# COMBINING AVAILABLE DATASETS FOR BUILDING NAMED ENTITY RECOGNITION MODELS OF CROATIAN AND SLOVENE

Nikola LJUBEŠIĆ, Marija STUPAR, Tereza JURIĆ, Željko AGIĆ
Department of Information and Communication Sciences, Faculty of Humanities and Social Sciences, University of Zagreb

*Ljubešić, N., Stupar, M., Jurić, T., Agić, Ž. (2013): Combining Available Datasets for Building Named Entity Recognition Models of Croatian and Slovene. Slovenščina 2.0, 1 (2): 35–57.*

URL: http://www.trojina.org/slovenscina2.0/arhiv/2013/2/Slo2.0_2013_2_03.pdf.

The paper presents efforts in developing freely available models for named entity recognition and classification in Croatian and Slovene text. Our experiments focus on the most informative set of linguistic features taking into account the availability of language tools and resources for the languages in question. Besides the classic linguistic features, distributional similarity features calculated from large unannotated monolingual corpora are exploited as well. We performed two batches on experiments, the first one on a self-built dataset on which the optimal set of features is sought, and a second batch with additional, much larger datasets obtained at a later point on which we verify the findings from the first batch. On the initial dataset using distributional information improves the results for 7-8 points in F1 while adding morphological information improves the results for additional 3-4 points in both languages. The second batch of experiments shows that morphosyntactic and distributional information lose importance as the dataset size significantly increases. The best performing models that use distributional information only, along with test sets for comparison with existing and future systems are made publicly available for both academic and non-academic use.

**Keywords:** named entity recognition, distributional similarity, Croatian language, Slovene language

## 1 INTRODUCTION

Named entity recognition and classification (NERC), nowadays often called

just named entity recognition (NER) is a subtask of the information extraction task. It aims to locate and classify text elements into predefined categories, and is regularly applied on more complex natural language processing problems, using statistical or rule-based models. State-of-the-art systems tend to be open-domain and language independent.

This paper presents efforts in creating NER models for Croatian and Slovene language, available for free academic and non-academic use. Besides performing initial experiments on datasets developed from our side, we experiment with multiple recently published datasets for both languages as well. Thereby we receive a clearer picture of the underlying phenomena and manage to publish models of greater robustness and higher accuracy than those built on our previous datasets only.

The tool we are using to build the models is the Stanford Named Entity Recognizer (StanfordNER), nowadays a frequently used tool for NER. It is an implementation of Conditional Random Fields (CRF) sequence models and is available under GNU GPL license and free for academic use (Finkel et al. 2005). Besides many feature extractors that come with this tool, it is designed to work with the clustering method proposed by (Clark 2003) which combines standard distributional similarity with morphological similarity to cover infrequent words for which distributional information alone is unreliable.

This paper is structured as follows: in Section 2 we give an overview of related work, in Section 3 we present the datasets used in our research. Section 4 gives the experimental setup, section 5 the results of our initial experiments and section 6 the results of the experiments on additional datasets. We lay out the main conclusions in section 7.

## 2 RELATED WORK

To our knowledge, there has been some effort in developing NER systems for South Slavic languages which were mostly rule-based. New statistic based

approaches have recently emerged.

A rule-based system for Croatian described in (Bekavac 2005) uses regular grammars for recognition and classification of names over annotated texts. The system contains a module for sentence segmentation, lexicon of common words, specialized lists of names and transducers for automatic recognition of certain word forms.

A statistical approach described in (Bošnjak 2007) uses a semi-supervised method based on lists of names and entity extraction system.

For Serbian a rule-based system (Vitas, Pavlović-Lažetić 2008) based on lexical recognition is developed. The authors point out certain differences between English and Serbian language that make the task of building a successful system for Serbian challenging, as well as all the other Slavic languages which require a more thorough preparation of the system due to rich inflection.

None of the presented systems are available for academic usage which hinders researchers investigating tasks that require NER as a preprocessing step. One of the main intentions of our research is to improve this situation. In the process of building a good NER system, features are considered as important as the selection of machine learning algorithms. The aim is to find an optimal set of features that will ensure the highest system accuracy with minimum complexity in classifier building. Several NER approaches use a very large number of features (Mayfield et al. 2003), but the inclusion of additional features after a certain point can even yield worse results.

In this paper we use Stanford NER property files obtained in our previous research (Filipić et al. 2012) along with the findings about best performing settings for Croatian and Slovene which include POS, MSD and distributional similarity features. The only work we are aware of that examines the usage of distributional features in Stanford NER is (Faruqui, Padó 2010). The paper describes the process of building and optimizing NER models for German and

by using distributional features F1 is improved for 6% in-domain and 9% out-of-domain.

