Initial explorations on using CRFs for Turkish Named Entity Recognition

Gökhan Akın Şeker[1] Gülşen Eryiğit[2]
(1) ITU Informatics Institute Graduate Program in Computer Science, Istanbul Technical University
Istanbul, 34469, Turkey
(2) Department of Computer Engineering, Istanbul Technical University Istanbul, 34469, Turkey
{sekerg, gulsen.cebiroglu}@itu.edu.tr

ABSTRACT
This paper reports the highest results (95% in MUC and 92% in CoNLL metric) in the literature for Turkish named entity recognition; more specifically for the task of detecting person, location and organization entities in general news texts. We give an in depth analysis of the previous reported results and make comparisons with them whenever possible. We use conditional random fields (CRFs) as our statistical model. The paper presents initial explorations on the usage of rich morphological structure of the Turkish language as features to CRFs together with the use of some basic and generative gazetteers.

KEYWORDS: Named Entity Recognition , Turkish , Conditional Random Fields , ENAMEX.

Proceedings of COLING 2012: Technical Papers, pages 2459–2474,
COLING 2012, Mumbai, December 2012.

1 Introduction

Named Entity Recognition (NER) can be basically defined as identifying and categorizing certain type of data (i.e. person, location, organization names, date-time expressions). NER is an important stage for several natural language processing (NLP) tasks including machine translation, sentiment analysis and information extraction. MUC (Sundheim, 1995; Chinchor and Marsh, 1998) and CoNLL (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003) conferences define three basic types of named entities; these are 1- ENAMEX (person, location and organization names), 2- TIMEX (date and time entities) and 3- NUMEX (numerical expressions like money and percentages). But NER research is not limited to only these types; different application areas concentrate to determining alternative entity types such as protein names, medicine names, book titles.

The NER research was firstly started in early 1990s for English. In 1995, with the high interest of the research community, the success rates for English achieved nearly the human annotation performance on news texts (Sundheim, 1995). Nadeau and Sekine (2007) gives a survey of the research for English NER between 1991 to 2006. The satisfaction on English NER task directed the field to new research areas such as multilingual NER systems (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003), NER on informal texts (LIU et al., 2011; Rüd et al., 2011; Mohit et al., 2012), transliteration (Zhang et al., 2012) and coreference (Na and Ng, 2009) of named entities .

Conditional Random Fields (Lafferty et al., 2001) is a very popular method used in NLP. It is also widely used for named entity recognition task in various domains (LIU et al., 2011; Ekbal and Bandyopadhyay, 2009; Settles, 2004). Stanford NER (Finkel et al., 2005) which is a well-known NER tool also uses CRFs as its machine learning method.

Morphologically rich languages pose interesting challenges for NLP tasks as it is the case for NER (Hasan et al., 2009). Turkish being one of such languages attracts attention of the NLP community. Nevertheless, the results for Turkish NER remain still very behind the reported accuracies for English. The first published work on this topic is Cucerzan and Yarowsky (1999) which is a language independent system tested on Romanian, English, Greek, Turkish and Hindi. This system is trained with a small training data and learns from unannotated text using a bootstrapping algorithm. It reports an F-Measure of 53.04% for Turkish. The other studies[1] on Turkish NER are as follows: Tür et al. (2003), Bayraktar and Temizel (2008), Yeniterzi (2011), Özkaya and Diri (2011), Tatar and Cicekli (2011) and Küçük and Yazıcı (2012). Although some of these studies try to use CRFs for Turkish NER task, this is the first study which introduces a successful CRF model which beats the state of the art results for this problem. Yeniterzi (2011) uses CRFs with an IG[2] based tokenization. Özkaya and Diri (2011) reports 84% F-measure on e-mail messages by using CRFs, but since they are using features specific to email domain only (such as from, subject fields) their work may not be extended to general texts.

This is an initial study which aims to propose a successful CRF model for Turkish NER. With this purpose, we firstly focused on general news domain where there exists many reported results for comparison. We obtained the highest scores in the literature on ENAMEX types. We made an initial exploration on the use of morphological features and gazetteers. But we believe there is still room for improvement by using more expansive gazetteers and using the

[1] The studies are investigated in more detail in section §5.3 for comparison.

[2] IG is used for inflectional units smaller than words.

morphological features more efficiently. As a future work, we want to work on TIMEX and NUMEX types as well as informal texts.

