Koc. J. Sci. Eng., 4(2): (2021) 86-92 https://doi.org/10.34088/kojose.871873

Kocaeli University

Kocaeli Journal of Science and Engineering

http://dergipark.org.tr/kojose

Named Entity Recognition in Turkish Bank Documents

Osman KABASAKAL [1], Alev MUTLU [2, *]

[1] *Department of Computer Engineering, Kocaeli University, Kocaeli, 41001, Turkey,* ***ORCID:*** *0000-0003-1187-5147*
[2] *Department of Computer Engineering, Kocaeli University, Kocaeli, 41001, Turkey,* ***ORCID:*** *0000-0003-0547-0653*

Article Info

Research paper

Received *: January 31, 2021*
Accepted *: April 13, 2021*

Keywords

Bank Document
Conditional Random Fields
Named Entity Recognition
Natural Language Processing
Turkish Documents

Abstract

Named Entity Recognition (NER) is the process of automatically recognizing entity names such as person, organization, and date in a document. In this study, we focus on bank documents written in Turkish and propose a Conditional Random Fields (CRF) model to extract named entities. The main contribution of this study is twofold: (i) we propose domain-specific features to extract entity names such as law, regulation, and reference which frequently appear in bank documents; and (ii) we contribute to NER research in Turkish document which is not as mature as other languages such as English and German. Experimental results based on 10-fold cross validation conducted on 551 real-life, anonymized bank documents show the proposed CRF-NER model achieves 0.962 micro average F1 score. More specifically, F1 score for the identification of law names is 0.979, regulation name is 0.850, and article no is 0.850.

1. Introduction

Named Entity Recognition (NER) is the process of automatically recognizing named entities in a text document and mapping them to semantic categories such as person, place, organization, time, and date. It is one of the most studied tasks of natural language processing and has applications in domains such as keyword extraction [1], question answering [2], text clustering [3], and text summarization [4].

The NER problem was first defined in MUC-6 [5], and since then several NER systems have been proposed. Based on the learning technique, these systems can be classified as rule-based [6, 7], unsupervised [8], supervised [9, 10], and semi-supervised [11]. Rule-based systems are among the earliest NER attempts and employ gazetteer information and hand-crafted rules to identify and classify named entities. Such systems have high precision and low recall due to domain-dependent and user-defined rules. Unsupervised approaches are generally based on the clustering of a large corpus. These systems identify and classify named entities based on the context similarity of clustered documents. In supervised techniques, NER is formulated as a multi-class classification problem. In supervised NER, words are represented using word-, document-, and corpus-level features and learning algorithms such as support vector machines (SVM) and decision trees; sequence labeling algorithms including conditional random fields (CRF) and Hidden Markov Models (HMM) are used. Supervised learning builds models on annotated data, which is labor-intensive and requires domain expertise to attain. Semi-supervised learning algorithms aim to overcome dpcument annotating problem of supervised learning algorithms by extending limited annotated data by self-labeled unlabeled data. More recently, NER systems based on deep learning models are also proposed [12, 13]. Such systems benefit from the ability to discover hidden features automatically [14].

NER systems can also be classified as generic and domain-specific based on the type of named entities they aim to recognize. The former focuses on identifying generic named entities such as organization, person, location, and percentage. The latter, on the other hand,

[*] Corresponding Author: alev.mutlu@kocaeli.edu.tr

aims at learning domain-specific named entities such as adverse drug reactions [15], legal norms [16], and chemical entities [17].

In this study, we propose a domain-specific CRF-based NER system to recognize named entities in bank documents. The proposed system aims to recognize domain-specific named entities such as money amount, law name, regularity law, law article, and reference. Moreover, the proposed system is modeled to recognize generic named entities including organization, date, time duration, abbreviation, and e-mail. This study does not focus on place and person categories as such entities are removed from documents due to privacy concerns.

To evaluate the performance of the proposed features' in identifying named entities in bank documents, we conducted experiments on real-life anonymized bank documents. The experimental results show a 0.962 micro-averaged F1-score. Considering the domain-specific named entities, 0.850, 0.979, 0.975, and 0.947 F1-scores are achieved, respectively, for regulation, law name, article number, and money amount.

