RECOGNIZING NAMED ENTITIES IN TURKISH TWEETS

Beyza Eken and A. Cüneyd Tantug

Department of Computer Engineering,
İstanbul Technical University, İstanbul, Turkey
[1]beyzaeken@itu.edu.tr
[2]tantug@itu.edu.tr

ABSTRACT

Named entity recognition (NER) is one of the well-studied sub-branch of natural language processing (NLP). State of the art NER systems give highly accurate results in domain of formal texts. With the expansion of microblog sites and social media, this informal text domain has become a new trend in NLP studies. Recent works has shown, social media texts are hard to process and the performance of the current systems substantially decrease when switched to this domain. We give our experience in improving named entity recognition on informal social media texts for the case of tweets.

KEYWORDS

Named Entity Recognition, Conditional Random Fields, Informal Domain, Tweet, Turkish

1. INTRODUCTION

Named entity recognition (NER) is a natural language processing (NLP) term that refers to the recognition of named entities in natural language. It is a way of extracting information by detecting and classifying named entities in texts. Must studied named entity types are person, location, organization which defined in MUC-6 [1] conference as ENAMEX type. Other mostly studied types are numeric entities like money, percentage as NUMEX type and date, time as TIMEX.

NER could take part in other NLP tasks like machine translation, sentiment analysis, and question-answering.

There have been a lot of studies in NER field in many languages and the state of the art performance has reached to nearly human annotation performance on formal texts [2]. But texts are not always formal like e-mails, microblog texts, social media texts, etc. But off the shelf NLP tools give low accuracy when they are applied to informal texts [3], because they may be ungrammatical and can have spelling mistakes unlike formal texts. As a consequence, the need arises to develop new methods which would work properly for informal structure of texts.

Dhinaharan Nagamalai et al. (Eds) : CCSEA, DKMP, AIFU, SEA - 2015
pp. 155–162, 2015. © CS & IT-CSCP 2015 DOI : 10.5121/csit.2015.50213

With the expansion of the web, information gathering and sharing via social media has become a rising trend. Twitter is one of the most used microblog site around the world, 500 million tweets are sent per day [4]. Tweets hold great amounts of statistics, they can give important information about a company, person etc. Therefore, at this point NER on tweet domain is holds crucial importance.

The aim of this study is to increase the performance of NER in Turkish tweets. Tweets are short texts that have maximum 140 characters, and most of the time they can contain grammar or spelling mistakes, slang words, smileys and so on. Unfortunately these irregular nature of tweets make it harder to process such data.

Turkish is a highly agglutinative language and it makes Turkish language morphologically rich. Morphological features hold meaningful importance in NLP tasks, they have important information about words. But in informal texts off the shelf morphological analysers do not give sufficient results. So, instead of morphological analysing process we prefer to use first and last four character of word in order to take advantage of morphological features. According to our results when first four characters of the word are used as an alternative to stem of the word, performance changes slightly.

Previous works [5]-[8] have shown that conditional random fields (CRF) method has reached to good performance at NER task. Consequently, in this work CRF have been chosen as the method to build named entity recognition model.

The rest of the paper follows with related works, then describes the method we used, after that gives and explains our results and lastly final section as conclusions.

2. RELATED WORKS

Named entity recognition is a well-studied field in many languages especially in English. First studies started in 1990s [2], now state of the art performance has reached nearly %95.

First NER study specific to Turkish used hidden markov models (HMM) on news data and reached %91.56 performance with person, location, organization types [9]. Bayraktar and Temizel [10] used patterns and word frequency to recognize Turkish person names on financial text domain. Küçük and Yazıcı [11] proposed a rule based system to recognize ENAMEX, TIMEX and NUMEX types, then they improved their system with rote learning algorithm and achieved %90.13 performance on Turkish news data [12]. Tatar and Çiçekli [13] created an automatic rule learning system and they achieved %91.08 performance on Turkish news data. Yeniterzi [6] got %88.94 performance on Turkish news data with CRF using morphological features. Şeker and Eryiğit [7] achieved %92 performance on Turkish news data with CRF using morphological features and gazetteers.

