

RuznamceNER: A Named Entity Recognition Dataset for Ottoman Turkish

Esmâ Fatıma Bilgin Taşdemir, Dilara Zeynep Gürer, Şaziye Betül Özates

Istanbul Medeniyet University, Boğaziçi University, Boğaziçi University
Kadıköy, İstanbul, 34720, Turkey, Bebek, İstanbul, 34342, Turkey, Bebek, İstanbul, 34342, Turkey
esmabilgin.tasdemir@medeniyet.edu.tr, dilara.gurer@std.bogazici.edu.tr, saziye.ozates@bogazici.edu.tr

Abstract

Named Entity Recognition (NER) in historical texts poses distinct challenges. Language change reflected in spelling variations, archaic vocabulary, and inconsistent orthography diminish the efficacy of models trained on contemporary corpora. The limited availability of annotated historical datasets constrains the development and evaluation of accurate, domain-specific NER systems, underscoring the necessity for specialized approaches and domain adaptation. In this work, we introduce Ottoman Turkish *ruzmançe* registers as a valuable digital historical resource with broad potential for diverse NLP applications. Our primary contribution is RuznamceNER, a manually annotated NER dataset derived from *ruzmançe* documents spanning two centuries. The dataset contains 2,138 sentences and a total of 8,730 annotated entities of types PERSON, LOCATION, and ORGANIZATION. We further report evaluation results using a BERT-CRF baseline model pre-trained with modern Turkish, highlighting the pivotal importance of in-domain training data for effective NER in historical contexts. Experimental results on the RuznamceNER test set under various training configurations show that even a small amount of supervised in-domain data can yield robust performance for well-structured texts, despite significant lexical and orthographic differences between historical and modern language forms.

Keywords: named entity recognition, historical Turkish, dataset annotation, transfer learning, fine-tuning

1. Introduction

Named Entity Recognition (NER) constitutes a fundamental component of most Natural Language Processing (NLP) systems. Its primary function is the identification and classification of named entities, such as individuals, organizations, and temporal expressions. NER plays a pivotal role in historical research, digital humanities, and archival exploration. It helps historians and researchers in efficiently analyzing extensive textual corpora by automatically detecting and categorizing significant entities. In addition, NER systems enable researchers to extract meaningful insights and relationships that might otherwise be challenging or time-consuming to uncover through manual examination.

Yet, NER in historical texts presents several unique challenges that distinguish it from contemporary text processing (Ehrmann et al., 2024). One major difficulty arises from the evolution of languages over time; spelling variations and obsolete vocabulary commonly found in historical documents can hinder the effectiveness of modern NLP models trained in contemporary corpora (Blouin et al., 2024). Additionally, historical texts often lack standardized formats, punctuation, or consistent grammatical structures, further complicating entity recognition. The scarcity of annotated historical datasets also limits the development and evaluation of accurate NER systems tailored to these texts. Finally, historical documents may contain ambiguous references or culturally specific terms

that require contextual knowledge for correct interpretation. These factors collectively contribute to lower rates of accuracy and robustness in applying standard NER techniques to historical corpora, necessitating specialized approaches and domain-specific adaptations (Sprugnoli, 2018; Todorov and Colavizza, 2020).

Kazasker Ruznamçesi (*Chronicals of Chief Judges*), which will be referred to as *ruzmançe* in this work, is a special type of historical documents that records the appointments of judges (*kadı*) and scholars (*müderri*) in the Ottoman Empire. The records are written in an archaic version of Turkish and they contain information about the appointment, dismissal, and promotion of judges and scholars, along with certain details of these appointments (Gündoğdu, 2009). The significance of these records lies in their value as a crucial source of information on the judicial system (*kaza sistemi*) and educational system (*medrese sistemi*) of the period. Through the examination and analysis of these records, it is possible to gain new insights into the organization and functioning of these institutional structures as well as the historiography of the Ottoman Empire. Existing works on *ruzmançe* documents are created almost always within the humanities domain, and they are restricted in the number of documents they cover (Kuru, 2020), since they resort to manual processing of the registers, which are available as digitized images of the original documents. There are hardly any studies handling *ruzmançe*

texts from the NLP and computational linguistic perspective, mainly due to the lack of necessary digital resources.

In this work, we introduce the ruznamçe registers as a valuable digital historical resource with strong potential for a wide range of NLP tasks. Our main contribution is a manually annotated named entity recognition (NER) dataset, *RuznamceNER*, constructed from ruznamçe documents spanning a period of two centuries. We further present evaluation results obtained using a strong baseline model, demonstrating the crucial role of in-domain training data for effective NER on ruznamçe texts. Both the dataset and the trained model are publicly available at <https://huggingface.co/datasets/BUCOLIN/RuznamceNER>. These resources constitute an important step toward enabling large-scale named entity recognition and automated processing of the extensive ruznamçe registers, as well as other historical Turkish texts that remain unprocessed.

