Named Entities Recognition of Tunisian Arabic Using the Bi-LSTM-CRF Model

Anonymous ACL-IJCNLP submission

Abstract

Named Entity Recognition (NER) is an NLP field that deals with recognizing and classifying entities in written text. Most Arabic NER research studies discuss the Arabic NER challenge for the Modern Standard Arabic (MSA) language. However, the presence of dialectal Arabic textual resources in social media, blogs, TV shows, etc. is increasingly progressive. Therefore, the treatment of named entities is rapidly becoming a necessity, particularly for dialectal Arabic. In this paper, we are interested in the collection and annotation of a corpus as well as the realization of a NER system for Tunisian Arabic (TA), named TUNER. To the best of the researchers' knowledge, this is the first study that uses the suggested method for this purpose. In the present study, we adopt a hybrid method based on a Bi-LSTM-CRF model and a rule-based method. The proposed TUNER system yields an F-measure of 91.43%. This is an interesting improvement over comparable related work dialectal Arabic NER systems.

1 Introduction

Nowadays, a significant amount of textual data is available on public web repositories such as online news, social media, blogs, etc. The textual information is provided as unstructured data without a predefined data model or annotation form. Indeed, working with textual data is challenging along with its variability as to how people express themselves using various writing styles (i.e. depending on words pronunciation or using forms which are closer to the standard form).

Information extraction is a process of extracting information automatically from unstructured data and is generally concerned with the processing of human language text by means of natural language processing (NLP). Named Entities Recognition (NER) is a sub-task of the process of information extraction. It includes searching for textual content (i.e. a word or a group of words), which is categorized in classes such as names of people, names of organizations or companies, names of locations, quantities, distances, values, dates, etc. However, Tunisian Dialect (TD) NLP is attracting more attention as a result of massive use of social networking and web blogs, requiring a higher demand for TD NER systems. These tools can help with a variety of tasks, including information retrieval, question answering, and even machine translation. Furthermore, employing NLP tools (such as NER) built for MSA on dialectal Arabic in general and TD in particular, results in significantly low-performance (Habash et al., 2012). Therefore, resources and tools specialized to TD are required.

In this paper, we propose a hybrid method for dialectal Arabic (DA) NER, in particular Tunisian Dialect (TD). We extract and classify four classes of Tunisian named entities (NEs): organization (ORG), person (PER), location (LOC) and date (DATE). First, we suggest to use the Bi-LSTM-CRF architecture (Huang et al., 2015) to generate a deep learning-based model. Then, we applied a set of handcrafted rules to improve the result. We used the TD pre-processing tools (i.e. orthographic normalizer, sentence segmentation (Zribi et al., 2016)) and syntactically annotated TTB (Mekki et al., 2020). We have also collected and preprocessed a textual dataset from various sources to carry out the necessary experiments. The collected corpus is POS tagged using the TD parser Stanford-TUN (Mekki et al., 2020). Then, we proceeded to annotate these corpora for the

NER task.

This paper is structured as follows. Section 2 presents a brief overview of the NER related works. In the second part, Tunisian Dialect characteristics are examined. In Section 3, the details of our hybrid method to prepare the TUNER system are explained. Section 4 is devoted to describe the data set used to train and evaluate our system. In Section 5, we present the different experiments carried out with TUNER system and finally we draw our conclusions and suggest future works.

2 Related work

2.1 Modern Standard Arabic

In the literature, the three most widely adopted approaches have been proposed for conducting Arabic NER (i.e. rule-based methods, machine learning-based methods and hybrid methods).

The rule-based approach for Arabic NER depends on various features combinations such as contextual lexical triggers (Mesmia et al., 2018), morphological analysis (Aboaoga and Ab Aziz, 2013; Elsebai et al., 2009), gazetteers (Zaghouani, 2012; Elsherif et al., 2019)), etc. However, most of these methods were tested using a very restricted test set. Therefore, the effectiveness of these tools remains to be examined using other corpora, especially when evaluating them on defined benchmarks for the task of Arabic NER. However, RENAR (Zaghouani, 2012) and (Elsherif et al., 2019) were tested using ANERcorp dataset.

Many researchers adopt machine-learning methods to handle this task. Several NER models were proposed using Conditional Random Fields (CRF) (Alsayadi and ElKorany, 2016; Abdul-Hamid and Darwish, 2010) while others are based on Support Vector Machines (SVM) (Abdelali et al., 2016; Pasha et al., 2014; Koulali and Meziane, 2012). The recent CAMeL tools (Obeid et al., 2020) applied Random Forest algorithm for Arabic NER. These methods used diverse combinations of features including lexical, gazetteer, contextual, morphological, POS and syntactic features. Moreover, deep learning method has been considered for this task by (Khalifa and Shaalan, 2019; Liu et al., 2019; Ali et al., 2019).

