SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

SEMINAR

Međujezično prepoznavanje imenovanih entiteta pomoću wikifikacije

Stipan Mikulić
Voditelj: *dr. sc. Jan Šnajder*

Zagreb, svibanj 2017.

SADRŽAJ

1. Uvod

U današnje vrijeme svjedočimo stalni eksponencijali porast svih vrsta podataka, naročito teksta. Zbog tog naglog porasta podataka ljudi više nisu u mogućnosti obraditi te podatke da bi prepoznali bitne i korisne informacije. Riješenje problema krije se u računalnoj obradi podataka. Ovim problemom se bavi područje računarske znanosti (engl. computer science), umjetne inteligencije (engl. artificial intelligence) i strojnog učenja (engl. machine learning) koje se naziva obrada prirodnog jezika (engl. natural language processing, NLP).

U ovom radu razvit će se model za međujezično prepoznavanje imenovanih entiteta. 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.[1]U potpunoj dominaciji engleskog jezika u svim vrstama podataka i NLP alata i leži motivacija za razvoj međujezičnog modela.

Rad je strukturiran tako da u drugom poglavlju opisuje problem, u trećem analizira podatke nad kojima treniramo, validiramo i testiramo model. Četvrto poglavlje opisat će sam model za prepoznavanje imenovanih entiteta, a peto implementciju tog modela. Rezultati i evaluacija će biti opisani u šestom poglavlju. Zadnjem poglavlje će dati kratki zaključak rada.

[1] `https://en.wikipedia.org/wiki/Languages_used_on_the_Internet`

2. Opis problema

Prepoznavanje imenovanih entiteta je zadatak ekstrakcije informacija kojem je cilj klasificirati i locirati elemente u predefinirane kategorije kao što su:

- Imena – Osobe, Organizacije, Lokacije
- Vremena – Vrijeme, Datum
- Brojevi – Novac, Postotci

Iako su kategorije unaprijed definirane i dalje se postavlja pitanje koliko općenite i obuhvatne trebaju biti. Ovisno o domeni za koju se koriste imenovani entiteti, moguće ih je prizvoljno definirati. Pogledajmo pobliže ovaj problem kroz primjer. Ako sustavu za prepoznavanje imenovanih entiteta damo sljedeći tekst kao ulaz:

Jim bought 300 shares of Acme Corp. in 2006.

na izlazu statava ćemo dobiti:

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

U ovom primjeru entitet OSOBA sadrži jedan token dok entitet ORGANIZACIJA sadrži dva tokena.[1] U ovom radu želimo prepoznati sljedeće entitete u tekstu:

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

[1] `https://en.wikipedia.org/wiki/Named-entity_recognition`

Najčešći pristup ovom problemu je uz pomoć metoda nadziranog strojnog učenja. Ovaj problem spada u kategoriju označavanja slijedova (engl. sequence labeling) gdje se svakom članu slijeda pridružuje neka oznaka tj. predefinirana kategorija. Oznake su ovisne o svim članovima oko njih u slijedu. Zbog toga se ovisnost izražava s lijeva na desno, s desna na lijevo ili zajednički. Radi boljeg razumjevanja problema u idućem poglavlju će se pisati o analizi podataka.

3. Analiza podataka

Model za međujezično prepoznavanje imenovanih entiteta razvijen je nad skupovima podataka iz CoNLL02 i CoNLL03 dijeljenog zadatka. Skup podataka uključuje podatke na engleskom, španjolskom i nizozemskom jeziku. CoNLL skup podataka je podskup novinskih članaka Reutersa iz 1996. Entiteti su označeni u 4 razreda: PER, ORG, LOC i MISC. Skup za treniranje su članci iz kolovoza 1996, dok je testni skup iz prosinca 1996. Imenovani entiteti u testnom skupu su znatno različiti od skupa za treniranje što ih čini značajno težim.(Ratinov i Roth, 2009)

