Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 4546–4551
Marseille, 20-25 June 2022

A Twitter Corpus for Named Entity Recognition in Turkish

Buse Çarık, Reyyan Yeniterzi
Sabanci University
Istanbul, Turkey
{busecarik, reyyan}@sabanciuniv.edu

Abstract

This paper introduces a new Turkish Twitter Named Entity Recognition dataset. The dataset, which consists of 5000 tweets from a year-long period, was labeled by multiple annotators with a high agreement score. The dataset is also diverse in terms of the named entity types as it contains not only person, organization, and location but also time, money, product, and tv-show categories. Our initial experiments with pretrained language models (like BertTurk) over this dataset returned F1 scores of around 80%. We share this dataset publicly.

Keywords: Twitter, Named Entity Recognition, Turkish

1. Introduction

Named Entity Recognition (NER), a subtask of information extraction is used to identify predefined named entities (NEs) such as temporal and numerical expressions alongside person, location, or organization names. Researchers have achieved outstanding results in well-studied languages such as English for the NER task, which has been used as part of several other NLP tasks such as summarization, question answering, and entity linking. However, it remains a subject of study for languages that lack sufficient research and resources, such as Turkish.

There are several reasons for this. First of all, the majority of previous studies in Turkish NER focused on formal writings that comply with grammatical and spelling rules (Tür et al., 2003; Tatar and Cicekli, 2011). In limited studies to date (Çelikkaya et al., 2013; Okur et al., 2018), the application of the developed NER models to informal texts such as mini-blogs has yielded poor results. Nevertheless, with the growth in the amount of social media content, the need to recognize NEs in these noisy texts has increased.

In addition, apart from the NE types defined by the Message Understanding Conference (MUC) series (Grishman and Sundheim, 1996), new NE types have not been adequately studied for Turkish except for a few studies (Küçük et al., 2014). However, investigating different types of NEs helps improving the results when these NE types are tailored to other NLP tasks.

Another significant issue in Turkish NER studies is that most of the datasets, especially the informal ones, are not publicly available. Only Küçük et al. (2014) and Küçük and Can (2019) released their datasets of tweets publicly by providing the tweet IDs. However, these two datasets are limited both in terms of diversity and size.

In this work, we introduce a new dataset for Turkish NER gathered from Twitter distributed uniformly across months over a long time. We included understudied NE types in our label set and obtained a high agreement score among multiple annotators. We also present some initial results on this dataset. Since transformer-based models outperform in many NLP tasks, we experimented with different variations of these models on our dataset as well. The dataset is publicly available at `https://github.com/SU-NLP/SUNLP-Twitter-NER-Dataset`.

The rest of this paper is organized as follows: Section 2 discusses the overview of NER in Turkish; Section 3 describes the details of the data collection and annotation processes; Section 4 presents our initial experiments on the developed NER model and discusses our results; and finally, Section 5 concludes the paper.

2. Related Work

There are several attempts to create formal or informal datasets in the Turkish NER. The first Turkish NER dataset, which is also the largest one with 500K tokens, is a dataset of news articles annotated with NE categories of person, organization, and location (Tür et al., 2003). Later, Tatar and Cicekli (2011) built a relatively small formal news dataset on terrorism with 55K tokens. In this study, both the money and percent NE types were included in the annotation process. With a rule-based system, they achieved an F1 score of 91.08% in this dataset. In a later study (Küçük et al., 2016), 89.85% was obtained with a rule-based method on a substantially small dataset of 20K words constructed using news.

Although the number of informal datasets is greater than the formal ones, none of them is close to the size of the Tür et al. (2003) dataset. The first study introduced three informal datasets from different sources (Çelikkaya et al., 2013). Their Twitter dataset contains 5K tweets with 54K tokens. A forum with hardware product evaluations provided the broadest dataset, which contained 54K words. Another dataset was created with text-to-speech data converted by a mobile assistant application, and all of them were annotated with the seven basic NE types (person, location, organization, time, date, money, percentage). The largest dataset on informal texts was created by Tantug (2015)

	Dataset	Source	Number of Tokens	Number of NEs	Availability
Formal	(Tür et al., 2003)	News	500K	40K	Available
	(Küçük et al., 2016)	News	20K	1,425	Not Available
	(Tatar and Cicekli, 2011)	News	55K	5,672	Not Available
Informal	(Çelikkaya et al., 2013)	Twitter	54K	1,437	Not Available
	(Tantug, 2015)	Twitter	108K	7,747	Not Available
	(Seker and Eryigit, 2017)	UGC	43K	1,162	Not Available
	(Küçük et al., 2014)	Twitter	21K	1,322	Only Tweet IDs
	(Küçük and Can, 2019)	Twitter	-	1,879	Only Tweet IDs

Table 1: Formal and informal NER datasets in Turkish.

from Twitter, labeling 9,358 tweets with seven basic categories. Unfortunately, these datasets are not publicly available.

