Named Entity Recognition: Exploring Features

Maksim Tkachenko
St Petersburg State University
St Petersburg, Russia
maksim.tkachenko@math.spbu.ru

Andrey Simanovsky
HP Labs Russia
St Petersburg
andrey.simanovsky@hp.com

Abstract

We study a comprehensive set of features used in supervised named entity recognition. We explore various combinations of features and compare their impact on recognition performance. We build a conditional random field based system that achieves 91.02% F_1-measure on the CoNLL 2003 (Sang and Meulder, 2003) dataset and 81.4% F_1-measure on the OntoNotes version 4 (Hovy et al., 2006) CNN dataset, which, to our knowledge, displays the best results in the state of the art for those benchmarks respectively. We demonstrate statistical significance of the boost of performance over the previous top performing system. We also obtained 74.27% F_1-measure on NLPBA 2004 (Kim et al., 2004) dataset.

1 Introduction

Recognition of named entities (e.g. people, organizations, locations, etc.) is an essential task in many natural language processing applications nowadays. Named entity recognition (NER) is given much attention in the research community and considerable progress has been achieved in many domains, such as newswire (Ratinov and Roth, 2009) or biomedical (Kim et al., 2004) NER. Supervised NER that uses machine learning algorithms such as conditional random fields (CRF) (McCallum and Li, 2003) is especially effective in terms of quality of recognition.

Supervised NER is extremely sensitive to selection of an appropriate feature set. While many features were proposed for use in supervised NER systems (Krishnan and Manning, 2006; Finkel and Manning, 2009; Lin and Wu, 2009; Turian et al., 2010), only limited studies of the impact of those features and their combinations on the effectiveness of NER were performed. In this paper we provide such a study.

Our contributions are the following:

- analysis of the impact of various features taken from a comprehensive set on the effectiveness of a supervised NER system;
- construction of a CRF-based supervised NER system that achieves 91.02% F_1-measure on the CoNLL 2003 (Sang and Meulder, 2003) dataset and 81.4% F_1-measure on the OntoNotes version 4 (Hovy et al., 2006) CNN dataset;
- demonstration of statistical significance of the obtained boost in NER performance on the benchmarks;
- application to NER of a DBPedia (Mendes et al., 2011) markup feature and a phrasal clustering (Lin et al., 2010) feature which have not been considered for NER in previous works.

The remainder of the paper is structured in the following way. In Section 2 we describe related work on feature analysis. In Section 3 we give a brief introduction to the benchmarks that we use. In Section 4 we discuss various features and their impact. Section 5 describes the final proposed system. Section 6 contains a summary of the performed work and future plans.

2 Related Work

The majority of papers on NER describe a particular method or feature evaluation and do not make a systematic comparison of combinations of features. In this paper those works are mentioned later when we discuss a particular feature or a group of features. In this section we present several works that deal with multiple features and thus are close to our study.

Design questions of NER systems were considered by (Ratinov and Roth, 2009). They used a perceptron-based recognizer with greedy inference and evaluated two groups of features: non-local dependencies (e.g. context aggregation) and external information (e.g. gazetteers mined from Wikipedia). Their recognizer was tested on the CoNLL 2003 dataset, a newswire dataset ($F_1 = 90.80\%$), the MUC7 dataset, and their own web pages dataset.

The authors of (Turian et al., 2010) systematically compared word representations in NER (Brown clusters, Collobert and Weston embeddings, HLBL embeddings). They ignored other types of features.

(Saha et al., 2009) presented a comparative study of different features in biomedical NER. They used a dimensionality reduction approach to select the most informative words and suffixes and they used clustering to compensate for the lost information. The MaxEnt tagger developed by them obtained $F_1 = 67.4\%$ on NLPBA 2004 data.

3 Benchmarks

In this paper we present the results obtained on three benchmarks: CoNLL 2003, OntoNotes version 4, and NLPBA 2004 dataset.

CoNLL 2003 is an English language dataset for NER. The data comprises Reuters newswire articles annotated with four entity types: person (PER), location (LOC), organization (ORG), and miscellaneous (MISC). The data is split into a training set, a development set (testa), and a test set (testb). Performance on this task is evaluated by measuring precision and recall of annotated entities combined into F_1-measure. We used BILOU (begin, inside, last, outside, unit) annotation scheme to encode named entities. Previous top performing systems also followed that scheme. We study feature behavior on this benchmark; our system is tuned on the test and development sets of it.

