Undefined 0 (2016) 1
IOS Press

Extending a CRF-based Named Entity Recognition Model for Turkish Well Formed Text and User Generated Content [1]

Editor(s): Name Surname, University, Country
Solicited review(s): Name Surname, University, Country
Open review(s): Name Surname, University, Country

Gökhan Şeker [a] Gülşen Eryiğit [b,*]
[a] *ITU Informatics Institute, Istanbul Technical University, Istanbul, 34469, Turkey*
E-mail: sekerg@itu.edu.tr
[b] *Department of Computer Engineering, Istanbul Technical University Istanbul, 34469, Turkey*
E-mail: gulsen.cebiroglu@itu.edu.tr

Abstract. Named entity recognition (NER), which provides useful information for many high level NLP applications and semantic web technologies, is a well-studied topic for most of the languages and especially for English. However the studies for Turkish, which is a morphologically richer and lesser-studied language, have fallen behind these for a long while. In recent years, Turkish NER intrigued researchers due to its scarce data resources and the unavailability of high-performing systems. Especially, the need to discover named entities occurring in Web datasets initiated many studies in this field. This article presents the enhancements made to a Turkish named entity recognition model [5] (based on conditional random fields (CRFs) and originally tailored for well formed texts) in order to extend its covered named entity types, and also to process extra challenging user generated content coming with Web 2.0. The article introduces the re-annotation of the available datasets to extend the covered named entity types, and a brand new dataset from Web 2.0. The introduced approach reveals an exact match F1 score of 92% on a dataset collected from Turkish news articles and ∼65% on different datasets collected from Web 2.0.

Keywords: Named entity recognition, Turkish, User generated content, CRF, Web data

1. Introduction

The second generation of the world wide web (Web 2.0), which is also referred as the Social Web, focuses on people and their social interactions by the use of attractive and easy-to-use applications. The volume of user generated content (UGC)[1] grows enormously every day as well as the need to semantically interpret this high volume of data. Semantic Web technologies focuses on interpreting automatically this ever-growing dynamic UGC (as well as static web pages) and convert it into a machine readable structured data [14]. Semantic enrichment (e.g., sentiment detection, polarity, named entity recognition) of UGC plays a key role for Semantic Web Technologies. Named entity recognition and linking (to external resources such as those in DBpedia) are of primary importance for

[1] This article is a revised and extended version of a paper that was presented at COLING-2012 [5]. This research is supported in part by a TUBITAK 1001 grant (no: 112E276) and is part of the ICT COST Action IC1207.

* Corresponding author.

[1] [28] defines UGC as "any form of content such as blogs, wikis, discussion forums, posts, chats, tweets, podcasts, digital images, video, audio files, advertisements and other forms of media that was created by users of an online system or service, often made available via social media websites".

extracting information and for populating knowledge bases [36].

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). Beside its value for semantic web technologies, NER is also an important stage for several natural language processing (NLP) tasks including machine translation, sentiment analysis and syntactic parsing. MUC (Message Understanding Conference [41,4]) and CoNLL (Conference on Computational Natural Language Learning [44,45]) conferences define three basic categories 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). However, NER research is not limited to only these types; different application areas concentrate on 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 [41]. [31] 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 [44,45], transliteration [50], coreference [30] of named entities and especially to NER on UGC [23,24,25,29, 35,37].

The use of Conditional Random Fields (CRFs) [22], which are reported to offer several advantages over hidden Markov models (HMMs), stochastic grammars and maximum entropy Markov models (MEMMs), became very dominant in the literature for the named entity recognition task. CRF-based NER models have been experimented for various domains and languages: [27] for English and German, [8] for Hindi and Bengali, [3] for Chineese, [39] for biomedical data, [24,35] for Tweets are some studies among many others.

Morphologically rich languages (MRLs) poses interesting challenges for NLP tasks (e.g., data scarcity, the representation of rich morphological features in different tasks) as it is the case for NER. Although there exist some studies reporting their approach for some MRLs, the usage of morphological information for the NER task is still an open research issue. [12] put word prefix and word suffix information as new features to their systems for Bulgarian. [15] uses the first and last 3 characters of the words as extra features in order to use them as prefix and suffix information for Bengali. [18] reports that their effort to add morphological features did not yield in any success improvement for Czech as well as [49] which reports similar findings for Turkish.

Turkish being a morphologically very rich language attracts the attention of the NLP community. Especially, the need to discover named entities occurring in Web datasets initiated many studies for Turkish NER in recent years. Nevertheless, the results for Turkish NER remain still very behind the reported accuracies for English. This article introduces an extended CRF-based Turkish NER model (firstly introduced in [5] on Turkish well formed texts for only ENAMEX types), the enhancements made in order to extend its coverage to also include TIMEX and NUMEX entity types and to process UGC[2] which poses extra challenges for NLP applications. The introduced system makes extensive use of morphological information and reports significant improvement by the use of this, differing from some of the previous NER studies on MRLs. The article introduces the re-annotation of the most commonly used datasets to extend the covered named entity types, and a brand new dataset from Web 2.0. The introduced approach reveals an exact match F1 score of 92% on a dataset collected from Turkish news articles and $\sim$65% on different datasets collected from Web 2.0.

The article is organized as follows. Section 2 gives brief information about Turkish, Section 3 gives a brief overview of the previous studies for Turkish NER, Section 4 gives information about existing and newly introduced language resources, Section 5 gives the details of the proposed framework, its extensions to TIMEX and NUMEX entities and to Web 2.0 domain, Section 6 gives our experiments and evaluates the results by comparing with related work and Section 7 gives the conclusion.

2. Turkish

This section briefly states the characteristics of the Turkish language which is treated to have influence for 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

[2]This manuscript focuses only to the textual content coming with UGC. [28].

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). In case of the proper nouns, the inflectional suffixes are separated from the lemma by an apostrophe in well formatted texts. 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. Section 6 investigates the impact of using lexical information for the named entity recognition task.

Although in well formed text, only the proper nouns, abbreviations and the initial words of the sentences start with an initial capital letter, this is most of the time not the case in social media domain. Turkish person (first) names are usually selected from common nouns such as İpek *(silk)*, Kaya *(rock)*, Pembe *(pink)*, Çiçek *(flower)*. This property of the language makes the recognition of such named entities very hard in UGC domain where the appropriate capitalization rules are frequently ignored.

Turkish is a free word order language. As a consequence of this property, the position of the word in a sentence doesn't provide 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".

