Chapter 6
Turkish Named-Entity Recognition

Reyyan Yeniterzi, Gökhan Tür, and Kemal Oflazer

Abstract Named-entity recognition is an important task for many other natural language processing tasks and applications such as information extraction, question answering, sentiment analysis, machine translation, etc. Over the last decades named-entity recognition for Turkish has attracted significant attention both in terms of systems development and resource development. After a brief description of the general named-entity recognition task, this chapter presents a comprehensive overview of the work on Turkish named-entity recognition along with the data resources various research efforts have built.

6.1 Introduction

Named-entity recognition (NER) can be defined as the process of identifying and categorizing the named-entities, such as person, location, product and organization names, or date/time and money/percentage expressions in unstructured text. This is an important initial stage for several natural language processing tasks including information extraction, question answering, and sentiment analysis.

Earlier approaches to this task in English relied on handcrafted rule-based systems but over time machine learning became the dominant paradigm (Nadeau and Sekine 2007). State-of-the-art NER systems have been developed for many

R. Yeniterzi
Özyeğin University, Istanbul, Turkey
e-mail: reyyan.yeniterzi@ozyegin.edu.tr

G. Tür
Google Research, Mountain View, CA, USA
e-mail: gokhan.tur@ieee.org

K. Oflazer (✉)
Carnegie Mellon University Qatar, Doha-Education City, Qatar
e-mail: ko@cs.cmu.edu

K. Oflazer, M. Saraçlar (eds.), *Turkish Natural Language Processing*,
Theory and Applications of Natural Language Processing,
https://doi.org/10.1007/978-3-319-90165-7_6

```
"Büyük Şehirlerin Deprem Güvenliği" konulu konferansın ilk günü
ağırlıkla bilimsel deprem senaryolarına ayrılmıştı. Bu çerçevede
<b_enamex Type="LOCATION"> Türkiye <e_enamex> 'den
<b_enamexTYPE="ORGANIZATION"> ODTÜ <e_enamex> öğretim üyelerinden
<b_enamex TYPE="PERSON"> Derin Ural <e_enamex>
<n_enamex TYPE="LOCATION"> İzmir <e_enamex> için hazırlanan böyle
bir senaryo çalışmasını aktardı.
```

Fig. 6.1 An example ENAMEX labeled Turkish text

languages and for widely studied languages like English, NER can be considered as a solved problem with an accuracy of around 95%.

Named-entity recognition task was initially introduced by DARPA, and evaluated as an understanding task in both the Sixth and Seventh Message Understanding Conferences (MUC) (Sundheim 1995; Chinchor and Marsh 1998). Later, CoNLL shared tasks (Tjong Kim Sang 2002; Tjong Kim Sang and De Meulder 2003) and Automatic Content Extraction (ACE) program (Doddington et al. 2004) stimulated further research and competition in NER system development.

These conferences defined three basic types of named-entities:

- ENAMEX (person, location, and organization names)
- TIMEX (date and time expressions)
- NUMEX (numerical expressions like money and percentages)

Depending on the application, additional types of entities can also be introduced such as proteins, medicines, etc., in medical text or particle names in quantum physics text. An example Turkish text annotated with ENAMEX entities can be seen in Fig. 6.1.

6.2 NER on Turkish

Initial studies on NER on Turkish texts started in the late 90s. Cucerzan and Yarowsky (1999) proposed a language independent bootstrapping algorithm that uses word-internal and contextual information about entities. They applied this approach to Turkish as well as four other languages. Tür (2000) and Tür et al. (2003) proposed an HMM-based NER system, that was specifically developed for Turkish, together with some other tools for similar information extraction related tasks. They also created the first widely used tagged Turkish newspaper corpora for the NER task. Later, Bayraktar and Temizel (2008) applied a local grammar approach to Turkish financial texts in order to identify person names. Küçük and Yazıcı (2009a,b) developed the first rule-based NER system for Turkish and applied it to Turkish news articles as well as to other domains like children's stories, historical texts, and speech recognition outputs. Dalkılıç et al. (2010) is another rule-based system for Turkish NER.

Recently, NER systems predominantly use machine learning approaches. Küçük and Yazıcı (2010, 2012) extended their rule-based system into a hybrid recognizer in order to perform better when applied to different domains. Yeniterzi (2011) explored the use of morphological features and developed a CRF-based NER system for Turkish. Other CRF-based systems were proposed by Özkaya and Diri (2011) and Şeker and Eryiğit (2012). Şeker and Eryiğit (2012) compared their system with other Turkish NER systems and among the ones that use the same data collection, their system outperformed other systems. Demir and Özgür (2014) developed a neural network based semi-supervised approach, which outperformed Şeker and Eryiğit (2012) over the same dataset (news articles) but without the use of gazetteers. Another notable approach (Tatar and Çiçekli 2011) proposed an automatic rule learner system for Turkish NER.

With the popularity and availability of social media collections, Turkish NER tools that can be used on more informal domains like tweets and forums have recently been developed (Küçük et al. 2014; Küçük and Steinberger 2014; Çelikkaya et al. 2013; Eken and Tantuğ 2015). Küçük et al. (2014) and Küçük and Steinberger (2014) applied rule-based NER systems to tweets. Çelikkaya et al. (2013) applied CRF-based approach of Şeker and Eryiğit (2012) to tweets, forums, and spoken data. Most recently Kısa and Karagöz (2015) applied NLP from Scratch approach to the NER task to propose more generalized models. They tested their system on both formal and informal texts.

