Named Entity Recognition on Turkish Tweets

Dilek Küçük, Guillaume Jacquet, Ralf Steinberger

European Commission, Joint Research Centre
Via E. Fermi 2749
21027 Ispra (VA), Italy
firstname.lastname@jrc.ec.europa.eu

Abstract

Various recent studies show that the performance of named entity recognition (NER) systems developed for well-formed text types drops significantly when applied to tweets. The only existing study for the highly inflected agglutinative language Turkish reports a drop in F-Measure from 91% to 19% when ported from news articles to tweets. In this study, we present a new named entity-annotated tweet corpus and a detailed analysis of the various tweet-specific linguistic phenomena. We perform comparative NER experiments with a rule-based multilingual NER system adapted to Turkish on three corpora: a news corpus, our new tweet corpus, and another tweet corpus. Based on the analysis and the experimentation results, we suggest system features required to improve NER results for social media like Twitter.

Keywords: named entity recognition, Turkish, tweets

1. Introduction

Named entity recognition (NER) is an information extraction task which targets at the recognition of person, location, and organization names. Although NER is considered a solved problem for well-studied languages like English, the topic still requires attention for languages like Turkish for which language processing tools, related resources, and data sets are comparatively rare. Additionally, most of the existing NER systems are built for well-formed text types such as news articles and these systems are reported to perform poorly on informal text types such as tweets. Since the need for information extraction from these informal text types increases significantly, there are several recent studies presenting NER approaches on tweets.

In this study, we present a tweet data set in Turkish annotated with named entities (NEs) and our first NER experiments on this data set. The experiments are performed using a rule-based system for Turkish which is an adaptation of a multilingual NER system (Pouliquen and Steinberger, 2009) to Turkish by integrating the required linguistic resources for Turkish to the system. We also present the peculiarities of the tweet texts in Turkish and an error analysis of the evaluation results. In the following sections, we present related work regarding NER on tweets and on Turkish texts (Section 2), we describe our NE-annotated Turkish tweet data set (Section 3) and we present the evaluation results of our NER experiments on the data set together with discussions (Section 4). Finally, we conclude the paper with a summary of main points (Section 5).

2. Related Work

There are several recent studies targeting NER on tweets. A two-phase NER system for tweets called T-NER is presented in (Ritter et al., 2011), which employs Conditional Random Fields (CRFs) and labeled latent Dirichlet allocation (Ramage et al., 2009). In (Liu et al., 2011), a hybrid NER approach, based on k-Nearest Neighbors and CRF classifiers, is proposed. A factor graph-based approach that performs NER and NE normalization is proposed in (Liu et al., 2012). An unsupervised NE segmentation approach called TwiNER for targeted tweets is presented in (Li et al., 2012). In (Jung, 2012), a clustering-based approach for NER on microtexts is described. A lightweight filter-based approach, called FS-NER, is described in (de Oliveira et al., 2013). NER experiments on targeted tweets in Polish are presented in (Piskorski and Ehrmann, 2013). Finally, an adaptation of the ANNIE component of the GATE framework to microblog texts, called TwitIE, is presented in (Bontcheva et al., 2013). The NER component of TwitIE is reported to outperform T-NER (Ritter et al., 2011).

Regarding NER on Turkish, a statistical NER system based on Hidden Markov Models is proposed in (Tür et al., 2003). In (Küçük and Yazıcı, 2009) a rule-based NER system is presented and a hybrid NER system based on this rule-based system is described in (Küçük and Yazıcı, 2012). A rule-learning approach for NER in Turkish is presented in (Tatar and Çicekli, 2011). Finally, CRF-based approaches for NER in Turkish are proposed in (Yeniterzi, 2011; Şeker and Eryiğit, 2012). These NER systems for Turkish are mostly proposed and tested on news articles and the CRF-based system in (Şeker and Eryiğit, 2012) is reported to outperform the other proposals. To the best of our knowledge, the only study addressing NER on Turkish tweets is presented in (Çelikkaya et al., 2013) where the CRF-based NER system trained on news articles (Şeker and Eryiğit, 2012) is evaluated on a tweet data set.

