T.C.

MARMARA UNIVERSITY

FACULTY OF ENGINEERING

COMPUTER ENGINEERING DEPARTMENT

CSE4078 Introduction to Natural Language Processing

Group 1 - Delivery #1

Group Members:

Leen I. A. Shaqalaih 150121921

Fatma Melisa Küçük 150119916

Ayşe Sena Aydemir 150119735

Ahmet Sinan Kalkan 150119747

Ahmet İkbal Adlığ 150120517

NLP Task - NER:

Named Entity Recognition (NER) is a core task in Natural Language Processing (NLP) that focuses on identifying and categorizing named entities in a text. These entities typically fall into predefined categories such as persons, organizations, locations, numerical values, and other domain-specific labels.

NER is generally approached as a sequence labeling problem, with models being trained to assign entity tags to words or phrases. Recent advancements in deep learning have led to the adoption of powerful architectures such as Transformer-based models (e.g., BERT) and recurrent neural networks (RNNs), which enhance the performance of NER systems. These models require annotated datasets containing labeled entities, which are crucial for model training and evaluation.

NER plays a significant role in numerous applications, including information extraction, question answering, and content analysis, enabling machines to derive structured information from unstructured text.

Turkish Datasets

For this project, we selected 11 publicly available datasets from different repositories, including Kaggle, Hugging Face, and GitHub. These datasets were converted into a standardized format suitable for NER tasks. Below is an overview of the datasets used:

Datasets	# of rows in training dataset	# of rows in test dataset	# of rows in validation dataset	Total # of rows in datasets	URL	Source	Description
Vitamins and Supplements NER	2072	200	200	2472	https://huggingface.co/datasets/turkish-nlp-suite/vitamins-supplements-NER	https://huggingface.co/turkish-nlp-suite	A collection of customer reviews after using supplements from Vitaminler.com. Reviews include reasons of buying, effectiveness, dosages, side effects, smell, taste, etc.
Turkish	1662532	-	-	1662532	https://huggi	https://huggi	Focuses on

Organization NER					ngface.co/datasets/STNM-NLPhoenix/turkish-org-ner	ngface.co/STNM-NLPhoenix	organization entities. It has 3 labels: B (Beginning), I (Inside), O (Outside) an organization entity.
Turkish Wiki-NER	18000	1000	1000	20000	https://github.com/turkish-nlp-suite/Turkish-Wiki-NER-Dataset	github/DuyguA	A dataset derived from Wikipedia sentences, re-annotated from Kuzgunlar NER.
ATISNER (Airline Travel Information System)	4,978	890	-		https://huggingface.co/datasets/ctoraman/atis-ner-turkish		ATISNER, includes airline spoken queries translated from English to Turkish, customized for Named Entity Recognition.
NER T5 Turkish	-	-	-	299,800	https://www.kaggle.com/datasets/binbirmetin/ner-t5-turkish	kaggle/binbirmetin	A large dataset leveraging the T5 (a text-to-text transfer transformer) model for NER applications.
Turkish NER	-	-	-	40,000	https://huggingface.co/datasets/erayyildiz/turkish_ner		An automatically labeled Turkish corpus using gazetteers.
PAN-X.tr	20000	10,000	10,000	40000	https://huggingface.co/datasets/xtreme/viewer/PAN-X.tr	Huggingface/xtreme	A crowd-sourced for the MultiNLI corpus.
NakbaNER	4032	-	-	4032	https://github.com/sb-b/NakbaTR/tree/main	github/sb-b	developed to capture narratives surrounding the Nakba—the mass displacement of Palestinians beginning in 1948. It's obtained from

							real testimonies and news.
SUNLP Twitter NER	7910	1697	1667		https://github.com/SU-NLP/SUNLP-Twitter-NER-Dataset/tree/main	github/busecarik	A Twitter Corpus for Named Entity Recognition in Turkish. The dataset consists of 5,000 randomly selected tweets published between June 2020 and June 2021.
TWNERTC & EWNERTC	450,000	200,000	50,000	700,000 (for TR)	https://data.mendeley.com/datasets/cdcztymf4k/1		TWNERTC (Turkish Wikipedia Named-Entity Recognition and Text Categorization) and EWNERTC (English Wikipedia NER and Text Categorization), consists of automatically categorized and annotated sentences sourced from both Turkish and English Wikipedia.
HisTR	13100	6540	5660	25306	https://huggingface.co/datasets/BUCOLIN/HisTR	Huggingface/ BUCOLIN	Ottoman Turkish NER dataset manually using a subset of sentences from issues of Servet-i Funun journal. It covers a wide range of topics including literature, science, daily life and world news.

1- Vitamins and Supplements NER:

This dataset is sourced from user reviews from Vitaminler.com related to vitamins and supplements, specifically from a variety of online product platforms and forums. The reviews are in Turkish, offering a diverse range of feedback from users who have used different vitamin and supplement products. It was released 8 months ago. It has a total of 2472 instances (2072 training, 200 validation, 200 test). The dataset also includes emojis.

```
{"text": "D vitamini eksikliğim nedeniyle aldım.  yaklaşık 1ay önce de almıştım indirime girmişken elimdeki
bitmeden yeni kutu siparisini verdim. kronik yorgunluk, halsizlik yaşayan birisiyim, yoğun iş temposunda
çalışıyorum kullandığımdan beri daha güçlü,daha enerjik hissediyorum faydasını gördüm tavsiye ederim 😊", "spans":
[{"val": "kronik yorgunluk, halsizlik yaşayan birisiyim", "label": "SAĞLIK_ŞİKAYETLERİ", "start": 136, "end":
181}, {"val": "D vitamini eksikliğim", "label": "HASTALIK", "start": 0, "end": 21}, {"val": "daha güçlü,daha
enerjik hissediyorum", "label": "ETKİ", "start": 235, "end": 271}]}
```

The dataset consists of user reviews of vitamins and supplement products in Turkish, and it includes both **entity** annotations and **semantic span** annotations.

- **Entity annotations** (NER) categorize mentions of diseases, biomolecules, users, brands, products, dosages, and other related categories.

- **Span annotations** provide additional semantic information about the relationships and effects of entities, such as health complaints, side effects, taste or smell descriptions, and reported effects (e.g., the product's impact on health or energy levels).
  This dataset is valuable for training models to recognize both structured and unstructured information within the domain of health and wellness, especially in the context of user-generated content about vitamins and supplements.

Tagset Description

This dataset includes annotations for both **named entities** and **semantic spans**, providing detailed information not only about the entities themselves but also about **what happens with those entities**, a common approach in medical NLP for capturing deeper semantics.

🔹 **Named Entity Recognition (NER) Tags:**

These tags identify concrete entities mentioned in user reviews related to vitamins and supplements. Below is a list of the NER tags and their frequencies in the dataset:

Tag	Count
Disease	1,875
Biomolecule	859
User	634
Other_product	543

Recommender	436
Dosage	471
Brand	275
User_demographics	192
Ingredient	175
Other_brand	121

- **Span Tags:**

Span annotations provide semantic insight into the effects and perceptions of the mentioned entities, such as how users feel after using a product or its sensory characteristics. These are especially useful in modeling subjective experiences like effects or side effects.

