CroNER: Recognizing Named Entities in Croatian Using Conditional Random Fields

Mladen Karan, Goran Glavaš, Frane Šarić, Jan Šnajder, Jure Mijić, Artur Šilić and Bojana Dalbelo Bašić
University of Zagreb
Faculty of Electrical Engineering and Computing
Text Analysis and Knowledge Engineering Lab
Unska 3, 10000 Zagreb, Croatia
E-mail: takelab@fer.hr and http://takelab.fer.hr

Keywords: named entity recognition, conditional random fields, natural language processing, information extraction, Croatian language

Received: February 27, 2013

In this paper we present CroNER, a named entity recognition and classification system for Croatian language based on supervised sequence labeling with conditional random fields (CRF). We use a rich set of lexical and gazetteer-based features and different methods for enforcing document-level label consistency. Extensive evaluation shows that our method achieves state-of-the-art results (MUC F1 90.73%, Exact F1 87.42%) when compared to existing NERC systems for Croatian and other Slavic languages.

*Povzetek: V pričujočem prispevku je predstavljen CroNER, sistem za prepoznavanje in klasifikacijo imenskih entitet za hrvaščino, ki temelji na nadzorovanemu označevanju s pomočjo pogojnih naključnih polj (*conditional random fields – *CRF).*

1 Introduction

Named Entity Recognition and Classification (NERC) is a well-known natural language processing (NLP) and Information Extraction (IE) task. NERC aims to extract and classify all names (*enamexes*), temporal expressions (*timexes*), and numerical expressions (*numexes*) appearing in natural language texts. The classes of named entities typically extracted by NERC systems are names of people, organizations, and locations as well as dates, temporal expressions, monetary expressions, and percentages.

In this paper we present CroNER, a supervised NERC for the Croatian language. We use sequence labeling with conditional random fields (CRF) [13] to extract and classify named entities from newspaper text. We use a rich set of features, including lexical and gazetteer-based features, with many of them incorporating morphological and lexical peculiarities of the Croatian language. We implemented two different methods for document-level consistency of NE labels: postprocessing rules (hard consistency constraint) and a two-stage CRF (soft consistency constraint). Postprocessing rules are hand-crafted patterns designed to extract or re-label named entities omitted or misclassified by the CRF model. Two-stage CRF [12] 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. We evaluate the performance of the system using standard MUC and Exact NERC evaluation schemes [19].

The rest of the paper is structured as follows. In Section 2 we present related work on named entity extraction for Croatian and other Slavic languages. Section 3 discusses the details of corpus annotation. In Section 4 we thoroughly describe the feature set and the extensions used (rule-based postprocessing and two-stage CRF). Section 5 presents experimental setup and evaluation results. In Section 6 we conclude and outline future work.

2 Related work

Identifying references to named entities in text was recognized as one of the important subtasks of IE, and it has been a target of intense research for the last twenty years. The task was formalized at the Sixth Message Understanding Conference in 1995 [10]. There is a large body of NERC work for English [18, 17, 6, 12] and other major languages [7, 26, 4, 22]. Substantially less research has targeted Slavic (especially South Slavic) languages; NERC systems have been reported for Russian [23], Polish [21, 16], Czech [11], and Bulgarian [5, 9]. In [9] it was shown that CRF-based NERC with a rich set of features outperforms all other methods for Bulgarian, as well as other Slavic languages.

The rule-based system from [2], which uses a cascade of finite-state transducers, is the only reported work on NERC for Croatian language. In [15] a method for generating a morphological lexicon of organizational names was proposed, a valuable resource for morphologically rich languages. We used a similar approach to expand morphological lexica with inflectional forms of Croatian proper names, but we include first names, surnames, and toponyms in ad-

dition to organization names.

To the best of our knowledge, we are the first to use supervised machine learning for named entity recognition and classification in Croatian language. Using a machine learning method, we avoid the need for specialized linguistic knowledge required to design a rule-based system. This way we also avoid the explicit modelling of complex dependencies between rules and their application order. We instead focus on designing a rich set of features and let the CRF algorithm uncover the dependencies between them.