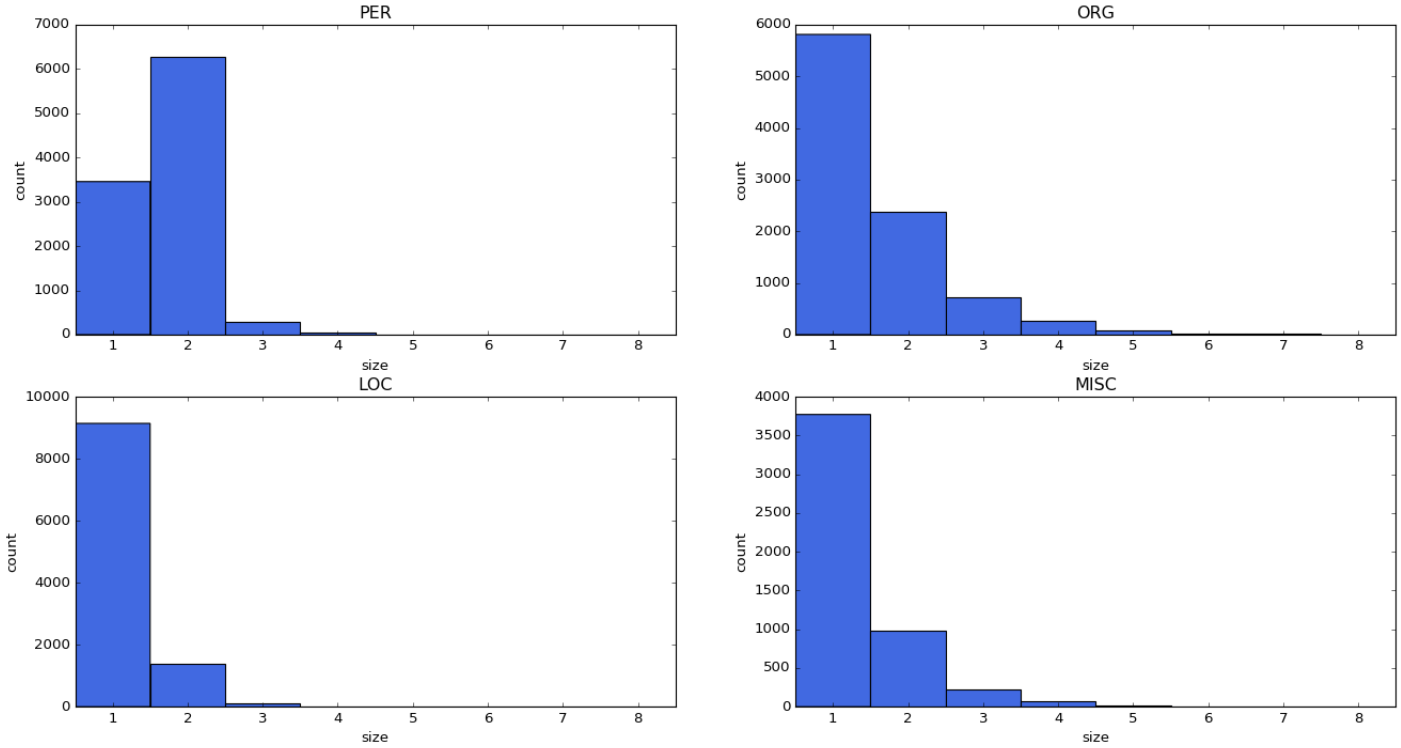
Španjolski i nizozemski skup podataka označen je BIO[1] formatom dok je engleski skup podataka označen IO formatom[2] koji je naknadno pretvoren u BIO.

