NakbaTR: A Turkish NER Dataset for Nakba Narratives

Esma F. Bilgin Taşdemir
Medeniyet University
İstanbul, Turkey
esmabilgin.tasdemir@medeniyet.edu.tr

Şaziye Betül Özateş
Boğaziçi University
İstanbul, Turkey
saziye.ozates@bogazici.edu.tr

Abstract

The narratives of the ongoing Nakba are of significant importance for both the Palestinian people and the global community. The creation of language resources is crucial to automating the processing of written content related to this historical event. This paper introduces an annotated Named Entity Recognition (NER) dataset derived from a collection of 182 news articles about the Nakba and its witnesses. Given their prominence as a primary source of information on the Nakba in Turkish, news articles were selected as the primary data source. An initial analysis on the constructed dataset is also presented.

1 Introduction

The Nakba, which is translated as "catastrophe" in Arabic, is a term used for the mass displacement and dispossession of Palestinians starting from the 1948 Arab-Israeli war. During many Nakba events suffered by the Palestinian people, hundreds of thousands were forced to flee their homes, property, and belongings. The expulsion of native people to the degree of ethnic cleansing has several catastrophic results, including a refugee crisis and generational trauma that continues to this day. Today, there are again genocide and Nakba events against the Palestinian people, which have been escalating since the end of 2023.

The narratives about the Nakba events are important in many aspects, including cultural, legal, and historical significance. There are different types of sources where the narratives can be found. The news content from different outlets is one of the main sources of narratives in the Turkish language. In this work, we generated a manually annotated Named Entity Recognition (NER) dataset from websites of two news agencies. Both annotations and collected text can be used for several NLP tasks such as relation extraction, sentiment analysis, and topic modeling related to the Nakba event. Furthermore, it will serve as a new language resource for Turkish, a language often considered underrepresented in NLP research.

2 Related Work

Early studies on Named Entity Recognition (NER) began in the 1990s. Since then, numerous researchers have explored various aspects of NER tasks (Li et al., 2022; Yadav and Bethard, 2018). A NER dataset can be generated as a general purpose dataset or it can be a domain-specific dataset. Among different sources of text corpora, news texts are one of the popular sources as they are easy to collect especially in digital format. PERSON, LOCATION, and ORGANIZATION are the most common entity types in news NER datasets (Zhang and Xiao, 2024).

Most studies on NER in Turkish texts have focused on modern texts, which can be categorized as either formal or informal (noisy). Formal texts adhere to standard grammatical and orthographic rules, while informal texts may exhibit variations and deviations from these norms. In Akkaya and Can (2021), a transfer learning approach is used for NER in informal data with rarely-seen entity types. They add a CRF layer that is trained on both formal text and a noisy text to a BiLSTM-CRF model. They report 61.53% F1 score on noisy text. Safaya et al. (2022) experiment with several datasets and architectures with a goal of providing a benchmarking platform for various NLP tasks. They achieve 93.8% F1 score on the WikiANN dataset (Pan et al., 2017) using BiLSTM-CRF model and 93.07% F1 score with BERTurk-CRF model. Kılıç et al. (2020) report 82.31% F1 score on formal Turkish texts which is obtained through a multilingual cased BERT model. Finally, Ozcelik and Toraman (2022) present results for several models evaluated on some public datasets. They report, an

Proceedings of the Proceedings of the 1st International Workshop on Nakba Narratives as Language Resources, pages 122–126

F1 score of 96.10% achieved by ELECTRA-tr on a dataset of news articles and 92.26% F1 score obtained by ConvBERTurk on the WikiANN dataset.

3 Methodology

3.1 Dataset

We created NakbaTR, a new and domain-specific dataset, using news texts published online. The content is limited to news containing testimonies from witnesses of the Nakba currently taking place. This section gives details on data collection and annotation processes we employed for creating the NakbaTR dataset.