This paper is organized as follows. §2 gives brief information about Turkish, §3 explains our framework, §4 gives information about datasets, evaluation metrics and features, §5 gives our experiments and evaluates the results by comparing with related work and §6 gives the conclusion.

2 Turkish

This section briefly states the characteristics of the Turkish language which we believe have impact on the NER task.

Turkish is a morphologically rich and highly agglutinative language. In most of the Turkish NLP studies, lemmas are used instead of word surface forms in order to decrease lexical sparsity. For example a Turkish verb "gitmek" (*to go*) may appear in hundreds of different surface forms[3] depending on the tense, mood and the person arguments whereas the same verb in English has only five different forms (going, go, goes, went, gone). But for the proper nouns, in formal texts, the inflectional suffixes are separated from the lemma by an apostrophe. As a result, although it seems that it is unnecessary to make an automatic morphological processing for the stemming of the proper nouns, the stemming of the surrounding words of the proper nouns has influence on the success of NER. §5 investigates the impact of using lexical information for the named entity recognition task.

Turkish person (first) names are usually selected from the words used in daily conversation such as İpek*(silk)*, Kaya*(rock)*, Pembe*(pink)*, Çiçek *(flower)*. Only the proper nouns and the initial words of the sentences start with an initial capital letter.

Turkish is a free word order language, so the position of the word in a sentence doesn't give us information about being a named entity or not. All of the three sentences: "Ahmet yarın Mehmet ile konuşmaya gidecek.", "Yarın Mehmet ile konuşmaya Ahmet gidecek" and "Yarın Ahmet, Mehmet ile konuşmaya gidecek." are valid Turkish sentences all with the English translation of "Tomorrow, Ahmet will go to talk to Mehmet".

3 Proposed Framework

This section describes our CRF based NER framework trained using morphological features and gazetteers. Figure 1 shows the outline of the framework.

3.1 Tokenization

We tokenized our data so that each word is represented as a token except for proper nouns which go under inflection. Since the suffixes separated by an apostrophe are not part of the named entities(NEs), we partitioned such proper nouns into two tokens (the tokens before and after the apostrophe.) All punctuation characters are considered as a token. Sentences are separated from each other by an empty line. Tokenization of a sample sentence can be seen in Table 2.

[3] Some surface forms of "gitmek" (only in simple present tense for different person arguments): gidiyorum, gidiyorsun, gidiyor, gidiyoruz, gidiyorsunuz, gidiyorlar.

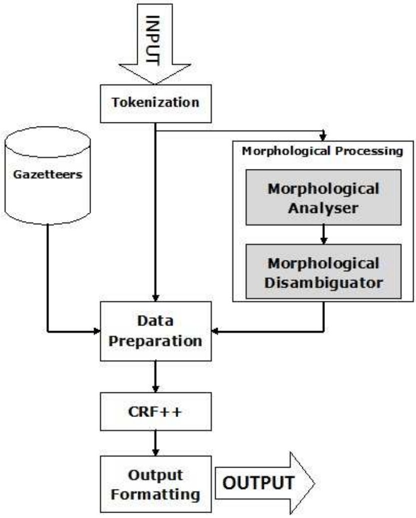

Figure 1: Proposed Framework

3.2 Morphological Processing

We used a two-level morphological analyzer (Oflazer, 1994) for producing the possible analyses for each word. We then give the output to a morphological disambiguator (Sak et al., 2008) in order to get the most probable analysis in the given context. For example, the analyzer produces three different possible analyses for the word "Teknik"(*Technical*) which corresponds to an adjective, a noun and a proper noun accordingly; the disambiguator selects most probable one for us :

Teknik teknik+Adj

Teknik teknik+Noun+A3sg+Pnon+Nom

Teknik teknik+Noun+Prop+A3sg+Pnon+Nom

The output of the analyzer both includes the stem of the word and the morphological features[4] which we use as features for our CRF model. One should keep in mind that, this is an automatic processing and it possesses its own error margin. Eryiğit (2012) gives the performance of this morphological pipeline on raw data.

[4] The abbreviations after the plus sign stand for: +Adj: Adjective, +Noun: Noun, +A3sg: 3sg number-person agreement, +Pnon: Pronoun (no overt possessive agreement), +Nom: Nominative case, +Prop: Proper noun

3.3 Gazetteers

In this work, we prepared two kind of gazetteers[5] which we call base and generator gazetteers. Table 1 gives the details for each one. Base gazetteers are the ones which include words with high probability of occurring in a named entity. These are large gazetteers with thousands of tokens. We collected person names from different sources. We split them into first name and surname gazetteers in order to be able to detect different combinations of these. We compiled the location gazetteer so that it includes all location names in Turkish postal code system[6] , all country names from international telephone code system[7], city and states of those countries[8] and geographical names from different sources.

