Name Entity Recognition System for Turkish Tweets

Buğse Erdoğan
Marmara University Faculty of Engineering
Department of Computer Engineering
Istanbul, Turkey
bugseerdogan@gmail.com

Fahriye Gün
Marmara University Faculty of Engineering
Department of Computer Engineering
Istanbul, Turkey
fahriyegun@gmail.com

Kübra Özgöç
Marmara University Faculty of Engineering
Department of Computer Engineering
Istanbul, Turkey
kubraozgoc@gmail.com

Abstract— **Named Entity Recognition which is an important subject of Natural Language Processing is a key technology of information extraction, information retrieval, question answering and other text processing applications. In recent years, Turkish NER intrigued researchers due to its scarce data resources and the unavailability of high-performing systems. Furthermore we propose a new association measure, and compare it with the other methods. The evaluation of these methods is performed by precision and recall measures. This paper reports the highest results (72% accuracy in MNB algorithm) in the literature for Turkish named entity recognition; more specifically for the task of detecting person, location and organization entities in general news texts. We believe, the paper draws light to the difficulty of these new domains for NER and the possible future work.**
Keywords-component; formatting; style; styling; insert (key words)

I. INTRODUCTION

Definition of Named Entity Recognition(NER) that classifies named entities in an unstructured text. NER classifies elements in text into predefined categories such as the names of people, organizations, locations etc. NER systems serve as an important pre-processing system for tasks such as information extraction, information retrieval, question answering and other text processing applications.

In NER systems, linguistic grammar-based techniques or statistical models have been used. Although grammar-based systems typically obtain better precision, these systems have lower recall and generally they are not independent from language. On the other hand, statistical NER systems require a large amount of manually annotated training data.

Also 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 in the literature for NER, the studies focused on real data is very limited and recent.

There have been a lot of work on NER systems on several languages especially on English. Although new models and techniques have been suggested, developing NER systems for Turkish is still a difficult process because of the agglutinative structure of the language however in recent years, there have been many studies for Turkish NER.

In this study, we evaluate previously well-established machine learning system in order to extract named entities in Turkish corpus.

Designing a more dynamic structured NER system for Turkish language by using existing methods, analyzing tweets by implementing a user interface and finding unknown named entities in unstructured text compose scope of our project.

We focus on the NER and raise the level of knowledge about NER with reading these articles. Later we examine English benchmark and learn to structures and tag standards in detail. After research about NER and related works in detail, we research about Turkish benchmark and read tweets from mongodb, establish to tag system to make tagging by improving a machine learning based system.

If we examine software part of our project, firstly we create interface using ASP.NET and C#. We take tweets with Twitter API and collect them in a dataset at Mongodb. After we read every tweet word by word and tag some words within the framework of some categories that we identify in C#. This tagging operation operates in the background of the code how implement tag into category of each word manually.

This project is proceed with dataset from Twitter. Tagging are limited in 7 categories that are shown below:

- Person
- Location
- Organization
- Date
- Time
- Money
- Percent

Tagging is the most important part of our project. For this reason, successful rate of this project is increased because of as much as possible caching true word in NER system. For this reason, every word cannot be tagged in this categories.

Based on the problem description and the objectives, the following assumptions are taken: Search, download, install and test the existing NER systems and a program need to be developed to test the tweets' data sets. In previous studies, many algorithms are proposed to solve this problem. Some of them give numerical results, and some of them even do not have any numerical results. The best solution approach proposed in the literature gives a above 85% success rate. However, since the problem considers human life, it is very important to get more accurate results. In our project, we try to find a better approach that gives higher accuracy. Moreover, we plan work on different databases.

The benefit of our project is that be contributed with easier, faster and more word tagging to science about NER surveys. We believe that it keeps light to many research like NER, Information Extraction, NLP.