3.1.1 Data Collection

Testimonies in Turkish related to the ongoing Nakba can be found mostly in news sources. The NakbaTR dataset was curated from websites of two state-owned news outlets; Anadolu Ajansı (AA)[1] and TRTHaber.[2]

Texts are scraped semi-automatically using a pipeline of searching, downloading and cleaning. We decided on two Turkish phrases; "anlattı" (told) and "konuştu" (spoke) which are used as searching keywords. We made searches for each keyword in each outlet, four searches in total. Web pages containing keywords along with "Gazze" (Gaza) are downloaded automatically. We downloaded 120 pages per website in this manner and 480 pages are downloaded in total. Pages that include video or photographic content are removed from the collection resulting a set of 369 pages (215 from TRTHaber and 181 from AA). The collected text are cleaned from irrelevant elements first automatically by exploiting the web page structure followed by manual cleaning of any remaining noise. The items of the dataset are comprised of text of the news, its URL and date it was published, and the source. Both sources are news agencies, so they use a similar, formal language.

We aimed to generate a collection of news in which expressions of Nakba witnesses are contained. We employed a rigorous manual filtering process based on existence of testimonies within the content. As a result, news texts that do not contain a witness testimony are removed from the collection. The final collection contains 182 news articles (74 news articles from TRTHaber and 107 news articles from AA). The testimonial expressions can be given with both direct and indirect speech. It should be noted that content other than the testimony part which gives contextual information regarding the testimony are kept in the dataset as well.

3.1.2 NER Dataset Generation

NakbaTR dataset is annotated with PERSON, LOCATION, and ORGANIZATION types while words of other type of tags are all marked with O. We use the following definitions of tags:

- PERSON : People, including fictional.
- LOCATION: GPE and Non-GPE locations including countries, cities, states, mountain ranges and bodies of water.
- ORGANIZATION: Collectives such as companies, political groups, government bodies, and public organizations.

Annotation of NakbaTR is done in two steps. A BERT-based language model, which is detailed in Section 3.2 is employed for automatic annotation in the first step. In the second step, its output is corrected and verified by two human annotators. The data is prepared in CoNLL-2003 format with multiple word entities marked with B- and I- prefixes. Figure 1 depicts two example annotated sentences from the dataset. To assess the reliability of the annotations, we calculated the inter-annotator agreement between human annotators on 100 randomly selected sentences which contain 295 named entities in total, achieving a Cohen's Kappa score of 0.968, indicating a high level of consistency in annotations.

The resulting NakbaTR NER dataset contains 3,957 sentences, 2,289 PERSON, 5,875 LOCATION, and 1,299 ORGANIZATION tags in total. Details regarding the sources are given in Table 1. The NakbaTR dataset can be accessed at https://github.com/sb-b/NakbaTR.

3.2 BERT-based Language Model

Various pre-trained Language Models (PLMs) were previously utilized for the NER task on Modern Turkish (see Section 2). We observed that among different architectures and models, BERTurk (Schweter, 2020), a Turkish language model utilizing the BERT architecture and pretrained on Turkish text, reached the highest F1

[1] http://aa.com.tr
[2] http://trthaber.com

Source	Number of News	Number of Sentences	Number of Tokens	Number of Person	Number of Location	Number of Organization
AA	107	2,482	70,188	1,457	3,878	893
TRTHaber	74	1,550	41,639	832	1,997	406
TOTAL	181	4,032	111,827	2,289	5,875	1,299

Table 1: Dataset statistics.

```
#doc_id = https://www.trthaber.com/haber/dunya/gazzede-israilin-saldirilarinda
          -yaralanan-filistinli-cocuklar-cektikleri-aciyi-anlatti-826779.html
#metadata = 06.01.2024 12:29
#sent_id = 375
#text = Gazze'de, İsrail'in saldırılarında yaralanan Filistinli çocuklar
        çektikleri acıyı anlattı
Gazze'de B-LOC
, O
İsrail'in B-LOC
saldırılarında O
yaralanan O
Filistinli O
çocuklar O
çektikleri O
acıyı O
anlattı O

#sent_id = 376
#text = Gazze Şeridi'nin güneyine yönelik İsrail saldırılarında yaralanan
        Filistinli 2 kız çocuğunun yaşadıkları, gözlerinin önünden gitmiyor.
Gazze B-LOC
Şeridi'nin I-LOC
güneyine O
yönelik O
İsrail B-LOC
saldırılarında O
yaralanan O
Filistinli O
2 O
kız O
çocuğunun O
yaşadıkları O
, O
gözlerinin O
önünden O
gitmiyor O
. O
```