Tablica 3.1: Broj entiteta u skupovima

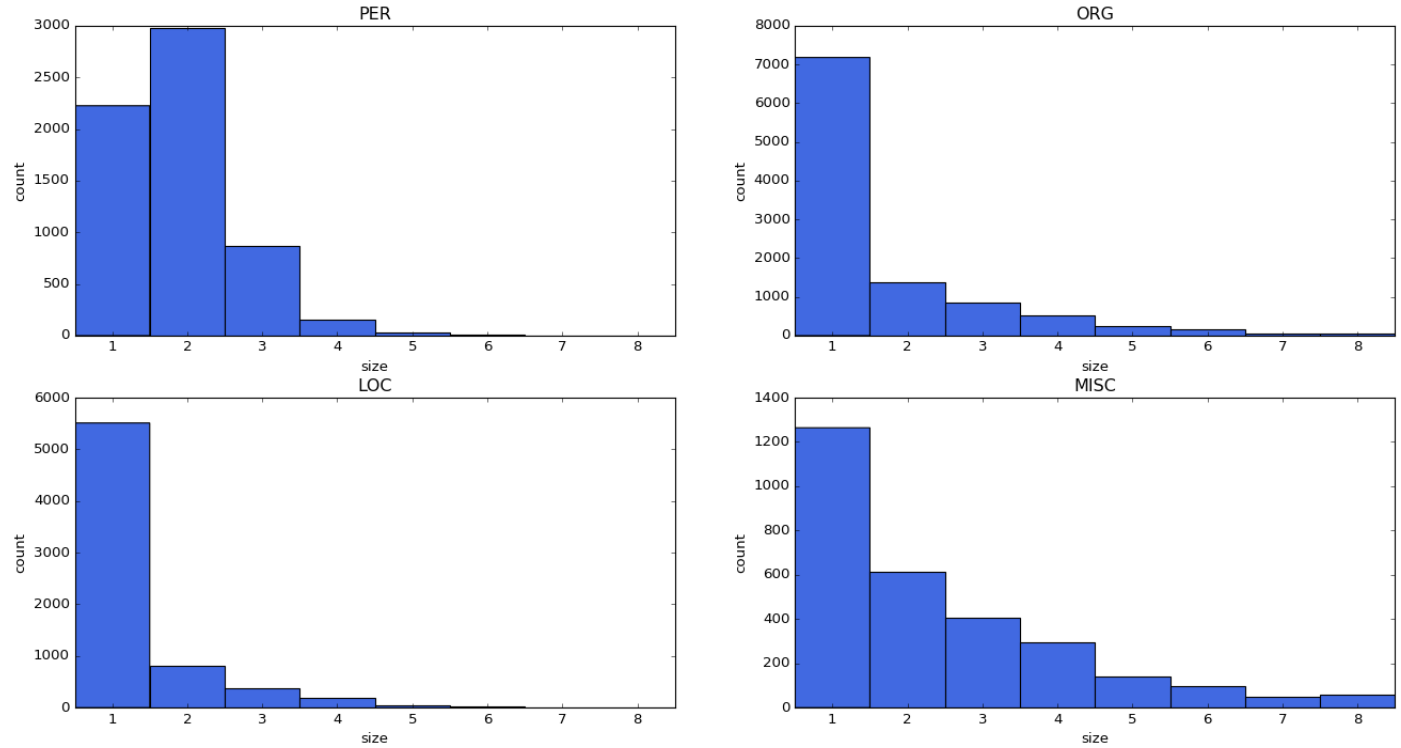
language	set	PER	ORG	LOC	MISC
eng	train	6600	6321	7140	3438
	validation	1842	1341	1837	922
	test	1617	1661	1668	702
esp	train	4321	7390	4913	2173
	validation	1222	1700	984	445
	test	735	1400	1084	339
ned	train	4716	2082	3208	3338
	validation	703	686	479	748
	test	1098	882	774	1187

[1]Format u kojem se s B (**B**egining) označavaju riječi na početku entiteta, I (**I**nside) označavaju riječi unutar entiteta i O (**O**utside) označavaju riječi koje ne pripadaju ni jednom entitetu.