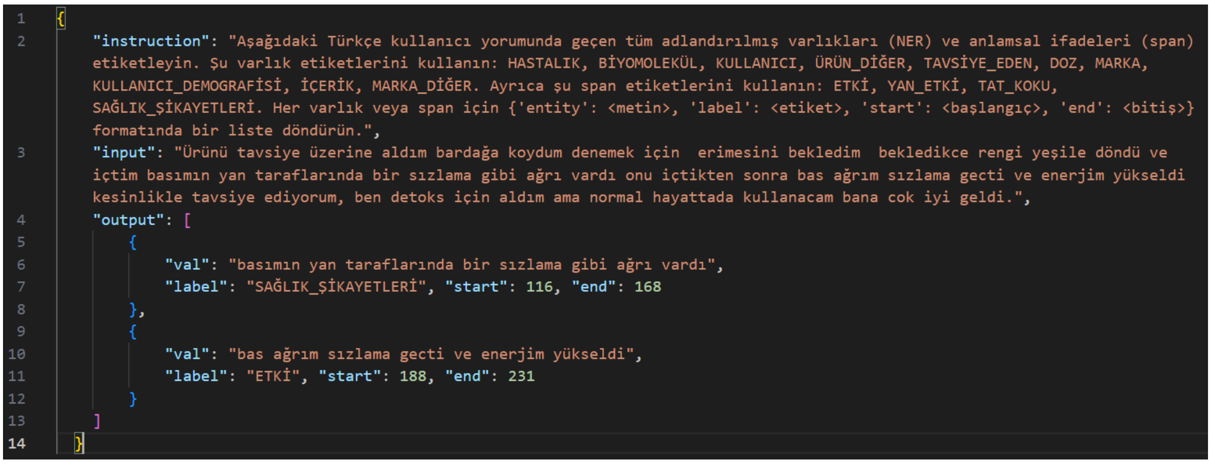
Tag	Count
Effect	2,562
Side_effect	608
Taste_smell	558
Health_complaints	858

Example instances:

```
{
      "text":
            "Ürünü tavsiye üzerine aldım bardağa koydum denemek için
            erimesini bekledim  bekledikce rengi yeşile döndü ve içtim
            basımın yan taraflarında bir sızlama gibi ağrı vardı onu
            içtikten sonra bas ağrım sızlama gecti ve enerjim yükseldi
            kesinlikle tavsiye ediyorum, ben detoks için aldım ama
            normal hayattada kullanacam bana cok iyi geldi.",
      "spans":
            [{"val": "basımın yan taraflarında bir sızlama gibi ağrı
            vardı",
            "label": "SAĞLIK_ŞİKAYETLERİ", "start": 116, "end": 168},
            {"val": "bas ağrım sızlama gecti ve enerjim yükseldi",
            "label": "ETKİ", "start": 188, "end": 231}]
}

{
      "text":
          "Yaklaşık yirmi gündür kullanıyorum,eskiye oranla saç
          dökülmem azaldı ve cildimin daha canlı olduğunu
          düşünüyorum.",
     "spans":
          [{"val": "saç dökülmem azaldı ve cildimin daha canlı
          olduğunu düşünüyorum",
          "label": "ETKİ", "start": 49, "end": 112},
          {"val": "saç dökülmem",
          "label": "HASTALIK", "start": 49, "end": 61}]
}
```

Alpaca format:



```
{
    "instruction": "Aşağıdaki Türkçe kullanıcı yorumunda geçen tüm adlandırılmış varlıkları (NER) ve anlamsal ifadeleri (span) etiketleyin. Şu varlık etiketlerini kullanın: HASTALIK, BİYOMOLEKÜL, KULLANICI, ÜRÜN_DİĞER, TAVSİYE_EDEN, DOZ, MARKA, KULLANICI_DEMOGRAFİSİ, İÇERİK, MARKA_DİĞER. Ayrıca şu span etiketlerini kullanın: ETKİ, YAN_ETKİ, TAT_KOKU, SAĞLIK_ŞİKAYETLERİ. Her varlık veya span için {'entity': <metin>, 'label': <etiket>, 'start': <başlangıç>, 'end': <bitiş>} formatında bir liste döndürün.",
    "input": "Ürünü tavsiye üzerine aldım bardağa koydum denemek için  erimesini bekledim  bekledikce rengi yeşile döndü ve içtim basımın yan taraflarında bir sızlama gibi ağrı vardı onu içtikten sonra bas ağrım sızlama gecti ve enerjim yükseldi kesinlikle tavsiye ediyorum, ben detoks için aldım ama normal hayattada kullanacam bana cok iyi geldi.",
    "output": [
        {
            "val": "basımın yan taraflarında bir sızlama gibi ağrı vardı",
            "label": "SAĞLIK_ŞİKAYETLERİ", "start": 116, "end": 168
        },
        {
            "val": "bas ağrım sızlama gecti ve enerjim yükseldi",
            "label": "ETKİ", "start": 188, "end": 231
        }
    ]
}
```

2- Turkish Organization NER:

This dataset consists of 1,662,532 rows. This dataset provides sentences split into tokens, with each labeled to indicate its role in an organization entity. Each token in a sentence is annotated with labels such as B-organization to denote the beginning of an organization name, I-organization for tokens inside an organization name, and O for tokens outside of any organization entity.

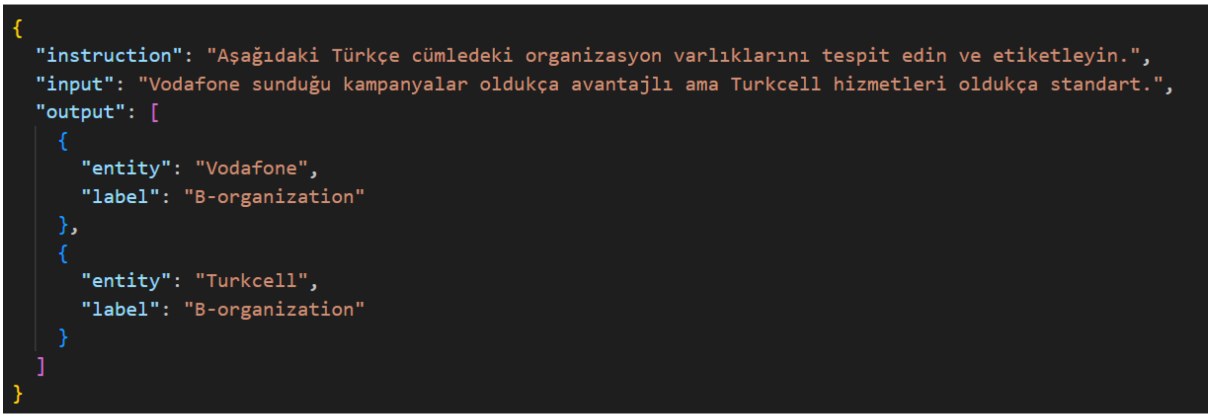
Example instances:

Token	Label
Vodafone	B-organization
sunduğu	O
kampanyalar	O
oldukça	O
avantajlı	O
ama	O
Turkcell	B-organization
hizmetleri	O
oldukça	O
standart.	O

Token	Label
Turk	B-organization
Telekom	I-organization
müşteri	O
hizmetleri	O
kötü	O
olsa	O
da	O

Türknet	B-organization
müşteri	O
hizmetleri	O
oldukça	O
ortalama.	O

Alpaca format:



```
{
  "instruction": "Aşağıdaki Türkçe cümledeki organizasyon varlıklarını tespit edin ve etiketleyin.",
  "input": "Vodafone sunduğu kampanyalar oldukça avantajlı ama Turkcell hizmetleri oldukça standart.",
  "output": [
    {
      "entity": "Vodafone",
      "label": "B-organization"
    },
    {
      "entity": "Turkcell",
      "label": "B-organization"
    }
  ]
}
```

3- Turkish Wiki-NER:

This dataset is sourced from Wikipedia texts and has been re-annotated from the Kuzgunlar NER dataset. It contains 20,000 instances (18,000 training, 1,000 validation, 1,000 test). The dataset was made available on GitHub and is designed to support Turkish NER research.