Within an hour, Küçük et al. (2014) collected 2300 tweets from Twitter and labeled them with person, location, organization, money, date, time, and percentage tags. They also put all TV shows, songs, and products under the MISC category. This dataset is limited due to covering a short period of time.

Another dataset was introduced on user-generated content from different domains, such as customer reviews, social media posts, blogs, and forums (Seker and Eryigit, 2017).

A recent study annotated 1,065 tweets about Turkish sports teams with person, location, and organization labels. In this dataset, the labeling was performed by a single annotator (Küçük and Can, 2019). Another limitation of this dataset is that tweets are about a very specific domain. The statistics about all these Turkish datasets are presented in Table 1.

The majority of studies on Turkish NER have been conducted with (Tür et al., 2003) dataset since it is the largest dataset available. Earlier studies concentrated on statistical and rule-based systems, whereas recent research has focused on deep learning approaches. As a statistical method, Tür et al. (2003) applied an approach based on the Hidden Markov Models. CRF-based methods were later proposed by Yeniterzi (2011; Şeker and Eryiğit (2012). Küçük and Yazıcı (2012; Tatar and Cicekli (2011) experimented with rule-based approaches in their small datasets. The first study to utilize a neural network was Demir and Özgür (2014) which developed a regularized averaged perceptron on the Tür et al. (2003) news dataset. Later studies have explored the BiLSTM model on top of CRF through different embedding settings, such as utilizing characters or morphological features (Kuru et al., 2016; Güneş and Tantuğ, 2018; Güngör et al., 2019). With the popularity of pretrained language models, recent Turkish NER studies have begun to use these models as well. The current state-of-the-art model was achieved by (Aras et al., 2021) with a 95.95% F1 score by implementing a CRF layer on top of the BERTurk[1] model.

Although the scores achieved in the formal datasets are considerably high, the results are significantly low when these methods are applied to the informal ones. When Çelikkaya et al. (2013) applied the same system presented in the (Şeker and Eryiğit, 2012) to their datasets, F1 scores of 19% on Twitter, 50.84% on speech, and 5.6% on the forum were obtained. One of the important factors causing this decrease from 91.94% to 19% in the transition from the news (Tür et al., 2003) to Twitter data is that they carried out the training process over the news dataset since there was not sufficient Twitter data for training. A multilingual rule-based approach developed by Küçük and Steinberger (2014) obtained 38.01% on (Çelikkaya et al., 2013) and 48.13% on (Küçük et al., 2014). The first study that used an informal dataset for training is (Tantug, 2015) and achieved a 64.03% F1 score with a CRF-based method. Okur et al. (2018) obtained 48.96% F1 score on (Çelikkaya et al., 2013) by utilizing a Word2Vec trained on a large informal dataset in a regularized averaged multi-class perceptron model.

3. NER Dataset

In this section, we describe the dataset collection and annotation steps in detail. We also provide an analysis of the collected annotations.

3.1. Data Collection

The data was collected through the Twitter streaming API from June 2020 to June 2021. We obtained approximately 65 million tweets in this period using the top trending topics in Turkey. Although the tweets covered a wide range of topics due to the broad time interval, hotly-debated events may dominate other subjects in several time intervals. Since it is beneficial to include varied topics to improve the generalizability of the current systems, we tried to generate a diverse dataset. Furthermore, since not all tweets contain a named entity, in order to get the most out of the annotation process, we used several heuristics while creating the dataset.

The following steps were performed for selecting tweets to be annotated. Firstly, tweets that have the same content without considering mentions, hashtags, and URLs were eliminated. After this near duplicate removal, in order to increase the chance of including a NE, we only kept the tweets with a character length

[1] https://huggingface.co/dbmdz/bert-base-turkish-cased

greater than 50 and removed the rest. Moreover, to ensure having at least one NE in the tweet, we fed our remaining tweets to an effective NER model and selected those that had at least one previously unseen NE in its predictions. For this model, we used a BERT (Devlin et al., 2018) model pretrained on large Turkish corpora[2] and fine-tuned it on a well-studied and largest Turkish NER corpus (Tür et al., 2003). This corpus contains only person, organization, and location entities, therefore it is limited but still better than no filtering at all. After this filtering, in order to guarantee a diversity of topics, we decided that any one hashtag can be in a maximum of 3 tweets. After this final filtering, we randomly selected 5,000 tweets from the remaining ones and manually annotated them. The dataset contains a total of 126,228 words, with an average of 25.24 words per tweet.

