TR-SEQ: Named Entity Recognition Dataset for Turkish Search Engine Queries

Berkay Topcu
AI Industry Solutions
Turkcell Technology
İstanbul, Turkey
berkay.topcu@turkcell.com.tr

İlknur Durgar El-Kahlout
NLP Technologies
Turkcell Technology
İstanbul, Turkey
ilknur.durgar@turkcell.com.tr

Abstract

Recognizing named entities in short search engine queries is a difficult task due to their weaker contextual information compared to long sentences. Standard named entity recognition (NER) systems that are trained on grammatically correct and long sentences fail to perform well on such queries. In this study, we share our efforts towards creating a cleaned and labeled dataset of real Turkish search engine queries (TR-SEQ) and introduce an extended label set to satisfy the search engine needs. A NER system is trained by applying the state-of-the-art deep learning model BERT to the collected data and its high performance on search engine queries is reported. Moreover, we compare our results with the state-of-the-art Turkish NER systems.

1 Introduction

Named entity recognition (NER) aims to identify the predetermined entity categories (person, organization, location, etc.) accurately in a given text. NER studies which is one of the most widely studied subjects in the field of natural language processing (NLP), has been started for English and spread over a range of languages including Turkish (Yeniterzi, 2011). Initial NER studies for Turkish used small labeled datasets that contain grammatically correct and typo-free sentences. Since manual labeling is a time-consuming and a costly process, gathering larger datasets and manually labeling them has not been feasible. With the increase of internet use, correctly classifying named entities has become a crucial need especially for social media entries and search engine (SE) queries. In search engines, identifying entities such as famous persons, places or dates in user queries is crucial in order to accurately determine the information to be displayed to the user and response with the right information. For example, if the user is searching for a famous person, SE should return the Wikipedia information about the person. Moreover if this famous person is an artist, the identified named entity assists in conveying the details of his/her movie/album/artwork details. Similarly, it is important to determine the location in the query for which we want to know the weather.

The state-of-the-art Turkish NER tools fail to perform well on search engine queries due to two main reasons. The first reason is that most of the NER tools that have been developed for Turkish so far generally carry out categorization studies with three classes (**Person**, **Organization** and **Location**). Naturally, a three-class model cannot detect all the named entities in a SE query that is required to create a successful SE response. For this reason, it has become a necessity to determine the entity types specific to the SEs and to make a special development for these types. The second reason is that NER tools are dependent on the texts and domains they are trained on. NER tools, which are usually trained in canonical and grammatically correct long sentences, do not succeed in short SE queries that contain abbreviations and spelling errors and mostly consist of two or three words. As the short inputs have poor contextual knowledge compared to long sentences and often include misspellings, existing NER tools are observed to perform poorly on these type of texts and the success rates are far below the desired levels.

Although the first attempts for Turkish NER systems are dominated by rule-based systems (Küçük and Yazıcı, 2009), CRF-based models (Yeniterzi, 2011; Seker and Eryiğit, 2012; Yeniterzi et al., 2018) came to the fore in the forthcoming studies. In all of these studies, a relatively small dataset consisting of about $30K$ sentences, clean in terms of spelling and grammar, is used (Tur et al., 2003) and models are based on three classes (person, location, organization). Recently, models using deep learning methods (Demir and Ozgur, 2014; Aras et al., 2021), especially Bidirectional Long Short-Term

Proceedings of Recent Advances in Natural Language Processing, pages 1417–1422
Sep 1–3, 2021.
https://doi.org/10.26615/978-954-452-072-4_158

Memory (BLSTM), became more popular (Güneş and Tantuğ, 2018).

Even though the accuracies of these NER studies are high in their test sets, the performance drop is dramatical on social media (e.g., Twitter) and SE queries (Çelikkaya et al., 2013; Küçük and Steinberger, 2014). For this reason, studies are carried out to collect dedicated data for social media texts (Okur et al., 2016). The most recent studies in Turkish NER field utilize Transformers and BERT based models (Yıldırım, 2019; Akdemir and Güngör, 2019; Aras et al., 2021).