	Gazetteer	# of tokens
Base	First names	44.048
	Surnames	138.844
	Location names	33.551
Generator	Location	44
	Organization	60
	Person	22

Table 1: # of distinct tokens in gazetteers

Our generator gazetteers are relatively small compared to the base gazetteers. They include the stems of some basic named entity generator words. To give an example: the stem "bakanlık" (*ministry*) which could come after some regular words such as spor, tarım (*sports, agriculture*) to construct organization NEs such as "Tarım Bakanlığı" (*Ministry of Agriculture*).

3.4 Data Preparation

In this stage, we use the information coming from the raw data, the gazetteers and the morphological processing in order to prepare the feature vectors for our training/test instances. For the related class labels at the training stage, we use "Raw Tags". In this format, we use the labels "PERSON", "ORGANIZATION", "LOCATION" and "O" (other - for the words which do not belong to a NE) without any position information (that is without any prefix). In §5.1, we also give the results of our experiments of using different formats for class labels.

3.5 Conditional Random Fields

Conditional random fields (CRFs) (Lafferty et al., 2001) is a framework for building probabilistic models to segment and label sequence data. CRFs offer several advantages over hidden Markov models (HMMs), stochastic grammars and maximum entropy Markov models (MEMMs). CRF is a discriminative model better suited to including rich, overlapping features focusing solely on the conditional distribution $p(\mathbf{y}|\mathbf{x})$. We use linear chain CRFs where $p(\mathbf{y}|\mathbf{x})$ is defined as:

$$p_\theta(\mathbf{y}|\mathbf{x}) = \frac{1}{Z_\theta(\mathbf{x})} \exp\left\{ \sum_{t=1}^{T} \sum_{k=1}^{K} \theta_k f_k(y_{t-1}, y_t, x_t) \right\} \tag{1}$$

[5] available from http://web.itu.edu.tr/gulsenc/ner.html

[6] https://interaktifkargo.ptt.gov.tr/posta_kodu/

[7] http://www.ttrehber.turktelekom.com.tr/trk-web/ulkekodlari.html

[8] mostly collected from wikipedia.com

where $f_k(y_{t-1}, y_t, x_t)$ is the function for the properties of transition from the state y_{t-1} to y_t with the input x_t and θ_k is the parameter optimized by the training. $Z_\theta(\mathbf{x})$ is a normalization factor calculated by:

$$Z_\theta(\mathbf{x}) = \sum_{\mathbf{y}\in Y^T} \exp\left\{ \sum_{t=1}^{T} \sum_{k=1}^{K} \theta_k f_k(y_{t-1}, y_t, x_t) \right\} \tag{2}$$

For the named entity task, each state y_t is the named entity label and each feature vector x_t contains all the components of the global observations x that are needed for computing features at time t. Sutton and McCallum (2011) gives detailed information on mathematical foundations and many examples about the usage of CRFs. In this study we used CRF++[9] which is an open source implementation of CRFs.

3.6 Output Formatting

Our system has the capability of labeling the output with two different type of tags: 1. Raw tags and 2. IOB2 tags. Raw tag format which is introduced in §3.4 is also used during the training. The experiments related to the selection of the training format is given in §5.1. IOB2 (Tjong Kim Sang, 2002) is one of the most common formats used in the literature for labeling named entities. I-O-B denotes **I**n, **O**ut and **B**egin. In this tagging format, the first token of a NE is labeled with a “B-” prefix while other words in NE are labeled with an “I-” prefix. Tokens that are not part of any NE are tagged with the label “O”. Table 2 shows a sample tokenized sentence (*Mustafa Kemal Atatürk went to Samsun in 1919.*) tagged in both formats.

Token	IOB2 Tags	RAW Tags
Mustafa	B-PERSON	PERSON
Kemal	I-PERSON	PERSON
Atatürk	I-PERSON	PERSON
1919	O	O
yılında	O	O
Samsun	B-LOCATION	LOCATION
’a	O	O
çıktı	O	O
.	O	O

Table 2: IOB2 tagging vs RAW tagging

4 Configuration

This section gives detailed information about the used datasets, evaluation metrics, feature categories and feature templates. §4.3 (Feature categories) presents the features which are provided for each token in our input file. §4.4 (Feature templates) presents the templates used for creating CRF feature vectors for each instance, using the given categories for the current token and its context together with some combinations of these.

