Medujezično prepoznavanje imenovanih entiteta pomoću wikifikacije

Stipan Mikulić

Mentor: doc. dr. sc. Jan Šnajder

Fakultet elektortehnike i računarstva
Sveučilište u Zagrebu

1. lipnja 2017.

2017-06-01

Seminar

Medujezično prepoznavanje imenovanih entiteta pomoću wikifikacije

Stipan Mikulić

Mentor: doc. dr. sc. Jan Šnajder

Fakultet elektortehnike i računarstva
Sveučilište u Zagrebu

1. lipnja 2017.

Dobar dan svima. Ja sam Stipan Mikulić. Tema moga seminara je medujezicko prepoznavanje imenovanih entiteta pomocu wikifikacije.

Sadržaj

Sadržaj

Opis problema i motivacija

- za razvoj dobrog modela potrebno je puno podataka
- više od 50% sadržaja na internetu je na engleskom jeziku
- Referencirani rad: Chen-Tse Tsai and Stephen D. Mayhew and Dan Roth, Cross-Lingual Named Entity Recognition via Wikification,

Opis problema i motivacija

- za razvoj dobrog modela potrebno je puno podataka
- više od 50% sadržaja na internetu je na engleskom jeziku
- Referencirani rad: Chen-Tse Tsai and Stephen D. Mayhew and Dan Roth, Cross-Lingual Named Entity Recognition via Wikification,

Za razvoj dobrog modela za klasifikaciju potrebno nam je puno podataka. Prema zadnjim procjenama više od 50% sadžaja na internetu je pisano na engleskom jeziku. U potpunoj dominaciji engleskog jezika u svim vrstama podataka i NLP alata i leži motivacija za razvoj medjezičnog modela.

Opis problema i motivacija

Prepoznavanje imenovanih entiteta je zadatak klasifikacije elemenata u predefinirane kategorije kao što su:

- Imena – Osobe, Organizacije, Lokacije
- Vremenske oznake – Vrijeme, Datum
- Brojevi – Novac, Postotci
- ...

$[\text{Jim}]_{\text{OSOBA}}$ bought $[300]_{\text{BROJ}}$ shares of $[\text{Acme Corp.}]_{\text{ORGANIZACIJA}}$ in $[2006]_{\text{VRIJEME}}$.

Opis problema i motivacija

Opis problema i motivacija

Prepoznavanje imenovanih entiteta je zadatak klasifikacije elemenata u predefinirane kategorije kao što su:

- Imena – Osobe, Organizacije, Lokacije
- Vremenske oznake – Vrijeme, Datum
- Brojevi – Novac, Postotci
- ...

[Jim]$_{\text{OSOBA}}$ bought [300]$_{\text{BROJ}}$ shares of [Acme Corp.]$_{\text{ORGANIZACIJA}}$ in [2006]$_{\text{VRIJEME}}$.

Prepoznavanje imenovanih entiteta je zadatak ekstrakcije informacija kojem je cilj klasificirati u predefinirane kategorije kao što su: • Imena – Osobe, Organizacije, Lokacije • Vremena – Vrijeme, Datum • Brojevi – Novac, Postotci

Ovisno o domeni za koju se koriste imenovani entiteti, moguće ih je prizvoljno definirati.

U ovom primjeru entitet OSOBA sadrži jedan token dok entitet ORGANIZACIJA sadrži dva tokena.

Opis problema i motivacija

Kategorije eniteta u ovom radu:

- PER – Osobe
- ORG – Organizacije
- LOC – Lokacije
- MISC – Razno

Analiza podataka

- CoNLL – Conference on Computational Natural Language Learning
- izvor podataka: CoNLL02 i CoNLL03
- španjolski i nizozemski označeni u BIO format.
- engleski prebačen iz IO u BIO format.

- CoNLL – Conference on Computational Natural Language Learning
- izvor podataka: CoNLL02 i CoNLL03
- španjolski i nizozemski označeni u BIO format.
- engleski prebačen iz IO u BIO format.