2. Related Work

The NLP of historical documents has been receiving increased attention from the research community due to the massive digitization campaigns at a global scale. While much of the historical NER research is mainly confined to the Latin alphabet-based languages (Blouin et al., 2024), (Ehrmann et al., 2024), the number of studies on other languages is increasing steadily (Shaalan, 2014; Patil et al., 2016).

Official and legal documents constitute a sub-domain of historical documents where relations between people and institutions are recorded in structured or semi structured styles. Orasmaa et al. (2022) present historical corpus of Estonian Parish Court records from the years 1821-1920. The corpus is manually annotated for seven entity types including PERSON, LOCATION and ORGANIZATION resulting in a dataset of 27.5K entities in total. They use the dataset in fine-tuning a modern Estonian BERT model for NER task and report a micro-average F1 score of 93.6%. The NER in historical documents is conventionally carried out in two steps; the digital version of a document is created by transcribing the content and then NER task is performed on the digital text. Rouhou et al. (2022) propose an end-to-end method where transcription and NER tasks are performed jointly for two handwritten datasets of marriage records.

Much of the work on NER in Turkish is dealing with modern Turkish (Aras et al., 2021; Çarık and Yeniterzi, 2022; Ozcelik and Toraman, 2022). The number of works on NER in historical Turkish is quite limited due to the lack of language resources. There are some recent attempts to provide neces-

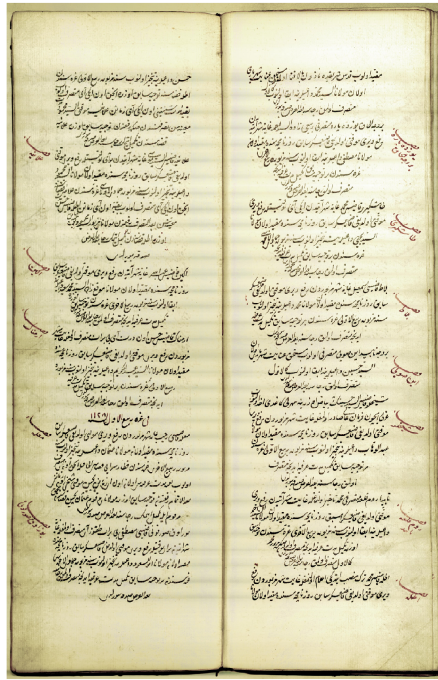
sary NLP resources in historical Turkish, such as (Karagöz et al., 2024; Özates et al., 2025). Karagöz et al. (2024) present a new text corpus for Ottoman Turkish generated from various resources and use it for continual pre-training of a BERTurk model (Karagöz et al., 2024). They evaluate the model performance on the named entity recognition task using a small annotated Ottoman Turkish dataset created from a journal. They report 0.86% F1 score for PERSON and LOCATION entities. First of its kind, a historical Turkish NER dataset, which includes 660 manually annotated sentences from late Ottoman texts, is introduced in Özates et al. (2025). A fine-tuned pre-trained language model achieved 91.21% F1 score on that dataset.

3. Motivation

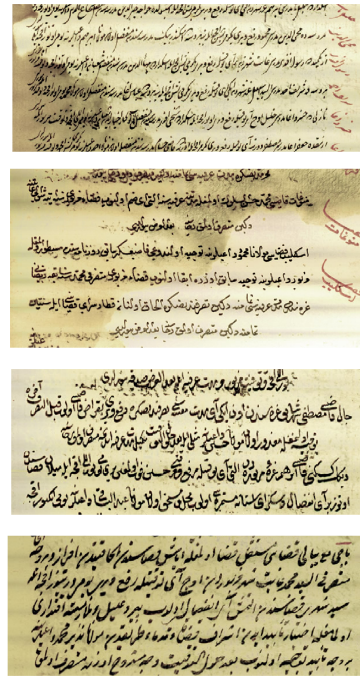
Ottoman Turkish represents a historical form of the Turkish language that was used for approximately six centuries prior to the emergence of modern Turkish. The language underwent significant changes, particularly in its vocabulary and, to a lesser extent, its syntax throughout this period. A language reform aiming at the removal of Arabic and Persian borrowed expressions made a significant contribution to the transformation of the language. Replacement of the Arabic alphabet-based Ottoman writing system with the Latin alphabet-based Turkish script in 1928 widened the gap between the two versions of the same language.

Many historical languages suffer from a lack of NLP resources, in contrast to modern languages that benefit from extensive linguistic corpora and annotated datasets. Ottoman Turkish is not an exception. Limited availability of annotated corpora, dictionaries, and linguistic tools hinders development of effective NLP models for Ottoman Turkish. One potential approach to mitigate this issue is leveraging existing resources developed for the modern Turkish, given the fact that there is a continuity between the two versions. However, the linguistic transformation due to changes in vocabulary, semantic interpretation, and grammatical structure (Özates et al., 2024) complicates the direct use of modern Turkish NLP tools for Ottoman Turkish texts. Yet, period- and domain-specific language resources can be used for adapting the modern tools.