OntoNotes version 4 is an English language dataset designed for various natural language processing tasks including NER. The dataset consists of several sub-datasets taken from different sources including Wall Street Journal, CNN news, machine-translated Chinese magazines, Web blogs, etc. We provide the results obtained by our final system on OntoNotes subsets in order to compare them with earlier works. It has its own set of named entity classes but it has a mapping of those to CoNLL classes. We use the latter for systems comparison. We used the same test/training split as in (Finkel and Manning, 2009).

NLPBA 2004 dataset (Kim et al., 2004) is an English language dataset for bio-medical NER. It consists of a set of PubMed abstracts nad has a correspoding set of named entites (protein, DNA, RNA, cell line, cell type).

4 Feature Set

We performed feature comparison using our system which is a CRF with Viterbi inference. We have also tested greedy inference and have found out that the system performs worse and its results are lower than those of a perceptron with greedy inference that we modeled after (Ratinov and Roth, 2009).

In each of the following subsections we consider a particular type of features. In Subsection 4.1 we deal with local knowledge features which can be extracted from a token (word) being labeled and its surrounding context. Subsection 4.2 describes evaluation of external knowledge features (part-of-speech tags, gazetteers, etc.). Discussion of non-local dependencies of named entities is included in Subsection 4.3. Subsection 4.4 contains further improvements of performance and specific features that do not fall into previous categories; they help to overcome common errors on the CoNLL 2003 dataset.

4.1 Local Knowledge

Our baseline recognizer uses only local information about a current token. It is not surprising that a token-based CRF performs poorly, espe-

Features	Dev	Test
CoNLL-2003		
w_0	25.24%	22.04%
w_{-1}, w_0, w_1	83.41%	74.82%
w_{-1}, w_0, w_1, $w_{-1}\&w_0, w_0\&w_1$	81.20%	72.26%
$w_{-2}, w_{-1}, w_0, w_1, w_2$	82.31%	73.73%
NLPBA 2004		
w_0	-	61.67%
w_{-1}, w_0, w_1	-	65.51%
w_{-1}, w_0, w_1, $w_{-1}\&w_0, w_0\&w_1$	-	66.01%
$w_{-2}, w_{-1}, w_0, w_1, w_2$	-	65.45%

Table 1: Evaluation of context in NER; w — token, $a\&b$ — conjunction of features a and b.

Features	Dev	Test
CoNLL-2003		
w_0	25.24%	22.04%
w_0 + suffixes and prefixes	87.41%	78.59%
$w_0 + s_0$	86.70%	79.16%
$w_0 + s_{-1}, s_0, s_1$, $s_{-1}\&s_0, s_0\&s_1, s_{-1}\&s_0\&s_1$	87.67%	81.37%
All Local Features	88.91%	82.89%
NLPBA 2004		
w_0	-	61.67%
w_0 + suffixes and prefixes	-	66.22%
$w_0 + s_0$	-	62.01%
$w_0 + s_{-1}, s_0, s_1$, $s_{-1}\&s_0, s_0\&s_1, s_{-1}\&s_0\&s_1$	-	65.85%
All Local Features	-	66.83%

Table 2: Evaluation of local features in NER; w — token, s — shape, $a\&b$ — conjunction of features a and b.

cially when we try to model non-unit named-entity chunks[1]. A system which only selects complete unambiguous named entities that appear in training data works better (Tjong Kim Sang and De Meulder, 2003). Table 1 contains performance results of context features. w_0 is a current token, w_1 is a following token and w_{-1} is a preceding one. The larger context we consider the worse F_1-measure we get. Such behavior indicates that token/class dependency statistics in the training corpus is not enough to deduce which context is important. THe quality maximum is observed when we use a sliding window of three tokens. The context can be smoothed by semi-supervised learning algorithms (Ando and Zhang, 2005) in order to compensate for the lack of statistics.

Better results are obtained if we ignore surrounding tokens but use more features based only on the current token. We used suffixes, prefixes[2] and orthographic features (shape[3]) of the current token (see Table 2). Different word-based features give us better evidence of a particular word being a part of a named entity (the gain in F_1 is about 4%) than the context does. It is also useful to extend the shape feature onto surrounding words. The token-based features do not outperform the context features in the biomedical domain but still provide useful information. Biomedical entities are different from newswire entities in terms of shape features; for instance, lower-cased entities (e.g. *persistently infected cells*) are common in the former domain. Domain-specific modifications are required for the shape function (e.g., the regular shapes of the proteins *CD4* and *CD28* are not the same).