[9] http://crfpp.googlecode.com/svn/trunk/doc/index.html

4.1 Data Sets

In our experiments we used the data from Tür et al. (2003). This data consists of 500K words and is annotated only for ENAMEX types with 24,101 person names, 16,105 location names and 13,540 organization names. We reserved one tenth of the data (47,344 words) for testing and used the remaining for training purposes by exactly the same way in Yeniterzi (2011).

4.2 Evaluation Metrics

There are two main metrics in the literature for the evaluation of NER systems: CoNLL and MUC. The MUC metric is the average F-Measure of MUC TEXT and MUC TYPE. MUC TYPE evaluates the performance of assigning the correct NE type to each word without taking into account if the NE boundaries are detected correctly. MUC TEXT makes evaluation only on NE boundaries without looking if the correct NE type is assigned or not. The CoNLL metric on the other hand evaluates an assignment to be correct if both the type and the boundary of a NE is determined correctly. The details of the calculation for these metrics may be investigated from Nadeau and Sekine (2007).

Although CoNLL metric seems to be preferred in recent studies, we evaluate[10] our results using both CoNLL and MUC in order to be able to make comparisons with previous works.

4.3 Feature Categories

In our **base model (BM_surf)** we used word tokens converted to lower case in their surface form. The idea behind converting tokens to lowercase is avoiding one of the major problems of the Turkish language studies; the sparse data problem. Other features added to this model can be grouped into three main categories: morphological, lexical and gazetteer lookup features.

4.3.1 Morphological Features

The morphological features are extracted from the analysis produced after the automatic morphological processing of each word.

Stem : The stem information. For the inflected proper nouns where the inflections after the apostrophe are treated as a separate token, the same surface form after the apostrophe is assigned as the stem of the token representing inflections.

Part of Speech Tag (POS) : The final part of speech category for each word. In Turkish, with the use of derivations, words may change their part of speech categories within a single surface form. The final form of the word determines its syntactic role within a sentence. Therefore, we use the final POS form of each word. We assigned a special POS tag ("APOST") to the tokens separated by an apostrophe from the proper nouns.

Noun Case (NCS) : The case argument. This feature is 0 for non nominal tokens and one of the following values for nominals: Nominative(NOM), Accusative/Objective(ACC), Dative (DAT), Ablative(ABL), Locative(LOC), Genitive(GEN), Instrumental(INS), Equative(EQU). Ex: the value will be NOM for the word "Teknik" with the morphological analysis "teknik+Noun+Prop+A3sg+Pnon+Nom".

[10] We use the evaluation script from CoNLL 2000 shared task (http://www.cnts.ua.ac.be/conll2000/chunking/output.html) for CoNLL and MUC TYPE scores (with the option "-r").

Proper Noun (PROP) : A binary feature indication that the "+Prop" tag exists (1) in the selected morphological analysis or not (0). Ex: The value will be 1 for the word "Teknik" given above. It is useful to mention that the morphological pipeline tags all unknown words as proper nouns.

All Inflectional Features (INF) : All inflectional tags after the POS category. If a derivation exists then the inflectional tags after the last derived POS category is used. Ex: the value will be "Prop+A3sg+Pnon+Nom" for the word "Teknik" with the above morphological analysis.

4.3.2 Lexical Features

Case Feature (CS) : The information about lowercase and uppercase letters used in the current token. This feature takes 4 different values: lowercase(0), UPPERCASE(1), Proper Name Case(2) and miXEd CaSe(3)

Start of the Sentence (SS) : A binary feature indicating that the current token is the beginning of a sentence (1) or not (0).

4.3.3 Gazetteer Lookup Features

Six different features used for each of the six gazetteers introduced in §3.3. Lookup features for base gazetteers (BG) have a 1 value if the token exists in the corresponding gazetteer and 0 otherwise. Generator gazetteer lookup features (GG) are binary features as well but this time the stem of the word is checked instead of the full surface form.