6.3 Task Description

6.3.1 Representation

There are several ways to represent the named-entities and the choice of representation can have a big impact on the performance of NER systems. The most basic and simple format is to just use the raw named-entity tags by marking each token of a named-entity with a tag indicating its type. While simple, this has the important problem that it is not possible to annotate two or more consecutive named-entities properly.

The most common representation scheme for named-entities is the IOB2 representation (Tjong Kim Sang 2002) (also known as BIO). It is a variant of the IOB scheme (Ramshaw and Marcus 1995). With this representation, the first token of any named-entity gets the prefix “B-” in the tag type, and the “I-” prefix is used in the rest of the tokens in the named-entity if it involves multiple tokens. Tokens that are not part of a named-entity are tagged with “O”.

Table 6.1 An example tagging for Turkish (Raw, IOB2, and BILOU tags)

Token	Raw tag	IOB2 tag	BILOU
Mustafa	PERSON	B-PERSON	B-PERSON
Kemal	PERSON	I-PERSON	I-PERSON
Atatürk	PERSON	I-PERSON	L-PERSON
23	DATE	B-DATE	B-DATE
Nisan	DATE	I-DATE	I-DATE
1920	DATE	I-DATE	L-DATE
'de	O	O	O
Türkiye	ORGANIZATION	B-ORGANIZATION	B-ORGANIZATION
Büyük	ORGANIZATION	I-ORGANIZATION	I-ORGANIZATION
Millet	ORGANIZATION	I-ORGANIZATION	I-ORGANIZATION
Meclisi	ORGANIZATION	I-ORGANIZATION	L-ORGANIZATION
'ni	O	O	O
Ankara	LOCATION	B-LOCATION	U-LOCATION
'da	O	O	O
kurdu	O	O	O
.	O	O	O

Another representation scheme which is not as popular as IOB2 is the BILOU (Ratinov and Roth 2009). In contrast to BIO (IOB2), the BILOU scheme identifies not only the beginning, inside or outside of a named-entity, but also the last token using the "L-" prefix, in addition to identifying the unit length named-entities with a "U-" prefix. This scheme has been shown to significantly outperform the IOB2 representation (Ratinov and Roth 2009).

The named-entities in the following example Turkish sentence

> Mustafa Kemal Atatürk 23 Nisan 1920'de Türkiye Büyük Millet Meclisi'ni Ankara'da kurdu.
> "Mustafa Kemal Atatürk established the Turkish Grand National Assembly on 23 Nisan 1920 in Ankara."

can be represented with all these three formats as shown in Table 6.1.[1]

Among these representations, the IOB2 scheme has been used most commonly by Turkish NER systems (Yeniterzi 2011; Şeker and Eryiğit 2012; Çelikkaya et al. 2013; Önal et al. 2014). Tür (2000) and Tür et al. (2003) used a different representation, which has been shown to reduce the performance compared to IOB2 representation (Şeker and Eryiğit 2012). They showed that using raw labels is also not as effective as the IOB2 representation. Demir and Özgür (2014) seem to be the only ones who have used the BILOU representation in Turkish NER.

[1]Note that any suffixes on the last word of a named-entity is split as a separate token.

6.3.2 Evaluating NER Performance

Evaluation of NER performance has used three metrics: (1) MUC, (2) CoNLL, and (3) ACE. For Turkish NER, researchers have used the first two therefore only those will be detailed in this section. Detailed information on all these three evaluation metrics is available in Nadeau and Sekine (2007).

The MUC metric was initially used when NER was part of the understanding task in both the Sixth and Seventh Message Understanding Conferences (Sundheim 1995; Chinchor and Marsh 1998). This metric has two components that evaluate different aspects of NER tasks. MUC TEXT evaluates only the boundaries of the identified entities, and MUC TYPE evaluates whether the identified type of the entity is correct or not. For each of these two criteria, the following values are computed:

- *Correct*: number of named-entities recognized correctly by the system
- *Actual*: number of segments of tokens the system has indicated as named-entities by marking boundaries
- *Possible*: number of named-entities manually annotated in the data.

These values are used in *Precision* and *Recall* calculations as follows:

$$Precision = \frac{CorrectType + CorrectText}{ActualType + ActualText} \tag{6.1}$$

$$Recall = \frac{CorrectType + CorrectText}{PossibleType + PossibleText} \tag{6.2}$$

Recall measures the percentage of actual existing named-entities in a text that a system correctly recognizes, while precision measures the percentage of the named-entities that are correct among all the named-entities recognized by the system. The *f-measure*, the weighted harmonic mean of the *Precision* and *Recall* defined as

$$f\text{-}measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{6.3}$$

combines these into one quantity.

While MUC evaluates the identification and classification steps of the NER task separately, the CoNLL metric (Tjong Kim Sang 2002) is more strict, and only accepts labelings in which both the boundary (start and end positions) and the type of entity recognized are correct. A named-entity is counted as correct only if it is an exact match of the corresponding entity both in terms of boundary and type. Similar to MUC, CoNLL metric also uses the *f-measure* to report the finalized score.