4.2 External Knowledge

Most NER systems use additional features like part-of-speech (POS) tags, shallow parsing, gazetteers, etc. Such kind of information requires external knowledge: unlabeled texts, trained taggers, etc. We consider POS tags (Section 4.2.1), words clustering (Section 4.2.2), phrasal clustering (Section 4.2.3), and encyclopedic knowledge (Section 4.2.4). F_1 measures obtained in the experiments covered in this section are shown in Tables 3 and 4; the discussion follows below.

4.2.1 Part-of-Speech Tagging

POS tags are widely used in NER but recently proposed systems omit this information (Ratinov

[1] BILOU scheme is not appropriate for one-token features; an adequate result should be around $F_1 = 52\%$ as in (Klein et al., 2003).

[2] We used prefixes and suffixes of the length up to 6 to reduce the number of features. For example, suffixes like *-burg*, *-land* are highly correlated with location entities (Marburg, Nederland)

[3] Informally, the shape feature is a result of mappings like Bill → Xxx, Moscow-based → Xxx-xx, etc.

and Roth, 2009; Lin and Wu, 2009). POS tagging is itself a challenge and this preprocessing task can take a lot of time. We find that the impact of these features depends on a POS tagger. We replace the original POS tag annotation with the annotation produced by OpenNLP tagger[4], however, even high-quality POS tags lead to a decrease of F_1-measure.

4.2.2 Words Clustering

The authors of (Ando and Zhang, 2005; Suzuki and Isozaki, 2008; Turian et al., 2010) showed that utilization of unlabeled data can improve the quality of NER. We divide the recognizers that use unlabeled text into two groups. The first group consists of semi-supervised systems which directly use labeled and unlabeled data in their training process (Ando and Zhang, 2005; Suzuki and Isozaki, 2008). The second group includes systems that use features derived from unlabeled data (Ratinov and Roth, 2009; Lin and Wu, 2009). (Turian et al., 2010) have shown that recognizers of the first group tend to perform better presumably because they have task-specific information during the training process. However, a simpler way to improve NER quality is to include word representations as features into learning algorithms.

Brown clusters were prepared by the authors of (Turian et al., 2010) by clustering the RCV1 corpus which is a superset of the CoNLL 2003 dataset[5]. Clark clusters were induced by us with the original Clark's code[6] on the same RCV1 corpus but without preprocessing step used in (Turian et al., 2010). Brown clusters were successfully applied in NER (Miller et al., 2004; Ratinov and Roth, 2009). We consider Clark's algorithm since it shows competitive results in unsupervised NLP (Christodoulopoulos et al., 2010; Spitkovsky et al., 2011) and it is also successfully used in NER (Finkel and Manning, 2009). A combination of different word representations (Turian et al., 2010) gives better results. We also applied latent Dirichlet allocation (LDA) to create probabilistic word clustering in the same way as

Features	Dev	Test
CoNLL-2003		
p_0	45.63%	43.98%
$w_0 + p_0$	83.07%	73.42%
b_0	80.98%	75.51%
$w_0 + b_0$	89.35%	82.17%
c_0	67.47%	64.06%
$w_0 + c_0$	86.47%	79.29%
l_0	45.20%	44.24%
$w_0 + l_0$	82.28%	72.63%
g_0	79.90%	76.72%
$w_0 + g_0$	88.36%	81.98%
$b_0 + c_0 + l_0$	86.40%	80.76%
$b_0 + c_0 + l_0 + g_0 + p_0$	89.26%	84.66%
$w_0 + b_0 + c_0 + l_0 + g_0 + p_0$	90.87%	87.00%
NLPBA 2004		
p_0	-	18.29%
$w_0 + p_0$	-	62.81%
b_0	-	30.65%
$w_0 + b_0$	-	63.70%
c_0	-	15.53%
$w_0 + c_0$	-	63.41%
l_0	-	12.81%
$w_0 + l_0$	-	63.42%
g_0	-	43.30%
$w_0 + g_0$	-	63.41%
$b_0 + c_0 + l_0$	-	40.65%
$b_0 + c_0 + l_0 + g_0 + p_0$	-	58.47%
$w_0 + b_0 + c_0 + l_0 + g_0 + p_0$	-	63.52%

Table 3: Evaluation of phrasal and word clusterings in NER; w — token, p — POS tag, c — Clark clusters, b — Brown clusters, l — LDA clusters, g — phrasal clusters. Subscript index stands for the token which clustering label is used. –1 stands for the previous token; +1 stands for the next token; 0 stands for the current token.