3.2. Named Entity Types

In addition to the most common three NE types, person, location, and organization, four other NE types have been annotated in this data set. We followed the definitions in the MUC (Grishman and Sundheim, 1996) for NE types *PERSON, ORGANIZATION, LOCATION, and MONEY*. The remaining two types are *PRODUCT* and *TV-SHOW*. We defined *PRODUCT* as an item produced or manufactured by people or corporations. Songs, books, movies, Instagram, and an iPhone can be given as examples for this class. We noticed that Turkish TV shows are often among the trending hashtag topics on Twitter. Therefore, we used a more specific type as the *TV Show* category for soap operas, reality programs, and other TV shows broadcast on TV. Besides, we have considered time and date expressions as part of the *TIME* class. We did not include percentages in numerical expressions as we could not see any significant number of samples in the annotation process.

3.3. Annotation Process

Our annotation team consists of four undergraduate students whose native language is Turkish. We distributed the selected 5000 tweets to these annotators and made sure that each tweet was annotated by two annotators. Label Studio[3], an open-source labeling tool, was used during the annotation process due to its user-friendly and easy-to-learn interface.

In addition to the context in tweets, annotators also labeled the hashtags if it is a NE as a whole (except for the # character). Hashtags in which the NE is only a part of, were not annotated as Named Entity. For example, if the hashtag is *#Fenerbahçe*, it was labeled as *ORGANIZATION*. However, if the hashtag is *#ŞampiyonFenerbahçe*, this token was labeled as *OTHER*.

[2]https://huggingface.co/dbmdz/bert-base-turkish-128k-uncased

[3]https://labelstud.io/

The inter-annotator agreement was measured for all tweets in our dataset. The Cohen kappa score is 0.94 when all tokens are included (including the OTHER label). It is 0.87 when only the seven NEs are considered (without the OTHER label). There were 845 disagreements among the 5,000 tweets. After a detailed examination of these conflicts, we observed that the annotators mostly disagreed in the following two situations: *ORGANIZATION* vs. *LOCATION* and *ORGANIZATION* vs. *PRODUCT*.

For the conflicts between *ORGANIZATION* and *LOCATION* annotators usually could not agree on whether countries were mentioned as a place or a state. For example, consider the following tweets:

- **LOCATION:** *We are going to the beautiful beaches of Turkey on vacation in summer.*
- **ORGANIZATION:** *Negotiations between Turkey and the USA continue.*

The first tweet refers to Turkey as a location since it is about its coasts, whereas in the second tweet, it is an organization since the tweet is about the negotiations between governments. Some annotators had a hard time differentiating these concepts in some tweets.

Another popular conflicting case is deciding whether a named entity is a *PRODUCT* or *ORGANIZATION*. Although our annotators accurately categorized the corporations as organizations, in some cases, their goods were annotated as organizations instead of a product. For instance, while the company *Apple* is an organization, *iPhone* is a product of this company. Unfortunately, this becomes more challenging when both the company and product share the same name. For instance, in the following tweets, *Google* is used as a search engine product in the first one and a company in the second one.

- **PRODUCT:** *If you are not sure, just ask it to Google.*
- **ORGANIZATION:** *I will start working at Google starting next month :)*

Our expert author on the NER task resolved these conflicts one-by-one manually. In the finalized dataset, we have a total of 11,081 NEs, with the largest classes being *PERSON*, *ORGANIZATION*, and *LOCATION*. The number of distinct NEs is 7,231. Table 2 illustrates the distribution of the NEs in our dataset.

According to Table 2, common NE types are also the most common ones here. *PERSON* is the most frequent type. It is followed by *ORGANIZATION* and *LOCATION*. Time expressions are very common on Twitter, hence the high frequency of *TIME* is also expected. *PRODUCT* and *TVSHOW* are low in frequency, but one should not forget that *TVSHOW* can be considered as a type of *PRODUCT*, therefore when considered together, it is quite high in frequency.

NE Type	Count
PERSON	5,526
ORGANIZATION	2,956
LOCATION	1,243
TIME	608
PRODUCT	334
TV-SHOW	255
MONEY	159
Total	**11,081**

Table 2: The distribution of NEs in our dataset.