The dataset consists of sentences from Wikipedia, annotated with named entity labels. It includes various categories such as geographical locations, organizations, numerical values, and personal names.

Tagset Description

This dataset includes multiple entity categories to support structured analysis of Wikipedia texts.

- **Named Entity Recognition (NER) Tags:**

This dataset consists of 19 tags. These tags are listed below:

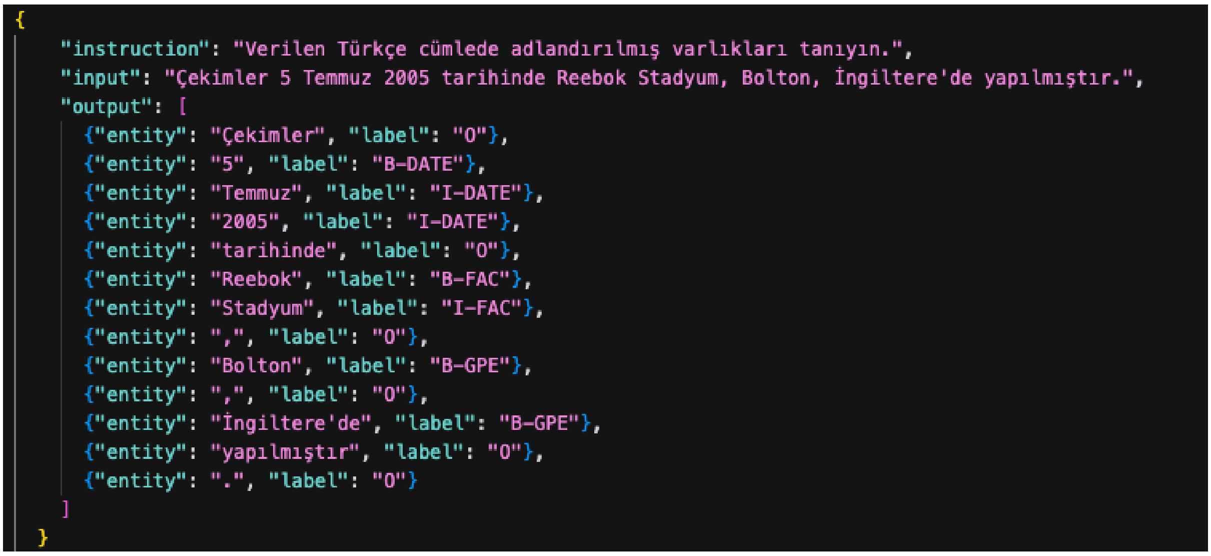
Tag	Description
CARDINAL	Numerical values representing counts or measurements.
DATE	Specific dates or periods.
EVENT	Named events such as festivals or wars.
FAC	Facilities like buildings, airports, or bridges.
GPE	Geopolitical entities, including countries, cities, or states.
LANGUAGE	Names of languages.
LAW	Legal documents or legislation.
LOC	Non-GPE locations, such as mountain ranges or bodies of water.
MONEY	Monetary values.
NORP	Nationalities or religious/political groups.
ORDINAL	Ordinal numbers (e.g., first, second).
ORG	Organizations like companies, institutions, or agencies.
PERCENTAGE	Percentage expressions.
PERSON	Individual names.
PRODUCT	Products, including vehicles, foods, or tools.

QUANTITY	Measurements such as weight or distance.
TIME	Specific times of day.
TITLE	Titles of people, like Mr., Dr., or President.
WORK_OF_ART	Names of creative works, including books, songs, or paintings.

Example instances:

```
Çekimler    O
5      B-DATE
Temmuz      I-DATE
2005  I-DATE
tarihinde   O
Reebok      B-FAC
Stadyum     I-FAC
,      O
Bolton      B-GPE
,      O
İngiltere'de      B-GPE
yapılmıştır O
.      O
```

Alpaca Format:



```
{
    "instruction": "Verilen Türkçe cümlede adlandırılmış varlıkları tanıyın.",
    "input": "Çekimler 5 Temmuz 2005 tarihinde Reebok Stadyum, Bolton, İngiltere'de yapılmıştır.",
    "output": [
      {"entity": "Çekimler", "label": "O"},
      {"entity": "5", "label": "B-DATE"},
      {"entity": "Temmuz", "label": "I-DATE"},
      {"entity": "2005", "label": "I-DATE"},
      {"entity": "tarihinde", "label": "O"},
      {"entity": "Reebok", "label": "B-FAC"},
      {"entity": "Stadyum", "label": "I-FAC"},
      {"entity": ",", "label": "O"},
      {"entity": "Bolton", "label": "B-GPE"},
      {"entity": ",", "label": "O"},
      {"entity": "İngiltere'de", "label": "B-GPE"},
      {"entity": "yapılmıştır", "label": "O"},
      {"entity": ".", "label": "O"}
    ]
  }
```

4- PAN-X.tr:

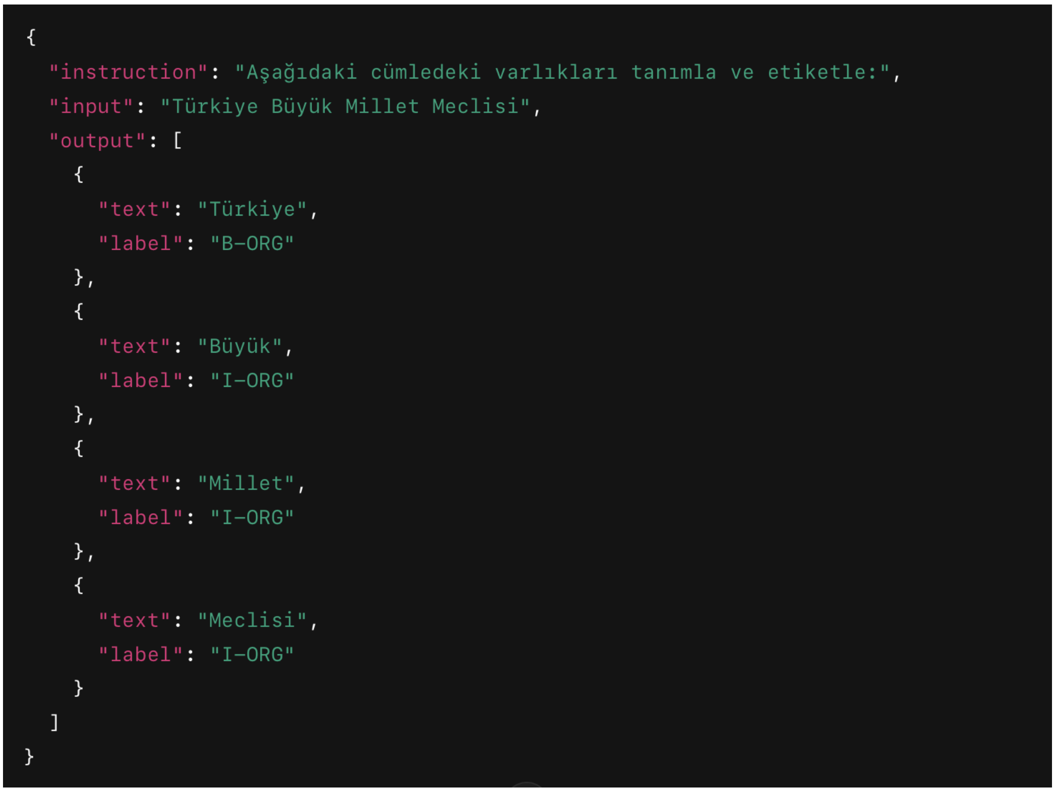
The Cross-lingual Natural Language Inference (XNLI) corpus is a crowd-sourced collection of 5,000 test and 2,500 dev pairs for the MultiNLI corpus. The pairs are annotated with textual entailment and translated into 14 languages: French, Spanish, German, Greek, Bulgarian, Russian, Turkish, Arabic, Vietnamese, Thai, Chinese, Hindi, Swahili and Urdu. This results in 112.5k annotated pairs. Each premise can be associated with the corresponding hypothesis in the 15 languages, summing up to more than 1.5M combinations. The corpus is made to evaluate how to perform inference in any language (including low-resources ones like Swahili or Urdu) when only English NLI data is available at training time. One solution is cross-lingual sentence encoding, for which XNLI is an evaluation benchmark.