The latest approach in Croatian NER research, called CroNER (Glavaš et al. 2012), is based on supervised learning using CRF. Observed classes include personal names, ethnics, percentages, locations, organizations, dates, and monetary and temporal expressions. The analysis of the Vjesnik corpus containing 310,000 tokens, reaches (exact) F1 measure of 87.42%. The system uses gazetteer-based features for personal names, ethnicity, city, state, street and organization names, and a rich set of lexical and morphological features specific for Croatian. Authors also defined a special named entity type to cover instances of possessive adjectives. According to the authors, two different methods for document-level consistency of NE labels are implemented: post-processing rules (hard consistency constraint) and a two-stage CRF (soft consistency constraint). Post-processing rules are hand-crafted patterns designed to extract or re-label named entities omitted or misclassified by the CRF model. Two-stage CRF aims to consolidate NE label predictions on document and corpus level by employing a second CRF model that uses features computed from the output of the first CRF model.

NER model for Slovene (Štajner et al. 2012), developed using Mallet (McCallum 2002), also implements a CRF supervised learning algorithm. Research has been made on Slovene SSJ500k[1] corpus annotated with morphosyntactic tags and three named entity classes. It has shown that the inclusion of morphosyntactic tag features benefits named entity extraction. The system reaches precision of 77% and recall of 76%, having stronger performance on personal and geographical named entities than on other entities, since the class of other entities (everything not being person or location) is very diverse and difficult to predict.

[1] http://www.slovenscina.eu/tehnologije/ucni-korpus

## 3 CORPORA

Two initial datasets used in the first batch of our experiments, one Croatian (HR) and one Slovene (SL) were built during a student project (Filipić et al. 2012) from data taken from specific Internet domains from the Croatian and Slovene web corpora hrWaC and slWaC (Ljubešić, Erjavec 2011). The Croatian corpus (HR) contains 59,212 tokens taken from four different Internet domains covering two general newspaper portals, nacional.hr and jutarnji.hr, one ICT portal bug.hr and the business news portal poslovni.hr. The data was annotated during a student project in which data diversity was given special emphasis. The Slovene corpus (SL) is at almost two thirds the size of the Croatian one, containing 37,032 tokens from one general news portal rtvslo.si. While selecting the Slovene data the main goal was to build a usable dataset with limited annotation capacities.

We obviously live in exciting times for natural language processing in both Croatia and Slovenia because after finishing our initial batch of experiments, three additional datasets – two for Croatian and one for Slovene – have emerged, all published under quite permissive licenses.

The additional Croatian datasets include the SETimes and Vjesnik corpora. SETimes is a newspaper domain corpus consisting of general news articles written in Croatian language, originally extracted from the "Southeast European Times" web portal. It contains 178,982 tokens and has the highest density of named entities in the text. The Vjesnik corpus contains a collection of two main text domains – internal affairs and other text domains, evenly distributed between culture, foreign affairs and other news, lifestyle and sports. The Corpus contains 104,494 tokens. Text collection was performed by using a custom crawler, texts were further processed, i.e., cleaned, sentence split and tokenized by using Apache OpenNLP[2] tools and POS/MSD annotated using the CroTag MSD-tagger (Agić et al. 2008). Related

[2] http://opennlp.apache.org/

experiments with Croatian NER using the Vjesnik corpus are described in (Agić, Bekavac 2013).

The additional resource obtained for Slovene language is a part of the SSJ500k corpus. It is available under the Creative Commons CC-BY-NC-SA license. It is manually lemmatized and morphosyntactically tagged, and in part dependency parsed and annotated for named entities. Named entities are classified into four classes: names of persons, locations and organizations and other entities. In our research we use only the part of the corpus annotated with named entities which is 118,609 tokens in size. The amount of data in all datasets for both languages is given in Tables 1 and 2.

All corpora were tagged using the IOB2 standard following the CoNLL 2003 annotation guidelines3 where each row represents a token in the text with its linguistic annotation and designated predefined named entity category. IOB2 labels show whether a word is at the beginning (B), inside (I) or outside (O) of a named entity. Both initial datasets (HR and SL) were annotated with the four traditional categories – location (LOC), organization (ORG), person (PERS) and miscellaneous (MISC). The additional Slovene dataset (SSJ) contains the same three categories while the two additional datasets in Croatian (SETimes and Vjesnik) have only the basic three categories annotated – location (LOC), organization (ORG) and person (PERS). Possessive adjectives indicating named entities are additionally annotated in the SETimes, Vjesnik and SSJ datasets as it is the case in the initial HR and SL datasets (Filipić et al. 2012).

Basic part-of-speech (first letter of the Multext-East MSD) (Erjavec et al. 2003) on the HR corpus was manually annotated since related work shows that these features are useful for the task. Slovene SL corpus was MSD tagged and lemmatized with the freely available ToTaLe tagger (Erjavec et al. 2005) trained on JOS corpus data (Erjavec et al. 2010).