3. Turkish Tweet Data Set

Our Turkish tweet data set comprises tweets that are dated July 26, 2013 between 12:00 and 13:00 GMT. It has a total of 2,320 tweets after some data cleaning operations like deleting non-Turkish tweets and retweets. The total number of words is 20,752 and hence the average number of words per tweet is about 8.9.

The data set is annotated with the NEs of types person, location, organization names (henceforth, PLOs), date/time (temporal) and money/percent (numeric) expressions. These seven types of NEs are denoted as `PERSON`, `LOCATION`, `ORGANIZATION`, `DATE`, `TIME`, `MONEY`, and `PERCENT` during the annotation process, following the types considered in the Message Understanding Conference series (Grishman and Sundheim, 1996). Additionally, some other prevalent NE types such as the names of TV programs/series, movies, music bands, and products within this data set are also annotated. The type of this latter set of NEs is denoted as `MISC` during the annotation process. The overall NE annotations on this tweet set together with the corresponding tweet ids are made publicly available at `http://optima.jrc.it/Resources/2014_JRC_Twitter_TR_NER-dataset.zip`.

Tweet that Include ...	**Number of Tweets (Percentage over the Total of 2,320 Tweets)**
PLOs	670 (28.9%)
Date/Time Expressions	187 (8.1%)
Money/Percent Expressions	22 (0.9%)
Any of Basic Seven Types	**814 (35.1%)**
Other NEs (of `MISC` type)	96 (4.1%)
Any of Annotated Types	**868 (37.4%)**

Table 1: Frequencies of Tweets Including NEs in the Turkish Tweet Data Set.

Table 1 presents statistical information on the Turkish tweet data set, as related to NEs. As demonstrated in the last row of the table, about 37.4% of the tweets include at least one of the considered eight types of NEs and about 35.1% of them include at least one of the basic seven types. Since a tweet may include more than one NE and each NE may be of a different type, the sets within rows 1–3 and 5 are not disjoint.

Table 2 presents the distribution of the annotated NEs in the data set. Deictic date expressions like *bugün* ('today'), *yarın* ('tomorrow'), and *şimdi* ('now') are also annotated, since they can be resolved to actual dates using the dates of the tweets as the reference dates. The annotation of such expressions in addition to those less frequent counterparts like *22.02.2012* accounts for the considerably large number of date expressions within the data set.

Another important aspect of the tweet data set as related to NEs is the appearance of some NE instances within hashtags. The data set comprises a total of 153 hashtags, 70 of which include annotated NEs (52 of them being PLOs) and hence covered by the statistics provided in Tables 1 and 2. Sample hashtags including NEs are *#Yenisezonforman**TurkTelekom**dan* and *#nobeticin**ankara**dayiz*, where the part written in boldface in the first example is an organization name and the one in the second example is a location name. In 14 of these 70 hashtags, the included NE covers the whole hashtag where a sample case is ***#MustafaKemalAtatürk*** which denotes a person name. A

NE Type	**Count**
Person	457
Location	282
Organization	241
Total for the PLOs	**980**
Date	201
Time	5
Money	16
Percent	9
Total for the Basic Seven Types	**1,211**
Other NEs (of `MISC` type)	111
Total for All Annotated Types	**1,322**

Table 2: Distribution of NEs in the Turkish Tweet Data Set.

related point to note is that there are 31 multi-token NEs annotated in the data set which are not part of any hashtags but appear like hashtags in tweets with their whitespaces removed, such as the person name instances of *MuratBoz* and *SabriSarıoğlu*, which should have been written as *Murat Boz* and *Sabri Sarıoğlu*, respectively. Hence, 101 of all 1,322 NEs (7.6%) in the tweet data set either appear in hashtags or hashtag-like with their whitespaces removed.

Below listed are some common phenomena observed in the tweets within the data set as related to NEs and usually as opposed to other well-formed text types like news articles. The ones peculiar to Turkish are written in boldface.

