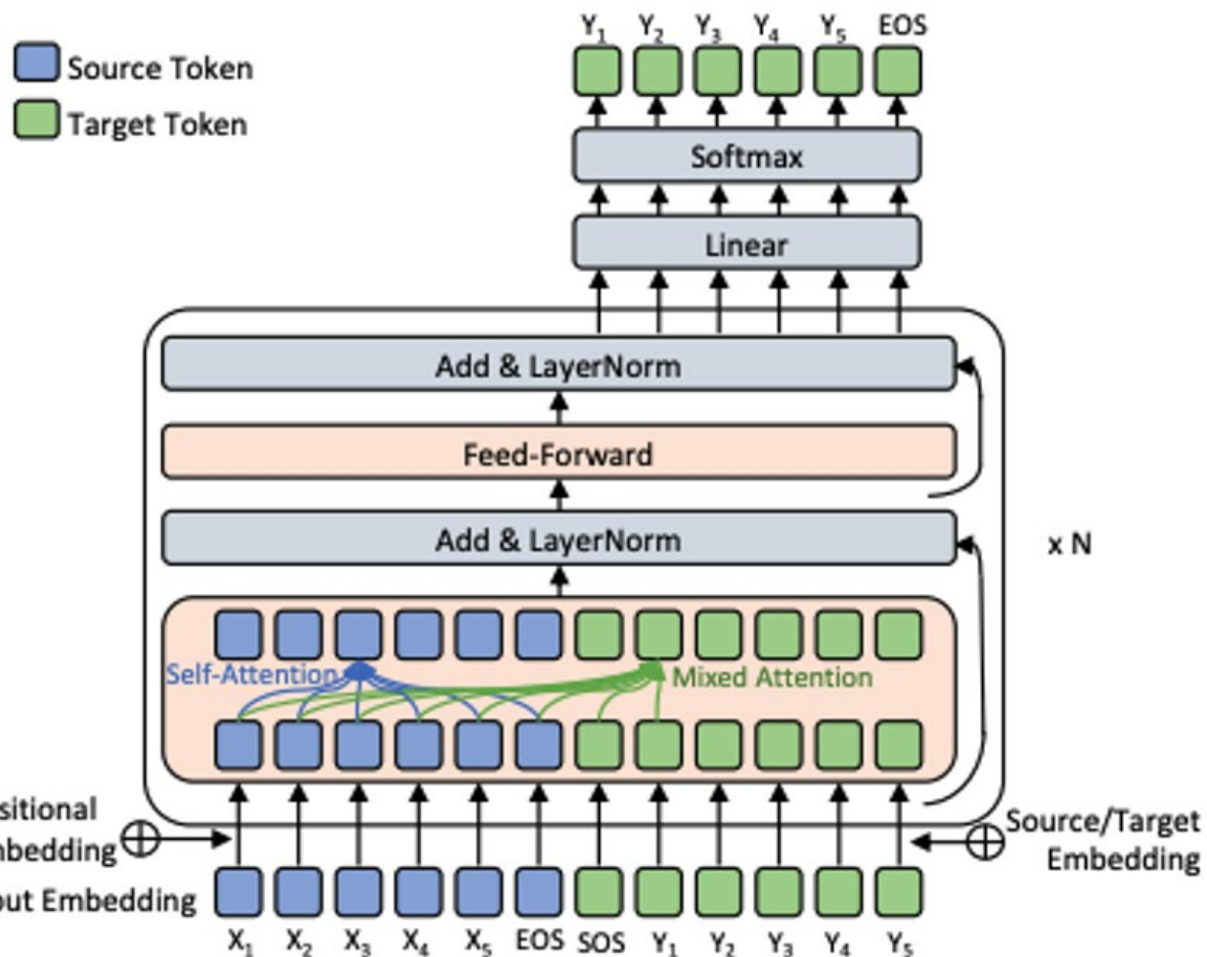
## Dodajanje komentarjev

Privzeti komentari

I



# Baza znanja (ORODJE)



23 tipov relacij

"{vhodni stavek} {relacija} [GEN]  
 {izhodni stavek}[EOS]"

Model	Jezik	Množica	BLEU-1	ROUGE-L
cjvt/gpt-si-base	Slo.	Avto. ALL	0.34	0.41
	Slo.	Avto. 10k	0.35	0.42
	Slo.	Roč. 10k	0.28	0.36
COMET (GPT2-XL)	Angl.	Original	0.41	0.49