A. Abbreviations and Acronyms

*NER:*Named Entity Recognition
*NE:*Named Entity
*NLP:*Natural Language Processing
IE:Information Extraction
NEEL:Named Entity Recognition and Linking
*POS:*Part of Speech
MI: Mutual Information
IG: Information Gain
TwitIE:An open-source NLP pipeline customised to text at every stage
MUC: Message Understanding Conference
CoNLL: Conference on Computational Natural Language Learning

II. RELATED WORKS

We study on machine learning based systems in NER for Turkish tweets because there are limited works on Turkish language but also there are many systems that apply NER with different systems on different language from all over the world. We will discuss details about other works related our project below. Named Entity Recognition (NER) can be basically defined as identifying and categorizing certain type of data (i.e. person, location, organization names, date-time expressions). NER is an important stage for several natural language processing (NLP) tasks including machine translation, sentiment analysis and information extraction. MUC (Sundheim, 1995; Chinchor and Marsh, 1998) and CoNLL (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003) conferences define three basic types of named entities; these are:

1- ENAMEX (person, location, organization)
2- TIMEX (date and time entities)
3- NUMEX (numerical expressions)

But NER research is not limited to only these types; different application areas concentrate to determining alternative entity types such as protein names, medicine names, book titles. The NER research was firstly started in early 1990s for English. In 1995, with the high interest of the research community, the success rates for English achieved nearly the human annotation performance on news texts (Sundheim, 1995). Nadeau and Sekine (2007) gives a survey of the research for English NER between 1991 to 2006. The satisfaction on English NER task directed the field to new research areas such as multilingual NER systems (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003), NER on informal texts (LIU et al., 2011; Rüd et al., 2011; Mohit et al., 2012), transliteration (Zhang et al., 2012) and co-reference (Na and Ng, 2009) of named entities.

Another of the first research papers in the field was presented by Lisa F. Rau [6]. Since then, many methods and strategies for automatic identification of named entities have been proposed. Methods for NER systems classify into 3 groups: the rule based approach, probabilistic approach, and the hybrid approach.

In the rule-based approach, the natural language descriptions and rules need to be formulated. The rules are used to define name entities using their syntactic and lexical structure with the help of manually annotated corpora. In addition to rules, rule-based approaches require gazetteers and general dictionary [7]. (Content of a gazetteer which includes a subject's location, dimensions of peaks and waterways, population and literacy rate.). Differently from rule-based approach, the probabilistic approach does not require any natural language information. This approach builds their models by learning patterns from the annotated corpora [7]; and the approach displays good enough performance with large corpora [8][9]. In [10] it is indicated that recent studies about NER are mostly based on probabilistic methods.

Although some studies address language independence or multilingualism in NER solutions, a large part of the NER studies are on English. NER studies on other languages than English have been also carried out; such as German [11] Spanish [12], Japanese [13] [14] [15], Chinese [16] [17] [18], French [19] [20], Greek [21], Italian [22] [23], Bulgarian [24], Hindi [25], Polish [26], Russian [27], Swedish [28], Portuguese [29].

There is a limited work on NER systems applied on Turkish texts. Study of Cucerzan and Yarowski [24] is the first study on Turkish NER. In [24], a language independent bootstrapping algorithm that learns from word internal and contextual information of entities is presented and the proposed algorithm is experimented on five languages including Turkish. Following this, in [30] a statistical approach (HMMs) for NER is used on Turkish texts. In [31], a huge database of person, organization, and location names is constructed instead of employing a complex name entity extraction scheme. As a recent study, Kucuk and Yazici [32] presented a rule-based NER system for Turkish. In this study, they presented a rule-based system for named entity recognition from Turkish texts. It is initially engineered for news texts, employs a set of lexical resources and pattern bases, being a rule-based system, needs no training data and evaluated on diverse text types including news texts, child stories, historical texts, and news video transcriptions

Finally, CRF-based approaches for NER in Turkish are proposed in [33]. These NER systems for Turkish are mostly proposed and tested on news articles and the CRF based system in [33] is reported to outperform the other proposals.