The hybrid method combines rule-based and machine-learning based techniques. Among the developed tools, (Abdallah et al., 2012; Oudah and Shaalan, 2012) integrated the output of the rule-based system as features to machine-learning classifier J48 (Quinlan, 2014). (Meselhi et al., 2014) applied the same idea with SVM algorithm. (Oudah and Shaalan, 2017) proposed independent rule-based and machine learning-based components that can operate separately. The machine learning method is based on Decision Tree algorithm proposed by WEKA[1]. (Alotaibi and Lee, 2014) presented two methods that rely on the output of the dependency parser and the clustering algorithm CRF. (Hkiri et al., 2017) generated a CRF model using bilingual NE lexicon and grammar rules. (Khalifa and Shaalan, 2019) adopted the existing LSTM neural tagging model for Arabic NER with a Convolutional Neural Network (CNN) for extracting the character-level features. The majority of existing Arabic NER research centered on extracting entities from MSA. However, due to the unstructured character of the dialectal Arabic utilized in social media, MSA NER systems perform poorly. Therefore, in the next section we present NER tools proposed for dialectal Arabic.

2.2 Dialectal Arabic

We present in the rest of this section the state-of-the-art methods proposed for dialectal Arabic.

(Zirikly and Diab, 2015) adopted a gazetteers-free method using CRF (Lafferty and Mccallum, 2001) for Egyptian NER. They used the corpora proposed by (Darwish, 2013) and (Zirikly and Diab, 2014). As a result, the train corpus contains 55K tokens with 1,950 NEs and 24K tokens as a test set with 485 NEs. (Zirikly and Diab, 2015) processed for normalization and preprocessing step. In addition, MADAMIRA was used for word tokenization and morphological features. The proposed method gives an overall F1 result of 72.68% using the corpus of (Zirikly and Diab, 2014).

(Gridach, 2016) introduced a neural network architecture using combination of bidirectional Long Short-Term Memory (LSTM) and CRF. CRF layer is employed on the top of the Bi-LSTM in order to identify contextual features in the form of neighboring NER tags. They

[1] https://www.cs.waikato.ac.nz/ml/weka.

also used the twitter dataset that contains MSA and Egyptian tweets (Darwish, 2013). The best obtained F1 result is equal to 85.71%.

(Sabty et al., 2019) annotated a code switched Egyptian-English corpus for NER. This corpus comprises 1,331 sentences (i.e. where 884 sentences (66.4%) contain NEs) gathered from three different sources (transcribed speech, Twitter and the Egyptian translation of ANERCorp (Benajiba et al., 2007)). To evaluate the collected corpus, they applied the method of (Gridach, 2016) using a test set composed of 33,890 words (Egyptian/English), two MSA corpora (Benajiba et al., 2007; Mohit et al., 2012) and CoNLL 2003 for English. The final NER system achieved a F1-score of 60%.

(Torjmen and Haddar, 2021) proposed a linguistic method based on a bilingual dictionary and an elaborated set of local grammars. First, they analyzed the TD corpus by the morphological analyzer (Torjmen and Haddar, 2018). Then, the ANER and translation systems recognize the NEs and translate them into MSA using bilingual dictionaries, gazetteers (1 863 words), and syntactic grammars. The proposed method gives an overall F1 result of 92.05% using 20K words as a test set.

3 Tunisian Dialect

The lexical dialectal system appears to be more open and reflects a rich particular vocabulary due to the mix of different cultures over several centuries when compared to MSA. TD vocabulary is characterized by the frequent usage of words imported from other languages. Thus, we find words from Turkish, Italian, French, Berber and other languages. For example, (منوبة, mnwbp) is a state name of Punic origin that means *on duty* in MSA. Moreover, in contrast to MSA the same TD word can be written in a variety of ways. All of these factors show the importance of TD NER system development. Attached clitics are also an important criterion when attached to the following proper nouns (e.g. كمنوبة, kmnwbp, *such as Manouba*). It is an agglutination case that needs to be resolved before NE can be detected.

3.1 Tunisian Dialect forms

In the literature, we can discern between three forms of TD (Mekki et al., 2018).