Our main goal in this work is to create a NER dataset tailored to historical texts, enabling accurate recognition and classification of named entities. In its current form, the dataset can serve as a complementary layer to existing models and corpora—facilitating adaptation through techniques such as transfer learning and domain-specific fine-tuning. By leveraging this dataset in conjunction with modern language models, researchers can



(a)



(b)

Figure 1: Images from ruznamçe volumes; a) image of two full pages: individual records are organized as paragraphs with relevant kaza names written with red ink as side notes b) details from different volumes: readability of the texts varies due to script density, calligraphy style and physical deterioration level.

more effectively bridge the gap between contemporary NLP tools and historical language data.

3.1. Ruznamçe Registers

The *kaza* (judicial district) was the smallest administrative unit in the Ottoman Empire. A judicial system based on *kaza* districts spread across vast territories and remained in place for several centuries, until the dissolution of the Ottoman Empire in 1922. Faced with the challenge of administering justice to a highly diverse population across millions of square kilometers, the Ottoman state developed one of the most extensive judicial institutions of its time, particularly expanding during the 16th and 17th centuries.

The *kadı* (judge) was the official responsible for both judicial and administrative matters within a *kaza*. *Kadis* functioned not only as judges but also as municipal administrators, wielding authority akin to that of a mayor or district governor. Their duties extended well beyond the courtroom, encompassing a wide range of civic and administrative responsibilities.

The official registers documenting judicial appointments made by the Anatolian and Rumelian *kazaskers* (chief judges) are known as *Kazasker Ruznamçesi* in Ottoman bureaucracy. Some vol-

umes also include records related to the appointments of provincial governors and religious scholars, along with other administrative notes. While the exact date of the inception of these records is unclear, it is generally accepted that they began in the 16th century. The earliest known *ruznamçe* dates to 1544. There are two distinct collections corresponding to the two main judicial district (*kazaskerlik*) of the Ottoman Empire: the Anatolian and Rumelian. To date, 167 *ruznamçe* volumes have been identified for Anatolia and 261 for Rumelia (Kütükçü, 2022).

The *ruznamçe* registers feature entries that are handwritten in the Arabic alphabet-based Ottoman alphabet. Sample images of the register pages are presented in Figure 1. The texts are created with various calligraphic styles and often densely arranged in a single-column format. Interspersed among the entries are date lines, occasionally written in red ink to enhance visual distinction. In addition to the main body of the text, the name of the relevant *kaza* is separately noted alongside each entry. Generally, individual records are organized as paragraphs, which make them easy to detect. However, elements like side notes, corrections and non-text parts can complicate the segmentation process.

The reading of these registers presents several

Antakya mutasarrıfı Hasan gayet-i şehri atiden iki ay tevkitle ref ve yeri ba-salise muvakkiti olduđu kazasker-i esbak ruznamçesinde mukayyed sezavar-ı inayet-i şehriarı ve şayeste-i refet-i tacdari Mevlana Ömer bin Mehmed el-İstanbuli da'ilerine ber-tevcih-i esbak İbka olunub sene-i atıyye Muharremü'l-haramı guresinden ancak altı ay mutasarrıf olub bakiyye-i müddet-i örfiyyesin kazay-ı ahardan tekmil eylemek ricasına ba'de'l-arz sadaka buyuruldu.

Antakya mutasarrıfı Hasan, gelecek ayın sonundan (itibaren) iki ay vakitlendirilerek görevden alınmış ve yerine üçüncü dereceden muvakkit olduđu eski kazaskerin ruznamçesinde kayıtlı olan, padişahın yardımına layık ve sultanlık korumasına uygun görülmüş, İstanbullu Mehmed ođlu Mevlana Ömer duacılarının, eskiden verilen (görev) üzerine devam ettirilip gelecek senenin Muharremü'l-haramı birinci gününden itibaren sadece altı ay (süreyle) mutasarrıf olup kalan mutad görev süresini başka bir kazada tamamlaması ricası arz edildikten sonra ihsan olarak kabul buyuruldu.

(a)

(b)

Hasan, the mutasarrıf (i.e. sanjak governor.) of Antakya, was dismissed from his position with a two-month notice starting from the end of next month, and it was approved that Mevlana Ömer, son of Mehmed from Istanbul—who is recorded in the diary of the former kazasker as possessing the third rank, deemed worthy of the Sultan's protection and imperial favor, and who had previously been appointed to this position—be reappointed as mutasarrıf for only six months beginning on the first day of the month of Muharrem of the coming year, after which he will complete the remainder of the usual term in another district.