III. METHODOLOGY AND APPROACH

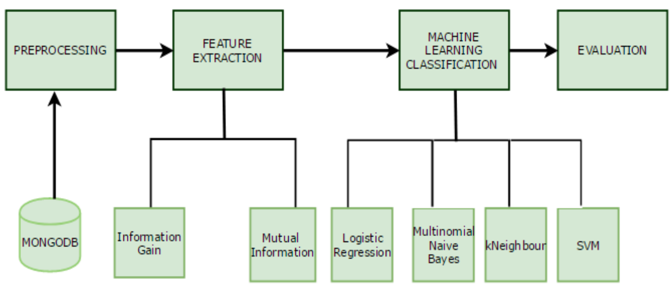

Figure 1. Schema Of Methodology

We told the way we watched on the above diagram. Firstly preprocessing part we did:

- Data gathering : We have collected tweets from people who tweeted on trend topics with using Twitter API.
- Data filtering: Python da We deleted repeated tweets and introduced tweets to understand our a system of tagging.
- Data bringing to tagged format: We develop a word tag system with using C# and ASP.NET. We label each word individually in Turkish data

We have approximately 5000 tweets in total and around 30,000 words. We have added some features to make class prediction easier once the data becomes available. These features are based on tweet and word. We evaluated the word bases by looking at words, words before and after. Our goal here is to make it easier to guess in special tags that contain more than one word. If we give an example: the word 'firat' is only person class in its own right, and 'firat nehri' class is location. We tried to guess the class by looking after the word to ensure this too. Our features are these:

#	Name	Abbrevation	Description
1		W_isCapital	does Word start with Capital Letter?
2	Is Capital	WA_isCapital	does After Word start with Capital Letter?
3		WB_isCapital	does Before Word start with Capital Letter?
4		W_isAllCapital	does Word includes all letters Capital?
5	Is All Capital	WA_isAllCapital	does After Word includes all letters Capital?
6		WB_isAllCapital	does Before Word includes all letters Capital ?
7		W_length	Length of word
8	Letter	WA_length	Length of word after
9		WB_length	Length of word before
10		W_hasEmoticon	does Word have emoticon?
11	Has Emoticon	WA_hasEmoticon	does After Word have emoticon?
12		WB_hasEmoticon	does Before Word have emoticon?
13		W_hasPunctuation	does Word have punctuation?
14	Has Punctuation	WA_hasPunctuation	does After Word have punctuation?
15		WB_hasPunctuation	does Before Word have punctuation?
16		W_hasHashtag	does Word have hashtag?
17	Hashtag	WA_hasHashtag	does After Word have hashtag?
18		WB_hasHashtag	does Before Word have hashtag?
19		W_hasURL	does Word have URL?
20	URL	WA_hasURL	does After Word have URL?
21		WB_hasURL	does Before Word have URL?

Table1: word based features
(W =word, WA=word after, WB=word before)

#	Name	Abbrevation	Description
1	has Punctiaton	T_hasPunctuation	does tweet have puntuation?
2	has Emoticon	T_hasEmoticon	does tweet have emoticon?
3	has Hashtag	T_hasHashtag	does tweet have hashtag?
4	has Mention	T_hasMention	does tweet have mention?
5	has RT	T_RT	does tweet have retweet?
6	has URL	T_hasURL	does tweet have URL link?
7	Word Count	T_wordCount	How many word has in tweet?
8	Length of Tweet	T_tweetLength	Length of tweet

Table2: tweet based features
(T=tweet)

After implementing these features in Python, the general usage distribution is as follows:

Abbrevation	Mode
T_hasPunctuation	TRUE
T_hasEmoticon	FALSE
T_hasHashtag	FALSE
T_hasMention	TRUE
T_RT	TRUE
T_hasURL	FALSE
T_wordCount	17
T_tweetLength	109

Table3: distribution of tweet based features

Attributes	Mode
W_isCapital	FALSE
WA_isCapital	FALSE
WB_isCapital	FALSE
W_isAllCapital	FALSE
WA_isAllCapital	FALSE
WB_isAllCapital	FALSE
W_length	5
WA_length	5
WB_length	5
W_hasEmoticon	FALSE
WA_hasEmoticon	FALSE
WB_hasEmoticon	FALSE
W_hasPunctuation	TRUE
WA_hasPunctuation	TRUE
WB_hasPunctuation	TRUE
W_hasHashtag	FALSE
WA_hasHashtag	FALSE
WB_hasHashtag	FALSE
W_hasURL	FALSE
WA_hasURL	FALSE
WB_hasURL	FALSE

Table4: distribution of word based features

After implementing them, we calculated Mutual Information (MI) and Information Gain (IG) values in python. Our aim was to look at the relationship between features and classes, and we got very interesting results.

