Named Entity Recognition on Real Data: A Preliminary Investigation for Turkish

Gökhan Çelikkaya
R&D Department
Huawei Technologies Co., Ltd
Istanbul, 34768, Turkey
Gokhan.Celikkaya@huawei.com

Dilara Torunoğlu
Institute of Science and Technology
Istanbul Technical University
Istanbul, 34469, Turkey
torunoglu@itu.edu.tr

Gülşen Eryiğit
Dep. of Computer Engineering
Istanbul Technical University
Istanbul, 34469, Turkey
gulsenc@itu.edu.tr

***Abstract*— Named Entity Recognition (NER) is a well-studied area in natural language processing (NLP) and the reported results in the literature are generally very high (~>%95) for most of the languages. Today, the focus area of most practical natural language applications (i.e. web mining, sentiment analysis, machine translation) is real natural language data such as Web2.0 or speech data. Nevertheless, the NER task is rarely investigated on this type of data which differs severely from formal written text. In this paper, we present 3 new Turkish data sets from different domains (on this focused area; namely from Twitter, a Speech-to-Text Interface and a Hardware Forum) annotated specifically for NER and report our first results on them. We believe, the paper draws light to the difficulty of these new domains for NER and the possible future work.**

Keywords — Named Entity Recognition, Turkish, Conditional Random Fields, ENAMEX, Speech Data, Twitter.

I. INTRODUCTION

NER is an important stage for several NLP tasks including machine translation, sentiment analysis and information extraction. As most of the NLP tasks, it is gaining importance with the rapid increase of the Internet usage and especially social media sites such as Twitter and Facebook. NER is a subtask of information extraction that seeks to locate and classify atomic elements in text into predefined categories such as the names of persons, organizations and locations.

Although, there are many important studies [1-4] in the literature for NER, the studies focused on real data is very limited and recent [5-7].

In recent years, there have been many studies for Turkish NER [8-14]. In this paper, we first replicated the results of the most state of the art study for Turkish (95% in MUC and 92% in CoNLL metric) [14] where the experiments have been conducted on News Media which is known to be very well-edited. The reader may check the cited paper for further related references.

In this study, we collected our data sets from three different domains (Twitter, Speech and Forum) and annotated them manually according to the NER guidelines of MUC 6 [15]. We used our replicated system for [14] on our new data sets under different scenarios. For comparison purposes, we evaluated only on ENAMEX (person, location and organization names) types and left TIMEX (date and time entities) and NUMEX (numerical expressions like money and percentages) evaluation for future studies.

The used machine learning algorithm is Conditional Random Fields (CRFs) [16] which is a very popular method used in Natural Language Processing and proven to be very effective on the NER task. In order to improve our first results on real data, we investigated the impact of using a text normalizer which we developed for Twitter data. We also made an initial exploration on the use of morphological features and gazetteers.

This paper is organized as follows: Section 2 gives brief information about Turkish and the challenges posed by real data, Section 3 presents our new datasets and Section 4 our NER approach, Section 5 gives our experimental results and discussions. We conclude with Section 6.

II. TURKISH AND THE CHALLENGES OF REAL DATA FOR NER

Turkish is a very characteristic agglutinative language and the studies for this language serve as a model for similar languages from the same family such as other Turkic languages (Azerbaijani, Turkmen, Qashqai, Gagauz, Balkan Gagauz Turkish and Oghuz) and agglutinative languages (Korean, Basque and so on). The writing rules of proper nouns serve as good features for NER on formal and edited texts. In most of the situations in written formal texts, the proper nouns are written with an initial capitalized letter and the suffixes to these are separated from the word by an apostrophe. However, people do not obey to these rules while writing to social media sites and the current speech-to-text interfaces do not produce their results as so. Another important challenge of this language is that most of the proper nouns (person and organization names, locations) are actually valid common nouns. And, in the formal text, when these words are written lower cased and without an apostrophe separator from the suffixes, they generally stand for a common noun which contradicts with the situation in real data. An example showing this situation is given below. In this Twitter example, which should actually be written as in the second line in formal writing, “Aydın” is a person name. The word when written with lowercase letters has also the meaning of a common noun “*enlightened”*. This makes very difficult to differentiate/identify this named entity (person name) from the word “enlightened”.