- Grammar/writing/spelling errors such as:
  - not capitalizing the initial letters of PLO tokens,
  - **not separating names from suffixes with apostrophes**,
  - modifying names to stress or show affection like repeating some characters,
  - **utilizing the corresponding non-accentuated characters (c, g, i, o, s, and u) instead of Turkish characters with their correct diacritics (ç, ğ, ı, ö, ş, and ü)**.
- Lack of important contextual clues for NER (such as person titles/professions), mostly due to the character limitation of the tweets.
- Referring to location and organization names in contracted forms or using metonymic expressions, like referring to *Facebook* as *face* or to *Boğaziçi Üniversitesi* ('Boğaziçi University') as *boğaziçi* only.
- Appearance of person names as single forenames, surnames, or nicknames.
- Appearance of some NEs in hashtags.

4. Named Entity Recognition Experiments

In this section, we present our NER experiments on our Turkish tweet data set, in addition to comparative evaluations on a news data set and another tweet data set. In the first subsection, we briefly describe the NER system

employed and present the evaluation results of the system on these three data sets. In the following subsection, we discuss these evaluation results and perform an error analysis with respect to related tweet-specific phenomena and finally outline some suggestions to improve NER performance on Turkish tweets in the last subsection.

4.1. Evaluation Results

We perform our NER experimentation using the NER software (Pouliquen and Steinberger, 2009) of the *Europe Media Monitor* (EMM) multilingual media analysis and information extraction system. EMM processes approximately 175,000 news articles per day in up to 75 languages from about 4,000 news sources (Steinberger, 2013). EMM's rule-based NER system mostly employs language-independent rules that make reference to language-specific dictionary lists to recognize PLOs and considers only those candidate tokens which have their initial letters capitalized. The system can be adapted to a new language by providing for that language separate word lists containing titles, professions, demonyms, modifiers, and various types of stop words, among others. It is tailored to Turkish by equipping it with the required lists for Turkish information extraction, including lists of common PLOs and organization endings in Turkish.

Table 3 presents the initial performance evaluation results of EMM's NER system for Turkish while Table 4 presents the corresponding results after the system is modified with the below extensions, to improve its overall performance, especially on tweets:

- The initial form of EMM's NER system considers only multi-token candidates during person name recognition because it focuses on high recall at document level only and at least two name parts are needed to ground the name mention to a real-world entity. But, both in formal text types like news articles and informal ones like tweets, person names can appear as single tokens such as single surnames (common in news articles) and single forenames (more common in tweets). Yet, relaxing this multi-token constraint of the NER system to increase its coverage can lead to many false positives as single names in Turkish are often homonymous to common nouns. Hence, we have extended the resources of the system with a list of 2,670 entries, which are individual tokens (name parts) of the person names that had previously appeared at least 30 times in Turkish news articles, as obtained from the EMM database of names. We should note that this resource includes both Turkish and foreign person names.

- We have also extended the list of organization names used by the system with a list of about 550 organization names obtained from the EMM database. This additional list includes Turkish and foreign organization names that have appeared in Turkish news articles at least 30 times, similar to the first extension procedure described above.

Data Set	Metric	Per.	Loc.	Org.	PLOs
News	**P (%)**	82.00	93.66	93.89	91.27
Data Set	**R (%)**	31.70	54.29	48.21	46.12
	F (%)	**45.72**	**68.74**	**63.70**	**61.28**
Tweet	**P (%)**	61.94	73.53	88.73	72.14
Data Set	**R (%)**	18.16	35.46	26.14	25.10
	F (%)	**28.09**	**47.85**	**40.38**	**37.24**
Tweet	**P (%)**	48.28	70.80	92.37	71.39
Data Set	**R (%)**	7.24	32.92	33.89	18.70
ÇTE2013	**F (%)**	**12.58**	**44.94**	**49.59**	**29.64**

Table 3: Initial Evaluation Results of EMM's NER System for Turkish on Different Data Sets.