[4] http://incubator.apache.org/opennlp/

[5] The resource is available at http://metaoptimize.com/projects/wordreprs/

[6] The code is available at http://www.cs.rhul.ac.uk/home/alexc/

was done in (Chrupala, 2011) and used the most probable cluster label of a word as a feature.

Brown's algorithm is a hierarchical clustering algorithm which clusters words that have a higher mutual information of bigrams (Brown et al., 1992). The output of the algorithm is a dendrogram. A path from the root of the dendrogram represents a word and can be encoded with a bit sequence. We have used prefixes of the length of 7, 11, 13 of such encodings as features (these numbers were selected on the CoNLL development set with a recognizer that used *token* + *Brown* feature), which gave us 10368 clusters.

Clark's algorithm groups words that have similar context distribution and morphological clues starting with most frequent words (Clark, 2003). We induced 200 clusters according to (Finkel and Manning, 2009) and used them as features.

We used LDA wih 100 clusters.

We experimented with a combination of the current token feature and its cluster representation as well as with the representation alone. One interesting observation is that the small space of features (without any tokens) gives results comparable to those of the system which uses tokens with context (on CoNLL 2003 dataset). This trend continues with phrasal clusters (see Table 3).

4.2.3 Phrasal Clustering

Word clustering can be extended onto phrases. Presumably, phrases are far less ambiguous than single words. That consideration was applied to NER by (Lin and Wu, 2009) who presented a scalable k-means clustering algorithm based on Map-Reduce. It is not possible to reproduce their result exactly because they employed private data. In our experiments we used a 1000 soft clusters derived from 10 million phrases from a large web n-gram corpus by a similar k-means algorithm (Lin et al., 2010). N-grams that have high entropy of context are considered phrases. The resource[7] contains phrases and their cluster memberships (up to twenty clusters) along with the similarity to each cluster centroid. We omit similarity information and treat cluster id's as features (with the corresponding prefixes of the BILOU-scheme) for each word in a phrase.

The combination of phrasal clusters and Brown clusters performs well and is only slightly worse than their combination with tokens. Thus, context-based clustering with enough statistical information is able to detect named entity mentions. The NLPBA 2004 dataset is from another domain and the above-mentioned effect is not fully preserved but the clustering still improves performance.

4.2.4 Encyclopedic Knowledge

A simple way to guess whether a particular phrase is a named entity or not is to look it up in a gazetteer. Look-up systems with large entity lists work pretty well if entities are not ambiguous. In that case the approach is competitive against machine learning algorithms (Nadeau, 2007). Gazetteer features are common in machine-learning approaches too and can improve performance of recognition systems (Nadeau and Sekine, 2007; Kazama and Torisawa, 2007). Nowadays there are a lot of web resources which are easily adaptable to NER such as Wikipedia[8], DBPedia[9], and YAGO[10]. We employed Wikipedia and DBPedia.

Wikipedia was successfully used for NER before in (Kazama and Torisawa, 2007; Joel Nothman, 2008; Ratinov and Roth, 2009). Wikipedia contains redirection pages which take the reader directly to a different page. They are often created for synonymous lexical representations of objects or they denote variations in spelling, e.g., the page entitled *International Business Machines* is a redirection page for *IBM*. Disambiguation pages is another kind of Wikipedia pages. Disambiguation pages contain links to other pages which titles are homonyms. For instance, the page *Apple (disambiguation)* contains links to *Apple Inc.* and *Apple (band)*.

We used the titles of Wikipedia pages with their redirects as elements of gazetteers. To get class labels for each Wikipedia page, we employed a classifier proposed by (Tkatchenko et

[7] http://webdocs.cs.ualberta.ca/~bergsma/PhrasalClusters/

[8] http://www.wikipedia.org/

[9] http://dbpedia.org/

[10] http://www.mpi-inf.mpg.de/yago-naga/yago/

Features	Dev	Test
CoNLL-2003		
w_0 + Wikipedia gaz.	56.35%	53.98%
w_0 + Wikipedia gaz. + disambig.	84.73%	77.72%
w_0 + DBpedia gaz.	84.06%	75.40%
w_0 + DBpedia gaz. + disambig.	83.62%	75.14%
w_0 + Wikipedia & DBPedia gaz.	85.21%	78.16%
NLPBA 2004		
w_0 + DBpedia gaz.	-	61.66%

Table 4: Evaluation of encyclopedic resources in NER; w — token.

al., 2011). We combined the semi-automatically derived training set with manually annotated data which enabled us to improve classification results (in order to omit redundant and noisy features we selected classes that correspond to named entity classes in the dataset: PERSON, GPE, ORGANIZATION, FACILITY, GEOLOGICAL REGION, EVENT). Since this markup does not contain NLPBA classes we used Wikipedia only with CoNLL classes.