The rest of the paper is organized as follows. In Section 2, we introduce CRF-based named entity recognition. In Section 3, we introduce the domain-specific features proposed for the bank domain. Section 4 presents the experimental findings, and the last section concludes the paper with possible future directions.

2. Background

This section introduces CRF-based named entity recognition and later provides a literature summary related to named entity recognition in Turkish, financial, and legal domains.

Conditional Random Fields are statistical methods for sequence modeling. They are undirected graphical models that consider context while determining a sequence label. CRF aims to learn the conditional probability of values on designated output nodes given values of the designated input nodes. More specifically, let o = <o_1, o_2, …, o_n> be observed input sequence and s = <s_1, s_2, …, s_T> be a state sequence corresponding to labels assigned to observations in o, the conditional probability is calculated as in Equation (1).

$$\mathrm{P}(\mathrm{s}|\mathrm{o}) = \frac{1}{\mathrm{Z_0}} \exp\left(\sum_{i=1}^{n}\sum_{j=1}^{m} \lambda_j f_j(s_{i-1}, s, o, i)\right) \tag{1}$$

In Equation (1), $\mathrm{Z_o}$ is a normalization factor, $f_j(s_{i-1}, s, o, i)$ is one of *m* functions that describe a feature, and λ_j is a learned weight. Training in CRF corresponds to maximize the learned weights such that a learned weight, λ_j, for each feature, f_j, should be positive when there is a correlation with the target label, negative when there is an anti-correlation, and close to 0 when uninformative.

One of the earliest studies on Turkish NER is presented in [18]. The authors of the study propose a method to model morphological and contextual patterns for named entity recognition.

A rule-based NER for generic named entities for Turkish is proposed in [19]. The authors contribute to the literature by evaluating a rule-based NER system on diverse types of text documents.

A recent study [20] investigates the effect of different features in CRF-based NER performance for Turkish. To this aim, the authors investigate the effect of different positional and semantic features on the performance of NER in generic Turkish named entities.

A rule-based NER system for financial documents was introduced in [21]. The authors aim to identify organization, person, and location names by defining hand-crafted rules.

A rule-based system that aims to extract specific information, namely current/previous financial factor, current/previous volume, change type, and volume from financial documents is introduced in [22]. The authors induce rules to extract such entity names using symbolic learning model trained by greedy and tabu search.

A CRF-based NER system to identify stock names, company names and their abbreviations from financial documents is introduced in [23]. To recognize company names, the authors define features regarding suffix keys and location names. The authors define features based on mutual information and information entropy between full names and abbreviations to identify abbreviated stock names.

A NER system to support credit risk assessment is introduced in [24]. The authors present a CRF-based NER model to extract person, organization, location, and miscellaneous entities from financial documents. The study contributes to the literature by enhancing domain-data with a large amount of out-of-domain annotated data to improve performance.

A NER system to identify person names from Turkish financial news is presented in [25]. The authors define local grammars based on reporting verbs to identify person names. The study concludes that the local grammars are successful in named entity extraction; however, they are difficult to construct due to Turkish word-formation.

A NER system to identify legal name entities such as legal norm, ordinance, court decision as well as generic named entities such as person, location, and organization is for German legal documents is introduced in [26]. The authors make use of gazetteers of persons, locations, and laws to improve performance further. The authors compare CRF- and BiLSTM-based NER systems and conclude that the BiLSTM model is superior over the CRF model.

A deep learning model for NER to recognize court names, judgment date petitioner, respondent, judge name, and act entities in legal documents is proposed in [27]. The authors propose a four-layer CNN model wherewith residual connections and the max out activation function to perform the encoding step.

The NER system proposed in this study differs from [21,24,25] as it does not require defining domain-specific rules, which require expert knowledge, are costly, and difficult to keep up-to-date. The proposed NER system differs from [25] as it works only with binary attributes. The proposed study also differs from [24] as it aims to identify a larger number of named entities. The proposed method differs from [26], as it does not need external resources such as gazetteers.

3. NER for Bank Documents

In Table 1, we list the categories that the proposed NER system aims to recognize. Table 1 also provides a brief description of these categories and gives some examples for each category. There are ten categories, the first four of which are generic, and the last six are domain-specific. In this study, we do not aim to recognize locations and person names, as the bank documents we are working on are anonymized.