3 Corpus annotation

The training and testing corpus consists of 591 news articles (about 310,000 tokens) from the Croatian newspaper *Vjesnik*, spanning years 1999 to 2009. The preprocessing of the corpus involved sentence splitting and tokenization. For annotation we used seven standard MUC-7 types: *Organization*, *Person*, *Location*, *Date*, *Time*, *Money*, and *Percent*. We also introduced five additional types: *Ethnic* (names of ethnic groups), *PersonPossessive* (possesive adjectives derived from person names), *Product* (names of branded products), *OrganizationAsLocation* (organization names used as metonyms for locations, as in *"The entrance of the PBZ bank building"*), and *LocationAsOrganization* (location names used as metonyms for organizations, as in *"Zagreb has sent a demarche to Rome"*). The additional types were introduced for experimental reasons; in this work only the *Ethnic* tag was retained, while other additional tags were not used (i.e., the *Product* tag was discarded, while the remaining three subtype tags were mapped to the corresponding basic tags). Thus, in the end we trained our models using eight types of named entities.

The annotation guidelines we used are similar to MUC-7 guidelines, with some adjustments specifically for the Croatian language. The corpus was independently annotated by six annotators. To ensure high annotation quality, the annotators were first asked to independently annotate a calibration set of about 10,000 tokens. On this set, all the disagreements have been resolved by consensus, the borderlines were discussed, and the guidelines revised accordingly. Afterwards, each of the remaining documents was annotated by two independent annotators, while a third annotator resolved the disagreements. For annotating we used an in-house developed annotation tool.

The inter-annotator agreement (calculated in terms of MUC F1 and Exact F1 score and averaged over all pairs of annotators) is shown in Table 1. The inter-annotated was measured on a subset of about 10,000 tokens that was annotated by all six annotators. Notice that the overall quality of the annotations improved after resolving the disagreements, but – because each subset was resolved by a single annotator – we cannot objectively measure the resulting improvement in annotation quality.

Table 1: Inter-annotator agreement

Tag	F1 Exact	F1 MUC
Person	98.05	98.55
Ethnic	97.19	97.19
Percent	92.00	96.77
Location	93.95	94.93
Money	91.95	94.15
Organization	89.35	93.58
Date	71.47	85.79
Time	67.55	71.04

4 CroNER

CroNER is based on sequence labeling with CRF. We use the CRFsuite [20] implementation of CRF. At the token level, named entities are annotated according to the Begins-Inside-Outside (B-I-O) scheme, often used for sequence labeling tasks. Following is a description of the features used for sentence-level label prediction and the techniques for imposing document-level label consistency.

4.1 Sentence-level features

Most of the features can be characterized as lexical, gazetteer-based, or numerical. Some of the features were *templated* on a window of size two, both to the left and to the right of the current word. This means that that the feature vector for the current word consists of features for this word, two previous words, and two following words.

Lexical features. The following is the list of the lexical features used (templated features are indicated as such).

1. Word, lemma, stem, and POS tag (*templated*) – For lemmatization we use the morphological lexicon described in [25]. For stemming, we simply remove the word's suffix after the last vowel (or the penultimate vowel, if the last letter is a vowel). Words shorter than 5 letters are not stemmed. For POS tagging, we use a statistical tagger with five basic tags.
2. Full and short shape of the word – describe the ordering of uppercased and lowercased letters in the word. For example, *"Zagreb"* has the shape *"ULLLLL"* and short shape *"UL"*, while *"iPhone"* has the shape *"LULLLL"* and short shape *"LUL"*.
3. Sentence start – indicates whether the token is the first token of the sentence.
4. Word ending – the suffix of the word taken from the last vowel till the end of the word (or the penultimate vowel, if the last letter of the word is a vowel).
5. Capitalization and uppercase (*templated*) – indicates whether the word is capitalized or entirely in uppercase (e.g., an acronym).