Data Set	Metric	Per.	Loc.	Org.	PLOs
News	**P (%)**	75.00	94.51	87.04	85.03
Data Set	**R (%)**	70.36	54.29	52.69	58.22
	F (%)	**72.61**	**68.97**	**65.64**	**69.12**
Tweet	**P (%)**	56.49	74.07	78.72	66.03
Data Set	**R (%)**	29.54	35.46	30.71	31.53
	F (%)	**38.79**	**47.96**	**44.18**	**42.68**
Tweet	**P (%)**	52.53	71.17	89.36	66.80
Data Set	**R (%)**	17.44	32.51	35.29	24.75
ÇTE2013	**F (%)**	**26.19**	**44.63**	**50.60**	**36.11**

Table 4: Evaluation Results of the Modified Version of EMM's NER System for Turkish on Different Data Sets.

The evaluation results in Tables 3-4 are provided in terms of precision (P (%)), recall (R (%)) and balanced F-Measure (F (%)) without giving credit to partial extractions. The columns 3-5 include the evaluation results for person, location and organization names, respectively, while the last column includes the overall results for the PLOs. The first three rows of the tables show the evaluation results on the news article data set employed in (Küçük and Yazıcı, 2009) where the articles within the set have been compiled from the METU Turkish corpus (Say et al., 2002). This news data set comprises about 20,130 tokens, with 1,613 annotated NEs belonging to one of the basic seven types, 1,405 of which are PLOs (Küçük and Yazıcı, 2009). Hence, this set can be considered as comparable in size and in the number of PLOs to our tweet data set. Rows 4-6 of Tables 3-4 include the corresponding results on our Turkish tweet data set presented in Section 3. Finally, the last three rows of the tables correspond to the evaluation results of the system on the tweet data set given in (Çelikkaya et al., 2013), which is denoted as *Tweet Data Set ÇTE2013*. This tweet data set includes about 54,000 tokens and 1,374 annotated PLOs.

4.2. Discussion of the Results

On all of the data sets, the precision rates of the initial and modified versions of the system are good while the corresponding recall and accordingly F-Measure values are comparatively lower. The results in Tables 3-4 show that the resource extensions to arrive at the modified system improve the recall rates significantly for all data sets while decreasing the precision rates. As the increases in recall are considerably larger than the corresponding decreases in precision, the F-Measure values of the modified system are higher.

On the news data set, the precision rates are particularly high with the overall values of 91.27% and 85.03% achieved for all PLOs by the initial and modified versions of the system, respectively. However, especially the initial version of the system suffers from poor coverage revealed as low recall rates, with an overall value of 46.12%, which in turn lead to an F-Measure rate of 61.28% for all PLOs. In the context of a NER evaluation campaign, these recall and F-Measure rates of the initial version of the system are not good but within the context of the EMM system, which daily processes a huge amount of redundant news articles and serves a large number of users, the main focus is to achieve high precision. Hence, the initial version can be considered more convenient in this application setting compared to the modified version which performs better in terms of recall, with an overall value of 58.22%, and in terms of F-Measure, with an overall value of 69.12%, on the news data set.

The F-Measure rate (69.12%) of the modified system on the news data set is also promising since the system is an initial adaptation of a generic multilingual NER system and does not perform deep language processing. A previously proposed rule-based NER system (Küçük and Yazıcı, 2009), which also considers some temporal/ numeric expressions in addition to PLOs, achieves an overall F-Measure of 78.7% on the same news data set. The main differences between EMM's NER system and the latter system are that the latter system extracts the temporal/numeric expressions with decent performance, it employs a morphological analyser to also recognize NE's missing apostrophes to separate them from suffixes, and it considers single tokens as person name candidates without restriction.

