Domain Dependence of Statistical Named Entity Recognition and Classification in Croatian Texts

Željko Agić[1], Božo Bekavac[2]
[1]*Department of Information and Communication Sciences*
[2]*Department of Linguistics*
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, HR-10000 Zagreb
{zeljko.agic, bbekavac}@ffzg.hr

Abstract. *Influence of text domain selection on statistical named entity recognition and classification in Croatian texts is investigated. Two datasets of Croatian newspaper texts of differing text domains were manually annotated for named entities and used for training and testing the Stanford NER system for named entity recognition based on sequence labeling with CRF. State of the art scores were observed in both domains. A strong preference for systems trained on mixed text domains is established by the experiment. The top-performing system was recorded with an overall F1-score of 0.876 on mixed-domain test sets, scoring 0.899 in one of the selected domains and 0.852 in the other. The single best domain F1-scores were recorded at 0.910 and 0.858.*

Keywords. text domain, domain dependence, named entity recognition, Croatian language

1. Introduction

Named entity recognition and classification (NER, NERC) is a very well-established task in natural language processing. It involves detecting and classifying named entities – names of persons, locations and organizations, along with temporal and monetary expressions, depending on the exact task-dependent specification – in natural language text.

As detecting names of people, locations and organizations is expectedly an important subtask of information extraction – as these entities are basically where natural language text and physical reality intersect – the task has been thoroughly researched, especially for English. As elaborated in [7], from task definition at the MUC-6 conference [8] and the CoNLL 2003 shared task on named entity recognition [11] to publicly available state-of-the-art systems as Stanford NER [10], detecting entities in English texts has matured, both performance-wise and scalability- and integration-wise.

Until recently, named entity detection in Croatian texts has only been addressed by a rule-based approach using finite state transducer cascades [3]. Just recently, two systems were presented that address Croatian NERC by a statistical approach using sequence labeling with conditional random fields (CRF) [7, 9]. 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 [3] uses the MUC-7 specification, CroNER [7] extends it with five additional named entity classes and finally adds one of them – namely, the *ethnic* class – into the respective models, while Stanford NER models of [9] annotate names of persons, locations and organizations and introduce a class for miscellaneous entities. All three systems are developed and tested in the general domain of Croatian newspaper text and, therefore, none of them explicitly addresses the influence of dataset alterations on system performance.

The experiment presented here attempts to isolate and address this single issue specifically and to provide datasets for joint testing of Croatian NERC systems. A dataset of Croatian newspaper texts is collected and manually annotated for named entities in the three overlapping and most frequently used entity classes – names of locations, organizations and persons, i.e., the MUC-7 ENAMEX classes. The dataset encompasses two general newspaper text domains: internal affairs and other texts, the latter one consisting of texts from cultural, sports, lifestyle and unclassified news domain. NERC models are trained and tested on texts from the dataset by using the Stanford NER tagger [10] in order to establish dependency relations between the training data and respective NERC model properties on one side and named

entity detection scores across different domains on the other side.

In the following sections, the datasets and the experiment setup are elaborated, obtained named entity detection results are analyzed and future work plans are briefly sketched, with special emphasis on the need for domain-specific testing of NERC systems.

2. Experiment setup

The texts in the dataset were collected from the Vjesnik newspaper. The collection was done by a custom crawler and the texts were further cleansed, sentence-delimited and tokenized by using Apache OpenNLP tools [2] 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 five expert annotators. Annotations were not overlapped and thus inter-annotator agreement was not observed as high agreement on these classes was expected, following what was previously observed in the process of developing the CroNER system [7], where the average inter-annotator agreement on ENAMEX was shown to be approximately 95%. Dataset stats are given in Table 1.