6. Acronym declension – indicates whether the word is a declension of an acronym (e.g., *"HOO-om"*, *"HDZ-a"*). Declension of acronyms in Croatian language follows predictable patterns [1].
7. Initials – indicates whether a token is an initial, i.e., a single uppercase letter followed by a period.
8. Cases – concatenation of all possible cases for the word, based on morpho-syntactic descriptors (MSDs) from the morphological lexicon. If the word has two or more MSDs with differing cases, we concatenate them in alphabetical order. We also add one Boolean feature for each individual case (*isNominative*, *isGenitive*, *isDative*, *isAcusative*, and *isInstrumental*).
9. Bigram features – concatenations of the previously described features computed for two consecutive tokens: *word bigram*, *lemma bigram*, *POS bigram*, *shape bigram*, and *cases bigram*.
10. Lemmas in window – all lemmas within a symmetric window of size 5 from the current token.
11. MSDs in window – all MSDs of the words within a symmetric window of size 5 from the current token.

Gazetteer-based features. Information about the presence of named entities from predefined gazetteers has been shown to be an important information for NERC [19]. We use several gazetteers: first names, surnames, ethnics, organizations, cities, streets, and countries gazetteers. The last four gazetteers have multi-word entries. The following is a list of gazetteer-based features.

1. Gazetteer match – indicates whether the lemma matches a gazetteer entry (used for gazetteers with single-word entries: names, surnames, and ethnics).
2. Starts gazetteer match – indicates whether there is any sequence of words starting with the current word that fully matches a gazetteer entry. E.g., in *"usluge Zavoda za javno zdravstvo"* (*services of the Public Health Department*), the word *"Zavoda"* would have this feature set to *true* because the organizations gazetteer contains *"Zavod za javno zdravstvo"*.
3. Stemmed gazetteer match – similar to the previous feature, but considers stems instead of lemmas. This feature is used only for the organizations gazetteer.
4. Gazetteer match length – the length (number of words) of the gazetteer entry whose first token matches the current token (e.g., for token *"Zavod"* in text *"usluge Zavoda za javno zdravstvo"*, the length would be 4).
5. Inside gazetteer match – indicates whether a word is inside the phrase that matches a gazetteer entry (e.g., true for tokens *"za"*, *"javno"*, and *"zdravstvo"* in organization entry *"Zavod za javno zdravstvo"*).

Both the text and the gazetteer entries were lemmatized before looking for matches. As gazetteers predominantly contain proper nouns, we needed to extend the morphological lexicon with the inflectional forms of proper names. We did this automatically with a set of rules following the paradigms for proper names declension [1]. We expanded both Croatian and foreign proper names.

Some simple preprocessing steps were applied for all gazetteers. All entries containing non-alphabetic characters were removed. We considered all words with more than 10% non-capitalized occurrences in the corpus to be common words and removed such entries. The rationale was to eliminate common word entries from the gazetteers in order to reduce the noise in the training set. For example,*"Luka"* is a very common personal name, but also a frequent common noun (*port*). Capitalization frequencies required for the above analysis were gathered from the *Vjesnik* corpus, a collection of 270,000 newspaper articles.

The major source of the Croatian names and surnames was the Croatian telephone directory. For English names, we used Stanford NER[1] to extract names from the NYT corpus[2] and Wikipedia. The compiled gazetteers for personal names and surnames contain 13,618 Croatian first names, 64,240 Croatian surnames, 70,488 foreign first names, and 228,134 foreign surnames. For locations we use three gazetteers – for streets, countries and cities. The street names (52,593 entries) were extracted from the Croatian telephone directory. Country names in Croatian (276 entries) were obtained from Wikipedia. The cities gazetteer (289,707 entries) was constructed using the telephone directory and internet sources. The organizations gazetteer (3035 entries) was created from several different sources, and includes names of institutions (e.g., *Ministry of Science*, *Louvre*), political parties (e.g., SDP, HDZ), international organizations (e.g., *UNESCO*, *NATO*), local and foreign companies, newspaper names, and sports teams. Finally, we compiled the ethnics gazetteer (940 entries) automatically from country names using the appropriate rules of Croatian grammar [1].

Numerical features. We used the following features to deal specifically with numbers (occurring in numexes and timexes):

1. Integer or decimal number – indicates whether the word is an integer or a decimal number;
2. Two/four digit integer – indicates whether the token is a two digit (useful for recognizing numexes) or a four digit integer (useful for recognizing years in dates);
3. Number followed by a period – indicates whether the token is an integer followed by a period (a good clue for dates and currencies);
4. Currency – indicates whether a token is a currency marker (e.g., *"$"* or *"EUR"*). We compiled a currency gazetteer that includes all major world currencies.