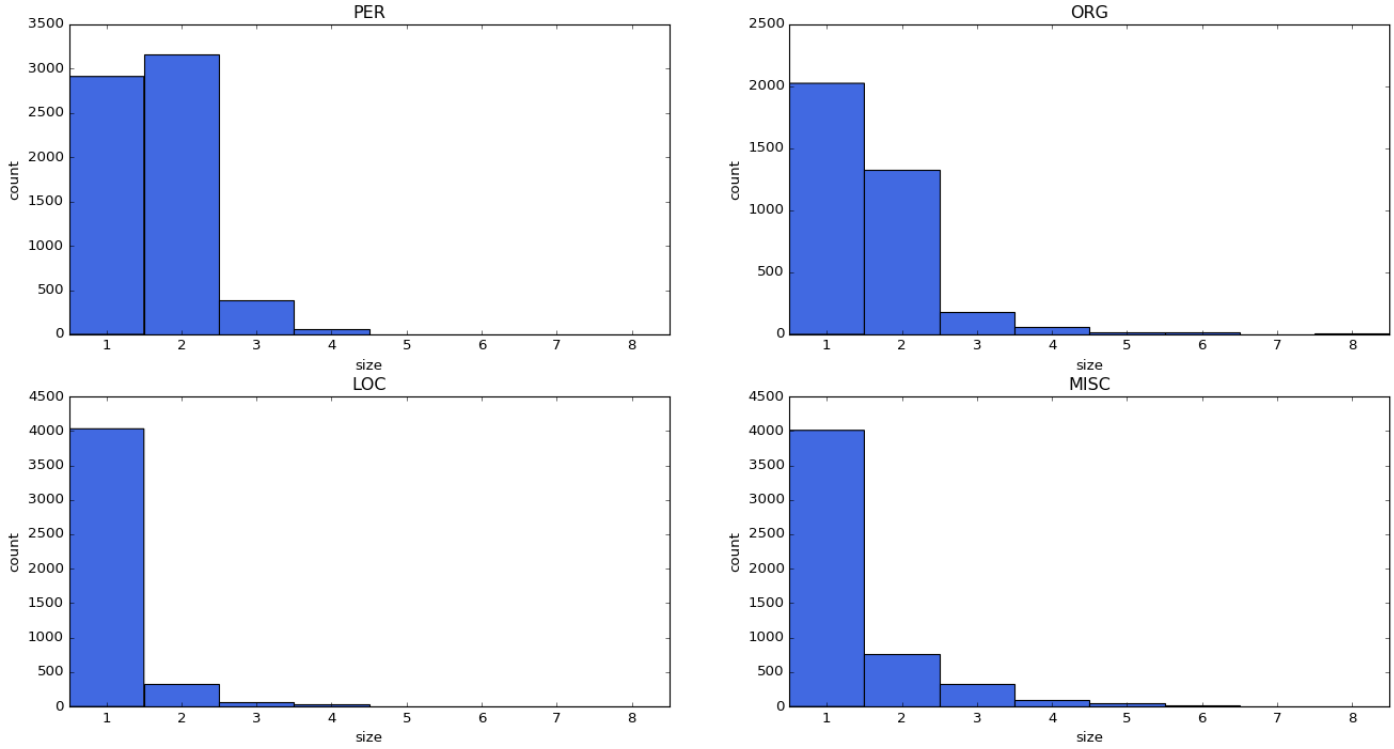
[2]Format u kojem se s I (**I**nside) označavaju riječi koje pripadaju nekom entitetu i O (**O**utside) označavaju riječi koje ne pripadaju ni jednom entitetu.



Slika 3.1: Veličina entiteta skupa podataka na engleskom jeziku



Slika 3.2: Veličina entiteta skupa podataka na španjolskom jeziku



Slika 3.3: Veličina entiteta skupa podataka na nizozemskom jeziku

4. Model

4.1. Perceptron

U svrhu prepoznavanja imenovanih entiteta u tekstu najčešće se koriste sljedeći modeli:

- HMM (Hidden Markov Model)
- CRF (Conditional random field)
- MaxEnt
- Perceptron

U ovom radu korišten je Perceptron. Iako prva dva navedena modela daju mogućnost zajedničkog učenja oznaka u slijedu, ipak je korišten Perceptron koji nam omogućuje korištenje raznih značajki.

Perceptron je matematički model neurona. U stvarnom neuronu dendriti dobivaju signale od aksona drugih neurona dok su u matematičkom modelu dendriti predstavljeni s brojčanim vrijednostima. Izlaz neurona, akson, predstavljen je aktivacijskom funkcijom koja kao argument prima težinsku sumu ulaza neurona zbrojenom s pomakom.[1]. Najčešće korištena aktivacijska funkcija je $sigmoid$[2].

