International Journal on Advanced Science Engineering Information Technology

Vol.10 (2020) No. 3
ISSN: 2088-5334

An Evaluation Methodology of Named Entities Recognition in Spanish Language: ECU 911 Case Study

Marcos Orellana[a,1], Andrea Trujillo[a,2], Juan-Fernando Lima[a,3], María-Inés Acosta[a,4], Mario Peña[a,5]

[a] *Laboratorio de Investigación y Desarrollo en Informática, Universidad del Azuay,Cuenca, Ecuador*
E-mail: [1]marore@uazuay.edu.ec; [2]atrujillo@uazuay.edu.ec; [3]flima@uazuay.edu.ec; [4]acosta@uazuay.edu.ec; [5]mario_pena@uazuay.edu.ec

***Abstract*— The importance of the gathered information in Integrated Security Services as ECU911 in Ecuador is evidenced in terms of its quality and availability in order to perform decision-making tasks. It is a priority to avoid the loss of relevant information such as event address, places references, names, etc. In this context it is present Named Entity Recognition (NER) analysis for discovering information into informal texts. Unlike structured corpus and labeled for NER analysis like CONLL2002 or ANCORA, informal texts generated from emergency call dialogues have a very wide linguistic variety; in addition, there is a strong tending to lose important information in their processing. A relevant aspect to considerate is the identification of texts that denotes entities such as the physical address where emergency events occurred. This study aims to extract the locations in which an emergency event has been issued. A set of experiments was performed with NER models based on Convolutional Neural Network (CNN). The performance of models was evaluated according to parameters such as training dataset size, dropout rate, location dictionary, and denoting location. An experimentation methodology was proposed, with it follows the next steps: i) Data preprocessing, ii) Dataset labeling, iii) Model structuring, and iv) Model evaluating. Results revealed that the performance of a model improves when having more training data, an adequate dropout rate to control overfitting problems, and a combination of a dictionary of locations and replacing words denoting entities.**

***Keywords*— named entity recognition; Spanish language; emergency calls; informal text.**

I. INTRODUCTION

The Ecuadorian government enacted the Decree N° 889 for the creation of the Integrated Security Service ECU911. It is an institution that integrates different emergency services, becoming the central axis for registering emergency information through telephone calls, video surveillance cameras, and a resource dispatch manager [1]. The call registry process is performed by dispatchers, who have the responsibility of: i) typing relevant information from emergencies, ii) identifying its priority, and iii) assigning the necessary resources through a computer platform. Moreover, an audio file of the conversation between the caller and the dispatcher is recorded, whose language is unstructured and informal. The language provides insights about typical emotions of emergency situations, and sometimes with unconcluded ideas or sentences. The steps of a traditional process were compiled from on-site visits that were carried out at the ECU911 emergency services in Cuenca-Ecuador.

In this context, the importance of the gathered information is evidenced in terms of its quality and availability in order to perform decision-making tasks. It is a priority to avoid the loss of relevant information. There are aspects that must be considerate such as: i) the perception of being in a different geographical location; ii) the use of colloquial language, iii) a possible faint of the person who issues the alert, iv) the dispatcher's bias during the transcription of messages.

This situation has been analyzed applying Named Entities Recognition (NER) models based on convolutional neural networks and Long Short-Term Memory (LSTM) neural networks. A Convolutional Neural Network (CNN) is a model that considers each dialogue as a sequence of characters and features, it converts them into a vector of continuous words in order to predict and analyze them [2]–[4]. This vector contains dimensions of a word such as the semantic or grammatical aspect of it, including the name of geographical places (location) and the identity of the people who intervene on the dialogue. LSTM neural networks manage the gradient descent problems computing a weighted sum of the input signal and applying a nonlinear activation function, taking information only from the past, but not knowing anything about the future [5].

NER models are considered an important research field in the Natural Language Processing (NLP) [6]. It is mainly focused on setting labels into unstructured data, and applying a process to extract information, combining not only geographic and linguistic analysis, but also including

algorithms to automatically classify entities into classes or types that have been defined previously [6]–[8]. Some of the areas where NER has been implemented include speech recognition, machine translation, information extraction, and question answering [4], [8], [9].

A NER model focuses on the identification of entities into the training dataset, their detection in new unstructured information and their classification [10]. The first step is the automatic transcription of audio files, converting them into unstructured texts while emergencies are occurring. In this context, it is necessary to incorporate a specific corpus related to emergency services which are specific for each country [11], or to apply lemmatization or tagging process [7].These techniques are used to identify continuous words that denotates location, and people in a specific context. The variation on the size of the dataset and the definition of the dropout rate have also been analyzed in order to obtain measures.