The Cross-lingual TRansfer Evaluation of Multilingual Encoders (XTREME) benchmark is a benchmark for the evaluation of the cross-lingual generalization ability of pre-trained multilingual models. It covers 40 typologically diverse languages (spanning 12 language families) and includes nine tasks that collectively require reasoning about different levels of syntax and semantics. The languages in XTREME are selected to maximize language diversity, coverage in existing tasks, and availability of training data. Among these are many under-studied languages, such as the Dravidian languages Tamil (spoken in southern India, Sri Lanka, and Singapore), Telugu and Malayalam (spoken mainly in southern India), and the Niger-Congo languages Swahili and Yoruba, spoken in Africa.

For example, here is a sample from the dataset:

Token	NER Tag
Türkiye	B-ORG
Büyük	I-ORG
Millet	I-ORG
Meclisi	I_ORG

Alpaca format:

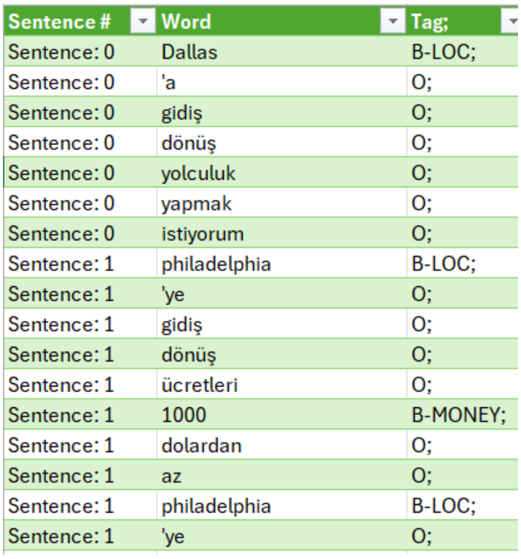


```
{
  "instruction": "Aşağıdaki cümledeki varlıkları tanımla ve etiketle:",
  "input": "Türkiye Büyük Millet Meclisi",
  "output": [
    {
      "text": "Türkiye",
      "label": "B-ORG"
    },
    {
      "text": "Büyük",
      "label": "I-ORG"
    },
    {
      "text": "Millet",
      "label": "I-ORG"
    },
    {
      "text": "Meclisi",
      "label": "I-ORG"
    }
  ]
}
```

5- ATISNER (Airline Travel Information System):

The ATIS (Airline Travel Information System) dataset consists of spoken queries annotated for the task of slot filling in conversational systems. The ATISNER dataset is a Turkish adaptation of ATIS, where airline-related spoken queries have been translated from English and customized for Named Entity Recognition (NER).

This dataset contains Turkish-language sentences related to airline travel, designed to help models recognize and categorize named entities in spoken dialogue systems. The train and test splits include 4,978 and 890 sentences, respectively.



Sentence #	Word	Tag;
Sentence: 0	Dallas	B-LOC;
Sentence: 0	'a	O;
Sentence: 0	gidiş	O;
Sentence: 0	dönüş	O;
Sentence: 0	yolculuk	O;
Sentence: 0	yapmak	O;
Sentence: 0	istiyorum	O;
Sentence: 1	philadelphia	B-LOC;
Sentence: 1	'ye	O;
Sentence: 1	gidiş	O;
Sentence: 1	dönüş	O;
Sentence: 1	ücretleri	O;
Sentence: 1	1000	B-MONEY;
Sentence: 1	dolardan	O;
Sentence: 1	az	O;
Sentence: 1	philadelphia	B-LOC;
Sentence: 1	'ye	O;

This dataset contains token-level annotations for Named Entity Recognition (NER) in airline travel queries. The table consists of three main columns:

Sentence #: Indicates which sentence the word belongs to (e.g., "Sentence: 0", "Sentence: 1").

Word: Each token (word) in the sentence.

Tag: The corresponding NER tag for each token. The tags follow the BIO format:

- **B-LOC** (Beginning of a Location): Marks the start of a location entity (e.g., city, country, airport).
- **B-MONEY** (Beginning of a Money Entity): Represents the beginning of a monetary value.
- **O** (Outside): Indicates that the token does not belong to any named entity.

Dataset Composition

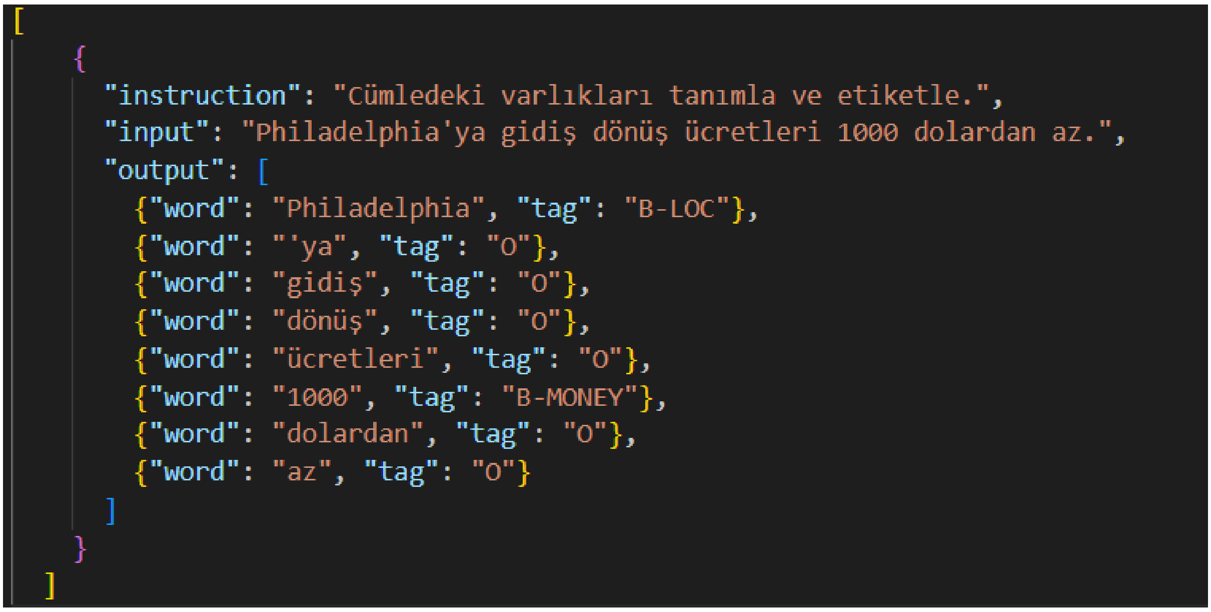
The dataset consists of tokenized sentences with each word labeled according to its entity type. The tags help identify key information in travel-related queries, making it valuable for Turkish-language conversational AI applications.