Slovene ▾

People &amp; Events ▾

Mojca je pojedla odličen sendvič

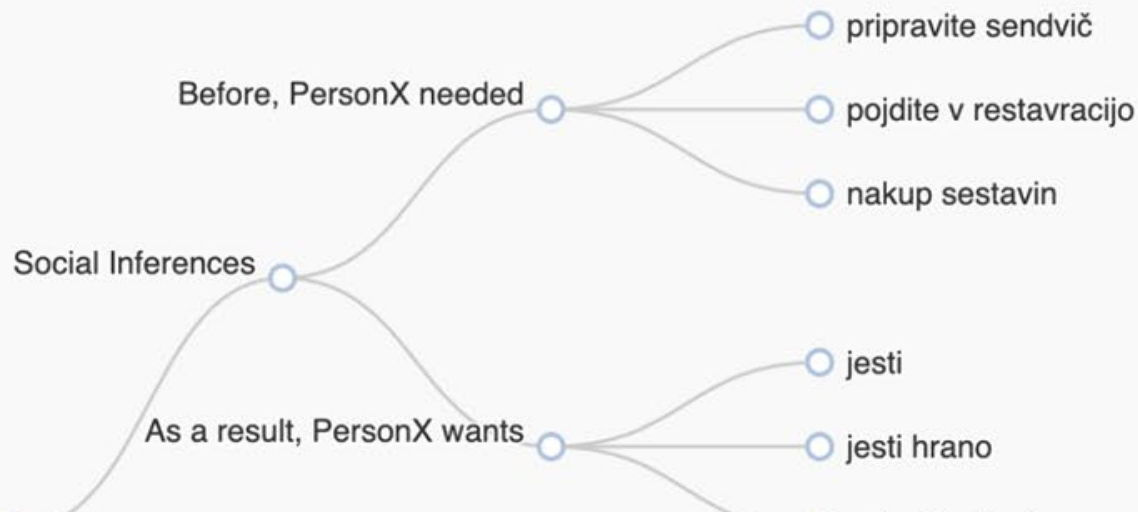
Submit

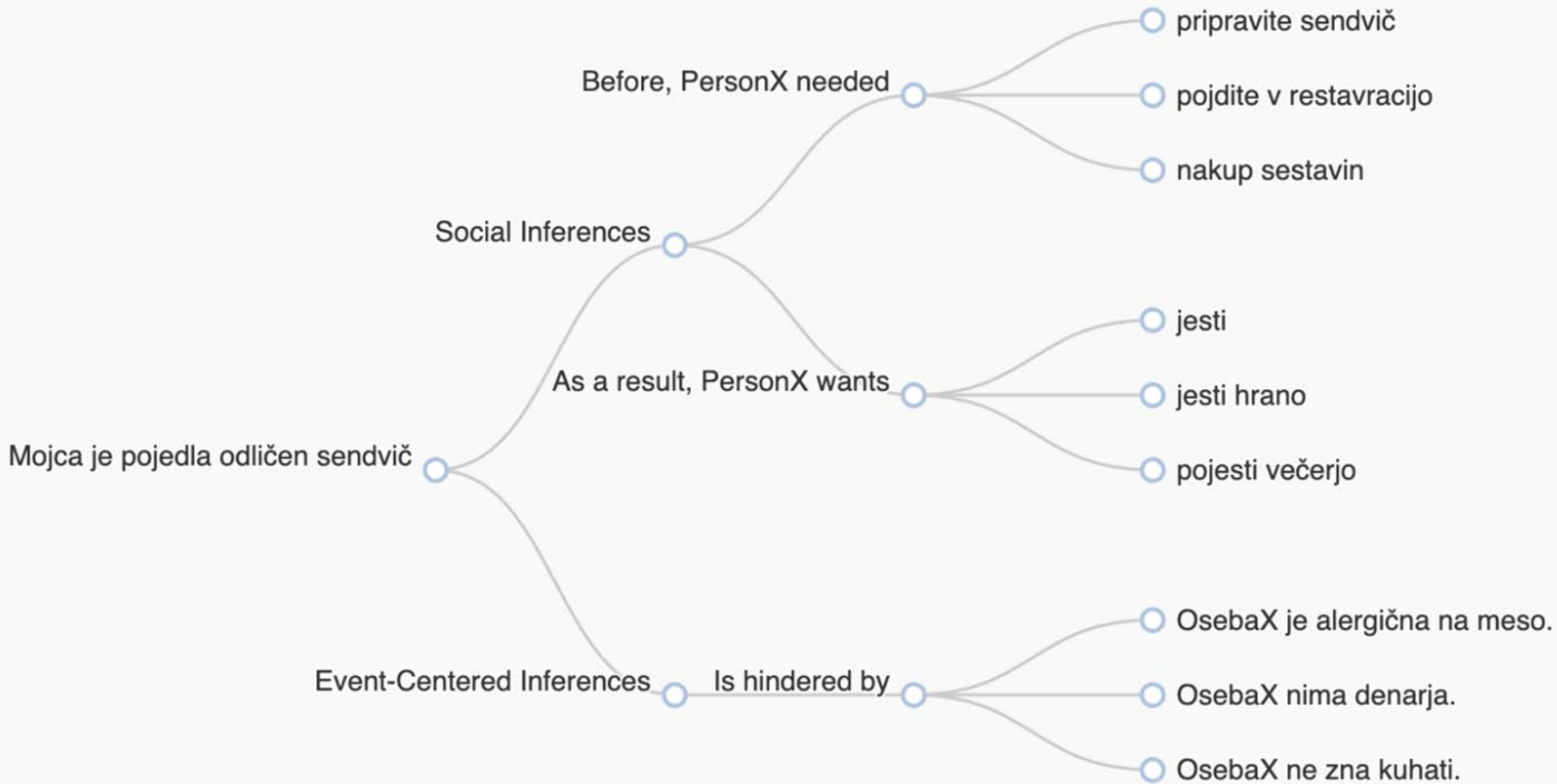
Select 1 to 4 descriptors:

- Before, PersonX needed  PersonX is seen as  As a result, PersonX feels  PersonX then  As a result, PersonX wants
- Because PersonX wanted  As a result, others feel  Others then  As a result, others want  Happens after
- A possible subevent is  Happens before  Is hindered by  Causes  As a result, PersonX reasons
- If there is a blank, it can be filled by

**Poskusi:** [OsebaX hitro reagira](#), [Janez je zelo pomemben](#), [Mojca je pojedla odličen sendvič](#)

**Model omejen le na napovedovanje 4 tipov relacij, napovedovanje traja cca. 30s.**





# Povzemanje besedil (ORODJE)

STA, AutoSentiNews, SURS, KAS, CNN-DM

**Metamodel** (prednapoved modela, ki bo dal najboljšo oceno ROUGE)

**Basic** (frekvence besed)

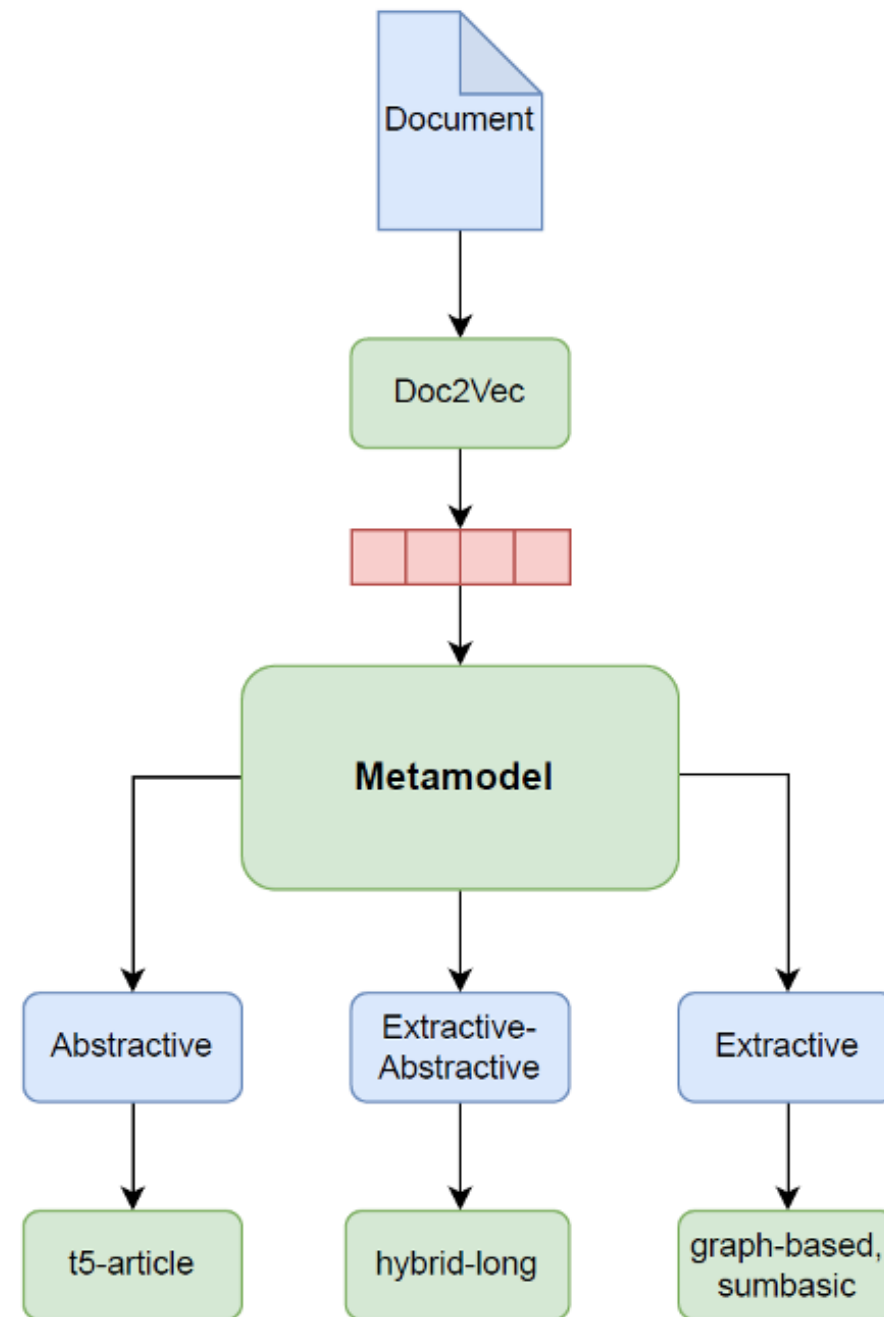
**Graph-based** (grafi + kodirnik LaBSE)

**Headline** (SloT5 + doučitev na prevodu CNN-DM)

**Article** (SloT5 + doučitev na prevodu CNN-DM)

**Hybrid-long** (T5 + graph-based; pomembni stavki, abstraktivno pokrajšani)

Žagar A., Robnik Šikonja M. (2022). **Cross-lingual transfer of abstractive summarizer to less-resource language**, in *Journal of intelligent information systems*, 58(1), str. 153-173.



- **Basic model** - nenadzorovan enostaven povzemalnik, ki uporablja frekvence besed in vrača N najbolj pomembnih povedi.
- **Hybrid-long model** - nenadzorovan hibridni (na osnovi grafov in modelov transformer) pristop, ki vrača kratke povzetke dolgih besedil.

Spletna storitev je namenjena izključno demonstracijskim namenom in je omejena s številom zahtevkov na časovno enoto ter vnosom dolžine besedila. Za uporabo storitve v okviru vaših aplikacij, si prenesite rezultate projekta, ki so objavljeni v repozitoriju Clarin.si.

#### BESEDILO

Slovenija se po večini glavnih kazalnikov evropskega stebra socialnih pravic uvršča visoko med državami EU, vseeno pa so pred njo še številni izzivi. Pri kazalniku neizpolnjene potrebe po zdravstveni oskrbi denimo zaostaja za večino članic, pri osnovnih digitalnih spretnostih pa je tik za povprečjem EU, ugotavlja Umar. Evropski steber socialnih pravic so leta 2017 razglasili Evropski parlament, Evropski svet in Evropska komisija, da bi se s primernimi, dostopnimi in finančno vzdržnimi sistemi socialne zaščite ter ustreznimi politikami trga dela in enakih možnosti zvišali socialni standardi v državah članicah EU. Podpisniki so evropski steber socialnih pravic utemeljili na 20 načelih s treh področij, in sicer enakih možnosti in dostopu do trga dela, poštenih delovnih pogojev ter socialni zaščiti in vključenosti. Stanje v Sloveniji na področju teh 20