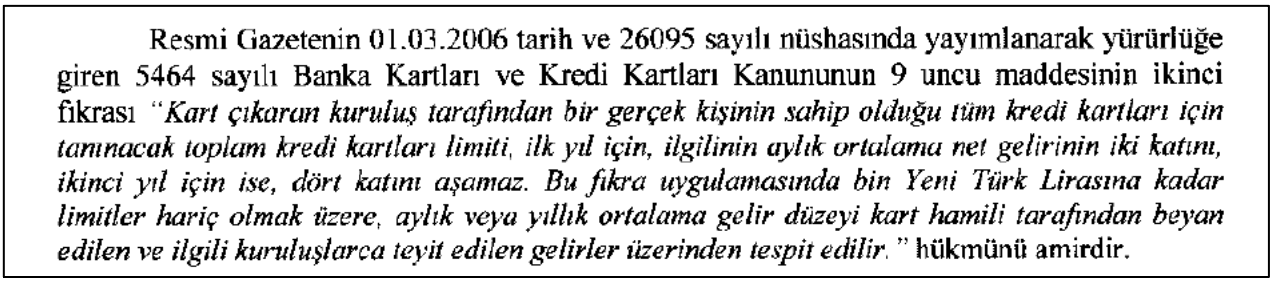
Table 1. Semantic Categories and their definition.

Category	Symbol	Explanation	Example
E-mail	EMA	E-mail addresses	yonetim@banka.com.tr
Time duration	ZAM	This entity indicates time durations or time intervals	yedi gün içinde, 1 yıllık dönem
Organization	ORG	Indicates an organization name	Türkiye Cumhuriyeti Merkez Bankası
Date	TAR	Indicates a date	11/05/2020, 13 Ocak 2020
Money amount	PAR	Indicates the amount of money	2 TL, 85 Avro
Article number	MAD	A paragraph or section of a legal document or statute	28 inci madde
Abbreviation	KIS	A shortened form of a name or phrase	IBAN, DAB
Law name	KAN	Name of law	Elektronik Para Kuruluşları hakkında kanun 5. maddesi
Reference	ILG		131161/A sayılı yatırım
Ordinance	YON	Rules and regulations created by federal and state agencies.	Bankaların Kredi İşlemlerine İlişkin Yönetmelikte

To recognize these named entities, we defined the following nine features.

- POS: Each word has a feature indicating its part-of-speech tag.
- isNumber: This is a boolean feature indicating either a token is a number or not.
- isCapital: This feature indicates either a token is all capitalized, letter case or small case. Our observation is that all capitalized tokens are generally abbreviations. Organization names are letter cases.
- containsAt: This is a boolean feature indicating if a term contains the “@” sign. This feature is used to help recognize e-mails.
- isTime: This is a boolean feature indicating if a token contains “:” sign or precedes the term “*saat*”.
- isDate: This is a boolean value indicating if a number is followed or preceded by a month name or includes the “/” sign.
- isTimeInterval: This is a boolean feature that indicates if a term is followed by “*gün*”, “*ay*” or “*yıl*”.
- precedesYonetmelik: This is a boolean feature that indicates if a term precedes a term that includes “yönet”. Our observation is that the length of regulation names ranges from 4 to 14 tokens. Hence, we look ahead to 4 to 10 tokens to set this feature to true.
- precedesCurrency: This is a Boolean feature indicating if a term precedes currency names such as “Avro”, “Euro”.

We implemented the proposed model using Python’s sklearn-crfsuite library. L-BFGS is used to train the model. Figure 1 is a snapshot of a sample input file. The named entities recognized for this file along with their tages are listed in Table 2.



Resmi Gazetenin 01.03.2006 tarih ve 26095 sayılı nüshasında yayımlanarak yürürlüğe giren 5464 sayılı Banka Kartları ve Kredi Kartları Kanununun 9 uncu maddesinin ikinci fıkrası *"Kart çıkaran kuruluş tarafından bir gerçek kişinin sahip olduğu tüm kredi kartları için tanınacak toplam kredi kartları limiti, ilk yıl için, ilgilinin aylık ortalama net gelirinin iki katını, ikinci yıl için ise, dört katını aşamaz. Bu fıkra uygulamasında bin Yeni Türk Lirasına kadar limitler hariç olmak üzere, aylık veya yıllık ortalama gelir düzeyi kart hamili tarafından beyan edilen ve ilgili kuruluşlarca teyit edilen gelirler üzerinden tespit edilir."* hükmünü amirdir.