When examining NER for informal domain Özkaya and Diri [8] reached %92.89 performance with ENAMEX types on Turkish e-mails, e-mail domain kind of informal. Çelikkaya et al. [14] normalized tweets and tested on a model trained with Turkish news data and achieved %19 performance, they used CRF with morphological features and gazetteers. Küçük and Steinberger [15] adapted Küçük's rule based NER system [11] to tweet domain and got %61 performance.

Ritter et al. [3] tailored NLP pipeline to tweet domain, and get %51 score on English tweets, they used CRF in part-of-speech tagging, chunking, named entity segmentation parts and they used LabeledLDA [16]. Liu et al. [17] created a semi supervised system using k-nearest neighbors algorithm and CRF, they achieved %80.2 performance on English tweets. Li et al. [18] created an unsupervised system for only segmentation of named entities in English tweets using Wikipedia and Web N-gram corpus. Oliveria et al. [19] created a filter based system for English tweets.

3. DATASETS AND METHOD

We aimed to develop a model that will recognize person, location, organization, date, time, money and percentage named entities in Turkish tweets. Tweets are short texts which can be solecistic. Lack of context, containing spelling errors on purpose or not, slangs, repeating characters to indicate exclamation make hard NER process on tweets. Two root ideas for NERin domain like tweets are to tailor texts to existing NER tools or tailor existing NER tools to fit informal texts.

CRF have been used to build our NER model. CRF are introduced by Lafferty et al. CRF are statistical machine learning techniques which aim is to be applied to sequential data to segment and label. We used CRF++ tool [20] for training and testing system.

We used news data to train a base model just to see our results on news data to make comparison of selected features. Then we used tweets to train a second model which is more feasible for tweet domain.

We used two main data set from two different domain, news data as formal texts and tweets as informal texts. News data set which are collected from Turkish newspapers and labelled by Tür et al. [9]. Tweet data set consisting of two parts, first part is labelled by us for this work and consists of nearly 9K tweets, second part is labelled by Çelikkaya et al. [14] and consists of nearly 5K tweets. Entity counts for all datasets are given in Table 1 and Table 2.

Table 1. News data set entity counts.

	Train	**Test**	**Total**
Token	444.475	47.343	491.818
Entity	38.388	3579	36.967
Person	14.492	1598	36.967
Location	10.538	1177	11.715
Organization	8358	804	9162

We divide news data and used %10 for testing and remain for training.

Table 2. Tweets data set entity counts.

	Tweets-1	Tweets-2	Train	Test	Total
Tweet	9.358	5.040	12.471	1.930	14.401
Token	108.743	46.620	137.345	21.300	158.645
Entity	7.838	1.689	5.511	901	6.412
Person	2.744	875	2.099	429	2.528
Location	1.419	277	1.168	172	1.340
Organization	2.935	389	1.733	236	1.969
Date	351	82	261	31	292
Time	86	33	90	21	111
Money	212	29	88	10	98
Percentage	91	4	72	2	74

Tweets-2 column in Table 2 represents entity counts for Çelikkaya et al.'s tweets data set [14], Tweet-1 column represents our tweets data set. We combine two tweets data sets to make balanced training and testing sets, that is to say we take %10 of each tweets data set to comprise final tweets data set, and remaining of each are combined to comprise final training tweets.

We trained our first news model as same way in Şeker and Eryiğit's work [7], we nearly get same results as this work. We apply morphological analyse and disambiguation processes on data after tokenization. Oflazer's tool [21] is used for morphological analyse and Sak's tool [22] for morphological disambiguation. Morphological process is used for to extract stem, inflectional suffixes, part of speech, noun case and proper name case information of tokens, all of these information are used when training the model.

Another encountered writing style for Turkish tweets is that instead of using Turkish characters (ö, ç, ş, ı, ğ, ü) equivalent of English characters (o, c, s, i, g, u) are used. Therefore we asciified all data sets, which means we replaced all Turkish specific characters with equivalent of English characters.

Hence Turkish is an agglutinative language last characters of words are generally suffixes of words so they hold meaningful information about word's morphology. On the other hand, morphological processing tools do not perform well on tweets, so in this second method instead of using morphological features we used first four and last four characters of the tokens as features to train models. This alternative model performs nearly same as first one, so we infer that there is no need to use morphological analysing and disambiguation processes for this work.