In this paper, we focus on the recognition of a selected group of named entities for Turkish search engine queries. In the study, Yaani[1] search engine entries are collected, the categories for the SE needs are determined and a labeling study is carried out using an in-house annotation tool. The data consisting of a total of $100K$ SE queries is labeled by more than one person and a verification study is carried out with the obtained outputs. In addition, a high-performance NER tool is developed by training the Transformers-based BERT (Devlin et al., 2018) model, which is one of the most successful approaches and widely used in current NLP studies, with our labeled dataset, and an accuracy of 90.41% is achieved.[2]

2 Dataset Creation Process

For data-driven NER systems to perform well on real world scenarios, it is of crucial importance to have training data that has a similar distribution with the real world data. In order to satisfy this constraint, we carried out an extensive data collection and labeling process for a NER system specific to SE queries. We collected a dataset of $100K$ queries that are submitted to Yaani search engine for a specific period of time without having a constraint related to short/long queries. Next, we cleaned the malicious content and normalized the spelling errors manually. Finally, we removed the duplicates which resulted in a dataset with **97,428** queries.

2.1 Determining Entity Classes

Active usage of widgets and snippets is a particular way of enhancing user experience in search engines. Currently, most of the well-known search engines respond to user queries via widgets and snippets which convey information requested by the user in a concise and compact manner. The main motivation behind this study is the extraction of relevant information obtained from user queries and activate related widgets and snippets accordingly. At the time of this study, recognition of named entities specific to particular widgets such as weather, currency converter, maps and info boxes (i.e., Wikipedia results) have been focused.

Due to the inadequacy of the standard three-class model for a SE specific NER model; we examined 48 different entity categories provided by Shrinked TWNERTC Turkish NER Data [3] (Şahin et al., 2017) and selected the following seven main categories to cover different agents such as weather, maps, and currency inquiries. These classes are as follows;

- **Person:** All type of proper names for persons **Ex:** *Mustafa Kemal Atatürk, Prof. Dr. Ali Ünal*
- **Organization;** All type of entities that have an organization or that are organized such as government institutions, firms, hospitals, universities, soccer clubs, festivals **Ex:** *Walter Reed Askeri Hastanesi (Walter Reed Military Hospital), İstanbul Üniversitesi (İstanbul University)*
- **Location:** All places that have a physical position like countries, cities, villages, etc. **Ex:** *İstanbul, Van Gölü (Van Lake), Hollywood*
- **Date:** All date/time entries plus important dates like "mother's day" etc. **Ex:** *15.02.2020, 15 Ekim (November, 15th), M.Ö 500 (500 B.C.), 1870*
- **Measure:** All entities that are in the global standard, the International System of Units (SI), such as, kg, lt, etc. **Ex:** *1700 metrekare (square meters), 2 lt (liters), 300 metre (meters)*
- **Currency:** All currency entities plus commodities such as gold, silver and cryptocurrency **Ex:** *19 milyon TL (milion Turkish Liras), 20 bitcoin, 10 gram altın (gram of gold)*
- **Production_art_music:** All entities that are produced like films, series, songs, books, etc.

[1] yaani.com.tr

[2] The data is freely available for academic purposes. Please contact the authors for dataset acquisition.

[3] https://www.kaggle.com/behcetsenturk/shrinked-twnertc-turkish-ner-data-by-kuzgunlar

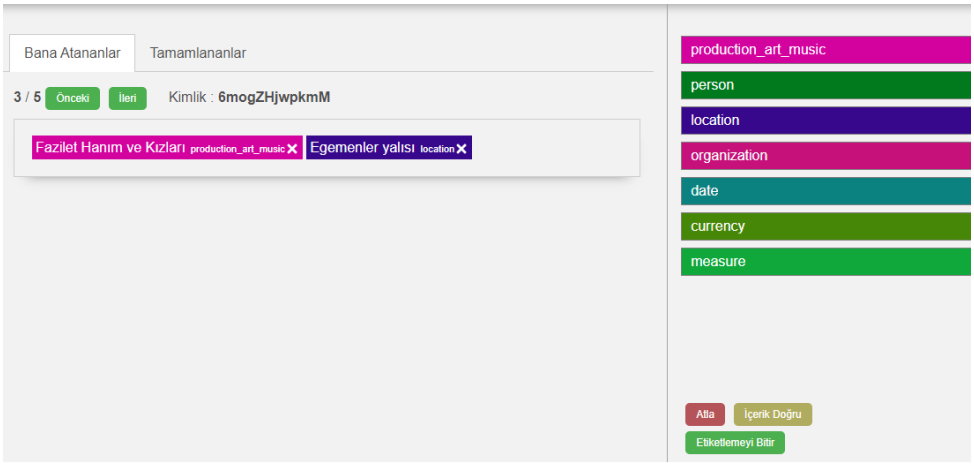

Figure 1: *Labeling for "Fazilet Hanım ve Kızları Egemenler Yalısı"*

2.2 Labeling Process

Following the category determination, we organized a team of twenty people for data labeling. We shared a labeling guide in which we itemized the tagging criteria and provided positive and negative examples. We also asked annotators to complete a demo task before initializing the main project. By the help of the demo outputs, we revised the labeling guide and finalized the rule set as follows:

- A NE should be labelled to include the longest phrase that can be retrieved. **Ex:** *Bursa Büyükşehir Belediyesi (Bursa Metropolitan Municipality)* should be labeled **Organization** as a whole instead of labeling only *Bursa* as **Location**.

- Successive NEs of the same type should be labeled individually if separated by "," or "and". **Ex:** *Washington DC White House* should be labelled seperately in *Washington DC'deki Beyaz Saray'a nasıl giderim (How can I go to the White House in Washington DC)*, instead of selecting it as a whole phrase such as *White House in Washington DC*.

- While labeling person NEs, if the title phrase does not have another NE, the whole title will be labeled as **Person**. **Ex:** *İçişleri Bakanı (Minister of Interior) Soylu* will be labeled as a whole as **Person** but as the query *Bursa Büyükşehir Belediyesi Başkanı (Mayor of Bursa Metropolitan Municipality) Alinur Aktaş* has an **Organization** NE in the phrase *Bursa Büyükşehir Belediyesi (Bursa Metropolitan Municipality)*, labeller will seperately label this phrase and only label *Alinur Aktaş* as **Person**.

- General names following a proper name should not be included in the NE except the location names including "Lake", "Street" and "Univerity" etc. **Ex:** *İstanbul* should be labeled as **Location** in the query *İstanbul ilinde hava durumu* (*Weather forecast in İstanbul city*), on the other side, location names such as *Van Gölü (Van Lake), İstiklal Caddesi (Istiklal Street)* should be labeled with the following location name.

- Due to the Turkish morphology, an apostrophe is attached at the end of the proper name if it is followed by a suffix. Such words with apostrophes should be labeled as a whole. **Ex:** *Yunanistan'ın (Greece's)* should be labeled instead of *Yunanistan (Greece)* in queries.

- Subordinates consisting more than two or more words should be labeled as a whole.

During the whole process, we employed an in-house labeling and verification tool. Figure 1 shows an example annotation screen used in the study. In the light of the given instructions, the annotator labeled the phrase *Fazilet Hanım ve Kızları* (a Turkish serie) as **Production_art_music** and *Egemenler Yalısı* (a location in this serie) as **Location**.

2.3 Consolidation Step: Annotator Disagreement

The first three weeks of the annotation process was dedicated to the labeling of the queries and the final week was used to check the mismatches between

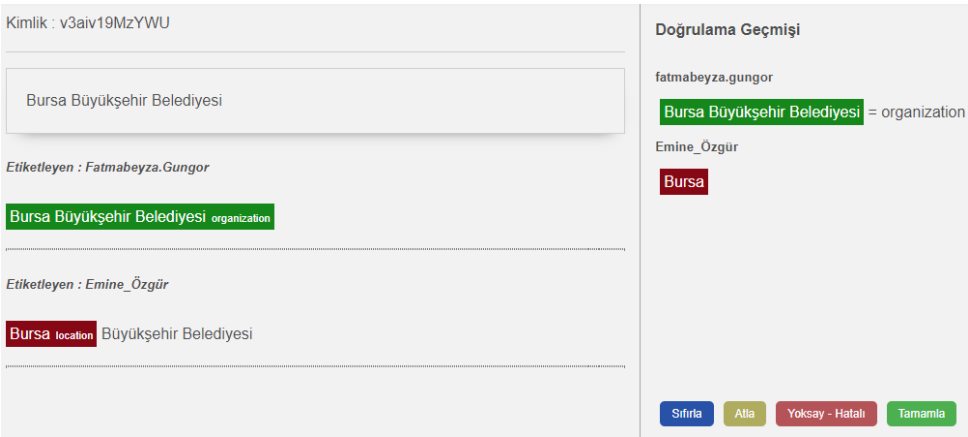

Figure 2: Consolidation example: *"Bursa Büyükşehir Belediyesi"*

annotators and consolidate the output. Almost half of the queries (**45,442**) are labeled by two different annotators. These queries are categorized into three groups; i) queries that both annotators agree on, ii) queries that two annotators do no agree on, and iii) queries that one of the annotators labels some words and the other annotator does not label any of the words.