- Social Media Dialect (SMD) is the most available form of dialect. It is characterized by the use of onomatopoeia, emojis, abbreviations, etc. Moreover, Internet users tend to use nicknames in their comments. For example, Tunisians are more inclined to call the Tunisian president (باجي قائد السبسي, bAjy qA}d Alsbsy, *Béji Caïd Essebsi*) with his nickname (بجبوج, bjbwj).
- Intellectualized Dialect (ID) can be used in interviews on TV and radio programs. It is the closest type of dialect to MSA. However, the NE is more likely to be used in its complete form without any abbreviations or nicknames. It is also noticeable that Tunisians frequently use introductory words before the NE such as (السيد رئيس الجمهورية باجي قائد السبسي, Alsyd r}ys Aljmhryp bAjy qA}d Alsbsy, *Mr. President of the Republic Béji Caïd Essebsi*).
- Spontaneous Dialect (SD) is the spoken form of TD. It is distinguished by code switching (TD, French), the presence of disfluencies, etc. SD is a combination of both previous forms of dialects. We can find some corpora where the NEs are quite similar to the ID with the frequent use of introducing words (e.g. titles prefixing a person's name) and full names. Other SD corpora are more informal such as the SMD form. These differences are highlighted according to the source of the transcribed corpus. For example, if the corpus was transcribed from a political TV interview, the conversation would be more formal than an entertainment show.

3.2 Tunisian Dialect characteristics

Given the particularity of Arabic Language in general and TD in particular, applying NER task is very challenging. Some challenges are specific for the Arabic language, while others are common to all languages. Even if the challenges were shared, the NER systems could not be applied to TD given the lexical differences between MSA and TD. Using MSA NLP tools on TD leads in significantly low quality results (Habash et al., 2012), necessitating the development of resources and tools targeted exclusively towards TD.

- **No capitalization** - Unlike Latin languages where the proper noun begins with a capital letter, Arabic script does not cover the capitalization feature.
- **The agglutinative nature** - Arabic has a very complex morphology owing to its high agglutinative nature. An Arabic word consists of different combinations of prefixes, stem and suffixes. For example, (وتعرفهاشي, wtErfhA$y, *and did you know her*) is a TD phrase composed of a prefix (و + stem (تعرف) + two suffixes (ها and شي).
- **No diacritics** - Pronunciation and disambiguation highly depend on the use of short vowels or diacritics. Nevertheless, DA texts do not include diacritics. Therefore, an Arabic phrase may refer to several meanings according to the context it appears in, creating a higher ambiguity for NER task. For instance, the word (براد, brAd) can be pronounced in TD as a noun (بَرَّاد, baraãAd, *Teapot*) or a NE (برَاد, braAd, *Brad*).
- **Confusion** - Many streets are named after scholars and politicians, after their historical and intellectual value. Therefore, locations and person names could be the same, which makes the automatic distinction between both classes a challenging task (e.g. شارع الحبيب بورقيبة, $ArE AlHbyb bwrkybp, *Habib Bourguiba*[2] *Street*). The same idea is applied on companies or business organizations, which took the name of their owners.
- **Orthographic errors** - In dialectal Arabic, people write what they want without any constraint. They may write words as they are pronounced or by making it quite similar to the one of MSA, which may render this task more complicated.
- **Nicknames** - Nicknames are often used by social media users as a familiar or ironic name assigned to a person or organization (e.g. a football team) instead of or in addition to the original name. Unfortunately, these names can also be used for bullying.
- **Percentage of Tunisian NEs in ANER Corpus** - NER corpus «ANERCorp» (Benajiba et al., 2007) is one of the most widely used Arabic NER resources (El-Haj and Koulali, 2013; Hkiri et al., 2017; El Bazi and Laachfoubi, 2019; Alsaaran and Alrabiah, 2021) . However, the percentage of NEs found in our corpus and seen in ANERCorp was only 27% compared to the dataset and 9.17% compared to the test set.
- **Lack of linguistic resources** - To the best of the researchers' knowledge, no TD NER resources (annotated corpora, gazetteers, etc.) are publicly available for research purposes.

4 Proposed method

In this paper, we proposed a hybrid method for the implementation of TUNER system. We started by the preparation of the dataset by preprocessing the text and annotating the instances. Then, we generated Bi-LSTM-CRF model. After that, we prepared a set of handcrafted rules to improve the results. In the rest of this section, we present in details the steps of our hybrid method.

4.1 Deep learning based component

In this section, we describe our method, which is based on Bi-LSTM-CRF architecture. To process NEs efficiently, (Huang et al., 2015) proposed this architecture. Indeed, it is among the most used treatments of NER. However, a huge dataset is required to apply language models like AraBERT (Antoun et al., 2020). As a result, we are unable to apply this method at this stage since TD is still a low-resource language.

4.1.1 Data collection

There are many textual resources proposed for TD. To select the best corpora to use, we have relied on certain factors such as the corpora accessibility, the availability of the annotation,

[2] Habib Bourguiba was the first president of Tunisian who led the country from 1957 to 1987

the size and the diversity of the sources from which it was collected. We present in the next section the textual resources that we have used for the creation of an annotated corpus for the NER.