DBPedia is a structured knowledge base extracted from Wikipedia. The DBpedia ontology consists of 170 classes that form a subsumption hierarchy. The ontology contains about 1.83 million of classified entities (Bizer et al., 2009). We used this hierarchy to obtain high quality gazetteers.

We developed a simple disambiguation heuristic to utilize resources provided by Wikipedia and DBPedia. A disambiguation page defines a set of probable meanings of a particular term. If an ambiguous term co-occurs with an unambiguous term from the same meaning set, the former will be resolved to the meaning of the latter and labeled as an element of the corresponding gazetteer. The results of application of gazetteers is shown in Table 4. Because ambiguity resolving (or wikification) is out of the scope of this paper we do not use it in further experiments including the final system.

We also applied *additional gazetteers* taken from Illinois University parser (Ratinov and Roth, 2009).

4.3 Non-local Dependencies

The same tokens often have the same labels. However, sometimes they may have different labels, for example, *HP* and *HP LaserJet 4200* are not entities of the same type (it is likely that we annotate them as COMPANY and PRODUCT respectively). The latter case is often governed by non-local information. Techniques like two-stage prediction (Krishnan and Manning, 2006), skip-chain CRF (Sutton and McCallum, 2006), and a recognizer which uses Gibbs sampling and penalizes entities with different labels (Finkel et al., 2005) were proposed to account for non-local dependencies in NER. Non-local information is also partially propagated by phrasal and word clusterings. We implemented two approaches which take into account non-local dependencies: two-stage prediction (Krishnan and Manning, 2006) and context aggregation (Ratinov and Roth, 2009).

Two-stage prediction is an approach in which we use the output of a first recognizer to train a second one. For instance, document-based and corpus-based statistic of given token labels is used to re-assign a label to a token (Krishnan and Manning, 2006).

The idea of context aggregation (Ratinov and Roth, 2009) is that if a current token occurs more than once within a window of 200 tokens, we add features to the current token. The features are previous, next, and current tokens of all those extra occurrences. We also performed aggregation of cluster labels for all word and phrasal clusterings that we considered.

We have not performed a separate evaluation of non-local dependencies and tested them only in the final system.

4.4 Minor Features

If we combine the features discussed above (except the non-local dependencies) we get a drastic performance improvement (see Table 5). However, we developed features which correct common errors found on the development and training sets of our benchmarks. Those features were (1) hyphen feature that indicates if a particular token contains a hyphen; (2) sub-tokens feature

Features	Dev	Test
CoNLL-2003		
All features	93.78%	91.02%
NLPBA 2004		
All features	-	74.27%

Table 5: Evaluation of feature combinations.

that adds all sub-tokens of a current token which is hyphened, e.g. *Moscow-based* has sub-tokens *Moscow* and *based*; (3) text-break (expected and unexpected line breaks) feature capturing splits in text; (4) numbers generalization feature, we considered masks of numbers instead of specific numbers according to (Ratinov and Roth, 2009), e.g. 10-12-1996 $\rightarrow$ *DD*-*DD*-*DDDD*, 1999 $\rightarrow$ *DDDD*; (5) a conjunction of the Brown cluster of a current token with the preceding token; (6) capitalized context, which captures additional context of capitalized words, namely, we add a feature that encodes two previous tokens. We use an umbrella term *minor features* to describe this error-fixing list of features.

We also added a two stage prediction in which the first CRF (tuned for a specified task), the boundary recognizer, tries to detect entity boundaries and the second CRF utilizes the output of the first CRF and considers all tokens inside an entity. For example, if we have a phrase ... *blah-blah the Test and County Board Cricket blah-blah-blah* ... and the boundary recognizer has detected that *Test and County Board Cricket* is a potential entity, the second CRF adds all tokens of the potential entity as features when it classifies each token within the entity.

The combination of all proposed features is shown in Table 5. We tested the two-stage prediction approach on this configuration but have not found improvements.