(c)

Figure 2: A ruznamçe record; a)transcribed text b) translation to modern Turkish c)translation to modern English. Equivalent expressions are highlighted with the same color.

Type of Evolution	Ottoman Turkish	Modern Turkish	English
Obsolete word(s)	Ba-salise	Üçüncü dereceden	Third rank
	Tevkit	Vakitlendirme	Scheduling
	Ref'	Görevden alma	Removal
	Sezavar	Layık	Worthy
	Da'i	Duacı	Prayerful well-wisher
Noun Phrase Order	İbka	Devam ettirme	To continue
	Gayet-i şehri-ı atı	Gelecek ayın sonu	The end of next month
	Sene-i atıyye	Gelecek sene	Next year
Modernization	Bakiyye-i müddet-i örfiyye	Kalan mutad süre	Remainder of the usual term
	Esbak	Eski	Preceding
	Mukayyed	Kayıtlı	Registered
	İnayet	Yardım	Favor
	Tevcih	Verme	Referral

Table 1: Examples from the RuznamceNER dataset showing language evolution.

challenges. These difficulties stem not only from the calligraphic styles and the physical deterioration common in historical documents but also from the complexity of interpreting personal names, which requires considerable expertise. An additional challenge arises in identifying place names; variations in spelling, the presence of multiple possible readings, and references to locations no longer traceable today make the accurate identification of place names a task requiring specialized knowledge.

The records are semi-structured in the sense that they contain almost a standard set of information. However, the language evolved significantly throughout the centuries, which led to variation in the vocabulary and writing style.

Most of the records deal with a single appointment while there are some records that contain

complex regulations regarding several interleaved appointments.

Figure 2 shows transcription of a record from 1807, along with its translations into modern Turkish and English. Correlated expressions are marked with the same colors. The highlighted words show different ways of vocabulary evolution. A frequent type of modification is replacement of obsolete words with newer ones. In some cases one obsolete word is replaced by a multi-word expression (*ref' → görevden alma (removal)*). Some of the words are replaced by more popular and widely known synonyms since they are mostly used in specific domains like law and religion. Finally, the noun phrase formation used in Ottoman Turkish usually follows the Persian pattern, whereas they are replaced by the Turkish pattern in modern Turk-

ish (*sene-i atiyye* → *gelecek sene* (next year)). A replacement word can be etymologically related to the old one (ex. *tevkîl* → *vakıtlendirme* (scheduling)) or a synonym (ex. *sezavar* → *layık* (worthy)). Table 1 explains various types of language evolution in the example given in Figure 2.

The NER dataset is created as a part of a larger project that aims to extract information contained in the Kazasker Ruznamçe records pertaining to districts in Anatolia between 1660 and 1895 through automated and semi-automated methods. Subsequently, the extracted data will be used for analysis of the Ottoman state from economic, administrative, and institutional perspectives within approximately a 250-year period. Additionally, the project envisions the semantic modeling of the collected data, its integration into a graph-based database, the discovery of new information through data mining, and the incorporation of these data into Linked Open Data sources.

4. The RuznamceNER Dataset

In order to create a new Ottoman NER dataset from ruznamçe records, we first started with collecting manually transcribed texts to form a corpus. After a character set normalization process, we manually annotated the text for three entity types: PERSON, LOCATION and ORGANIZATION. This section gives details about the dataset creation process.

4.1. The Text

The original script of the ruznamçe documents is the Ottoman script, which uses an extended version of the Arabic alphabet. It is a common practice for scholars to transcribe them into the modern Turkish alphabet before using the texts for research purposes. In this work, we generated a corpus from transcripts of three ruznamçe volumes from the years 1603, 1720 and 1807.

There are slight variations between transcription schemes of the three volumes since they are transcribed by three different people. We applied character normalization to unify some accented characters such as "â" and "î" before proceeding to the annotation process.

4.2. Annotation Process

The annotation process is carried out by two human annotators using a simple annotation tool¹. The ruznamçe dataset is annotated with PERSON, LOCATION, and ORGANIZATION types while words of other types of tags are all marked with O. We use the following definitions of tags:

¹<https://arunmozhi.in/ner-annotator/>

- PERSON : Human beings regarded as individuals.
- LOCATION: Geopolitical entity (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.