Also, it is necessary the use of a corpus in Spanish language, designed especially using Ecuadorian dialect, because many of the words acquire different meanings depending on the context of each geographical region and with many possible connotations [12]. This set of dialogues is considered the training dataset to be tested in a NER model using a SpaCy library.

SpaCy includes a statistical model for the tagging, parsing, and entity recognition in CNN [7], [13], [14]. It is a free library that can be configurated depending on specific requirements of each domain including features to process the Spanish language. Moreover, it is considered as one of best well-established open-source NER tools because of its high performance [15].

Informal and unstructured texts present challenges for NER analysis due to their dynamics and variety of dialects. It is considered a noisy domain as it involves spelling errors, diverse vocabulary, unstructured and incomplete grammar, slang and colloquial nature [16]. A considerable amount of work has been done for NER in formal texts. In recent years, machine learning-based approaches have captured the attention of researchers and within this, artificial neural networks are techniques that offer superior performance over feature-engineered models [2]. An effective combination incorporates a bidirectional LSTM [17], [18] and a convolutional neural network (CNN) [8] with labeling words [19]–[21]. Nowadays, the word and character-level deep learning hybrid models have presented the highest performance in NER tasks, and include key features of rule-based approaches that improve the overall models performance [2]. Nevertheless, when moving from the formal text to informal text domain, the performance in NER tasks decreases considerably (55% on average) [22].

Different well-known free tools for NER tasks have been developed in the last decade. In order to prove the performance of such tools, the contribution developed in [23] evaluates Stanford NER, SpaCy, Alias-i LingPipe and Natural Language Toolkit (NLTK), and then generate a hybrid tool for NER using Stanford NER and SpaCy which were the best tools. The discriminative capacity of SpaCy and Stanford NER applied computer-automated verbal deception detection was compared in [15]. In [24] Stanford NER, SpaCy, OpenNLP and NLTK were evaluated using HAREM corpus. The research included a hyper-parameter study (e.g., cutoff, iterations, tolerance, entropy_cutoff) to improve the performance reached with the default configurations. Therefore, each case study should determine the hyperparameters that yield the best performance. SpaCy is a tool based on neural network architectures, while Stanford NER uses the rule-based approach.

To evaluate the performance of the NER models, there are many corpus in various areas, for example, multilingual corpus (CoNLL2002 – 2003 , ANCORA [6], OntoNotes 5.0 [25], SemEval 2013 [18], Wikipedia gold standard corpus [23], Twitter NER Shared task dataset [16]). However, it has not yet dealt with dialogs in the domain of emergency calls.

The most used evaluation measures of NER performance are *precision* (i.e., the correctly classified entities divided by the total named entities detected., *recall* (i.e., the relevant correctly classified entities divided by the total, named entities detected) and F1-measure is the is the harmonic average of *precision* and *recall* [26].

In general, models based on neural network architectures require a considerable number of samples for an effective fit. With scarce sources for training, the neural network just memorizes the samples and the generalization start to get worse. In a simple way, dropout rate can be considered as a neural networks regularization method to prevent overfitting. For NER tasks, some studies are presented with neural networks architecture using a default probability around 0.5 for the dropout rate [8], [17]. Nevertheless, using high values generates negative impacts on performance, while low values leads to more training time [17]. Likewise, the change in the dropout rate parameter does not achieve significant performance improvement when there is a small number of samples. In [27] the authors determined the impact of the training dataset size produced on the performance of several neural network architectures for answer selection tasks. The study finds out that the performance is not determinant for improving when the training dataset size increases. Conversely, more discriminative information can be extracted from informal texts when more data is used for training [28]. Hence, there is a trade-off between the dropout rate and the number of samples for training that lead to getting better performance of the model [29]. At the majority of studies, it is used a common partition of 70% for training and 30% for testing [24].

On the other hand, rule-based approaches using lexical syntax patterns, as well as lists of information or dictionaries to identify and classify named entities and can be used on the top of machine learning approaches in order to improve the overall performance [19], [23]. The use of specific knowledge of the language or other resources can contribute to improve the performance of NER systems [17], [20], such as dictionaries [30] and external tagged data such as gazetteers [28].