4.4 Feature Templates

CRFs are log-linear models. In order to take advantage of the useful feature combinations, one needs to provide these as new features to the CRFs. In some studies, it is shown that the useful feature conjunctions may be determined incrementally and provided to the system automatically (McCallum, 2003). But, in this study, we used the approach proposed in Sha and Pereira (2003) and selected useful features manually for our initial explorations. Although this approach generally results with a huge number of features, we didn't have any memory problem by using the combinations.

We provided our atomic features within a window of {-3,+3} and some selected combinations of these as feature templates to CRF++. Two sample feature templates are given in the below example. The templates are given in [pos,col] format, where pos stands for the relative position of the token in focus and col stands for the feature column number in the input file.

$$U15 : \%x[-2,2]$$

$$U50 : \%x[0,10]/\%x[0,6]$$

U15 is the template for using the 2nd feature (part-of-speech tag) of the second previous word. U50 is the template for using the conjunction of the existence of the current word in the location name gazetteer (LG) (col=10) and its case feature (col=6) such as *exists in LG written in lowercase; exists in LG and the first letter is capitalized;...*

We use the bigram option of the CRF++ in order to automatically generate the edge features using the previous label y_{-1} and the current label y_0.

As a result, for the 14 feature categories presented in the previous section, we formulated 92 feature templates[11] in our final model which resulted to $\sim$ 20M binary features.

5 Experiments & Evaluation

This section comprises the experiments we conducted to make the decision for our training format (§5.1), to measure the impact of our proposed models (§5.2) and to compare our best model with previous studies (§5.3).

5.1 Training format experiments

We experimented with different training data formats. These are IOB, IOB2, raw labels and fictitious boundary model of Tür et al. (2003). In all of these experiments, we converted the test output to IOB-2 style and evaluated in the same manner. This conversion is made in a straightforward manner by assuming the consecutive tokens labeled with the same NE type as the part of the same NE. This is an acceptable assumption since in Turkish, words in a sentence are separated with a comma if they have the same syntactic function. And also, the subject of a sentence is written separately by the use of a comma if it appears in a confusing context. Tür et al. (2003) gives the probability of two consecutive tokens to have both "B-PERSON" tags as 0.006076. Our training and test sets do not include such NEs at all.

Our experiments show that, we obtain the highest performance by using the RAW labels whereas using the IOB formats reduces the performance by 0.4% and the fictitious boundary format by 2% in all metrics in our base model.

5.2 Feature-related experiments

	MUC TYPE				MUC TEXT	MUC
Feature	PER	ORG	LOC	Overall		
BM_stem	85.31	79.89	86.87	84.03	83.95	83.54
BM_surf	83.83	82.71	86.67	84.19	85.81	85.00
+STEM	85.62	83.26	87.26	85.30	87.08	86.19
+POS	87.34	83.08	87.47	86.06	88.11	87.09
+NCS	87.46	83.95	87.27	86.33	88.85	87.59
+PROP	88.87	85.12	88.68	87.68	90.98	89.33
+INF	89.65	85.32	89.79	88.38	91.60	89.99
+CS	92.76	89.09	89.85	90.92	94.73	92.83
+SS	92.75	89.01	90.15	90.97	94.68	92.83
+BG	94.00	89.82	92.20	92.27	95.50	93.89
+GG	**94.81**	**91.09**	**93.35**	**93.29**	**95.89**	**94.59**

Table 3: F-measure in MUC TYPE, MUC TEXT and MUC Metrics

In our base model (BM_surf) we trained CRF using only one feature; the surface form of the word that appeared in the sentence. Tokens are converted to lowercase to avoid sparse data

[11] The used templates and our NER tool is available from http://web.itu.edu.tr/gulsenc/ner.html

Feature	PER	ORG	LOC	Overall
BM_stem	81.84	74.94	86.82	81.78
BM_surf	80.77	77.86	87.66	82.28
+STEM	82.76	78.78	87.95	83.47
+POS	84.75	78.16	88.39	84.33
+NCS	84.65	79.08	88.00	84.38
+PROP	85.90	80.61	89.20	85.69
+INF	86.71	81.97	89.88	86.59
+CS	90.65	86.12	90.74	89.59
+SS	90.46	85.95	91.01	89.55
+BG	92.23	87.28	92.14	91.02
+GG	**92.94**	**88.77**	**92.93**	**91.94**

Table 4: F-measure in CONNL Metric

problem. i.e. Our training data includes the name of a city "AKSARAY" as a token only one time and doesn't include "Aksaray". This usage converted both to the same token "aksaray". The disadvantage of this usage is losing the effect of an important property of proper nouns in Turkish; the first letter of a proper noun is capitalized so that the noun "gül" (rose) differs from the person name "Gül". Effect of converting all tokens to lowercase is about 1% F-Measure loss in base model in our experiments, but this loss is recoverable with added case (CS) feature. We also wanted to see the performance of using only the word stems instead of the surface form in the sentence so generated a second model (BM_stem).