Vstavi vzorčno besedilo

1358/3000

#### POVZEMALNIK

Metamodel



Povzemi



Metamodel

[Povzemi](#)

## Rezultat

Povzetek

### Article model



#### Povzetek

Evropski steber socialnih pravic so leta 2017 razglasili Evropski parlament, Evropski svet in Evropska komisija. Na področju enakih možnosti in dostopa do trga dela so ugotovili visoko vključenost mladih v izobraževanje in nizko neenakost v izobraževanju. Javni sistem izobraževanja zagotavlja dobro dostopnost izobraževanja za skoraj vse skupine prebivalcev.

Graph-based model



Povzemi

## Rezultat

Povzetek

### Graph-based model



#### Povzetek

Slovenija se po večini glavnih kazalnikov evropskega stebra socialnih pravic uvršča visoko med državami EU, vseeno pa so pred njo še številni izzivi. Evropski steber socialnih pravic so leta 2017 razglasili Evropski parlament, Evropski svet in Evropska komisija, da bi se s primernimi, dostopnimi in finančno vzdržnimi sistemi socialne zaščite ter ustreznimi politikami trga dela in enakih možnosti zvišali socialni standardi v državah članicah EU.

Vstavi vzorčno besedilo

1558/3000

POVZEMALNIK

Headline model



Povzemi

## Rezultat

Povzetek

### Headline model



**Povzetek**

Slovenija po večini kazalnikov evropskega stebra socialnih pravic visoko med državami EU

Basic model



Povzemi

## Rezultat

Povzetek

### Basic model

 **Povzetek**

Evropski steber socialnih pravic so leta 2017 razglasili Evropski parlament, Evropski svet in Evropska komisija, da bi se s primernimi, dostopnimi in finančno vzdržnimi sistemi socialne zaščite ter ustreznimi politikami trga dela in enakih možnosti zvišali socialni standardi v državah članicah EU. Na področju enakih možnosti in dostopa do trga dela je ugotovil visoko vključenost mladih v izobraževanje in nizko neenakost v izobraževanju.