- Tunisian TreeBank (TTB) (Mekki et al., 2017) is a set of corpora syntactically annotated by Tunisian experts. This Treebank is fully normalized according to the orthographic convention CODA-TUN (Zribi et al., 2014). All corpus sentences are well segmented and the clitics are manually tokenized. TTB contains 10,000 SMD sentences from (Younes et al., 2015), 928 ID sentences (12K tokens) from (El Klibi et al., 2014) and 1,072 SD sentences (10K tokens) from STAC (Zribi et al., 2015).
- (Younes et al., 2015) collected 151K words written in Arabic letters from Facebook comments, mobile phone messages, etc. This corpus is characterized by a high use of nicknames. Furthermore, people frequently change certain letters of the word or even an entire word, in negative comments, to intimidate people or organizations. It was not normalized according to any conventional orthography. Only a part of the corpus was normalized by (Mekki et al., 2020) as mentioned above.
- Tunisian Media Corpus (TMD) (Boujelbane et al., 2014) is a SD corpus collected from TV news and political debate broadcasts. Therefore, the NEs are presented in their complete form without any abbreviation and are generally preceded by gazetteers. TMD contains 38K words (5 hours and 20 minutes of transcripts). This corpus is normalized according to the CODA-TUN orthographic convention (Zribi et al., 2014).

The TD treebank TTB (Mekki et al., 2020) includes 122K tokens annotated manually and POS tagged. In this paper, we use this corpus as a training set using the POS tags as one of our features set. For development and test, we collected randomly 30K tokens (20% of the train set) from TAD (Younes et al., 2015) and TMD (Boujelbane et al., 2014) corpora. We started by normalizing this dataset according to the orthographic convention CODA-TUN (Zribi et al., 2014). Then, we corrected the segmentation using STAr-TUN system (Zribi et al., 2016) in order to simplify the parsing step afterwards. Moreover, we tokenized all the prefixes and suffixes of the collected corpus. Table 1 details the number of instances for each class per dataset.

Table 1: Data set size statistics per classes

Class	Train set	Dev. set	Test set
O	114,618	13,159	13,566
PER	4,143	1,036	1,066
PLC	1,109	470	329
ORG	2,266	339	304
DATE	139	46	52

4.1.2 Annotation process

In this part, we expose the steps of annotation of the collected corpora with the exception of the TTB already annotated (see section 4.1.1).

NER classification. There are different categories to be considered as named entities and how broad those categories should be. Message Understanding Conference[3] defined three subtasks for NER: ”ENAMEX” tags are used for names (i.e. organization, person and location), ”NUMEX” tags are used for money and percent entities, and ”TIMEX” tags are used for temporal entities (i.e. time and date). However, the categories selected for a given NER project depend on the project’s goals. For instance, if geographical categorization is important, then the categories dealing with location data may need to be more refined such as city, country, state, river, etc. We noticed that the majority of works proposed for Arabic language such as (Zirikly and Diab, 2015; Gridach, 2016; Sabty et al., 2019; Alsayadi and ElKorany, 2016) are based only on the ”ENAMEX” labels. In our study, we propose to apply both ”ENAMEX” and ”TIMEX” tags. However, ”NUMEX” and time instances are very rare in our training corpus. Thus, for this reason we focused on four classes: organization (ORG), person (PER), location (LOC) and date (DATE) for the classification of Tunisian NEs.

Feature extraction. In the following part, we present the features that we used for the

[3] https://cs.nyu.edu/cs/faculty/grishman/muc6.html

machine learning-based component.

- Part Of Speech (POS) tags - In the literature, using the POS tag as a feature showed its convenience for many NER systems such as (Zirikly and Diab, 2015; Alsayadi and ElKorany, 2016). For example, (Zirikly and Diab, 2015) benefited from the POS tags generated by MADAMIRA (Pasha et al., 2014). (Alsayadi and ElKorany, 2016), also, employed Stanford POS to extract POS tags. Therefore, we used the TD parser (Mekki et al., 2020) to annotate the test set automatically. For the training set, we took advantage of the Tunisian Treebank TTB (Mekki et al., 2017).
- Lexicon-based Features - As far as we know, there is no TD lexicon proposed in the state-of-the-art for this purpose. Thus, we extracted a lexicon for the NEs from the training corpus. This lexicon is made up of four sub-lexicons, comprising 628 location (LOC), 358 organization (ORG), 809 person (PER) and 88 date (DATE). We followed this strategy for lexicon lookup. The exact match means that the word sequence matches completely an entry in the lexicon. Otherwise, we accept the partial match for words starting by the definite article (ال, Al, *the*).