[3] http://www.cnts.ua.ac.be/conll2003/ner/annotation.txt

| | HR | SETimes | Vjesnik |
|---|---|---|---|
| **Token #** | 59,212 | 178,982 | 104,494 |
| **ORG** | 839 | 4,686 | 1,875 |
| **PERS** | 602 | 3,761 | 2,317 |
| **LOC** | 590 | 5,746 | 2,055 |
| **MISC** | 632 | - | - |
| **ALL** | 2,663 | 14193 | 6247 |
| **Density** | 0.045 | 0.0793 | 0.0598 |

**Table 1:** Size of the Croatian corpora and the number of annotated named entities.

| | SL | SSJ |
|---|---|---|
| **Token #** | 37,032 | 118,609 |
| **ORG** | 311 | 804 |
| **PERS** | 1,086 | 2,008 |
| **LOC** | 716 | 1,284 |
| **MISC** | 378 | 406 |
| **ALL** | 2,491 | 4,502 |
| **Density** | 0.067 | 0.038 |

**Table 2:** Size of the Slovene corpora and the number of annotated named entities.

To be able to use POS information on unseen Croatian data, we trained a model for the HunPos tagger (Halácsy et al. 2007) from the initial Croatian dataset. We performed a simple test of the resulting model by dividing the dataset into training and test set with a 9:1 ratio. Accuracy obtained on the test set was 95.1%. We publish the tagger trained on all available data along with the NER models and the benchmark datasets. To our knowledge, this is the first freely available part-of-speech tagger for Croatian.4

The expected difference in diversity of the initial datasets can be clearly observed from the amount of annotated named entities for each corpus. First of all, although the SL corpus has 37% less textual material than HR, it has just 6% less named entities showing a higher density of named entities one

[4] A full MSD tagger and lemmatizer were developed and published recently on http://nlp.ffzg.hr/resources/models/.

would expect from a straightforward newspaper dataset. Furthermore, when we look at the type of named entities, we can observe that the SL dataset contains many more person names and slightly more locations while the Croatian dataset contains more organization names and named entities labelled with the miscellaneous category. This data confirms our assumption that the Croatian dataset is much more diverse and will thereby present a harder task for supervised classification in the initial experiments.

The additional Croatian datasets are newspaper datasets and show a non-surprising high named entity density in the text. The density is even higher than the one in the initial Croatian dataset regardless of the fact that additional datasets do not contain the miscellaneous category.

The additional Slovene dataset is rather named-entity-sparse having half of the initial Slovene dataset density. This should not come as a surprise since the SSJ dataset is a reference corpus and therefore does not contain only newspaper texts as all other datasets do.

We divided the initial corpora into development and test sets by shuffling documents and producing test sets of similar size for both languages. The decision to build test sets of similar size was guided by the idea of publishing those test sets as benchmark datasets for both languages. For that reason the HR development set contains 53,142 tokens while the SL one contains 29,686 tokens, i.e. 56% of the amount of Croatian data.

An additional insight into the features and thereby specificities of the two initial datasets is given by calculating vocabulary transfer between identical portions of development and test sets. The numbers are given in Table 3. Vocabulary transfer is calculated as token and type percentage of named entities in the test set being already present in the development set. Two interesting properties can be observed here. First of all, the Slovene vocabulary transfer is higher than the Croatian one pointing at the expected lower content diversity of Slovene data. Secondly, there is almost no

difference between token and type transfer on Croatian data showing that the diversity of named entities is really high. Namely, this points to the fact that almost none of the named entities from the development set present in the test set appears more than once in the Croatian test set which is not the case in Slovene data.

| Corpus | Token transfer | Type transfer |
|---|---|---|
| **HR** | 10.7% | 10.6% |
| **SL** | 17.3% | 12.4% |

**Table 3:** Vocabulary transfer for initial corpora on identical portions of development and test set.

We built additional test sets once we obtained the additional datasets by adding similar amount of information from each dataset to the joint test set. The Croatian test set includes 6,730 tokens from the Vjesnik corpus, and 6,736 tokens from the SETimes corpus along with the previously mentioned initial HR test set. Since the MISC category is not present in the additional Croatian datasets, the extended Croatian dataset naturally does not contain that category. Slovene additional test set contains 6,981 tokens from SSJ corpus, along with the initial SL test dataset. The MISC category is retained in that test set since both datasets contain it.

For calculating distributional similarity of tokens from large monolingual corpora, portions of hrWaC and slWaC web corpora were used. For Croatian we built a 100Mw corpus and for Slovene a 50Mw corpus, both containing data from large news portals.

## 4 EXPERIMENTAL SETUP

Since different annotation levels on initial Croatian and Slovene datasets were available, in the first batch of experiments we evaluated different settings for each language on the HR and SL corpora. Besides part-of-speech information for both languages, on SL data MSD and lemma information was present as well.