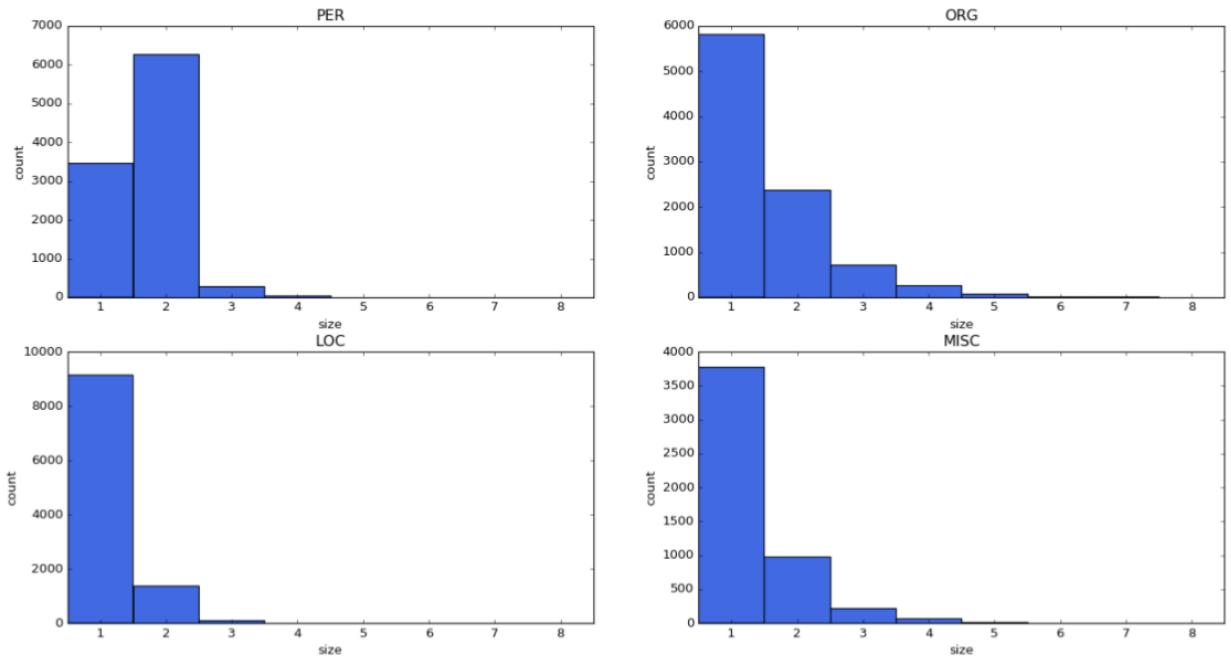
Format u kojem se s B označavaju riječi na početku entiteta, I označavaju riječi unutar entiteta i O označavaju riječi koje ne pripadaju ni jednom entitetu.

Format u kojem se s označavaju riječi koje pripadaju nekom entitetu i O označavaju riječi koje ne pripadaju ni jednom entitetu.