The entity tags in the dataset are:

NER Tag	Description	Count
O	Non-entity words	32,167
B-LOC	Beginning of a location name	7,16
I-LOC	Inside a location name	1,072
B-DATE	Beginning of a date expression	1,523
I-DATE	Inside a date expression	130
B-TIME	Beginning of a time expression	541
I-TIME	Inside a time expression	25
B-ORG	Beginning of an organization name	679

I-ORG	Inside an organization name	197
B-CODE	Beginning of an airline or airport code	245
I-CODE	Inside an airline or airport code	2
B-MONEY	Beginning of a price-related term	47
I-MONEY	Inside a price-related term	1

Alpaca format:



```
[
    {
      "instruction": "Cümledeki varlıkları tanımla ve etiketle.",
      "input": "Philadelphia'ya gidiş dönüş ücretleri 1000 dolardan az.",
      "output": [
        {"word": "Philadelphia", "tag": "B-LOC"},
        {"word": "'ya", "tag": "O"},
        {"word": "gidiş", "tag": "O"},
        {"word": "dönüş", "tag": "O"},
        {"word": "ücretleri", "tag": "O"},
        {"word": "1000", "tag": "B-MONEY"},
        {"word": "dolardan", "tag": "O"},
        {"word": "az", "tag": "O"}
      ]
    }
  ]
```

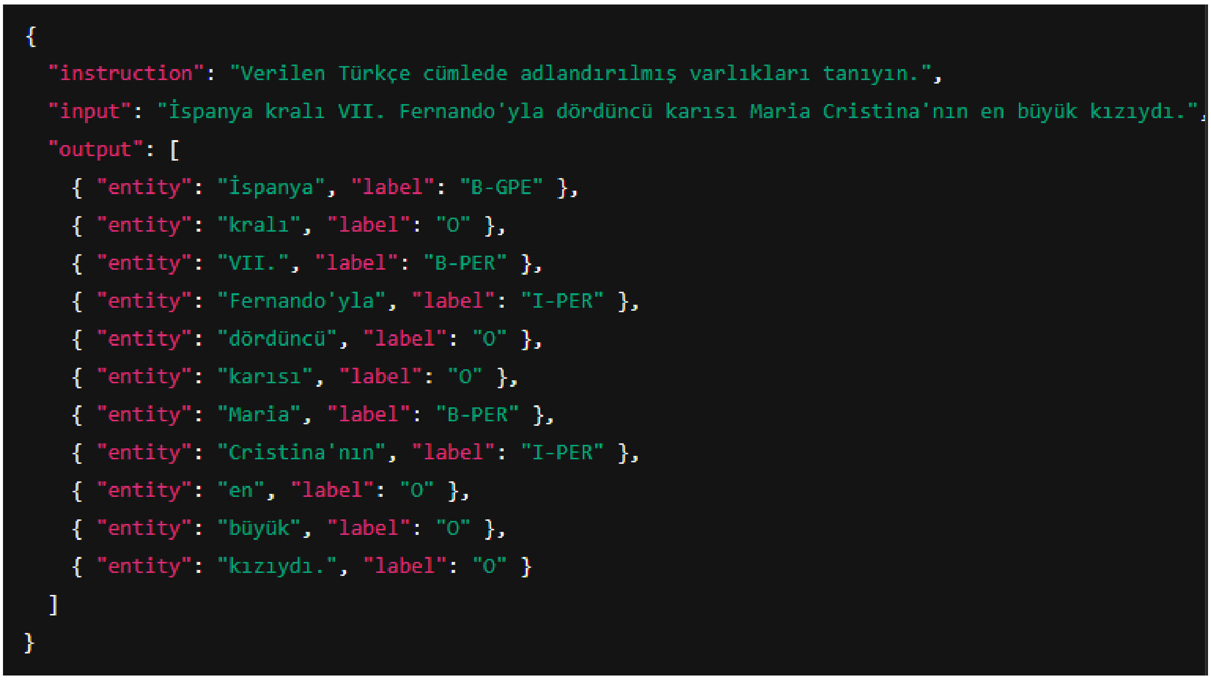
6- NER T5 TURKISH:

The NER-t5-Turkish dataset contains 299,833 unique instances derived from various texts, articles, and documents spanning multiple domains. Each instance is annotated with Named Entity Recognition (NER) tags such as PER for persons, LOC for locations, and ORG for organizations. All instances have been converted into the Alpaca-style format. The conversion into the Alpaca format ensures consistency and ease of integration with other datasets and models.

Below is a sample token/NER tag table:

Token	**Label**
Rodos’u	LOC
Weston	PER
Joost	ORG

The Alpaca format:



```
{
  "instruction": "Verilen Türkçe cümlede adlandırılmış varlıkları tanıyın.",
  "input": "İspanya kralı VII. Fernando'yla dördüncü karısı Maria Cristina'nın en büyük kızıydı.",
  "output": [
    { "entity": "İspanya", "label": "B-GPE" },
    { "entity": "kralı", "label": "O" },
    { "entity": "VII.", "label": "B-PER" },
    { "entity": "Fernando'yla", "label": "I-PER" },
    { "entity": "dördüncü", "label": "O" },
    { "entity": "karısı", "label": "O" },
    { "entity": "Maria", "label": "B-PER" },
    { "entity": "Cristina'nın", "label": "I-PER" },
    { "entity": "en", "label": "O" },
    { "entity": "büyük", "label": "O" },
    { "entity": "kızıydı.", "label": "O" }
  ]
}
```

7- Turkish NER:

The Turkish_ner dataset is an automatically annotated Turkish corpus for named entity recognition and text categorization. It leverages large-scale gazetteers containing approximately 300K entities with thousands of fine-grained entity types distributed. This rich set of annotated entities supports both NER and text categorization tasks.