On HR data we experimented with POS information (”POS”), distributional information (”DISTSIM”) calculated from 10Mw, 50Mw and 100Mw corpora while on Slovene data we experimented with POS, MSD (”MSD”) and lemma (”LEMMA”) information and distributional information obtained from 10Mw and 50Mw corpora. Thereby we performed 8 initial experiments on HR data and 11 initial experiments on SL data (we eliminated the experiments varying with availability of lemma information once it proved to be non-informative).

All the experiments were performed on development sets of both datasets via 5-fold cross-validation that takes into account document borders. By respecting document borders we were trying to keep the vocabulary transfer as low as possible and thereby obtain the most realistic results, i.e., differences between different experimental settings. Distributional similarity was calculated by using Clark's cluster_neyessen tool (Clark 2003) with default settings (numberStates=5, frequencyCutoff=5, iterations= 10). The number of resulting clusters was set on best-performing values in (Faruqui, Padó 2010), i.e., for 10Mw corpora 100 clusters and for 50Mw and 100Mw corpora 400 clusters were built. First twenty elements of example clusters calculated from the Croatian 100Mw and Slovene 50Mw corpora are given in Table 4. The Croatian cluster contains exclusively country and city names in the locative (or dative) case. The Slovene cluster contains first names of people in the nominative case of both Slovene and English origin. These examples show very clearly how the cluster ID can be used as a very informative feature in the supervised training procedure.

After identifying the best performing settings on the development sets we calculated our final results by training a system on the whole development set and testing it on the left-out initial test set.

| |
|---|
| njemačkoj rijeci londonu sarajevu osijeku italiji zadru francuskoj haagu austriji parizu dubrovniku vukovaru španjolskoj milanu bruxellesu rimu beču moskvi berlinu |
| tomaž simon goran martina dejan jan nina tom saša mojca vesna jurij eva nataša maria jernej daniel richard thomas damjan žiga |

**Table 4:** First 20 elements of sample clusters obtained with Clark's tool on the 100Mw Croatian and 50Mw Slovene corpus.*

* The Croatian cluster contains exclusively country and city names, and the Slovene cluster contains first names of people of both Slovene and English origin.

Obtaining additional datasets for both languages at a later point enabled us to perform an additional batch of experiments and re-examine our findings in the initial experiments. We built additional test sets containing left-out information from all datasets as described in the previous section. We performed calculations on the few most promising settings from the initial batch of experiments. The important difference between the two languages in the second batch of experiments is that the Croatian dataset does not contain the miscellaneous category while the Slovene dataset does.

Finally we compared results obtained with different amounts of annotated data on both languages with the best performing settings to identify the gain we can expect from adding more annotated data.

## 5 RESULTS OF THE FIRST BATCH OF EXPERIMENTS

The results obtained by 5-fold cross-validation on both development sets are presented for Croatian in Figure 1 and for Slovene in Figure 2. The results of each cross-validation are averaged by calculating their harmonic mean. Regarding the statistical significance of the results, we perform a one-tailed paired t-test over pairs of results we find interesting. On Croatian results we can observe already in the second experiment (POS) that basic morphological information in this simple setting improves F1 for 4.5% ($p = 0.002$). Our third experiment (DISTSIM 10M) shows that using distributional information

obtained from a 10 million token corpus improves the result as much as the part-of-speech information with similar significance ($p = 0.005$). By combining both features we improve our results for 8.5%, more significantly in comparison with using only one of those features ($p < 0.001$).

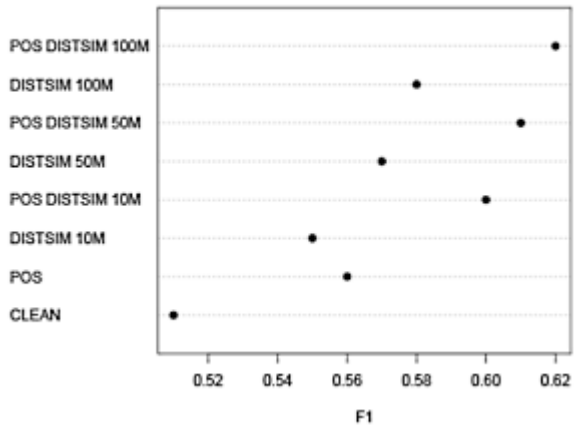


**Figure 1:** F1 results obtained via 5-fold cross-validation on Croatian development set.

By calculating distributional information on five and ten times more data we get improvements of 2% and 3% when not using part-of-speech information and 1% and 2% when using part-of-speech information. The differences between neighbouring corpus sizes (10 and 50; 50 and 100) are not statistically significant, but the differences between using 10Mw and 100Mw corpora are ($p = 0.007$). We see a steady rise in performance as the unlabelled monolingual corpus size increases, motivating us to perform similar calculations on much larger datasets in the future.