## Rezultat

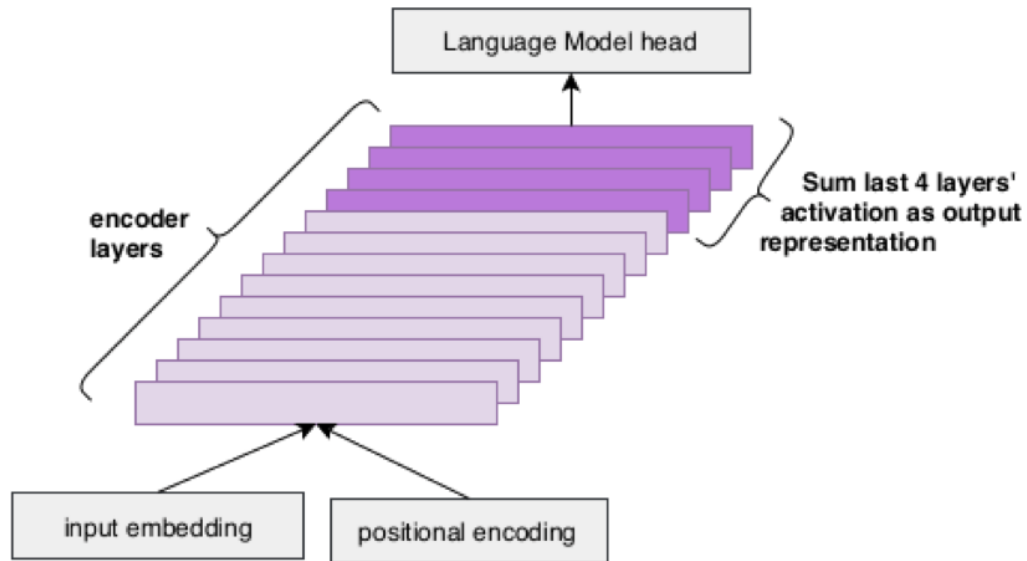
Povzetek

### Hybrid-long model

#### Povzetek

Slovenija se po večini glavnih kazalnikov socialnega varstva uvršča visoko med državami EU. Slovenija zaostaja za večino članic, pri osnovnih digitalnih spretnostih pa zaostaja za povprečjem EU. Evropski steber socialnih pravic so leta 2017 razglasili Evropski parlament, Evropski svet in Evropska komisija. Podpisniki so evropski steber socialnih pravic utemeljili na 20 načelih. Vključujejo enake možnosti in dostop do trga dela, poštenih delovnih pogojev. Stanje v Sloveniji na področju 20 načel je analiziral Urad RS za ekonomske analize in razvoj. Na področju enakih možnosti in dostopa do trga dela so ugotovili visoko vključenost mladih v izobraževanje in nizko neenakost v izobraževanju. Javni sistem izobraževanja v Sloveniji zagotavlja dobro dostopnost izobraževanja za skoraj vse skupine prebivalcev. Število vključenih v sekundarno in terciarno izobraževanje je močno nad povprečjem EU.