Although much work has already been developed in the NER tasks domain, there is no work that addresses both informal and unstructured text using a corpus for the Spanish language. In addition, it is important to determine the optimal values of dropout rate parameter and training data size.

This study aims to reach the following objectives: i) to extract relevant information as locations and their references

in the emergency call domain; ii) to analyze the best configuration for working with informal texts in Spanish. These tasks are based on parameters (i.e., training data size, dropout rate, locations dictionary, replacing of phrases denoting locations). According to Blumer et al., [31], it is important to maintain a simple model for real-life applications to improve the performance.

The results of the study look forward to support ECU911 dispatchers, who need tools to help in the treatment of the informal texts, which shows the location and frequency tags automatically; and thus, finding a better allocation of resources and efficiency in the management of emergencies

The structure of this paper is as follows: Section 1 introduces and presents the motivation for this study. Section 2 presents related studies and the differences with this proposal. Section 3 provides the methodology and proposes NER models to train dialogues in Spanish language. Finally, the results and evaluation metrics are presented in Section 4.

II. Materials and Method

This study presents a methodology represented with the Software & Systems Process Engineering Meta-Model 2.0 (SPEM 2.0) [32] in order to find named entities by using Named Entities Recognition (NER) techniques. The proposed methodology is divided into the following four main activities each of which has input and output artifacts and guidelines as it is shown in Fig. 1: i) Data Preprocessing, ii) Dataset Labeling, iii) Model Structuring, and iv) Model Evaluating.

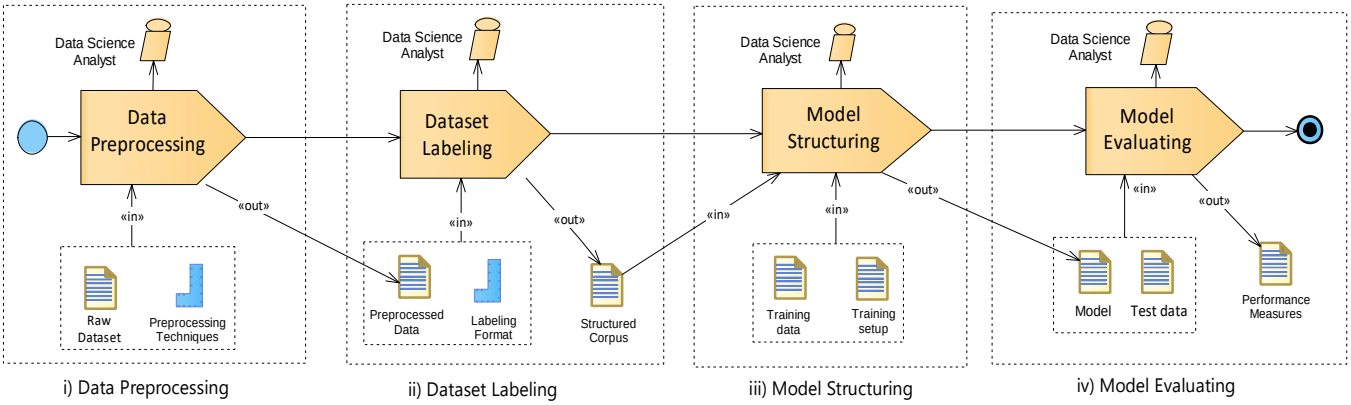

Fig. 1 Experimentation Methodology

The activity of Data Preprocessing describes the raw dataset, the processing tasks and techniques applied to obtain a clean text for the experimentation. The Dataset Labeling activity performs the labeling format, which is done manually in order to generate the training and testing corpus. The Model Structuring activity describes the general architecture and parameters considered to work with informal texts in Spanish. The Model Evaluating activity explains the performance measures and configurations for evaluating the generated models. The next subsections describe each activity, considering that each of them uses the output artifact obtained from its previous activity.

A. Data Preprocessing

The inputs of this activity are: i) the raw dataset to be manually labeled with the named entities, and ii) the preprocessing techniques that are represented as guidelines by SPEM. The raw dataset will be used as input for training and testing the model, and preprocessing techniques allow removing unnecessary text and to reduce its dimensionality.

1) Raw Dataset: The corpus to be analyzed by NER can be formal (e.g., news, medicals, historical) or informal (e.g., tweets, dialogues, websites). The corpus in Spanish is addressed by datasets such as ConLL2002 [8], [17] and Ancora [33], [34], and in English by medical corpus such as BC2GM, JNLPBA, BC4CHEMD [35], [36]. These corpuses have multiple entities (e.g., localization, person, organization, date, anatomy, chemical, diseases).