For both of the tweet data sets, a comparison of the performance rates in Tables 3 and 4 reveals that the performance of EMM's modified NER system is considerably better than the performance of its initial version. This observation confirms that extending the related resources of the system with frequently appearing person name parts and organization names in news articles clearly improves the NER performance on tweets. Hence, we discuss the evaluation results of the modified version of the NER system on the tweet data sets in the rest of this section. The overall F-Measure of EMM's modified NER system drops from 69.12% to 42.68% (Table 4) when ported from news articles to our tweet data set presented in Section 3. This is an expected result due to the following characteristics of the tweet data set:

- For 246 of the 980 PLOs (25.1%), the initial letters of each of their tokens are not properly capitalized and hence these NEs are missed by the system.
- Only 147 of the 457 person names (32.17%) are composed of forename-surname pairs while the remaining ones are single names. This leads to a considerable drop in the recall of person name recognition (70.36% to 29.54%) since EMM's NER system mostly requires person names to span more than one token. It only recognizes single names if they are included within the list of 2,670 entries corresponding to the names and surnames of frequently appearing person names in Turkish news articles, as described earlier in this section.
- Common writing errors, like using the corresponding non-accentuated characters instead of Turkish characters with proper diacritics, are observed in 10% of the 980 PLOs and these PLOs cannot be extracted by the system. PLOs appearing in hashtags or hashtag-like and PLOs lacking apostrophes that should separate them from the suffixes are again missed by the system.

The F-Measure of 36.11% achieved by the system on *Tweet Data Set ÇTE2013*[1] can be considered comparable to that of 42.68% obtained on our tweet data set. The CRF-based NER system for Turkish (Şeker and Eryiğit, 2012), which achieves an F-Measure of 91.64% on news articles, achieves that of 19.28% on *Tweet Data Set ÇTE2013* (Çelikkaya et al., 2013). Major differences between *Tweet Data Set ÇTE2013* and our data set are that the former includes some non-Turkish tweets and it is not a publicly available data set. Non-Turkish tweets have been manually filtered out from our tweet data set.

4.3. Suggestions for Improvement

Based on the above discussions, the following suggestions can be employed to improve NER performance on Turkish tweets.

- NE candidates without proper capitalization can also be considered since such NEs are common in tweets.
- As single person names are frequent in tweets, the NER procedure can further be relaxed to cover them. Yet, as several person names in Turkish are also common nouns, combined with the capitalization relaxation item above, this may lead to many false positives.
- Since apostrophes separating the PLOs and the attached suffixes are usually missed in Turkish tweets, a morphological analysis procedure can be performed to detect such NEs.
- The effects of writing/spelling errors and contractions can be alleviated by employing a normalization/ correction procedure on the input prior to the NER procedure, as pointed out in studies such as (Liu et al., 2011; Liu et al., 2012).
- As NEs can appear in hashtags or hashtag-like with their whitespaces removed within tweets, a NER system for tweets can consider such cases to increase its coverage on tweets.

[1] In *Tweet Data Set ÇTE2013*, about 50 location names have metonymic readings in which organization names are referred to, like referring to a sports club with the name of its home city. When organization names, instead of location names, are considered as the correct types for these NEs, the performance of the NER system on this data set is 33.56% in F-Measure.

- Following the NER systems with significant accuracies on tweets in English (like T-NER (Ritter et al., 2011) and TwitIE (Bontcheva et al., 2013)), learning/rule-based approaches utilizing tools like part-of-speech taggers tailored to tweets can be implemented.

5. Conclusion

In this study, we present a tweet data set in Turkish which is annotated with named entities and the corresponding annotations are made publicly available for research purposes. After providing statistical information on the tweet data set, we present the results of our first NER experiments on this set using a NER system initially engineered for news articles, in addition to comparative experiments on a news corpus and another tweet data set. We perform similar NER experiments on these three data sets after extending the system's resources with lists of person and organization names frequently appearing in recent news articles in Turkish. Based on these evaluation results and our detailed analysis of the tweet-specific linguistic phenomena with their frequencies, we suggest some system features to improve the NER performance on social media texts in Turkish.

6. Acknowledgements

This study is supported in part by a postdoctoral research grant from TÜBİTAK.

7. References

Kalina Bontcheva, Leon Derczynski, Adam Funk, Mark Greenwood, Diana Maynard, and Niraj Aswani. 2013. TwitIE: An Open-Source Information Extraction Pipeline for Microblog Text. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*.