The inter-annotator agreement is measured using Cohen's kappa, yielding $\kappa = 0.94$ computed over 40K tokens with labels {PERSON, ORGANIZATION, LOCATION, O}. There are various honorifics appearing in ruznamçe records. We adopted the following approach when dealing with titles, appellations etc. attached to person names:

- titles such as *el-Hac*², *es-Seyyid*³, *eş-Şeyh*⁴, *Hafız*⁵, *Mevlana*⁶, *Hattat*⁷ are not included in the PERSON tags.
- consecutive titles containing non-specific lineage reference like *Şeyhulislam bin/ibn Şeyhulislam* (the Chief Islamic Legal Authority who is also son of the Chief Islamic Legal Authority) are not included in the PERSON tags.
- titles occurring in lineage reference of a person such as *Ahmed bin es-Seyyid Mehmed* are included in the PERSON tags.
- gentilic appellations like *Ahmed İstanbuli* or *İstanbuli Ahmed* (Ahmed of İstanbul) are included in the PERSON tags.
- in case of Turkish gentilic appellations like *Ahmed İstanbul'dan* (Ahmed from İstanbul), the place name is not included in the PERSON tags; it is tagged as LOCATION.

The data is prepared in the CoNLL-2003 format with multiple word entities marked with B- and I- prefixes. The resulting dataset contains 2,138 sentences; 4,442 PERSON, 3,787 LOCATION, and 501 ORGANIZATION tags in total. Details regarding the dataset are given in Table 2.

5. Methodology

We conducted a series of experiments for evaluation of the NER dataset. Our goals are 1) to

²a honorific title for visitors of Mecca

³honorific title for descendants of the Islamic prophet Muhammad

⁴a honorific title showing affiliation to a sufi order

⁵a title for Qoran memorizers

⁶a title showing affiliation to Mevlevi order

⁷a title for calligraphers

Split	Number of					
	Sentences	Tokens	Tokens/Sent	PER Tags	LOC Tags	ORG Tags
Train	1,497	80,213	53.58	3,157	2,665	357
Dev	214	10,641	49.72	415	376	53
Test	427	22,633	53.00	870	746	91
Total	2,138	113,487	53.08	4,442	3,787	501

Table 2: Some statistics for the RuznamceNER dataset.

explore the usability of the base model for the specific dataset at hand, 2) to measure the effect of in-domain data on the recognition performance, 3) to understand how much in-domain data is needed to achieve effective NER performance on the ruznamçe documents. We trained models using two other NER datasets: one in modern Turkish and another in historical Turkish from another domain in addition to the training portion of the RuznamceNER dataset. This section presents the details about the datasets and the NER architecture we employed.

5.1. Datasets

We utilized three datasets in the experiments: **MilliyetNER**, a large modern Turkish NER dataset; **HisTR**, a small historical Turkish NER dataset sourced from late Ottoman text and the newly generated **RuznamceNER** dataset.

MilliyetNER. MilliyetNER (Tür et al., 2003) was collected from Turkish news articles and manually annotated for three entity types: PERSON, LOCATION, and ORGANIZATION.

It is a large-scale dataset containing over 27K sentences and 500K tokens, which makes it one of the most frequently used benchmarks for Turkish NER tasks. In our experiments, we used a subset of 1,500 randomly selected sentences from the Milliyet-NER dataset to compare the effects of different datasets on model performance with similar dataset sizes.

HisTR. HisTR (Özates et al., 2025) is the first named entity recognition dataset specifically developed for historical Turkish.

The dataset comprised of sentences drawn from the Servet-i Fünun (SF) journal, a prominent Ottoman magazine published between 1896 and 1901 that covered topics such as literature, science, daily life, and world news. HisTR was manually annotated with three entity types: PERSON, LOCATION, and ORGANIZATION.

The dataset includes 660 sentences and 18K tokens, making it a valuable resource for evaluating NER models on historical Turkish text.

RuznamceNER. The newly introduced dataset in Section 4 comprises 2,138 ruznamçe records transcribed in the Latin alphabet, extracted from three fascicles spanning the years 1603–1807. Each record contains one sentence. It was manually annotated with PERSON, LOCATION, and ORGANIZATION tags and includes more than 113K tokens.

5.2. Model

As a baseline NER model, we opted to utilize a BERT-CRF hybrid model, which reflects an architecture widely used for sequence labeling tasks, including NER (Nie et al., 2021; Chang et al., 2021). BERT-CRF combines contextualized embeddings from a pre-trained BERT architecture with a Conditional Random Field (CRF) layer to model label dependencies across tokens (Nie et al., 2021).

BERT. Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) is a transformer-based model pre-trained on large-scale text corpora using a masked language modeling and next sentence prediction objectives. For our experiments, we employed BERTurk (Schweter, 2020), a monolingual BERT model developed specifically for Turkish and trained exclusively on extensive Turkish text corpora. BERTurk has demonstrated state-of-the-art performance across a range of natural language understanding tasks in Turkish (Uzunoglu and Şahin, 2023; Uludoğan et al., 2024), making it a strong choice for our NER experiments.

CRF. Conditional Random Fields (CRFs) (Lafferty et al., 2001) model the conditional probability of a label sequence given an input sequence. They remain strong baselines for structured prediction (Mao et al., 2024; Qiu et al., 2025). In this study, CRF layers were integrated with transformer encoders (e.g., BERT-CRF) to better capture sequential dependencies and improve entity boundary consistency.