3.4. Annotation Format

The adapted annotation format for our dataset is the **IOB2** tagging scheme, also known as **BIO** (Sang and Veenstra, 1999). In this format, **B-** stands for the NE beginning with that token. And if the entity is followed by more tokens, they take **I-** tags, which stand for INSIDE. An example of a **IOB2** format is illustrated in Table 3.

Tokens	IOB2 tags
Sergen	B-PERSON
Yalçın	I-PERSON
Beşiktaş	B-ORGANIZATION
'	O
ta	O
kaldı	O
Bülent	B-PERSON
Uslu	I-PERSON
çarpıcı	O
değerlendirmelerde	O
bulundu	O

Table 3: Example of the IOB2 format

4. Named Entity Recognition Model

In this section, we present our baseline NER models built with our Twitter dataset described in Section 3.

4.1. Experimental Setup

Firstly, we replaced the URL links with *$URL* special token, as they do not add any knowledge to the context of tweets. In addition, *@USER* token was used instead of mentions in the tweets in order to ensure privacy. Using these specific tokens is also useful for modeling since the tokenizers of the pretrained models we use, probably do not know the representation of these words anyway.

We conducted our experiments on validation and test sets, each consisting of 750 randomly selected tweets. The remaining 3,500 tweets were used for training. The results were reported with Precision, Recall, and F1 metrics computed for the entire NE spans.

4.2. Models

Since transformer-based pre-trained models outperform in a variety of NLP tasks and datasets, we investigated variations of these models as a baseline in this paper as well.

BERTurk, BERT_loodos[4], and ALBERT_loodos[4] transformer models which were pretrained on Turkish corpora were used. Similarly, various multilingual models mBERT[5] and XLM-RoBERTa[6] were applied to our task.

For the Turkish models, the type of text utilized during pretraining is different. While the BERTurk model was pre-trained on the Turkish Wikipedia dump, the OSCAR[7], and the OPUS[8] datasets, which contain fewer spelling and grammatical errors, the data used in Loodos' training includes informal text such as Twitter and online blogs. The same corpora were utilized in the training of both BERT_loodos and ALBERT_loodos as well.

All the BERT models listed above are base models, and each feed-forward layer has 12 encoder layers and 768 hidden units. The XLM-RoBERTa model consists of 24 layers and 1024 hidden units.

4.3. Experiments and Results

The results obtained on the test and validation sets are summarized in Table 4. As shown in the table, all models pretrained on Turkish except for ALBERT gave better results than the multilingual models, as expected. Among the Turkish BERT models, BERT_loodos consistently outperforms other models in both validation and test sets. This shows the positive impact of texts' domain in the pretraining phase of these large LM models.

In order to observe the effect of the training set on performance clearly, we trained the BERT models on (Tür et al., 2003) since it is the only available dataset and has enough instances to perform training. Scores are presented in Table 5. In the test data set, our results were calculated over the *PERSON*, *LOCATION*, and *ORGANIZATION* tags because only these three NE types were labeled in the (Tür et al., 2003). As expected, the models that were trained using our training set outperformed the models trained on (Tür et al., 2003). Even though (Tür et al., 2003) dataset is a larger one, it is comprised of properly written media articles, and the sources utilized were from the years 1997-1998, which are somewhat ancient. Among the BERT models trained with our data, BERT_loodos again achieved better scores than the other model across all metrics.

We also explored the performance of models for each named entity category. The scores are listed in Table 6. Not surprisingly, BERT_loodos outperformed in all classes except for *PRODUCT*. In this category, multilingual models achieved a better result. Since non-Turkish songs and foreign products are popular in dif-

[4]https://github.com/Loodos/turkish-language-models
[5]https://huggingface.co/bert-base-multilingual-cased
[6]https://huggingface.co/xlm-roberta-base
[7]https://oscar-corpus.com/
[8]https://opus.nlpl.eu/

Model	Recall		Precision		F1 Score	
	Val Set	Test Set	Val Set	Test Set	Val Set	Test Set
BERTurk	84.31	**85.02**	80.24	78.63	83.12	81.37
BERT_loodos	**84.99**	80.00	**83.56**	**84.49**	**84.27**	**82.18**
ALBERT_loodos	71.81	74.05	74.73	69.80	73.24	71.86
mBERT	78.95	76.61	74.15	73.41	76.48	74.98
XLM-RoBERTa	81.39	82.76	77.42	73.89	82.76	79.36

Table 4: Results of Transformer-based Models on Validation and Test Sets.