[2] makes a preliminary investigation on the problems caused by UGC for Turkish NER. The following example from [2] shows the complexity caused by the omission of the above mentioned rules for proper nouns. In this Twitter example, which should actually be written as in the second line in formal writing, "Aydın" is a person name. The word when written with lowercase letters has also the meaning of a common noun; "enlightened". This makes very difficult to differentiate/identify this named entity (person name) from the word "enlightened".

"**aydınlara** gidiyoruz."
"*Aydın'lara* gidiyoruz."
(*We are going to **Aydın**'s house*)

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

Another problem for real data is the spelling errors produced either by mistake or on purpose for exaggeration, interjection or ASCIIfication (removal of accent, cedilla, etc) of special Turkish letters (öüçşığİ). In the first line of the below example, the letter "ı" is written with its ascii counterpart and repeated multiple times for specifying exclamation. The second line of the same example shows the case where all the letters are capitalized and it is again very difficult to detect the named entity and alleviate the ambiguity caused by the common sense of the proper noun.

"**aydiiiiiiiiin** nerdesin?"
"**AYDIIIIIIIIIN** NERDESİN?"
(**Aydın**, where are you?)

And finally, the following example again from [2] examplifies the foreign words inflected with Turkish suffixes by omitting the required apostrophe sign as shown in the second line. In the following Tweet: "Bieber" is used in accusative case without the required apostrophe.

"**Justin Bieber**i sevmem."
"**Justin Bieber**'i sevmem."
(*I don't like **Justin Bieber***)

3. Previous Turkish NER Studies

[5] compiles the performances of all the previous Turkish NER studies and tries to make comparisons with them whenever possible although none of the previous systems nor most of the used data sets were publicly available. The authors interacted with the owners of the previous systems which sometimes allowed to obtain a performance score by sending their test data to be tested with the prior system or obtaining the test data set used in the prior work.

Table 1 (from [5]) gives an overview of the previously reported Turkish NER performances. The performances listed in Table 1 is organized in decreasing order of credit given to partial matches during evaluation. The outputs of the NER systems are generally evaluated in comparison with human annotations. [31] exemplifies the different types of errors which may occur during the automatic recognition of the named entities; e.g. a totally missed NE, an identified entity with wrong entity type or wrong boundaries or both. [31] gives the details of the three main scoring techniques

Table 1
Overview of the previously reported Turkish NER results (compiled in [5])

Related work	Best Result	Eval. Method	Domain	NE Types
[32]	84.24	***n/a***	E-mail texts	ENAMEX
[20]	90.13	OTHER	General news	ENAMEX,TIMEX,NUMEX
[48]	91.56	MUC	General news	ENAMEX
[1]	81.97	MUC	Financial Texts	PERSON NAMES
[5]	94.59	MUC	General news	ENAMEX
[43]	91.08	CoNLL	Terrorism news	ENAMEX,TIMEX
[49]	88.94	CoNLL	General news	ENAMEX
[7]	91.85	CoNLL	General news	ENAMEX
[5]	91.94	CoNLL	General news	ENAMEX

(ACE[4], MUC and CONLL) which have been used in previous entity recognition studies.

Similar to most NER studies in the literature, MUC and CoNLL evaluations were widely used in Turkish NER studies. In the MUC evaluation approach, a system is scored on two axes: its ability to find the correct type (TYPE) and its ability to find exact text (TEXT). MUC TEXT makes evaluation only on NE boundaries without looking if the correct NE type is assigned or not. MUC TYPE evaluates the performance of assigning the correct named entity (NE) type to each word without taking into account if the NE boundaries are detected correctly. The final MUC score is the micro-averaged F-measure, which is the harmonic mean of precision and recall calculated over all entity slots on both axes. On the other hand, the CoNLL scoring technique uses an exact-match evaluation; it evaluates an assignment to be correct if both the type and the boundary of a NE is determined correctly. The calculated score is again the micro-averaged F-measure, but this time calculated on the exactly matched named entities. Most of the results in Table 1 are given as MUC and CoNLL scores. Note that the test sets, evaluation methods (3^{rd} column), working domain (4^{th} column) and entity types (5^{th} column) in focus of each work are different from each other.

The first published work on Turkish NER is [6] 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. The first NER work specific to Turkish is [48]. 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 use a training and test set collected from news articles which will be be introduced in the following sections. The authors use the same training data with [5], but a different test data which is not available. Their performance is reported as 91.56%. In order to be able to have an idea (although not strictly comparable), [5] also provides their MUC score (94.59%) as well as the CoNLL score of 91.94%.

[1] works on financial texts to find only person names. They apply the local grammar based approach of [47] to Turkish. [1] initially identified common reporting verbs in Turkish and then used these reporting verbs to generate patterns for locating person names. The study reports an CoNLL score of 81.97% which is not directly comparable with none of the related work given in this section due to the difference in the used datasets.

[49] uses CRFs and exploits the impact of morphology for Turkish NER. This work is the one which is most similar to ours except the usage of morphological features and gazetteers. [49] and [5] use the same training and test data. Table 1 gives the reported performance by [49]. In order to be able to make a strict comparison, the results of its replication and evaluation under our settings are provided under Section 6.2.

[32] also uses CRFs for NER on email messages, but since they are using features specific to email domain

[4] In the ACE evaluation method, each entity type has a parameterized weight and contributes up to a maximal proportion of the final score (e.g., if each person is worth 1 point and each organization is worth 0.5 point then it takes two organizations to counterbalance one person in the final score).

only (such as from, subject fields) their work may not be extended to general texts. They do not provide their evaluation metrics and their overall results, but overall precision, recall and F-measure values are calculated as 92.89%, 77.07% and 84.24% respectively using the token counts provided in their paper.

The automatic rule learning system of [43] starts with a set of seeds selected from the training set, and then extracts rules over these examples. The named-entities are generalized by using contextual, lexical, morphological and orthographic features. Although the authors do not namely mention that they use the CoNLL evaluation method, the evaluation strategy of looking for the exact match seems compatible with it. Their reported accuracy is 91.08% on ENAMEX and TIMEX types. The relevant F-measure for only ENAMEX types is calculated as 90.63%.

[20] uses rote-learning [11] in order to extend their rule-based recognizer [19] into a hybrid recognizer so that it can learn from the available annotated data and extend its knowledge resources. They evaluate their system on general news texts, financial news texts, historical texts and child stories. In Table 1 we took the results on general news texts domain which sounds similar to our domain. Their evaluation strategy gives more credit to partial matches and is not similar to neither CoNLL nor MUC scoring techniques. They work on ENAMEX, TIMEX and NUMEX entity types but they do not provide the scores for each of these. After measuring this system performance on their own dataset, [5] reports a CoNLL score of 69.78% on ENAMEX types for [20].