Figure 1. Sample Input File

Table 2. Sample output of the named entites

Resmi - ORG
Gazetenin - ORG
01.03.2006 - TAR
26095 - ILG
5464 - ILG
Banka - KAN
Kartları - KAN
ve - KAN
Kredi - KAN
Kartları - KAN
Kanununun - KAN
9 - MAD
uncu - MAD
maddesinin - MAD
ikinci - ZAM
yıl – ZAM
Yeni - PAR
Türk - PAR
Lirasına - PAR
aylık - ZAM
yıllık - ZAM

4. Experiments

This section introduces the dataset used in the experiments, the data preprocessing procedure, and the experimental setting. Next, we present and discuss the experimental findings.

4.1. Experimental Setting

The dataset used to evaluate the proposed NER system's performance consists of 551 real-life, anonymized bank documents. The documents' length ranges from 7 to 633 words, with an average length of 99 words. In Table 3, we list the distribution of the categories over the documents. As the categories' distribution indicates, domain-related categories such as law name, article number, and regulation law frequently appear within the dataset. On the other hand, instances related to e-mail and abbreviation categories are less frequent.

Table 3. Distribution of the semantic categories

Category name	**#**	**%**
E-mail	11	0.12
Time duration	497	5.56
Ordinance	560	6.27
Date	1044	11.69
Money amount	388	4.34
Organization	2688	30.09
Article no	1605	17.97
Abbreviation	250	2.80
Law name	994	11.13
Reference	897	10.04
Total	9430	100

The documents we conducted experiments on are in PDF format. To convert the files into text format, we used PyTesseract[†], Python's optical character recognition tool. The documents consist of three sections: title, main text, and signature section. As the documents are structured, we wrote a script that extracts the main text from the documents. To tokenize the main text and determine the POS-tags, we used Zemberek[‡]. Named entity recognition is applied to the main text of the documents.

To evaluate category-based performance, precision, recall, and F1-score are used. Precision in NER is formulated in (2). It refers to the fraction of the correctly predicted named entities for a category (COR) over the total number of the category predictions (ACT). Recall refers to the number of correct predictions for a category (COR) over the number of instances that belong to that category (POS). Recall is formulated in (3). F1-score refers to the harmonic mean of precision and recall and is formulated in (4).

$$\text{Precision} = \frac{\mathrm{COR}}{ACT} \tag{2}$$

$$\text{Recall} = \frac{\mathrm{COR}}{POS} \tag{3}$$

[†] https://pypi.org/project/pytesseract/
[‡] https://github.com/ahmetaa/zemberek-nlp

$$\mathrm{F1-score} = 2 * \frac{\mathrm{Precision} * \mathrm{Recall}}{Precision + Recall} \tag{4}$$

To evaluate the overall performance of the proposed system, we employ the micro-averaged F1-score. Micro-averaged F1-score is preferred when there is a class imbalance. Micro-averaged F1 is formulated in (5), where *n* indicates the number of classes, TP_i indicates the number of true positive predictions for class *i*, and FP_i indicates the false positive predictions for class *i*.

$$\mathrm{MicroAveraged-F1} = \sum_{i=1}^{n} \frac{TP_i}{TP_i + FP_i} \tag{5}$$

The experimental results presented below are based on 10-fold cross validation. At each fold, 90% of the data is used for training and the remaining 10% of the data is used for testing.

4.2. Experimental Results

In Table 4, we report class-based precision, recall, and F1-scores. As the results indicate, the highest precision and the lowest recall values are achieved for the e-mail category. This is an expected result as there are only 11 instances that belong to the e-mail category. When the confusion matrix given in Table 5 is examined, one can observe that 8 out of 11 e-mail instances are correctly predicted, while the remaining 3 are misclassified as other. This may due to the fact that, no e-mail instances are present in training sets during some folds. Category ordinance ranks the second-worst in terms of F1-score. The rest of categories achieve F1-scores around 0.97. The system achieves 0.962 micro-averaged F1-score.

