Named Entity Recognition in Croatian Tweets

Krešimir Baksa, Dino Dolović, Goran Glavaš, Jan Šnajder

University of Zagreb, Faculty of Electrical Engineering and Computing
Text Analysis and Knowledge Engineering Lab
Unska 3, 10000 Zagreb, Croatia
{kresimir.baksa, dino.dolovic, goran.glavas, jan.snajder}@fer.hr

Abstract

Existing named entity extraction tools, typically designed for formal texts written in standard language (e.g., news stories, essays, or legal texts), do not perform well on user-generated content (e.g., tweets). In this paper we present a supervised approach for named entity recognition and classification for Croatian tweets. Comparison of three different sequence labeling models (HMM, CRF, and SVM) revealed that CRF is the best model for the task, achieving a micro-averaged F_1-score of over 87%. We also demonstrate that the state-of-the-art NER model designed for Croatian standard language texts performs much worse than our Twitter-specific NER models.

Prepoznavanje imenskih entitet v hrvaških tvitih

Obstoječa orodja za prepoznavanje imenskih entitet, ki so tipično izdelana za formalna besedila, napisana v standardnem jeziku (npr. novice, eseji ali pravna besedila), ne delujejo dobro nad vsebinami, ki jih ustvarjajo uporabniki (npr. tviti). V prispevku predstavimo voden način za prepoznavanje in klasifikacijo imenskih entitet v hrvaških tvitih. Primerjava treh različnih modelov za označevanje zaporedij (HMM, CRF in SVM) je pokazala, da je najboljši model za to nalogo CRF, ki doseže za mikropovprečeno mero F_1 rezultat prek 87 %. Pokažemo tudi, da najboljši model za prepoznavanje hrvaških imenskih entitet v standardnem jeziku deluje mnogo slabše kot naši modeli za prepoznavanje imenskih entitet v tvitih.

1. Introduction

Named Entity Recognition (NER) is a well-known task in information extraction (IE) and natural language processing (NLP), which aims to extract and classify names (personal names, organizations, locations), temporal expressions, and numerical expressions appearing in natural language texts. For many applications (e.g., journalism, intelligence, historical research) named entities carry the piece of information that is crucial for understanding and interpreting the text. Robust named entity recognition is also essential for other IE and NLP tasks (e.g., relation extraction and sentiment analysis). For example, to identify towards whom the sentiment is expressed in news analysis, one first needs to identify people and organizations mentioned in news stories.

NER systems typically extract named entities from documents written in standard language (e.g., news stories, essays, manuals, legal documents, police reports), i.e., documents for which the correctness of language (spelling, grammar, vocabulary) is typically checked prior to their publishing. In contrast, a lot of textual content on the web that may contain valuable information (e.g., forums, blogs, posts on social networks) is user-generated, which means that it is written in informal and colloquial language. Such language is often orthographically and grammatically incorrect, and abounds with social-media jargon. This makes user-generated text very challenging for automated processing. It has been shown (Liu et al., 2011) that the performance of the standard NER systems drops significantly when applied to informal text

In this paper we address the task of named entity extraction from tweets in Croatian. Tweets are messages from a micro-blogging service Twitter in which users post anything from news and trending events to personal information. The approach taken in this work is a supervised one: we first manually annotate tweets with named entities and then train supervised machine learning models to automatically recognize named entities in tweets. We experiment with three different supervised models – a Hidden Markov Model (HMM), Conditional Random Fields (CRF), and Support Vector Machines (SVM) – and compare their performance in a relaxed and strict evaluation settings. To the best of our knowledge, this is the first work on named entity extraction from tweets for Croatian or a Slavic language in general.

The rest of the paper is structured as follows. In the next section, we give an overview of work on NER from tweets and NER for Croatian. In Section 3, we describe the dataset and the annotation process in more detail. In Section 4, we describe the different models and features used for the task, whereas in Section 5 we present and discuss the performance for all models. Finally, we conclude and outline ideas for future work in Section 6.

2. Related work

While there is an immense body of work on named entity recognition from texts written in standard language for various languages (Finkel et al., 2005; Faruqui et al., 2010; Cucchiarelli and Velardi, 2001; Poibeau, 2003), the work on named entity extraction from tweets is rather recent and so far virtually limited to English (Finin et al., 2010; Liu et al., 2011; Ritter et al., 2011; Li et al., 2012).