Table 3 and 4 give detailed evaluation results in all metrics. A plus(+) sign before the feature name indicates that this feature is added (together with suitable feature templates) to the model at the above line. These tables also show the contribution of each feature. While adding a new feature to the model, we also made experiments using different feature template combinations for this new feature to determine the best n-gram relations and their interaction with the previously added features. These experiments are not added to the paper due to the space constraints.

From Table 3 and 4, it can be seen that BM_surf has higher performance than BM_stem. But using the stem information (+STEM) together with the BM_surf raises the performance. This supports our claim about the influence of the stemming of the surrounding words in §2.

All added morphological features raised the performance. But the performance gain acquired by the +INF feature is surprising. One should keep in mind that the +NCS and +PROP features are the atomic units extracted from the +INF feature. Since Turkish is an agglutinative language, the possible number of different values for the +INF feature is very high. For this reason, in many recent studies the usage of the inflectional features as a block[12] is not a preferred approach. We think this result indicates that there is still room for improvement using more fine grained usage of these inflectional features.

The high increase obtained by the +CS feature is as expected due to the known disadvantage of our lowercased base model. But the low performance (and some negative effects) of the feature SS is surprising.

Our base gazetteer features (+BG) performed under our expectations but added a gain of

[12] many atomic features concatenated to each other

about 1% in all metrics. An interesting observation about these features is the gain on organization name identification performance. We don't have any gazetteer of organization names, but using our gazetteers raised the number of identified person and location names. As a result, the number of false positive identifications of organization names decreased.

Impressive performance of the generator gazetteers (+GG) with their modest sizes encourages us adding more of such gazetteers for future work.

5.3 Comparison with related work

This section tries to make a detailed analysis on the related studies: At the time of writing of this paper none of the tools were publicly available so that it wasn't possible to train and test them on the same data set. Table 5 gives the reported results of each related work. We give the results of our pairwise comparisons in the running text whenever possible.

Related work	Best Result	Ev.Metr.	Domain	NE Types
Özkaya and Diri (2011)	84.24	*n/a*	E-mail texts	ENAMEX
Küçük and Yazıcı (2012)	90.13	OTHER	General news	ENAMEX,TIMEX,NUMEX
Tür et al. (2003)	91.56	MUC	General news	ENAMEX
Bayraktar and Temizel (2008)	81.97	MUC	Financial Texts	PERSON NAMES
OURS	**94.59**	**MUC**	**General news**	**ENAMEX**
Tatar and Cicekli (2011)	91.08	CoNLL	Terrorism news	ENAMEX,TIMEX
Yeniterzi (2011)	88.94	CoNLL	General news	ENAMEX
OURS	**91.94**	**CoNLL**	**General news**	**ENAMEX**

Table 5: Comparison with related work (The reported results in each paper)

The performances listed in Table 5 is organized in decreasing order of credit given to partial matches during evaluation. Most of the results are on MUC and CoNLL metrics, therefore we listed our results twice in both of these. Note that the test sets, evaluation metrics (3^{rd} column), working domain (4^{th} column) and entity types[13] (5^{th} column) in focus of each work are different from each other. Table 5 tries to give an overview of these features for each work.

The first NER work specific to Turkish is Tür et al. (2003). The study focuses on three Information Extraction (IE) tasks, namely, sentence segmentation, topic segmentation and name tagging. For name tagging task they use lexical, morphological and contextual features of the words to generate an HMM based model. They evaluate their results in MUC metrics. They use the same training data, but different test data which is not available. Their performance (91.56%) given in Table 5 are comparable with our result in MUC metrics (94.59%)

Bayraktar and Temizel (2008) work on financial texts to find person names. They apply the local grammar based approach of Traboulsi (2006) to Turkish. They construct a list of Turkish reporting verbs. Since they work on a different domain focusing only to person names, their results are not comparable with none of the related work given in this section.