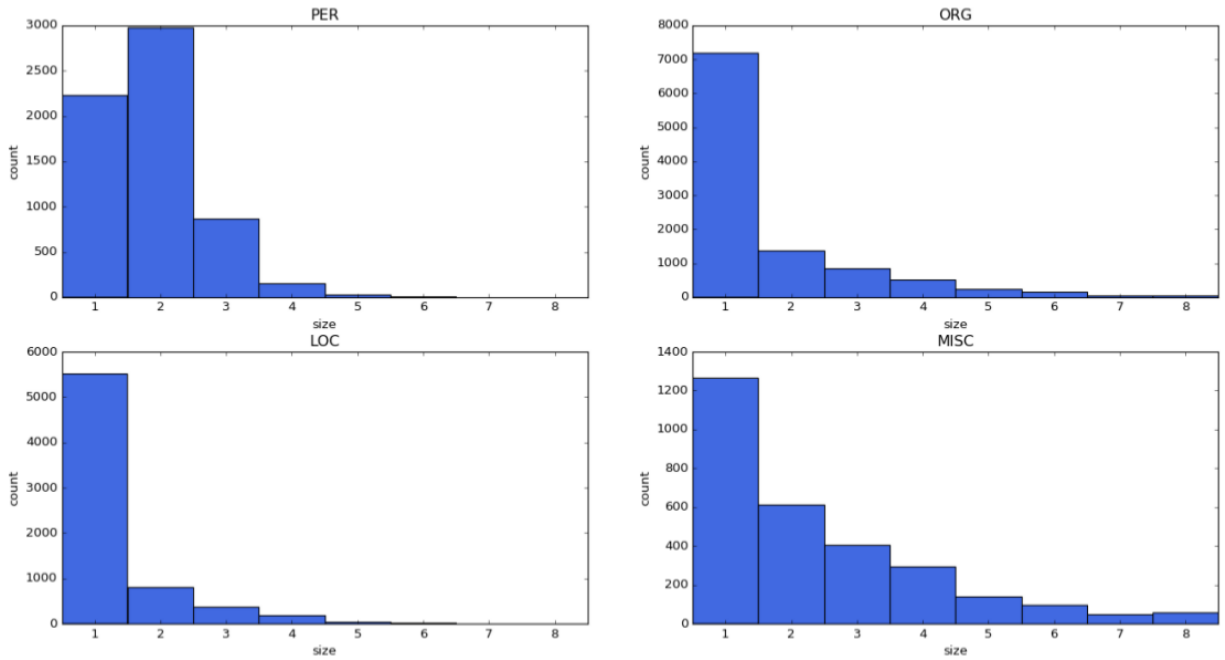
Broj entiteta u skupovima

jezik	skup	PER	ORG	LOC	MISC
eng	train	6600	6321	7140	3438
	validation	1842	1341	1837	922
	test	1617	1661	1668	702
esp	train	4321	7390	4913	2173
	validation	1222	1700	984	445
	test	735	1400	1084	339
ned	train	4716	2082	3208	3338
	validation	703	686	479	748
	test	1098	882	774	1187

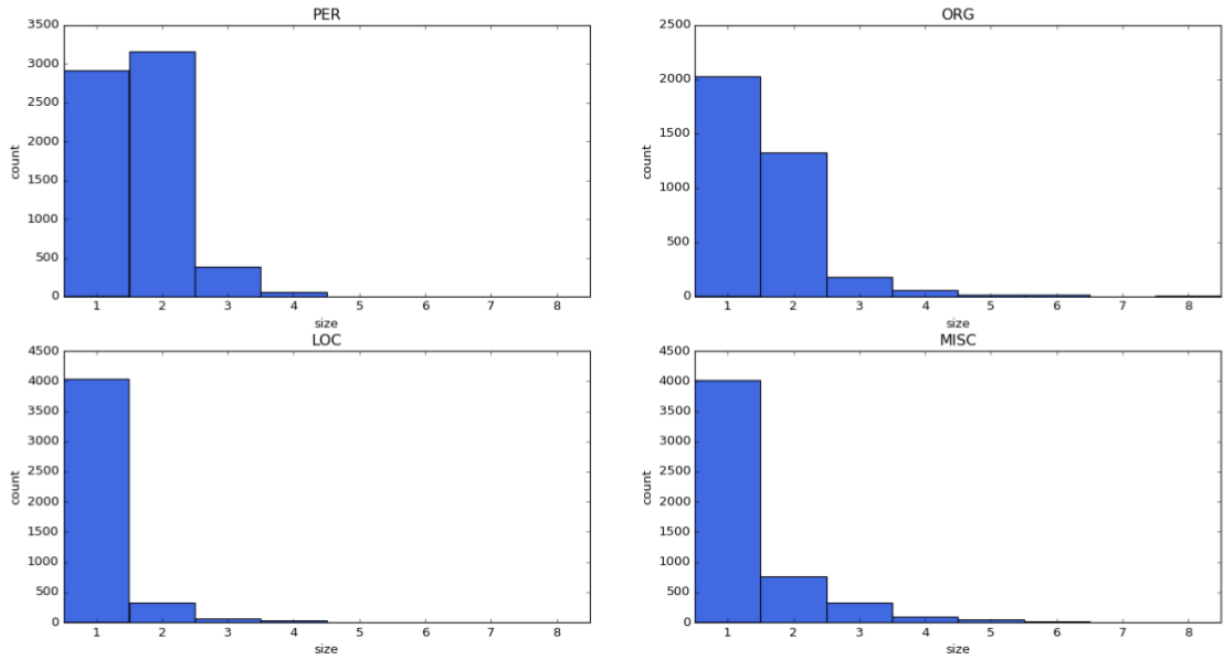
Veličina entiteta skupa podataka na engleskom jeziku



Veličina entiteta skupa podataka na španjolskom jeziku



Veličina entiteta skupa podataka na nizozemskom jeziku



Modeli

- Modeli iz scikit-learn knjižnice:
- Perceptron
- Logistička regresija
- modeli trenirani na 200 iteracija
- balansirane klase pri treniranju

parametar je postavljen na "balanced" što znači da model klase koje se rijetko pojavljuju unutar skupa za treniranje kažnjava više za grešku u predikciji jer sve klase smatra jednakima. Modeli su trenirani na 200 iteracija.