[1] `http://nlp.stanford.edu/software/CRF-NER.shtml`
[2] The New York Times Annotated Corpus, (2008), LDC.

4.2 Document-level consistency

The CRF model predicts the sequence of B-I-O labels on the sentence level. It is therefore possible to have at the document level differing labels for the same named entity. The goal of the document-level label consistency postprocessing is to unify the labels of named entities on the document level. We experimented with incorporating document-level consistency into our model as both soft constraints (two-stage CRF) and hard constraints (handcrafted postprocessing rules).

Two-stage CRF. The two-stage CRF [12] is a model that accounts for non-local dependencies between named entities. The main idea is to employ a second CRF that uses both local features (same features the first CRF uses) and non-local features computed on the output of the first CRF. We use three document-level features computed from the output of the first CRF:

1. The most frequent lemma label – the most frequent label assigned to a given lemma in the document (e.g., *B_Person* or *I_Organization*);
2. The most frequent NE label – the most frequent label assigned to a given NE mention in the document;
3. The most frequent superentity label – a superentity is a mention of the same entity that contains two or more tokens (e.g., *"Ivan Horvat"* vs. *"Horvat"*, or *"Zavod za javno zdravstvo"* vs. *"Zavod"*). This feature represents the most frequent label assigned to all the superentities of a given entity within the document.

Postprocessing rules (PPR). We created two sets of postprocessing rules: one to enforce document-level consistency (hard constraint) and another one to improve the recall on numexes and timexes. The rules for enforcing document-level label consistency work as follows. First, we collect all the different named entities recognized by the CRF model and identify the most frequent label assigned to each of them. Then we correct (i.e., re-label) NE instances that were assigned a different label from the most frequently assigned one. In the second step, we search for the potential false negatives (i.e., mentions of named entities from the collection that were omitted by the CRF model). If found, omitted mentions are also assigned the most frequent label for the corresponding named entity.

The rules for improving the recall for numexes are in fact token-level regular expressions. For currencies and percentages the rules are defined as follows:

1. $[num][num|prep|conj]^*[currencyMarker]$ – the currency expression starts with a number, followed by either numbers, prepositions, or conjunctions, and ends with a currency clue. When written in words, numbers often contain conjunctions. E.g., in *"trideset i pet"* (*thirty five*), word *"i"* is a conjunction. Ranges are often expressed using prepositions; e.g., *"30 do 50 milijuna kuna"* (*30 to 50 million kuna*);
2. $[num][num|prep|conj]^*[percentClue]$ – the rule for percentages is similar to the rule for currencies, except for requiring that the phrase ends with a percent clue (*"posto"* or *"%"*) instead of a currency marker.

For timex (time) class we use the following three rules:

1. $[\text{u}][number][timeword]$ – captures phrases like *"u 12.30 sati"* (*at 12.30 o'clock*), where *number* is an appropriately formatted number and *timeword* is a word from a predefined list of time-related words, e.g., *"sati"* (*o'clock*);
2. $[mod]?[preposition]?[daytimeword][mod]?$ – captures phrases like *"rano u jutro"* (*early in the morning*). Here *mod* represents a modifying word, e.g.. *"rano"* (*early*);
3. $[modGen][daytimeword]$ – captures phrases like *"tijekom podneva"* (*during the afternoon*), where *modGen* is a modifier that governs a noun in genitive case; e.g., *"prije"* (*before*).

5 Evaluation

We measured the performance of four different models: single CRF (1-CRF), two-stage CRF (2-CRF), single CRF with postprocessing rules (1-CRF + PPR), and two-stage CRF with postprocessing rules (2-CRF + PPR). In Tables 2 and 3 we report the performance in terms of precision, recall, and F1 for MUC (allows for extent overlap instead of an exact extent match) and Exact (requires that both extent and class match) evaluation schemes [19], respectively. Results are reported separately for each NE class. We also report both micro- and macro-averaged overall performance for each of the four models. The results were obtained with 10-fold cross validation on the entire annotated corpus.

