Journal of Computing and Information Technology - CIT 21, 2013, 3, 195–209
doi:10.2498/cit.1002190

Domain-aware Evaluation of Named Entity Recognition Systems for Croatian

Željko Agić[1] and Božo Bekavac[2]

[1] Department of Information and Communication Sciences, Faculty of Humanities and Social Sciences, University of Zagreb, Croatia
[2] Department of Linguistics, Faculty of Humanities and Social Sciences, University of Zagreb, Croatia

We provide an evaluation of the currently available named entity recognition systems for Croatian. The evaluation puts special emphasis on domain dependence. To this goal, we manually annotated a dataset of approximately 1 million tokens of Croatian text from various domains within the newspaper text genre. The dataset was annotated using a three-class named entity tagset – denoting personal names, locations and organizations. We give insight to feature selection, domain sensitivity and effects of increase in training set size for statistical named entity recognition using the state-of-the-art Stanford NER system. We also sketch a comparison of publicly available named entity recognition systems for Croatian considering domain dependence, regardless of their underlying paradigms. Our top-performing system achieved an F_1-score of 0.884 in a mixed-domain testing scenario, scoring 0.925 and 0.843 in the two domains separated for the experiment. The system shows consistency in state-of-the-art scores for detecting names of persons, locations and organizations.

Keywords: named entity recognition, Croatian language, text domain, domain dependence, evaluation

1. Introduction and related work

Named entity recognition (NER, sometimes also NERC – named entity recognition and classification) is a subtask of information extraction and a well-established task in natural language processing (NLP). It involves detecting and classifying named entities in natural language text. A named entity is loosely defined as a phrase that clearly identifies one item from a set of other items with similar attributes. More specifically, in our interpretation, named entities are basically phrases in which natural language text and physical reality intersect. As with most information extraction tasks, defining named entity classes is goal-oriented, but it usually includes such classes as names of persons, locations and organizations (the three canonical ENAMEX classes), followed by temporal (TIMEX) and other numerical (NUMEX) expressions, such as monetary and percentage phrases. Since the usefulness of detecting, e.g., names of people and organizations in relation to certain properties of texts in which they are immersed – such as text domain, genre and topic, opinion or sentiment conveyed – was evident even in the early stages of information retrieval development, the first named entity recognition systems appeared as early as 1991 [19, 21]. Further development, especially regarding English language processing, included elaboration of the task definition and named entity classes at the MUC-6 conference [12] and the CoNLL 2003 shared task on named entity recognition [23] to the point of supporting NER through publicly available state-of-the-art systems such as Stanford NER [10] or generic sequence labeling solutions such as CRF++ [14] or CRFSuite [20]. Based on these facts, it is safe to state that detecting and classifying named entities in English text has reached a point of performance-, scalability- and integration-wise maturity.

Evaluation and adaptation of English NER systems regarding text genre and domain is also a well-elaborated and currently very active field of study in natural language processing, as domain adaptation in NLP is a very propulsive

research topic in general at this specific point in time. Research highlights for NER include various approaches to semi-supervised learning and domain adaptation [9, 25] and entity recognition in non-standard text such as blog posts, newspaper article comments, Twitter posts and other user-generated content from web sources in which text standardness is not guaranteed or expected [13, 22].

In contrast, research in advancing NER – both genre- or domain-wise and otherwise – for Slavic languages is not nearly as abundant as for English [11]. Since Croatian is a generally under-resourced South Slavic language, the Croatian-oriented NLP community is still dealing with basic NLP resource and tool development [26]. Ensuring public availability of NLP resources for Croatian still poses a challenge [2, 17].

More specifically, until recently, named entity detection in Croatian texts has only been addressed by a rule-based approach using finite state transducer cascades [5, 6]. Just recently, two systems were presented that address Croatian NER by statistical approaches using sequence labeling with conditional random fields (CRF) [11, 16]. All three systems report state-of-the-art performance on their respective datasets, which differ in size, text domains and specifications of named entity classes. The rule-based system OZANA [5, 6] uses the MUC-7 specification [7]. It organizes named entities into three top-level classes or seven classes overall:

1. named entities (ENAMEX): names of persons (PERSON), politically or geographically defined locations (LOCATION) and named corporate, governmental, or other organizational entities (ORGANIZATION),
2. temporal expressions (TIMEX): absolute or relative, complete or partial date (DATE) and time (TIME) expressions,
3. numeric expressions (NUMEX): monetary (MONEY) and percentage (PERCENT) expressions in numeric or alphabetic form.

CroNER [11] extends the MUC-7 specification with five additional experimental named entity classes and finally adds one of them – namely, the ethnic class – into the respective models. Stanford NER models of [16] annotate names of persons (PERS), locations (LOC) and organizations (ORG) respecting the MUC-7 guidelines for the ENAMEX top class and introduce a class for miscellaneous entities (MISC), following the guidelines of the CoNLL 2003 shared task [23]. OZANA and CroNER are developed and tested in the general domain of Croatian newspaper text and, therefore, they do not explicitly address the influence of dataset alterations on system performance. In contrast, the third line of research [16] does implicitly denote general domains included in the training set by stating the used sources – general, information technology, business, newspaper – but does not provide domain-dependent evaluation.

As documented, being a rule-based system, no training set or development set was used while constructing OZANA. CroNER was trained on approximately 310 000 tokens (310 kw) of unclassified Croatian newspaper text from a single source – the daily newspaper *Vjesnik*. The mixed-domain or domain-aware Stanford NER models were trained on a much smaller dataset of approximately 60 kw. The latter line of research also addresses some issues in linguistic preprocessing and feature selection in statistical NER, as it investigates the usefulness of lemmatization, part-of-speech tagging (POS tagging) and distributional similarity features for Croatian NER.

In this paper, we attempt to address four aspects regarding Croatian NER that were left largely unaddressed by previous research.

1. We investigate the effects of introducing domain-dependent training and testing data on statistical NER for Croatian.
2. We create a 1 million token NER-annotated corpus of Croatian by collecting and classifying newspaper text by domain, aiming at observing the effects of enlarging the training data for Croatian NER beyond the current 300 kw limit.
3. We provide a comparison of available NER systems for Croatian by conducting an empirical evaluation using domain-sensitive test sets created from our 1 Mw dataset.
4. We make the 100 kw domain-dependent test sets available to the public, seeking to enable a more uniform approach to testing Croatian NER systems. As a side-effect, we provide this dataset for possible training of new NER

systems for Croatian, improved in terms of domain adaptability.

Together with the test sets, we also make a selection of documented NER models publicly available for research purposes.

2. Experiment setup

In this section, we present the setup of our experiment with domain-dependent evaluation of Croatian NER systems. We define the experiment objectives and its workflow, which is divided into four experimental batches. We describe the datasets and their subdivision into training and testing samples, as well as the NER tools used in each of the experimental batches.