#	Attributes	IG Value	IG Rank	MI Value	MI Rank
1	T_tweetLength	0,983801444	1	0,729977407	1
2	T_WordCount	0,981349577	2	0,729707143	2
3	WB_Length	0,978992876	3	0,537996621	3
4	WA_Length	0,978991698	4	0,537972361	5
5	W_Length	0,978991698	5	0,537986351	4
6	W_hasPunctuation	0,978991555	6	0,483368045	6
7	WA_hasPunctuation	0,978991555	7	0,483336804	7
8	WB_hasPunctuation	0,978991512	8	0,483320607	8
9	W_isCapital	0,978991499	9	0,474989448	13
10	WA_isCapital	0,978991499	10	0,474989448	14
11	WB_isCapital	0,978991498	11	0,475011868	12
12	W_isAllCapital	0,978881591	12	0,469017201	15
13	WB_isAllCapital	0,978881591	13	0,469017201	16
14	WA_isAllCapital	0,978413063	14	0,469014163	17
15	W_hasHashtag	0,749356617	15	0,47731818	9
16	WA_hasHashtag	0,749356617	16	0,47731818	10
17	WB_hasHashtag	0,749356617	17	0,47731818	11
18	T_RT	0,675005776	18	0,46787835	18
19	W_hasEmoticon	0,672973079	19	0,466469392	19
20	W_hasURL	0,672973079	20	0,466469392	20
21	WA_hasEmoticon	0,672973079	21	0,466469392	21
22	WA_hasURL	0,672973079	22	0,466469392	24
23	WB_hasEmoticon	0,672973079	23	0,466469392	22
24	WB_hasURL	0,672973079	24	0,466469392	23
25	T_hasHashtag	0,537564196	25	0,372611107	25
26	T_hasMention	0,515733238	26	0,35747904	26
27	T_hasPunctuation	0,315733238	27	0,331979552	27
28	T_hasEmoticon	0,277367585	28	0	28
29	T_hasURL	0,245261235	29	0	29

Table5: IG and MI ranks of features between class

This table shows us that our most useful class predictor feature is our tweet length. This is actually a very interesting result. Other than this, word length, words have punctuation marks, etc. Word-based features were expected values at the top. The class distribution of the tags of words is as follows:

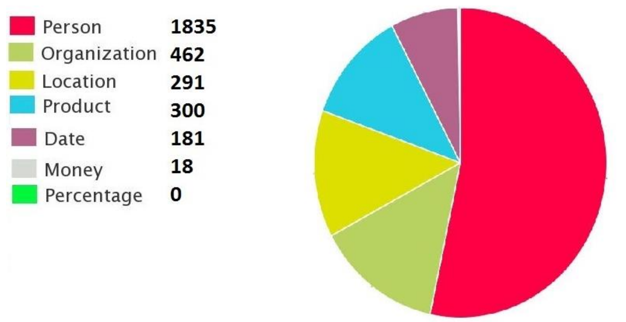

Figure 2. Class Distribution

After finishing the Feature Extraction part, we have been using ML algorithms in three different ways. These are test-on-training data set, train 60%- test40%, 10-fold CV. The algorithms we use are Logistic Regression (LR), Multinomial Naive Bayes (MNB), k-Nearest Neighbors (k-NN), Support Vector Machine (SVM). Our results are as follows:

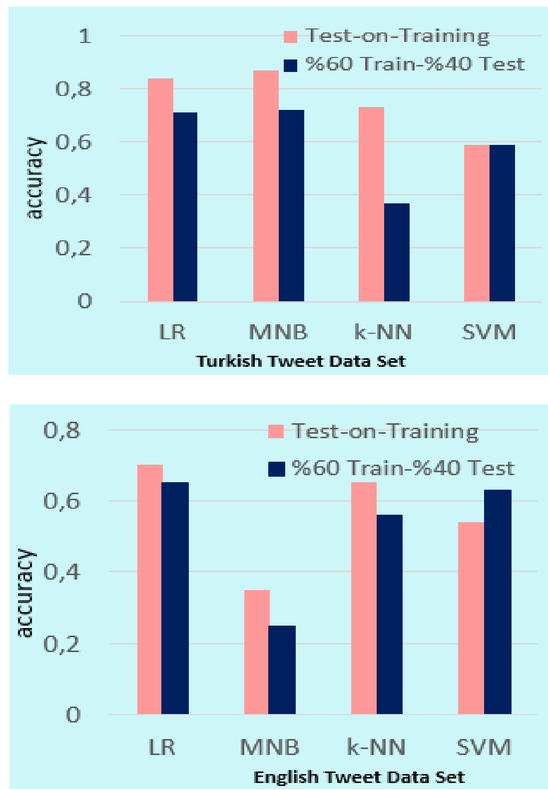

Figure 3. Our Result in ML Algorithms

We compared our own results with those of previous years.

