SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

SEMINAR

Međujezično prepoznavanje imenovanih entiteta pomoću wikifikacije

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

Zagreb, lipanj 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)

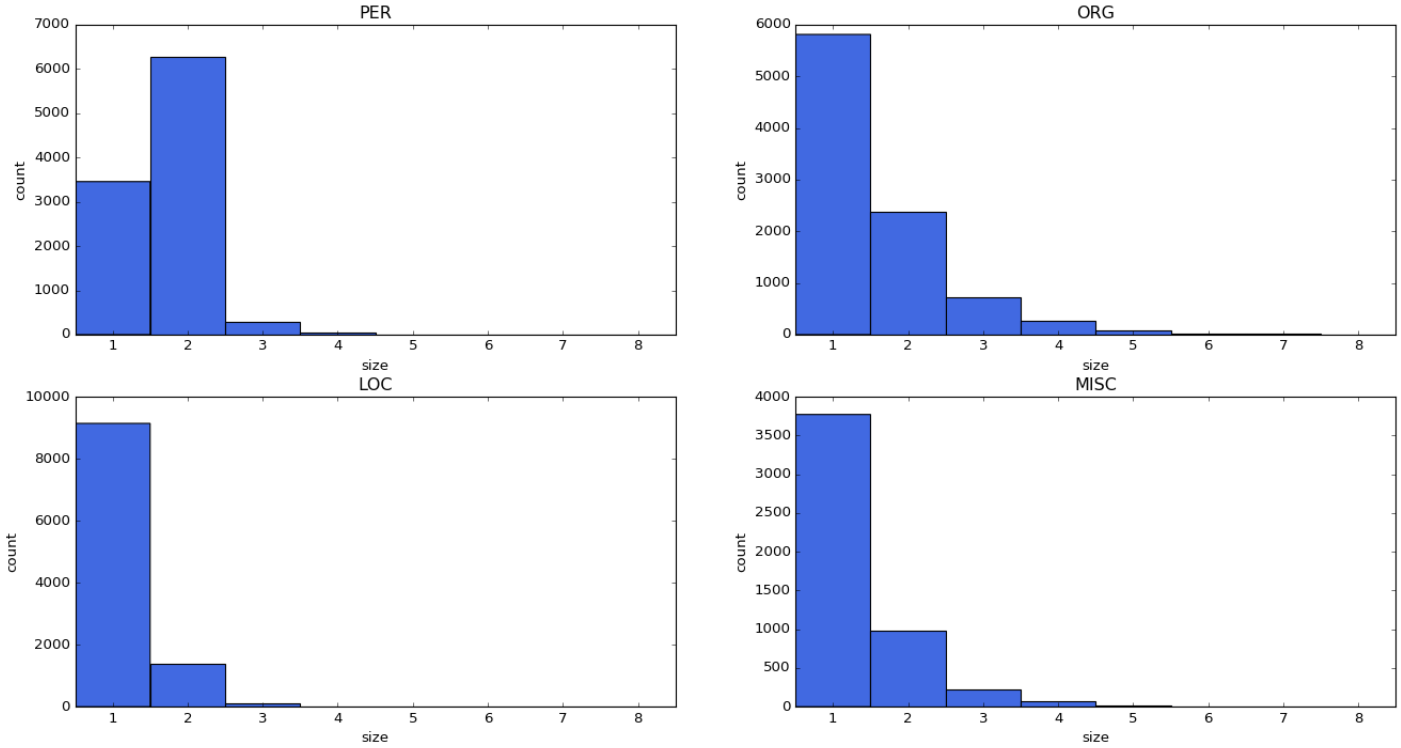
Španjolski i nizozemski skup podataka označen je BIO[1] formatom dok je engleski skup podataka označen IO formatom[2] koji je naknadno pretvoren u BIO.