5.1 Result analysis

Regarding the enamex classes, the performance for organizations is significantly (5–7%) worse than for persons and locations. This is expected, because in Croatian many organization instances are multi-word expressions, whereas person and location mentions more often consist of only one or two words. The lower inter-annotator agreement (cf. Table 1) for organizations supports this assumption.

The results show that 2-CRF outperforms 1-CRF consistently on main enamex classes (*Person*, *Organization*, and *Location*); the improvement is between half a point (*Location*) and a full point (*Organization*). The 1-CRF + PPR model similarly outperforms 1-CRF (e.g., 0.8 point increase for *Person*). However, the 2-CRF + PPR model brings negligible gain when compared to either 2-CRF or 1-CRF + PPR (on average 0.1 point for enamex classes). This indicates that both the second stage CRF and postprocessing rules ensure document-level consistency in a similar fashion, hence combining them does not lead to significant performance improvements.

Table 2: CroNER MUC evaluation results

	1-CRF			2-CRF			1-CRF + PPR			2-CRF + PPR		
NE Class	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Person	91.31	92.12	91.71	91.76	93.26	92.50	91.13	93.58	92.34	91.62	93.68	**92.64**
Location	89.27	89.77	89.52	89.83	90.30	**90.06**	88.30	91.00	89.63	89.00	90.46	89.72
Organization	88.15	81.65	84.78	88.66	82.94	**85.71**	85.51	84.74	85.13	86.43	84.11	85.25
Ethnic	96.82	90.56	93.59	97.73	90.55	94.01	97.74	90.56	94.01	98.29	90.56	**94.27**
Date	93.72	82.35	87.67	93.48	82.02	87.38	93.55	83.05	**87.99**	93.56	82.47	87.67
Time	91.86	50.22	64.94	91.74	49.33	64.16	76.96	78.67	77.80	77.06	79.11	**78.07**
Currency	99.54	87.30	93.02	99.32	88.10	93.37	99.20	99.20	**99.20**	99.20	99.20	**99.20**
Percent	100.00	96.43	98.18	100.00	96.21	98.07	99.54	97.77	**98.65**	99.54	97.77	**98.65**
Overall Micro	90.67	87.21	88.91	91.07	87.99	89.51	89.48	89.43	89.45	90.09	89.09	**89.59**
Overall Macro	93.84	83.80	88.78	94.06	84.08	88.79	91.49	89.82	90.65	91.83	89.67	**90.73**

Table 3: CroNER Exact evaluation results

	1-CRF			2-CRF			1-CRF + PPR			2-CRF + PPR		
NE Class	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Person	89.42	90.22	89.81	89.92	91.38	**90.64**	89.06	91.46	90.24	89.62	91.64	90.62
Location	87.60	88.09	87.84	88.11	88.57	**88.34**	86.58	89.21	87.87	87.34	88.74	88.03
Organization	80.79	74.83	77.70	81.05	75.82	**78.35**	77.26	76.94	77.10	78.58	76.57	77.56
Ethnic	96.82	90.56	93.59	97.74	90.56	94.01	97.73	90.56	94.01	98.29	90.56	**94.27**
Date	86.19	75.73	80.62	85.98	75.44	80.37	85.73	76.10	**80.63**	85.95	75.77	80.54
Time	87.80	48.00	62.07	88.43	47.55	61.85	66.08	67.55	66.81	66.23	68.00	**67.10**
Currency	95.93	84.13	89.64	95.75	84.92	90.01	96.45	97.22	**96.84**	96.27	97.22	96.74
Percent	95.60	92.19	93.86	95.82	92.19	93.97	98.86	97.09	**97.97**	98.86	97.10	**97.97**
Overall Micro	86.84	83.53	85.15	87.19	84.24	**85.69**	85.30	85.36	85.33	86.08	85.17	85.62
Overall Macro	90.00	80.47	84.97	90.35	80.80	85.31	87.21	85.76	86.49	87.64	87.20	**87.42**

For numexes, the second CRF model seems not to improve the performance, whereas the postprocessing rules significantly improve the performance. This improvement is to be attributed to the use of extraction rules for numexes, implying that document-level consistency is not an issue for numexes. Postprocessing rules for currencies and percents increase the recall and keep the precision on the same level. For temporal expressions, however, increase in recall is accompanied by a proportional decrease in precision. Deeper inspection reveals that this is mostly due to inconsistent annotations of timexes, as confirmed by the very low inter-annotator agreement for these classes (cf. Table 1).