"**aydınlara** gidiyoruz."
"**Aydın'lara** gidiyoruz."
(*We are going to **Aydın's** house*)

Another problem for real data is the spelling errors produced either by mistake or on purpose for exaggeration, interjection or ASCIIfication (removal of accent, cedilla, etc) of special letters (öüçşığİ). In the below example, the character "ı" is written with its ascii counterpart and written several times for specifying exclamation.

"aydiiiiiiiiin nerdesin?"
(*Aydın, where are you?*)**

And finally, many foreign words are used in this data set with normal Turkish suffixes such as in the following Tweet: "Bieber" is used in accusative case without the required apostrophe.

"**Justin Bieber**i sevmem."
"I don't like **Justin Bieber**"

The given examples are also valid for the used speech-to-text interfaces such as Google API.

III. DATA SETS

We created three new data sets from different domains in order to create a representative model for the focused area: namely, social media and spoken language data. The data is collected from three different resources and the data size for each resource (except speech data) is chosen to be approximately 50K tokens which is also the test set size used in previous studies [8, 12, 14]. All of these new data sets are annotated manually according to the NER guidelines of MUC6 [15]. Although we also annotated TIMEX and NUMEX entities, in this study we only evaluate for ENAMEX types given in the first three rows (Person, Location, and Organization) for comparison purposes.

TABLE I. reports the token counts and the counts for each of the named entity types. The first and second column of the table states the counts for the training and the test data from [14], which is also used in our experiments. This data set which is collected from an online Turkish Newspaper site is well written and edited. Thus, the spelling rules are almost always obeyed within the text. It consists of ~500K tokens where ~450K is used for training and ~50K is used for testing purposes on previous studies.

The third column gives the statistics for our 1st new resource which is from a very popular online forum (http://www.donanimhaber.com) dedicated for hardware products' reviews. An important feature of this data set is that it contains mostly trademarks (generally company names), their products together with a related model. An example to this may be given as: "Apple" or "IPhone 5". Although, this type of named entities are categorized under more specific named entity classes in extended NE classifications [17], the most relevant category in MUC6[1] for these is the "Organization". This forum data is full of spelling errors and capitalization is not properly used or not used at all in most of the cases.

The fourth column of TABLE I. is our spoken language resource. Nowadays, one of the most popular areas for NLP is the mobile personal assistant applications. Recognizing the named entities such as Locations, Person Names and Date/Time expressions is especially important in order to accomplish the related mobile operations such as adding meeting reminders, sending sms or email messages. Since there is no such an existing resource for Turkish, we first developed a mobile application on Android OS. This application takes the spoken utterance via a mobile phone and converts it into written text by using Google Speech Recognition Service. We then asked different people to talk to this phone and give some relevant orders. The data set size is small when compared to the other two datasets due to our time constraints. But we plan to increase this data size for our future work. The most important characteristic of this data set is that there is no capitalization or punctuation at all in the produced text message.

The last column presents our data set from Twitter. As before, the written text do not obey to spelling rules and the data consists many exceptions when compared to formal texts.

TABLE I. DESCRIPTION OF THE DATA SETS

Data Set	News Media Train Data	News Media Test Data	Forum Test Data	Speech Test Data	Twitter Test Data
Data Size	457K	48K	54451	1451	54283
Person	22700	1400	21	79	676
Location	13750	2260	34	90	241
Organization	12322	1218	858	64	419
Date	n/a	n/a	7	70	60
Time	n/a	n/a	2	34	23
Money	n/a	n/a	67	27	14
Percentage	n/a	n/a	11	26	4

IV. APPROACH

We adopted a similar approach to the cited previous work[14] which firstly tokenize the data and then prepare training/testing instances by the use of the features coming from an automatic morphological analysis process and some other predefined features: The morphological features (TABLE III.) are mainly the stems of the tokens within the context,

[1] MUC 6 suggests either to assign the "organization" markup or no markup at all to the product categories.

their main POS-tags, the case marker (for nouns) and the availability of the proper noun tag within the related morphological analysis of the token. The remaining features are the surface form of the tokens, their lower/upper case status and some binary features stating the sentence beginning and the existence in the prepared gazetteers. As that work is mainly prepared for formal written media, the gazetteers (TABLE II.) are also prepared with this in mind. In other words the generator gazetteers consists of only formal words which are used to generate some NE categories: To give an example[14]: the stem "bakanlık" (ministry) which could come after some regular words such as "spor", "tarım" (sports, agriculture) to construct organization NEs such as "Tarım Bakanlığı" (Ministry of Agriculture).