Study	Precision	Recall	F1
A feature based approach performing Stanford NER [3]	0.729	0.626	0.674
Stanford NER, MITIE, twitter_nlp and TwitIE, [4]	0.587	0.287	0.386
TwitIE (CRF Model), [2]	0.435	0.459	0.447
Logistic Regression,5 features + 7 word2vec features, 7 NER classes), [1]	0.71	0.56	0.58
Our approach(MNB,30 features,7 NER Types)	0,68	0,59	0,61

Figure 4. Comparison of the performance with respect to the studies presented in NEEL 2016 workshop [5]

IV. CONCLUSION

Our contribution in this paper is mainly the creation of new NE datasets from different real data and the presentation of the first NER results on them. In this study, we present a tweet data set in Turkish which is annotated with named entities. 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 tweet data set.

As a result of our experiments with four different machine learning algorithms, we took best accuracy from Multinomial Naive Bayes. When we compare our solutions with a tweet data set in English, Logistic Regression is the best machine learning classifier for English data set.

REFERENCES

[1] 1. Taşpınar,M.,Ganiz,M.C.,Acarman,T., A Feature Based Simple Machine Learning Approach with Word Embeddings to Named Entity Reconition on Tweets,NLDB 2017

[2] 2. P. Torres-Tramon, H. Hromic, B. Walsh, B. Heravi, and C. Hayes. Kanopy4Tweets: Entity Extraction and Linking for Twitter.In 6 th International Workshop on Making Sense of Microposts (#Microposts), 2016

[3] 3. S. Ghosh, P. Maitra, and D. Das. Feature Based Approach to Named Entity Recognition and Linking for Tweets. In 6 th International Workshop on Making Sense of Microposts (#Microposts), 2016

[4] 4. K. Greenfield, R. Cacares, M. Coury, K. Geyer, Y. Gwon, J. Matterer, A. Mensch, C. Sahin, and O. Simek.A Reverse Approach to Named Entity Recognition and Linking in Microposts. In 6 th International Workshop on Making Sense of Microposts, 2016

[5] 5. G. Rizzo, M. Van Erp, J. Plu, and R. Troncy. Making Sense of Microposts (#Microposts2016) Named Entity Recognition and Linking (NEEL) Challenge. .In 6 th International Workshop on Making Sense of Microposts (#Microposts2016), pages 50-59, 2016.

[6] L. F. Rau, 1991. Extracting Company Names from Text. In Proc. Conference on Artificial Intelligence Applications of IEEE.

[7] H. N. Traboulsi, “Named Entity Recognition: A Local Grammarbased Approach,” unpublished Ph.D. dissertation, Department of Computing School of Electronics and Physical Sciences,University of Surrey Guildford, Surrey GU2 7XH, U.K, 2006

[8] D. M. Bikel, S. Miller, R. Schwartz, and R. Weischedel, “Nymble: a High-Performance Learning Name-finder,” in Proceedings of the fifth conference on Applied natural language processing, pp. 194 - 201, 1997.

[9] Borthwick, J. Sterling, E. Agichtein, and R. Grishman, “NYU: Description of the MENE Named Entity System as Used in MUC-7,” in Message Understanding Conference (MUC-7), 1998.

[10] D. Nadeau & S. Sekine, “A survey of named entity recognition and classification,“ Ed. S. Sekineand E. Ranchhod, unpublished Technical Report, 2007.

[11] C. Thielen,. 1995. An Approach to Proper Name Tagging for German. In Proc. Conference of European Chapter of the Association for Computational Linguistics.

[12] X. Carreras; L. Márques, L. Padró. 2003. Named Entity Recognition for Catalan Using Spanish Resources. In Proc. Conference of the European Chapter of Association for Computational Linguistic.