Table 1. Dataset stats

domain	*sent's*	*tokens*	*NEs*	*NE tok's*
internal	13.209	346.886	21.752	31.809
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
lifestyle	1.810	43.222	1.342	2.129
foreign	1.999	39.979	2.262	3.508
total	20.861	515.792	30.695	45.665

The data collection is split into two main text domains: (1) internal affairs or internal politics and (2) other text domains, evenly distributed between culture, foreign affairs and other news, lifestyle and sports. Table 1 shows that there is ca 347 kw in the internal affairs domain and ca 169 kw of text in other domains. Other features in Table 1 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 sentence lengths differ as an average sentence from the internal affairs domain holds 26 words, while there are 22 words on average in other domains. The token to named entity token ratio also slightly differs between domains. 9.17% 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 sub-domains of the latter domain, with culture and lifestyle texts having longer sentences and lifestyle texts having as few as 4.93% named entity tokens.

The dataset is split into training and testing sets for five-fold cross-validation by random sampling and respecting document boundaries. The split was done separately for the internal affairs domain and for the other texts domain. Approximately 50 kw was left out for testing purposes for each of the two domains, leaving out a test set of approximately 100 kw, evenly spread between five internal affairs samples of 10 kw each and five other texts samples of 10 kw each. A mixed test set of ten 10 kw samples was also created, randomly combining sentences from these two domain test sets. The remaining 300 kw of internal affairs texts and 100 kw of other domains texts was used for training the NERC models. Two batches of experiments were designed.

The first batch was used to establish the optimal feature set for NERC models and to investigate the functional dependency of training set size and named entity detection accuracy. It was done by using only the internal affairs texts for training the models. The following features were investigated: tokens, part-of-speech (POS) annotation, morphosnytactic (MSD) annotation and distributional similarity features which were calculated and shared freely in [9] by utilizing the clustering tool described in [5] over a 100 Mw sample of Croatian from the hrWaC corpus [6] with 400 clusters. 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} × {plain text, POS, MSD} × {no distsim, distsim}. Precision, recall and F1-scores were calculated for the three named entity classes and overall on all available test sets.

In the second testing batch, the internal affairs training sets from the first batch are 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 feature set established in the first experiment batch is used. The training set sizes are 50 kw (internal) + 50 kw (other), 100 kw (internal) +

100 kw (other) and 200 kw (internal) + 100 kw (other). The 100 kw training set containing only the text from other domains was also tested. The results of detection using the mixed models are compared with those of the internal-only models from the first batch that used the same feature set.

3. Results and discussion

The results of the first experiment batch in terms of F1-scores achieved by Stanford NER models trained on internal affairs training sets and tested on all test sets are given in Table 2 with training set sizes and features used.

Table 2. Overall F1-scores for in-domain models

in-domain test set, without distsim			
kw	***plain***	***POS***	***MSD***
100	0.858	0.864	0.868
200	0.889	0.890	0.893
300	0.899	0.897	0.902
in-domain test set, with distsim			
100	0.876	0.879	0.877
200	0.900	0.900	0.898
300	0.910	0.906	0.908
out-of-domain test set, without distsim			
100	0.561	0.597	0.610
200	0.586	0.606	0.624
300	0.598	0.622	0.632
out-of-domain test set, with distsim			
100	0.621	0.646	0.661
200	0.644	0.663	0.673
300	0.664	0.674	0.686
mixed-domain test set, without distsim			
100	0.709	0.731	0.739
200	0.737	0.748	0.758
300	0.748	0.759	0.767
mixed-domain test set, with distsim			
100	0.748	0.763	0.769
200	0.772	0.781	0.785
300	0.787	0.790	0.797

Overall F1-scores from Table 2 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 F1-score on the internal affairs test set (in-domain 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 text domains test set (out-of-domain test set) is the 300 kw MSD and distsim model with an F1-score of 0.686. This is also reflected in the mixed-domain test scenario, where the same 300 kw MSD distsim system scored an overall F1-score of 0.797. It should be noted that these small differences between the in-domain scores are mostly not statistically significant.

Models using distributional similarity features consistently outperform the respective models without these features. The overall difference in F1-scores between these two groups of models increases with the complexity of the test set: from less than 1% increase for the in-domain test set to 5% increase for the out-of-domain scenario which is in turn reflected in the mixed-domain test set increase of ca 3% in favor of the models using distributional similarity features.