The results on Slovene data regarding the categories present in Croatian data are rather similar backing up those findings. There are two types of information in Slovene data we did not have for Croatian – MSD and lemma. By using MSD and not only POS information the results do improve for additional 1%, but statistically insignificantly ($p = 0.21$). On the contrary, by adding lemma information to the MSD decreases the result significantly for

5.5% (p = 0.007). One could expect such an outcome since lemmatization performs worst on named entities. Adding more distributional information by moving from a 10Mw to a 50Mw corpus we achieve an improvement of 5% which is even steeper than the one obtained on Croatian data, now highly significant (p < 0.001).

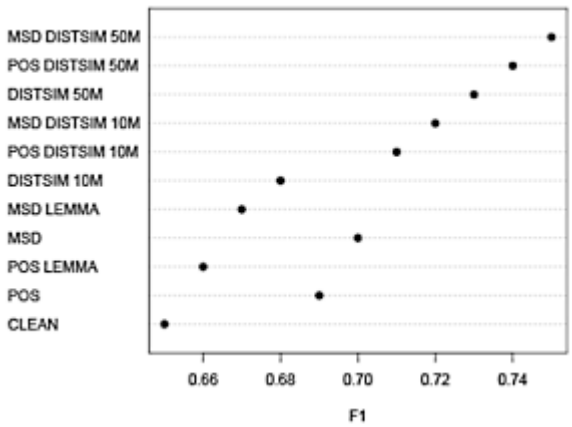


**Figure 2:** F1 results obtained via 5-fold cross-validation on Slovene development set.

This could be explained by the higher simplicity and similarity of this dataset to the monolingual corpus used for distributional similarity calculation, pointing to a conclusion that for datasets of narrower domains additional data sources such as this one give more improvement. We can observe on both datasets that, when using distributional similarity from larger corpora, including additional features like POS or MSD accounts for a lower increase in the results.

When comparing results on Croatian and Slovene datasets one observes right away that the results on Slovene data are much better although the size of the dataset is below half the size. This can be traced back to the fact that the Slovene dataset contains a narrower domain, has a higher vocabulary transfer and a higher amount of named entities like person and location which are considered easier to recognize and classify. On the other hand the resulting

Croatian module is expected to be more robust and should perform better on different domains.

| HR DISTSIM 100Mw | | | | | | |
|---|---|---|---|---|---|---|
| Entity | P | R | F1 | TP | FP | FN |
| LOC | 0.8049 | 0.7021 | 0.7500 | 33 | 8 | 14 |
| MISC | 0.7436 | 0.3867 | 0.5088 | 29 | 10 | 46 |
| ORG | 0.6742 | 0.6250 | 0.6486 | 60 | 29 | 36 |
| PERS | 0.9032 | 0.5185 | 0.6588 | 28 | 3 | 26 |
| Totals | 0.7500 | 0.5515 | 0.6356 | 150 | 50 | 122 |
| HR POS DISTSIM 100Mw | | | | | | |
| Entity | P | R | F1 | TP | FP | FN |
| LOC | 0.8293 | 0.7234 | 0.7727 | 34 | 7 | 13 |
| MISC | 0.7778 | 0.4667 | 0.5833 | 35 | 10 | 40 |
| ORG | 0.6989 | 0.6771 | 0.6878 | 65 | 28 | 31 |
| PERS | 0.8500 | 0.6296 | 0.7234 | 34 | 6 | 20 |
| Totals | 0.7671 | 0.6176 | 0.6843 | 168 | 51 | 104 |
| SL DISTSIM 50Mw | | | | | | |
| Entity | P | R | F1 | TP | FP | FN |
| LOC | 0.7423 | 0.7273 | 0.7347 | 72 | 25 | 27 |
| MISC | 0.5000 | 0.2143 | 0.3000 | 15 | 15 | 55 |
| ORG | 0.8947 | 0.3617 | 0.5152 | 17 | 2 | 30 |
| PERS | 0.8966 | 0.8509 | 0.8731 | 234 | 27 | 41 |
| Totals | 0.8305 | 0.6884 | 0.7528 | 338 | 69 | 153 |
| SL MSD DISTSIM 50Mw | | | | | | |
| Entity | P | R | F1 | TP | FP | FN |
| LOC | 0.7957 | 0.7475 | 0.7708 | 74 | 19 | 25 |
| MISC | 0.4688 | 0.2419 | 0.3191 | 15 | 17 | 47 |
| ORG | 0.8947 | 0.3617 | 0.5152 | 17 | 2 | 30 |
| PERS | 0.8619 | 0.8400 | 0.8508 | 231 | 37 | 44 |
| Totals | 0.8180 | 0.6977 | 0.7531 | 337 | 75 | 146 |

**Table 5:** Test results on the four best performing models (P - precision, R - recall, F1 - F1 measure, TP - true positives, FP - false positives, FN - false negatives).