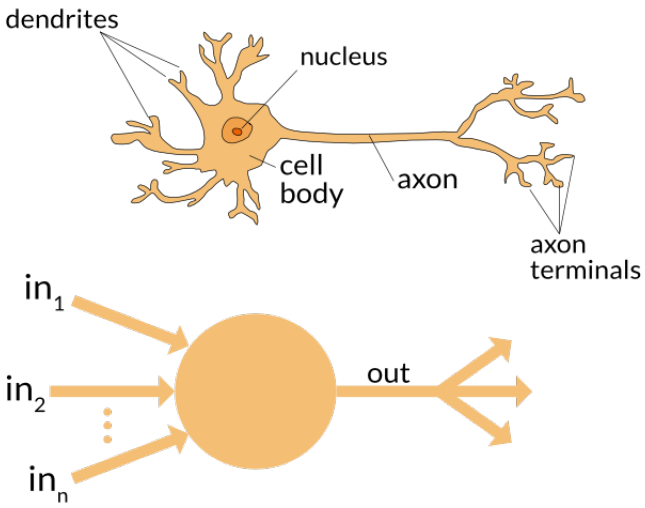
$$suma = \sum_{i=1}^{n} w_i x_i + b$$

$$izlaz = sigmoid(suma)$$

Perceptron pokušava naučiti težine na način da ih ne ažurira nakon prolaska kroz cijeli skup podataka već nakon svakog primjera. Ovakav način učenja se zove online učenje.

[1] https://cs.stanford.edu/people/eroberts/courses/soco/projects/neural-networks/Neuron/index.html

[2] https://en.wikipedia.org/wiki/Sigmoid_function



Slika 4.1: Matematički model biološkog neurona

[3]

4.2. Logistička regresija

[3] https://appliedgo.net/perceptron/

5. Implementacija modela

cross validacija, baseline, obicne znacajke, gazetteri-grupe, wikifikacija, 9-class clasification, klase Sustav za prepoznavanje imenovanih entiteta je razvijen u programskom jeziku Python3. Korištena je implementacija Perceptrona iz scikit-learn[1] knjižnice. Glavni program runner.py pokreće cijeli cjevovod sustava. Moguće je specificirati jezike skupa za treniranje, validaciju i testiranje preko argumenata komandne linije. Naredba za pokretanje programa koji trenira i validira sustav engleskim i nizozemskim jezikom a testira nad španjolskim jezikom.

python3 runner.py -train eng,ned -validation eng,ned -test esp

Nakon parsiranja argumenata dohvaćaju se navedeni skupovi za treniranje, validiranje i testiranje. Nakon toga se izvlače značajke iz podataka te se pretprocesiraju u brojčani oblik pogodan za algoritam. Prilikom odabira najboljeg modela radi se unakrsna validacija nad skupovima za treniranje i validiranje.

Razvijena su tri modela koji se razlikuju u značajkama:

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

Gazeteri su unaprijed prikupljeni skupovi entiteta. 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. Za dohvaćanja teme za neku riječ korišten je pomični prozor veličine 4. Ovisno o poziciji na kojoj se nalazi riječ u prozoru dodaje se prefiks B- ili I- temi kojoj pripada. Ako neka riječ ima više tema kojima pripada biramo prvu nađenu.(Tsai et al., 2016)

[1] `http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Perceptron.html`

opisati wikifikaciju

Tablica 5.1: Statistika gazetera i wiki-a

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

5.1. Značajke

U tablici je naveden popis značajki podjeljen prema modelima u kojima su korištene.

Tablica 5.2: Značajke sustava

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 veika slova	$allcaps(w_i)$
sadrži samo brojke	$iscapitalized(w_{i-2}, w_{i-1}, w_i, w_{i+1}, w_{i+2})$
3-gram	suma pojavljivanja 3-grama za svaku klasu
Gazeteri	
naziv kategorije gazetera	$topic(w_i, w_{i+1}, w_{i+2}, w_{i+3})$
Međujezične značajke	

detaljnije opisati znacajke, dodati trigram znacajke

5.2. Pretprocesiranje značajki

Većina korištenih značajki su kategoričke stoga su kodirane Onehot [2] metodom. Brojčane značajke kojima je definiran poredak skalirane su na interval $[0, 1]$. Za skaliranje je korišten MinMaxScaler[3].

[2] Svaka kategorija neke značajke se kodira u vektor duljine $broj\_kategorija$ tako da je jedan element vektora 1 a ostali 0. `http://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html`

[3] `http://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html#sklearn.preprocessing.MinMaxScaler`

5.3. Unakrsna validacija

Unaksrsna 5-struka validacija je korištena radi dobivanja najboljeg mogućeg modela za dane podatke. Parametri perceptrona koji su optimizirani unakrsnom validacijom:

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

$$penalty = (l2, l1)$$

oba parametra se koriste u svrhu regularizacije, alpha je konstanta koja množi regularizacijski faktor a penalty je regularizacijska funkcija.