Küçük and Yazıcı (2012) adds statistical methods (Rote learning- (Freitag, 2000)) to their rule based study (Küçük and Yazıcı, 2009) raising the F-measure on general news text from 87.96

[13] Detailed information on the terms ENAMEX(person, organization and location names), TIMEX(dates and times) and NUMEX(currency values and percentages) included in NE Types column can be found at (Sundheim, 1995).

to 90.13. This study is the only published work of Turkish NER comparing performances for different domains. They evaluate their system on general news texts, financial news texts, historical texts and child stories. In Table 5 we took the results on general news texts domain which sounds similar to our domain. Their evaluation metric gives more credit to partial matches and not comparable with none of our metrics. They work on ENAMEX, TIMEX and NUMEX entity types but they do not provide the scores for each of these. In order to be able to make a fair comparison between the two studies, we measure the performance of their system on our test data and calculate the overall ENAMEX performance (F-Measure) as 69.78% in CoNLL metrics and 74.59% in MUC TYPE metrics. We think the reasons of the observed difference between the performances reported in their work and on our tests are the evaluation criteria, the working test domain (our dataset consists of older news texts) and the performance drop due to the lack of TIMEX and NUMEX types (where they have higher performances).

Tatar and Cicekli (2011) propose an automatic rule learning system exploiting morphological features. Although they don't namely mention that they use the CoNLL metric, the evaluation strategy of looking for the exact match is compatible with the CoNLL metric. Their overall score includes the performance on dates and time expressions which is higher than the performance for the NE types of our interest. Their reported accuracy is 91.08% on ENAMEX and TIMEX types. The relevant F-measure for only ENAMEX types is calculated as 90.63%; this result is comparable with our reported F-measure 91.94% in CoNLL metric (except the fact that the evaluations are made on different test sets).

Yeniterzi (2011) uses CRFs and exploits the impact of morphology for Turkish NER. In this work, she uses the inflectional units (IG) as tokens. This work is the one which is most similar to ours but we use morphological features in a different way and add the use of gazetteers. We use the same training and test data, so our results given in CoNLL metrics are fully comparable with this work. One should note that our performance before adding the gazetteers (89.55%) is still higher than her best result (88.94%) which shows that the increase may not be credited to only to the use of gazetteers.

Özkaya and Diri (2011) also uses CRFs on informal texts so it is not fair to compare the results with any of the work discussed in this section all working on formal texts. They do not provide their evaluation metrics and their overall results, but we calculate overall precision, recall and F-measure values as 92.89%, 77.07% and 84.24 respectively using the token counts provided in their paper.

6 Conclusion & Future Work

In this work we presented a Turkish NER model using conditional random fields trained with morphological and lexical features. We compiled large scale person and location names gazetteers which will be available for the researchers. We obtained state of the art results which we believe will act as the baseline for future Turkish NER research. We also tried to compare the results and the evaluation metrics of recent NER work in Turkish. We believe this is also an important contribution since the results given in previous works were not comparable because of first, different evaluation metrics giving different credits to partial matches were used and second, the studies focused to different sets of named entity types and provided their results as the average of these. As future work, we will test our model in different formal and informal domains, investigate the ways to better use the morphological properties collected in the inflectional (INF) feature such as using them as atomic units. We aim to improve our model extending and adding new generator gazetteers. We also aim to add NUMEX and TIMEX entity types to our system. The automatic determination of the useful features and their combinations is another subject of our future work.

Acknowledgments

The authors want to thank the following people without whom it would be impossible to produce this work: Reyyan Yeniterzi and Ilyas Çiçekli for providing their datasets, Gökhan Tür for the helpful discussions, Dilek Küçük and Adnan Yazıcı for processing the test data with their NER tool.

References

Bayraktar, O. and Temizel, T. T. (2008). Person Name Extraction From Turkish Financial News Text Using Local Grammar Based Approach. In *23rd International Symposium on Computer and Information Sciences (ISCIS'08)*, Istanbul. ISBN 978-1-4244-2880-9 electronic version (4 pp.).

Chinchor, N. A. and Marsh, E. (1998). Muc-7 information extraction task definition. In *Proceeding of the Seventh Message Understanding Conference (MUC-7), Appendices*.

Cucerzan, S. and Yarowsky, D. (1999). Language independent named entity recognition combining morphological and contextual evidence. In *In Proceedings of the joint SIGDAT conference on empirical methods in natural language processing and very large corpora.*

Ekbal, A. and Bandyopadhyay, S. (2009). A conditional random field approach for named entity recognition in Bengali and Hindi. *Linguistic Issues in Language Technology*, 2(1).

