State of the art in Turkish Named Entity Recognition [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 state of the art in Turkish named entity recognition both on well formed texts and user generated content, and introduces the details of the best-performing system so far. The introduced approach uses conditional random fields and obtains the highest results in the literature for Turkish NER with 92% CoNLL score on a dataset collected from Turkish news articles and ∼65% on different datasets collected from Web 2.0. The article additionally introduces the re-annotation of the available datasets to extend the covered named entity types, and a brand new dataset from Web 2.0.

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

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 [32,3] and CoNLL [35,36] 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 [32]. [25] 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 [35,36], NER on informal texts [21,28,23], transliteration [41] and coreference [24] of named entities .

[1] This article is a revised and extended version of a paper that was presented at COLING-2012 [4]. This research is supported in part by a TUBITAK 1001 grant (no: 112E276) and is part of the ICT COST Action IC1207.

[*] Corresponding author.

Morphologically rich languages poses interesting challenges for NLP tasks as it is the case for NER [14]. Turkish being one of such languages attracts the attention of the NLP community. Nevertheless, the results for Turkish NER remain still very behind the reported accuracies for English. Previous studies on Turkish NER are investigated in the following sections. Although some of these studies try to use Conditional Random Fields (CRF)[1] [20] for Turkish NER task, the work given in this article is the first study which introduces a successful CRF model which outstrips the state of the art results for this problem.

This article introduces a state of the art Turkish named entity recognizer (firstly introduced in [4] as the best-performing system so far 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 user generated content (UGC) which poses extra challenges coming with Web 2.0. The introduced system, which uses conditional random fields, obtains the highest results in the literature for Turkish NER with 92% CoNLL score on a dataset collected from Turkish news articles and ∼65% on different datasets collected from UGC. An important contribution of the article is the overview of the previously published literature on Turkish NER and their comparison with the introduced system in commonly used evaluation metrics whenever possible. The article additionally 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 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 example a Turkish verb "gitmek" (*to go*) may appear in hundreds of different surface forms[2] 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

[1] CRF is a very popular method used in NLP. It is also widely used for named entity recognition task in various domains [21,8,30]. Stanford NER [11] which is a well-known NER tool also uses CRFs as its machine learning method.

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

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*)

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

The first published work on Turkish NER is [5] 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 [39]. 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.

[1] works on financial texts to find only person names. They apply the local grammar based approach of [38] to Turkish. [40] uses CRFs and exploits the impact of morphology for Turkish NER. [26] also uses CRFs for NER on email messages, 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. They do not provide their evaluation metrics and their overall results. [34] proposes an automatic rule learning system exploiting morphological features and works on terrorism news.

[18] adds statistical methods (Rote learning [12]) to their previous rule based study [17] raising the F-measure on general news text from 87.96 to 90.13. They evaluate their system on general news texts, financial news texts, historical texts and child stories.

[6] 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. [19] presents an automatic approach to compile language resources for named entity recognition (NER) in Turkish by utilizing Wikipedia article titles.

[2] is the first study which investigates the NER success on UGC; they test on 3 different domains, namely on datasets collected from Twitter, a Speech-to-Text Interface and a Hardware Forum. [2] follows the work of [4] and try to adapt a similar system to UGC domains. [15], [16] and [9] follow this trend and report their approaches on Twitter datasets.

Unfortunately, most of the listed studies are on different datasets and evaluated their performances with different metrics. In some cases, either the resources or the tools are not accessible. This situation makes hard to start research in the field. Following a previously started effort [4] of creating a benchmark for Turkish NER studies, this article tries to collect the available resources and make comparison with the previous works whenever possible using similar evaluation metrics.