As expected, Exact evaluation results are generally lower than MUC results. However, for most classes the decrease in performance is not significant. Exceptions to this are *Organization*, *Date*, and *Time* classes, for which the decrease in performance is 7%, 7%, and 11%, respectively. Many organization instances consist of four or more words, and in such cases our models – though able to recognize the mention – often fail to exactly match its extent. The most common errors include omitting the last word or adding an extra word at the end. The performance on the three mentioned classes is also limited by the annotation quality; these classes are in fact the ones on which human annotators agreed the least (cf. Table 1).

Table 4 shows the performance of the best-performing model (2-CRF + PPR) depending on the size of the training set. (25%, 50%, 75%, and 100% of the training data). Expectedly, the performance generally improves as the size of the training set increases. However, the improvement from using 75% data to using 100% data is relatively small, suggesting that no significant increase in performance could be gained from annotating a larger corpus.

5.2 Discussion

Unfortunately, our results are not directly comparable to other reported results because of the differences in (1) language (though very similar, all Slavic languages have their own peculiarities), (2) NE types (e.g., some use only four classes: *Person*, *Location*, *Organization*, and *Miscellaneous*), or (3) evaluation methodology (non-adherence to standard evaluation methodology, such as in the work from [2]). Nonetheless, the comparison might still be informa-

Table 4: CroNER performance depending on the size of the training set (CRF-2 + PPR)

Evaluation	Size (tokens)	Person	Loc.	Org.	Ethnic	Date	Time	Curr.	Perc.	Micro	Macro
MUC	25% (75k)	92.51	82.69	79.95	92.30	79.46	78.74	100.00	98.99	86.01	88.08
	50% (155k)	92.56	87.56	82.60	93.70	85.01	76.40	99.62	98.64	88.05	89.51
	75% (230k)	92.19	88.81	85.00	94.87	87.30	76.84	99.59	98.77	89.07	90.42
	100% (310k)	92.64	89.72	85.25	94.27	87.67	78.07	99.20	98.65	**89.59**	**90.73**
Exact	25% (75)	90.17	79.50	69.53	92.30	71.57	59.84	96.97	98.32	80.65	82.28
	50% (155k)	90.59	85.04	73.66	93.70	76.47	62.17	97.51	97.74	83.35	84.61
	75% (230k)	90.06	86.71	77.25	94.87	79.45	65.40	97.24	97.84	84.80	86.10
	100% (310k)	90.62	88.03	77.56	94.27	80.54	67.10	96.74	97.97	**85.62**	**87.42**

tive to some extent. In [2], a 79% F1-score on persons, 89% on organizations, and 95% on locations is reported, although it must be noted that for the latter two classes the evaluation was limited to selected subsets of NE instances. Our results seem to be better than those reported for other Slavic languages: Polish – 82.4% F1, [21], Czech – 76% F1, Russian – 70.9% F1 [23]. Only the best reported results for Bulgarian are comparable to our results: 89.6% overall F1, persons 92.79%, locations 90.06%, organizations 89.73% [9]. These comparisons suggest that CroNER is a state-of-the-art NERC system when considering the Slavic languages.

5.3 Experiments with distributional features

It has been demonstrated [8, 24, 7] that NERC can benefit from distributional modelling of lexical semantics. Distributional semantics is based on the hypothesis that semantically similar words occur in similar contexts, therefore the meaning of a word can be represented by its context. Distributional representations can be used to compare and cluster together similar words, improving NERC performance. To determine if this also holds in our case, we performed preliminary experiments with semantic cluster features.