Gökhan Çelikkaya, Dilara Torunoğlu, and Gülşen Eryiğit. 2013. Named Entity Recognition on Real Data: A Preliminary Investigation for Turkish. In *Proceedings of the 7th International Conference on Application of Information and Communication Technologies*.

Gökhan A. Şeker and Gülşen Eryiğit. 2012. Initial Explorations on Using CRFs for Turkish Named Entity Recognition. In *Proceedings of the International Conference on Computational Linguistics*, pages 2459–2474.

Diego Marinho de Oliveira, Alberto HF Laender, Adriano Veloso, and Altigran S da Silva. 2013. FS-NER: A Lightweight Filter-Stream Approach to Named Entity Recognition on Twitter Data. In *Proceedings of the 22nd International Conference on World Wide Web Companion*, pages 597–604.

Ralph Grishman and Beth Sundheim. 1996. Message Understanding Conference–6: A Brief History. In *Proceedings of the International Conference on Computational Linguistics*, volume 96, pages 466–471.

Jason J. Jung. 2012. Online Named Entity Recognition Method for Microtexts in Social Networking Services: A Case Study of Twitter. *Expert Systems with Applications*, 39(9):8066–8070.

Dilek Küçük and Adnan Yazıcı. 2009. Named Entity Recognition Experiments on Turkish Texts. In T. Andreasen et al., editor, *Proceedings of the International Conference on Flexible Query Answering Systems*, volume 5822 of *LNCS*, pages 524–535.

Dilek Küçük and Adnan Yazıcı. 2012. A Hybrid Named Entity Recognizer for Turkish. *Expert Systems with Applications*, 39(3):2733–2742.

Chenliang Li, Jianshu Weng, Qi He, Yuxia Yao, Anwitaman Datta, Aixin Sun, and Bu-Sung Lee. 2012. TwiNER: Named Entity Recognition in Targeted Twitter Stream. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 721–730.

Xiaohua Liu, Shaodian Zhang, Furu Wei, and Ming Zhou. 2011. Recognizing Named Entities in Tweets. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 359–367.

Xiaohua Liu, Ming Zhou, Furu Wei, Zhongyang Fu, and Xiangyang Zhou. 2012. Joint Inference of Named Entity Recognition and Normalization for Tweets. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 526–535.

Jakub Piskorski and Maud Ehrmann. 2013. On Named Entity Recognition in Targeted Twitter Streams in Polish. In *Proceedings of the ACL Workshop on Balto-Slavic Natural Language Processing*.

Bruno Pouliquen and Ralf Steinberger. 2009. Automatic Construction of Multilingual Name Dictionaries. In C. Goutte et al., editor, *Learning Machine Translation*, Advances in Neural Information Processing Systems Series, pages 59–78. MIT Press.

Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D Manning. 2009. Labeled LDA: A Supervised Topic Model for Credit Attribution in Multi-labeled Corpora. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 248–256.

Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. 2011. Named Entity Recognition in Tweets: An Experimental Study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1524–1534.

Bilge Say, Deniz Zeyrek, Kemal Oflazer, and Umut Özge. 2002. Development of a Corpus and a Treebank for Present-Day Written Turkish. In *Proceedings of the 11th International Conference of Turkish Linguistics*.

Ralf Steinberger. 2013. Multilingual and Cross-Lingual News Analysis in the Europe Media Monitor (EMM). In M. Lupu et al., editor, *Proceedings of the 6th Information Retrieval Facility Conference*, volume 8201 of *LNCS*, pages 1–4.

Serhan Tatar and İlyas Çicekli. 2011. Automatic Rule Learning Exploiting Morphological Features for Named Entity Recognition in Turkish. *Journal of Information Science*, 37(2):137–151.

Gökhan Tür, Dilek Hakkani-Tür, and Kemal Oflazer. 2003. A Statistical Information Extraction System for Turkish. *Natural Language Engineering*, 9(2):181–210.

Reyyan Yeniterzi. 2011. Exploiting Morphology in Turkish Named Entity Recognition System. In *Proceedings of the ACL Student Session*, pages 105–110.