A dataset into the *Emergency calls* domain was used in this study. The Integrated Ecuadorian Security Service ECU 911 provided a dataset collected in December 2015, where a sample of 400 transcript texts (calls) were used. The main entities found in this dataset were people and locations. The entities of type PERSON represent the caller, and the entities of type LOCATION represent the address of an emergency event.

The person's name is important, but due to its confidential nature, that entity should be protected; and therefore, it cannot be extracted. However, the main objective is the extraction of location entities. Getting this type of entities help to improve the geographic location of calls, with which the time of an emergency attention is substantially reduced [37]. Finally, the dataset shows that the arithmetic mean is equal to 119 words for each text, the minimum is 31 and the maximum is 545 words.

1) Preprocessing Techniques: Working with informal texts during their transcription can lead to noise problems due to writing errors or colloquialisms proper of the natural language. Errors generally reduce the data quality. Therefore, it is important to remove those errors for decreasing the dimensionality of texts and to get cleaner texts that contain only the necessary information. Several studies propose the replacement of digits with specific symbols [8], [17], tokenizer [38] and post tagging [8], [10]. Thus, in this study, three preprocessing techniques are integrated to be used:

- Lowercase: This technique is considered to standardize texts, originating from transcription software.
- Stop-words removal: This technique is focused on reducing the dimensionality of texts. However, there are stop words that connect entities (e.g., “a”, “la”, “en”, “el”, “y”) that should be considered. In this contribution, a dictionary was built according to the needs of the text.
- Reduce vocabulary: Firstly, the tokenization technique was applied for each text; secondly, those tokens were grouped into three grams to obtain the most frequent sequences of consecutive words used. It has been stated a threshold with a frequency of five apparitions, where 85 sequences of words were obtained. In particular, for a given text (e.g. “estoy aquí, en la avenida 24 de mayo…”), three grams were replaced with a denotation of a location with special label denominated LOC* (e.g. “estoy LOC* avenida 24 de mayo …”) to eliminate several colloquialisms that show the presence of an entity in the text. The LOC* label was selected because it denotes insights related to the presence of a location entity. It is important to emphasize that in this preprocessing task, only entities of type LOCATION were considered due to its large number of coincidences. Also, this technique was only applied with the model evaluation purpose.

B. Dataset Labeling

A NER model requires a corpus for training and testing into the dataset. Additionally, the corpus must be previously labeled. This activity is performed manually considering the entities to be extracted. This task can be improved using annotation (e.g., Doccano [39], Prodigy [40], and NeuroNer [41]) because they facilitate the manual labeling of texts. The sequence text labeling was made using Doccano tool that allows downloading data in several formats. LOC label was selected to represent a place reference (e.g., street names, avenues, neighborhoods, cities, parishes, parks, institutions).

The process of labeling is tied to the entities to be extracted and the format of the corpus. The labeling formats most used are IOB [24], [42], [43] and BILOU [44], [45]. IOB has the nomenclature: *inside, outside,* and *beginning*. *Beginning* tag indicates the starting of an entity; *inside* tag indicates that it is part of an entity, but it is not the first token and *outside* is any other token. In the BILOU format, there are added the following labels: i) *last* that represents the last token in the entity, and ii) *unit* that represents an entity with a single token. Finally, the result of manual labeling by using the BILOU format shows a total of 2230 LOC.

C. Model Structuring

For this study, models were structured with different configurations. A statistical model was used with an overall structure based on CNN and LSTM networks of three layers. Firstly, the convolutional layer involves the feature map with a window of three tokens (trigram) [46]; secondly, max-pooling layer, decreases the examples and computational cost, through the reduction of the parameters to learn [47]. Also, LSTM is an artificial recurrent neural network, which reads the words with a bidirectional model, generating a vector representation of each word [17]; finally, the output layer adds the results of the LSTM layers and the best tag or prediction found on the tokens is achieved [47].

There are four parameters to configure the model (i.e., dropout rate, size batch, number of iterations, optimization algorithm). A default minibatch was used, and number of iterations equal to one hundred, value suggested by related studies[13], [24].

Also, it was used the optimization Stochastic Gradient Descent (SGD) algorithm and Spacy library for the implementation of the methodology. This library is an open-source written in Python and Cython programming languages.