TABLE II. DESCRIPTION OF THE GAZETTEERS

Data Set	Gazetteer	# of tokens
BASE	First names	44.048
	Surnames	138.844
	Location names	33551
GENERATOR	Location	44
	Organization	60
	Person	22

But, an observed problem in the case of real data is that people uses many informal generator words in daily language other than the ones included in the above gazetteers. To give an idea, some of the words from the Person Entity Generator gazetteer are the words like "President", "Minister" so on. On the other hand, on Twitter data we mostly see generator words like the following informal sayings "Abla" (*elder sister*), "*Abi*" (older brother), "Hoca" (*professor*) and so on. For the time, we used exactly the same gazetteers to have a first impression on our problem, but it is obvious that we will need to extend these gazetteers while adapting our work to the real data domain.

As stated earlier, the new domains that we work on contain many spelling errors occurred either by mistake or on purpose. And, it is impossible for our automatic morphological analysis to process these erroneous tokens. A first effort to normalize this data is to create a text normalizer and process the input data by this tool before other stages. Fig. 1. presents the flow in our approach.

Our text normalizer mainly works on the normalization of the following cases:

1. Slang words (i.e. "nbr" for "ne haber?" -"*what's up?*")
2. Repeated characters for ejaculation (i.e. "çoooook" for "çok"-"*many*")
3. Hash tags, mentions, smiley icons and vocatives
4. Emo style writings (i.e. "$eker 4you" instead of "şeker senin için" – "Sweety! for you"
5. Capitalization (i.e. "aydın" for "Aydın")

As a result, the following input sample takes the form of the second line after normalization.

input	" @dida_ss nbr yaaaaa s3n b€ni hiç aramıon #regex harika dimi hahahahaha :))) "
output	" @mention[@dida_ss] ne haber ya sen beni hiç aramıon @hashtag[#regex] harika değil mi @vocative[ha] @smiley[:)] "

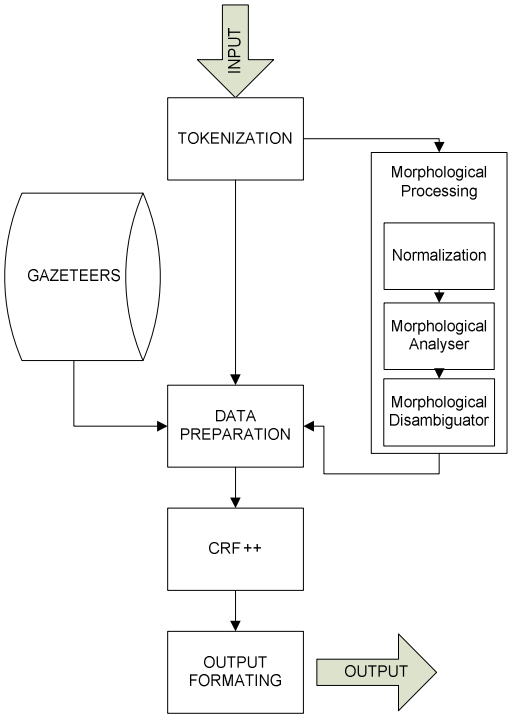

Fig. 1. Design of the Framework.

We used Conditional random fields (CRFs) as the machine learning algorithm. CRFs are a framework for building probabilistic models to segment and label sequence data. Advantages of hidden Markov Models (HMMs), stochastic grammars and maximum entropy Markov models (MEMMs) are used in CRFs. CRF is a discriminative model better suited to including rich, overlapping features focusing solely on the conditional distribution $p(y|x)$. We use linear chain CRFs where $p(y|x)$ is defined as:

$$p_\theta(y|x) = \frac{1}{Z_\theta(x)} \exp\left\{\sum_{t=1}^{T}\sum_{k=1}^{K} \theta_k f_k(y_{t-1}, y_t, x_t)\right\}$$

Where $f_k(y_{t-1}, y_t, x_t)$ is the function for the properties of transition from the state y_{t-1} to y_t with the input x_t and θ_k is the parameter optimized by the training. We used the same feature set of the previous work and provided the CRF++ library [18] with the atomic features of TABLE III. within a window of [-3,+3] and some selected combinations of these.