4.1.3 Experimental step

Batch size. The batch size controls the accuracy of the error gradient estimation when training the neural network. The batch size can be one of the following three alternatives:

1. **The batch mode** where the batch size equals the total dataset, render iteration, and epoch equivalents.
2. **The mini-batch mode** in which the batch size is less than the total size of the dataset but greater than one.
3. **The stochastic mode** where the batch size is one.

Batch size influences the speed and stability of the learning process. For the NER, we have experimented with multiple sizes to work at the best batch size. In the first experiment, we obtained a F-measure value of 83.2%. Subsequently, the result increased by 4.43% using a batch size equal to 16. Then, the result continued its variability until reaching a new threshold equal to 88.35% (with the batch size 64). However, although we have continued to increase the batch size to find a better result, the value of F-measure continues to decline. The table 2 details the recall, precision and F-measure values of the final model of this experiment.

Table 2: Results of NER model using batch size of 64.

Class	R	P	F1
O	99%	99,28%	99,14%
ORG	88,75%	68,93%	77,59%
PER	85,56%	92,77%	89,02%
PLC	84,28%	91,34%	87,67%
Macro AVG	**89,4%**	**88,08%**	**88.35%**

Number of epochs. In terms of artificial neural networks, an epoch refers to a cycle through the entire learning corpus. In other words, if we feed a neural network with training data for more than one epoch in different models, we hope for a better prediction when given new input (test data). In our case, we are passing the batch training data of 64 instances. To complete an epoch, all 122,275 instances of the learning corpus must be passed back and forth through the neural network. In order to generalize the model, we use more than one epoch to adjust the lattice weights. However, too many eras could lead to over fitting. For that, we tested several numbers of epochs between 10 and 300. The best result was by applying 100 epochs (number of epochs by default), from where, the result was not modified.

4.2 Rule based component

In section 4.1, we have generated a machine-learning model for classifying ENAMEX classes. However, NEs classified as DATE are limited as shown in Table 1, which makes them hard to identify using a machine learning method. By adding the DATE class to the previous method, we get an evaluation result below the MACRO average with more than 60%. Hence, we notice that adding the class «DATE» to the «ENAMEX» classes decreased the overall result with 11.55%. TD is an under-resourced language in terms of the availability of anno-

Table 3: Examples of proposed rules

Nº	Rule

1	**if** word **IS IN** sub_lexicons **:**
	word ← DATE

2	**if** word **IS** DATE **AND**
	context **IS** valid_month() **:**
	context ← DATE

3	**if** word **IS** DATE **AND**
	context **IS** valid_year() **:**
	context ← DATE

4	**if** word **IS** DATE **AND**
	context **IS** valid_day() **:**
	context ← DATE

5	**if** word **IS** digit **AND** word **IS** valid_digit_day() **OR** word **IS** valid_digit_month() **OR** word **IS** valid_digit_year() **:** word ← DATE

tated textual resources and tools. Thus, we have decided to incorporate a rule-based step annotate these instances more effectively. Using handcrafted rules and lexicons created by studying the train set, the system uses the information extracted during the previous step to perform NEs DATE for specific target entities.

The first step to elaborate this method was to define 11 rules to extract the required entities. These rules are based on the sub-lexicon created for this category. For example, based on rule number «1» in Table 3, all instances found in the sub-lexicon as months (e.g. (فيفري, fyfry, *February*) and days (e.g. (خميس, xmys, *Thursday*) are classified as DATE instances. Thereafter, we have identified the contextual instances for each selected DATE instance. For example, numerical instances that may be a valid day or a valid year are labelled as DATE NEs (see rules 2 to 5 in Table 3). The same procedure is applied for the written form of digits which can take until +/- eight instances (e.g. الف و تسعة مية و تسعة و تسعين, Alf w tsEp myp w tsEp w tsEyn, *one thousand nine hundred ninety-nine*).

The results presented in Table 4 show an improvement in the overall F-measure score of 2.2%. Additionally, we have noticed an improvement of 1.35% in the result of the PER NE where six instances were incorrectly predicted as PER with the CRF model and then annotated as DATE using the linguistic method. For example, the word (رمضان, rmDAn, *Ramadan*) was incorrectly predicted by the CRF model to a PER NE due to its use as a proper name. However, this word was corrected by applying the rule-based method.

Table 4: Final evaluation results with five classes using the development set.