2.1. Objectives

Our experiment aims at investigating properties of named entity recognition in Croatian text from a viewpoint of text domain sensitivity. Keeping that in mind, we set several main objectives for the experiment:

- investigating general properties of state-of-the-art approaches to statistical NER when applied to Croatian text varying in domain, including an investigation of benefits of introducing certain linguistic features to these approaches;
- inspecting the effects of adding large(r) amounts of manually NER-annotated data to statistical NER training procedure for Croatian;
- investigating domain-sensitivity of statistical NER for Croatian by introducing domain-dependent data in training and testing stages and
- providing a survey of currently available NER systems for Croatian by performing a domain-dependent empirical evaluation.

Respecting these general guidelines and objectives, along the lines of ample research in domain-dependence for named entity recognition in English text, we seek to provide a new viewpoint for observing Croatian NER. By making the test sets publicly available, we hope to enable the research community with a more uniform evaluation approach for Croatian NER in the future, following a previous line of research in Croatian NER [16].

2.2. Training and testing sets

The texts in our experimental dataset were collected from the Vjesnik newspaper in the period of 2008-2010. The collection was done by a custom crawler and the texts were further cleansed, sentence-delimited and tokenized by using Apache OpenNLP tools [4] trained on manually delimited Croatian data and POS/MSD annotated using CroTag MSD tagger [1]. Manual annotation for named entities from the MUC-7 ENAMEX category (locations, organizations, persons) was done by expert annotators. The annotations were not overlapped and inter-annotator agreement was therefore not observed. High agreement on these classes was expected, following what was observed in previous research, e.g., in the process of developing the CroNER system [11], where the average inter-annotator agreement on ENAMEX was shown to be approximately 95%.

The reason we selected only the three ENAMEX classes from MUC-7 for our dataset is mainly a practical one. From one viewpoint, inter-annotator agreement is shown to be very high for these categories [11] and from another, instances of NUMEX and TIMEX classes are easily detectable automatically, i.e., by simple rule-based modules [5], enabling the focus switch to ENAMEX for the human annotators.

The data collection consists of texts separated into two main text domains:

1. internal affairs, i.e., articles mainly regarding Croatian internal politics and
2. other newspaper text domains, evenly distributed between culture, foreign affairs and other news, lifestyle and sports.

Basic dataset stats are given in Table 2. It shows that there is approximately 760 kw in the internal affairs domain and approximately 170 kw of text in other domains. Other features in the table illustrate certain shared properties, but also certain differences between the domains. For example, the ratio of approximately 1.5 tokens per single named entity is maintained across domains, while the token to named entity token ratio also slightly differs between domains.

9.05% tokens in the internal affairs domain belongs to named entities, while it holds for 8.20% in other domains. The differences are also evident within subdomains of the latter domain, e.g., with texts belonging to the lifestyle subdomain having as few as 4.93% named entity tokens.

Regarding the absolute size of the dataset, to our knowledge, this is the largest manually annotated corpus of Croatian text used in any experiment with Croatian NER to this point. It has approximately 55 thousand manually annotated ENAMEX entities or approximately 82 thousand named entity tokens.

Domain	Sent's	Tokens	NEs	NE tokens
Internal	28 665	758 034	45 654	68 652
Other	7 652	168 906	8 943	13 856
Sports	2 266	42 947	3 050	4 173
Culture	1 577	42 758	2 289	4 046
Life	1 810	43 222	1 342	2 129
Foreign	1 999	39 979	2 262	3 508
Total	36 317	926 940	54 597	82 508

Table 1. Dataset stats.

Domain	LOC	ORG	PERS
Internal	15 514	15 880	14 260
Other	3 079	1 833	4 031
Sports	988	547	1 515
Culture	593	498	1 198
Life	361	137	844
Foreign	1 137	651	474
Total	18 593	17 713	18 291

Table 2. Distribution of named entity classes.

Table 2 provides additional insight as to the spread of the named entities in the dataset across the three ENAMEX classes. In the internal affairs domain, the entities are pretty evenly spread between the classes, with approximately 15 thousand entities per class. In text from the other domains, however, the distribution is governed by names of persons, as there are twice as much person mentions than mentions of organizations and 30% more than instances of location names. Moreover, there are interesting differences between the subdomains of the other domains test set. For example, while names of persons are predominant in sports, culture and lifestyle texts (49%, 52% and 63% of all named entities, respectively), distribution of named entities in foreign affairs texts is governed by locations (50%), in which names of organizations are also relatively more frequent than in other subdomains (29%).

The dataset was split into training and testing sets for five-fold cross-validation by random sampling and respecting the document boundaries to avoid the bias of syntactic transfer [16]. The split was done separately for the internal affairs domain and for the other domains. Approximately 50 kw was left out for testing purposes for each of the two top domains, leaving out a test set of approximately 100 kw, evenly spread between five internal affairs samples of 10 kw each and five samples of 10 kw each for the other domains. A mixed test set of ten 10 kw samples was also created, randomly combining sentences from these two domain test sets to provide an uniform evaluation across the domains. The remaining 700 kw of internal affairs texts and 100 kw of texts from the other domains was used for training the models using Stanford NER.

It should be noted that splits or groupings of the test sets other than the internal vs. other domains could also be implemented. We chose to split the dataset in this manner mainly because of the prevailing size of the internal domains dataset and the fact that its annotation predates the annotation of the other domains dataset. Additionally, we envisioned the usage of the internal affairs models in certain real-world applications of named entity detection regarding relation discovery between Croatian political entities.

2.3. Workflow

In order to achieve the previously set research objectives, we define four interlinked batches of experiments with Croatian NER. Their workflows are exposed in the following four subsections.

2.3.1. Feature selection

Previous research in Croatian statistical NER [11, 16] has indicated the need for careful feature selection with the goal of achieving the state