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 [39]. 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 [34]. 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 [29], the most relevant category in MUC6 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 [15], [16] and [9]. Unfortunately the annotations on this new domain was arguable and this resulted with the emergence of re-annotated versions[3] of the same dataset simultaneously by different groups ([15], [16] and [9] as well as this study). Additionally, [15] and [16] introduces a Twitter dataset of 20K tokens 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 fair 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 [39] (which is also the biggest dataset) and one from UGC domain [2] and re-annotated with the following two main purposes:

1- to extend the covered named entity types which were priorly limited to ENAMEX types only (in [39]).

2- to improve the quality of the annotations by strictly following the MUC-6 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) [27]. IWT is specifically selected for the NER annotation due to its representativeness on UGC. Its composition 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[4]. 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 1 shows some sample annotations.

Table 2 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. [40,2]) 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.

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

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

Table 1
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 2
Entity distributions in newly introduced datasets

		News articles [39]	Tweets [2]	IWT [27]
Group	Type	492K	50K	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

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 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 3.

Table 3
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

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

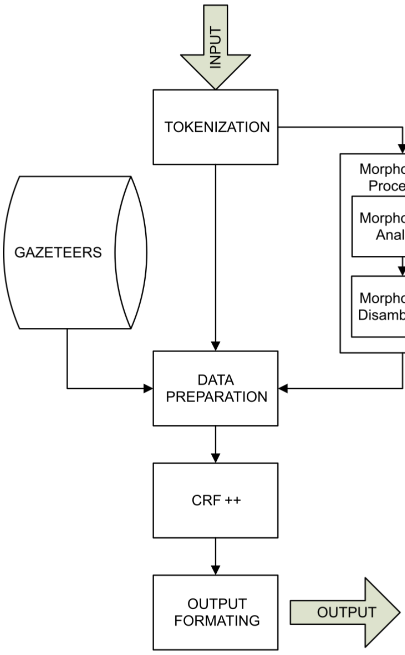

Fig. 1. Proposed Framework

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[5] 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.

5.1.3. Gazetteers

In this work, we basically add two small gazetteers to the ones introduced in [4] in order to be able to identify TIMEX and NUMEX types. There exists mainly two kind of gazetteers which we call base and generator gazetteers. Table 4 gives the details for each one. Base gazetteers are the ones which include words with high probability of occurrence in a named entity. These are large gazetteers with thousands of tokens except the recently added month gazetteer in order to catch time expressions. We collected person names from different sources. We split them into first name and surname gazetteers in order to both anonymize our gazetteers and to be able to detect different combinations of these. For example, Ahmet Yılmaz and Mehmet Demir are two person names, we include Ahmet and Mehmet in names gazetteer and Yılmaz and Demir in surnames gazetteer, so the person names Ahmet Demir and Mehmet Yılmaz would automatically be detected as person names. 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.

Table 4
# 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

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*). Similarly, location generator gazetteer includes words like "cadde" (Eng:street) which generate location names together with the previous tokens. Person generator gazetteer includes titles or relations usually occurring before or after person names such as “bey” (Mr.) or “profesör” (professor). Currency units generator gazetteer includes currency unit names of different countries generating currency expressions with the previous numerals.

5.1.4. Data Preparation

At this stage, we use the information coming from the raw data, the gazetteers and the morphological pro-

[5] 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

[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

cessing 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). [4] experiments with different training data formats. These are IOB, IOB2, raw labels and fictitious boundary model of [39] 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 3 gives tagging examples with both IOB2 and raw tags.

5.1.5. *Conditional Random Fields*

Conditional random fields (CRFs) [20] 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:

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

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. [33] 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.

5.1.6. *Feature Templates*

CRFs are log-linear models. In order to get 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 [22]. But, in this study, we used the approach proposed in [31] 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.

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.

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

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.

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 couldn't produce good results, was to use a Turkish text normalizer [37] 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

There exist 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 named entity (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. On the other hand, the CoNLL metric 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 [25]. In recent studies, the CoNLL metric became a de facto standard for the evaluation of NER systems. In this article, we follow this trend[10] and use this metric on almost all of our evaluations. We also provide MUC scores in related sections in order to be able to make comparisons with previous works.