Figure 1: Annotations of the first two sentences of a news document in the NakbaTR dataset. Sentence translations: *(First sent.)* "In Gaza, Palestinian children injured in Israeli attacks described their suffering." *(Second sent.)* "The experiences of two Palestinian girls injured in Israeli attacks on the southern Gaza Strip remain vivid in their minds."

score for the NER task on the Turkish split of the WikiANN NER dataset. Hence, we opted to utilize BERTurk that is fine-tuned on a large Turkish NER dataset (Tür et al., 2003) as the Turkish NER model in the automatic annotation process.

The utilized BERTurk model has 12 transformer layers each consisting of 12 attention heads. The number of hidden units is 768. A total of 110 million parameters are fine-tuned during the pre-training phase on a large corpus of Turkish text data, allowing the model to learn contextual representations that capture intricate syntactic and semantic relationships within the language.

In the automatic annotation phase, the BERT-based model achieves an impressive F1 score of 87% in predicting named entities. Its performance on individual entity types is as follows: 89% for LOCATION entities, 76% for ORGANIZATION entities, and 90% for PERSON entities. The precision score of the model on all entity types is 100%, indicating that whenever the model predicts a named entity, it is always correct. However, the recall scores are 81% for LOCATION entities, 61% for ORGANIZATION entities, and 81% for PERSON entities. The low recall indicates that there are named entities that the model cannot detect.

It is important to note that these model performance metrics were derived from aligned sentences between the output of the BERTurk NER model and the manual annotation step. Segmentation and tokenization errors that occurred while preparing the news data in CoNLL-2003 format propagated throughout the dataset, necessitating significant effort during the manual correction phase.

4 Dataset Analysis

We conducted some basic analyses regarding the annotated named entities on the NakbaTR dataset. We plotted the mention frequencies of frequently occurring location names over time. Figure 2 illustrates these plots for the TRTHaber and AA sections of the dataset, respectively. Note that, we excluded the locations *Gazze (Gaza), Filistin (Palestine)*, and *İsrail (Israel)* from this plot since the mention frequency of these location names are much more higher than any other location in the dataset. The figure is helpful in understanding the changing focus of the news. For instance, the coverage of the intensive assault of Israeli army on Northern Gaza civilian areas like Jabaliya and Beit Lahia is clearly traceable from both plots. Although there are some common patterns between the two news sources, they have different coverage rates. This can be attributed to the difference in their focus, target audience, and perspective since AA provides a broader, more international one, while TRTHaber has more local perspective.

We also conducted an analysis of the co-occurrence patterns of named entities within the dataset to explore relationships between entities. Specifically, we counted the frequency of each named entity pair appearing in the same sentence throughout the dataset. To enhance the clarity of