Class	R	P	F1
O	99.77	98.17	98.97
ORG	82.51	94.02	87.89
PER	89.92	95.94	92.84
PLC	76.57	98.39	86.12
DATE	89.8	92.63	91.19
Macro AVG	**87.71**	**95.83**	**91.4**

5 Evaluation results

The final evaluation results for the proposed hybrid method are presented in details in Table 5. Table 1 describes the test set used in this section. The performance of TUNER system is encouraging with a macro AVG F-measure equal to 94%. Indeed, the NER of all classes (ORG, PER, PLC and DATE) exceeds 91%. The best performance given to PER class (F-measure= 94.43%) However, PLC class gave the lowest F-measure score compared to all other classes. Therefore, we studied the failure cases to recognize its causes. For example, we the name of the Tunisian village (سيدي بو سعيد, sydy bw sEyd) was incorrectly annotated by the model where the first two words (سيدي بو, sydy bw) were classified as O, meaning that they are not classified as a NEs. In addition, t . The last term (سعيد, sEyd) was annotated as a PER. This is due to the absence of this village name and other similar examples such as (سيدي بو زيد, sydy bw zyd) in the train corpus. Likewise, the word (سعيد, sEyd) only exists as a PER in the train set, which favors the prediction given by the model.

We compared our method to the only TD NER proposed method in the related works. Table 5 presents the evaluation results.Indeed, we find that the use of a deep learning approach is more appealing than (Torjmen and Haddar, 2021)'s linguistic method. The overall analysis

Table 5: Comparison of the performance of the TUNER system and (Torjmen and Haddar, 2021)'system.

Class	**(Torjmen and Haddar, 2021)**			**TUNER**		
	R	**P**	**F1**	**R**	**P**	**F1**
O	99.06	89.55	94.07	99.71	99.08	99.39
ORG	0.03	45.71	0.06	86.67	89.66	88.14
PER	22.77	69.22	34.27	82.9	96.21	89.06
PLC	35.21	94.61	51.32	84.45	95	89.41
DATE	54.08	88.33	67.09	93.89	92.78	93.33
Macro AVG	42.23	77.48	54.67	**89.52**	**94.55**	**91.97**

result is much better with a difference of 37.3%. However, (Torjmen and Haddar, 2021)'s system correctly annotated 16 ORG NEs out of 629 ORG NEs in the test set. For PER NEs, they annotated 479 against 2104. These two classes gave the lowest result, especially for the recall values. For the DATE NEs class, we both used a linguistic method. The rule-based component proposed in this article has outperformed (Torjmen and Haddar, 2021). For example, their method failed to detect month and day names (such as رمضان, rmDAn, *ramadan*; الخميس, Alxmys, *Thursday*) as well as numeric dates (such as 2010, 2013). Moreover, we used the Wilcoxon matched-pairs signed-ranks test proposed by (Demšar, 2006). In this test, if the computed p-value ¡ 0.05 means that the suggested system has achieved a statistically significant improvement. The comparison returns a p-value of 0.00512, indicating that there are notable differences between the two systems.

All the other work described in section 2 is devoted to different Arabic languages (MSA or other Arabic dialects). Thus, they are using different corpora for learning and evaluating their systems. Therefore, by comparing our system to the results found in the state-of-the-art dialectal Arabic work, we notice a higher score by 22% compared to (Zirikly and Diab, 2015), which applied almost the same machine learning method using CRF toolkit (Lafferty and Mccallum, 2001). Also, we found a 9% higher result than the NER system of (Gridach, 2016) based on a combination of LSTM and CRF architecture. However, our train set represents more than the double of the corpora used by (Zirikly and Diab, 2015; Gridach, 2016), which is a very important factor, particularly in the machine learning-based methods.

6 Conclusion

To extract named entities from several sources of unstructured textual data, we propose a hybrid named-entity recognition method for Tunisian Dialect information extraction, called TUNER. It is based on Bi-LSTM-CRF model and a rule-based method using a diverse corpus that covers all dialect types (intellectualized dialect, spontaneous dialect and social media dialect). To the best of the researchers' knowledge, it is the best NER system proposed for studying the TD. Indeed, its evaluation showed that TUNER achieved very promising results with 92% of F-measure.

In future work, we have planned to explore the possibility of enhancing the system by adding more classes such as telephone number, price, percent, etc. We will also consider the use of other deep learning techniques.

References

Sherief Abdallah, Khaled Shaalan, and Muhammad Shoaib. 2012. Integrating rule-based system with classification for arabic named entity recognition. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 311–322. Springer.

Ahmed Abdelali, Kareem Darwish, Nadir Durrani, and Hamdy Mubarak. 2016. Farasa : A Fast and Furious Segmenter for Arabic. In *North American Chapter of the Association for Computational Linguistics*.

Ahmed Abdul-Hamid and Kareem Darwish. 2010. Simplified feature set for arabic named entity recognition. In *Proceedings of the 2010 Named Entities Workshop*, pages 110–115.