V. Experimental Results and Discussions

For our experiments we have developed three different feature models (TABLE III.) and tested them (TABLE IV.) on 4 different test sets. We used the CoNLL metric on the evaluation. This metric evaluates an assignment to be correct if both the type and the boundary of a NE is determined correctly. The details of the calculation for these metrics may be investigated from [4].

TABLE III. Description of the Feature Models

Features	Model A	Model B	Model C
Word	x	x	x
Stem	x	x	x
Pos Tag	x	x	x
Noun Case	x	x	x
Proper Noun	x	x	x
Inflectional Features		x	x
Lower/Upper Case			x
First Word			x
Gazetteers	x	x	x

Model C in TABLE IV. is the best model given in [14]. We also tested with Model A and Model B which are actually some reduced versions of Model C where some of the features (such as Case Letter State, Sentence Beginning) becomes useless and meaningless in our real data test sets. We also experiment these models with and without normalization. To see the impact added by the capitalization alone, we also created two experiment sets: normalization without Capitalization (TABLE IV. the middle row block) and with added Capitalization (TABLE IV. the last row block). Our capitalization strategy is very naïve for now; it capitalizes only the words of which the surface forms exactly occur within the gazetteers. In the future, this strategy needs to be ameliorated with more intelligent selections: the use of stems instead of surface forms and the use of statistical information for most frequently used proper names and common nouns in order to differentiate between these two categories.

TABLE IV. Experiment Results on Data Sets on ConLL Metric

		News Media Data	Twitter Data	Speech Data	Forum Data
No Normalization	Model A	14,70%	0,93%	0,82%	0,40%
	Model B	88,56%	13,88%	50,69%	2,86%
	Model C	91,64%	12,23%	6,92%	5,62%
Normalized	Model A	14,10%	0,92%	0,82%	0,39%
	Model B	86,72%	15,27%	50,84%	2,41%
	Model C	82,59%	19,28%	6,90%	2,16%
Normalized + Capitalized	Model A	14,06%	1,33%	3,95%	0,38%
	Model B	83,60%	14,57%	50,26%	2,18%
	Model C	75,48%	15,00%	32,88%	1,49%

For news media data, our reimplementation "Model C without normalization" obtained almost the same result (91.64%) with the original work (91.94%) [14]. We observe the negative impact of normalization and capitalization approaches on this data set with 82.59% and 75.48% accordingly. Since this data is already written in the appropriate writing style, the change on the data causes severe problems such as the treatment of common nouns as proper nouns. We also see that the reduced models (Model A and Model B) also cause a very high decrease in the performance for this data set.

Since the system is trained on formal text data, the results on informal data is very low as expected. For Twitter data, normalization has a promising impact since the tweets are including extensive OOV words that the normalization approach can easily solve. The best results obtained for this data set was the Model C with normalization without capitalization with 19.28%.

We could not achieve any important improvement on forum data since our training data was not proper for this dataset at all. Forum data contains almost only organization entities as shown in TABLE I. The organization entities in this data set consists of brands and models on the contrary to our train data.

Although, capitalization improved the results on Model C for speech data, we obtained the best result with Model B. Since speech data consists of all lower case letters, removing lower/upper case feature has a good impact on the data and improved the results. We obtained the best results for this data set with Model B with normalization without capitalization as 50.84%. We also tried training by lowercased version of our training set (instead of automatic capitalization of the test data). This increased the results of speech data a little bit from 6.90% to 14.77%. But not as much as not using this feature (lower/upper case) in the model.

During our experiments, we observed the difficulty of NER on these new domains which differ severely from our training data both by the mentioned NEs and the writing style. In order to adapt the NER work into these new domains, and improve the results to the range of ~90%, we urgently need to create new models. For now, we plan to focus on three possible ways: 1. To try different feature sets specific to these new domains 2. To use active learning and extend our training data with the data coming from these new domains. 3. To extend the generator gazetteers with generator words coming from these new domains and the name gazetteers with especially popular foreign names from other languages (such as "Obama", "Justin" so on.).