6. Evaluacija

evaluacijske mjere, tablice s rezultatima, exact evaluacija, corelation matrix za evaluaciju

Evaluacija sustava rađena je nad testnim skupom podataka. Model koji evaluiramo je najbolji model dobiven unakrsnom validacijom. Mjere koje su korištene za evaluaciju su f1-score, precision, recall, micro and macro accuracy.

Tablica 6.1: Evaluacijske mjere

	Perceptron				Log. regresija			
	ENG	ESP	NED	AVG	ENG	ESP	NED	AVG
Entiteti (treniranje)	–	–	–	–	–	–	–	–
Entiteti (testiranje)	–	–	–	–	–	–	–	–
Jednojezični eksperimenti								
Osnovne značajke	–	–	–	–	–	–	–	–
+Gazeteri	–	–	–	–	–	–	–	–
+Wikifikacija	–	–	–	–	–	–	–	–
Međujezični eksperimenti								
Osnovne značajke	–	–	–	–	–	–	–	–
+Gazeteri	–	–	–	–	–	–	–	–
+Wikifikacija	–	–	–	–	–	–	–	–

6.1. Poboljšanja

presjek tema za gazettere

7. Zaključak

zakljuciti

8. Literatura

Lev Ratinov i Dan Roth. Design challenges and misconceptions in named entity recognition. U *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, CoNLL '09, stranice 147–155, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-29-9. URL http://dl.acm.org/citation.cfm?id=1596374.1596399.

Erik F. Tjong Kim Sang. Introduction to the conll-2002 shared task: Language-independent named entity recognition. U *Proceedings of CoNLL-2002*, stranice 155–158. Taipei, Taiwan, 2002.

Erik F. Tjong Kim Sang i Fien De Meulder. Introduction to the conll-2003 shared task: Language-independent named entity recognition. U Walter Daelemans i Miles Osborne, urednici, *Proceedings of CoNLL-2003*, stranice 142–147. Edmonton, Canada, 2003.

Chen-Tse Tsai i Dan Roth. Cross-lingual wikification using multilingual embeddings. U *NAACL*, 6 2016. URL http://cogcomp.cs.illinois.edu/papers/TsaiRo16b.pdf.

Chen-Tse Tsai, Stephen D. Mayhew, i Dan Roth. Cross-lingual named entity recognition via wikification. U *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, CoNLL 2016, Berlin, Germany, August 11-12, 2016*, stranice 219–228, 2016. URL http://aclweb.org/anthology/K/K16/K16-1022.pdf.

Međujezično prepoznavanje imenovanih entiteta pomoću wikifikacije

Sažetak

Zbog stalnog rasta svih vrsta podataka, naročito teksta ljudi više nisu u mogućnosti obraditi te podatke da bi prepoznali bitne i korisne informacije. Zbog toga posežemo za računalnom obradom podataka. U ovom radu razvijen je model za međujezično prepoznavanje imenovanih entiteta. 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. Motivacija za razvoj međujezičnog modela leži u potpunoj dominaciji engleskog jezika u svim vrstama podataka i NLP alata. Razvojem takvog modela jezici sa skromnim izvorima podataka bi napredovali ne samo u prepoznavanju imenovanih entiteta već u analizi teksta općenito.

Ključne riječi: Strojno učenje, Procesiranje prirodnog jezika, Prepoznavanje imenovanih entiteta, Perceptron

Cross-Lingual Named Entity Recognition via Wikification

Abstract

Because of the steady growth of all kinds of data, especially text, people are no longer able to process this data to recognize essential and useful information. That's why we reach for computer data processing. In this paper, a model for cross-lingual named entity recognition was developed. To develop a good model for classification we need a lot of data. According to the latest estimates, more than 50% of web content is written in English. Motivation for the development of an cross-lingual model lies in the overall dominance of English in all types of data and NLP tools. By developing such a model, languages with modest data sources would advance not only in the recognition of named entities, but in text analysis in general.

Keywords: Machine learning, Natural language processing, Named entity recognition, Perceptron