The results given in Figures 1 and 2 are obtained via cross-validation by

evaluating five models built on different data on five different evaluation sets. We chose two settings per dataset for final testing on the left-out test set. The first one uses distributional information, but leaves out the need for morphological annotation of the data while the second one uses both distributional and morphological information. We present the results of precision, recall, F1, true positives and false positives and negatives by category in Table 5.

The number of false negatives shows to be on both datasets and settings higher than the number of false positives with higher percentage of precision than recall as a direct consequence. On Slovene data the best performing categories are PERS, LOC, ORG and then MISC. On Croatian data LOC tends to perform best, ORG and PERS forming a tie and MISC being traditionally the worst category. The somewhat unexpected order of category performance on the Croatian dataset can probably be followed to the wider domain of that dataset.

## 6 RESULTS OF THE SECOND BATCH OF EXPERIMENTS

In the second batch of experiments we used the secondary test sets consisting of left-out parts of all datasets used for training the models. We experimented with using part of speech and distributional information since these features showed to be most promising in the first batch of experiments. An additional reason not to include full MSD information is the result of an experiment where we assessed the usability of the MSD information on larger datasets such as the SETimes corpus which is partially manually and partially automatically annotated with full morphosyntactic information. We used 5-fold cross-validation on the dataset and the differences between using POS and MSD were consistently below 1%. This goes in line with our overall findings in the initial batch of experiments where additional linguistic features were losing importance when increasing the amount of annotated data.

The results for specific datasets are given in Table 6. The distributional

information proves to be of greater importance than the part-of-speech information on all datasets. Combining those two does just slightly improve the results when compared to using only distributional information. This is a very usable finding since it enables us to build final models that do not rely on part-of-speech information and thereby do not require such pre-processing.

| | **HR** | **Vjesnik** | **SETimes** | **SL** | **SSJ** | **HARMN** |
|---|---|---|---|---|---|---|
| **CLEAN** | 0.525 | 0.721 | 0.801 | 0.579 | 0.598 | 0.630 |
| **POS** | 0.577 | 0.732 | 0.811 | 0.573 | 0.587 | 0.642 |
| **DIST** | 0.624 | 0.796 | 0.844 | 0.635 | 0.666 | 0.702 |
| **POS DIST** | 0.663 | 0.786 | 0.846 | 0.641 | 0.647 | 0.707 |

**Table 6:** F1 test results on all available corpora based on secondary test sets.

Distributional similarity shows better performance on smaller datasets, with the difference on HR being 9.95%, on the Vjesnik dataset 7.53% and on the largest and densest SETimes dataset just 4.35%. POS features seem to lose on their significance when using bigger datasets as well.

The SSJ corpus, although almost 4 times bigger than the SL corpus, shows just slightly greater performance in recognizing named entities. The reason for this result is the low density of named entities in that dataset showing that newspaper corpora are better reference corpora for annotating and modelling this phenomenon.

The overall lower results on the Slovene datasets can be followed to their smaller size, inclusion of the miscellaneous category and the lower usefulness of the SSJ dataset for the task at hand.

The harmonic mean of all F1 measures on all corpora actually sums up our main findings – POS information has a small positive impact while DIST information has a very significant impact by improving the result for 7 points. Combining those two does not yield enough improvement to justify the pre-processing step of part-of-speech tagging.

We performed one final experiment where we combined all Croatian and all Slovene datasets into one dataset per language. Those results are shown in Figure 3 which depicts the F1 evaluation result as a function of the number of annotated named entities in each dataset. This plot shows the possible impact of adding more data for training the model as well.

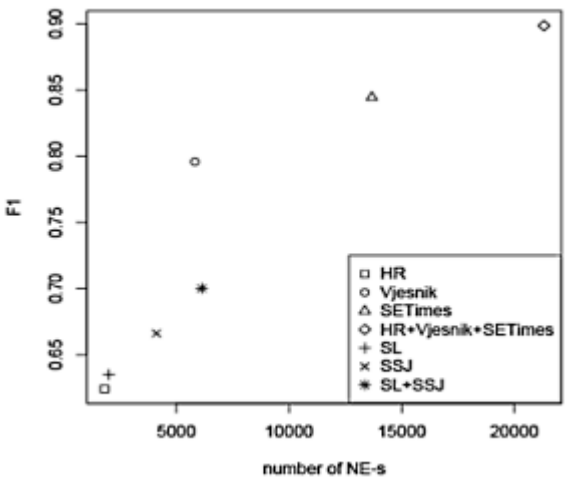


**Figure 3:** F1 measure as a function of the number of named entities.

On Croatian datasets we can observe a typical logarithmic behaviour with obvious room for improvement by annotating even more data. The Slovene datasets, when compared to the Croatian ones, are all small and near to each other regarding the amount of annotated named entities and are obviously still in the strong growth phase so building a larger Slovene dataset should be an even higher priority than for Croatian. The slower rise of the Slovene learning curve can be followed to two specificities of those datasets – inclusion of the miscellaneous category and smaller density of named entities providing a smaller amount of positive examples in the training set and leaving more room for errors in the test set.