A recent study [26] that focuses on legal documents reports an F1-score 0.867 for the ordinance category, 0.966 for law names. Compared to that system, the proposed features achieve similar results, 0.850 and 0.979 for ordinance and law name, respectively. The proposed NER system is superior over [26] when compared with respect to the organization category. The proposed NER system achieves 0.973 F1-score while [26] reports at a rate of 0.908.

Table 4. Category-based results

Category	Precision	Recall	F1-Score
E-mail	1.0	0.73	0.842
Time duration	0.97	0.84	0.900
Ordinance	0.92	0.79	0.850
Date	0.98	0.99	0.986
Money amount	0.97	0.93	0.947
Organization	0.97	0.98	0.973
Article no	0.98	0.99	0.985
Abbreviation	0.93	0.92	0.926
Law name	0.97	0.99	0.979
Reference	0.99	0.98	0.984

In Table 5, we present the confusion matrix. As the results indicate, most of the misclassifications are assigned to the category *other*.

Table 5. The confusion matrix

		Predicted										
		ORG	**PAR**	**KAN**	**YON**	**ILG**	**TAR**	**ZAM**	**KIS**	**EMA**	**MAD**	**OTHER**
Actual	**ORG**	2631	0	0	0	0	0	0	1	0	0	56
	PAR	0	360	0	0	2	4	0	5	0	0	17
	KAN	0	0	981	0	0	0	0	3	0	0	10
	YON	0	0	0	441	0	0	0	0	0	0	119
	ILG	0	1	3	0	879	8	0	0	0	0	6
	TAR	0	0	0	0	1	1034	2	0	0	0	7
	ZAM	0	0	0	0	0	0	418	0	0	0	79
	KIS	1	3	0	0	3	0	0	231	0	0	12
	EMA	0	0	0	0	0	0	0	0	8	0	3
	MAD	0	0	2	0	0	0	0	0	0	1593	10
	OTHER	87	8	25	37	4	8	12	9	0	35	35117

5. Conclusion

In this study, we focused on bank documents to recognize domain-specific categories such as law-name, ordinance, and article no. To this aim, we trained a CRF-based NER system with nine features. The experimental results show that the achieved results are comparable to those reported in the literature.

As a future work, we plan to annotate data using the IOB standard and reevaluate the proposed features' performance.

Declaration of Ethical Standards

The authors of this article declare that the materials and methods used in this study do not require ethical committee permission and/or legal-special permission.

Conflict of Interest

The authors declare that they have no known Table 4competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work is partially supported by TUBITAK under grant number 5190074. The authors would like to thank Dr. Kerem Kucuk and Mahir Temur for their valuable comments on this work.

References

[1] Nagy I., Berend G., Vincze V., 2011. Noun compound and named entity recognition and their usability in keyphrase extraction. International Conference Recent Advances in Natural Language Processing, Hissar, Bulgaria, 12-14 September.

[2] Rodrigo A., Perez-Iglesias J., Penas A., Garrido G., Araujo L., 2013. Answering questions about European legislation. Expert Systems with Applications, **40**(15), pp. 5811-5816.

[3] Cao T. H., Tang T. M., Chau C. K., 2012. Text clustering with named entities: a model, experimentation and realization. In Data mining: Foundations and intelligent paradigms, Springer, Berlin, Heidelberg.

[4] Hassel M., 2003. Exploitation of named entities in automatic text summarization for Swedish. 14th Nordic Conference on Computational Linguistics, Reykjavik, Iceland, 30-31 May.

[5] Grishman R., Sundheim B. M., 1996. Message Understanding Conference – 6: A brief history. The 16th International Conference on Computational Linguistics, Copenhagen, Denmark, 5-9 August.

[6] Black W. J., Rinaldi F., Mowatt D., 1998. FACILE: Description of the NE System Used for MUC-7. 7th Message Understanding Conference, Fairfax, Virginia, 29 April – 1 May.

[7] Aone C., Halverson L., Hampton T., Ramos-Santacruz M., 1998. SRA: Description of the IE2 system used for MUC-7. 7th Message Understanding Conference, Fairfax, Virginia, 29 April – 1 May.

[8] Nadeau D., Turney P. D., Matwin S., 2006. Unsupervised named-entity recognition: Generating gazetteers and resolving ambiguity. 19th Canadian Conference on Artificial Intelligence, Quebec, Canada, 7-9 June.