Proper name's suffixes should be separated with apostrophe, therefore containing an apostrophe gives important clue about being a named entity, so this is also added as a feature.

Also we applied distance based matching to extract gazetteer features, because of twitter domain peculiarities exact matching can lead to missed out entities. Since tweets contain spelling errors, some named entities can be contracted like "İstnbul" instead of writing correct form of entity which is "İstanbul". Exact matching of input tokens and gazetteers will miss out contracted entities, in order to not miss out these entities we applied distance based matching with Levenshtein distance algorithm [23]. Levenshtein distance algorithm calculate distance between two strings, calculated distance between two strings represents minimum number of edits which are necessary to change one word into the other. For this work we calculate distances between

input token and each token in gazetteer. Zero distances are already named entities, distances closer to zero are candidate named entities. So we give a chance to tokens like "İstnbul" for being a named entity.

4. EVALUATION AND RESULTS

We evaluated our results according to CoNLL metric using CoNLL evaluation script [24], this metric calculates f-measure considering entity type and boundaries. In system output, if both of type and boundaries of a named entity are labelled correctly this entity counted as correct.

We labelled entities in data sets using NER annotation tool from [11] and we represent entities with IOB2 representation style introduced in [25].

Results are on our first model based on this work [7], it trained with news data and named with *N1_model*. We used morphological features, letter case features, start of sentence features and gazetteers to build this model. We tested all our test data on this model and results are in Table 3. We have got nearly same result as in [7] for news test dataset.

Tweets Test Set-1 results in Table 3, 4 and Table 5 are from final tweet test set which is a combination of our tweets and tweets from this [7] work. Tweet Test Set-2 results are represent results of tweets data set from this work [14].

Table 3. Results on first news model (N1_model)

Model	News Test Set	Tweets Test Set-1	Tweets Test Set-2
Surface	82.93	32.30	18.80
Stem	83.36	13.71	14.76
+Surface	84.30	32.43	15.16
+Part of speech	84.85	33.53	15.75
+Noun case	85.59	35.18	17.55
+Proper noun	86.91	21.87	9.58
+Inflections	87.14	22.39	9.79
+Case	90.01	34.66	20.38
+Start of sentence	89.91	34.83	20.27
+Gazetteers	90.38	41.22	25.45
+Distance-based Matching	90.47	41.79	25.62

Second news model named *N2_model* based on some different features instead of morphological features, results in Table 4. Our primary objective is improving tweets data performance for NER but we also trained and tested on news datasets to see and compare results.

Table 4. Results on second news model (N2_model)

Model	News Test Set	Tweets Test Set-1	Tweets Test Set-2
Surface	82.93	32.30	18.80
+First 4 characters	84.07	34.75	19.12
+Last 4 characters	85.34	36.39	20.80
+Apostrophe	86.11	37.43	22.70
+Case	89.41	41.58	24.49
+Start of sentence	89.50	41.57	24.15
+Gazetteers	90.17	46.97	27.90
+Distance-based Matching	90.23	46.57	28.53

When we look at Table 3 and 4 it can be seen we get nearly same results for news test data in both news models. It shows we can capture significant features with second way. Beside that f-measures are improved for tweets on *N2_model.*

Then we trained third model with second way using tweets as training data, which name is *T_model.* We got highest scores for tweets on this model.

Table 5. Results on tweets model (T_model)

Model	Tweets Test Set-1
Surface	49.32
+First 4 characters	56.41
+Last 4 characters	57.00
+Apostrophe	57.98
+Case	61.82
+Start of sentence	61.97
+Gazetteers	64.03
+Distance-based Matching	63.77

5. CONCLUSIONS

We studied on improving performance of NER on Turkish tweets. Although NER is almost a solved problem in formal texts domain, when switch domain to informal texts performance decreases in respectable amount. There are two main way in literature to handle this decrease, tailoring systems to adapt to informal texts or tailoring data to adapt to existing systems. We proposed a NER system for tweets without normalization of tweets.