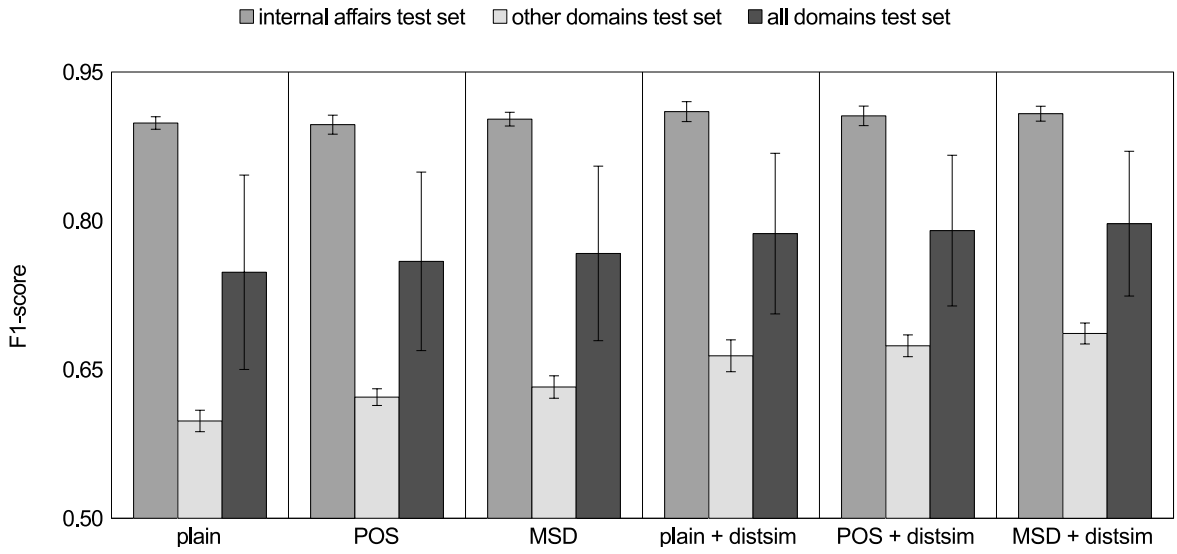

Figure 1. Overall F_1-scores for models trained on 300 000 tokens internal affairs training set with respect to features selected (plain, POS, MSD, with or without the use of distributional similarity features). The 95% confidence intervals are indicated by the error bars.

of the art in named entity detection. We focus on four specific sets of linguistic features: tokens, their part-of-speech annotation (POS tags) and full morphosyntactic annotation (MSD tags) and distributional similarity features (distsim). The POS and MSD annotation was provided by the CroTag tagger [1]. Lemmatization was disregarded as it was shown to introduce noise by previous research [16]. Distributional similarity features were calculated and shared freely in [16] by utilizing the clustering tool described in [8] over a 100 Mw sample of Croatian from the hrWaC corpus [15] with 400 clusters.

Since its sole purpose was to establish the optimal feature set for further experimentation, this batch of experiments was done by using only the internal affairs texts for training the models as the text from other domains was still under development at the time of conducting the batch. Eighteen 5-folded models were built in this batch, defined by a product of training set sizes and features used: (100 kw, 200 kw, 300 kw) $\times$ (plain text, POS, MSD) $\times$ (no distsim, distsim). Precision, recall and F_1-scores were calculated for the three named entity classes and overall on all available test sets.

2.3.2. Domain sensitivity

For the second testing batch, the internal affairs training sets from the first batch were injected with texts from the other domains training sets and tested on all test sets. In these experiments, only the size of the training sets varies, as only the best linguistic preprocessing feature set established in the first experiment batch is used. The linguistic feature sets are addressed in the previous paragraph, i.e., the description of the first experiment batch. The training set sizes with text domain denotation were as follows: 50 kw (internal) + 50 kw (other), 100 kw (internal) + 100 kw (other) and 200 kw (internal) + 100 kw (other). The results of detection using the mixed models are compared with those of the internal-affairs-only models from the first batch that used the same feature set.

2.3.3. Training set enlargement

Having created a 1 Mw manually ENAMEX-annotated corpus of Croatian newspaper text, in this testing batch, we implemented an experiment with increasing the amount of training data beyond the limit of approximately 300 kw from the previous two batches of our experimentation and previous research in statistical NER for Croatian [11].

Namely, we extend the experiment from the second batch by using the abundance of annotated

text from the internal affairs domain. We introduce more text to the training procedure, stepping by 100 kw: from 300 kw to 700 kw. Following the previous scheme, we introduce four new models trained on internal affairs text only (400, 500, 600 and 700 kw) and four new models trained by mixing internal affairs text with text from other domains (300 + 100 kw,. . . , 600 + 100 kw). Since text from the other domains has a size cap of 100 kw for training and since we wanted to evaluate it against the internal affairs models, the last remaining option (700 + 100) was not used in the experiment as it would be 100 kw larger than the largest internal affairs model and thus not directly comparable.

2.3.4. System comparison

In the final batch of experiments, we compare the top-performing model, selected within the previous three batches, with the other available NER systems for Croatian. We use our ENAMEX-annotated and domain-sensitive test sets to provide this comparison. Since the other systems were either trained on manually annotated data following other annotation schemes and using other feature sets or were otherwise designed to match other annotation schemes, this comparison might introduce a bias towards our top-performing system. However, this being to our knowledge the first such comparative evaluation of Croatian NER systems, we chose to provide it as a reference point for comparing the existing systems in terms of underlying paradigms, while separating our top-performing model. Additionally, extracting the ENAMEX categories from text is generally considered as the most important subtask of NER [19]. Thus we consider measuring the performance of all existing Croatian NER systems on these specific categories to be of major importance regardless of the underlying named entity models implemented by specific systems.

2.4. Tools

For the first three batches of experiments, we utilized the Stanford NER system for detecting named entities using sequence labeling with conditional random fields (CRF) [10]. It is a well-documented and supported state-of-the-art system ubiquitously used in today's NLP research. In the fourth batch, we compared the best model obtained by using Stanford NER on our dataset in the previous batches with other existing named entity recognizers for Croatian:

1. CroNER [11], a named entity recognition and classification system for Croatian language based on supervised sequence labeling with conditional random fields (CRF), developed from scratch using CRFSuite [20] and a rich set of features,
2. OZANA [5, 6], a rule-based NER system based on finite state transducer cascades, additional supporting heuristics and heavy use of specialized Croatian inflectional lexica, implemented using the Intex linguistic development environment [24] and
3. Stanford NER models trained on approximately 60 kw of Croatian mixed-domain text by Ljubešić et al. (2012) [16].

As previously discussed, the systems differ with respect to the named entity models they implement. CroNER implements the MUC-7 categories and introduces the category of ethnic. Ljubešić et al. (2012) use the four categories from the CoNLL 2003 shared task, ENAMEX and MISC, while OZANA closely follows the MUC-7 specification. Our models use the three-class ENAMEX model encoded in our dataset. Since the four systems' named entity models overlap precisely in the three-class ENAMEX model, we observe only these three classes throughout the experiment.

Other than in underlying paradigms, the systems also differ in levels of public availability. The rule-based system OZANA is not publicly available. CroNER was made available for purposes of our experiment as a web service. The 60 kw Stanford NER models of [16] were freely available online via permissive licensing at the time of conducting this experiment. To the best of our knowledge, no other Croatian NER systems existed at that point in time.

3. Results and discussion

In this section we provide a discussion of the results obtained in the four batches of our experiment.