# Semantični premiki in diahrone analize (ORODJE)

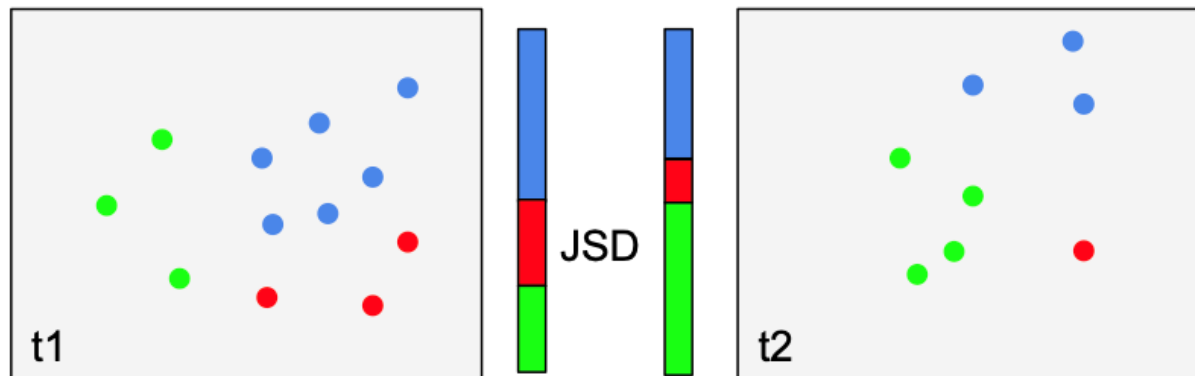


## Gigafida

2017, 2002, 2007,  
2013, 2018

## Doučen SloBERTa (MLM)

4 nivoji -> gručenje  
normalizacija  
primerjava porazdelitev (JSD)



Montariol S., Martinc M., Pivovarova L. (2021). **Scalable and Interpretable Semantic Change Detection**, in *Proceedings of the 2021 Conference of the NAACL: Human Language Technologies*, str. 4642-4652.

## SEMANTIC CHANGE DETECTION FROM 1997 TO 2018

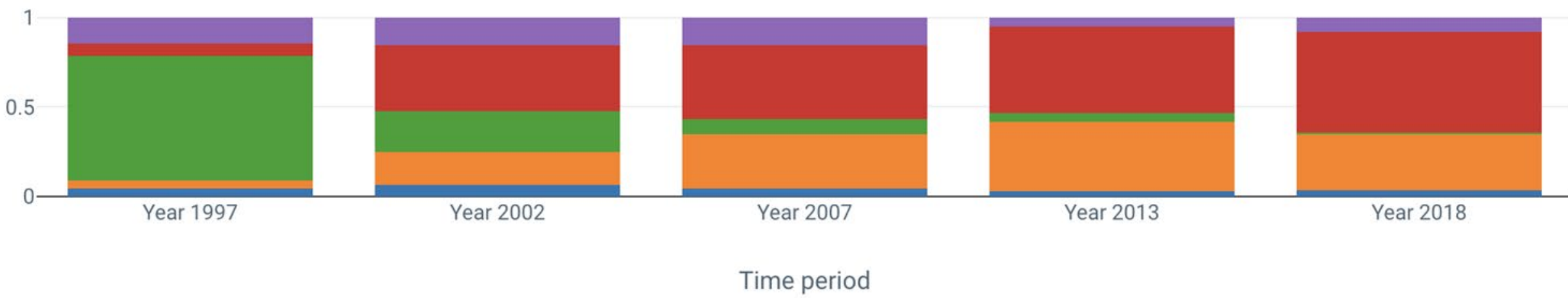
The words are by default sorted according to the JSD K5 semantic change score (bigger number indicates bigger change in word usage across two time periods). If there are more than two time periods, the default sorting is by JSD K5 All (semantic change between first and last time period). JSD K5 Avg measures average change across all consecutive time periods. Click on a specific word to get details.