Our BERT-CRF implementation is based on the approach described in Souza et al. (2019). The model stacks a pre-trained BERT encoder, a token-level linear layer that produces per-tag emission scores, and a linear-chain CRF with learned start/end transitions. BERT encodes the

WordPiece sequence into contextual representations; the linear layer projects each token to the BIO tag set; and the CRF models local label dependencies and decodes the most likely tag sequence using Viterbi. Training maximizes the conditional log-likelihood of the gold sequence, and—consistent with WordPiece practice (Devlin et al., 2019)—losses and predictions are computed only for the first sub-token of each word.

6. Experiments

The experiments were designed to assess the effectiveness of domain adaptation using different NER datasets and to understand how modern Turkish and historical Turkish training data from other domains contribute to model performance on the named entity recognition of ruznamçe documents.

This section provides experimental settings, results, and an error analysis on the predictions of the best-performing model.

6.1. Experimental Settings

We performed five separate fine-tuning experiments using various dataset combinations: (1) Milliyet-NER, (2) HisTR, (3) Milliyet-NER + HisTR, (4) RuznamceNER, and (5) MilliyetNER + RuznamceNER. All models were evaluated on the same RuznamceNER test set to ensure comparability.

The BERTurk-based BERT-CRF model was optimized using the AdamW optimizer with a learning rate of $5e-5$, trained for five epochs with a batch size of two, and gradient accumulation over eight steps to efficiently utilize GPU memory. The training was conducted using the Hugging Face `Trainer` API and all experiments were run on a single NVIDIA A100 GPU (40 GB memory) using mixed precision (`bfloat16`) for efficiency. Preprocessing, token alignment, and evaluation were handled via custom Python scripts with the `nervaluate` library for entity-level metrics.

6.2. Results

Table 3 shows the model’s performance on the RuznamceNER test set with different training setups. Precision, recall, and F1-score were computed using the *strict* matching scheme, where a prediction is considered correct only if both the entity boundaries and the entity type exactly match the gold annotation. This approach ensures a rigorous assessment of each model’s ability to identify and classify entities accurately, avoiding inflated scores from partial overlaps.

The results in Table 3 reveal several key patterns. The supervised BERT-CRF models perform strongly when fine-tuned on in-domain data. Models trained with the RuznamceNER dataset (either

alone or combined with Milliyet-NER) achieve near-perfect performance ($F_1 \approx 0.98$), underscoring the importance of domain-specific supervision. In contrast, using only modern Turkish data (Milliyet-NER) yields lower performance ($F_1 = 0.52$), reflecting the domain mismatch between contemporary and historical language. Training only with HisTR results in an improved performance ($F_1 = 0.61$) and combining HisTR with Milliyet-NER further improves results moderately ($F_1 = 0.62$), suggesting that historical corpora closer in style and period to RuznamceNER help bridge the gap. These results suggest that in-domain training data is necessary to achieve an acceptable NER performance on the ruznamce documents.

To address the question of how large the in-domain training data should be, we investigated the effect of training set size on our model’s performance through an ablation study, progressively increasing the number of training sentences from the RuznamceNER dataset.

Figure 3 presents the results of an ablation study illustrating how the model’s performance changes as the training data size increases from 100 to 1,497 sentences (evaluated at 100, 500, 1,000, and 1,497 sentences from the training set of the RuznamceNER dataset). We observe that the performance reaches a saturation point at a training size of 1,000 sentences and does not improve further when the full training set (1,497 sentences) is used.

Considering the results presented in Table 3 and Figure 3, we draw two main conclusions: (1) in-domain training data is essential for performing NER tagging on ruznamce documents, and (2) even a small amount of supervised in-domain data is sufficient to achieve strong performance. This is primarily because ruznamçe texts are well structured, although their lexicon differs significantly from that of modern Turkish and even late Ottoman texts (as in HisTR). Once the model learns the structural patterns and domain-specific vocabulary from a limited number of examples, the NER task becomes relatively straightforward. However, in the absence of such supervised in-domain data, the performance does not exceed an F1 score of 0.62, as shown in Table 3.

6.3. Error Analysis

The BERT-CRF model achieves a remarkably high performance when trained on the RuznamceNER training set, reaching an F1 score of 0.98. To better understand the remaining sources of error, we conducted a manual error analysis to identify where the model tends to make mistakes.

The errors are more frequent for LOCATION and ORGANIZATION type entities. We observe that multi-word place names like *Seferihisar Günyüzü*

Table 3: Performance comparison of BERT-CRF models fine-tuned on different datasets.