MUC metric was commonly used in earlier Turkish NER research (Tür 2000; Tür et al. 2003; Bayraktar and Temizel 2008; Şeker and Eryiğit 2012) while more recent studies have preferred the CoNLL metric (Yeniterzi 2011; Şeker and Eryiğit

2012; Demir and Özgür 2014; Önal et al. 2014; Eken and Tantuğ 2015; Kısa and Karagöz 2015).[2]

6.4 Domain and Datasets

This section describes some commonly used data resources used for Turkish NER system development and evaluation. In general there is a distinction between formal and informal texts. While some basic preprocessing schemes like basic tokenization and morphological processing, etc., are usually enough for the formal datasets such as news texts, informal texts such as those found in social media abound with misspelled forms, incomplete or fragment sentences, require many additional preprocessing steps for more accurate NER.

6.4.1 Formal Texts

Formal texts include but are not limited to news articles and books. They can be defined as well-formed texts with correct spellings of words, proper sentence structure, and capitalization.

One of the first datasets for Turkish NER was created by Tür (2000). This dataset consists of news articles from Milliyet newspaper, covering the period between January 1, 1997 and September 12, 1998. This dataset was annotated with ENAMEX type entities and divided into a training set of 492,821 words containing 16,335 person, 11,743 location, and 9199 organization names, for a total of 37,277 named-entities, and a test set of about 28,000 words, containing 924 person, 696 location, and 577 organization names for a total of 2197 named-entities. Parts of this dataset have been widely used in other Turkish NER studies as well (Yeniterzi 2011; Şeker and Eryiğit 2012; Çelikkaya et al. 2013; Demir and Özgür 2014; Eken and Tantuğ 2015; Kısa and Karagöz 2015).[3]

Another newspaper dataset was constructed by Küçük and Yazıcı (2010) using METU Turkish corpus (Say et al. 2004) as the source. A total of 50 news articles were labeled in MUC style with ENAMEX, NUMEX, and TIMEX tags. This dataset contains 101,700 words with 3280 person, 2470 location, 3124 organization names along with 1413 date/time and 919 money/percent expressions. A subset of this dataset with ten news articles has been also used in several other studies (Küçük and Yazıcı 2009a,b; Kısa and Karagöz 2015).

[2]The evaluation scripts from the CONLL 2000 shared task can be found at github.com/newsreader/evaluation/tree/master/nerc-evaluation (Accessed on Sept. 14, 2017).

[3]The entity type counts are different in these studies due to either using different subsets or counting multiple token entities as one or not.

A financial news articles dataset was compiled by Küçük and Yazıcı (2010, 2012). This dataset contains 350 annotated financial news articles retrieved from a news provider, Anadolu Agency, with only person and organization names annotated. It comprises 84,300 words and has 5635 named-entities, 1114 are person names and 4521 are organization names.

Another Turkish newspaper text dataset is the TurkIE dataset (Tatar and Çiçekli 2011), which consists of 355 news articles on terrorism, with 54,518 words. The collection includes 1335 person, 2355 location, 1218 organization names, and 373 date and 391 time expressions for a total of 5672 named-entities.

Texts from two books have also been used in several Turkish NER studies (Küçük and Yazıcı 2009a,b, 2010, 2012). The first one consists of two children's stories, with around 19,000 words which contains manually annotated 836 person, 157 location, 6 organization names, and 65 date/time and 20 money/percent expressions. The second dataset comprises of the first three chapters of a book on Turkish history. It contains about 20,100 words and 387 person, 585 location, 122 organization names, and 79 date/time expressions, all manually annotated.

6.4.2 Informal Texts

Following the general trend in NLP, social media text has become a popular domain for the NER research in recent years. Özkaya and Diri (2011) have used an informal email corpora for NER. Çelikkaya et al. (2013) compiled two social media collections, one from an online forum and another from tweets. The first was from a crawl of a popular online forum for hardware product reviews www.donanimhaber.com (Accessed Sept. 14, 2017). With 54,451 words, this collection contains 21 person, 858 organization, 34 location names and 7 date, 2 time, 67 money, 11 percentage expressions (Çelikkaya et al. 2013). Kısa and Karagöz (2015) present some results from this dataset. The tweet dataset includes 54,283 words with around 676 person, 419 organization, 241 location names and 60 date, 23 time, 14 money, 4 percentage expressions. This tweet dataset has been also used in other NER studies (Küçük et al. 2014; Küçük and Steinberger 2014; Eken and Tantuğ 2015; Kısa and Karagöz 2015).

Another Turkish twitter dataset was compiled by Küçük et al. (2014) and Küçük and Steinberger (2014). Tweets posted on June 26, 2013 in between 12:00 and 13:00 GMT were crawled and after removing non-Turkish tweets, the total number of words was 20,752. In addition to the regular ENAMEX, TIMEX, and NUMEX tags, the authors also annotated TV program series, movies, music bands, and products (Küçük et al. 2014). This dataset includes 457 person, 282 location, 241 organization names, 206 date/time, 25 money/percent expressions, and 111 other named-entities. This collection was also used by Kısa and Karagöz (2015).

Finally, Eken and Tantuğ (2015) crawled and tagged around 9358 tweets consisting of 108,743 tokens with 2744 person, 1419 location, 2935 organization names, and 351 date, 86 time, 212 money expressions.