Leaving out the queries that both annotators agree on, **10,800** queries are considered in the consolidation step in which four experts (different from the annotators) checked the mismatching labels. If any of the labels is correct, it is selected as the final label. If both labels are wrong, the true label is determined by the expert. Figure 2 illustrates the consolidation screen used in order to analyze the mismatch between the annotations of a query. Regarding the rule that the longest phrase should be retrieved, the label from the first annotator is kept and the label from the second annotator is discarded.

2.4 Dataset Statistics

The collected dataset consists of **97,428** search engine queries and includes **294,100** words. Figure 3 illustrates the distribution of query lengths. The dataset mostly contains short word sequences of length between one and four. The average length of the queries is **3.018**. The longest query has **35** words whereas the shortest one is a single word.

Table 1 provides the number of different named entity types used in this study and the percentage of multi-word entities for each type. The most frequent named entity type is **Organization**. **Person**, **Production_art_music** and **Location** are three other named entity types that occur too often. **Measure** is the rarest type whereas **Date** and **Currency** are not observed very frequently.

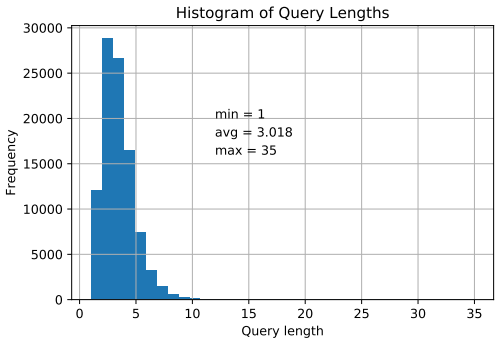

Figure 3: Distribution of search engine query lengths

NER Type	Count	Multi-word
Organization	27,633	%63.02
Person	18,728	%76.83
Production_art_music	13,024	%89.99
Location	10,580	%18.23
Date	3,450	%30.87
Currency	1,466	%45.77
Measure	468	%51.50
Other	186,759	—

Table 1: Distribution of named entity types

3 Experiments and Results

Using this new dataset, we trained a Turkish NER system and demonstrated the improved NER performance of using short search queries during training. Our Turkish NER system is trained by fine-tuning the BERTurk (Schweter, 2020) model using our dataset. BERTurk is a community driven BERT (Bidirectional Encoder Representations for Trans-

Model	Accuracy	Precision	Recall	F-measure
TC-BERT	0.9041	0.7302	0.8131	0.7695
TC-ELECTRA	0.8982	0.7056	0.8077	0.7532

Table 2: Seven class NER performance

Model	Accuracy	Precision	Recall	F-measure
Turkish-Bert-NLP (Yıldırım, 2019)	0.5055	0.2007	0.5552	0.2948
Char-BiLSTM-CRF (Aras et al., 2021)	0.8649	0.4968	0.3951	0.4402
BERTurk-CRF (Aras et al., 2021)	0.8804	0.5844	0.4843	0.5296
TC-BERT	**0.9457**	**0.7491**	**0.8561**	**0.7991**

Table 3: Comparision of three class NER performances

formers) (Devlin et al., 2018) model for Turkish. Its training corpus has a size of 35GB and 44B tokens with a vocabulary size of 128K. For fine-tuning the BERTurk model for NE classification, **max_seq_length** is set to 128, **train_batch_size** and **eval_batch_size** are set to 32, **learning_rate** is set to $2e^{-5}$ and the model is trained for 5 epochs.

In order to evaluate the performance of the NER system, the dataset is split into three: train, validation and test. For the test set, 1% of the data is selected randomly from the whole dataset by preserving the distribution of named entity types. Similarly, 1% is randomly selected as validation data and the remaining 98% is used for training. Using 1% of the whole dataset for testing purposes resulted about 1000 test queries and all seven classes are included proportionally to their distribution in the dataset.

In addition to the BERT model, we have also experimented with a pre-trained version of ELECTRA (Clark et al., 2020) model (also publicly available within the BERTurk study), which allowed us to compare the performance of the collected dataset on different deep architectures. Table 2 illustrates the performance of the NER systems (TC-BERT and TC-ELECTRA) fine-tuned using our dataset.