Skupovi značajki

1. Osnovni model (engl. baseline)
2. Osnovni model + Gazeteri
3. Osnovni model + Gazeteri + Wikifikacija

Table: Broj gazetera za svaki entitet

	PER	LOC	ORG	MISC
Gazeteri	2 972 k	3 106 k	977 k	2 991 k

- Osnovni model (engl. baseline)
- Osnovni model + Gazeteri
- Osnovni model + Gazeteri + Wikifikacija

Table: Broj gazetera za svaki entitet

	PER	LOC	ORG	MISC
Gazeteri	2 972 k	3 106 k	977 k	2 991 k

Gazeteri su unaprijed prikupljeni skupovi entiteta. **Wikifikacija** je proces prepoznavanja entiteta u tekstu te povezivanja istih s nasličnijim stranicama na wikipediji.

Engleska wikifikacija

Barack Hussein Obama is an American politician who served as the 44th President of the United States from 2009 to 2017. He is the first African American to have served as president. He previously served in the U.S. Senate representing Illinois from 2005 to 2008, and in the Illinois State Senate from 1997 to 2004.

Wikify! Clear

The Wikification system has identified the following entities with Wikipedia articles. Click on an entity to visit the corresponding Wikipedia page. Hover over links to view the categories associated with each entity.

Barack Hussein Obama is an American politician who served as the 44th President of the United States from 2009 to 2017. He is the first African American to have served as president. He previously served in the U.S. Senate representing Illinois from 2005 to 2008, and in the Illinois State Senate from 1997 to 2004.

Španjolska wikifikacija

Barack Hussein Obama en inglés americano; Honolulu, Hawái, 4 de agosto de 1961) es un político estadounidense que fue el 44º presidente de los Estados Unidos de América desde el 20 de enero de 2009 hasta el 20 de enero de 2017.4 [actualizar] Fue senador por el estado de Illinois desde el 3 de enero de 2005 hasta su renuncia el 16 de noviembre de 2008.5 Además, es el quinto legislador afroamericano en el Senado de los Estados Unidos, tercero desde la era de reconstrucción. También fue el primer candidato afroamericano nominado a la presidencia por el Partido Demócrata y es el primero en ejercer el cargo presidencial.6a of Tranquility.

Wikify! Clear

The Wikification system has identified the following entities with Wikipedia articles. Click on an entity to visit the corresponding Wikipedia page. Hover over links to view the categories associated with each entity.

Barack Hussein Obama en inglés americano; Honolulu, Hawái, 4 de agosto de 1961) es un político estadounidense que fue el 44º presidente de los Estados Unidos de América desde el 20 de enero de 2009 hasta el 20 de enero de 2017.4 [actualizar] Fue senador por el estado de Illinois desde el 3 de enero de 2005 hasta su renuncia el 16 de noviembre de 2008.5 Además, es el quinto legislador afroamericano en el Senado de los Estados Unidos, tercero desde la era de reconstrucción. También fue el primer candidato afroamericano nominado a la presidencia por el Partido Demócrata y es el primero en ejercer el cargo presidencial.6a of Tranquility.

Nizozemska wikifikacija

Značajke

Osnovne značajke	
prethodni tag entiteta	(t_{i-1}, t_{i-2})
sadrži samo brojke i slova	$alphanumeric(w_i)$
sadrži samo brojke	$alldigits(w_i)$
sadrži samo velika slova	$allcaps(w_i)$
sadrži samo brojke	$iscap(w_{i-2}, w_{i-1}, w_i, w_{i+1}, w_{i+2})$
3-gram	suma pojavljivanja za svaku klasu
Gazeteri	
naziv kategorije gazetera	$topic(w_i, w_{i+1}, w_{i+2}, w_{i+3})$
broj pojavljivanja riječi u kategoriji	$cat\_count(w_i, \{PER, LOC, ORG\})$
Medujezične značajke	
—	—

Osnovne značajke	
prethodni tag entiteta	(t_{i-1}, t_{i-2})
sadrži samo brojke i slova	$alphanumeric(w_i)$
sadrži samo brojke	$alldigits(w_i)$
sadrži samo velika slova	$allcaps(w_i)$
sadrži samo brojke	$iscap(w_{i-2}, w_{i-1}, w_i, w_{i+1}, w_{i+2})$
3-gram	suma pojavljivanja za svaku klasu
Gazeteri	
naziv kategorije gazetera	$topic(w_i, w_{i+1}, w_{i+2}, w_{i+3})$
broj pojavljivanja riječi u kategoriji	$cat\_count(w_i, \{PER, LOC, ORG\})$
Medujezične značajke	
—	—