We improved performance on tweets and get %64 f-measure with some basic features that are first and last 4 characters of the word, capitalization and apostrophe information and gazetteers. We asciified data sets and gazetteers before building our model and apply a little normalization. We employ distance based matching with Levenshtein distance algorithm when extracting gazetteer look up features, we will work on enhance gazetteer look up techniques.

REFERENCES

[1] R. Grishman & B. Sundheim (1996) "Message Understanding Conference-6: A Brief History", In Proceedings of 16th International Conference on Computational Linguistics, pp. 466-471.

[2] D. Nadeau & S. Sekine (2007) "A Survey of Named Entity Recognition and Classification", Linguisticae Investigationes, 30(1):3-26.

[3] A. Ritter et al. (2011) "Named Entity Recognition in Tweets: An Experimental Study", In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1524-1534.

[4] (2014, Dec 21). https://about.twitter.com/company.

[5] J. R. Finkel et al. (2005) "Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling", In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics, pp. 363-370.

[6] R. Yeniterzi (2011) "Exploiting Morphology in Turkish Named Entity Recognition System", In Proceedings of the ACL 2011 Student Session, pp. 105-110.

[7] G. A. Şeker & G. Eryiğit (2012) "Initial Explorations on using CRFs for Turkish Named Entity Recognition", In Proceedings of the 24th International Conference on Computational Linguistics, pp. 2459-2474.

[8] S. Özkaya & B. Diri (2011) "Named Entity Recognition by Conditional Random Fields from Turkish Informal Texts" In Proceedings of the IEEE 19th Signal Processing and Communications Applications Conference, pp. 662-665.

[9] G. Tür et al. (2003) "A Statistical Information Extraction System for Turkish", Natural Language Engineering, vol. 9, pp. 181-210.

[10] O. Bayraktar & T. T. Temizel (2008) "Person Name Extraction From Turkish Financial News Text Using Local Grammar Based Approach", In 23rd International Symposium on Computer and Information Sciences.

[11] D. Küçük & A. Yazıcı (2009) "Named Entity Recognition Experiments on Turkish Texts" In Proceedings of the 8th International Conference on Flexible Query Answering Systems, pp. 524-535.

[12] D. Küçük & A. Yazıcı (2012) "A Hybrid Named Entity Recognizer for Turkish", Expert Systems With Applications, vol. 39, pp. 2733-2742.

[13] S. Tatar & İ. Çiçekli (2011) "Automatic Rule Learning Exploiting Morphological Features for Named Entity Recognition in Turkish", Journal of Information Sciences, vol. 37, pp. 137-151.

[14] G. Çelikkaya et al. (2013) "Named Entity Recognition on Real Data", In Proceedings of the 7th International Conference on Application Information and Communication Technologies, pp. 1-5.

[15] D. Küçük & R. Steinberger (2014) "Experiments to Improve Named Entity Recognition on Turkish Tweets", In Proceedings of the 5th Workshop on Language Analysis for Social Media, pp. 71-78.

[16] D. Ramage et al. (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, vol. 1, pp. 248-256.

[17] X. Liu et al. (2011) "Recognizing Named Entities in Tweets", In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pp. 359-367.

[18] C. Li et al. (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, pp. 721-730.

[19] D. Oliveira et al. (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, pp. 597-604.

[20] (2014, Dec 21). http://crfpp.googlecode.com/

[21] K. Oflazer (1994) "Two-Level Description of Turkish Morphology", Literary and Linguistic Computing, vol. 9, pp. 137-148.

[22] H. Sak et al. (2008) "Turkish Language Resources: Morphological Parser, Morphological Disambiguator and Web Corpus", 6th International Conference on Natural Language Processing, vol. 5221, pp. 417-427.

[23] V. Levenshtein (1966) “Binar Codes Capable of Correcting Deletions, Insertions, and Revelsals”, Soviet Physics Doklady, vol. 10, pp. 707-710.
[24] (2014, Dec 21). http://www.cnts.ua.ac.be/conll2000/chunking/output.html
[25] E. F. Tjong Kim Sang & J. Veenstra (1999) “Representing Texting Chunks”, In Proceedings of the 7th Conference of the European Association for Computational Linguistics, pp. 173-179.