Model	Train Data	Recall		Precision		F1 Score	
		Val Set	Test Set	Val Set	Test Set	Val Set	Test Set
BERTurk	Our Train	86.84	86.90	84.53	80.04	85.67	83.33
BERTurk	(Tür et al., 2003)	68.87	69.01	69.17	70.87	69.02	69.92
BERT_loodos	Our Train	89.64	87.51	85.70	81.05	87.63	84.15
BERT_loodos	(Tür et al., 2003)	68.43	68.26	68.99	70.75	68.71	69.48

Table 5: Comparison of Formal and Informal Dataset on Person, Location, and Organization.

NE Class	BERTurk	BERT_loodos	ALBERT_loodos	mBERT	XLM-RoBERTa
PERSON	0.87	**0.88**	0.78	0.80	0.83
LOCATION	0.77	**0.81**	0.64	0.64	0.67
ORGANIZATION	0.77	**0.80**	0.72	0.72	0.75
TIME	0.89	**0.90**	0.83	0.86	0.88
PRODUCT	0.32	0.37	0.43	**0.52**	0.46
TV-SHOW	0.49	**0.57**	0.35	0.52	0.49
MONEY	0.88	**0.93**	0.88	0.75	0.85

Table 6: F1 Score on Test Set for Each NE

ferent languages and datasets (including our Turkish dataset), these multilingual models might be exposed to more of these entities during pre-training. Hence, they were more successful in this category.

5. Conclusion

In this paper, we introduced and made publicly available a new Twitter dataset for NER with high agreement scores in Turkish. Besides the common NE types, we also included new categories, *PRODUCT*, and *TV-SHOW*. We obtained initial scores with various transformer-based models on our validation and test sets. A BERT model pre-trained on a blend of formal and informal texts yielded the highest score. Besides, on the validation and test sets, we compared our training set with (Tür et al., 2003), which is the most studied data set in the literature. Models that used our training set during the fine-tuning phase achieved significantly higher scores than other models.

6. Bibliographical References

Aras, G., Makaroğlu, D., Demir, S., and Cakir, A. (2021). An evaluation of recent neural sequence tagging models in turkish named entity recognition. *Expert Systems with Applications*, 182:115049.

Çelikkaya, G., Torunoğlu, D., and Eryiğit, G. (2013). Named entity recognition on real data: a preliminary investigation for turkish. In *2013 7th International Conference on Application of Information and Communication Technologies*, pages 1–5. IEEE.

Demir, H. and Özgür, A. (2014). Improving named entity recognition for morphologically rich languages using word embeddings. In *2014 13th International Conference on Machine Learning and Applications*, pages 117–122. IEEE.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Grishman, R. and Sundheim, B. M. (1996). Message understanding conference-6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

Güneş, A. and Tantuğ, A. C. (2018). Turkish named entity recognition with deep learning. In *2018 26th Signal Processing and Communications Applications Conference (SIU)*, pages 1–4. IEEE.

Güngör, O., Güngör, T., and Üsküdarli, S. (2019). The effect of morphology in named entity recognition with sequence tagging. *Natural Language Engineering*, 25(1):147–169.

Küçük, D. and Can, F. (2019). A tweet dataset annotated for named entity recognition and stance detection. *arXiv preprint arXiv:1901.04787*.

Küçük, D. and Steinberger, R. (2014). Experiments to improve named entity recognition on turkish tweets.

arXiv preprint arXiv:1410.8668.

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

Küçük, D., Jacquet, G., and Steinberger, R. (2014). Named entity recognition on turkish tweets. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 450–454.

Küçük, D., Küçük, D., and Arıcı, N. (2016). A named entity recognition dataset for turkish. In *2016 24th Signal Processing and Communication Application Conference (SIU)*, pages 329–332. IEEE.

Kuru, O., Can, O. A., and Yuret, D. (2016). Charner: Character-level named entity recognition. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 911–921.

Okur, E., Demir, H., and Özgür, A. (2018). Named entity recognition on twitter for turkish using semi-supervised learning with word embeddings. *arXiv preprint arXiv:1810.08732*.

Sang, E. F. and Veenstra, J. (1999). Representing text chunks. *arXiv preprint cs/9907006*.

Şeker, G. A. and Eryiğit, G. (2012). Initial explorations on using crfs for turkish named entity recognition. In *Proceedings of COLING 2012*, pages 2459–2474.

Seker, G. A. and Eryigit, G. (2017). Extending a crf-based named entity recognition model for turkish well formed text and user generated content. *Semantic Web*, 8(5):625–642.

Tantug, B. E. A. C. (2015). Recognizing named entities in turkish tweets.

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

Tür, G., Hakkani-Tür, D., and Oflazer, K. (2003). A statistical information extraction system for turkish. *Natural Language Engineering*, 9(2):181–210.

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