[13] M. Asahara, Y. Matsumoto, 2003. Japanese Named Entity Extraction with Redundant Morphological Analysis. In Proc. Human Language Technology conference – North American chapter of the Association for Computational Linguistics.

[14] S. Sekine, 1998. Nyu: Description of the Japanese NE System Used For Met-2. In Proc. Message Understanding Conference.

[15] S. Sekine, Isahara, H. 2000. IREX: IR and IE Evaluation project in Japanese. In Proc.Conference on Language Resources and Evaluation.

[16] L-J Wang, W-C Li, C-H, Chang. 1992. Recognizing Unregistered Names for Mandarin Word Identification. In Proc. International Conference on Computational Linguistics.

[17] H. H. Chen, J. C. Lee, 1996. Identification and Classification of Proper Nouns in Chinese Texts.In Proc. International Conference on Computational Linguistics.

[18] S. Yu, S. Bai, P. Wu. 1998. Description of the Kent Ridge Digital Labs System Used for MUC-7. In Proc. Message Understanding Conference.

[19] G. Petasis, F. Vichot, F. Wolinski, G.Paliouras, V. Karkaletsis, C.D. Spyropoulos. 2001. Using Machine Learning to Maintain Rule-based Named-Entity Recognition and Classification Systems. In Proc. Conference of Association for Computational Linguistics.

[20] T. Poibeau. 2003. The Multilingual Named Entity Recognition Framework. In Proc.Conferenceon European chapter of the Association for Computational Linguistics.

[21] S. Boutsis; I. Demiros, V. Giouli, M. Liakata, H. Papageorgiou, S. Piperidis. 2000. A System for Recognition of Named Entities in Greek. In Proc. International Conference on Natural Language Processing.

[22] W.J. Black, F. Rinaldi, D. Mowatt. 1998. Facile: Description of the NE System used for Muc-7.In Proc. Message Understanding Conference.

[23] A. Cucchiarelli, P. Velardi, P. 2001. Unsupervised Named Entity Recognition Using Syntacticand Semantic Contextual Evidence. Computational Linguistics 27:1.123- 131,Cambridge: MITPress.

[24] S. Cucerzan, D. Yarowsky. 1999. Language Independent Named Entity Recognition Combining Morphological and Contextual Evidence. In Proc. Joint Sigdat Conference on Empirical Methods in Natural Language Processing and Very Large Corpora.

[25] J. May, A. Brunstein, P. Natarajan, R.M. Weischedel. 2003. Surprise! What’s in a Cebuano or Hindi Name? ACM Transactions on Asian Language Information Processing 2:3.169-180, NewYork: ACM Press.

[26] J. Piskorski. 2004. Extraction of Polish Named-Entities. In Proc. Conference on Language Resources an Evaluation.

[27] B. Popov, A. Kirilov, D. Maynard, D. Manov. 2004. Creation of reusable components and language resources for Named Entity Recognition in Russian. In Proc. Conference on Language Resources and Evaluation.

[28] D. Kokkinakis. 1998., AVENTINUS, GATE and Swedish Lingware. In Proc. of Nordic Computational Linguistics Conference.

[29] D. D. Palmer, S. Day. 1997. A Statistical Profile of the Named Entity Task. In Proc. ACL Conference for Applied Natural Language Processing.

[30] G. Tür, D. Hakkani-Tür, and K. Oflazer. 2003. A Statistical Information Extraction System for Turkish. Natural Language Engineering. Vol. 9 No 2 181-210

[31] K. Oflazer, Ö. Çetinoglu, B. Say, “Integrating Morphology with Multiword Expression Processing in Turkish,” in Second ACL Workshop on Multiword Expressions: Integrating Processing, pp. 64 - 71, 2004.

[32] D. Kucuk, A. Yazici. 2009. Named entity recognition experiments on Turkish texts. In Proceedings of the 8th International Conference on Flexible Query Answering Systems, FQAS’09, pages 524–535, Berlin, Heidelberg. Springer-Verlag.

[33] Gokhan A. Seker and Gulsen Eryigit. 2012.Initial Explorationson Using CRFs for Turkish Named Entity Recognition.In Proceedings of the International Conference on Computational Linguistics, pages 2459–2474.