Tablica 3.1: Broj entiteta u skupovima

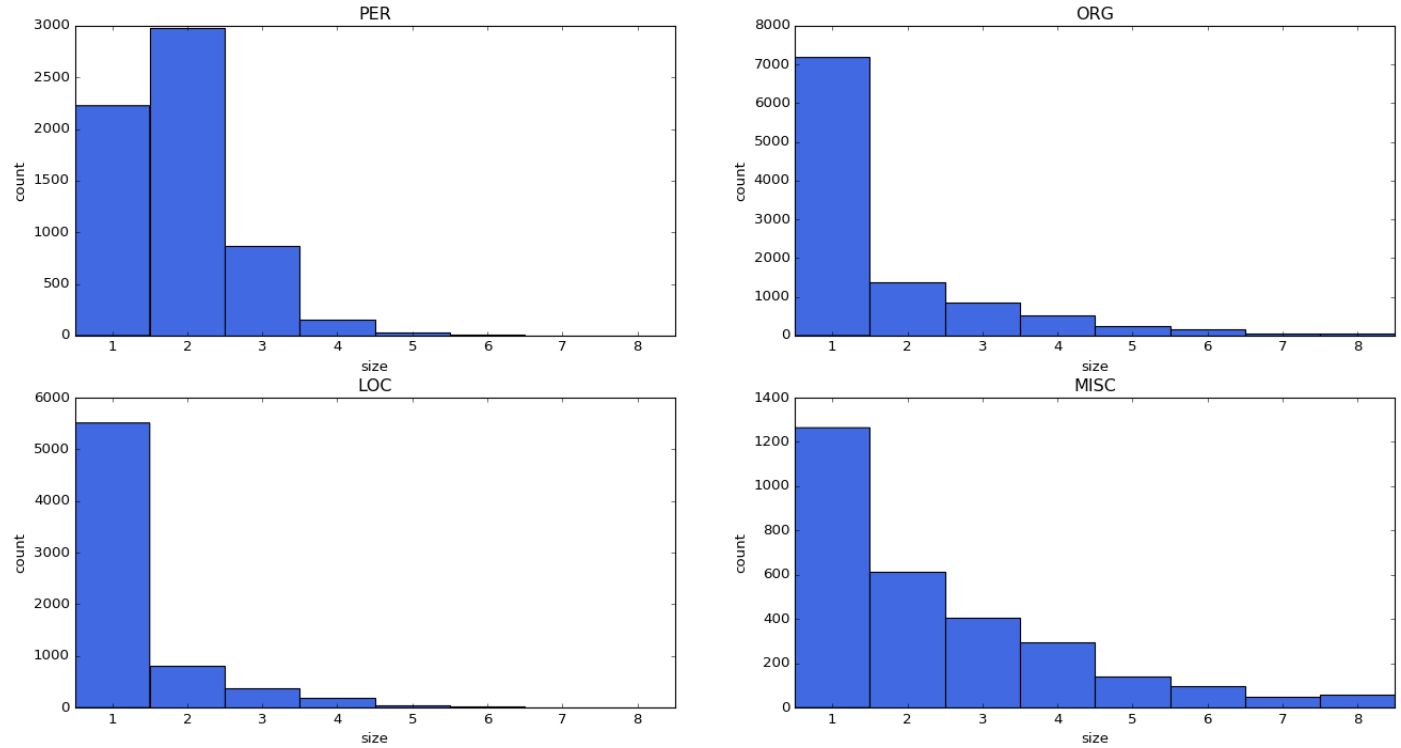
language	set	PER	ORG	LOC	MISC
eng	train	6600	6321	7140	3438
	validation	1842	1341	1837	922
	test	1617	1661	1668	702
esp	train	4321	7390	4913	2173
	validation	1222	1700	984	445
	test	735	1400	1084	339
ned	train	4716	2082	3208	3338
	validation	703	686	479	748
	test	1098	882	774	1187

[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.