[7] addresses NER task for morphologically rich languages by employing a semi-supervised learning approach based on neural networks. They adopt a fast unsupervised method for learning continuous vector representations of words, and uses these representations along with language independent features. They test their work on the data set of [5] and reports a CoNLL score of 91.85%.

The Turkish NER studies on UGC domain are very recent and limited in number compared to well formed text domain. [2] is the first study which investigates the NER success on Turkish UGC; they test on 3 different domains, namely on datasets collected from Twitter, a Speech-to-Text Interface and a Hardware Forum. [16], [17] and [9] follow this trend and report their approaches on Twitter datasets. The outputs of our extended CRF-Model are compared with the mentioned studies in Section 6.2.

4. Language Resources

This section firstly gives the features of the existing and freely available Turkish datasets tagged with named entities. Then, it introduces the newly annotated ones within this work.

4.1. Available Language Resources

The most widely used dataset for Turkish NER research is introduced by [48]. This data, consists of nearly 500K words collected from newspaper articles and is annotated only for ENAMEX types. Another available dataset from well-written text genre comes from [43]. This dataset is rather small (∼55K) compared to the previous one and as a result is less preferable for supervised machine learning systems which mostly needs high volume of human-annotated data. The dataset consists of news articles on terrorism from both online and print news sources in Turkish. The annotated types on this corpus are ENAMEX and TIMEX categories.

The datasets from the UGC domain are brand new and the available ones are as follows:
[2] introduces three datasets annotated by ENAMEX, TIMEX and NUMEX types; 1- a 55K dataset which is from a very popular online forum dedicated for hardware products' reviews. An important feature of this dataset is that it contains mostly trademarks (generally company names), their products together with a related model. Although, this type of named entities are categorized under more specific named entity classes in extended NE classifications [38], the most relevant category in MUC-6 for these is the "Organization". This forum data is full of spelling errors and capitalization is not properly used or not used at all in most of the cases. 2- a very small corpus (∼1.5K) collected from Speech-to-Text Interface of a mobile assistant application. The most important characteristic of this dataset is that there is no capitalization or punctuation at all in the produced text message. 3- a 55K Twitter corpus which is used for testing purposes in many of the follow up studies [16], [17] and [9]. Unfortunately the annotations on this new domain were arguable and this resulted with the emergence of re-annotated versions[5] of the same dataset simultaneously by different groups ([16], [17] and [9] as well as this study). Additionally, [16] and [17] introduces a Twitter dataset of 20K to-

[5] Although not detailed in the cited references, the update information was obtained via personal communication with the authors.

kens whereas [9] introduces another one with 108K tokens.

4.2. Newly Introduced Language Resources

As known, human annotation of language resources is a costly process. The creation of benchmark datasets is very valuable to speed-up progress in a specific research area. As may be noticed from the previous subsections, early Turkish NER studies mostly evaluated their success on their own datasets which makes hard to make a comparison between the proposed approaches. In this study, we selected two mostly used datasets from the Turkish NER literature; one from well-written text domain [48] (which is also the biggest dataset) and one from UGC domain [2] and re-annotated them with the following two main purposes:

1- to extend the covered named entity types (to also cover TIMEX and NUMEX types) which were priorly limited to ENAMEX types only (in [48]).

2- to improve the consistency and hence the quality of the annotations by strictly following a specific guideline (namely the MUC-6 the Sixth Message Understanding Conference guidelines [13]).

Previous annotations were also carefully investigated during this second round of annotation. In addition to these two datasets, we also annotated a brand new Turkish treebank from the social media domain: ITU Web Treebank (IWT) [33]. IWT is specifically selected for the NER annotation due to its representativeness on UGC. Its composition (free from duplicates and re-tweets) includes UGC from different Web 2.0 domains (namely news story comments, personal blog comments, customer product reviews, social network posts and discussion forum posts) which we believe eliminates the dependency of the recent works towards the Twitter content only. Two human annotators served during the annotation process. The strength of agreement is considered to be 'very good' using Kappa statistics[6]. In all of the three datasets, we used MUC-6 style SGML tag elements: ENAMEX, TIMEX, and NUMEX; and the subcategorization is captured by a SGML tag attribute called TYPE, which is defined to have a different set of possible values for each tag element. Table 2 shows some sample annotations.

Table 3 gives the distribution of the named entities for each annotated datasets. One should note that the reported number of named entities may differ significantly from some of the previous studies (e.g. [49,2]) which report the number of tokens (conforming a named entity) instead of the actual number of named entities (consisting of one or more tokens) provided in here.

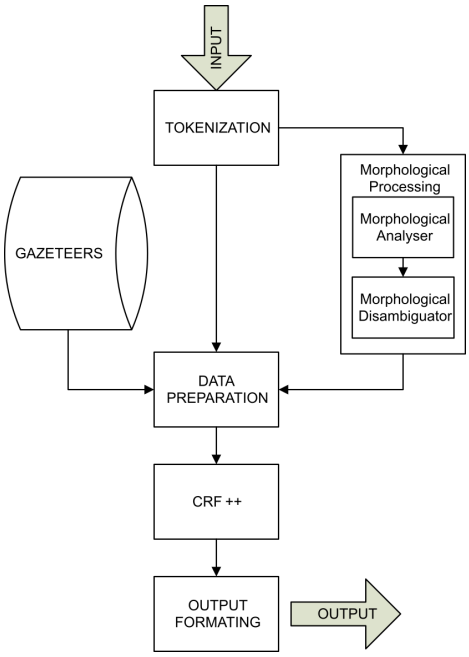

Fig. 1. Proposed Framework

5. a CRF-based Turkish Named Entity Recognizer

This section introduces the highest performing system for Turkish NER, its used features for ENAMEX types and newly added TIMEX and NUMEX types, and its adaptation for UGC.

5.1. Proposed Framework

Figure 1 shows the architecture of the used framework. The following subsections provides the details of each module.