5 Final System

Our final system utilizes all above mentioned features except the two-stage prediction. Each feature set improves performance of the recognizer. We tried to perform optimization by deleting features one by one in order to get the best performing configuration with a smaller set of features. We find that the sequence of deletion steps depends on the initial search space (e.g. if we start optimization procedure without Clark clusters, it will delete the text-break feature; otherwise, it will delete hyphen and sub-tokens features). Table 6 shows the quality of the system with particular features omitted. You can see that the performance of the recognizer is not dramatically reduced in most cases. We believe that it is possible to come up with a smaller feature space or to do feature reweighing (Jiang and Zhai, 2006) in order to improve NER quality and processing speed.

Most of considered features are local and are extracted from a token or its local context. First of all, the behavior of context tokens as features is preserved for both datasets. A small sliding window of three tokens is good enough. Second, the word-based features behavior is not persistent and depends on the specificity of entities. Nevertheless, names contain morphological clues that distinguish them from common words. Comparing token-based with word-based features you might see that token-derived information gives a gain of at least four points of F_1-measure for newswire corpus and can be on the same level for biomedical domain. Third, clustering could be considered as feature reduction process (Saha et al., 2009); it helps to overcome the lack of statistics. Using only clustering representations hypothesis on the reduced space of features can be useful in recognition and works even better than token-based features. Last but not least, gazetteers are still useful for NER, especially when we have such freely available resources as Wikipedia and DBPedia. Disambiguation approaches in gazetteer matching could bring radical improvements.

Two tables compare our results with the best reported systems on the CoNLL 2003 (Table 7), OntoNotes version 4 (Table 8), and NLPBA (Table9) datasets.

We used approximate randomization test (Yeh, 2000) with 100000 iterations to compare our system to (Ratinov and Roth, 2009). The test checks if a randomely sampled mixture of the outputs of the baseline algorithm and the one being tested performs better than the basline algorithm. Our improvement over the top performing competitor is statistically significant with p-value 0.0001. Unfortunately, we could not compare with (Lin and Wu, 2009) because their system uses propri-

Feature	Dev	Test	Test+
—	93.78	**91.02**	74.27
Capitalized context*	**93.82**	90.66	74.24
Clark aggregation	93.80	90.66	74.01
Hyphen*	93.78	90.84	74.11
Brown + token*	93.66	90.79	74.22
Sub-tokens*	93.66	90.74	74.20
POS tag	93.65	90.82	74.07
Numbers gen.*	93.65	90.65	74.25
Text break features*	93.60	90.68	74.06
Additional gazzetters	93.56	90.17	74.24
Context aggregation	93.55	90.42	74.10
Brown aggregation	93.52	90.29	74.27
Current token	93.51	90.74	74.21
Clark cluster	93.49	90.35	74.23
Affixes	93.47	90.63	74.08
DBPedia gazetteers	93.46	90.53	74.27
Tokens in window	93.35	90.37	74.30
LDA cluster	93.34	90.46	74.18
Brown cluster	93.26	90.25	74.20
Phrasal cluster	93.12	90.43	**74.75**
Tokens in entity	93.12	90.43	74.11
Wikipedia gazetteers	93.12	90.43	n/a
Shape	92.83	89.75	74.02

Table 6: Evaluation of omitting of features on the CoNLL 2003 development (Dev) and test (Test) sets and on NLPBA test set (Test+). All F_1 values are in %. "*" indicates minor feature

System	F_1-measure
Our system	**91.02%**
(Lin and Wu, 2009)	90.90%
(Ratinov and Roth, 2009)	90.80%
(Ciaramita and Altun, 2005)	90.80%
Tjong Kim Sang 2003	90.30%
(Suzuki and Isozaki, 2008)	89.92%
(Ando and Zhang, 2005)	89.31%
(Florian et al., 2003)	88.76%

Table 7: Comparison of recognizers on the CoNLL 2003 benchmark. Tjong Kim Sang stands for (Tjong Kim Sang and De Meulder, 2003).