Domain selection influence on overall named entity detection accuracy is substantial. F1-score decrease of 0.224 is observed between the best NERC systems when comparing in-domain and out-of-domain test sets. 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 in-domain models does not provide a substantial increase in F1-scores for in-domain texts, the out-of-domain scores clearly benefit both from increasing the training set size and from adding POS, MSD and distributional similarity features.

Functional dependencies of training set sizes, selected training features and overall system performances are additionally illustrated by Figure 1. It clearly indicates both the strength of feature selection influence and the learning rates for the in-domain models in the in-domain and out-of-domain test scenario. The groupings of learning curves also illustrate the significance of differences in the results. Moreover, the learning curves indicate that the in-domain 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.

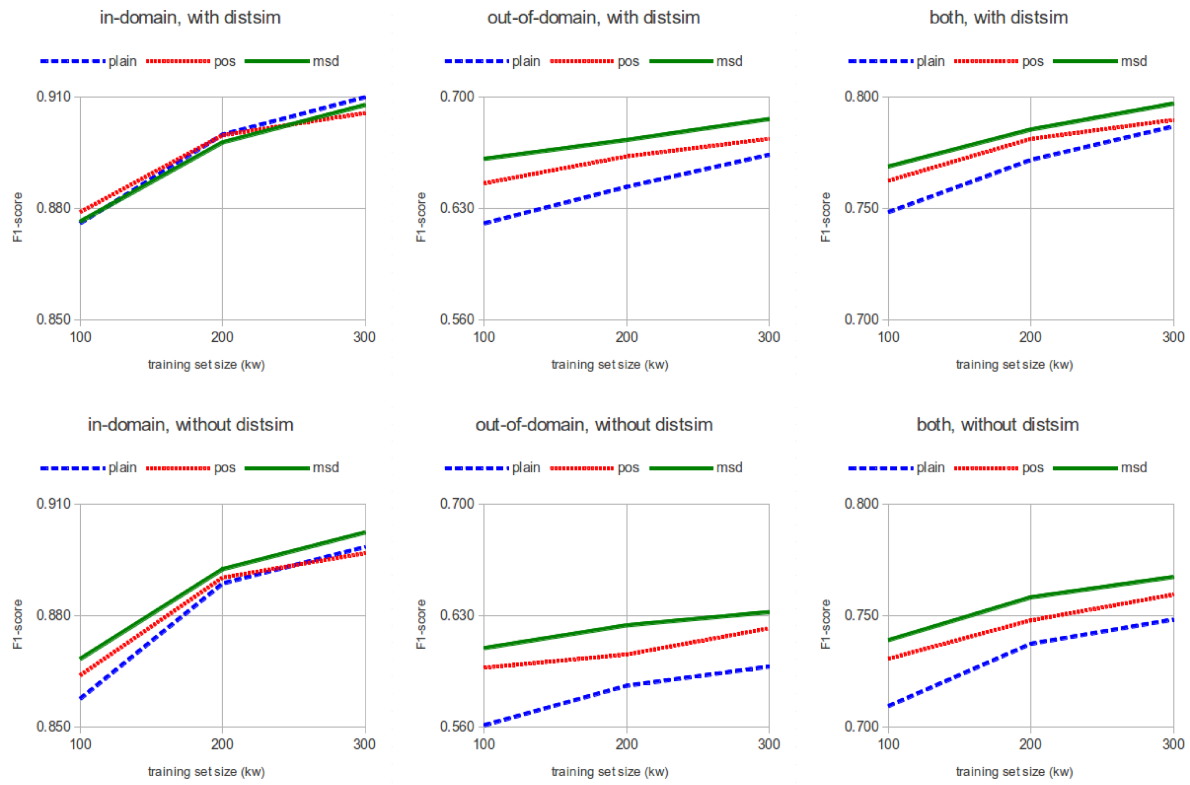

Figure 1. Learning curves for in-domain models

Statistical significance exploration using t-test 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 smaller and faster to train and use, they are further observed in more detail in the experiment batch 1 discussion. For the same reasons, the POS and distsim feature set is the only feature set used in the second batch of experiments.