Finin et al. (2010) experimented with annotating named entities in tweets in English using crowdsourcing, which showed to be rather effective, fast, and cheap. Liu et al. (2011) use a semi-supervised approach to recognize and classify named entities in English tweets. They employ k-nearest neighbors (k-NN) classifier to pre-label the tweets and sequence labeling with CRF to capture fine-grained information encoded in tweets. Ritter et al. (2011) develop a POS-tagger, a shallow parser, and a named entity recognizer for English tweets by considering both in-domain and

out-of-domain data. Their NER system exploits the output of a tweet-adjusted POS-tagger, but also employs distant supervision by applying topic modeling with constraints based on a Freebase dictionary of entities. Unlike aforementioned supervised attempts, Li et al. (2012) introduce an unsupervised, two-step NER system for targeted Twitter streams. In the first step they partition the tweets into NE candidates, which they then rank using a random-walk model based on the intrinsic properties of Twitter streams.

A number of NER systems for standard Croatian have been developed, both rule-based (Bekavac and Tadić, 2007) and statistical ones (Ljubešić et al., 2012; Karan et al., 2013). Ljubešić et al. (2012) train the Stanford NER model (Finkel et al., 2005) on Croatian data manually annotated with basic classes of named entities (PERSON, ORGANIZATION, LOCATION, MISC). Karan et al. (2013) developed CroNER, a supervised NER system using sequence labeling with conditional random fields (CRF). CroNER employs a rich set of lexical and gazetteer-based features and enforces document-level consistency of individual classification decisions. CroNER annotates nine classes of named entities and is considered to be a state-of-art NER system for Croatian (Agić and Bekavac, 2013; Karan et al., 2013).

Like CroNER, in this work we also use sequence labeling algorithms for named entity recognition and classification. However, our models are trained on manually annotated tweets instead of standard-language texts. Similarly to Ljubešić et al. (2012), we focus on three main classes of named entities: PERSON, ORGANIZATION, and LOCATION. To confirm that extracting named entities from tweets is different from extracting named entities from standard text, we evaluated CroNER on the tweets dataset, where it exhibited a significant drop in performance.

3. Dataset and annotations

In our work we use the corpus of Croatian tweets compiled by Ljubešić et al. (2014) with the open-source tool TweetCaT. TweetCaT is designed to construct Twitter corpora for smaller languages like Croatian and Slovene by collecting the URLs of web pages from seed terms. The Croatian tweet corpus contains approximately 26 million tweets. However, a fairly large portion of tweets is in Serbian language. To ease filtering, each tweet has been automatically tagged with a language identification tag. From tweets tagged as Croatian, we selected a sample 5.000 tweets for manual annotation. We subsequently removed some tweets because they were informationally irrelevant (e.g., *"Ivana Ivana Ivana Ivana"*), leaving us with the final dataset of 4.667 tweets. Further inspection revealed that roughly 30% of tweets tagged as Croatian are actually written in Serbian, and that additional manual filtering would be required to obtain a clean dataset. Because of the considerable effort involved, we decided not to perform additional filtering, but instead decided to use the corpus with mixed Croatian and Serbian tweets.[1]

To speed up the annotation process, we performed semi-automated instead of fully manual annotation. Before initiating the semi-automated annotation, we compiled the annotation guidelines, some of which adopted from Finin et al. (2010):

- Annotate each token separately, following the B-I-O annotation scheme (e.g., *Hrvatska [B-ORG] narodna [I-ORG] banka [I-ORG]*);
- Annotate names, surnames, and nicknames but not their titles (e.g., *doc. dr. sc.* as instances of the PERSON class (e.g., *Marko [B-PER]*; *dr. Ivo [B-PER] Josipović [I-PER]*);
- Annotate names of concrete organizations, institutions, state authorities, sport clubs, national teams, but not generic terms like *government* or *party* as instances of the ORGANIZATION class (e.g., *NK [B-ORG] Rijeka [I-ORG]*);
- Annotate mentions of places, regions, states, rivers, mountains, squares, streets, etc. as instances of the LOCATION class (e.g., *Velika [B-LOC] Gorica [I-LOC]*);
- Do not annotate tokens starting with "@";
- Do annotate named entities preceded by "#";
- Annotate words according to the tweet context (e.g., token *"Rijeka"* may denote the location but it may also be part of the organization mention *"NK Rijeka"*);
- When in doubt whether to annotate the word as an instance of LOCATION or ORGANIZATION class, prefer ORGANIZATION.