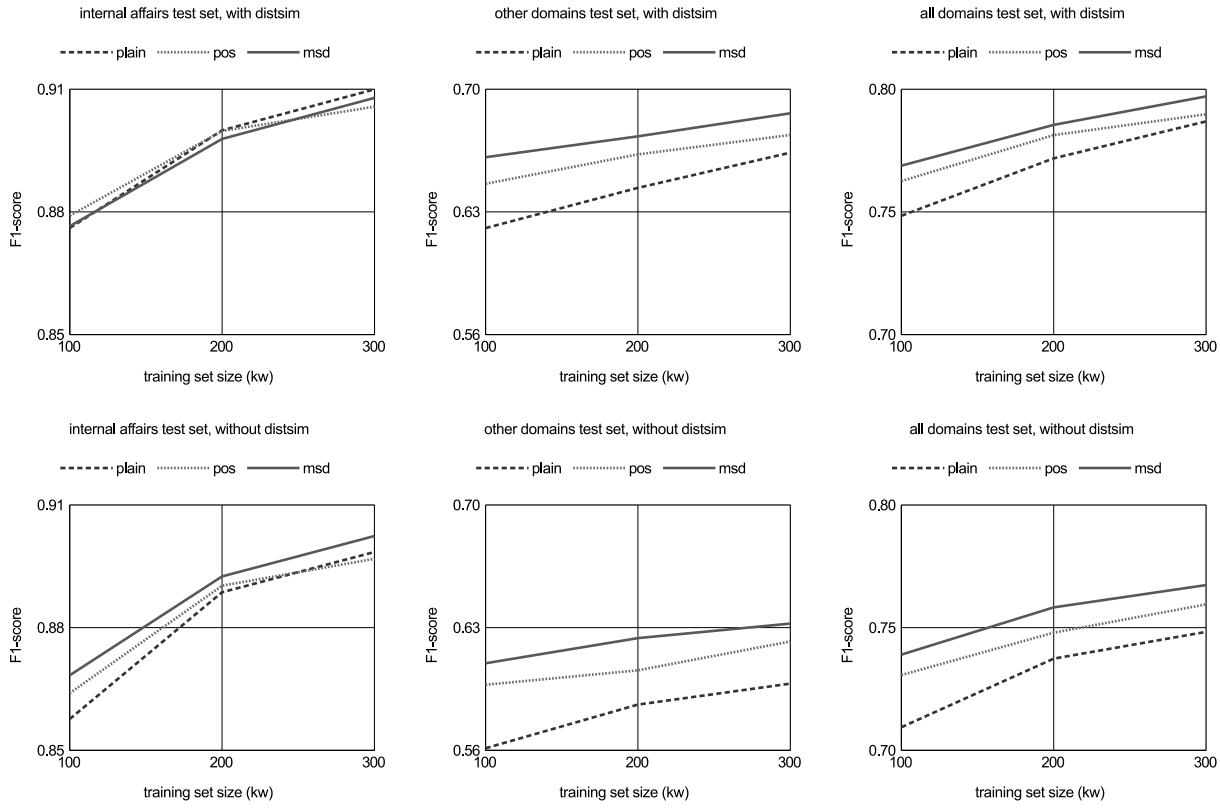

Figure 2. F_1-score learning curves for models trained on internal affairs training set with respect to features selected (plain, POS, MSD, distributional similarity) and test set used (internal affairs, other domains, all domains).

3.1. Feature selection

The results of the first experiment batch in terms of F_1-scores achieved by Stanford NER models trained on internal affairs 300 kw training sets and tested on internal affairs and other domains test sets are given in Figure 1 with test sets, feature sets and overall scores indicated.

Overall F_1-scores show that the top-performing models are consistently using a training set size of 300 kw, MSD features and distributional similarity (distsim) features. The highest observed F_1-score on the internal affairs test set is, however, achieved by the model trained on unannotated text (plain) and distsim features and it amounts to 0.910. The top-performer for the other domains test set is the 300 kw MSD and distsim model with an F_1-score of 0.686. This is also reflected in the all domains test scenario, where the same 300 kw MSD distsim system scored an overall F_1-score of 0.797. It should be noted that the small differences between the in-domain scores are not statistically significant.

Models using distributional similarity features consistently outperform the respective models without these features. The overall difference in F_1-scores between these two groups of models increases with the complexity of the test set: from less than 1% increase for the internal affairs test set to 5% increase for the other domains test scenario which is in turn reflected in the all domains test set increase of approximately 3% in favor of the models using distributional similarity features.

Domain selection influence on overall named entity detection accuracy is substantial. F_1-score decrease of 0.224 is observed between the best NER systems when comparing internal affairs and other domains test set. Domain dependence is also reflected by the impact of feature selection on the results across domains. While introducing additional training set data and additional features to the internal affairs models does not provide a substantial increase in F_1-scores for internal affairs texts, the scores on other domains benefit both from increasing the training set size and from adding POS, MSD and distributional similarity features.

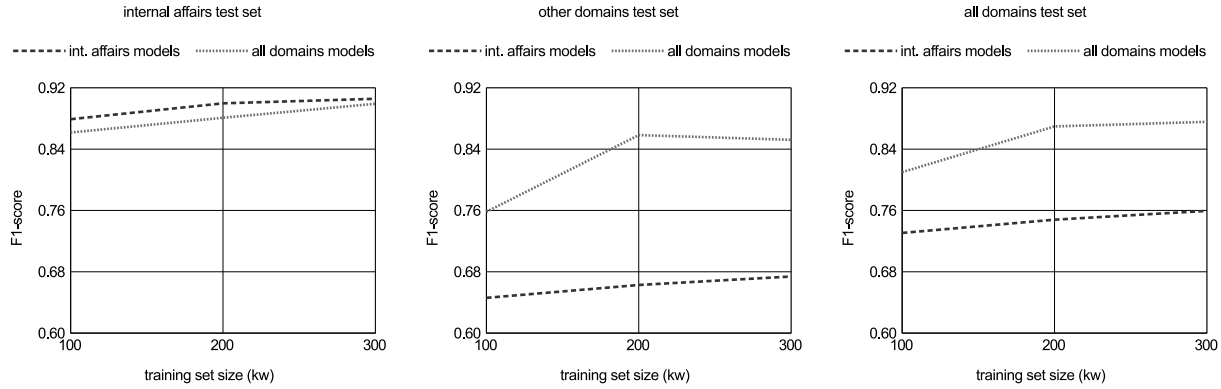

Figure 3. Comparison of F_1-score learning curves for internal affairs models and all domains models on all test sets, using POS and distributional similarity features.

Model	Internal affairs test set			Other domains test set			All domains test set		
	P	R	F_1	P	R	F_1	P	R	F_1
Internal affairs	**0.906**	**0.906**	**0.906**	0.725	0.629	0.674	0.816	0.767	0.790
All domains (200 + 100)	0.900	0.899	0.899	**0.869**	**0.837**	**0.852**	**0.884**	**0.868**	**0.876**

Table 3. Overall F_1-scores for internal affairs system and all domains system trained on 300 000 tokens with POS and distributional similarity features.