Regarding the observed absence of statistical significance of differences between the F1-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 ca 85% for 20% unknown words [1]. 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.

Table 3 shows the overall F1-scores achieved by the selected POS and distsim NER models for the three MUC-7 ENAMEX classes (location, organization, person) in all three test sets. What was previously observed for overall F1-scores is decomposed in Table 3 into three named entity classes and their respective F1-scores follow a similar pattern of difference between in-domain and out-of-domain data.

Perhaps most notably, the organization class F1-score difference for the top-performing systems is shown to be 0.415, which is considered to be a substantial decrease. The systems are consistently better at detecting names of people (0.857 in mixed-domain) than location names (0.805) and organization names (0.648). This 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. Complementing Table 1, taking the out-of-domain dataset as an example, there are 3.079 locations (3.858

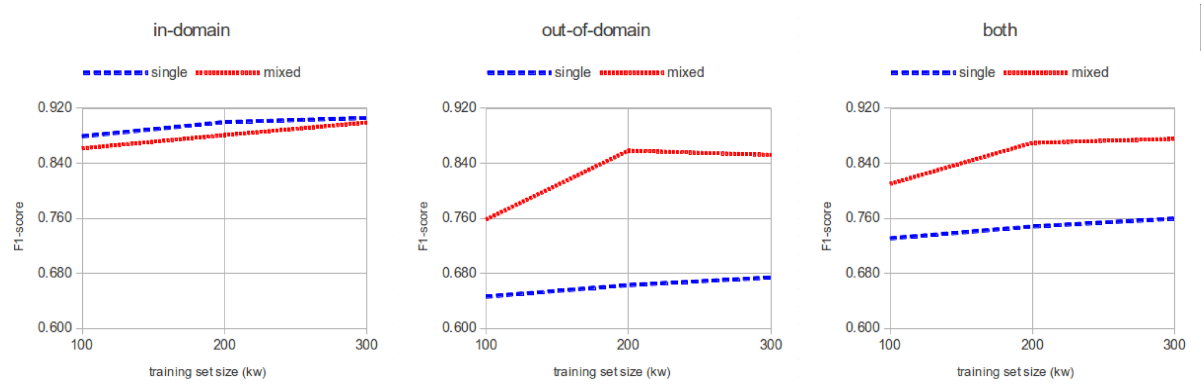

Figure 2. Learning curve comparison for in-domain and mixed-domain models

location tokens), 4.031 persons (6.733 tokens) and 1.833 organizations (3.265 tokens). 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.

The second batch of experiments included creating mixed-domain Stanford NER models by combining in-domain and out-of-domain training sets. On all models from the second batch, POS and distributional similarity features are used. The mixed-domain models are compared with the respective in-domain models. Table 4 shows their overall F1-scores in comparison. 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 training set, 200 kw equals the 100 kw + 100 kw mix and 300 kw amounts to 200 kw of in-domain text and 100 kw of out-of-domain text.

Table 3. F1-scores on ENAMEX classes for the models using POS and distsim

in-domain test set			
kw	***LOC***	***ORG***	***PER***
100	0.869	0.824	0.946
200	0.902	0.844	0.957
300	0.905	0.855	0.960
out-of-domain test set			
100	0.679	0.397	0.723
200	0.704	0.406	0.742
300	0.705	0.440	0.753
mixed-domain test set			
100	0.774	0.611	0.834
200	0.803	0.625	0.849

Table 4 shows 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 accuracy between the best in-domain model and the top-performing mixed-domain model is not significant for the in-domain test sets, while being substantially significant – with an F1-score difference of 0.178 and 0.117 – for the out-of-domain and the mixed-domain test sets. The top-performing mixed-domain model still maintains the state-of-the-art accuracy of 0.899 in the in-domain test scenario and 0.876 overall, i.e., for the mixed-domain test.