To obtain semantic word clusters we use Brown's algorithm [3, 14]. The algorithm takes as input a sequence of words (a corpus) and outputs semantic clusters for each word. The number of clusters k is a parameter of the algorithm. The algorithm works by assuming the probability of a word sequence is given as follows:

$$P(w_1, \ldots, w_n) = \prod_{i=1}^{n} p(w_i|C(w_i))p(C(w_i)|C(w_{i-1})) \quad (1)$$

where $C(w_i)$ is the cluster to which the i-th word w_i is assigned. It is assumed that the probability of an occurrence of a particular word w_i at position i depends only on its cluster $C(w_i)$, which, in turn, depends only on the cluster of the previous word $C(w_{i-1})$. The quality of a clustering is measured by how well the classes of adjacent words in the sequence predict each other. This is achieved by maximizing the log probability of the input sequence given by (1). In [14] it was demonstrated that this optimisation is equivalent to maximizing the sum of mutual information

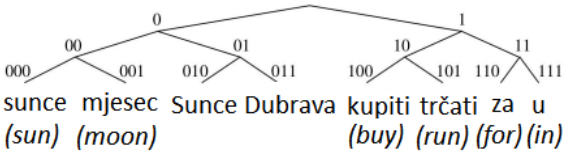

Figure 1: A clustering example for Croatian words (*"Sunce"* and *"Dubrava"* are proper names)

weights between all pairs of classes, and presented an efficient algorithm for computing the clusters.

In order to generate the sequence required as input to the algorithm, we took a sample from the HrWaC corpus (it was not possible to use the entire corpus due to its size). The sample consists of texts from three large internet news portals: *monitor.hr*, *slobodnadalmacija.hr*, and *vecernji.hr*. We chose news portals because they are of the same genre as our training and test data. Additionally, we expect the language used in news portals to be more standard and clean. The chosen texts were tokenized and lemmatized. The final input set for the algorithm had 351M tokens. To compute the clustering we used the freely available implementation from [14] with k set to 100. As a result, we obtained classes for each word as a bit string. The bit strings represent paths to each word in a binary tree whose leaves are clusters. An example of a good clustering is given in Fig. 1. An interesting property of this clustering is that we can control the generality of the clustering by looking only at a fixed length prefix of the bit string (e.g., it has been noted that prefixes of length four often correspond to POS tags).

We use clusters (in form of bit strings) as additional features for the CRF model. For each word w_i there are five features representing distributional clusters of words w_{i-2} to w_{i+2}. The number of possible distinct values for each of these features equals the number of clusters (100 in our case). We use the same procedure to include information about cluster prefixes of length two and four; in these cases the number of possible distinct values is smaller than the total number of clusters because all clusters beginning with the same prefix are merged. This approach is along the

Table 5: Comparison of CroNER performance with and without distributional features

Evaluation	Model	Person	Loc.	Org.	Ethnic	Date	Time	Curr.	Perc.	Ov. Macro
MUC	2-CRF + PPR	92.35	89.31	83.88	95.53	87.69	79.40	99.53	98.57	89.78
	2-CRF + PPR + dist.	93.37	89.17	85.21	95.46	87.26	78.5	99.4	98.57	**89.86**
Exact	2-CRF + PPR	90.20	87.82	76.81	95.53	81.19	67.15	96.93	98.0	86.00
	2-CRF + PPR + dist.	91.33	87.61	77.49	95.46	80.93	67.29	97.64	98.04	**86.25**

lines of the one proposed in [24].

Table 5 gives a comparison of performance with and without using distributional features, averaged over five cross validation folds on the entire data set. The use of distributional features leads to consistent improvements for Person and Organization classes. However, results for some of the other classes showed slight deterioration. This suggest that the distributional features are beneficial, but further experiments (with respect to the number of clusters and corpus size/choice) are required.

6 Conclusion and future work

We have presented CroNER, a NERC system for Croatian based on sequence labeling with CRF. CroNER uses a rich set of lexical and gazetteer-based features achieving good recognition and classification results. We have shown how enforcing document-level label consistency (either through postprocessing rules or a second CRF model capturing non-local dependencies) can further improve NERC performance. The experimental results indicate that, as regards the Slavic languages, CroNER is a state-of-the-art named entity recognition and classification system.

The work presented here could be extended in several ways. First, the annotated set should be revised, considering that the inter-annotator agreement is rather low on some classes. Secondly, a systematic feature selection (e.g., wrapper feature selection) may be performed in order to select an optimal subset of features. Thirdly, we plan to employ classification using more fine-grained NE labels. Finally, we intend to further explore the use of distributional semantic features.