Eryiğit, G. (2012). The impact of automatic morphological analysis & disambiguation on dependency parsing of turkish. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC)*, Istanbul, Turkey.

Finkel, J. R., Grenager, T., and Manning, C. (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.

Freitag, D. (2000). Machine learning for information extraction in informal domains. *Machine Learning*, 39(2/3):169–202.

Hasan, K. S., Rahman, A., and Ng, V. (2009). Learning-based named entity recognition for morphologically-rich, resource-scarce languages. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 354–362.

Küçük, D. and Yazıcı, A. (2009). Named entity recognition experiments on turkish texts. In *Proceedings of the 8th International Conference on Flexible Query Answering Systems*, FQAS '09, pages 524–535, Berlin, Heidelberg. Springer-Verlag.

Küçük, D. and Yazıcı, A. (2012). A hybrid named entity recognizer for turkish. *Expert Syst. Appl.*, 39(3):2733–2742.

Lafferty, J. D., McCallum, A., and Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, pages 282–289.

LIU, X., ZHANG, S., WEI, F., and ZHOU, M. (2011). Recognizing named entities in tweets. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 359–367, Portland, Oregon, USA. Association for Computational Linguistics.

McCallum, A. (2003). Efficiently inducing features of conditional random fields. In *UAI*, pages 403–410.

Mohit, B., Schneider, N., Bhowmick, R., Oflazer, K., and Smith, N. A. (2012). Recall-oriented learning of named entities in arabic wikipedia. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 162–173, Avignon, France. Association for Computational Linguistics.

Na, S.-H. and Ng, H. T. (2009). A 2-poisson model for probabilistic coreference of named entities for improved text retrieval. In *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009*, pages 275–282.

Nadeau, D. and Sekine, S. (2007). A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30(1):3–26. Publisher: John Benjamins Publishing Company.

Oflazer, K. (1994). Two-level description of Turkish morphology. *Literary and Linguistic Computing*, 9(2):137–148.

Özkaya, S. and Diri, B. (2011). Named entity recognition by conditional random fields from turkish informal texts. In *Proceedings of the IEEE 19th Signal Processing and Communications Applications Conference (SIU 2011)*, pages 662–665.

Rüd, S., Ciaramita, M., Müller, J., and Schütze, H. (2011). Piggyback: Using search engines for robust cross-domain named entity recognition. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 965–975, Portland, Oregon, USA. Association for Computational Linguistics.

Sak, H., Güngör, T., and Saraçlar, M. (2008). Turkish language resources: Morphological parser, morphological disambiguator and web corpus. In *GoTAL 2008*, volume 5221 of *LNCS*, pages 417–427. Springer.

Settles, B. (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.

Sha, F. and Pereira, F. C. N. (2003). Shallow parsing with conditional random fields. In *HLT–NAACL*.

Sundheim, B. (1995). Overview of results of the muc-6 evaluation. In *MUC*, pages 13–31.

Sutton, C. and McCallum, A. (2011). An introduction to conditional random fields. *Foundations and Trends in Machine Learning*. To appear.

Tatar, S. and Cicekli, I. (2011). Automatic rule learning exploiting morphological features for named entity recognition in turkish. *Journal of Information Science*, 37(2):137–151.

Tjong Kim Sang, E. F. (2002). Introduction to the conll-2002 shared task: Language-independent named entity recognition. In *Proceedings of CoNLL-2002*, pages 155–158.

Tjong Kim Sang, E. F. and De Meulder, F. (2003). Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of CoNLL-2003*, pages 142–147.

Traboulsi, H. N. (2006). *Named Entity Recognition: A Local Grammar-based Approach*. PhD thesis, Department of Computing School of Electronics and Physical Sciences University of Surrey.

Tür, G., Hakkani-Tür, D., and Oflazer, K. (2003). A statistical information extraction system for turkish. *Natural Language Engineering*, 9:181–210.

Yeniterzi, R. (2011). Exploiting morphology in turkish named entity recognition system. In *Proceedings of the ACL 2011 Student Session*, pages 105–110, Portland, OR, USA.

Zhang, M., Li, H., Liu, M., and Kumaran, A. (2012). News 2012 shared task on machine transliteration. In *Proceedings of the 4th Named Entities Workshop 2012 at ACL 2012*.