Functional dependencies of training set sizes, selected training features and overall system performances are additionally illustrated by Figure 2. It clearly indicates both the strength of feature selection influence and the learning rates for the internal affairs models in the two test scenarios. The groupings of learning curves also illustrate the significance of differences in the results. Moreover, the learning curves indicate that the internal affairs training set size of 300 kw is sufficient to achieve in-domain state-of-the-art performance in comparison with other NER systems for Croatian.

Statistical significance exploration using t-tests indicates that the difference between POS and MSD models using distributional similarity features is in fact not significant in this specific five-fold cross-validation testing scenario. Respecting this fact and considering that NER models using POS and distsim features are also substantially smaller and faster to train and use, they are further observed in more detail. For the same reasons, the POS and distsim feature set is the only feature set used in the following batches of experiments. Regarding the observed absence of statistical significance of differences between the F_1-scores of NER models using POS and MSD features, it should be noted that the POS tagging accuracy of CroTag can be estimated at 95%, while its MSD tagging accuracy with the full tagset depends on the number of out-of-vocabulary word forms and peaks at approximately 85% for 20% unknown words [1, 2]. Thus the difference between models using POS and MSD would probably be more significant if the models were trained and tested using perfect tagging. However, as perfect tagging is almost never available in real-life scenarios for natural language processing systems, accuracy and speed of POS taggers paired with speed of training and using the resulting NER systems and a statistically insignificant decrease in named entity detection accuracy should be considered to be the most feasible choice, at least judging from the results presented here.

3.2. Domain sensitivity

The second batch of experiments included creating mixed-domain Stanford NER models by combining the internal affairs and the other domains training sets. As previously elaborated, with respect to training set size, the mixed-domain models are denoted as follows: the 100 kw model used 50 kw from the in-domain training set and 50 kw from the out-of-domain

Model	Internal affairs test set			Other domains test set			All domains test set		
	LOC	ORG	PER	LOC	ORG	PER	LOC	ORG	PER
Internal affairs	**0.905**	**0.855**	0.960	0.705	0.440	0.753	0.805	0.648	0.857
All domains (200 + 100)	0.895	0.841	**0.964**	**0.869**	**0.733**	**0.890**	**0.882**	**0.787**	**0.927**

Table 4. Named entity classification F_1-scores for internal affairs system and all domains system trained on 300 000 tokens with POS and distributional similarity features.

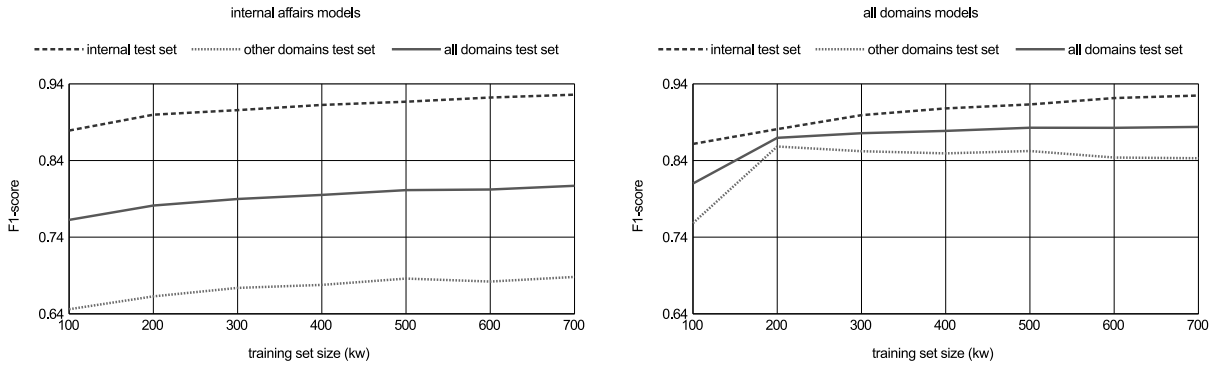

Figure 4. Overall F_1-score learning curves for enlarged internal affairs systems and all domains systems with POS and distributional similarity features.

training set, 200 kw equals the 100 kw + 100 kw mix and 300 kw amounts to 200 kw of internal affairs text and 100 kw of other domains text. On all models from the second batch, POS and distributional similarity features are used. The all domains models are compared with the respective internal affairs models. Figure 3 provides a comparison of learning curves for the two sets of models, while the top-performing internal affairs model and the top-performing all domains model are compared for overall accuracy in Table 3 and for accuracy on the three ENAMEX classes in Table 4.

The results show that joining data from differing text domains into single named entity detection models creates more robust and less error-prone NER systems. Standard statistical tests show that the difference in accuracy between the best internal affairs model and the top-performing mixed-domain model is not significant for the internal affairs test sets, while being substantially significant – with a high F_1-score difference of 0.178 and 0.086 – for the other domains and the all domains test sets. The top-performing mixed model still maintains the state-of-the-art accuracy of 0.899 in the internal affairs domain test scenario and 0.876 overall.

Table 3 is complemented with learning curves for the two sets of models in Figure 3. They indicate a large difference in the observed F_1-scores. The decline in the learning curve of the mixed-domain model when advancing from 200 kw to 300 kw training samples is subject to interpretation, as it might represent saturation of the model with text from the other domains as well as a local inflection point. This observation is further investigated by enlarging the training set in the following subsection.

Table 4 is a breakdown of the overall scores of the top-performing internal affairs models and all domains models into three named entity classes. The best mixed-domain model (300 kw) is significantly better at detecting personal and organizational names in the internal affairs test set than in the other domains test set – with overall difference of 0.074 and 0.108 in F_1-scores – and to some extent also at detecting locations. This might indicate that enlarging the out-of-domain training set might improve the mixed-domain model accuracy. What was previously observed when comparing the overall F_1-scores is decomposed in this table into three named entity classes and their respective F_1-scores follow a similar pattern of difference between the internal affairs and the other domains data. The systems are consistently better at detecting names of people than location

names and organization names. This effect is most likely due to the frequency of these entities in the training data and the data for distributional similarity modeling, as well as due to the inherently higher linguistic complexity of organizational names when compared with names of persons and locations. It should be taken into account that person tokens are most often regular in terms of capitalization while this does not hold for organization tokens.

Inspecting the differences between precision and recall in Table 3 also reveals room for improvement by introducing additional training data from other domains, as the difference between the two metrics is non-existent for the internal affairs test set and substantial (approximately 0.1) for the other domains test set. It should be noted that specific utilizations of NER systems might favor precision over recall and vice versa. (For example, OZANA, the rule-based system we use in the fourth batch, favors precision over recall.) This should be taken into account in a more elaborated system optimization and fine-tuning for specific tasks. Here, we focus mainly on optimizing F_1-scores by increasing both precision and recall and we thus consider them to be of equal importance.