Table 4. Overall F1-scores for in-domain and mixed-domain POS and distsim models

in-domain test set		
kw	***in-domain***	***mixed model***
100	0.879	0.862
200	0.900	0.881
300	0.906	0.899
out-of-domain test set		
100	0.646	0.758
200	0.663	0.858
300	0.674	0.852
mixed-domain test set		
100	0.731	0.810
200	0.748	0.870
300	0.759	0.876

Table 4 is complemented with learning curves for in-domain and mixed-domain models in Figure 2 on all test sets. The significance of observed differences in their F1-scores is clearly

indicated. 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 out-of-domain texts as well as a local inflection point. This observation could be further investigated by enlarging the training set, providing additional out-of-domain texts.

Table 5. F1-scores on ENAMEX classes for the mixed-domain POS and distsim models

in-domain test set			
kw	***LOC***	***ORG***	***PER***
100	0.851	0.795	0.942
200	0.874	0.816	0.955
300	0.895	0.841	0.964
out-of-domain test set			
100	0.789	0.565	0.810
200	0.874	0.745	0.893
300	0.869	0.733	0.890
mixed-domain test set			
100	0.820	0.680	0.876
200	0.874	0.781	0.924
300	0.882	0.787	0.927

Table 5 is a breakdown of the overall scores of the mixed-domain models into three named entity classes, similar to what Table 3 provided for the in-domain models and it follows a similar pattern. The best mixed-domain model (300 kw) is significantly better at detecting personal and organizational names in the in-domain test set than on the out-of-domain test set – with overall difference of 0.074 and 0.108 in F1-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.

4. Conclusions and future work

In this contribution, text domain dependence of statistical named entity recognition and classification in Croatian texts was investigated. A strong preference for models trained on mixed domain text was observed, where state-of-the-art accuracy in terms of overall F1-scores and F1-scores on all ENAMEX categories (detecting personal names, names of locations and names of organizations) was observed in all test scenarios.

Future work plans include enlarging the used datasets and introducing datasets for other text genres and domains. Experiments are underway which include comparing the models presented in this experiment with other publicly available NERC systems for Croatian. A subset of used Stanford NER models and domain-specific datasets for testing Croatian NERC is made available (http://zeljko.agic.me/resources/).

5. References

[1] Agić Ž, Tadić M, Dovedan Z. (2008). Improving Part-of-Speech Tagging Accuracy for Croatian by Morphological Analysis. Informatica. 32:4, pp. 445-451.

[2] Apache OpenNLP. The Apache Software foundation, URL http://opennlp.apache.org/

[3] Bekavac B. (2005). Strojno prepoznavanje naziva u suvremenim hrvatskim tekstovima. PhD thesis, University of Zagreb,.

[4] Bekavac B, Tadić M. (2007). Implementation of Croatian NERC System. In Proceedings of BSNLP, ACL, pp. 11-18.

[5] Clark A. (2003). Combining Distributional and Morphological Information for Part of Speech Induction. In Proceedings of EACL.

[6] Ljubešić N, Erjavec T. (2011). hrWaC and slWac: Compiling Web Corpora for Croatian and Slovene. In TSD 2011, Springer.

[7] Glavaš G, Karan M, Šarić F, Šnajder J, Mijić J, Šilić A, Dalbelo Bašić B. (2012). CroNER: A State-of-the-Art Named Entity Recognition and Classification for Croatian. Proceedings of IS-LTC, pp. 73-78.

[8] Grishman R, Sundheim B. (1996). Message Understanding Conference 6: A brief history. In Proceedings of COLING, pp. 466-471.

[9] Ljubešić N, Stupar M, Jurić T. (2012). Building Named Entity Recognition Models for Croatian and Slovene. In Proceedings of IS-LTC, pp. 129-134.

[10] Finkel J R, Grenager T, Manning C. (2005). Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In Proceedings of ACL 2005, pp. 363-370.

[11] Tjong Kim Sang E F, De Meulder F. (2003). Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In Proceedings of HLT-NAACL, ACL, pp. 142-147.