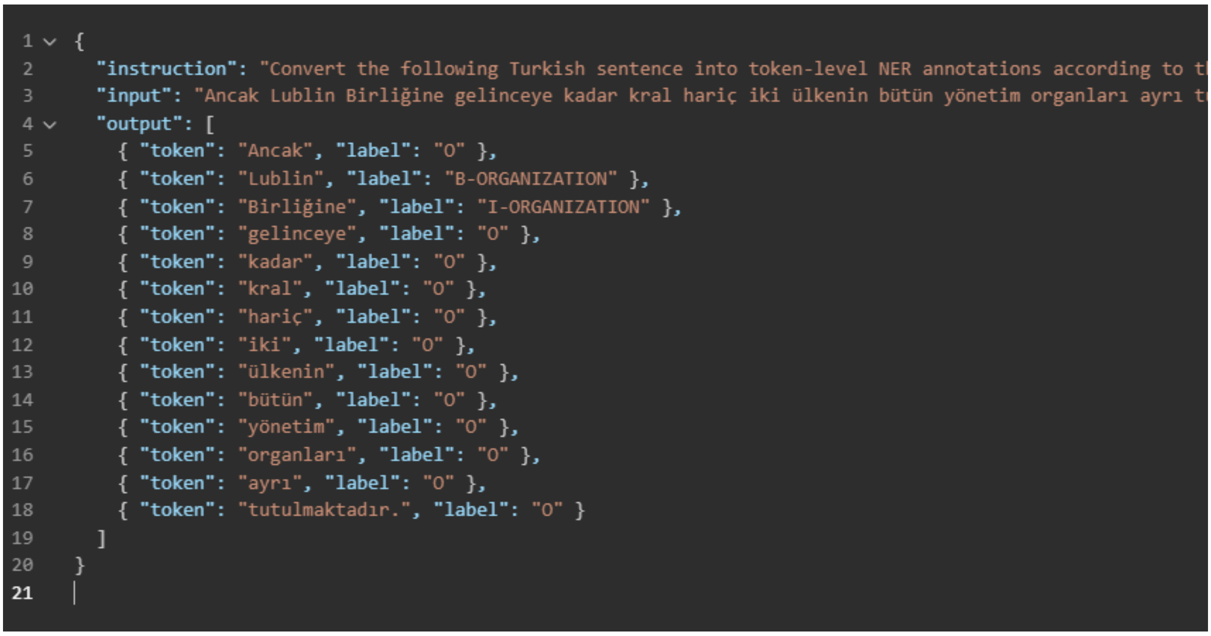
This dataset consists of 25 domains: architecture, basketball, book, business, education, fictional_universe, film, food, geography, government, law, location, military, music, opera, organization, people, religion, royalty, soccer, sports, theater, time, travel, tv

This dataset consist of 10 NER tags: O, B-PERSON, I-PERSON, B-ORGANIZATION, I-ORGANIZATION, B-LOCATION, I-LOCATION, B-MISC, I-MISC

Example instances:

Token	**Label**
Ancak	O
Lublin	B-ORGANIZATION
Birliğine	I-ORGANIZATION
gelinceye	O
kadar	O

The Alpaca format:



```
{
  "instruction": "Convert the following Turkish sentence into token-level NER annotations according to t
  "input": "Ancak Lublin Birliğine gelinceye kadar kral hariç iki ülkenin bütün yönetim organları ayrı t
  "output": [
    { "token": "Ancak", "label": "O" },
    { "token": "Lublin", "label": "B-ORGANIZATION" },
    { "token": "Birliğine", "label": "I-ORGANIZATION" },
    { "token": "gelinceye", "label": "O" },
    { "token": "kadar", "label": "O" },
    { "token": "kral", "label": "O" },
    { "token": "hariç", "label": "O" },
    { "token": "iki", "label": "O" },
    { "token": "ülkenin", "label": "O" },
    { "token": "bütün", "label": "O" },
    { "token": "yönetim", "label": "O" },
    { "token": "organları", "label": "O" },
    { "token": "ayrı", "label": "O" },
    { "token": "tutulmaktadır.", "label": "O" }
  ]
}
```

8- SUNLP Twitter NER:

This dataset is sourced from Turkish tweets and focuses on social media entity recognition. It includes informal language and abbreviations commonly found on Twitter. The dataset contains multiple instances covering various domains, including sports, politics, and entertainment.

The dataset consists of user-generated tweets annotated for entity recognition, making it useful for NLP applications in social media analysis.

Tagset Description

This dataset includes multiple entity categories to support structured analysis of tweets.

- **Named Entity Recognition (NER) Tags:**

This dataset consists of 7 tags. These tags are listed below:

Tag	Description
PERSON	Names of individuals mentioned in tweets.
LOCATION	Cities, countries, and regions referenced in posts.
ORGANIZATION	Sports clubs, companies, and government institutions.
MONEY	Monetary values expressed in the text.
TIME	Temporal expressions such as dates, times of the day, or event durations.
PRODUCT	Consumer products, including technology, food, and fashion items.
TV-SHOW	Names of television programs or series.

Example Instances:

tweet_id	start_pos	end_pos	named_entity_type
1275635208189075457	0	6	PERSON
1351089913861648388	13	20	PERSON
1270134830647345153	0	5	PERSON
1352955129520119809	13	17	ORGANIZATION
1352955129520119809	64	72	PERSON
1281305161181298694	33	39	LOCATION

1305964153451032577	0	5	ORGANIZATION
1305964153451032577	46	64	TVSHOW

5000 tweets used in this dataset were kept with id and labeled. However, the text of the tweets is not included in the dataset. That's why we couldn't create an alpaca for this dataset.

9- NakbaNER:

The Nakba, meaning "catastrophe" in Arabic, refers to the mass displacement and dispossession of Palestinians during the 1948 Arab-Israeli war, when Israel expelled approximately 750,000 Palestinians, forcing them to flee from their homes. The dataset focuses on narratives about the Palestinian Nakba. It comprises 181 news articles from Turkish news agencies Anadolu Ajansı and TRTHaber, totaling 4,032 sentences. These articles include testimonies from Palestinian refugees, highlighting the human impact of the Nakba. The dataset is annotated with 2,289 PERSON, 5,875 LOCATION, and 1,299 ORGANIZATION entities.

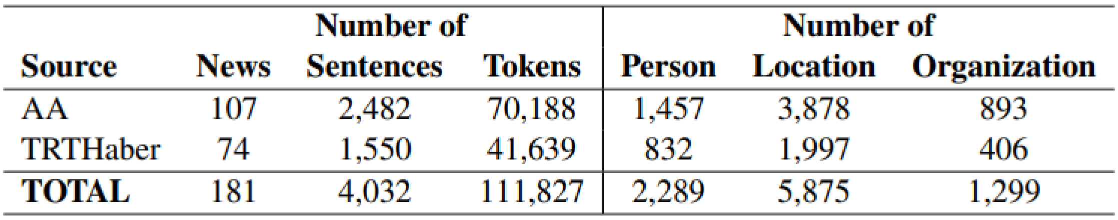
It uses the following entity tags for Named Entity Recognition (NER):

1. **PERSON**: Refers to individuals or groups of people, including both real and fictional persons.

2. **LOCATION**: This encompasses any geographical locations, including **GPE (Geopolitical Entities)**: Countries, cities, and states and **Non-GPE**: Mountain ranges, bodies of water, etc.

3. **ORGANIZATION**: Includes organizations such as companies, political groups, government bodies, and public organizations.

4. **O (Other)**: Words that do not fall into any of the above categories are marked with **O**, indicating that they are not part of any named entity.