[9] Özkaya S., Diri B., 2011. Named entity recognition by conditional random fields from Turkish informal texts. 19th Signal Processing and Communications Applications Conference, Antalya, 20-22 April.

[10] Lin W., Ji D., Lu Y., 2017. Disorder recognition in clinical texts using multi-label structured SVM. BMC bioinformatics, **18**(1), 75, pp. 1-11.

11] Zhang M., Geng G., Chen J., 2020. Semi-Supervised Bidirectional Long Short-Term Memory and Conditional Random Fields Model for Named-Entity Recognition Using Embeddings from Language Models Representations. Entropy, **22**(2), pp. 252.

[12] Zhu Q., Li X., Conesa A., Pereira C., 2018. GRAM-CNN: a deep learning approach with local context for named entity recognition in biomedical text. Bioinformatics, **34**(9), pp. 1547-1554.

[13] Korvigo I., Holmatov M., Zaikovskii A., Skoblov M., 2018. Putting hands to rest: efficient deep CNN-RNN architecture for chemical named entity recognition with no hand-crafted rules. Journal of cheminformatics, **10**(1), pp. 1-10.

[14] Li J., Sun A., Han J., Li C., 2020. A survey on deep learning for named entity recognition. IEEE Transactions on Knowledge and Data Engineering.

[15] Chen Y., Zhou C., Li T., Wu H., Zhao X., Ye K., Liao J., 2019. Named entity recognition from Chinese adverse drug event reports with lexical feature based BiLSTM-CRF and tri-training. Journal of Biomedical Informatics, **96**, pp. 103252.

[16] Leitner E., Rehm G., Moreno-Schneider J., 2019. Fine-grained Named Entity Recognition in Legal Documents. 15th International Conference on Semantic Systems, Karlsruhe, Germany, 9-12 September.

[17] Leaman R., Wei C. H., Lu Z., 2015. tmChem: a high performance approach for chemical named entity recognition and normalization. Journal of cheminformatics, **7**(S1), S3.

[18] Cucerzan S., Yarowsky D., 1999. Language independent named entity recognition combining morphological and contextual evidence. In 1999 joint SIGDAT conference on empirical methods in natural language processing and very large corpora.

[19] Küçük D., 2009. Named entity recognition experiments on Turkish texts. In International

Conference on Flexible Query Answering Systems. Springer, Berlin, Heidelberg.

[20] Cekınel R. F., Ağriman M., Karagöz P., Yilmaz B., 2019. Named Entity Recognition with Conditional Random Fields on Turkish News Dataset: Revisiting the Features. 27th Signal Processing and Communications Applications Conference, Sivas, Turkey, 24-26 April.

[21] Farmakiotou D., Karkaletsis V., Koutsias J., Sigletos G., Spyropoulos C. D., Stamatopoulos P., 2000. Rule-based named entity recognition for Greek financial texts. In Proceedings of the Workshop on Computational lexicography and Multimedia Dictionaries (COMLEX 2000).

[22] Sheikh M., Conlon S., 2012. A rule-based system to extract financial information. Journal of Computer Information Systems, **52**(4), pp. 10-19.

[23] Wang S., Xu R., Liu B., Gui L., Zhou Y., 2014. Financial named entity recognition based on conditional random fields and information entropy. In 2014 International Conference on Machine Learning and Cybernetics, Lanzhou, China, 13-16 July

[24] Alvarado J. C. S., Verspoor K., Baldwin T., 2015. Domain adaption of named entity recognition to support credit risk assessment. In Proceedings of the Australasian Language Technology Association Workshop 2015

[25] Bayraktar O., Temizel T. T., 2008. Person name extraction from Turkish financial news text using local grammar-based approach. In 2008 23rd International Symposium on Computer and Information Sciences (pp. 1-4). IEEE.

[26] Leitner E., Rehm G., Moreno-Schneider J., 2019. Fine-grained Named Entity Recognition in Legal Documents. In International Conference on Semantic Systems (pp. 272-287). Springer, Cham.

[27] Vardhan H., Surana N., Tripathy B. K., 2020. Named-Entity Recognition for Legal Documents. In International Conference on Advanced Machine Learning Technologies and Applications (pp. 469-479), Jaipur, India. Springer, Singapore, 13-15 February.