In addition to these social media texts, a spoken language dataset was also complied through a mobile application (Çelikkaya et al. 2013), by recording and converting spoken utterances into written text by using Google Speech Recognition Service. This dataset has 1451 words and contains 79 person, 64 organization, 90 location names and 70 date, 34 time, 27 money, 26 percentage expressions. This collection has been used in other studies as well (Kısa and Karagöz 2015).

Küçük and Yazıcı (2012) constructed two news video transcriptions data collections. The first one includes 35 manually transcribed news videos of around 4 h broadcast by Turkish Radio and Television (TRT). The second video data collection includes 19 videos with a total duration of 1.5 h. Unlike the first one, this video collection has been transcribed automatically using a sliding text recognizer.

6.4.3 Challenges of Informal Texts for NER

The switch from formal domains to these informal ones brings several challenges which cause significant reductions in NER performance (Ritter et al. 2011). As in other similar NLP tasks, the state-of-the-art NLP tools which assume properly constructed input texts may not perform as expected when applied to text in informal domains which contains a lot of misspelled words, ungrammatical constructs and extra-grammatical tokens such as user handles or hashtags.

For instance, Küçük et al. (2014) identified several peculiarities in informal texts especially in tweets. These include but not limited to grammar and spelling errors like incorrect use of capitalization, not using apostrophes to separate suffixes from named-entities, repeating letters for emphasis, using ASCII characters instead of proper Turkish characters. There are also some challenges due to size limitation in tweets leading to lack of useful contextual clues like person titles, professions, or using contracted forms of words or just using single forenames, surnames instead of the full names (Küçük et al. 2014). For instance, wrong use of capitalization and apostrophe makes it harder to recognize proper nouns which are also valid common nouns. Other spelling and grammar errors cause some language analysis tools like morphological analyzers to fail. Therefore, NER systems that depend on significant linguistic analysis of the texts may not perform as expected in such conditions.

In tweets there is also the case of named-entities occurring within a single hashtag but as a single token, for example, *#Istanbuldabahar*, or they can cover the whole hashtag like *#MustafaKemalAtaturk* (Küçük et al. 2014). Clearly these cases impose significant challenges for NER.

6.5 Preprocessing for NER

Depending on the NER system, there can be several data preprocessing steps that come before the identification of named-entities. These are tokenization, morphological analysis, and normalization.

6.5.1 Tokenization

Most NER systems use a word-level tokenizer. The apostrophe symbol that is used in standard formal Turkish orthography to indicate the boundary of the stem and suffixes in proper nouns can be used to split such tokens so that those suffixes appear as a separate token. Other punctuation characters that are not legitimate parts of tokens (e.g., decimal points) are considered as separate tokens (Şeker and Eryiğit 2012). Of course, other tokenization schemes are also possible: Yeniterzi (2011) has considered a morpheme-level tokenization where roots and connected morphemes were considered as separate tokens. The idea was to introduce explicit morphological information to the model, which, while not degrading the performance, did not produce a significant improvement. In her experiments, morpheme-level tokenization outperformed word-level tokenization in identification of person and location named-entities but caused drops for others.

6.5.2 Morphological Analysis

Morphological analysis is among the commonly used preprocessing steps. In order to deal with data sparsity issues, some NER systems use stems or root words in addition to the lexical form of the words. Also, some feature-based systems use inflectional morphemes to identify named-entities. Most Turkish NER systems (Yeniterzi 2011; Şeker and Eryiğit 2012; Eken and Tantuğ 2015) used Oflazer's two-level morphological analyzer (Oflazer 1994) to construct the morphological analysis of the word. A morphological disambiguator (Sak et al. 2011) was also used to resolve the morphological ambiguity.

Küçük and Yazıcı (2009a) also used their own morphological analyzer for their rule-based system. Their analyzer only considers the noun inflections on tokens which exist in the dictionaries and match an existing pattern.

In informal texts, like tweets, morphological analyzers do not work as expected because of spelling errors, capitalization errors, use of nonstandard orthographical forms or not using proper Turkish characters. In order to deal with these, some systems (Çelikkaya et al. 2013; Küçük and Steinberger 2014) have attempted normalizing text as described in the next section. Eken and Tantuğ (2015) proposed using the first and the last four characters instead of the root and inflectional

morphemes. Their experiments over tweets showed using such a heuristic provides similar results compared to using a morphological analyzer.

6.5.3 Normalization

As alluded to before, one way to deal with text in informal domains is to tailor the text so that NER systems developed over formal datasets can work with them. Several authors (Çelikkaya et al. 2013; Küçük and Steinberger 2014; Kısa and Karagöz 2015; Eken and Tantuğ 2015) have looked at this as a normalization procedure and applied steps to deal with the following:

- *Slang words*: Slang words are replaced with their more formal usage. For instance, *nbr* is replaced with *ne haber?—what's up?* (Çelikkaya et al. 2013)
- *Repeated characters*: Characters that are repeated for emphasis purposes but lead to a misspelled form are removed. (i.e., *çoooook* for *çok—many*) (Çelikkaya et al. 2013; Küçük and Steinberger 2014)
- *Special tokens*: Hash tags, mentions, smiley icons, and vocatives are replaced with certain tags (Çelikkaya et al. 2013)
- *Emo style writing*: Emo style writing and characters are replaced with their correct characters (i.e., *$eker 4 you* instead of *Seker senin için—Sweety! for you* (Çelikkaya et al. 2013)
- *Capitalization*: All characters are lowercased. (i.e., "aydin" for "Aydin") (Çelikkaya et al. 2013; Kısa and Karagöz 2015)
- *Asciification*: Special Turkish characters *(ç, ğ, ı, ö, ş, ü)* are replaced with equivalent nearest ASCII characters *(c, g, i, o, s, u)*. (Eken and Tantuğ 2015)