	Finkel	Ratinov	Our system
ABC	74.91%	72.84%	**76.75%**
CNN	78.70%	79.27%	**81.40%**
MNB	66.49%	**73.10%**	71.52%
NBC	**67.96%**	65.78%	67.41%
PRI	**86.34%**	79.63%	83.72%
VOA	**88.18%**	84.93%	87.12%

Table 8: Comparison of F_1-measures of recognizers on the OntoNotes version 4 benchmark. Finkel — (Finkel and Manning, 2009); Ratinov — (Ratinov and Roth, 2009)

System	F_1-measure
(Wang et al., 2008)	77.6%
Our system	74.27%
(Zhou and Su, 2004)	72.6%
(Finkel et al., 2004)	70.1%
(Settles, 2004)	69.8%
(Saha et al., 2009)	67.4%

Table 9: Comparison of recognizers on NLPBA 2004 benchmark.

etary data and its output is also unavailable.

6 Conclusions

In this paper we have analyzed a comprehensive set of features used in supervised NER. We have considered the impact of various individual features and their combinations on the effectiveness of NER. We have also built a CRF-based supervised NER system that achieves 91.02% F_1-measure on the CoNLL 2003 dataset and 81.4% F_1-measure on the OntoNotes version 4 CNN dataset and demonstrated that the performance boost over the earlier top performing system is statistically significant on the benchmarks. We have also considered novel features for NER, namely a DBPedia markup and a phrasal clustering from Google n-grams corpus.

We plan to extend the work on clustering features which we find very promising for NER. We have obtained a large proprietary newswire corpus from a media corporation and plan to utilize it in our further experiments on enhancing NER in the newswire domain. We also consider exploring features useful for specific entity classes.

References

Rie Kubota Ando and Tong Zhang. 2005. A high-performance semi-supervised learning method for text chunking. In Kevin Knight, Hwee Tou Ng, and Kemal Oflazer, editors, *ACL*. The Association for Computer Linguistics.

Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. 2009. Dbpedia - a crystallization point for the web of data. *J. Web Sem.*, 7(3):154–165.

Peter F. Brown, Vincent J. Della Pietra, Peter V. Desouza, Jennifer C. Lai, and Robert L. Mercer. 1992. Class-Based n-gram Models of Natural Language. *Computational Linguistics*, 18(4):467–479.

Christos Christodoulopoulos, Sharon Goldwater, and Mark Steedman. 2010. Two decades of unsupervised pos induction: how far have we come? In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 575–584, Stroudsburg, PA, USA. Association for Computational Linguistics.

Grzegorz Chrupala. 2011. Efficient induction of probabilistic word classes with LDA. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 363–372, Chiang Mai, Thailand, November. Asian Federation of Natural Language Processing.

M. Ciaramita and Y. Altun. 2005. Named-entity recognition in novel domains with external lexical knowledge. In *Proceedings of the NIPS Workshop on Advances in Structured Learning for Text and Speech Processing*.

Alexander Clark. 2003. Combining distributional and morphological information for part of speech induction. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics - Volume 1*, EACL '03, pages 59–66, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jenny R. Finkel and Christopher D. Manning. 2009. Joint parsing and named entity recognition. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, pages 326–334, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jenny Finkel, Shipra Dingare, Huy Nguyen, Malvina Nissim, Christopher Manning, and Gail Sinclair. 2004. Exploiting context for biomedical entity recognition: From syntax to the web. In *In Proceedings of the Joint Workshop on Natural Language Processing in Biomedicine and its applications (JNLPBA-2004*.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 363–370, Stroudsburg, PA, USA. Association for Computational Linguistics.

Radu Florian, Abe Ittycheriah, Hongyan Jing, and Tong Zhang. 2003. Named entity recognition through classifier combination. In Walter Daelemans and Miles Osborne, editors, *Proceedings of CoNLL-2003*, pages 168–171. Edmonton, Canada.

Eduard H. Hovy, Mitchell P. Marcus, Martha Palmer, Lance A. Ramshaw, and Ralph M. Weischedel. 2006. Ontonotes: The 90 In Robert C. Moore, Jeff A. Bilmes, Jennifer Chu-Carroll, and Mark Sanderson, editors, *HLT-NAACL*. The Association for Computational Linguistics.

Jing Jiang and Chengxiang Zhai. 2006. Exploiting domain structure for named entity recognition. In *In Human Language Technology Conference*, pages 74–81.

James R. Joel Nothman. 2008. Transforming Wikipedia into Named Entity Training Data. pages 124–132.

Jun'ichi Kazama and Kentaro Torisawa. 2007. Exploiting Wikipedia as External Knowledge for Named Entity Recognition. In *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 698–707.

Jin D. Kim, Tomoko Ohta, Yoshimasa Tsuruoka, Yuka Tateisi, and Nigel Collier. 2004. Introduction to the bio-entity recognition task at JNLPBA. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, JNLPBA '04, pages 70–75, Stroudsburg, PA, USA. Association for Computational Linguistics.