Following the previous work [40,4], in all of the provided experiments, we used 440K tokens of the news articles [39] (Table 2) as the training set and the remaining 47K tokens as the test set (**WFS**) for well formed text domain. The test datasets used in the following experiments are named as follows:

- **WFS3**: the original version of the news article test data [39,40,4] with ENAMEX only,
- **WFS7**: the re-annotated version of the news article test data [39] with 7 entity types (ENAMEX, NUMEX and TIMEX) (Table 2),
- **Tweet_Dataset_1 Ver 1**: Tweet dataset introduced in [2],
- **Tweet_Dataset_1 Ver 2**: Tweet dataset introduced in [2] re-annotated version from [15],
- **Tweet_Dataset_1 Ver 3**: Tweet dataset introduced in [2] re-annotated version from [9],
- **Tweet_Dataset_1 Ver 4**: Tweet dataset introduced in [2] re-annotated version from this article,
- **Tweet_Dataset_2**: Tweet dataset introduced in [15,16],
- **Tweet_Dataset_3**: Tweet dataset introduced in [9],
- **IWT**: ITU Web Treebank [27].

6.1. Evaluation of the Selected Features

Following the work of [4], our first experiment is to investigate the impact of each selected feature for the identification of ENAMEXs. Table 5 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 [4]), has an important impact on the overall system by causing a 2.11% decrease with its absence.

Table 5
Contribution of each feature for ENAMEX types (on 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

[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”).

that would be carried with individual words in a morphologically poor language.

Table 6 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 5 (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).

Table 6
Extension to 7 NE types

Type	[4] on WFS3	Base Model on WFS7	Best Model on 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**

The most attractive result in Table 6 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 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 of 91.53 with a standard error of ±0.50.

Table 7 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 7 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 Tweet_Dataset_1 v1, we obtain 22.57%. The re-annotation of this dataset (Tweet_Dataset_1 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 8. 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 Tweet_Dataset_1 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. *Comparison with Related Work*

This section tries to compare the provided approach with the related works.

6.2.1. *Comparison with the works on well formed texts*

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 dataset. Table 10 gives the reported results of each related work. We give the results of our pairwise comparisons in the running text whenever possible.

The performances listed in Table 10 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 (5^{th} column) in focus of each work are different from each other. Table 10 tries to give an overview of these features for each work (discussed in detail in Section 3).

The first NER work specific to Turkish [39] evaluates its results in MUC metrics. The authors use the same training data with this article, but a different test

Table 7

Contribution of each feature for NUMEX & TIMEX types (on 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 8

Contribution of each feature for UGC adaptation (on Tweet_Dataset_1 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 9

Performance on 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

Table 10

Comparison with related work on well formed texts (The reported results in each paper)

Related work	Best Result	Ev.Metr.	Domain	NE Types
[26]	84.24	***n/a***	E-mail texts	ENAMEX
[18]	90.13	OTHER	General news	ENAMEX,TIMEX,NUMEX
[39]	91.56	MUC	General news	ENAMEX
[1]	81.97	MUC	Financial Texts	PERSON NAMES
[4]	**94.59**	**MUC**	**General news**	**ENAMEX**
[34]	91.08	CoNLL	Terrorism news	ENAMEX,TIMEX
[40]	88.94	CoNLL	General news	ENAMEX
[7]	91.85	CoNLL	General news	ENAMEX
[4]	91.94	CoNLL	General news	ENAMEX
this article	**92.15**	**CoNLL**	**General news**	**ENAMEX**
this article	**92.34**	**CoNLL**	**General news**	**ENAMEX,TIMEX,NUMEX**

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), [4] also provides their results on MUC metric (94.59%). The CoNLL score of the same work was 91.94%. [1] which works on financial texts to find person names reports an F-measure of 81.97% (CoNLL) which is not directly comparable with none of the related work given in this section due to the difference of the used datasets. [18] evaluate their system on general news texts, financial news texts, historical texts and child stories. In Table 10 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).

