A Twitter Corpus for Named Entity Recognition in Turkish
Buse Çarık, Reyhan Yeniterzi
Sabancı 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/SUNLP/SUNLP-Twitter-NER-Dataset.

Related Work

- The first Turkish NER dataset (Tüर 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 dataset.
- In another work (Tantūğ, 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

<table>
<thead>
<tr>
<th>NE Type</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person</td>
<td>5,526</td>
</tr>
<tr>
<td>Organization</td>
<td>2,956</td>
</tr>
<tr>
<td>Location</td>
<td>1,243</td>
</tr>
<tr>
<td>Time</td>
<td>608</td>
</tr>
<tr>
<td>Product</td>
<td>334</td>
</tr>
<tr>
<td>TV-Show</td>
<td>255</td>
</tr>
<tr>
<td>Money</td>
<td>159</td>
</tr>
<tr>
<td>Total</td>
<td>11,081</td>
</tr>
</tbody>
</table>
- 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 evaluation 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

<table>
<thead>
<tr>
<th>Model</th>
<th>Recall</th>
<th>Precision</th>
<th>F1-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERTurk</td>
<td>84.31</td>
<td>80.24</td>
<td>81.32</td>
</tr>
<tr>
<td>BERT_loodos</td>
<td>84.99</td>
<td>80.00</td>
<td>84.49</td>
</tr>
<tr>
<td>ALBERT_loodos</td>
<td>71.81</td>
<td>74.05</td>
<td>73.24</td>
</tr>
<tr>
<td>mBERT</td>
<td>78.95</td>
<td>76.61</td>
<td>77.46</td>
</tr>
<tr>
<td>XLM-RoBERTa</td>
<td>81.39</td>
<td>82.76</td>
<td>82.76</td>
</tr>
</tbody>
</table>
- All Turkish pre-trained models except ALBERT gave better results than multilingual models.
- BERT_loodos model consistently outperformed.

Formal vs. Informal Training Sets

<table>
<thead>
<tr>
<th>Model</th>
<th>Train Data</th>
<th>Recall</th>
<th>Precision</th>
<th>F1-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERTurk</td>
<td>Our Train</td>
<td>86.90</td>
<td>84.53</td>
<td>85.68</td>
</tr>
<tr>
<td>BERTurk</td>
<td>[Tür et al., 2003]</td>
<td>68.87</td>
<td>69.02</td>
<td>69.92</td>
</tr>
<tr>
<td>BERT_loodos</td>
<td>Our Train</td>
<td>89.64</td>
<td>85.70</td>
<td>85.63</td>
</tr>
<tr>
<td>BERT_loodos</td>
<td>[Tür et al., 2003]</td>
<td>68.43</td>
<td>68.82</td>
<td>69.48</td>
</tr>
</tbody>
</table>
- Models trained with our dataset outperformed the larger and well-studied Turkish NER dataset (Tür et al., 2003).

FI-Score on Each NE

<table>
<thead>
<tr>
<th>NE Class</th>
<th>BERTurk</th>
<th>BERT_loodos</th>
<th>ALBERT_loodos</th>
<th>mBERT</th>
<th>XLM-RoBERTa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person</td>
<td>0.87</td>
<td>0.86</td>
<td>0.80</td>
<td>0.83</td>
<td>0.83</td>
</tr>
<tr>
<td>Location</td>
<td>0.77</td>
<td>0.81</td>
<td>0.64</td>
<td>0.67</td>
<td>0.67</td>
</tr>
<tr>
<td>Organization</td>
<td>0.77</td>
<td>0.80</td>
<td>0.72</td>
<td>0.75</td>
<td>0.75</td>
</tr>
<tr>
<td>Time</td>
<td>0.89</td>
<td>0.90</td>
<td>0.86</td>
<td>0.88</td>
<td>0.88</td>
</tr>
<tr>
<td>Product</td>
<td>0.33</td>
<td>0.37</td>
<td>0.32</td>
<td>0.46</td>
<td>0.46</td>
</tr>
<tr>
<td>TV-Show</td>
<td>0.49</td>
<td>0.37</td>
<td>0.52</td>
<td>0.49</td>
<td>0.49</td>
</tr>
<tr>
<td>Money</td>
<td>0.88</td>
<td>0.93</td>
<td>0.75</td>
<td>0.85</td>
<td>0.85</td>
</tr>
</tbody>
</table>
- 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