BERT-CRF	Precision	Recall	F1
Milliyet-NER	0.48	0.55	0.52
HisTR	0.60	0.62	0.61
Milliyet-NER+HisTR	0.60	0.64	0.62
RuznamceNER	0.98	0.98	0.98
Milliyet-NER+RuznamceNER	0.98	0.98	0.98

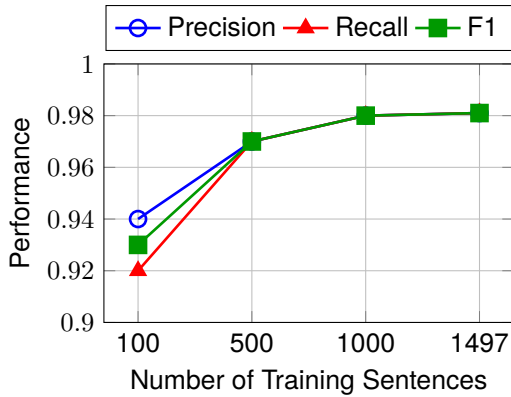


Figure 3: Performance of the BERT-CRF model on the RuznamceNER test set when trained with different amounts of in-domain training data.

confuse the model and it labels them as separate entities. Similarly, genitive (construct) phrases (i.e., *ezafe* construction) lead to errors in recognition of location entities. For example, the phrase *Medine-i Niğde'de* (in the City of Niğde) is in locative form and the model partly recognized it such that *Niğde* is correctly recognized as a location but the group of words preceding *Niğde* (i.e., *Medine-i*) are incorrectly labeled. However, a similar location name, *Mekke-i Mükerrerme* (Mecca the Honored), is in nominative form and completely misrecognized. A close examination reveals that the model makes use of the locative case suffix in Turkish (i.e., *-de/da/te/ta*), which marks location or positions, in partial recognition as in the case with *Medine-i Niğde'de*.

A similar pattern is observed in the detection of organizations, which are typically preceded by a place name bearing a dative, locative, or ablative case suffix. In the example *...İstanbul'da Musa Bey'in Medresesinden...* (...from the madrasa of Musa Bey in Istanbul...), *Musa Bey'in Medresesi* constitutes a multi-word organization name. Although the model failed to recognize the word *Medrese*—which means “theological school” and denotes an organization—it nevertheless identified the locative case suffix in *İstanbul'da* and labeled only the word *Musa* as an organization, while failing to detect the rest of the organization name. However, the model sometimes appears to rely more on vocabulary lists than on contextual pattern analysis.

In a similar organization example, *...İstanbul'da Anbar Kadısı Medresesinden...* (...from the madrasa of the judge of Anbar in Istanbul...), the model labels *Anbar* as a location, most likely because it corresponds to a modern-day place name in Iraq.

PERSON type entities are usually correctly labeled with a few examples where the model fails to recognize gentilic appellations as in *Kırımı* (from Crimea) and *İstanbuli* (from İstanbul).

Nevertheless, these errors remain rare, and exposing the model to a greater number of similar instances can substantially reduce the likelihood of their occurrence.

7. Conclusion

We have introduced the ruznamçe registers, a valuable historical resource about judge appointments in the Ottoman Empire, and presented RuznamceNER, a manually annotated NER dataset spanning two centuries of these documents. Our evaluation of the dataset using a strong baseline model showed that in-domain training data plays a crucial role in achieving high accuracy in NER in such texts. Based on this result, it is possible to estimate importance of in-domain training data for processing other Ottoman documents like court records (i.e. *sicil*) and official decrees (i.e. *ferman*, which have larger lexicons and more complex structures).

By making both the dataset and the trained model publicly available, we provide resources that can support large-scale NER and automated analysis of ruznamçe registers, as well as further research on other unprocessed historical Turkish texts.

The dataset currently contains annotations for PERSON, LOCATION and ORGANIZATION entities. But the potential of ruznamçe records is not limited to them. As future work, extraction of other information like the reasons for change or removal of a judge and scheduled chain appointments can be targeted since they will help to provide detailed insights into the social, political, and economic life of the Ottoman Empire.

This work was supported by the Scientific and Technological Research Council of Türkiye (TÜBİTAK) under the Innovative Solutions in Social Sciences and Humanities Research Projects Funding Program (3005) with Grant No. 124K247. The scientific content is solely the responsibility of the

authors and does not necessarily reflect the views of TÜBİTAK.