		Number of			Number of	
Source	**News**	**Sentences**	**Tokens**	**Person**	**Location**	**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

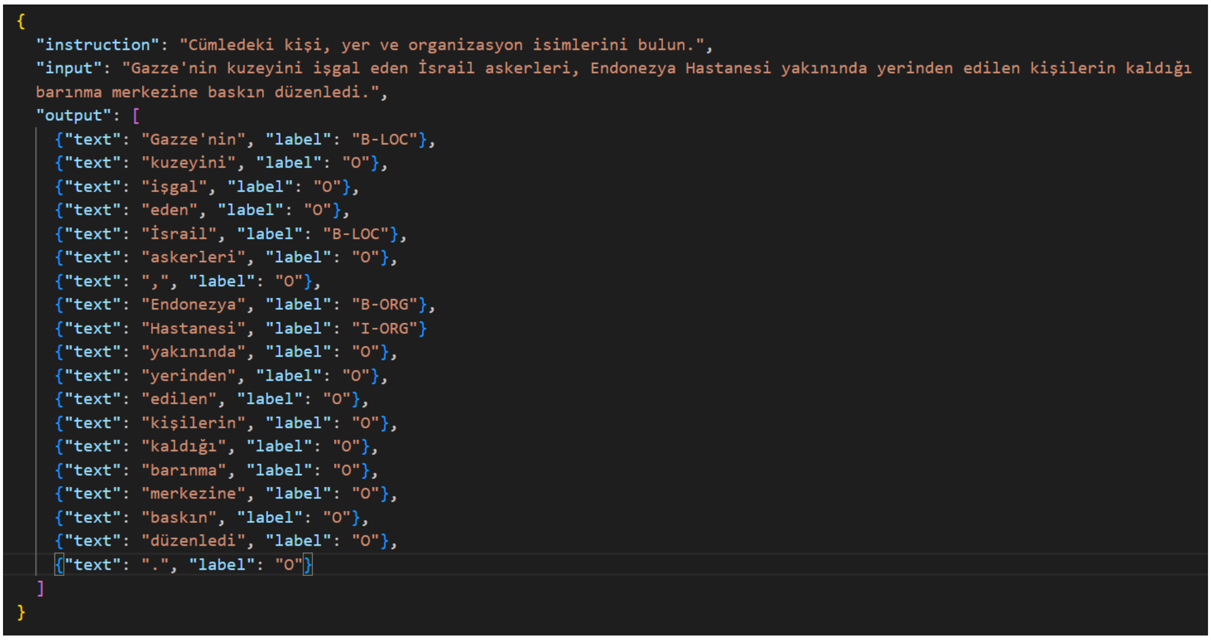
Example instances:

#doc_id =
https://www.aa.com.tr/tr/-haberici/israil-gazzenin-kuzeyindeki-yuzlerce-filistinliyi-silah-tehdidiyle-goce-zorladi-/3369340
#metadata = 21.10.2024
#sent_id = 1
#text = Gazze'nin kuzeyini işgal eden İsrail askerleri, Endonezya Hastanesi yakınında yerinden edilen kişilerin kaldığı barınma merkezine baskın düzenledi.
Gazze'nin B-LOC
kuzeyini O
işgal O
eden O

İsrail B-LOC
askerleri O
, O
Endonezya B-ORG
Hastanesi I-ORG
yakınında O
yerinden O
edilen O
kişilerin O
kaldığı O
barınma O
merkezine O
baskın O
düzenledi O
. O

#sent_id = 2
#text = Askerler buraya sığınan çok sayıda Filistinli erkeği alıkoydu.
Askerler O
buraya O
sığınan O
çok O
sayıda O
Filistinli O
erkeği O
alıkoydu O
. O

Alpaca format:



```
{
  "instruction": "Cümledeki kişi, yer ve organizasyon isimlerini bulun.",
  "input": "Gazze'nin kuzeyini işgal eden İsrail askerleri, Endonezya Hastanesi yakınında yerinden edilen kişilerin kaldığı
  barınma merkezine baskın düzenledi.",
  "output": [
    {"text": "Gazze'nin", "label": "B-LOC"},
    {"text": "kuzeyini", "label": "O"},
    {"text": "işgal", "label": "O"},
    {"text": "eden", "label": "O"},
    {"text": "İsrail", "label": "B-LOC"},
    {"text": "askerleri", "label": "O"},
    {"text": ",", "label": "O"},
    {"text": "Endonezya", "label": "B-ORG"},
    {"text": "Hastanesi", "label": "I-ORG"}
    {"text": "yakınında", "label": "O"},
    {"text": "yerinden", "label": "O"},
    {"text": "edilen", "label": "O"},
    {"text": "kişilerin", "label": "O"},
    {"text": "kaldığı", "label": "O"},
    {"text": "barınma", "label": "O"},
    {"text": "merkezine", "label": "O"},
    {"text": "baskın", "label": "O"},
    {"text": "düzenledi", "label": "O"},
    {"text": ".", "label": "O"}
  ]
}
```

10- TWNERTC and EWNERTC

This dataset is sourced from English and Turkish Wikipedia articles and is designed for Named-Entity Recognition (NER) and Text Categorization tasks. It consists of a large collection of sentences automatically categorized and annotated for NER, which includes named entities across various domains.

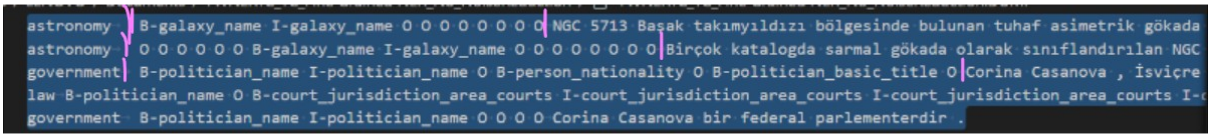
The dataset has around 700,000 sentences for the Turkish version (varies by the version), and 7 million sentences for the English version. There are 2 types of data files, coarse grained and fine-grained. Coarse grained categorises entities into broader categories, while fine-grained is more specific. The dataset for each file is divided into three versions: raw, domain-dependent post-processed, and domain-independent post-processed. There are 2 types of data files, coarse grained and fine-grained. Coarse grained categorises entities into broader categories, while fine-grained is more specific.

Tagset Description

This dataset includes coarse-grained and fine-grained NER tags:

- **Coarse-Grained Named Entity Recognition (NER) Tags:**

For more general classification, the fine-grained types are reduced to broader categories in these tags.



```
astronomy  B-galaxy_name I-galaxy_name O O O O O O O O NGC 5713 Başak takımyıldızı bölgesinde bulunan tuhaf asimetrik gökada
astronomy  O O O O O O B-galaxy_name I-galaxy_name O O O O O O O O Birçok katalogda sarmal gökada olarak sınıflandırılan NGC
government B-politician_name I-politician_name O B-person_nationality O B-politician_basic_title O Corina Casanova , İsviçre
law B-politician_name O B-court_jurisdiction_area_courts I-court_jurisdiction_area_courts I-court_jurisdiction_area_courts I-
government  B-politician_name I-politician_name O O O O Corina Casanova bir federal parlementerdir .
```