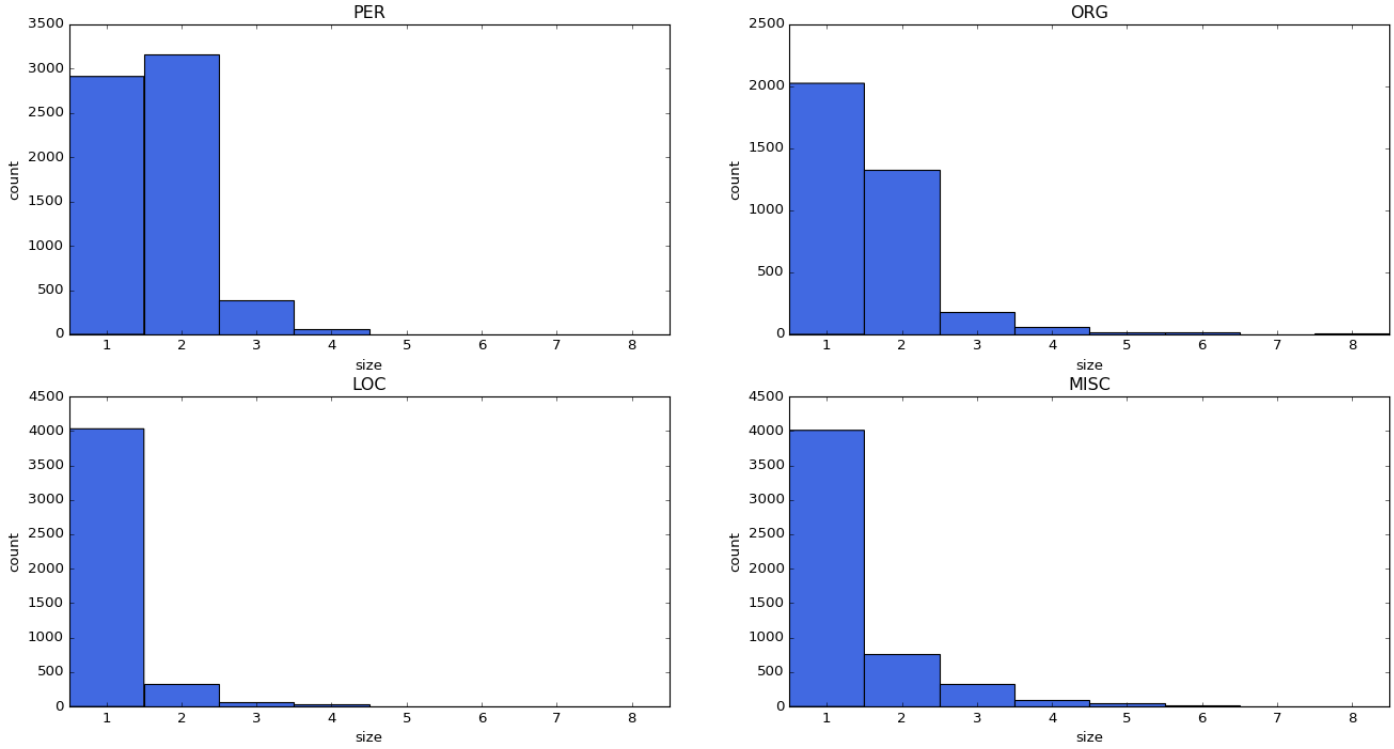
[2]Format u kojem se s I (**I**nside) označavaju riječi koje pripadaju nekom entitetu i O (**O**utside) označavaju riječi koje ne pripadaju ni jednom entitetu.



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

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

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

Iako prva dva navedena modela daju mogućnost zajedničkog učenja oznaka u slijedu, ipak su korišteni Perceptron i MaxEnt modeli koji nam omogućuje korištenje raznih značajki.