Semi-automated annotation. The semi-automated annotation consists of two steps: (1) automated annotation of all mentions found in precompiled gazetteers and (2) manual correction of errors (both false positives and false negatives) made by the automated gazetteer-based annotation. This automated gazetteer-based annotation was also used as a baseline for the evaluation of supervised models. To perform the first step of the semi-automated annotation, we first needed to compile the set of gazetteers. Gazetteers with personal names (2413 entries) and locations (71 entries) were obtained from individual web resources.[2,3] A gazetteer with organization names (109 entries) was compiled from several different web resources. Following the automated gazetteer-based annotation, we manually corrected all errors introduced by the automated annotator. We also labeled named entity mentions omitted by the automated annotator. Organization mentions were most frequently omitted by the automated annotator because of (1) the limited size of organizations gazetteer and (2) the fact that the organizations gazetteer contained only single-word entries and organizational mentions quite often consists several words.

[1] Arguably, from a machine learning perspective, using a mixed Croatian-Serbian corpus as the train set introduces some noise in all cases in which the differences between the two languages are reflected in the feature values. On the other hand, our preliminary experiments, carried out on separate Croatian and Serbian test sets, have shown that the model performs equally well on both test sets. Thus, the upside of using a noisy dataset in this case is that one gets a model that works reasonably well for both languages.

[2] http://www.croatian-genealogy.com

[3] http://goo.gl/79ddLr

Class	MUC F_1 (%)	Exact F_1 (%)
PERSON	94.7	92.8
ORGANIZATION	85.7	81.2
LOCATION	86.6	85.2
Micro-average	91.3	88.8

Table 1: Inter-annotator agreement.

Many locations were also omitted because only names of Croatian cities were in the location gazetteer. Person names were omitted rather rarely, primarily due to the size of the corresponding gazetteer.

Manual annotation. The manual annotation step was performed by two annotators (the first two authors). Initially, both annotators independently annotated the same set of 500 tweets to measure the inter-annotator agreement (IAA) and assess how well the annotation guidelines are followed. The IAA was measured by computing both MUC and Exact F_1-scores between the annotations of the two annotators. In the MUC scheme two annotations are considered the same if they have the same class and their extents overlap in at least one token. In the Exact evaluation scheme, the match is only counted when the two annotation are exactly the same (same class and exactly the same extent). IAA scores for all NE classes are given in Table 1. After annotating the same initial 500 tweets, each of the annotators annotated a separate set of approximately $2,230$ tweets. These tweets were used for training and testing the models.[4]

4. NER models

4.1. Machine learning models

We used three different supervised machine learning models to extract and classify named entities in tweets: (1) a Hidden Markov Model (HMM), (2) Conditional Random Fields (CRF), and (3) a Supported Vector Machine (SVM). For all three models, we used the implementation in NLTK,[5] a popular Python library for natural language processing.

Hidden Markov Model. This model is an extension of Markov process where each state has all observations joined by the probability of the current state generating observation (Blunsom, 2004). Formally, HMM is defined as a tuple:

$$HMM = \big(S, O, A, B, \pi\big), \tag{1}$$

where *S* denotes hidden states (in our case labels of tokens), *O* stands for outputs in each state (in our case all words seen in tweets) and three parameters that denote probabilities computed from the annotated corpus: the starting probability π, transition probabilities A of going from one state to the other, and output probabilities *B*, in other words the probability of seeing a word when in one particular state.

Support Vector Machines. The standard SVM is a binary classification algorithm, which performs classification by maximizing the margin between the examples of the two classes. The binary SVM formulation can be easily extended to account for multi-class classification problems. However, in this work we employ a structured, sequence labeling variant of the SVM, proposed by Altun et al. (2003). Sequence labeling formulation of the SVM is very similar to the multi-class SVM formulation with exponentially many classes.

Conditional Random Fields. CRF is a discriminative probabilistic graphical model that can model overlapping, non-independent features in a sequence of data. A special case, linear-chain CRF, can be thought of as the *undirected graphical model* version of the HMM. Unlike HMM, CRF allows to extract arbitrary features for the current token as well as for preceding and following tokens. We used a window of size five for extracting the features, i.e., all of the features were computed for the current token and the two tokens preceding and succeeding it.