VI. Conclusion and Future Work

Our contribution in this paper is mainly the creation of new NE datasets from different real data domains and the presentation of the first NER results on them. These first results showed the difficulty of this new domain for regular named entity recognizers.

In our study we focused on developing a Turkish NER model using conditional random fields trained with

morphological and lexical features influenced by the work done in [14]. We developed the same baseline approach and tested it on forum data, speech data and Twitter under different preprocessing scenarios: these are normalization and capitalization approaches. As future work, we plan to work deeply on this research are by using the methods explained in the previous section and also add NUMEX and TIMEX entity types to our system which are crucial for many practical NLP applications.

ACKNOWLEDGMENT

This work is part of two ongoing research projects:
1. Gökhan Çelikkaya is supported by SANTEZ (Industrial Thesis) project (grant no: 0073.STZ.2013-1) between Istanbul Technical University and Huawei Istanbul.
2. Dilara Torunoğlu is supported by ICT COST Action IC1207 TUBITAK 1001 (grant no: 112E276) project "Parsing Web 2.0 Sentences".

REFERENCES

[1] B. Sundheim, "Overview of Results of the MUC-6 Evaluation," presented at the MUC, 1995.
[2] N. A. Chinchor and E. Marsh, "Muc-7 information extraction task definition," in *Proceeding of the Seventh Message Understanding Conference (MUC-7), Appendices*, ed, 1998.
[3] E. F. Tjong Kim Sang and F. De Meulder, "Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition," presented at the CoNLL-2003, 2003.
[4] D. Nadeau and S. Sekine, "A Survey of Named Entity Recognition and Classification," *Linguisticae Investigationes,* vol. 30, pp. 3-26, January 2007.
[5] X. Liu, S. Zhang, F. Wei, and M. Zhou, "Recognizing Named Entities in Tweets," presented at the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, Oregon, USA, 2011.
[6] A. Ritter, S. Clark, O. Etzioni, and others, "Named Entity Recognition in Tweets: an Experimental Study," presented at the Conference on Empirical Methods in Natural Language Processing, 2011.
[7] J. J. Jung, "Online named entity recognition method for microtexts in social networking services: A case study of twitter," *Expert Syst. Appl.,* vol. 39, pp. 8066-8070, 2012.
[8] R. Yeniterzi, "Exploiting Morphology in Turkish Named Entity Recognition System," presented at the ACL 2011 Student Session, Portland, OR, USA, 2011.
[9] S. Tatar and I. Cicekli, "Automatic Rule Learning Exploiting Morphological Features for Named Entity Recognition in Turkish," *Journal of Information Science,* vol. 37, pp. 137-151, April 2011.
[10] S. Özkaya and B. Diri, "Named Entity Recognition by Conditional Random Fields from Turkish Informal Texts," presented at the IEEE 19th Signal Processing and Communications Applications Conference (SIU 2011), 2011.
[11] D. Küçük and A. Yazıcı, "A Hybrid Named Etity Recognizer for Turkish," *Expert Syst. Appl.,* vol. 39, pp. 2733-2742, 2012.
[12] G. Tür, D. Hakkani-tür, and K. Oflazer, "A Statistical Information Extraction System for Turkish," *Natural Language Engineering,* vol. 9, pp. 181-210, June 2003.
[13] Ö. Bayraktar and T. T. Temizel, "Person Name Extraction From Turkish Financial News Text Using Local Grammar Based Approach," in *23rd International Symposium on Computer and Information Sciences (ISCIS'08)*, ed. Istanbul, 2008.
[14] G. A. Şeker and G. Eryiğit, "Initial explorations on using CRFs for Turkish Named Entity Recognition.," presented at the In Proceedings of the 24th International Conference on Computational Linguistics, COLING 2012, Mumbai, India, 2012.
[15] Muc6. *6th conference on Message understanding, Named entity task definition*. Available: http://cs.nyu.edu/faculty/grishman/NEtask20.book_6.html#HEADING17
[16] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," in *ICML*, ed, 2001, pp. 282-289.
[17] S. Sekine, K. Sudo, and C. Nobata, "Extended named entity hierarchy," in *Proceedings of LREC*, 2002.
[18] (CRF++). (2003). *CRF++: Yet Another CRF toolkit*.