The mixed-domain system outperforms the internal affairs system on all classes but locations and organizations in the internal affairs test set, with the latter difference not being statistically significant. As with the overall results in Table 3, the subdivision of results into the ENAMEX classes shows statistically significant differences in results in favor of the mixed-domain model in the other domains and all domains test sets. This fully supports the argument for domain-aware design of statistical NER training sets for Croatian.

3.3. Training set enlargement

In this experiment batch, we dealt with enriching the internal affairs and the all domains training sets from 300 kw to 700 kw in order to observe the overall improvements when more than doubling the training set size. Figure 4 provides two sets of learning curves: three for the internal affairs models (one curve for each of the three test sets) and three for the all domains models. The learning curves for the internal affairs models all rise steadily, but their increase is unsubstantial, as supported by the data in Table 5. Namely, the accuracy gain for moving from the 300 kw to the 700 kw training set amounts to 0.020 and 0.014 for the internal affairs test set and the other domains test set, respectively. However, introducing new internal affairs data benefits the mixed-domain models as their increase is observed at 0.026 for the internal affairs test set when compared with the 300 kw mixed-domain model. On the other hand, they yield a statistically insignificant decrease in accuracy in the other domains test set. This is most likely due to bias introduction considering the 6:1 ratio in favor of one

Model	Internal affairs test set			Other domains test set			All domains test set		
	P	R	F_1	P	R	F_1	P	R	F_1
Internal affairs	**0.924**	0.928	**0.926**	0.734	0648	0.688	0.829	0.788	0.807
All domains (600 + 100)	0.922	0.928	0.925	**0.854**	**0.832**	**0.843**	**0.888**	**0.880**	**0.884**

Table 5. Overall F_1-scores for internal affairs system and all domains system trained on 700 000 tokens with POS and distributional similarity features.

Model	Internal affairs test set			Other domains test set			All domains test set		
	LOC	ORG	PER	LOC	ORG	PER	LOC	ORG	PER
Internal affairs	**0.922**	**0.885**	0.973	0.701	0.446	0.785	0.811	0.665	0.879
All domains (600 + 100)	0.918	0.884	**0.974**	**0.847**	**0.719**	**0.895**	**0.883**	**0.802**	**0.934**

Table 6. Named entity classification F_1-scores for internal affairs system and all domains system trained on 700 000 tokens with POS and distributional similarity features.

System	Internal affairs test set			Other domains test set			All domains test set		
	P	R	F_1	P	R	F_1	P	R	F_1
CroNER	0.797	0.853	0.824	0.719	0.700	0.709	0.758	0.777	0.767
Ljubešić et al. (2012)	0.782	0.715	0.747	0.598	0.510	0.551	0.690	0.612	0.649
OZANA	0.897	0.735	0.808	**0.867**	0.460	0.601	0.882	0.597	0.712
Our best system	**0.922**	**0.928**	**0.925**	0.854	**0.832**	**0.843**	**0.888**	**0.880**	**0.884**

Table 7. Overall F_1-scores for system comparison.

System	Internal affairs test set			Other domains test set			All domains test set		
	LOC	ORG	PER	LOC	ORG	PER	LOC	ORG	PER
CroNER	0.835	0.743	0.922	0.691	0.552	0.811	0.763	0.648	0.867
Ljubešić et al. (2012)	0.777	0.661	0.835	0.650	0.277	0.680	0.714	0.469	0.757
OZANA	0.777	0.748	0.900	0.632	0.314	0.698	0.704	0.531	0.799
Our best system	**0.918**	**0.884**	**0.974**	**0.847**	**0.719**	**0.895**	**0.883**	**0.802**	**0.934**

Table 8. Named entity classification F_1-scores for system comparison.

text domain over the other in the 700 kw mixed-domain test set. The learning curve for the all domains models in the other domains test set indicates an inflection point at 200 kw: introducing new data beyond the 200 kw (100 kw + 100 kw) dataset slowly biases the system towards the internal affairs test set, making the task of detecting named entities in text from other domains slightly more difficult. These figures should by all means be taken into account when designing statistical NER systems for Croatian and targeting a specific text domain, as we show an importance of establishing the optimal ratio in mixing data from different domains. Document weighting schemes might also be considered in future research.

Overall scores from Table 5 are further supported by the illustration in Figure 5, which clearly indicates the positive effects of introducing other domain data into the training set across test set domains.

In terms of overall scores on the all domains test set as a single reference point for overall system accuracy, however, the 6:1 ratio mixed domain system (700 kw; 600 kw internal + 100 kw other) is the top-performing system for all three batches of our experiments with Stanford NER. Therefore, we select this system as our entry point to the fourth batch of experiments, i.e., for comparison with other existing systems for Croatian NER. Differences between precision and recall that we observed in the all domains 300 kw models in Table 3 are evened out in the 700 kw models as they are shown to be virtually non-existent by Table 5. This might serve as another indicator of this model reaching a high point in overall accuracy.

Table 6 provides the split of accuracy for the 700 kw internal affairs model and the 600 kw + 100 kw mixed-domain model on the three ENAMEX classes. The patterns observed for the 300 kw systems in the second experimental batch are maintained and follow the pattern of overall scores from Table 5 as mixed-domain model scores increase on the internal affairs test set and slightly decrease on the other domains test set in comparison to the scores in Table 4, amounting to a slight general increase across domains.

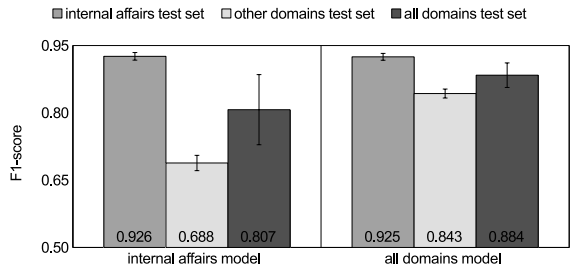

Figure 5. Overall F_1-scores for internal affairs system and all domains system trained on 700 000 tokens with POS and distributional similarity features. The 95% confidence intervals are indicated by the error bars.

3.4. System comparison

For the fourth and final experimental batch of this research plan, we compared the performance of our best 700 kw mixed-domain Stanford NER model with other Croatian NER systems: CRF-based system CroNER, rule-based OZANA and 60 kw Stanford NER model of Ljubešić et al. (2012). This evaluation serves not only as a comparison against our top-performing model, but to our knowledge the first uniform evaluation of the three systems on a single test set.

The overall scores – precision, recall and F_1-score on the two domains and the joint test set – are given in Table 7. As the training sets used for our system and test sets are derived from the same dataset, even with respecting document boundaries and randomizing the samples accordingly, introduces a positive bias of the results towards our system. Thus we present its scores in separation from the scores of the three other systems. It should also be noted that the three systems – CroNER, Ljubešić et al. (2012) and OZANA – significantly vary in training set domains and sizes and even underlying paradigms. However, as their implementations and licensing schemes prevent us from providing uniform datasets at training, we are only able to provide their black-box comparison using existing models.