5.1.1. 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 apos-

[6]Confidence intervals were calculated using the GraphPad QuickCalcs Web site: http://graphpad.com/quickcalcs/kappa1.cfm (accessed December 2015)

Table 2
Some sample annotations from the formal news text dataset

<ENAMEX TYPE="ORGANIZATION">Ankara 26. Asliye Hukuk Mahkemesi</ENAMEX> ,
<TIMEX TYPE="DATE">2 Temmuz 1997</TIMEX> 'de okuduğu şiirde dönemin
<ENAMEX TYPE="ORGANIZATION">Deniz Kuvvetleri</ENAMEX> Komutanı
<ENAMEX TYPE="PERSON">Erkaya</ENAMEX> 'nın kişilik haklarına
hakaret ettiği gerekçesiyle <ENAMEX TYPE="PERSON">Hatipoğlu</ENAMEX> 'nu
<NUMEX TYPE="MONEY">3 milyar lira</NUMEX> manevi tazminat cezasına çarptırdı .
<ENAMEX TYPE="LOCATION">Türkiye</ENAMEX> 'nin kirlenmesinin
<NUMEX TYPE="PERCENT">yüzde 30</NUMEX> 'u sanayiden geliyor.

Table 3
Entity distributions in newly introduced datasets

		News articles [48]	Tweets [2]	IWT [33]
Group	Type	492K	55K	43K
ENAMEX	Person	15,352	681	380
ENAMEX	Location	10,404	240	260
ENAMEX	Organization	9,571	428	401
TIMEX	Date	1,486	57	59
TIMEX	Time	169	20	9
NUMEX	Money	638	24	45
NUMEX	Percentage	710	5	8
TOTAL		38,330	1,455	1,162

Table 4
IOB2 tagging vs RAW tagging

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

trophe 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 4.

5.1.2. Morphological Processing

We used a two-level morphological analyzer [10] for producing the possible analyses for each word. We then give the output to a morphological disambiguator [10] 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 the most probable analysis within the given context:

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[7] 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.

[7]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

5.1.3. Gazetteers

Our preliminary work [5] has introduced two kind of gazetteers called base and generator gazetteers which have been compiled from different sources without taking the test corpora into consideration. Base gazetteers are composed of large lists of person and location names (∼261K tokens). The collected person names have been split into first name and surname gazetteers in order to both anonymize our gazetteers and to be able to detect different combinations of these. The location gazetteer has been collected so that it includes all location names in Turkish postal code system[8] , all country names from international telephone code system[9], city and states of those countries[10] and geographical names from different sources. The derivative gazetteers, which are rather small compared to the base gazetteers, consists of some frequently observed generator words (e.g. "Mr", "Professor", "Ministry of", "Street") having impact on the probability of next or previous words being part of a NE. Table 5 provides the number of tokens in each of these gazetteers. In this work, we basically add two small gazetteers (62 tokens in total) to the ones introduced in [5] in order to be able to identify TIMEX and NUMEX types. These are one base gazetteer (given as "Months" in Table 5) and one generator gazetteer (given as "Currency Units" in Table 5). Currency units generator gazetteer includes currency unit names of different countries generating currency expressions with the previous numerals.

Table 5
# of distinct tokens in gazetteers

	Gazetteer	# of tokens
Base	First names	44.048
	Surnames	138.844
	Location names	33.551
	Months	12
Generator	Location	44
	Organization	60
	Person	22
	Currency Units	50

[8]https://interaktifkargo.ptt.gov.tr/posta_kodu/
[9]http://www.ttrehber.turktelekom.com.tr/trk-web/ulkekodlari.html
[10]mostly collected from wikipedia.com

5.1.4. Data Preparation

At 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 such as "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). [5] experiments with different training data formats. These are IOB, IOB2, raw labels and fictitious boundary model of [48] and reports that the highest performance is obtained by using the RAW labels whereas using the IOB formats reduces the performance by 0.4% and the fictitous boundary format by 2%. Thus, in this article we follow the same approach and use the raw tags during the training stage. Table 4 gives tagging examples with both IOB2 and raw tags.

5.1.5. Conditional Random Fields

Conditional random fields (CRFs) [22] is a framework for building probabilistic models to segment and label sequence data. 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:

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

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:

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

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. [42] gives detailed information on mathematical foundations and many examples about the usage of CRFs. In this study

we used CRF++[11] which is an open source implementation of CRFs.

5.1.6. Feature Templates

The features (f_k) are based on some number of hand-crafted atomic observational tests (such as the token is capitalized or appears in a gazetteer) and a large collection of features is formed by making conjunctions of the atomic tests in certain user-defined patterns. "Conjunctions are important because the model is log-linear, and the only way to represent certain complex decision boundaries is to project the problem into a higher-dimensional space comprised of other functions of multiple variables" [26].

In some studies, it is shown that the useful feature conjunctions may be determined incrementally and provided to the system automatically [26]. But, in this study, we used the approach proposed in [40] 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.

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

5.2. Features used for ENAMEX Types

In our **base model** 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.

5.2.1. 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".

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.

As stated in the introduction, the usage of morphological information in NER modeling for MRLs is still an open research issue. In the literature, some studies use the first and last n characters of a word as extra features to CRF in order to include prefix and suffix information. However, Turkish is an MRL which is very rich and the possible suffix combinations could not be limited with a predefined length. Since the affixes in Turkish appear almost always as suffixes (except for some very rare foreign words), no extra feature is needed to represent prefixes in this language. A uniform representation (+INF feature) for suffixes free from variances due to vowel harmony is treated to be more appropriate for Turkish. As a result, we model the morphological information with the above features where +NCS and +PROP features are the atomic units extracted from the +INF feature.

5.2.2. 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).

5.2.3. Gazetteer Lookup Features:

Eight different features used for each of the eight gazetteers introduced in Section 5.1.3. **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.

5.3. Extra Features for TIMEX and NUMEX Types

Numeric Value (NV) : The numeric class argument. This feature is 0 for non-numeric tokens, 1 for integer tokens between [1-12], 2 for integer tokens between [13-31], 3 for integer tokens between [31-2020] and 4 for orher integer tokens and 5 for all other numeric values.

Percentage Sign (PS) : A binary feature indicating that the token is a percentage sign (%) or the word "yüzde" (*percent*) or not.

O'clock Term (OT) : A binary feature indicating that the token is the word "saat" (*o'clock*) or not.

Column Indicator (CI) : A binary feature indicating that the token includes the character ":" or not.

Month Gazetteer (MG) : A binary feature indicating that the token is included in the months gazetteer or not.

Currency Gazetteer (CG) : A binary feature indicating that the token is included in the currency units gazetteer or not.