7 Acknowledgments

This work has been supported by the Ministry of Science, Education and Sports, Republic of Croatia, under Grant 036-1300646-1986.

References

[1] S. Babić, B. Finka, and M. Moguš. *Hrvatski pravopis*. Školska knjiga, 1996.

[2] B. Bekavac and M. Tadić. Implementation of Croatian NERC system. In *Proc. of the Workshop on Balto-Slavonic Natural Language Processing: Information Extraction and Enabling Technologies*, pages 11–18, 2007.

[3] P.F. Brown, P.V. Desouza, R.L. Mercer, V.J.D. Pietra, and J.C. Lai. Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–479, 1992.

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

[5] J.F. Da Silva, Z. Kozareva, and GP Lopes. Cluster analysis and classification of named entities. In *Proc. Conference on Language Resources and Evaluation*, pages 321–324, 2004.

[6] O. Etzioni, M. Cafarella, D. Downey, A.M. Popescu, T. Shaked, S. Soderland, D.S. Weld, and A. Yates. Unsupervised named-entity extraction from the web: An experimental study. *Artificial Intelligence*, 165(1):91–134, 2005.

[7] M. Faruqui and S. Padó. Training and evaluating a German named entity recognizer with semantic generalization. *Semantic Approaches in Natural Language Processing*, page 129, 2010.

[8] Dayne Freitag. Trained named entity recognition using distributional clusters. In *Proceedings of EMNLP*, volume 4, pages 262–269, 2004.

[9] G. Georgiev, P. Nakov, K. Ganchev, P. Osenova, and K. Simov. Feature-rich named entity recognition for Bulgarian using conditional random fields. In *Proc. of the Conference on Recent Advances in Natural Language Processing (RANLP'2009)*, pages 113–117, 2009.

[10] R. Grishman and B. Sundheim. Message Understanding Conference-6: A brief history. In *Proc. of COLING*, volume 96, pages 466–471, 1996.

[11] J. Kravalová and Z. Žabokrtskỳ. Czech named entity corpus and SVM-based recognizer. In *Proc. of the 2009 Named Entities Workshop: Shared Task on Transliteration*, pages 194–201, 2009.

[12] V. Krishnan and C.D. Manning. An effective two-stage model for exploiting non-local dependencies in named entity recognition. In *Proc. of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL*, pages 1121–1128, 2006.

[13] J. Lafferty, A. McCallum, and F.C.N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. of ICML01*, 2001.

[14] P. Liang. *Semi-supervised learning for natural language*. PhD thesis, Massachusetts Institute of Technology, 2005.

[15] N. Ljubešić, T. Lauc, and D. Boras. Generating a morphological lexicon of organization entity names. In *Proc. of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, 2008.

[16] M. Marcińczuk and M. Janicki. Optimizing CRF-based model for proper name recognition in Polish texts. *Computational Linguistics and Intelligent Text Processing*, pages 258–269, 2012.

[17] A. McCallum and W. Li. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proc. of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 188–191, 2003.

[18] A. Mikheev, C. Grover, and M. Moens. Description of the LTG system used for MUC-7. In *Proc. of 7th Message Understanding Conference (MUC-7)*, 1998.

[19] D. Nadeau and S. Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007.

[20] N. Okazaki. CRFsuite: a fast implementation of conditional random fields (CRFs), 2007.

[21] J. Piskorski. Extraction of Polish named entities. In *Proc. of the Fourth International Conference on Language Resources and Evaluation, LREC*, pages 313–316, 2004.

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

[23] B. Popov, A. Kirilov, D. Maynard, and D. Manov. Creation of reusable components and language resources for named entity recognition in Russian. In *Proc. of the Fourth International Conference on Language Resources and Evaluation, LREC*, pages 309–312, 2004.

[24] L. Ratinov and D. Roth. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pages 147–155. Association for Computational Linguistics, 2009.

[25] J. Šnajder, B.D. Bašić, and M. Tadić. Automatic acquisition of inflectional lexica for morphological normalisation. *Information Processing & Management*, 44(5):1720–1731, 2008.

[26] S. Yu, S. Bai, and P. Wu. Description of the Kent Ridge Digital Labs system used for MUC-7. In *Proc. of the Seventh Message Understanding Conference*, 1998.