Detailed results on the combined corpora are given in Table 7. Combined

corpora for Croatian with DISTSIM features and three named entity classes (ORG, PERS, LOC) yielded an F1 score of 89.8%. The results showed high recall for all named entity categories. The best performing category was LOC, followed by PERS and lastly ORG.

| **HR + Vjesnik + SETimes DISTSIM 100Mw** | | | | | | |
|---|---|---|---|---|---|---|
| **Entity** | P | R | F1 | TP | FP | FN |
| **LOC** | 0.9056 | 0.9467 | 0.9257 | 355 | 37 | 20 |
| **ORG** | 0.8875 | 0.8282 | 0.8568 | 347 | 44 | 72 |
| **PERS** | 0.9083 | 0.9269 | 0.9175 | 317 | 32 | 25 |
| **Totals** | 0.9002 | 0.8970 | 0.8986 | 1019 | 113 | 117 |
| **SL + SSJ DISTSIM 50Mw** | | | | | | |
| **Entity** | P | R | F1 | TP | FP | FN |
| **LOC** | 0.6794 | 0.8114 | 0.7396 | 142 | 67 | 33 |
| **MISC** | 0.2917 | 0.1538 | 0.2014 | 14 | 34 | 77 |
| **ORG** | 0.7391 | 0.3493 | 0.4744 | 51 | 18 | 95 |
| **PERS** | 0.8224 | 0.8674 | 0.8443 | 301 | 65 | 46 |
| **Totals** | 0.7341 | 0.6693 | 0.7002 | 508 | 184 | 251 |

**Table 7:** Test results on combined corpora (P - precision, R - recall, F1 - F1 measure, TP - true positives, FP - false positives, FN - false negatives).

Combined corpora for Slovene with DISTSIM features and four named entity classes (ORG, PERS, LOC, MISC) showed a lower improvement, due to corpora size and the number of observed named entity classes. The overall F1 result was 70.02%. The hardest class observed is expectedly MISC, and the best results are obtained for the PERS class. A low recall for ORG and especially MISC class could indicate the system's partial inability to distinguish between those two classes.

## 7 CONCLUSION

In this paper we have presented the process of building freely available models for named entity recognition and classification for Croatian and Slovene. We have built two initial datasets, one for Croatian which is larger

and covers a broader domain and one for Slovene which is smaller but covers just the general news domain. We were searching for the optimal set of features on the development set via five-fold cross-validation. Lemmata have shown to be of no use for a morphologically complex language such as Slovene since lemmatization tends to work worst on word classes such as named entities. On the other hand morphological information such as POS tags or full MSD tags proved to be valuable with the latter being more informative. That type of information improved the F1 measure in a 3-5% window. Clustering tokens from a large monolingual corpus by features such as contextual and morphological properties has proven to be beneficial improving the results for 3-4% by using 10Mw corpora. With clustering results from larger corpora the results continue to improve steadily. Combining both morphological and clustering information proved to be the winning combination with an overall improvement of 10% on datasets of both languages. By omitting morphological information for which pre-processing is required we still get an improvement of 8%.

The second batch of experiments included two additional datasets for Croatian and one for Slovene. By repeating the most promising settings from the first batch on this collection of datasets we managed to gain a better insight in the best performing settings. The results have shown that the impact of part-of-speech information is much lower than the one of distributional similarity. Both features lose importance as the dataset size increases. Combining these two features proved to be very similar to using distributional information only and this is the reason why the final models we publish do not require part-of-speech tagging, but have the distributional information included. By analysing the relation between dataset size and the obtained results we conclude that for both languages additional annotated data would yield improvement. The Slovene model could especially be easily improved with additional data of higher named entity density than the SSJ corpus.

Finally we are releasing three models – two for Croatian and one for Slovene, all of them using only distributional information as an additional feature and thereby not relying on any pre-processing but tokenization. Out of the two Croatian models one does cover the MISC category, but is trained on a much smaller amount of data and the other does not cover the MISC category, but is trained on a much larger amount of data and thereby more accurate. The Slovene model covers all four traditional categories. The models – together with the initial and the extended test sets – can be obtained from http://nlp.ffzg.hr/resources/models/ner/.

For the future our plan is to increase the amount of annotated data for training by exploiting semi-supervised approaches and add the MISC category to the whole dataset. Additionally we plan to calculate distributional similarity on larger corpora and take under consideration variations of the distributional similarity method used in this paper.