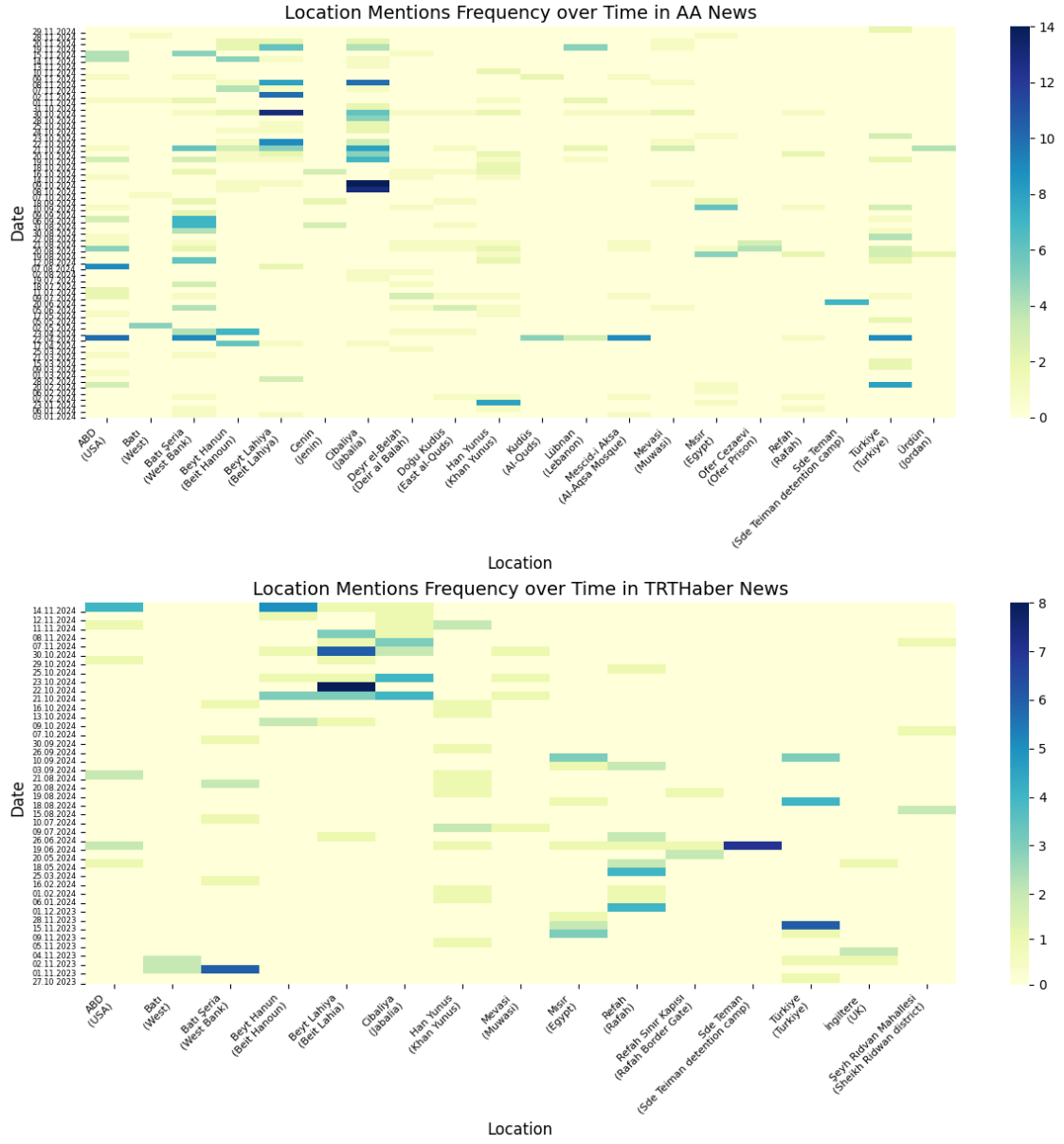

Figure 2: Mention frequency plots of frequently occuring location names in the AA and TRTHaber sections of the dataset over time.

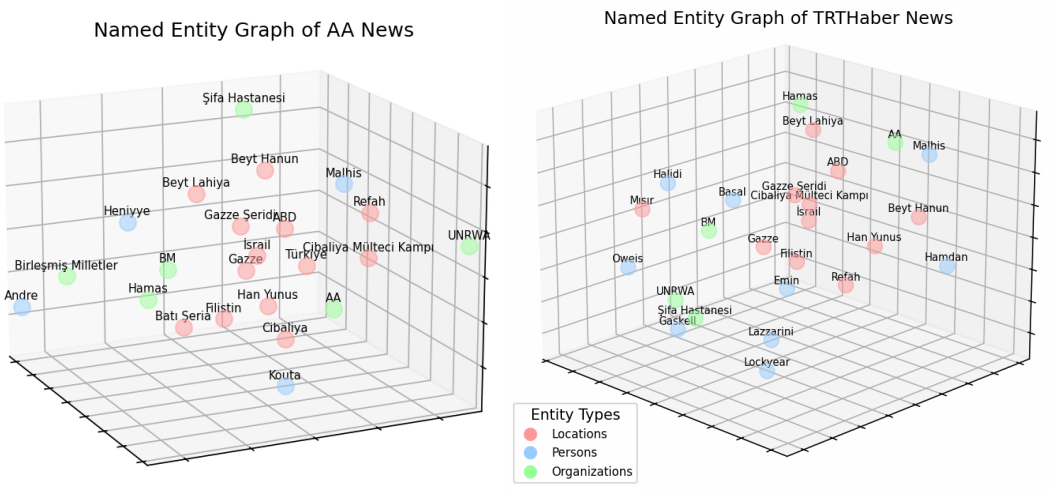

Figure 3: Co-occurence graph of named entities in AA and TRTHaber sections of the dataset.