4.2. Features

Due to the nature of the models, slightly different feature sets were used for each of them. The following list is the union of the features used for all three models:

- f^1 – The lemma of the token;
- f^2 – The length of the token;
- f^3 – The shape of the token encodes the lower/upper casing of the word (e.g., the shape of the word *Ana* is ULL);
- f^4 – A feature indicating whether the token contains a non-alphanumeric character (e.g., *Lovrić-Merzel*);
- f^5 – A feature indicating whether the token contains only non-alphanumeric characters (e.g., *?!*);
- f^6 – Features indicating whether the token is the first or the last token in the tweet
- f^7 – A feature indicating whether the token contains any lower-cased letters;
- f^8 – A feature indicating whether the token contains any upper-cased letters;
- f^9 – A feature indicating whether the token contains any alphanumeric characters;
- f^{10} – A feature indicating whether the token contains digits (e.g., *sk8*);
- f^{11} – Features indicating whether the token matches a gazetteer entry (one feature per gazetteer, as a token can match multiple gazetteer entries).

For HMM, we used only one feature - the lemma of the word (f^1) – as other features cannot be incorporated into the standard HMM model. For the other two models – CRF and structured SVM – we used all above-mentioned features (f^1–f^{11}).

[4]The annotated dataset is available under the Creative Commons BY-NC-SA license from
http://takelab.fer.hr/cronertweet

[5]http://www.nltk.org/

	Baseline			HMM			SVM			CRF		
NE class	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
PERSON	96.47	84.38	90.02	93.83	81.46	87.21	90.74	88.25	89.46	94.83	92.72	93.76
LOCATION	50.00	27.30	35.32	90.00	16.02	27.20	52.16	39.47	44.93	78.35	68.77	70.33
ORGANIZATION	74.26	45.56	56.48	87.64	45.86	60.22	73.33	44.66	55.51	76.94	75.80	76.37
Overall macro	73.58	52.42	60.60	**90.49**	47.78	58.21	72.08	57.46	63.31	83.37	**79.10**	**81.13**
Overall micro	88.38	68.37	77.10	**92.63**	65.21	76.54	83.77	72.11	77.50	89.01	**86.10**	**87.53**

Table 2: MUC evaluation results.

	Baseline			HMM			SVM			CRF		
NE class	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
PERSON	64.48	55.90	59.89	84.64	73.90	78.90	82.22	80.91	81.56	89.18	88.00	88.58
LOCATION	46.20	25.22	32.63	86.67	15.43	26.20	50.20	37.98	43.24	71.65	62.90	66.99
ORGANIZATION	38.13	23.91	29.39	69.50	35.64	47.12	55.80	34.74	42.82	66.08	65.43	65.76
Overall macro	49.60	35.01	40.64	**80.27**	41.66	50.74	62.74	51.21	55.87	75.64	**72.11**	**73.78**
Overall micro	57.87	44.67	50.42	**82.10**	57.85	67.88	74.43	64.85	69.31	82.09	**79.99**	**81.03**

Table 3: Exact evaluation results.

5. Evaluation

5.1. Experimental setup

We split the tweets dataset into two or three sets, depending on the learning algorithm. Since HMM only uses lemmas as features, we did not have to perform feature selection as for the other two algorithms. Thus, for HMM we split the tweets into two sets: train set (3399 tweets) and test set (1268) tweets. We trained HMM on the train set and we report the performance of the model on the test set. For SVM and CRF we performed greedy backward feature selection to identify the best subset of features for the task. Thus, we split the dataset into three subsets: train set (3399 tweets), validation set (423 tweets), and test set (845) tweets. For both algorithms we optimized the set of features according to the performance on the validation set. We report the performance for CRF and SVM with optimal feature subsets on the test set. As the baseline we used the same automated method that we employed as the first step of the semi-automated annotation process – the token is tagged as a named entity of some type if it can be found in the gazetteer for that NE type. Additionally, the baseline merges adjacent tokens found in the same gazetteer into a single named entity mention.

5.2. Results

The performance for all three models and the baseline, measured for MUC and Exact setting, is given in Table 2 and Table 3, respectively. The performance is reported for each of the NE classes, along with both micro-averaged and macro-averaged performance.

The CRF model outperforms the other two models by a wide margin in both evaluation settings. This is the consequence of CRF taking into account features of the preceding and following tokens as well. Thus, it is able to learn the patterns of named entity occurrence much better then the other models. Interestingly, HMM exhibits best precision but very low recall in both evaluation settings. In the MUC setting, HMM model does not even outperform the baseline in terms of F_1-score.

The structured SVM consistently outperforms the baseline and the HMM model, but is also consistently outperformed by the CRF model. The most common cause of errors for the structured SVM model are tokens labeled as inside of a named entity (e.g., I-PER) even when the preceding token was not the beginning of a named entity (e.g., B-PER). In contrast, CRF assigns very low probabilities for the "I-" labels when previous label in the sequence is not "B-".