As expected, our system is the top-performer of the group by a large margin on all test sets. The difference in overall F_1-scores is maintained at at least 0.1 in all the test scenarios. Out of the three remaining systems, CroNER consistently outperforms OZANA and Ljubešić's Stanford NER model – this fact is maintained across the domains. CroNER performs the best on the internal affairs test set, with a 0.113 decrease in its F_1-score when moving to other domains. Its precision-recall ratio remains even across domains. OZANA scores a second place in the group of existing systems. Its precision is consistently higher than recall by approximately 0.1, as it is a rule-based system that was tuned specifically for higher precision by design. Its precision is actually higher than the one of our top-performing model for the other domains test set. Its performance is significantly higher on the internal affairs test set, with a difference of approximately 0.2 in F_1 over the other domains test set. This is also due to system design, since it was developed to target Croatian text from this specific domain. We believe that fine-tuning the OZANA system by applying a set of relaxed rule cascades for increased recall, possibly even simple orthography-based rules, would further enhance its overall performance. Ljubešić's 2012 system peaks at approximately 0.75 F_1 on internal affairs and 0.55 on other domains, thus displaying the lowest performance in the group. This is due to its limited 60 kw training set size with mixed data from differing domains and sources from the Croatian web.

The overall scores are separated by ENAMEX classes in Table 8. As expected, the system rankings are maintained. All systems perform the best in the subtask of personal name detection, followed by names of locations and finally names of organizations on all test sets. Decreases when moving from the internal affairs test set to the other domains test set are substantial for all systems. CroNER suffers an F_1-score decrease of approximately 0.11 for personal names, 0.14 for locations and almost 0.2 for organizational names. OZANA maintains this pattern, but with a much steeper decrease for names of persons and organizations – 0.2 and almost 0.45 in F_1, respectively, once again most likely due to targeting its design towards one text domain. Ljubešić's system suffers a 0.4 decrease in F_1 for detecting organizations in texts from other domains and maintains the pattern of its statistical sibling CroNER for the other two classes.

Keeping in mind the positive bias introduced by syntactic transfer, our top-performing Stanford NER model outperforms the other systems on all ENAMEX classes by a large margin. The margin enlarges when moving from the easiest class (personal names, at least 0.05 difference in F_1 – inferred by comparing our system to CroNER) to names of locations (0.12) and finally names of organizations (0.15). It scores a state-of-the-art F_1-score of 0.934 for detecting personal names across domains, followed by 0.883 for locations and 0.802 for names of organizations. The scores we observed for the internal affairs test set are even higher.

Finally, we provide a brief overview of our manual error analysis of the statistical NER systems – CroNER, Ljubešić et al. (2012) and our top-

performing mixed-domain system – in comparison with OZANA as a rule-based NER system.

All the systems, especially statistical ones, exhibit an issue with detecting ending words of organizational names as they are of higher complexity and may vary in length. Examples include ***Muzej*** *suvremene umjetnosti* (en. ***Museum*** *of contemporary art*) and ***Kazalište*** *Kerempuh* (en. ***Theater*** *Kerempuh*) – only boldfaced words are recognized by the systems.

OZANA underperforms in detecting names of foreign, e.g., non-Croatian persons due to its underlying lexical resources. Namely, its foreign persons and foreign organizations lexica are of lower quality than the ones for Croatian names. Thus it also exhibits lower accuracy in detecting foreign organizations. On the other hand, it sometimes annotates multi-word expressions as organization names even if they do not refer to specific organizations, but rather to generic entities.

Statistical systems err by falsely classifying certain occurrences of the preposition *u* (en. *in*) followed by words with uppercased first letter as names of locations, e.g., *u* ***Nedinu*** *ili u* ***Pamelinu*** (en. *in* ***Neda's*** *or in* ***Pamela's*** – names of persons, not locations) and *u* ***Gavelli*** (en. *in* ***Gavella*** – organization, not location). They also exhibit and issue with detecting entities linked by conjunctions, as they are usually of the same type and most frequently indicate names of locations. Thus, in some of the occurrences, they annotate, e.g., ***Federera*** i ***Nadala*** (en. ***Federer*** and ***Nadal***) as names of locations instead of personal names.

4. Conclusions and Future Work

In this contribution, we have addressed text domain dependence of statistical named entity recognition and classification in Croatian texts. We have observed a strong preference for models trained on datasets encompassing multiple domain text, where state-of-the-art accuracy in terms of overall scores and scores on all ENAMEX categories (detecting personal names, names of locations and names of organizations) was observed in all test scenarios. We have also provided what we believe is the first empirical comparative evaluation of the available NER systems for Croatian. Three systems were compared to our top-performing model and evaluated for overall scores and scores on the ENAMEX classes. We made a subset of the developed Stanford NER models and domain-specific datasets for testing Croatian NER publicly available for research purposes[1].

Our future work plans include enlarging the used datasets and introducing datasets for other text genres and domains. It should be noted that our experiment addressed only issues with detecting named entities in various domains of the newspaper genre with domains collapsed into two disjoint classes. Experiments with fine-grained domain separation, genre variation and non-standard text processing have yet to be conducted for Croatian NER. Following the experience of [11], we should attempt to utilize a wider selection of features in pair with our dataset, possibly moving from the Stanford NER platform towards a more generic sequence labeling solution such as CRF++ [14] or CRFSuite [20]. Semi-supervised learning for NER [18] seems like another feasible line of research, as well as implementing voting NER systems, keeping in mind the observed differences between systems, especially across the underlying paradigms. Domain adaptation strategies for NER in Croatian texts should be thoroughly investigated.

As a final remark, we take note of Stanford NER models for Croatian that were developed upon completion of our experiment [17]. The two systems use the three-class ENAMEX and the four-class CoNLL 2003 named entity models, respectively. The former one is trained by combining the dataset of [16], our mixed-domain 100 kw test sets and the newly-developed 180 kw NE-annotated SETimes.HR corpus and dependency treebank of Croatian [2, 3], amounting for an aproximately 340 kw large training set. The authors report an overall named entity recognition score of 0.899 [17] and provide the models to the general public with a very permissive licensing scheme. Along with the positive bias of our test set documented in our experiment setup, this further establishes the need for a newer and more unbiased genre-, domain-

[1] The data is freely downloadable from `http://zeljko.agic.me/` and `http://nlp.ffzg.hr/` under the Creative Commons BY-NC-SA 3.0 license.

and text-standardness-aware test set to compare this newly-developed system with the systems presented in this paper to provide a more up-to-date uniform comparative evaluation of Croatian NER systems.

5. Acknowledgement

We would like to thank the anonymous reviewers for their valuable feedback. Many thanks to the team of annotators – Ana Agić, Daša Berović, Danijela Merkler, Matea Srebačić and a number of students from the Department of Linguistics – for their hard work in creating our 1 Mw ENAMEX gold standard for Croatian NER.