Çelikkaya et al. (2013) applied the CRF-based approach of Şeker and Eryiğit (2012) to one formal and three types of informal texts with different subsets of features. While normalization provided observable improvements when applied to tweets, it degraded the performance when applied to formal news dataset, and did not result in an improvement with forum and speech datasets (Çelikkaya et al. 2013). Overall, for informal domains, there is still room for improvement.

Apart from normalizing informal texts like tweets, normalization can also be applied to formal texts to make generalizations. For instance, Demir and Özgür (2014) normalized all numerical expressions into a generic number pattern so that unseen number tokens during testing could be handled properly.

6.6 Approaches Used in Turkish NER

The approaches for NER task can be divided into three main categories: (1) hand-crafted rule-based systems, (2) machine learning based systems, and (3) combination of the first two, hybrid systems. In this section, we review Turkish NER

systems and categorize them with respect to these approaches and describe some in detail. Even though it is impossible to make a fair comparison among these systems due to the differences between datasets used, their highest performance scores are nevertheless reported in order to give the reader some idea of the state-of-the-art performance.

6.6.1 Rule-Based Approaches

Küçük and Yazıcı (2009a,b) developed the first rule-based NER system for Turkish. They used two types of information sources: (1) lexical resources and (2) patterns. Lexical resources consists of a gazetteer of person names and lists of well-known people, organizations, and locations. Pattern bases include manually constructed patterns for identifying location names, organization names, and temporal and numerical expressions. Example patterns are as follows:

- Patterns for location names:

  X Sokak/Yolu/Kulesi/Stadyumu/...
  X Street/Road/Tower/Stadium/...

- Patterns for organization names:

  X Grubu/A.Ş./Partisi/Üniversitesi/...
  X Group/Inc./Party/University/...

- Patterns for temporal and numeric expressions:

  X başı/ortası/sonu...
  X start/middle/end...
  'The start/middle/end... of X'

While the authors targeted news text, they also tested their system over different text genres, including children's stories, historical texts, and news video transcriptions. Since not all these (like video transcriptions) have proper capitalization and punctuation, they were not able to exploit these clues for NER. The f-measures for their system were 78.7% on news articles, 69.3% on children's stories, 55.3% on historical texts, and 75.1% on video transcriptions. Even though their results were not even close to the state-of-the-art systems at that time, this study can be considered as a good baseline point for rule-based Turkish NER systems.

This system has been also applied to informal text like tweets with some simple modifications, in order to deal with the peculiarities of the data (Küçük and Steinberger 2014). Due to lack of proper use of capitalization in such texts, the authors initially relaxed the capitalization constraint of the system. They also extended their lexical resources to include both diacritic and non-diacritic variants of the entries. Several tweet normalization techniques were also applied. Experiments over two different tweet collections showed that these modifications were useful

(Küçük and Steinberger 2014). Önal et al. (2014) also applied a rule-based approach inspired by Küçük and Yazıcı (2009a,b) to tweets in order to identify locations.

Küçük et al. (2014) used Europe Media Monitor (EMM) multilingual media analysis and information extraction system (Pouliquen and Steinberger 2009) for Turkish NER. EMM is a language independent rule-based system which uses dictionary lists which contain language-specific words for titles, professions, etc. The EMM system can be adapted to a language by using these lists together with some capitalization related rules. Küçük et al. (2014) identified frequently mentioned person and organization names from news articles and used them to extend the existing resources of the system and applied it to Turkish tweets. On news domain they got an f-measure of 69.2% while on tweets the f-measure was 42.7%.

Dalkılıç et al. (2010) proposed another rule-based system where tokens and morphemes that frequently occur close to person, organization, and location entities can be used to classify other entities. This system was tested over economics, politics, and health domain texts and the best performance was observed in identifying locations with an f-measure of 87.0% on average. Unlike location, person and organization identification performances are lower with f-measures of 80.0% and 81.0%, respectively.

Bayraktar and Temizel (2008) used a system with several manually constructed patterns to identify person named-entities. They applied a local grammar approach (Traboulsi 2006) to recognize person names from Turkish financial texts.[4] Bayraktar and Temizel (2008) initially identified common reporting verbs in Turkish, such as *dedi* (said), *sordu* (asked), then they used these reporting verbs to generate patterns for locating person names. This approach returned an f-measure of 82.0% on news articles.