4.1. Perceptron

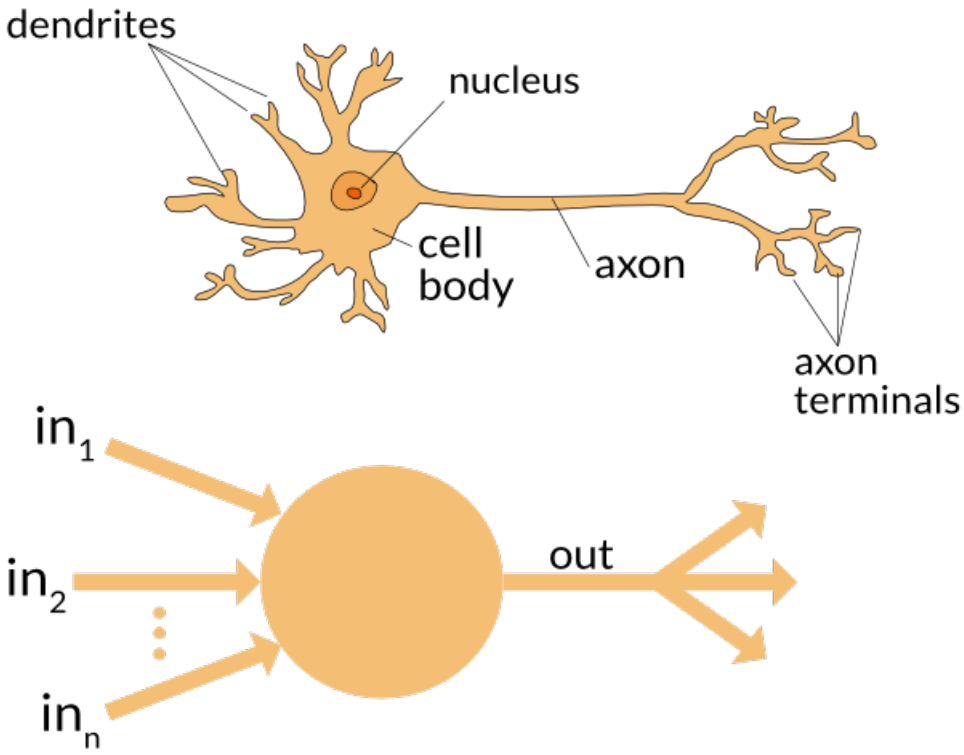
Perceptron je matematički model neurona. U stvarnom neuronu dendriti dobivaju signale od aksona drugih neurona dok su u matematičkom modelu dendriti predstavljeni s brojčanim vrijednostima. Izlaz neurona, akson, predstavljen je aktivacijskom funkcijom koja kao argument prima težinsku sumu ulaza neurona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

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

$$izlaz = activation\_function(suma)$$

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

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.



Slika 4.1: Matematički model biološkog neurona

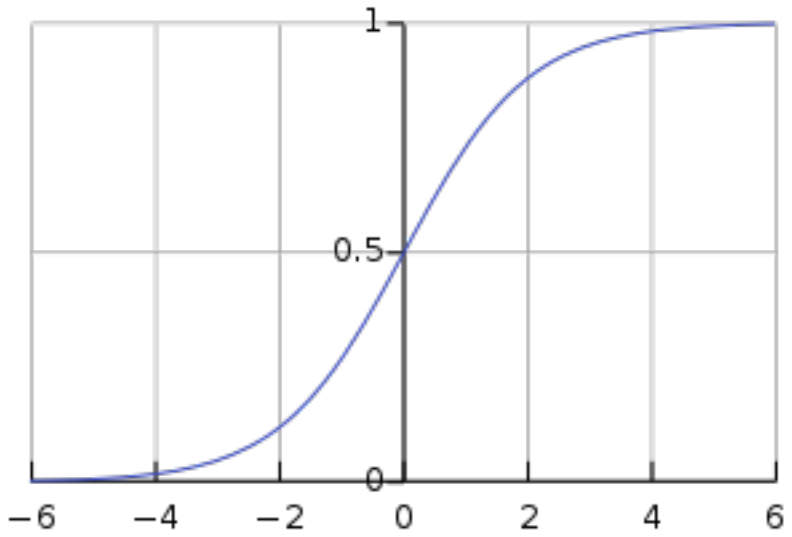
[2]