5.4. Adaptation for UGC

A widely used approach while adapting the NER systems to UGC domain is to use text normalization prior to the NE identification. Similarly, in this work, the first approach that has been tried, but could not produce good results, was to use a Turkish text normalizer [46] specifically developed for Web 2.0 domain. As a result, instead of using such a comprehensive normalizer as a pre-processor, different error-tolerant gazetteer lookup scenarios are investigated. Similar to our investigations, [2] and [9] reported unsuccessful trials with their minimum-edit-distance based approaches. In our work, the highest performing method (**ASC**) is found to be the toleration of the replacement of a single Turkish special character ('ı', 'ü', 'ş', 'ö', 'ç', 'ğ') with its ascii counterpart ('i', 'u', 's', 'o', 'c', 'g') at a time. Our observations show that allowing a more flexible error tolerance yields at very high number of false matches (of input tokens) with gazetteer items.

Auto Capitalization Gazetteer (CAP) : As exemplified in Section 2, it is very hard to detect proper names with a common noun meaning when written in lowercase letters. Although this still remains as a challenging issue for Turkish NER studies, in this work, we manually selected the names from our gazetteers with a very little chance of being used as a common noun in Turkish texts. We then add a new binary CRF feature (CAP) indicating that the current token exists in this auto capitalization gazetteer or not.

Mention (MEN) : A binary feature indicating if the given token conforms to a specific pattern (the Twitter mention tags).

6. Experimental Results

In recent years, the CoNLL evaluation method became a de facto standard for the evaluation of NER

systems. In this article, we follow this trend and use this method on all of our evaluations. The produced output of the testing stage with RAW labels is converted automatically to IOB-2 style and then evaluated by the evaluation script from CoNLL 2000 shared task[12]

Following the previous work [49,5], in all of the provided experiments for the well formed text domain, we used 445K tokens of the news articles [48] (Table 3) as the training set (to be referred as **TRAIN3** - the training data version annotated with 3 ENAMEX types only [49,5] and **TRAIN7** - the re-annotated training data version with 7 entity types) and the remaining 47K tokens as the test set. The overall test datasets used in the following experiments are named as follows:

- **WFS3**: the original version of the news article test data [48,49,5] with 3 ENAMEX (PERSON, LOCATION, ORGANIZATION) types only,
- **WFS7**: the re-annotated version of the news article test data [48] with 7 entity types (ENAMEX, NUMEX and TIMEX) (Table 3),
- **TDS1_v1**: Tweet dataset introduced in [2],
- **TDS1_v2**: Tweet dataset introduced in [2] re-annotated version from [16],
- **TDS1_v3**: Tweet dataset introduced in [2] re-annotated version from [9],
- **TDS1_v4**: Tweet dataset introduced in [2] re-annotated version from this article,
- **TDS2**: Tweet dataset introduced in [16,17],
- **TDS3**: Tweet dataset introduced in [9],
- **IWT**: ITU Web Treebank [33].

In the following sections, the training and test set couples are provided for each experiment separated with a slash sign and between parentheses such as (TRAIN3/WFS3).

6.1. Evaluation of the Selected Features

Following the work of [5], our first experiment is to investigate the impact of each selected feature for the identification of ENAMEXs. Table 6 shows the impact of each selected feature to the best model by leaving out one feature at a time. The results show that even the SS feature (which was treated to have a slight impact with an incremental addition approach of each feature in [5]), has an important impact on the overall system by causing a 2.11% decrease with its absence.

[12]http://www.cnts.ua.ac.be/CoNLL2000/chunking/output.html.

Table 6
Contribution of each feature for ENAMEX types (TRAIN3/WFS3)

Excluded Feature	PER	ORG	LOC	Overall
best model	**92.94**	**88.77**	**92.93**	**91.94**
base model	80.77	77.86	87.66	82.28
-STEM	90.03	86.30	90.61	89.31
-POS	90.00	87.31	91.00	89.66
-NCS	90.31	87.11	90.97	89.74
-PROP	90.39	87.18	91.00	89.81
-INF	90.63	86.55	91.35	89.88
-CS	89.73	83.16	90.97	88.57
-SS	90.36	87.16	91.11	89.83
-BG	90.11	86.53	91.24	89.60
-GG	92.23	87.28	92.14	91.02

The impact of the inflectional features (INF) is also not surprising in such an agglutinative language since most of the time these features carry some information that would be carried with individual words in a morphologically poor language. All added morphological features have important impact on the performance. 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 (many atomic features concatenated to each other) is not a preferred approach. We observe from Table 6, the INF feature has an important impact despite this fact.

Table 7 gives the evaluation results of our second set of experiments conducted on the extended news dataset (WFS7). The first column are the results provided in Table 6 (best model on WFS3). The second column provides the results when exactly the same ENAMEX features (Section 5.2) are applied on WFS7. The last column of the table provides the results of our best model which includes the extra features included for TIMEX and NUMEX categories (Section 5.3). The last two rows give the average performances on the ENAMEX category and overall categories (ENAMEX, TIMEX, NUMEX).

The most attractive result in Table 7 is that the base model's average success (92.33%) on ENAMEX types is better than the system trained only on ENAMEX types (91.94%). The investigations show that the reason for this is the alleviation of miss-classification of some named entities with the annotation of these in TIMEX categories: e.g. "Eylül" (September) and "Ekim" (October) are at the same time very common female names in Turkish but also the name of some

Table 7
Extension to 7 NE types

Type	[5] (TRAIN3/WFS3)	Base Model (TRAIN7/WFS7)	Best Model (TRAIN7/WFS7)
Person	**92.94**	92.19	91.47
Location	92.93	94.28	**94.34**
Organization	88.67	89.56	**89.88**
Date	-	54.79	**89.25**
Time	-	51.85	**91.89**
Money	-	86.36	**100.00**
Percentage	-	65.67	**98.41**
on ENAMEX	91.94	92.33	92.15
Overall	-	89.27	**92.34**

months. The new annotations prevent the tendency of the classifier to annotate these as person names as it was the case when trained on WFS3. The results of the third column show that the new features improve the results on almost all NE types except person names. We also executed the same experiments with 10 fold cross validation and obtained an average F-measure (CoNLL evaluation) of 91.53 with a standard error of $\pm$0.50.

Table 8 evaluates the impact of newly added TIMEX and NUMEX features similarly to our initial experiments. -OT (O'clock term) and -PS (percentage) lines in Table 8 give exactly the same performances due their impact to the same instances in the test data. When these 2 features are excluded at the same time the performance drop on NUMEX categories is almost 11 percentage points.