Za potrebe ovog modela prikupljeni gazeteri su podijeljeni u teme čiji su naslovi korišteni kao značajke modela. Neke od tema su: ArtWork, Building, Clothes, Films, Parks, Vehicles itd. Dodatno su prikupljeni skupovi za entitete Osoba, Organizacija i Lokacija te su za značajke korišteni kao broj pojavljivanja riječi u pojedinom skupu.

Pretprocesiranje značajki

- kategoričke značajke – Onehot coding
- brojčane značajke – MinMaxScaler(0, 1)

Većina korištenih značajki su kategoričke stoga su kodirane Onehot metodom.
Brojčane značajke kojima je definiran poredak skalirane su na interval $[0, 1]$. Za skaliranje je korišten MinMaxScaler

Unakrsna validacija

Parametri perceptrona:

$$alpha = (10^{-10}, 10^{-9}, ..., 10^{-2})$$

$$penalty = (l2, l1)$$

Parametri logističke regresije:

$$c = (10^{-7}, 10^{-9}, ..., 10^{2})$$

$$penalty = (l2, l1)$$

Evaluacija

Sustav je evaluiran na dva način:

1. Standardne mjere na razini svakog tokena.
2. Točno podudaranje gdje se entitet smatra dobro predvidenim ako se svaki token podudara po tipu s označenim podatcima.

Table: Veličina skupova podataka

	ENG	ESP	NED
Entiteti (treniranje)	29 441	23 148	15 960
Entiteti (testiranje)	5 648	3 558	3 941

Evaluacija

Evaluacija

Sustav je evaluiran na dva način:

- Standardne mjere na razini svakog tokena.
- Točno podudaranje gdje se entitet smatra dobro predvidenim ako se svaki token podudara po tipu s označenim podatcima.

Table: Veličina skupova podataka

	ENG	ESP	NED
Entiteti (treniranje)	29 441	23 148	15 960
Entiteti (testiranje)	5 648	3 558	3 941

Standardne evaluacijske mjere

	Perceptron				Log. regresija			
	ENG	ESP	NED	AVG	ENG	ESP	NED	AVG
Jednojezični eksperimenti								
Osnovne značajke	0.66	0.64	0.55	0.62	0.62	0.62	0.59	0.61
+Gazeteri	0.67	0.64	0.58	0.63	0.70	0.64	0.68	0.67
+Wikifikacija	–	–	–	–	–	–	–	–
Medujezični eksperimenti								
Osnovne značajke	–	0.59	0.56	0.575	–	0.63	0.66	0.645
+Gazeteri	–	0.61	0.54	0.575	–	0.64	0.63	0.635
+Wikifikacija	–	–	–	–	–	–	–	–

Evaluacija metodom točnog podudaranja

	Perceptron				Log. regresija			
	ENG	ESP	NED	AVG	ENG	ESP	NED	AVG
Jednojezični eksperimenti								
Osnovne značajke	0.46	0.43	0.38	0.42	0.73	0.75	0.67	0.72
+Gazeteri	0.51	0.44	0.43	0.46	0.78	0.77	0.72	0.76
+Wikifikacija	–	–	–	–	–	–	–	–
Medujezični eksperimenti								
Osnovne značajke	–	0.29	0.31	0.30	–	0.46	0.47	0.465
+Gazeteri	–	0.33	0.28	0.305	–	0.44	0.41	0.425
+Wikifikacija	–	–	–	–	–	–	–	–