Although the authors of [34] 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 may be compared with our reported F-measure 91.94% in CoNLL metric (except the fact that the evaluations are made on different test sets).

[40] 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.

[26] also uses CRFs on Email messages. 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.

6.2.2. *Comparison with the works on UGC*

Table 11 presents the comparison with the related works on UGC domain. All the provided results in the table are in CoNLL metric. The first set of the table provides the results on Tweet_Dataset_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. The impact of Tweet_Dataset_1's re-annotation was provided in the previous section.

[9] reports that the results are increased from 47% to 64% by changing the training set from the one used in here (news articles) to another Tweeter dataset. We observed the same behavior and obtained an improvement from 52% to 68% by using their dataset for training although it is relatively small in size when compared to [39]. But we consider that the claims deducted from here would not be trustworthy due to the high number retweets occurring in both training and test datasets.

7. Conclusion & Future Work

This article presents the state of the art in Turkish named entity recognition and tries to make an empirical comparison between recent work on both well formed text and user generated content as far as possible. The article gives the details of the best performing system which uses a sequential classifier "Conditional Random Fields" for the supervised learning of named entity types. Extensive feature engineering is conducted in order to select appropriate feature representations to create a successful NER system for the morphologically very rich language in focus; i.e. Turkish. The proposed system is tailored to classify ENAMEX, TIMEX and NUMEX categories on Turkish well formed texts and user generated content. The work introduces three newly annotated datasets for the researchers on both of these domains. Although the results obtained on well formed text 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.

Table 11
Comparison with related work on UGC (The reported results in each paper all in CoNLL metric)

Related work	Best Result	Dataset	NE Types
[2]	24.91	***Tweet_Dataset_1 V1***	ENAMEX,TIMEX,NUMEX
[15]	36.11	***Tweet_Dataset_1 V2***	ENAMEX
[16]	46.93	***Tweet_Dataset_1 V2***	ENAMEX,TIMEX,NUMEX
[9]	28.53	***Tweet_Dataset_1 V3***	ENAMEX,TIMEX,NUMEX
this article	67.96	***Tweet_Dataset_1 V4***	ENAMEX,TIMEX,NUMEX
Below are comparable results on ENAMEX types			
[15]	42.68	***Tweet_Dataset_2***	ENAMEX
[16]	48.13	***Tweet_Dataset_2***	ENAMEX
this article	49.02	***Tweet_Dataset_2***	ENAMEX
Below are comparable results on all 7 NE types			
[16]	54.81	***Tweet_Dataset_2***	ENAMEX,TIMEX,NUMEX
this article	56.02	***Tweet_Dataset_2***	ENAMEX,TIMEX,NUMEX
Below are comparable results on all 7 NE types			
[9]	46.97	***Tweet_Dataset_3***	ENAMEX,TIMEX,NUMEX
this article	51.61	***Tweet_Dataset_3***	ENAMEX,TIMEX,NUMEX

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

[4] 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.

[5] 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.

[6] Hakan Demir and Arzucan Ozgur. Redefinition of Turkish morphology using flag diacritics. In *Proceedings of The Tenth Symposium on Natural Language Processing (SNLP-2013),Phuket, Thailand*, 2013.

[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] Gülsen Eryigit. Itu turkish nlp web service. In *EACL*, pages 1–4, 2014.

[11] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 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, 2005. Association for Computational Linguistics.

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

[13] Ralph Grishman. Muc - 6, 1996. Last accessed : Aug 14th 2014.

[14] 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.

[15] 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).

[16] 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.

[17] 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.

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

[19] Dilek Küçük. Automatic compilation of language resources for named entity recognition in turkish by utilizing wikipedia article titles. *Computer Standards & Interfaces*, 41:1–9, 2015.

[20] 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.

[21] 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.

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

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

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

[30] 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.

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

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

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

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

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

[40] 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.

[41] 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.