Dan Klein, Joseph Smarr, Huy Nguyen, and Christopher D. Manning. 2003. Named entity recognition with character-level models. In Walter Daelemans and Miles Osborne, editors, *Proceedings of CoNLL-2003*, pages 180–183. Edmonton, Canada.

Vijay Krishnan and Christopher D. Manning. 2006. An effective two-stage model for exploiting non-local dependencies in named entity recognition. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, ACL-44, pages 1121–1128, Stroudsburg, PA, USA. Association for Computational Linguistics.

Dekang Lin and Xiaoyun Wu. 2009. Phrase Clustering for Discriminative Learning. In *Proceedings of*

the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, pages 1030–1038, Suntec, Singapore, August. Association for Computational Linguistics.

Dekang Lin, Kenneth Church, Heng Ji, Satoshi Sekine, David Yarowsky, Shane Bergsma, Kailash Patil, Emily Pitler, Rachel Lathbury, Vikram Rao, Kapil Dalwani, and Sushant Narsale. 2010. New Tools for Web-Scale N-grams.

Andrew McCallum and Wei Li. 2003. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pages 188–191, Stroudsburg, PA, USA. Association for Computational Linguistics.

Pablo N. Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. 2011. Dbpedia spotlight: Shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems (I-Semantics)*.

Scott Miller, Jethran Guinness, and Alex Zamanian. 2004. Name tagging with word clusters and discriminative training. In Daniel Marcu Susan Dumais and Salim Roukos, editors, *HLT-NAACL 2004: Main Proceedings*, pages 337–342, Boston, Massachusetts, USA, May 2 - May 7. Association for Computational Linguistics.

David Nadeau and Satoshi Sekine. 2007. A Survey of Named Entity Recognition and Classification.

David Nadeau. 2007. *Semi-supervised named entity recognition: learning to recognize 100 entity types with little supervision*. Ph.D. thesis, Ottawa, Ont., Canada, Canada. AAINR49385.

L. Ratinov and D. Roth. 2009. Design challenges and misconceptions in named entity recognition. In *CoNLL*, 6.

Sujan Kumar Saha, Sudeshna Sarkar, and Pabitra Mitra. 2009. Feature selection techniques for maximum entropy based biomedical named entity recognition. *J. of Biomedical Informatics*, 42:905–911, October.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*, pages 142–147, Morristown, NJ, USA. Association for Computational Linguistics.

Burr Settles. 2004. Biomedical named entity recognition using conditional random fields and rich feature sets. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, JNLPBA '04, pages 104–107, Stroudsburg, PA, USA. Association for Computational Linguistics.

Valentin I. Spitkovsky, Hiyan Alshawi, Angel X. Chang, and Daniel Jurafsky. 2011. Unsupervised dependency parsing without gold part-of-speech tags. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP 2011)*.

Charles Sutton and Andrew McCallum. 2006. An Introduction to Conditional Random Fields for Relational Learning. In L. Getoor and B. Taskar, editors, *Introduction to Statistical Relational Learning*. MIT Press.

Jun Suzuki and Hideki Isozaki. 2008. Semi-Supervised Sequential Labeling and Segmentation Using Giga-Word Scale Unlabeled Data. In *Proceedings of ACL-08: HLT*, pages 665–673, Columbus, Ohio, June. Association for Computational Linguistics.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In Walter Daelemans and Miles Osborne, editors, *Proceedings of CoNLL-2003*, pages 142–147. Edmonton, Canada.

Maksim Tkatchenko, Alexander Ulanov, and Andrey Simanovsky. 2011. Classifying wikipedia entities into fine-grained classes. In *ICDE Workshops*, pages 212–217.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 384–394, Stroudsburg, PA, USA. Association for Computational Linguistics.

Haochang Wang, Tiejun Zhao, Hongye Tan, and Shu Zhang. 2008. Biomedical named entity recognition based on classifiers ensemble. *IJCSA*, 5(2):1–11.

Alexander Yeh. 2000. More accurate tests for the statistical significance of result differences.

GuoDong Zhou and Jian Su. 2004. Exploring deep knowledge resources in biomedical name recognition. In Nigel Collier, Patrick Ruch, and Adeline Nazarenko, editors, *COLING 2004 International Joint workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP) 2004*, pages 99–102, Geneva, Switzerland, August 28th and 29th. COLING.