To asses the performance of the NER system built for texts written in standard language, we evaluated CroNER (Karan et al., 2013) on the test portion of the annotated Twitter dataset. The results for Croatian are in line with the observations for English (Liu et al., 2011) – the performance of the tagger built for texts written in standard language drops significantly when applied to tweets. CroNER exhibited micro-averaged performance of 35.8% F_1-score in the MUC setting, and merely 27.4% F_1-score of in the Exact evaluation setting.

6. Conclusion

Traditional IE and NLP tools have been shown ineffective when applied to user-generated content. This is especially true for tweets, micro-blogging messages filled with jargon vocabulary and abbreviations. In this paper we presented the work on named entity recognition for Croatian tweets. We semi-automatically annotated the collection of almost 5.000 tweets in Croatian and Serbian. We compared three different sequence labeling models, demonstrating that CRF, being able to incorporate context features, outperforms HMM and structured SVM as well as the gazetteer-based baseline. The overall performance of the CRF model (87% micro-averaged MUC F_1-score) is comparable to the performance of the state-of-the-art NER system for Croatian

standard language (90% micro-averaged MUC F_1-score; Karan et al. (2013)), which we consider very encouraging considering the lack of POS and syntactic information in current models. We also demonstrated that a NER system built for standard language texts performs poorly on tweets.

There are several possible extensions of the work presented in this paper. Firstly, we intend to extend the models with part-of-speech and syntactic information. This means that a designated POS-tagger and (shallow) parser for tweets need to be created for Croatian and Serbian as, similar to NER, standard tools have been shown inefficient. Secondly, a Twitter dataset could be enlarged in order to determine how the dataset size influences the performance of the tagger. Finally, we believe that enforcing consistency of named entity annotations across tweets of the same thread (re-tweets) would improve the overall performance.

7. References

Ž. Agić and B. Bekavac. 2013. Domain-aware evaluation of named entity recognition systems for Croatian. *CIT. Journal of Computing and Information Technology*, 21(3):185–199.

Y. Altun, I. Tsochantaridis, T. Hofmann, et al. 2003. Hidden markov support vector machines. In *ICML*, volume 3, pages 3–10.

B. Bekavac and M. Tadić. 2007. Implementation of Croatian NERC system. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing: Information Extraction and Enabling Technologies*, pages 11–18. Association for Computational Linguistics.

P. Blunsom. 2004. Hidden markov models. *Lecture notes, August*, 15:18–19.

A. Cucchiarelli and P. Velardi. 2001. Unsupervised named entity recognition using syntactic and semantic contextual evidence. *Computational Linguistics*, 27(1):123–131.

M. Faruqui, S. Padó, and M. Sprachverarbeitung. 2010. Training and evaluating a german named entity recognizer with semantic generalization. In *Proc. of KONVENS*, pages 129–133.

T. Finin, W. Murnane, A. Karandikar, N. Keller, J. Martineau, and M. Dredze. 2010. Annotating named entities in Twitter data with crowdsourcing. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 80–88. Association for Computational Linguistics.

J. R. Finkel, T. Grenager, and C. Manning. 2005. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 363–370. Association for Computational Linguistics.

M. Karan, G. Glavaš, F. Šarić, J. Šnajder, J. Mijić, A. Silić, and B. D. Bašić. 2013. CroNER: recognizing named entities in Croatian using conditional random fields. *Informatica (Slovenia)*, 37(2):165–172.

C. Li, J. Weng, Q. He, Y. Yao, A. Datta, A. Sun, and B.-S. Lee. 2012. Twiner: named entity recognition in targeted Twitter stream. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 721–730. ACM.

X. Liu, S. Zhang, F. Wei, and M. Zhou. 2011. Recognizing named entities in tweets. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 359–367. Association for Computational Linguistics.

N. Ljubešić, M. Stupar, and T. Jurić. 2012. Building named entity recognition models for Croatian and Slovene. In *Proceedings of the Eighth Information Society Language Technologies Conference*, pages 117–122.

N. Ljubešić, D. Fišer, and T. Erjavec. 2014. TweetCaT: a tool for building Twitter corpora of smaller languages. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

T. Poibeau. 2003. The multilingual named entity recognition framework. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 2*, pages 155–158. Association for Computational Linguistics.

A. Ritter, S. Clark, O. Etzioni, et al. 2011. Named entity recognition in tweets: an experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1524–1534. Association for Computational Linguistics.