## REFERENCES

Agić, Ž., and Bekavac, B. (2013): Domain Dependence of Statistical Named Entity Recognition and Classification in Croatian Texts. *Proceedings of the 35th International Conference on Information Technology Interfaces (ITI 2013)*: 277–282. Cavtat.

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

Bekavac, B. (2005): *Strojno prepoznavanje naziva u suvremenim hrvatskim tekstovima: Ph.D. Thesis*. Zagreb: University of Zagreb.

Bošnjak, M. (2007): *Strojno prepoznavanje naziva tehnikama strojnog učenja: Master's Thesis*. Zagreb: University of Zagreb.

Clark, A. (2003): Combining Distributional and Morphological Information for Part of Speech Induction. *Proceedings of the Conference of the*

*European Chapter of the Association for Computational Linguistics:* 59–66. Budapest.

Erjavec, T., Fišer, D., Krek, S., and Ledinek, N. (2010): The JOS Linguistically Tagged Corpus of Slovene. *International Conference on Language Resources and Evaluation*: 1806–1809. Valetta.

Erjavec, T., Ignat, C., Poliquen, B., and Steinberger, R. (2005): Massive Multilingual Corpus Compilation: Acquis Communautaire and ToTaLe. *The 2nd Language & Technology Conference - Human Language Technologies as a Challenge for Computer Science and Linguistics*: 32–36. Poznań.

Erjavec, T., Krstev, C., Petkevič, V., Simov, K., Tadić, M., and Vitas, D. (2003): The MULTEXT-East Morphosyntactic Specifications for Slavic Languages. *MorphSlav '03 Proceedings of the 2003 EACL Workshop on Morphological Processing*: 25–32. Stroudsburg.

Faruqui, M., and Padó, S. (2010): Training and Evaluating a German Named Entity Recognizer with Semantic Generalization. *Proceedings of the 10th Conference on Natural Language Processing (KONVENS) 2010*. Saarbrücken.

Finkel, J. R., Grenager, T., and Manning, C. (2005): Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005):* 363–370. Stroudsburg.

Filipić, L., Jurić, T., and Stupar, M. (2012): *Strojno prepoznavanje naziva u tekstovima pisanima hrvatskim jezikom*. Zagreb: Sveučilište u Zagrebu.

Glavaš, G., Karan, M., Šarić, F., Šnajder, J., Mijić, J., Šilić, A., and Dalbelo Bašić, B. (2012): CroNER: A State-of-the-Art Named Entity Recognition and Classification for Croatian. *Proceedings of the Eighth Language*

*Technologies Conference:* 73–78. Ljubljana.

Halácsy, P., Kornai, A., and Oravecz, C. (2007): HunPos: an Open Source Trigram Tagger. *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, ACL '07*: 209–212. Stroudsburg.

Ljubešić, N., and Erjavec, T. (2011): hrWaC and slWac: Compiling Web Corpora for Croatian and Slovene**.** *Text, Speech and Dialogue –14th International Conference, TSD 2011:* 395–402. Pilsen.

McCallum, A. K. (2002): *Mallet: A Machine Learning for Language Toolkit.* Available at: http://mallet.cs.umass.edu ( 5 June 2013).

Štajner, T., Erjavec T., and Krek, S. (2012): Razpoznavanje imenskih entitet v slovenskem besedilu. *Proceedings of 15th International Multiconference on Information Society – Jezikovne tehnologije:* 191–197. Ljubljana.

Vitas, D., and Pavlović-Lažetić G. (2008): Resources and Methods for Named Entity Recognition in Serbian. *INFOtheca – Journal of Informatics and Librarianship*, 9(1–2): 35a.

# IZGRADNJA MODELOV ZA PREPOZNAVANJE IMENSKIH ENTITET ZA HRVAŠČINO IN SLOVENŠČINO

Prispevek predstavlja razvoj prosto dostopnih modelov za prepoznavanje in klasifikacijo imenskih enot za hrvaški in slovenski jezik. Poskusi se osredotočajo na najbolj informativne jezikovne lastnosti, pri čemer upoštevajo dostopnost jezikovnih orodij za oba jezika. Poleg standardnih jezikovnih lastnosti so upoštevane tudi distribucijske lastnosti, ki so bile izračunane iz velikih neoznačenih enojezičnih korpusov. Uporaba distribucijskih lastnosti izboljša rezultate za 7–8 točk v meri F1, uporaba oblikoslovnih informacij pa dodatno za 3–4 točke, in to pri obeh jezikih. Najboljši naučeni model skupaj s testno množico za primerjavo z obstoječimi in bodočimi sistemi ter model za oblikoslovno označevanje hrvaščine s programom HunPos so dostopni za prenos za uporabo v znanstvene in komercialne namene.

**Ključne besede:** prepoznavanje imenskih entitet, distribucijske lastnosti, hrvaščina, slovenščina