This work was partly funded by the European Commission within the Seventh Framework Programme (FP7/2007-2013 STREP) under the grant number 288342 (project XLike).

References

[1] Ž. AGIĆ, M. TADIĆ, Z. DOVEDAN, Improving Part-of-Speech Tagging Accuracy for Croatian by Morphological Analysis. *Informatica*, **32**(4) (2008), 445–451.

[2] Ž. AGIĆ, N. LJUBEŠIĆ, D. MERKLER, Lemmatization and Morphosyntactic Tagging of Croatian and Serbian. In *Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing (BSNLP 2013)*, ACL; (2013).

[3] Ž. AGIĆ, D. MERKLER, Three Syntactic Formalisms for Data-Driven Dependency Parsing of Croatian. In *Text, Speech and Dialogue, Lecture Notes in Artificial Intelligence*, **8082** (2013), 560–567.

[4] APACHE OPENNLP, The Apache Software foundation; 2010. URL `http://opennlp.apache.org/`

[5] B. BEKAVAC, Strojno prepoznavanje naziva u suvremenim hrvatskim tekstovima. PhD thesis, University of Zagreb, 2005.

[6] B. BEKAVAC, M. TADIĆ, Implementation of Croatian NERC System. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing (BSNLP 2007)*, ACL; (2007), pp. 11–18.

[7] N. CHINCHOR, P. ROBINSON, MUC-7 Named Entity Task Definition. In *Proceedings of the 7th Conference on Message Understanding (MUC 7)*; 1997.

[8] A. CLARK, Combining Distributional and Morphological Information for Part of Speech Induction. In *Proceedings of the 10th Annual Meeting of the European Association for Computational Linguistics (EACL 2003)*, EACL; (2003), pp. 59–66.

[9] H. DAUMÉ III, Frustratingly Easy Domain Adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL 2007)*, ACL; (2007), pp. 256–263.

[10] J. R. FINKEL, T. GRENAGER, C. MANNING, Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, ACL; (2005), pp. 363–370.

[11] G. GLAVAŠ, M. KARAN, F. ŠARIĆ, J. ŠNAJDER, J. MIJIĆ, A. ŠILIĆ, B. DALBELO BAŠIĆ, CroNER: A State-of-the-Art Named Entity Recognition and Classification for Croatian. In *Proceedings of the 8th Language Technologies Conference (IS-LTC 2012)*, Jožef Stefan Institute, Ljubljana, Slovenia; (2012), pp. 73–78.

[12] R. GRISHMAN, B. SUNDHEIM, Message Understanding Conference 6: A brief history. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING 1996)*, The COLING Organizing Committee; (1996), pp. 466–471.

[13] V. JIJKOUN, M. ALAM KHALID, M. MARX, M. DE RIJKE, Named Entity Normalization in User Generated Content. In *Proceedings of the 2nd Workshop on Analytics for Noisy Unstructured Text Data (AND 2008)*, ACM; (2008), pp. 23–30.

[14] T. KUDO, CRF++: Yet Another CRF Toolkit; 2005. URL `https://code.google.com/p/crfpp/` (accessed 2013-07-18)

[15] N. LJUBEŠIĆ, T. ERJAVEC, hrWaC and slWac: Compiling Web Corpora for Croatian and Slovene. In *Text, Speech and Dialogue, Lecture Notes in Computer Science*, **6836** (2001), 395–402, Springer.

[16] N. LJUBEŠIĆ, M. STUPAR, T. JURIĆ, Building Named Entity Recognition Models for Croatian and Slovene. In *Proceedings of the 8th Language Technologies Conference (IS-LTC 2012)*, Jožef Stefan Institute, Ljubljana, Slovenia; (2012), pp. 129–134.

[17] N. LJUBEŠIĆ, M. STUPAR, T. JURIĆ, Ž. AGIĆ, Combining Available Datasets for Building Named Entity Recognition Models of Croatian and Slovene. *Slovenščina 2.0: empirical, applied and interdisciplinary research*, (2013), in press.

[18] S. MILLER, J. GUINNESS, A. ZAMANIAN, Name Tagging with Word Clusters and Discriminative Training. In *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: Proceedings of the Main Conference*, ACL, (2013), pp. 337–342.

[19] D. NADEAU, S. SEKINE, A Survey of Named Entity Recognition and Classification. *Lingvisticae Investigationes*, **30**(1) (2007), 3–26.

[20] N. OKAZAKI, CRFsuite: A Fast Implementation of Conditional Random Fields (CRFs); 2007. URL `http://www.chokkan.org/software/crfsuite/` (accessed 2013-07-18)

[21] L. F. RAU, Extracting Company Names from Text. In *Proceedings of the 7th IEEE Conference on Artificial Intelligence Applications, IEEE*, (1991), pp. 29–32.

[22] A. RITTER, S. CLARK, M. ETZIONI, O. ETZIONI, Named Entity Recognition in Tweets: An Experimental Study. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP 2011, ACL*, (2013), pp. 1524–1534.

[23] E. F. TJONG KIM SANG, F. DE MEULDER, Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 (CONLL 2003), ACL*, (2003), pp. 142–147.

[24] M. SILBERZTEIN, Text Indexing with INTEX. *Computer and the Humanities*, **33**(3) (1999), Kluwer.

[25] A. SØGAARD, *Semi-Supervised Learning and Domain Adaptation in Natural Language Processing*. Morgan & Claypool Publishers, 2013.

[26] M. TADIĆ, D. BROZOVIĆ-RONČEVIĆ, A. KAPETANOVIĆ, *The Croatian Language in The Digital Age*. Springer, 2012.

Received: July, 2013
Revised: September, 2013
Accepted: September, 2013

Contact addresses:

Željko Agić
Department of Information and Communication Sciences
Faculty of Humanities and Social Sciences
University of Zagreb
Ivana Lučića 3
10000 Zagreb
Croatia
e-mail: zagic@ffzg.hr

Božo Bekavac
Department of Linguistics
Faculty of Humanities and Social Sciences
University of Zagreb
Ivana Lučića 3
10000 Zagreb
Croatia
e-mail: bbekavac@ffzg.hr

ŽELJKO AGIĆ received his BSc (2005) in computer science and engineering from the University of Split and his PhD (2012) in information and communication science from the University of Zagreb. He is currently a postdoctoral researcher at the Department of Information and Communication Sciences, Faculty of Humanities and Social Sciences, University of Zagreb. His research interests lie in the field of language technology.

BOŽO BEKAVAC received his BA (1997) in linguistics and information science, MA (2001) and PhD (2005) in linguistics from the Faculty of Humanities and Social Sciences, University of Zagreb. He is currently an assistant professor at the Department of Linguistics, Faculty of Humanities and Social Sciences, University of Zagreb. His scientific work focuses on corpora and computational tools for processing of Croatian.