- **Fine-Grained Named Entity Recognition (NER) Tags:**

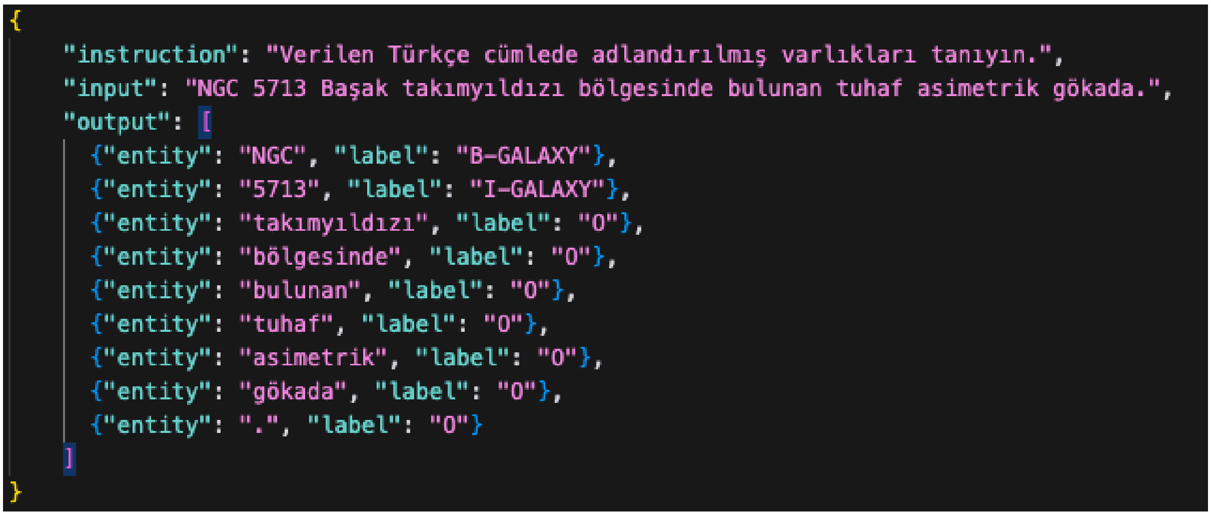
These tags cover a comprehensive set of more than 1000 fine-grained entity types across 77 different domains, with examples including:

Tag	Description
PERSON	Names of individuals.
LOCATION	Cities, countries, and regions.
ORGANIZATION	Sports clubs, companies, institutions.
MISC	Miscellaneous types (e.g., events, titles)

Example Instances:

DUMP FORMAT:
astronomy	B-galaxy_name I-galaxy_name O O O O O O O O	NGC 5713 Başak takımyıldızı bölgesinde bulunan tuhaf asimetrik gökada .

Alpaca format:



```
{
    "instruction": "Verilen Türkçe cümlede adlandırılmış varlıkları tanıyın.",
    "input": "NGC 5713 Başak takımyıldızı bölgesinde bulunan tuhaf asimetrik gökada.",
    "output": [
      {"entity": "NGC", "label": "B-GALAXY"},
      {"entity": "5713", "label": "I-GALAXY"},
      {"entity": "takımyıldızı", "label": "O"},
      {"entity": "bölgesinde", "label": "O"},
      {"entity": "bulunan", "label": "O"},
      {"entity": "tuhaf", "label": "O"},
      {"entity": "asimetrik", "label": "O"},
      {"entity": "gökada", "label": "O"},
      {"entity": ".", "label": "O"}
    ]
}
```

11- HisTR

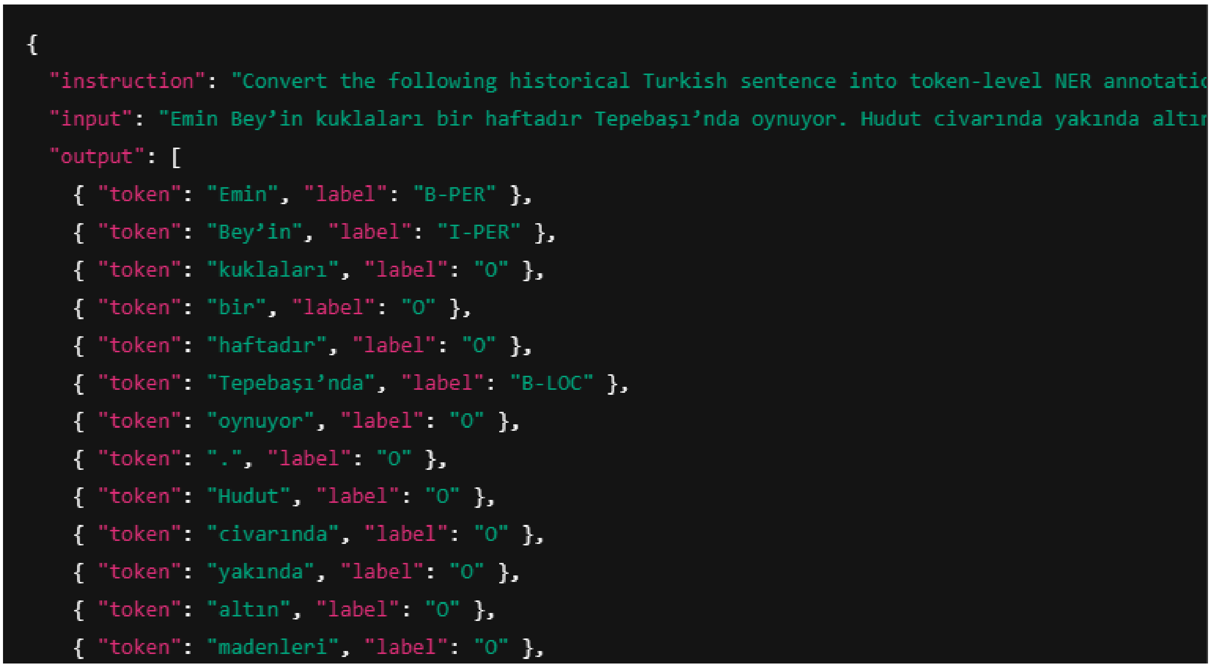
This dataset is Ottoman Turkish NER dataset manually using a subset of sentences from issues of Servet-i Funun journal. It covers a wide range of topics including literature, science, daily life and world news.The original script used in the journal is based on Arabic alphabet while the transcriptions of the sentences are written with the modern Turkish alphabet.

HisTR contains annotations that capture historical named entities, including persons, locations, organizations, and occasionally other entity types that are characteristic of historical narratives. The annotations adhere to a BIO scheme (e.g., B-PER for the beginning of a person name and I-PER for its continuation).

For example, here is a sample from the dataset:

Token	**NER Tag**
Emin	B-PER
Bey’in	I-PER
kuklaları	O
bir	O
haftadır	O
Tepebaşı’nda	B-LOC
oynuyor	O

The Alpaca Format:



```
{
  "instruction": "Convert the following historical Turkish sentence into token-level NER annotatic
  "input": "Emin Bey’in kuklaları bir haftadır Tepebaşı’nda oynuyor. Hudut civarında yakında altır
  "output": [
    { "token": "Emin", "label": "B-PER" },
    { "token": "Bey’in", "label": "I-PER" },
    { "token": "kuklaları", "label": "O" },
    { "token": "bir", "label": "O" },
    { "token": "haftadır", "label": "O" },
    { "token": "Tepebaşı’nda", "label": "B-LOC" },
    { "token": "oynuyor", "label": "O" },
    { "token": ".", "label": "O" },
    { "token": "Hudut", "label": "O" },
    { "token": "civarında", "label": "O" },
    { "token": "yakında", "label": "O" },
    { "token": "altın", "label": "O" },
    { "token": "madenleri", "label": "O" },
```