6.6.2 Hybrid Approaches

The problem with the rule-based systems is that they require the addition of more and more rules and their performance degrades when ported to new domains. In order to overcome this problem, Küçük and Yazıcı (2009a,b) extended their rule-based NER tool into a hybrid recognizer (Küçük and Yazıcı 2010, 2012), so that in a new domain, it can learn from the available annotated data and extend its knowledge resources. They used rote learning (Freitag 2000), which basically groups and stores available named-entities in the training set. When applying this system on different domains, the system starts with the same set of patterns and lexicons, but in the learning stage, it adapts itself to the particular domain by learning

[4] A local grammar is "a way of describing the syntactic behavior of groups of individual elements, which are related but whose similarities cannot be easily expressed using phrase structure rules" (Mason 2004).

from the new domain's training data. Küçük and Yazıcı (2009a,b) used their rule-based NER system originally targeted for news texts, and applied it to financial news texts, historical texts, and children's stories. In these experiments, the hybrid entity recognizer outperformed the rule-based system with an f-measure of 86.0% on news, 74.2% on financial news, 85.0% on child stories, and 67.0% on historical texts.[5] These scores were improved further (up to 90.1% on news domain) when they turned on the capitalization feature.

Yavuz et al. (2013) proposed another hybrid approach where they use a Bayesian learning together with the rule-based system by Küçük and Yazıcı (2009a,b).

6.6.3 Machine Learning Approaches

Due to their ability to learn from annotated data and not relying on hand-crafted-rules, and easy adaptability to new domains, machine learning approaches have been used widely in developing NER systems. These approaches however depend on having datasets where named-entities of interest are properly annotated.

The first work on Turkish NER describes a language independent EM-style bootstrapping algorithm that learns from word internal and contextual information of entities (Cucerzan and Yarowsky 1999). The bootstrapping algorithm is a semi-supervised learning algorithm, which starts with a seed set of examples or patterns and iteratively learns new patterns using the clues seeds provide. The authors used hierarchically smoothed trie structures for modeling the word internal (morphological) and contextual probabilities. The first set of clues refers to the patterns of prefixes or suffixes which are good indicators of a named-entity. For instance, for Turkish, '*-oğlu*' (*son of*) is a strong surname indicator. The contextual patterns either preceding or following a named-entity can also help identify them: for example, "*Bey*" (*Mr.*) or "*Hanım*" (*Mrs.*) can help identify preceding words as person names. Turkish was one of the five languages evaluated (along with English, Greek, Hindi, and Romanian). With a training size of 5207 tokens and 150 seeds, an f-measure of 53.0% was reported for Turkish.

Tür (2000) and Tür et al. (2003) developed a statistical name tagger system specifically for Turkish which depends on n-gram language models embedded in HMMs. They used four information sources and augmented lexical model with contextual, morphological, and tag models. In their lexical model, which can be considered as a baseline, they only used the lexical forms of the tokens. A word/tag combination HMM was built and trained, where a tag represents whether the word is part of a named-entity and if so its type. In the contextual model, in order to deal with words that do not appear in training data, they built another model with named entities tagged as *unknown*. This model provided useful clues regarding

[5]The data collection used in this study is not exactly the same with data used in Küçük and Yazıcı (2009a,b).

Table 6.2 F-measure results from Tür et al. (2003)

Model	Text	Type	F-measure
Lexical	80.87%	91.15%	86.01%
Lexical + Contextual	86.00%	91.72%	88.86%
Lexical + Contextual + Morphological	87.12%	92.20%	89.66%
Lexical + Contextual + Tag	89.54%	92.13%	90.84%
Lexical + Contextual + Morphological + Tag	90.40%	92.73%	91.56%

Table 6.3 F-measure results from Yeniterzi (2011)

Model	Person	Location	Organization	Overall
Lexical	80.88%	77.05%	88.40%	82.60%
Lexical + Root	83.32%	80.00%	90.30%	84.96%
Lexical + Root + POS	84.91%	81.63%	90.18%	85.98%
Lexical + Root + POS + Prop	86.82%	82.66%	90.52%	87.18%
Lexical + Root + POS + Prop + Case	88.58%	84.71%	91.47%	88.71%

the preceding and following tokens inside and around the named entities. Their morphological model captures information related to the morphological analysis of the token. The name tag model ignores the lexical form of the words and only captures the name tag information (like person, location, organization) of the words. Using only the tags and boundary information is useful for identifying multi-token named entities.

Tür (2000) and Tür et al. (2003) used MUC scoring to evaluate these four models and their combinations. The experimental results including both *text* and *type* and the overall *f-measure* scores of these models are summarized in Table 6.2. The baseline lexical model starts with 86.01% f-measure. Using the contextual cues in recognizing unknown words returned improvements up to 5.13% in *text* score. Furthermore, incorporating the tag model increased the *text* score by more than 3% points due to decreasing the improbable tag sequences. Combination of all these four models provided the best performance with 91.6% f-measure.

Conditional Random Fields (CRF) (Lafferty et al. 2001) have been used in several Turkish NER tools. Yeniterzi (2011) built a CRF-based NER tool for Turkish where she used features like stem, part-of-speech, proper noun markers, and case markers, in addition to the lexical form of the token. The individual effects of these features are summarized in Table 6.3. As a morphologically rich language, even adding the root (stem) as a feature to the lexical model improved the system by 2–3%. Other exploited features provided 1–2% improvements to the system individually, which at the end resulted in around 6% improvement in overall f-measure.

In order to see the effects of morphology more clearly, Yeniterzi (2011) also employed a morpheme level tokenization in which a word is represented in several states in the CRF: one state for the root and one state for each morphological feature.

Morpheme-level tokenization model, exploiting the same set of features as in the case of word-level model, improved the overall f-measure to 88.94%.