Mohammed Aboaoga and Mohd Juzaiddin Ab Aziz. 2013. Arabic person names recognition by using a rule based approach. *Journal of Computer Science*, 9(7):922.

Mohammed Nadher Abdo Ali, Guanzheng Tan, and Aamir Hussain. 2019. Boosting arabic named-entity recognition with multi-attention layer. *IEEE Access*, 7:46575–46582.

Fahd Alotaibi and Mark Lee. 2014. A hybrid approach to features representation for fine-grained arabic named entity recognition. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 984–995.

Norah Alsaaran and Maha Alrabiah. 2021. Arabic named entity recognition: A bert-bgru approach. *CMC-COMPUTERS MATERIALS & CONTINUA*, 68(1):471–485.

Hamzah A Alsayadi and Abeer M ElKorany. 2016. Integrating semantic features for enhancing arabic named entity recognition. *Int. J. Adv. Comput. Sci. Appl.(IJACSA)*, 7(3):2016.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104*.

Yassine Benajiba, Paolo Rosso, and José Miguel Benedíruiz. 2007. Anersys: An arabic named entity recognition system based on maximum entropy. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 143–153. Springer.

Rahma Boujelbane, Mariem Ellouze, Frédéric Béchet, and Lamia Belguith. 2014. De l'arabe standard vers l'arabe dialectal : projection de corpus et ressources linguistiques en vue du traitement automatique de l'oral dans les médias tunisiens. *TAL. 2. Traitement automatique du langage parlé*, 55:73–96.

Kareem Darwish. 2013. Named entity recognition using cross-lingual resources: Arabic as an example. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1558–1567.

Janez Demšar. 2006. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research*, 7:1–30.

Ismail El Bazi and Nabil Laachfoubi. 2019. Arabic named entity recognition using deep learning approach. *International Journal of Electrical & Computer Engineering (2088-8708)*, 9(3).

Mahmoud El-Haj and Rim Koulali. 2013. Kalimat a multipurpose arabic corpus. In *Second workshop on Arabic corpus linguistics (WACL-2)*, pages 22–25.

Salsabil El Klibi, Salwa El Hamzaoui, Hana Ben Abda, Chawki Kaddes, Farhat El Horcheni, and Anouar Maalla. 2014. *La constitution en dialecte tunisien*. Association tunisienne de droit constitutionnel, Tunisie.

Ali Elsebai, Farid Meziane, Fatma Zohra Belkredim, et al. 2009. A rule based persons names arabic extraction system. *Communications of the IBIMA*, 11(6):53–59.

Hatem M Elsherif, Khaled Mohammad Alomari, Ahmad Qasim Mohammad AlHamad, and Khaled Shaalan. 2019. Arabic rule-based named entity recognition system using gate. In *MLDM (1)*, pages 1–15.

Mourad Gridach. 2016. Character-aware neural networks for arabic named entity recognition for social media. In *Proceedings of the 6th workshop on South and Southeast Asian natural language processing (WSSANLP2016)*, pages 23–32.

Nizar Habash, Mona T. Diab, and Owen Rambow. 2012. Conventional orthography for dialectal arabic. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012), Istanbul, Turkey, May 23-25, 2012*, pages 711–718. European Language Resources Association (ELRA).

Emna Hkiri, Souheyl Mallat, and Mounir Zrigui. 2017. Integrating bilingual named entities lexicon with conditional random fields model for arabic named entities recognition. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 609–614. IEEE.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Muhammad Khalifa and Khaled Shaalan. 2019. Character convolutions for arabic named entity recognition with long short-term memory networks. *Computer Speech & Language*, 58:335–346.

Rim Koulali and Abdelouafi Meziane. 2012. A contribution to arabic named entity recognition. In *2012 Tenth International Conference on ICT and Knowledge Engineering*, pages 46–52. IEEE.

John Lafferty and Andrew Mccallum. 2001. Conditional Random Fields : Probabilistic Models for Segmenting and Labeling Sequence Data Conditional Random Fields : Probabilistic Models for Segmenting and. In *Proceedings of the eighteenth International Conference on Machine Learning, ICML*, volume 1, pages 282–289.

Liyuan Liu, Jingbo Shang, and Jiawei Han. 2019. Arabic named entity recognition: What works and what's next. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 60–67.

Asma Mekki, Inès Zribi, Mariem Ellouze, and Lamia Hadrich Belguith. 2020. Treebank creation and parser generation for Tunisian Social Media text. In *17th ACS/IEEE International Conference on Computer Systems and Applications AICCSA 2020*, Antalya, Turkey. IEEE.