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

4.2. Logistička regresija

Logistička regresija je algoritam za binarnu klasifikaciju podataka. Treniranjem modela dobiva se vektor težina. Treniranje se obavlja korištenjem iterativnih algoritama kao sto je gradijentni spust. Za predikciju klase kojoj primjer pripada koristi se logistička ili sigmoid funkcija.[3] Kao argument te funkcije se dobiva skalarni produkt vektora značajki i vektora težina.[4]

$$h_\theta(x) = \frac{1}{1 + e^{-x}}$$



Slika 4.2: Logistička ili sigmoid funkcija
[5]

[3] https://en.wikipedia.org/wiki/Sigmoid_function
[4] http://ufldl.stanford.edu/tutorial/supervised/LogisticRegression/
[5] https://upload.wikimedia.org/wikipedia/commons/thumb/8/88/Logistic-curve.svg/320px-Logistic-curve.svg.png

5. Implementacija modela

Sustav za prepoznavanje imenovanih entiteta je razvijen u programskom jeziku Python3. Korištena je implementacija Perceptrona[1] i Logističke regresije[2] iz scikit-learn knjižnice. U oba modela class_weight 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. 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

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

[2] http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

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)
Implementirani model za svaki token predviđa jednu od 9 klasa (O, B-PER, I-PER, B-ORG, I-ORG, B-LOC, I-LOC, B-MISC, I-MISC). Oznake B i I predstavljaju pocetak ili unutarnji dio entiteta prema BIO kodiranju imenovanih entiteta.

Wikifikacija je proces prepoznavanja entiteta u tekstu te povezivanja istih s naslič-nijim stranicama na wikipediji. U referenciranom članku wikifikacija je iskorištena na način da se riječi i fraze iz tekstova koji nisu na engleskom jeziku mapiraju na stranice engleske wikipedije. Na taj način se dobivaju značajke koje nisu ovisne o jeziku na kojem je tekst pisan. Jedini uvjet je pristup stranicama Wikipedije na odabranom jeziku. U modelu razvijenom u članku wikifikacija je korištena tako da su prvo označili svaki 4-gram u tekstu. Nakon toga su svakoj riječi dali 3 značajke od kojih je svaka zapravo kategorija stranice na wikipediji i Freebase [3] tip (w_{i-1}, w_i, w_{i+1}). (Tsai et al., 2016)

Tablica 5.1: Broj gazetera za svaki entitet

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

[3] `https://developers.google.com/freebase/guide/basic_concepts`

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	$topic(w_i, w_{i+1}, w_{i+2}, w_{i+3})$
broj pojavljivanja riječi u kategoriji	$category\_count(w_i, \{PER, LOC, ORG\})$
Međujezične značajke	
—	—

5.2. Pretprocesiranje značajki

Većina korištenih značajki su kategoričke stoga su kodirane Onehot [4] metodom. Brojčane značajke kojima je definiran poredak skalirane su na interval $[0, 1]$. Za skaliranje je korišten MinMaxScaler[5].

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:

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

$$penalty = (l2, l1)$$

[4] 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`

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

oba parametra se koriste u svrhu regularizacije, alpha je konstanta koja množi regularizacijski faktor a penalty je regularizacijska funkcija.
Parametri logističke regresije korišteni za unakrsnu validaciju su:

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

$$penalty = (l2, l1)$$

6. Evaluacija

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, preciznost, odziv i točnost. Sustav je evaluiran na dva način:

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

Ovisno o primjeni sustava u obzir se uzima jedna od dvije navedene metode evaluacije koja više odgovara primjeni. Metoda točnog podudaranja je dosta stroža od standardne evaluacijske metode. Prikazat ćemo način evaluacije na primjeru. Ako imamo sljedeću rečenicu u testnom skupu:

$[\text{Leo}]_{\text{B-PER}}$ $[\text{Messi}]_{\text{I-PER}}$ played great match against $[\text{Real}]_{\text{B-ORG}}$ $[\text{Madrid}]_{\text{I-ORG}}$ in $[\text{Barcelona}]_{\text{B-LOC}}$.

a sustav je predvidio sljedeće oznake:

$[\text{Leo}]_{\text{B-PER}}$ $[\text{Messi}]_{\text{I-PER}}$ played great match against $[\text{Real}]_{\text{B-ORG}}$ $[\text{Madrid}]_{\text{B-LOC}}$ in $[\text{Barcelona}]_{\text{B-LOC}}$.

Standardna točnost ove rečenice je 80% dok se metodom točnog podudaranja dobije točnost 67% jer cijeli entitet Real Madrid nije točan, dok se entitet Leo Messi broji kao jedan točan primjer.

Tablica 6.1: Veličina skupova podataka

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

U sljedećim tablicama dane su f1-score mjere za jednojezične modele i međujezične modele trenirane na engleskom jeziku a testirane na drugima.

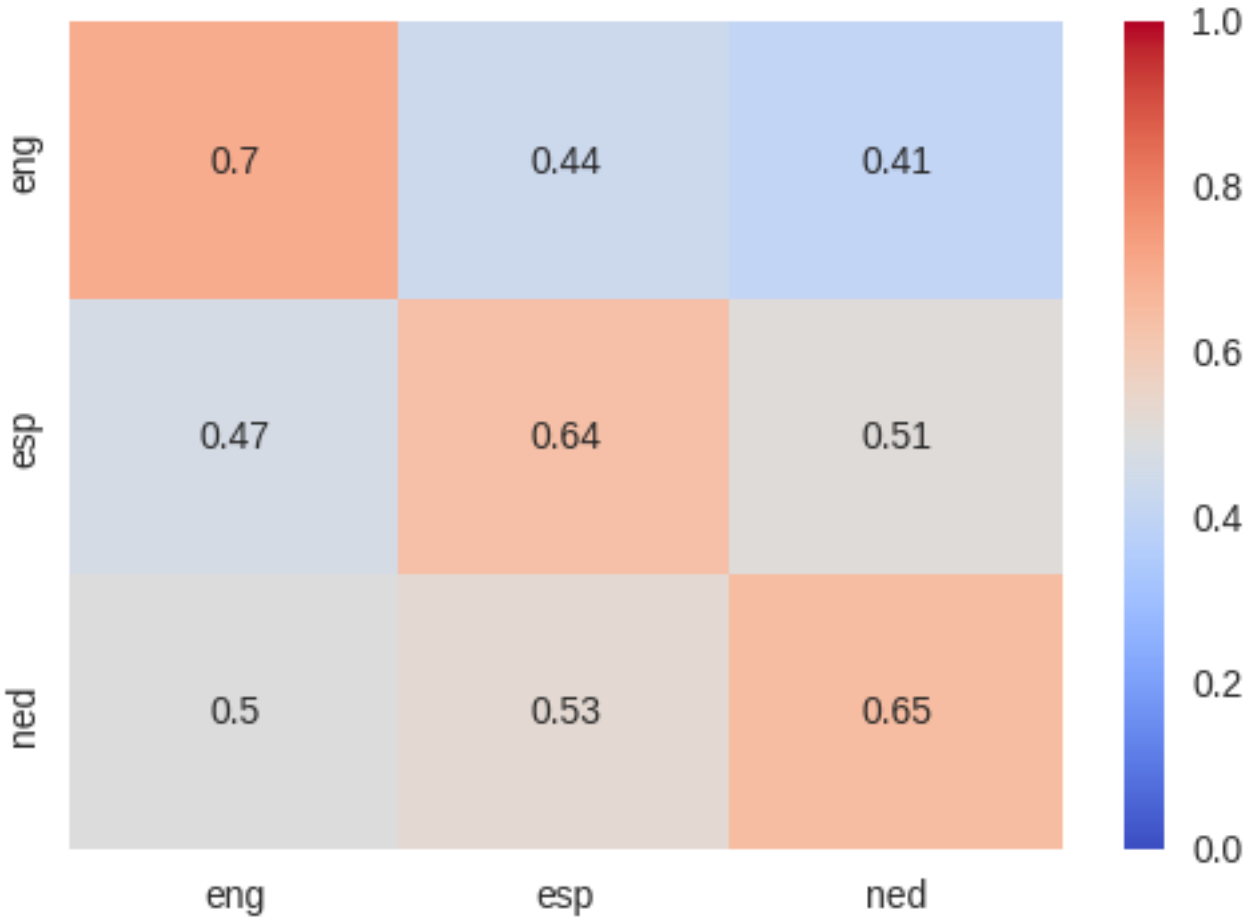
Tablica 6.2: Standardne evaluacijske mjere