Özkaya and Diri (2011) applied a CRF-based NER system to emails. They used features like capitalization, punctuation, context, and email field related features like whether the token belongs to *from*, *to*, or other similar fields. The system showed the highest performance in the identification of person named-entities with an 95% f-measure. The authors did not explore the impact of specific features over the results, so it is possible that the field related feature can be the determining component.

Şeker and Eryiğit (2012) also proposed a CRF-based system. Similar to Yeniterzi (2011), they employed lexical and morphological features like stem, part-of-speech, noun case, proper noun markers, various inflectional features. They applied the approach in Sha and Pereira (2003) to these features and manually selected the useful ones. All these added features improved the performance of the system to an f-measure of 91.9%, and outperformed some of the prior work. Çelikkaya et al. (2013) applied a similar approach to tweets, forum, and speech datasets. For training they used the same news dataset used by Şeker and Eryiğit (2012). As expected the CRF model performed at a much worse level when tested on these informal domains with f-measures 6.9% with speech dataset, 5.6% with forum dataset, and 12.2% with tweets. Even though the performance of tweets increased to 19.3% after normalizing them, the performance level was not comparable to that on formal datasets. Önal et al. (2014) also applied this approach to tweets just to recognize locations. Eken and Tantuğ (2015) compared the approach of Şeker and Eryiğit (2012) by using a simpler preprocessing used the first and last four characters of tokens instead of features extracted from morphological analysis of words. Their model exhibited a similar performance to the morphological model. This model which was trained on news articles was tested over tweets with low performance as expected but when training was performed over tweets, the test provided an f-measure 64.03 on tweets.

Tatar and Çiçekli (2011) proposed an automatic rule learning system for NER task. They started with a set of seeds selected from the training set, and then extracted rules over these examples. They generalized the named-entities by using contextual, lexical, morphological, and orthographic features. During this generalization procedure, they used several rule filtering and refinement techniques in order to keep their accuracy high with an f-measure of 91.1%.

Yavuz et al. (2013) were the first to apply the Bayesian Learning approach to Turkish NER. They employed a modified version of the BayesIDF approach (Freitag 2000) with features like case sensitivity, case, token length, etc., which exhibited an f-measure of 88.4%. Two hybrid systems were also constructed by combining this system with a rule-based system (Küçük and Yazıcı 2009a,b). In the first system the training data was used to train the Bayesian learner, and then the rule-based tagged NER data was used as additional training data to update the system. In the second system, the tagged output of the rule-based system was used as an additional feature by the Bayesian learner. Both hybrid systems outperformed the Bayesian learner alone, the first one with an f-measure of 90.0% and the second with 91.4%.

Another semi-supervised approach to Turkish NER was recently proposed by Demir and Özgür (2014). Their neural network based approach had two stages.

In the unsupervised stage, neural networks were used to obtain continuous vector representation of words by using large amounts of unlabeled data. In the supervised stage, these feature vectors and additional language independent features like capitalization patterns of previous tag predictions were used in another neural network to train the NER system. These word representations were also clustered to identify semantically similar words, and cluster ids were used as additional feature. This system has an f-measure of 91.85.

Another recently published semi-supervised approach to NER has also used word embeddings (Kısa and Karagöz 2015). The author have applied NLP from Scratch method (Collobert et al. 2011) to NER on social media texts. Initially a language model and word embeddings were learned from a large unannotated dataset and later these word embeddings were used as features to train a neural network classifier on labeled data. The authors have experimented with different datasets and domains: On formal text their approach outperformed the rule-based system of Küçük and Yazıcı (2009a) but was not better than the CRF-based system by Şeker and Eryiğit (2012) or neural network-based approach of Demir and Özgür (2014). However, when applied to informal texts, this system also outperformed a CRF-based system (Çelikkaya et al. 2013).

6.7 Conclusions

This section presented an overview of Turkish NER systems that have been developed in the last two decades, covering their salient aspects and performance, in addition to pointing out some of the datasets used for developing such systems. It is clear that there is significant room for improvement for Turkish NER systems especially in informal text domains and while performance of these systems is reasonably high on formal texts, further improvements and quick adaptability are the ongoing concerns.

References

Bayraktar Ö, Temizel TT (2008) Person name extraction from Turkish financial news text using local grammar based approach. In: Proceedings of ISCIS, Istanbul

Çelikkaya G, Torunoğlu D, Eryiğit G (2013) Named entity recognition on real data: a preliminary investigation for Turkish. In: Proceedings of the international conference on application of information and communication technologies, Baku

Chinchor N, Marsh E (1998) Appendix D: MUC-7 information extraction task definition (version 5.1). In: Proceedings of MUC, Fairfax, VA

Collobert R, Weston J, Bottou L, Karlen M, Kavukçuoğlu K, Kuksa P (2011) Natural language processing (almost) from scratch. J Mach Learn Res 12:2493–2537

Cucerzan S, Yarowsky D (1999) Language independent named entity recognition combining morphological and contextual evidence. In: Proceedings of EMNLP-VLC, College Park, MD, pp 90–99

Dalkılıç FE, Gelişli S, Diri B (2010) Named entity recognition from Turkish texts. In: Proceedings of IEEE signal processing and communications applications conference, Diyarbakır, pp 918–920