Asma Mekki, Inès Zribi, Mariem Ellouze Khmekhem, and Lamia Hadrich Belguith. 2018. Critical description of TA linguistic resources. In *The 4th International Conference on Arabic Computational Linguistics (ACLing 2018) & Procedia Computer Science, November 17-19 2018*, Dubai, United Arab Emirates.

Asma Mekki, Inès Zribi, Mariem Ellouze Khemakhem, and Lamia Hadrich Belguith. 2017. Syntactic analysis of the tunisian arabic. In *International Workshop on Language Processing and Knowledge Management*.

Mohamed A Meselhi, Hitham M Abo Bakr, Ibrahim Ziedan, and Khaled Shaalan. 2014. Hybrid named entity recognition-application to arabic language. In *2014 9th International Conference on Computer Engineering & Systems (ICCES)*, pages 80–85. IEEE.

Fatma Ben Mesmia, Kais Haddar, Nathalie Friburger, and Denis Maurel. 2018. Casaner: Arabic named entity recognition tool. In *Intelligent Natural Language Processing: Trends and Applications*, pages 173–198. Springer.

Behrang Mohit, Nathan Schneider, Rishav Bhowmick, Kemal Oflazer, and Noah A Smith. 2012. Recall-oriented learning of named entities in arabic wikipedia. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 162–173.

Ossama Obeid, Nasser Zalmout, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadhl Eryani, Alexander Erdmann, and Nizar Habash. 2020. Camel tools: An open source python toolkit for arabic natural language processing. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 7022–7032.

Mai Oudah and Khaled Shaalan. 2012. A pipeline arabic named entity recognition using a hybrid approach. In *Proceedings of COLING 2012*, pages 2159–2176.

Mai Oudah and Khaled Shaalan. 2017. Nera 2.0: Improving coverage and performance of rule-based named entity recognition for arabic. *Natural Language Engineering*, 23(3):441–472.

Arfath Pasha, Mohamed Al-badrashiny, Mona Diab, Ahmed El Kholy, Ramy Eskander, Nizar Habash, Manoj Pooleery, Owen Rambow, and Ryan M Roth. 2014. MADAMIRA : A Fast , Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1094–1101.

J Ross Quinlan. 2014. *C4. 5: programs for machine learning*. Elsevier.

Caroline Sabty, Mohamed Elmahdy, and Slim Abdennadher. 2019. Named entity recognition on arabic-english code-mixed data. In *2019 IEEE 13th International Conference on Semantic Computing (ICSC)*, pages 93–97. IEEE.

Roua Torjmen and Kais Haddar. 2018. Morphological Aanalyzer for the Tunisian Dialect. In *International Conference on Text, Speech, and Dialogue (TSD 2018)*, pages 180–187. Springer, Cham.

Roua Torjmen and Kais Haddar. 2021. The automatic recognition and translation of tunisian dialect named entities into modern standard arabic. In *Formalising Natural Languages: Applications to Natural Language Processing and Digital Humanities: 14th International Conference, NooJ 2020, Zagreb, Croatia, June 5–7, 2020, Revised Selected Papers 14*, pages 206–217. Springer International Publishing.

Jihene Younes, Hadhemi Achour, and Emna Souissi. 2015. Constructing Linguistic Resources for the Tunisian Dialect Using Textual User-Generated Contents on the Social Web. In *Current Trends in Web Engineering - 15th International Conference, ICWE 2015 Rotterdam, The Netherlands*, pages 3–14.

Wajdi Zaghouani. 2012. Renar: A rule-based arabic named entity recognition system. *ACM Transactions on Asian Language Information Processing (TALIP)*, 11(1):1–13.

Ayah Zirikly and Mona Diab. 2014. Named entity recognition system for dialectal arabic. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 78–86.

Ayah Zirikly and Mona Diab. 2015. Named entity recognition for arabic social media. In *Proceedings of the 1st workshop on vector space modeling for natural language processing*, pages 176–185.

Inès Zribi, Rahma Boujelbane, Abir Masmoudi, Mariem Ellouze, Lamia Hadrich Belguith, and Nizar Habash. 2014. A conventional orthography for tunisian arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland, May 26-31, 2014.*, pages 2355–2361. European Language Resources Association (ELRA).

Inès Zribi, Mariem Ellouze, Lamia Hadrich Belguith, and Philippe Blache. 2015. Spoken Tunisian Arabic Corpus STAC: Transcription and Annotation. *Research in computing science*, 90.

Inès Zribi, Inès Kammoun, Mariem Ellouze, Lamia Hadrich Belguith, and Philippe Blache. 2016. Sentence boundary detection for transcribed tunisian arabic. In *12th Edition of the Konvens Conference*, Bochum, Germany.