The next experiment set is to evaluate the UGC adaptation introduced in Section 5.4. When we evaluate the system extended to 7 entity types (without any UGC adaptation) on TDS1_v1, we obtain 22.57%. The re-annotation of this dataset (TDS1_v4) alone results with an increase of 15.79 percentage points (from 22.57% to 38.36%). The baseline success (38.36%) on this dataset is provided in the second line of Table 9. After the introduced adaptation, our best model obtains 67.96% on this dataset.

We also evaluate the final system on IWT. It is noticeable that the performance on monetary and percentage expressions are lower than the one obtained on TDS1_v4. When we investigate the produced outputs for error analysis, we notice that the recall for these two types are very low due to the unusual usage of these expressions in social media domain (e.g. monetary expressions without providing any currency unit).

6.2. Discussions & Comparison with Related Work

As stated in the previous sections and in many studies in the literature, CRFs are proven to perform well on the NER task. However its modeling for MRLs, in other words for the rich morphology and the sparse data problem due to the high number of possible word surface forms appearing with such languages, is still an active research area. In addition, the need of normalization for the textual content appearing in Social Web is also complicated for MRLs [46] and it is not clear how the orchestration should be designed with normalization and higher level tasks aiming to extract structured data from such content. For the NER task, there exists studies (given in previous sections) reporting negative results by applying a sophisticated text normalization prior to the NER task. The reason for this may be explained as the mutual need of both tasks.

This article which introduces a NER model for Turkish, which is a morphologically very rich language, reports some improvements over the previous trials to its modeling for both well-formed texts and UGC. Although the results are very promising, we believe this area still needs more in-depth investigations in order to improve the performances especially on UGC.

This section give comparisons with some prior works related to the morphological modeling of Turkish for the NER task and its adaptation to the UGC domain. With this purpose, [49] which exploits the impact of morphology for Turkish NER is selected as the baseline comparative study for our morphological modeling. The work is replicated for a reliable comparison and discussed in more detail within the remaining of this section. For the UGC adaptation, the proposed model is compared with two CRF-based models ([2] and [9]) and two rule-based models ([16] and

Table 8
Contribution of each feature for NUMEX & TIMEX types (TRAIN7/WFS7)

Excluded Feature	PER	ORG	LOC	DATE	TIME	MONEY	PERC	Overall
best model	91.47	94.34	89.88	89.25	91.89	100.00	98.41	92.34
base model	92.19	94.28	89.56	54.79	51.85	86.36	65.67	89.27
-NV	91.24	94.12	89.62	59.18	91.89	95.00	98.41	90.54
-PS	91.17	94.34	89.62	77.25	91.89	95.00	98.41	91.38
-OT	91.17	94.34	89.62	77.25	91.89	95.00	98.41	91.38
-CI	91.17	94.34	89.62	77.25	64.29	95.00	98.41	91.18
-MG	91.10	94.12	89.62	62.69	91.89	100.00	98.41	90.74
-CG	91.17	94.34	89.62	77.25	90.29	80.00	98.41	91.42

Table 9
Contribution of each feature for UGC adaptation (TRAIN7/TDS1_v4)

Excluded Feature	PER	ORG	LOC	DATE	TIME	MONEY	PERC	Overall
best model	75.98	69.54	59.86	39.03	41.23	54.55	94.12	67.96
base model	47.88	56.48	22.86	11.32	33.33	54.55	94.12	38.36
-Asc	75.98	58.39	52.44	39.03	41.23	54.55	94.12	63.94
-Cap	58.63	63.27	23.37	15.37	34.67	54.55	94.12	47.15
-Men	66.74	69.54	59.86	39.03	41.23	54.55	94.12	63.63

Table 10
Performance on IWT (TRAIN7/IWT)

Excluded Feature	PER	ORG	LOC	DATE	TIME	MONEY	PERC	Overall
best model	67.22	77.17	53.87	70.31	80.00	27.45	50.00	64.96

[17]). [2], which is the pioneering study for Turkish, is selected as the baseline comparative study for UGC adaptation and the performances of its reimplementation with our experimental settings are also provided below.

[49] tries to include the morphological information into its CRF model by a new tokenization approach instead of the word-based tokenization. In this approach, each atomic morphological feature is provided as a seperate token to the system and tried to be labeled. [49] states no significant improvement with this tokenization over the word-based one. As explained in the previous sections, our approach to include morphology consists of adding two atomic features (parts-of-speech tag POS and the noun case information NCS) extracted from a word's morphological analysis and the full analysis (INF) which are all shown to have positive impact on the overall performance (Table 6). In both of the works, the stem information extracted from the morphological analysis is also added to the feature model in order to reduce data sparsity.

Another difference of [49] is the usage of letter case information. While our case feature (CS) takes 4 possible values (lower-case(0), UPPERCASE(1), Proper Name Case(2) and miXEd CaSe(3)), it takes only 2 values (lowercase and uppercase) in [49]. [49] reports the impact of this feature as 1.53 percentage points whereas in our experiments we obtain a 3.37 percantage points impact (Table 6). In this section, we replicate the model of [49] with our settings ((TRAIN3/WFS3), datasets represented with RAW labels, the same feature templates used for the used CRF features) and obtained 89.13% CoNLL score when we used a binary CS feature as opposed to its reported 88.71% in [49]. We also tested the model with our 4-valued CS feature and obtained 89.21%. When we test our model by omitting all the gazetteer features (BG and GG) and the SS feature to see the impact of our morphological modeling, we obtain 89.70% which results in a statistically significant improvement (according to the McNemar chi-squared test with $p<0.05$). From these experiments, we may conclude that our proposed model of using morphological features as CRF features seems better suited to Turkish than using them as separate tokens. However, there could still be room for improvement with further feature engineering on the selection of atomic morphological features.