Publicly available Turkish NER models and other studies in the literature usually take three named entity types (**Person**, **Organization** and **Location**) into account. The effort towards increasing the number of named entity types is very limited for Turkish and to the best of our knowledge there is no other study that focuses on the named entity types discussed in this paper. Therefore, comparing the performances of our system with previous studies on the same test set is not immediately clear.

For a fair comparison, we removed the remaining labels from our test set by keeping three base named entity types (**Person**, **Organization** and **Location**). Since our model is based on seven named entity types, we filtered out the output labels that belong to other types (by masking as **Other**) and generated the final results. We compare our results with two recent Turkish NER studies: i) the NER model of "Turkish-Bert-NLP-Pipeline" (Yıldırım, 2019) that is known for its high performance on well-formed Turkish texts and ii) recent neural sequence tagging models discussed in (Aras et al., 2021). As depicted in Table 3, the NER system trained using our dataset of short queries outperforms the systems trained using grammarly correct Turkish sentences, both in terms of accuracy and F-measure.

4 Conclusion

This study presents a new Turkish NER dataset which is created specifically for short word sequences that comprises a large portion of the search engine queries. Our aim is to improve the NER performance of the models that are trained using grammatically correct sentences and fail to perform well on search engine queries. The dataset is created by cleaning and labeling $100K$ search queries and provides a rich resource for Turkish NER studies. In addition, our transformer based NER system presents useful baseline accuracies for future studies.

Our next step is to train a BERT-based language model from scratch by using search engine queries and replace the pre-trained BERT model provided by BERTurk. This will allow us to develop a NER system specific to short queries. Furthermore, creating a knowledge graph by utilizing the named entity labels will be an important step in enhancing

overall search engine performance.

References

Arda Akdemir and Tunga Güngör. 2019. Joint learning of named entity recognition and dependency parsing using separate datasets. *Computación y Sistemas*, 23(3).

Gizem Aras, Didem Makaroğlu, Seniz Demir, and Altan Cakir. 2021. An evaluation of recent neural sequence tagging models in turkish named entity recognition. *Expert Systems with Applications*, 182:115049.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Hakan Demir and Arzucan Ozgur. 2014. Improving named entity recognition for morphologically rich languages using word embeddings. In *Proceedings - 2014 13th International Conference on Machine Learning and Applications, ICMLA 2014*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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

Dilek Küçük and Adnan Yazıcı. 2009. Named entity recognition experiments on turkish texts. In *Flexible Query Answering Systems*, pages 524–535, Berlin, Heidelberg. Springer Berlin Heidelberg.

Dilek Küçük and Ralf Steinberger. 2014. Experiments to improve named entity recognition on Turkish tweets. In *Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM)*, pages 71–78, Gothenburg, Sweden. Association for Computational Linguistics.

Eda Okur, Hakan Demir, and Arzucan Özgür. 2016. Named entity recognition on Twitter for Turkish using semi-supervised learning with word embeddings. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 549–555, Portorož, Slovenia. European Language Resources Association (ELRA).

Stefan Schweter. 2020. BERTurk - BERT models for Turkish. https://github.com/stefan-it/turkish-bert.

Gökhan Akın Seker and Gülşen Eryiğit. 2012. Initial explorations on using CRFs for Turkish named entity recognition. In *Proceedings of COLING 2012*, pages 2459–2474.

Gokhan Tur, Dilek Hakkani-Tur, and Kemal Oflazer. 2003. A statistical information extraction system for Turkish. *Natural Language Engineering*, 9:181 – 210.

Reyyan Yeniterzi. 2011. Exploiting morphology in Turkish named entity recognition system. In *Proceedings of the ACL 2011 Student Session*, pages 105–110, Portland, OR, USA. Association for Computational Linguistics.

Reyyan Yeniterzi, Gökhan Tür, and Kemal Oflazer. 2018. *Turkish Natural Language Processing*, chapter Turkish Named-Entity Recognition. Springer International Publishing.

Savaş Yıldırım. 2019. Turkish-bert-nlp-pipeline. https://github.com/savasy/Turkish-Bert-NLP-Pipeline.

G. Çelikkaya, D. Torunoğlu, and G. Eryiğit. 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.

Bahadır Şahin, Mustafa Tolga Eren, Cağlar Tırkaz, Ozan Sönmez, and Eray Yıldız. 2017. English/Turkish Wikipedia named-entity recognition and text categorization dataset. doi:10.17632/cdcztymf4k.1.