8. Bibliographical References

- 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.
- Baptiste Blouin, Cécile Armand, and Christian Henriot. 2024. A dataset for named entity recognition and entity linking in Chinese historical newspapers. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 385–394. ELRA and ICCL.
- Buse Çarık and Reyhan Yeniterzi. 2022. A twitter corpus for named entity recognition in Turkish. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022*, pages 4546–4551. European Language Resources Association.
- Yuan Chang, Lei Kong, Kejia Jia, and Qinglei Meng. 2021. [Chinese named entity recognition method based on BERT](#). In *2021 IEEE International Conference on Data Science and Computer Application (ICDSCA)*, pages 294–299.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2024. Named entity recognition and classification in historical documents: A survey. *ACM Computing Surveys*, 56(2):27:1–27:47.
- İsmail Gündoğdu. 2009. The daybooks of chief justice's as a historical sources of the Ottoman empire and their significance. *Journal of Human Sciences*, 6(2):697–722.
- Fatih Karagöz, Berat Doğan, and Şaziye Betül Özateş. 2024. Towards a clean text corpus for Ottoman Turkish. In *Proceedings of the First Workshop on Natural Language Processing for Turkic Languages (SIGTURK 2024)*, pages 62–70.
- Levent Kuru. 2020. XVIII. yüzyılın İlk yarısında Rumeli kazaları ve statüleri (kazasker ruznamçelerine göre). *Turkish Studies – Historical Analysis*, 15(1):161–191.
- Büşra Aktaş Kütükçü. 2022. *1 Numaralı Anadolu Kazaskerliği Ruznamçesi*. Yeditepe Akademi, İstanbul.
- John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning, ICML '01*, page 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.
- Hongli Mao, Xian-Ling Mao, Hanlin Tang, Yu-Ming Shang, and Heyan Huang. 2024. Span graph transformer for document-level named entity recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18769–18777.
- Binling Nie, Ruixue Ding, Pengjun Xie, Fei Huang, Chen Qian, and Luo Si. 2021. Knowledge-aware named entity recognition with alleviating heterogeneity. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13595–13603.
- Siim Orasmaa, Kadri Muischnek, Kristjan Poska, and Anna Edela. 2022. Named entity recognition in Estonian 19th century parish court records. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5304–5313.
- Şaziye Betül Özateş, Tarık Tıraş, Efe Genç, and Esmâ Bilgin Tasdemir. 2024. Dependency annotation of Ottoman Turkish with multilingual BERT. In *Proceedings of the 18th Linguistic Annotation Workshop (LAW-XVIII)*, pages 188–196, St. Julians, Malta. Association for Computational Linguistics.
- Saziye Betül Özates, Tarık Emre Tiras, Ece Elif Adak, Berat Dogan, Fatih Burak Karagöz, Efe Eren Genç, and Esmâ F. Bilgin Tasdemir. 2025. Building foundations for natural language processing of historical Turkish: Resources and models. *CoRR*, abs/2501.04828.
- Oguzhan Ozcelik and Cagri Toraman. 2022. Named entity recognition in Turkish: A comparative study with detailed error analysis. *Information Processing and Management*, 59(6):103065.

- Nita Patil, Ajay S Patil, and BV Pawar. 2016. Survey of named entity recognition systems with respect to Indian and foreign languages. *International Journal of Computer Applications*, 134(16).
- Yunfei Qiu, Libo Dong, Wenwen Zhang, Haoran Xing, and Junwei Huang. 2025. A diffusion enhanced crf and bilstm framework for accurate entity recognition. *Scientific Reports*, 15(1):19670.
- Ahmed Cheikh Rouhou, Marwa Dhiyf, Yousri Kessentini, and Sinda Ben Salem. 2022. Transformer-based approach for joint handwriting and named entity recognition in historical document. *Pattern Recognition Letters*, 155:128–134.
- Stefan Schweter. 2020. [BERTurk - BERT models for Turkish](#).
- Khaled Shaalan. 2014. A survey of Arabic named entity recognition and classification. *Computational Linguistics*, 40(2):469–510.
- Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2019. [Portuguese named entity recognition using BERT-CRF](#).
- Rachele Sprugnoli. 2018. Arretium or arezzo? A neural approach to the identification of place names in historical texts. In *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Torino, Italy, December 10-12, 2018*, volume 2253 of *CEUR Workshop Proceedings*. CEUR-WS.org.
- Konstantin Todorov and Giovanni Colavizza. 2020. Transfer learning for named entity recognition in historical corpora. In *Working Notes of GLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020*, volume 2696 of *CEUR Workshop Proceedings*. CEUR-WS.org.
- Gökhan Tür, Dilek Hakkani-Tür, and Kemal Oflazer. 2003. [A statistical information extraction system for Turkish](#). *Natural Language Engineering*, 9(2):181–210.
- Gökçe Uludoğan, Zeynep Balal, Furkan Akkurt, Meliksah Türker, Onur Gungor, and Susan Üsküdarlı. 2024. [TURNA: A Turkish encoder-decoder language model for enhanced understanding and generation](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10103–10117, Bangkok, Thailand. Association for Computational Linguistics.
- Arda Uzunoglu and Gözde Şahin. 2023. [Benchmarking procedural language understanding for low-resource languages: A case study on Turkish](#). In *Proceedings of the 13th International Joint*
- Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 804–819, Nusa Dua, Bali. Association for Computational Linguistics.