Table 11
Comparison with related work on UGC

Related work	Best Result	Test Set	NE Types
[2]	19.28	**TDS1_v1**	ENAMEX,TIMEX,NUMEX
[16]	36.11	**TDS1_v2**	ENAMEX
[17]	46.93	**TDS1_v2**	ENAMEX,TIMEX,NUMEX
[9]	28.53	**TDS1_v3**	ENAMEX,TIMEX,NUMEX
this article	67.96	**TDS1_v4**	ENAMEX,TIMEX,NUMEX
Below are comparable results on ENAMEX types			
[16]	42.68	**TDS2**	ENAMEX
[17]	48.13	**TDS2**	ENAMEX
this article	49.02	**TDS2**	ENAMEX
Below are comparable results on all 7 NE types			
[17]	54.81	**TDS2**	ENAMEX,TIMEX,NUMEX
this article	56.02	**TDS2**	ENAMEX,TIMEX,NUMEX
Below are comparable results on all 7 NE types			
[9]	46.97	**TDS3**	ENAMEX,TIMEX,NUMEX
this article	51.61	**TDS3**	ENAMEX,TIMEX,NUMEX

Table 11 presents the comparison with the related works on UGC domain. The first set of the table provides the results on TDS1 (which has been reannotated by several groups as explained in Section 4.1). Since each group worked on a different version of this dataset these results are only provided to give an idea but essentially they are not comparable with each other. The remaining of the table compare the performances of our model on the test sets of the previous studies.

[2] follows the work of [5] and try to adapt a similar CRF based NER model to UGC domains. The authors test with different feature models which were the reduced versions (as the omission lexical features given in Section 5.2.2 and treated to be useless and meaningless for UGC domain) of the feature model used in [5], but couldn't obtain any performance increase. Their best model which only adds a normalization stage prior to the decoding stage performed very low with a 19.28% CoNLL score on TDS1_v1. One should notice that their training data was TRAIN3 although TDS1_v1 also consists TIMEX and NUMEX expressions. As stated in Section 6.1, our system (extended to 7 entity types) prior to the UGC adaptation obtains a performance score (22.57% on TDS1_v1) which is slightly better than their reported success. In order to make a fair comparison, we also replicate their work (with their best model) but this time trained on TRAIN7 instead of TRAIN3 and obtained an improvement of 5.63 percentage points (from 19.28% to 24.91%).

As given previously, our baseline score on the reannotated version (TDS1_v4) of the same data set was 38.36%. Table 11 provides our score (67.96%) obtained with our newly introduced model.

[9] also uses CRFs but with a different feature model which basically consists of the surface form, first and last 4 characters of the words instead of morphological features, lexical features, gazetteer lookups. Their reported accuracy on TDS1_v3 is 28.53% and 46.97% on TDS3. Our systems performance on TDS3 is calculated as 51.61%. [9] reports that the results are increased from 47% to 64% (on TDS3) by changing the training set from the one used in here (TRAIN7[13]) to another Tweeter dataset. We observed the same behavior and obtained an improvement from 52% to 68% by using their training dataset although it is relatively small in size when compared to TRAIN7. But we consider that the claims deducted from here would not be trustworthy due to the high number of retweets occurring in both training and test datasets.

[16] applies a rule-based multilingual NER system [34] to Turkish tweets. The system mostly employs language-independent rules that make reference to language-specific dictionary lists to recognize ENAMEX types and considers only those candidate

[13]The dataset TRAIN7 has been shared with the authors before the submission of this article.

tokens which have their initial letters capitalized. The system can be adapted to a new language by providing for that language separate word lists. [16] tailors it for Turkish by equipping it with the required lists for Turkish information extraction, including lists of common person, location and organization names as well as organization endings in Turkish. The work focuses only on ENAMEX types. In order to be able to make a comparison with this work, Table 11 provides an extra set of scores on TDS2 measuring the performances obtained only for ENAMEX types.

[17] adapt the rule based system of [21] to better fit Twitter language by relaxing its capitalization constraint and by diacritics-based expansion of its lexical resources. They employ a simplistic normalization scheme on tweets to observe the effects of these on the overall named entity recognition performance on Turkish tweets. Table 11 provides the comparisons on TDS2 and TDS3.

7. Conclusion & Future Work

This article presents an extended CRF-model for the named entity recognition of Turkish which is a morphologically very rich language. Extensive feature engineering is conducted in order to select appropriate feature representations in order to improve the performances both on well formed texts and user generated content. The re-annotation of the available datasets (from well formed text domain) to extend the covered named entity types (ENAMEX, TIMEX and NUMEX) as well as two newly annotated datasets from Web 2.0 are introduced. The compiled gazetteers, datasets and the used feature template are made available for future research from http://tools.nlp.itu.edu.tr/Datasets. The proposed system is available as a SaaS from http://tools.nlp.itu.edu.tr/Ner.

The introduced approach reveals an exact match F1 score of 92% on a dataset collected from Turkish news articles and ∼65% on different datasets collected from Web 2.0. Although the results obtained on well formed texts are in acceptable levels now, the field still needs new research in order to increase the results for non-canonical social media content. Especially the detection of proper nouns, with also a common noun meaning, written in lowercase letters needs special focus as the future work. The impact of normalization also needs to be investigated more. In this new UGC domain, named entity recognition and normalization becomes two NLP layers which are hard to orchestrate; one needing the outputs of the other one to produce better results. As a result, joint systems of these two layers may be a good research topic in the future.

Acknowledgements

We would like to acknowledge that this work is part of a research project entitled “Parsing Web 2.0 Sentences” subsidized by the TUBITAK (Turkish Scienti fic and Technological Research Council) 1001 program (grant number 112E276) and part of the ICT COST Action IC1207. 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 and Memduh Gokirmak for helping during the annotation process.

References

[1] Özkan Bayraktar and Tuğba Taşkaya Temizel. 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, 2008. ISBN 978-1-4244-2880-9 electronic version (4 pp.).

[2] Gökhan Çelikkaya, Dilara Torunoğlu, and Gülşen Eryiğit. Named entity recognition on real data: A preliminary investigation for Turkish. In *Proceedings of the 7th International Conference on Application of Information and Communication Technologies, AICT2013*, Baku, Azarbeijan, October 2013. IEEE.

[3] Wenliang Chen, Yujie Zhang, and Hitoshi Isahara. Chinese named entity recognition with conditional random fields. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, pages 118–121, 2006.

[4] Nancy A. Chinchor and Elaine Marsh. MUC-7 information extraction task definition. In *Proceeding of the Seventh Message Understanding Conference (MUC-7), Appendices*, 1998.

[5] Gökhan Akın Şeker and Gülşen Eryiğit. Initial explorations on using CRFs for Turkish named entity recognition. In *COLING 2012, 24th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, 8-15 December 2012, Mumbai, India*, 2012.

[6] Silviu Cucerzan and David Yarowsky. 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.*, 1999.

[7] Hakan Demir and Arzucan Ozgur. Improving named entity recognition for morphologically rich languages using word embeddings. In *The 13th International Conference on Machine Learning and Applications (ICMLA'14) , Detroit, Michigan, USA, December, 2014*, 2014.