the visualization, we applied a filtering step to exclude entity pairs with a co-occurrence frequency lower than 10 for the TRTHaber section and a co-occurrence frequency lower than 18 for the AA section. The resulting co-occurrence graph provides a visual representation of frequently co-occurring entities, highlighting key connections within the data. Figure 3 illustrates co-occurrence graphs, showing the relations between named entities in the AA and TRTHaber news in the NakbaTR dataset. In the figure, different entity types are represented in different colors. Although only the entity pairs that are most frequently observed in the dataset are included for clarity, this co-occurrence graph provides insights into the relationships among the entities in the narratives. For instance, the organization entity `UNRWA` and the person entity `Lazzarini` are located near in the TRTHaber graph which makes sense as Philippe Lazzarini is the commissioner-general of the UNRWA organization. Other visible relationships are between the person entity `Oweis` (Saleem Oweis, communications specialist at UNICEF Middle East and North Africa) and the organization entity `BM` (United Nations) in the TRTHaber graph and districts in the north Gaza located together in the AA graph. Some real life entities can have multiple names as in the case with `BM` (UN) and `Birleşmiş Milletler` (United Nations). This situation is clearly visible in the AA graph in the figure. Unifying multiple names for a single entity will therefore be beneficial before extracting relationships between entities.

5 Conclusion

In this work, we introduced a novel, manually annotated, Turkish NER dataset. The dataset comprises 3,957 news sentences collected from the websites of two prominent news agencies. We applied a filtering process to make sure that only the news which contain witness testimonies regarding the ongoing Nakba are included in the dataset. After a semi-automatic annotation for entities of types Person, Location, and Organization, we obtained a NER dataset of 2,289 PERSON, 5,875 LOCATION, and 1,299 ORGANIZATION tags. The dataset can be extended to be useful in several NLP tasks such as relation extraction or sentiment analysis for the Nakba event while providing a new language resource for Turkish. As future work, we aim to improve the dataset by increasing the number of news and entity types.

References

Emre Kagan Akkaya and Burcu Can. 2021. Transfer learning for Turkish named entity recognition on noisy text. *Nat. Lang. Eng.*, 27(1):35–64.

Yüksel Pelin Kılıç, Duygu Dinç, and Pınar Karagöz. 2020. Named entity recognition on morphologically rich language: exploring the performance of bert with varying training levels. IEEE International Conference on Big Data (2020).

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2022. A survey on deep learning for named entity recognition. *IEEE Trans. Knowl. Data Eng.*, 34(1):50–70.

Oguzhan Ozcelik and Cagri Toraman. 2022. Named entity recognition in Turkish: A comparative study with detailed error analysis. *Inf. Process. Manag.*, 59(6):103065.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

Ali Safaya, Emirhan Kurtuluş, Arda Goktogan, and Deniz Yuret. 2022. Mukayese: Turkish NLP strikes back. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 846–863, Dublin, Ireland. Association for Computational Linguistics.

Stefan Schweter. 2020. BERTurk - BERT models for Turkish. `https://zenodo.org/records/3770924`. [Online; accessed 21-10-2023].

Gökhan Tür, Dilek Hakkani-Tür, and Kemal Oflazer. 2003. A statistical information extraction system for Turkish. *Natural Language Engineering*, 9(2):181–210.

Vikas Yadav and Steven Bethard. 2018. A survey on recent advances in named entity recognition from deep learning models. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 2145–2158. Association for Computational Linguistics.

Ying Zhang and Gang Xiao. 2024. Named entity recognition datasets: A classification framework. *Int. J. Comput. Intell. Syst.*, 17(1):71.