Demir H, Özgür A (2014) Improving named entity recognition for morphologically rich languages using word embeddings. In: Proceedings of the international conference on machine learning and applications, Detroit, MI, pp 117–122

Doddington G, Mitchell A, Przybocki M, Ramshaw L, Strassel S, Weischedel R (2004) The automatic content extraction (ACE) program–tasks, data, and evaluation. In: Proceedings of LREC, Lisbon, pp 837–840

Eken B, Tantuğ C (2015) Recognizing named-entities in Turkish tweets. In: Proceedings of the international conference on software engineering and applications, Dubai

Freitag D (2000) Machine learning for information extraction in informal domains. Mach Learn 39(2–3):169–202

Kısa KD, Karagöz P (2015) Named entity recognition from scratch on social media. In: Proceedings of the international workshop on mining ubiquitous and social environments, Porto

Küçük D, Steinberger R (2014) Experiments to improve named entity recognition on Turkish tweets. Arxiv – computing research repository. arxiv.org/abs/1410.8668. Accessed 14 Sept 2017

Küçük D, Yazıcı A (2009a) Named entity recognition experiments on Turkish texts. In: Proceedings of the international conference on flexible query answering systems, Roskilde, pp 524–535

Küçük D, Yazıcı A (2009b) Rule-based named entity recognition from Turkish texts. In: Proceedings of the international symposium on innovations in intelligent systems and applications, Trabzon

Küçük D, Yazıcı A (2010) A hybrid named entity recognizer for Turkish with applications to different text genres. In: Proceedings of ISCIS, London, pp 113–116

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

Küçük D, Jacquet G, Steinberger R (2014) Named entity recognition on Turkish tweets. In: Proceedings of LREC, Reykjavík, pp 450–454

Lafferty JD, McCallum A, Pereira F (2001) Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Proceedings of ICML, Williams, MA, pp 282–289

Mason O (2004) Automatic processing of local grammar patterns. In: Proceedings of the annual colloquium for the UK special interest group for computational linguistics, Birmingham, pp 166–171

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

Oflazer K (1994) Two-level description of Turkish morphology. Lit Linguist Comput 9(2):137–148

Önal KD, Karagöz P, Çakıcı R (2014) Toponym recognition on Turkish tweets. In: Proceedings of IEEE signal processing and communications applications conference, Trabzon, pp 1758–1761

Özkaya S, Diri B (2011) Named entity recognition by conditional random fields from Turkish informal texts. In: Proceedings of IEEE signal processing and communications applications conference, Antalya, pp 662–665

Pouliquen B, Steinberger R (2009) Automatic construction of multilingual name dictionaries. In: Goutte C, Cancedda N, Dymetman M, Foster G (eds) Learning machine translation. The MIT Press, Cambridge, MA, pp 266–290

Ramshaw LA, Marcus MP (1995) Text chunking using transformation-based learning. In: Proceedings of the workshop on very large corpora, Cambridge, MA, pp 82–94

Ratinov L, Roth D (2009) Design challenges and misconceptions in named entity recognition. In: Proceedings of CONLL, Boulder, CO, pp 147–155

Ritter A, Clark S, Mausam, Etzioni O (2011) Named entity recognition in tweets: an experimental study. In: Proceedings of EMNLP, Edinburgh, pp 1524–1534

Sak H, Güngör T, Saraçlar M (2011) Resources for Turkish morphological processing. Lang Resour Eval 45(2):249–261

Say B, Zeyrek D, Oflazer K, Özge U (2004) Development of a corpus and a treebank for present-day written Turkish. In: Proceedings of the international conference on Turkish linguistics, Magosa, pp 183–192

Şeker GA, Eryiğit G (2012) Initial explorations on using CRFs for Turkish named entity recognition. In: Proceedings of COLING, Mumbai, pp 2459–2474

Sha F, Pereira F (2003) Shallow parsing with conditional random fields. In: Proceedings of NAACL-HLT, Edmonton, pp 134–141

Sundheim BM (1995) Overview of results of the MUC-6 evaluation. In: Proceedings of MUC, Columbia, MD, pp 13–31

Tatar S, Çiçekli İ (2011) Automatic rule learning exploiting morphological features for named entity recognition in Turkish. J Inf Sci 37(2):137–151

Tjong Kim Sang EF (2002) Introduction to the CoNLL-2002 shared task: language-independent named entity recognition. In: Proceedings of CONNL, Taipei, pp 1–4

Tjong Kim Sang EF, De Meulder F (2003) Introduction to the CoNLL-2003 Shared Task: language-independent named entity recognition. In: Proceedings of CONLL, Edmonton, pp 142–147

Traboulsi HN (2006) Named entity recognition: a local grammar-based approach. PhD thesis, Surrey University, Guildford

Tür G (2000) A statistical information extraction system for Turkish. PhD thesis, Bilkent University, Ankara

Tür G, Hakkani-Tür DZ, Oflazer K (2003) A statistical information extraction system for Turkish. Nat Lang Eng 9:181–210

Yavuz SR, Küçük D, Yazıcı A (2013) Named entity recognition in Turkish with Bayesian learning and hybrid approaches. In: Proceedings of ISCIS, Paris, pp 129–138

Yeniterzi R (2011) Exploiting morphology in Turkish named entity recognition system. In: Proceedings of ACL-HLT, Portland, OR, pp 105–110