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

[9] Beyza Eken and Ahmet Cüneyd Tantuğ. Recognizing named entities in Turkish Tweets. In *Proceedings of the Fourth International Conference on Software Engineering and Applications*, Dubai, UAE, January 2015.

[10] Gülşen Eryiğit. ITU Turkish NLP web service. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Gothenburg, Sweden, April 2014. Association for Computational Linguistics.

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

[12] Georgi Georgiev, Preslav Nakov, Kuzman Ganchev, Petya Osenova, and Kiril Simov. Feature-rich named entity recognition for bulgarian using conditional random fields. In *RANLP*, pages 113–117, 2009.

[13] Ralph Grishman. MUC - 6. In *http://cs.nyu.edu/faculty/grishman/muc6.html*, 1996. Last accessed : Aug 14th 2014.

[14] Tom Gruber. Collective knowledge systems: Where the social web meets the semantic web. *Web semantics: science, services and agents on the World Wide Web*, 6(1):4–13, 2008.

[15] Kazi Saidul Hasan, Altaf Rahman, and Vincent Ng. 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, 2009.

[16] Dilek Küçük, Guillaume Jacquet, and Ralf Steinberger. Named entity recognition on Turkish Tweets. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may 2014. European Language Resources Association (ELRA).

[17] Dilek Küçük and Ralf Steinberger. Experiments to improve named entity recognition on Turkish Tweets. In *Proceedings of the EACL'2014 workshop Language Analysis in Social Media (LASM)*, pages 71–78, Gothenburg, Sweden, april 2014.

[18] Michal Konkol and Miloslav Konopík. Crf-based czech named entity recognizer and consolidation of czech ner research. In *International Conference on Text, Speech and Dialogue*, pages 153–160. Springer, 2013.

[19] Dilek Küçük and Adnan Yazıcı. 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, 2009. Springer-Verlag.

[20] Dilek Küçük and Adnan Yazıcı. A hybrid named entity recognizer for Turkish. *Expert Systems with Applications*, 39(3):2733–2742, 2012.

[21] Dilek Küçük and A Yazici. Rule-based named entity recognition from turkish texts. In *Proceedings of the International Symposium on Innovations in Intelligent Systems and Applications*, 2009.

[22] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, pages 282–289, 2001.

[23] Chenliang Li, Jianshu Weng, Qi He, Yuxia Yao, Anwitaman Datta, Aixin Sun, and Bu-Sung Lee. Twiner: named entity recognition in targeted twitter stream. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 721–730. ACM, 2012.

[24] Xiaohua LIU, Shaodian ZHANG, Furu WEI, and Ming ZHOU. 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, June 2011. Association for Computational Linguistics.

[25] Xiaohua Liu, Ming Zhou, Furu Wei, Zhongyang Fu, and Xiangyang Zhou. Joint inference of named entity recognition and normalization for tweets. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 526–535. Association for Computational Linguistics, 2012.

[26] Andrew McCallum. Efficiently inducing features of conditional random fields. In *UAI*, pages 403–410, 2003.

[27] Andrew McCallum and Wei Li. 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*, pages 188–191. Association for Computational Linguistics, 2003.

[28] Marie-Francine Moens, Juanzi Li, and Tat-Seng Chua. *Mining User Generated Content*. Chapman & Hall/CRC, 2014.

[29] Behrang Mohit, Nathan Schneider, Rishav Bhowmick, Kemal Oflazer, and Noah A. Smith. 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, April 2012. Association for Computational Linguistics.

[30] Seung-Hoon Na and Hwee Tou Ng. 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, 2009.

[31] David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30(1):3–26, January 2007. Publisher: John Benjamins Publishing Company.

[32] Serap Özkaya and Banu Diri. 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, 2011.

[33] Tugba Pamay, Umut Sulubacak, Dilara Torunoglu-Selamet, and Gülsen Eryigit. The annotation process of the ITU Web Treebank. In *The 9th Linguistic Annotation Workshop held in conjuncion with NAACL 2015*, page 95, 2015.

[34] Bruno Pouliquen and Ralf Steinberger. Automatic construction of multilingual name dictionaries. *Learning machine translation. MIT Press, NIPS series*, 2009.

[35] Alan Ritter, Sam Clark, Oren Etzioni, et al. Named entity recognition in Tweets: an experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1524–1534. Association for Computational Linguistics, 2011.

[36] Giuseppe Rizzo, Marieke van Erp, and Raphaël Troncy. Benchmarking the extraction and disambiguation of named entities on the semantic web. In *LREC*, pages 4593–4600, 2014.

[37] Stefan Rüd, Massimiliano Ciaramita, Jens Müller, and Hinrich Schütze. 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, June 2011. Association for Computational Linguistics.

[38] Satoshi Sekine, Kiyoshi Sudo, and Chikashi Nobata. Extended named entity hierarchy. In *LREC*, 2002.

[39] Burr Settles. 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, 2004. Association for Computational Linguistics.

[40] Fei Sha and Fernando C. N. Pereira. Shallow parsing with conditional random fields. In *HLT–NAACL*, 2003.

[41] Beth Sundheim. Overview of results of the MUC-6 evaluation. In *MUC*, pages 13–31, 1995.

[42] Charles Sutton and Andrew McCallum. An introduction to conditional random fields. *Foundations and Trends in Machine Learning*, 2011. To appear.

[43] Serhan Tatar and Ilyas Cicekli. Automatic rule learning exploiting morphological features for named entity recognition in Turkish. *Journal of Information Science*, 37(2):137–151, April 2011.

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

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

[46] Dilara Torunoğlu and Gülşen Eryiğit. A cascaded approach for social media text normalization of Turkish. In *5th Workshop on Language Analysis for Social Media (LASM) at EACL*, Gothenburg, Sweden, April 2014. Association for Computational Linguistics.

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

[48] Gökhan Tür, Dilek Hakkani-Tür, and Kemal Oflazer. A statistical information extraction system for Turkish. *Natural Language Engineering*, 9:181–210, June 2003.

[49] Reyyan Yeniterzi. Exploiting morphology in Turkish named entity recognition system. In *Proceedings of the ACL 2011 Student Session*, pages 105–110, Portland, OR, USA, June 2011.

[50] Min Zhang, Haizhou Li, Ming Liu, and A Kumaran. News 2012 shared task on machine transliteration. In *Proceedings of the 4th Named Entities Workshop 2012 at ACL 2012*, 2012.