Word/Beseda	JSD K5 1997- 2002	JSD K5 2002- 2007	JSD K5 2007- 2013	JSD K5 2013- 2018	JSD K5 All	JSD K5 Avg	FREQ 1997	FREQ 2002	FREQ 2007	FREQ 2013	FREQ 2018
diagonalen	0.00003	0.003856	0.382575	0.00527	0.585142	0.097933	46	61	139	216	649
pogovoren	0.368549	0.065874	0.042726	0.007298	0.509551	0.121112	2396	473	553	238	438
evro	0.26191	0.068198	0.001052	0.004374	0.499824	0.083884	413	7203	56720	45911	3700

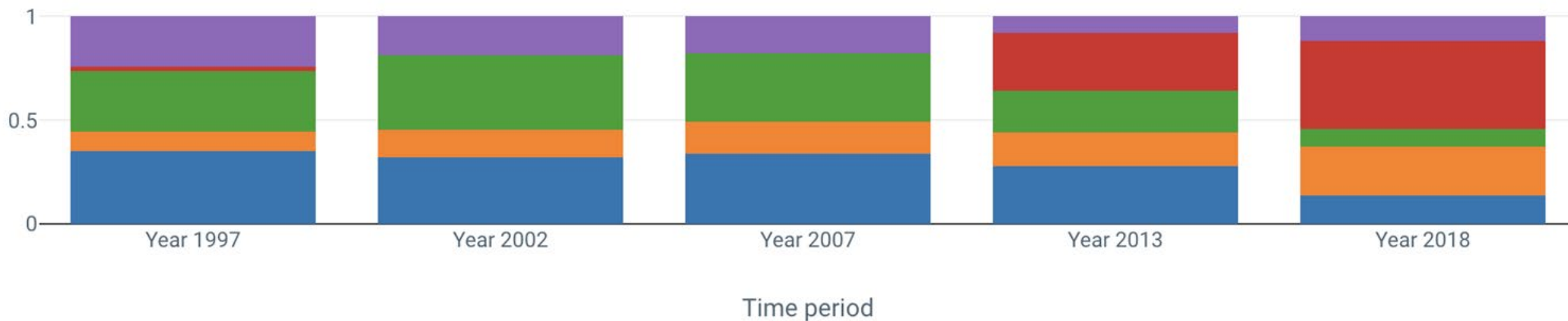


Usage distribution



- igralen plošček, konzola, blu ray plošček, knjiga, format, brezžičen plošček, biti plošček
- biti plošček, vratar plošček, strel, tretjina plošček, tekma plošček, poslati plošček, zadeti plošček
- skladba, pesem, glasben plošček, tehno, plošček slišati, ime, biti plošček
- strel plošček, biti plošček, vratar plošček, poslati plošček, tretjina, zadeti plošček, odbiti plošček
- zadeti, odbiti, palica, strel, vratar, drsalka, črta

Usage distribution



- hlev, krava, bikec, imeti, telica, prodati teliček, živina
- krava, hlev, dan, skotiti teliček, ležati, žival, moči
- imeti teliček, krava teliček, teliček biti, pustiti, dedek, hlev, mama

- dallas, tekma, točka, dončič, trener teliček, teliček rick carlisle, sezona
- teliček biti, točka, zmaga, imeti teliček, nowitzki, končnica, zabava

# Rezultati delovnega sklopa 3



Viri (avt., roč., vložitve): **9**

Orodja: **12**

Smernice za označevanje: **4**

Jezikovnotehnološka infrastruktura

Načrt in zasnova splošnega orodja za obdelavo naravnega jezika

<https://github.com/RSDO-DS3>

# Sodelujoči

## UL FRI

Marko Robnik-Šikonja

Slavko Žitnik

## UM FERI

Milan Ojsteršek

## IJS

Senja Pollak

Erik Novak

## UL FF, UNG, ZRC SAZU

Simon Krek

