A Twitter Corpus for Named Entity Recognition in Turkish

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

Introduction

❖ Named Entity Recognition (NER) is the identification of predefined named entities (NEs), such as *Person*, *Location*, and *Organization*.

❖ NER in social media is a subject of study for Turkish since:
- ➢ Previous research is limited to formal texts
- ➢ New NE types have not been studied enough
- ➢ Most of the datasets, especially informal ones, are not publicly available

❖ Our contributions:
- ➢ A new dataset for Turkish NER from Twitter is introduced.
- ➢ Tweets were collected for a year-long period.
- ➢ Understudied named entity (NE) types, namely *Product* and *TV-Show* were included in our label set.
- ➢ High agreement score among multiple annotators was obtained.
- ➢ Initial results on this dataset were reported.
- ➢ The dataset is shared publicly at *https://github.com/SU-NLP/SUNLP-Twitter-NER-Dataset.*

Related Work

❖ The first Turkish NER dataset (Tür et al., 2003), the largest one with 500K tokens, consists of news articles.
❖ A later study (Çelikkaya et al., 2013) introduced the relatively small first informal datasets.
❖ In another work (Tantuğ, 2015), the authors created the largest dataset in informal texts from Twitter.
❖ A recent work (Şeker and Eryiğit, 2017) introduced a dataset on user-generated content, such as customer reviews and blogs.

Data Collection

❖ Tweets were collected through the Twitter streaming API from June 2020 to June 2021.
❖ Using top trends in Turkey, we obtained 65 million tweets.

❖ Steps in selecting tweets to be annotated:
- ❖ Tweets have the same content without Twitter-specific artifacts were removed.
- ❖ Only tweets longer than 50 characters were kept.
- ❖ An effective NER model was used to select tweets that have at least one unseen NE.
- ❖ To ensure diversity of topics, any one hashtag can be in a maximum of 3 tweets.
- ❖ Finally, random 5000 tweets were selected to be annotated.

❖ The dataset contains 126,228 words, with an average of 25,24 words per tweet.

Annotation Process

❖ Annotated NE types in this dataset:
Person, Location, Organization, Time, Money, Product, TV-Show

❖ Our annotation team is 4 undergraduate students whose native language is Turkish.
❖ Each tweet was annotated by 2 annotators.
❖ A hashtag was also annotated if it is a NE as a whole.
❖ The inter-annotator agreement is 0.87 Cohen Kappa score (w/o *Other*).
❖ The most disagreed situations are:

1. Organization vs. Location
   Loc: *We are going to the beautiful beaches of Turkey on vacation in summer.*
   Org: *Negotiations between Turkey and the USA continue.*

2. Organization vs. Product
   Product: *If you are not sure, just ask it to Google.*
   Org: *I will start working at Google starting next month ☺*

NE Distributions

NE Type	Count
Person	5,526
Organization	2,956
Location	1,243
Time	608
Product	334
TV-Show	255
Money	159
Total	**11,081**

❖ Total number of NE is 11,081 with unique 7,231.
❖ Common NE types are also the most frequent ones.
❖ *Person* is the most frequent one.
❖ *Organization* and *Location* are second and third, respectively.
❖ *Product* and *TV-Show* categories are low in frequency, but *TV-Show* can be considered as a type of *Product*.

NER Model

❖ Results were reported on validation and test set, consisting of 750 randomly sampled tweets.
❖ The remaining 3,500 were used for training.
❖ Our evaluation metrics are *F1-score*, *Precision*, and *Recall* computed for the entire NE spans.
❖ Our baseline models are variations of transformer-based models:
- ❖ Turkish models: BERTurk, BERT_loodos, and ALBERT_loodos
- ❖ Multilingual models: mBERT, XLM-RoBERTa

❖ Turkish models differ according to the corpora used.
- ❖ BERTurk corpora: contains structural texts that have fewer grammatical and spelling errors.
- ❖ Loodos corpora: contains informal texts such as Twitter and online blogs.

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

Experiments

Transformer-based models

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

❖ All Turkish pre-trained models except for ALBERT gave better results than multilingual models.
❖ BERT_loodos model consistently outperformed.

Formal vs. Informal Training Sets

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

❖ Models trained with our dataset outperformed the larger and well-studied Turkish NER dataset (Tür et al., 2003).

F1-Score on Each NE

NE Class	BERTurk	BERT_loodos	mBERT	XLM-RoBERTa
Person	0.87	**0.88**	0.80	0.83
Location	0.77	**0.81**	0.64	0.67
Organization	0.77	**0.80**	0.72	0.75
Time	0.89	**0.90**	0.86	0.88
Product	0.32	0.37	**0.52**	0.46
TV-Show	0.49	**0.57**	0.52	0.49
Money	0.88	**0.93**	0.75	0.85

❖ BERT_loodos outperformed in all classes except *Product*.
❖ The multilingual models achieved better results in this category.

Conclusion

❖ A new Turkish NER dataset was introduced and shared publicly.
❖ New categories were included, such as *Product* and *TV-Show*.
❖ A BERT model pre-trained on a blend of formal and informal texts in Turkish obtained the highest score.
❖ Model trained with our dataset outperformed the most studied and largest dataset (Tür et al., 2003)

References

❖Tür, G., Hakkani-Tür, D., and Oflazer, K. (2003). A statistical information extraction system for Turkish. *In Natural Language Engineering*, pages 181–210.
❖Çelikkaya, G., Torunoğlu, D., and Eryiğit G. (2013). Named Entity Recognition on Real Data: A Preliminary Investigation for Turkish. In *2013 7th International Conference on Application of Information and Communication Technologies*, pages 1-5. IEEE.
❖Eken B., Tantuğ, A.C. (2015). Recognizing Named Entities in Turkish Tweets. In *Proceedings of the Fourth International Conference on Software Engineering and Applications*.
❖Şeker. G. A. and Eryiğit G. (2017). Extending a CRF-based Named Entity Recognition Model for Turkish Well-Formed Text and User Generated Content. *Semantic Web*, pages 625-642