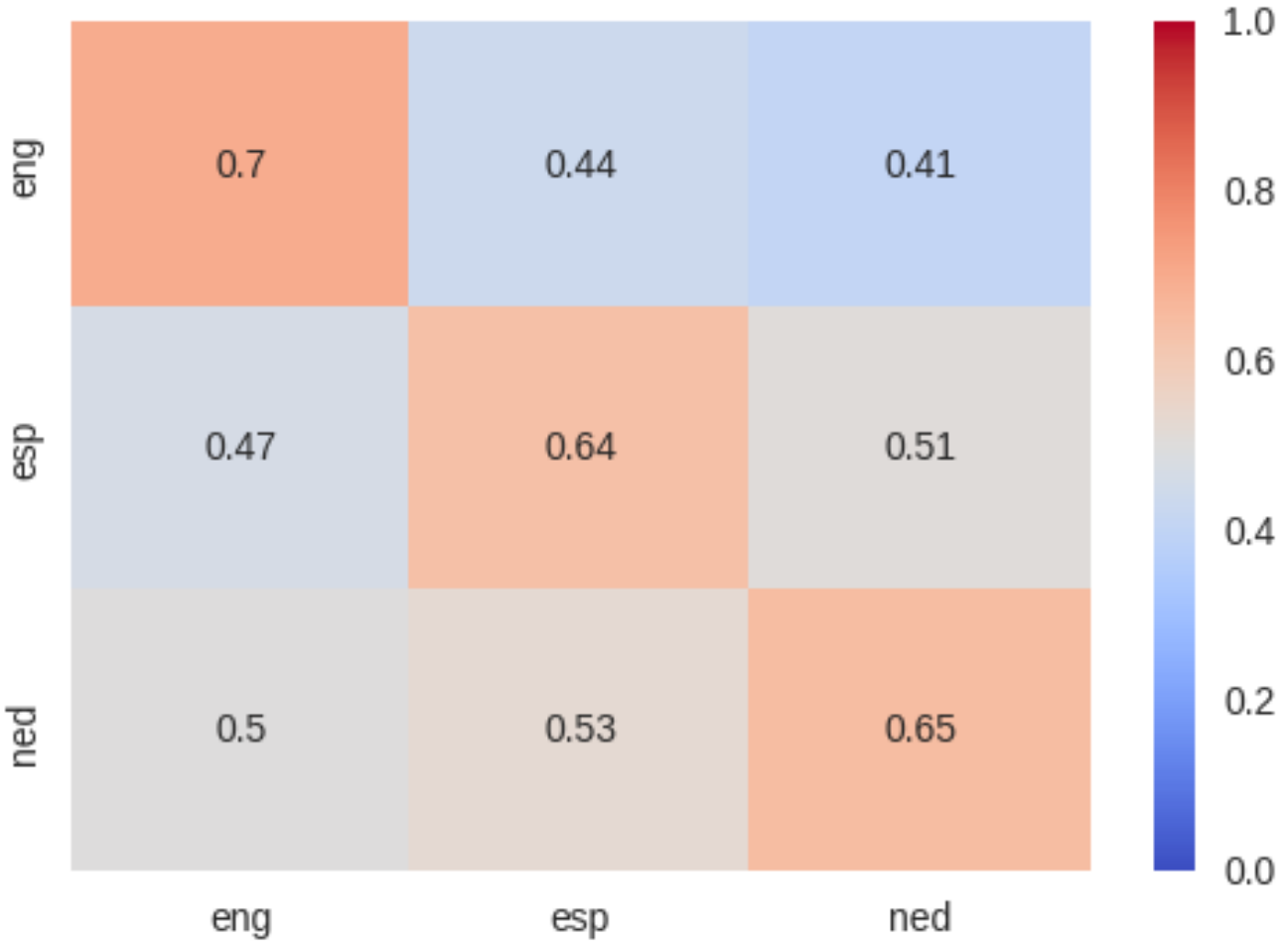


Figure: Evaluacijska matrica modela s različitim jezicima za treniranje i testiranje. Korištena je f1 mjera. Evaluacija je izvršena na modela koji uključuje baseline i gazetere.

Poboljšanja

- bolji odabir tema za gazetere
- dodavanje novih značajki
- podešavanje wikifiera

- bolji odabir tema za gazetere
- dodavanje novih značajki
- podešavanje wikifiera

Poboljšanja se kriju u boljoj kvaliteti podataka i otkrivanju nekih bolji značajki. Konkretno za razvijeni model u ovom radu pri odluci koji tema gazettera će biti dodjeljena trenutno promatranoj riječi dobije se pronalaskom te riječi u skupu teme. Poboljšanje možemo ostvariti presjekom tema s okolnim riječima jer ne pripada neka riječ samo jednoj temi. Uključivanje word embeddinga kao značajke za svaku riječ bi moglo re- zultirati poboljšanjem. Metoda Wikifikacije se može poboljšati boljom distribucijom kategorija i obogaćivanjem wikipedije za jezike s manjim resursima.

Hvala na pažnji!

Pitanja?