	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	–	–	–	–	–	–	–	–
Međujezični 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	–	–	–	–	–	–	–	–

Tablica 6.3: 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	–	–	–	–	–	–	–	–
Međujezični 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	–	–	–	–	–	–	–	–

U sljedećem grafu prikazane su f1 mjere modela treniranog na jezicima na y-osi, a testiranog na jezicima na x-osi. Crvena boja označava bolju mjeru. Ako treniramo model na engleskom jeziku a testirano na nizozemskom dobivamo f1 score 0.5.



Slika 6.1: Evaluacijska matrica modela s različitim jezicima za treniranje i testiranje. Korištena je f1 mjera. Evaluacija je iyvršena na modela koji uključuje baseline i gazetere.

6.1. Poboljšanja

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 rezultirati poboljšanjem. Metoda Wikifikacije se može poboljšati boljom distribucijom kategorija i obogaćivanjem wikipedije za jezike s manjim resursima.

7. Zaključak

Potreba za komunikacijom i razumijevanjem svih jezika je veća nego ikad uslijed naglog razvoja interneta i mogućnosti povezivanja ljudi iz raznih dijelova svijeta. Također, dijeljenje znanja je lakše nego ikad. Većina zapisa na internetu je na engleskom jeziku stoga za engleski jezik postoje najbolji modeli za analizu jezika. Prepoznavanje imenovanih entiteta u tekstu je zadataka primjenjiv u raznim područjima. Za jezike s malo resursa nemamo dovoljno dobre modele za prepoznavanje. Zbog toga posežemo za malo drugačijim pristupom. Želimo iskoristiti dobre modele na engleskom jeziku za poboljšanje modela resursno siromašnijih jezika.

U ovom radu je razvijen model po uzoru na (Tsai et al., 2016) gdje je iskorištena wikifikacija za povezivanje entiteta na wikipedijama različitih jezika. Nažalost, zbog nemogućnosti iskorištavanja već gotovih programa za wikifikaciju te zbog nedostatka vremena i težine zadatka razvijanja vlastitog modela wikifikacija nije korištena te međujezični model nije davao dobre rezultate. Zbog toga razvijeni model radi puno bolje na jednojezičnim postavkama. Iako je hrvatska Wikipedija siromašna, zanimljivo bi bilo primjeniti ovaj pristup na hrvatskom jeziku.

Prostora za poboljšanje modela ima još puno kako u ovom tako i korištenjem nekih drugačijim pristupa.

8. Literatura

L. Ratinov, D. Roth, D. Downey, i M. Anderson. Local and global algorithms for disambiguation to wikipedia. U *ACL*, 2011. URL http://cogcomp.cs.illinois.edu/papers/RRDA11.pdf.

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