Several models based on the four parameters were generated to measure performance, and to find the optimal configuration to be used with informal texts for Spanish language. The parameters used are detailed in the following paragraphs:

- Training dataset size: Neural network models need large datasets to get a good performance. Therefore, the amount of texts at this stage can be of influence in the final results [27].
- Dropout rate: This parameter reduces the overfitting; consequently, the model does not lose the ability to generalize the problem.
- Vocabulary reduction: The training dataset can influence the performance of the model. Standardize words that denote locations in an informal text are possible when the preprocessing technique to reduce vocabulary is applied. This technique alloes to reduce the dimensionality.
- Locations dictionary: Combining a location dictionary with deep learning can help the model to find entities overlooked during the training. Therefore, it was created a dictionary with localizations of the area, and it was added to the trained model.

D. Model Evaluating

This model has been evaluated by the application of four approaches: i) training data size, ii) dropout rate, iii) vocabulary reduction through denoting locations, and iv) locations dictionary. The general performance measure used was the *f-measure* because it maintains a harmonic average between *precision* and *recall*. Furthermore, this evaluation measure is the best for unbalanced datasets [6], [16]. This study measures the performance in the extraction of information for type LOC entity as in Table I.

TABLE I
SETUPS OF EVALUATION

Training Data Size	Dropout Rate	Locations Dictionary	Vocabulary Reduction	F-measure
60	0.1	0	0	60.53
60	0.1	0	1	60.84
60	0.1	1	0	62.97
60	0.1	1	1	62.63
60	0.2	0	0	62.01
90	0.9	1	1	40.74

During the dataset training, it is divided into several percentages (i.e., 60, 70, 80, 90%) and the testing dataset is formed by the remaining data. Each dataset percentage was tested with a dropout rate from 0.1 to 0.9. Subsequently, it was combined with a locations dictionary, a vocabulary reduction, or both. The evaluation generated 144 possible combinations. Table 1. details the setups of evaluation.

III. Results and Discussion

Section 3.4 stated that there are 144 combinations based on the evaluation parameters. Therefore, given the big amount of values; the most general results were evaluated first, which in this case is the effect of the training dataset size. Then, the effect in terms of dropout rate, location dictionaries and entity denotation were analyzed according to the dataset size with the best performance.

A. Training dataset size.

Fig. 2 shows the f-measure and the training dataset percentage. The best value of the model is which uses 90% of the dataset. The 1st and 3rd quartile are higher with 90% of training data size. Thus, the f-measure value increases according to the size of the training dataset.

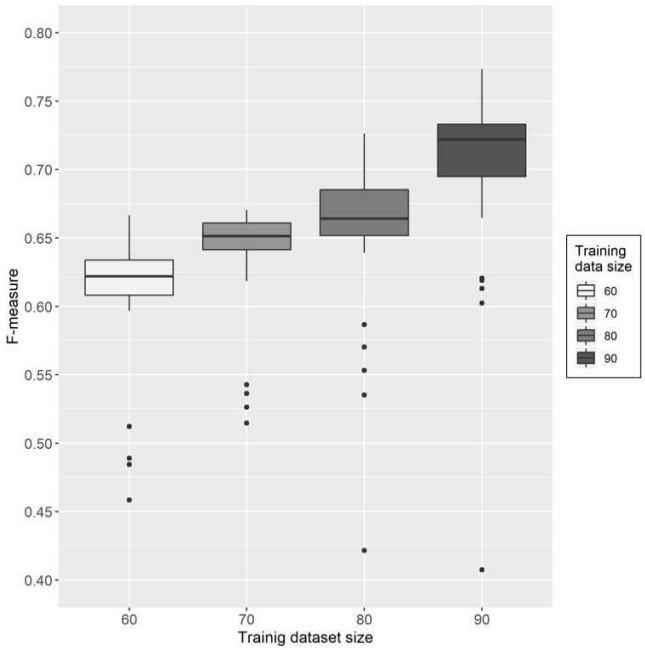

Fig. 2 Training data size vs F-measure

B. Dropout rate, entities denotation and locations dictionary.

The column with 90% of training data was selected for this study. Different types of combinations have been proposed by different techniques as show in Table 2. In each case can note, one or all techniques can be used. The number one into the matrix is equivalent to use the technique, and zero denotes that this technique has not been used. Fig. 3. shows that the f-measure reaches its highest value with a dropout rate equal to 0.3, in contrast, this value falls when dropout rate value increases. The combination *all* in several cases gives a higher performance (i.e., 0.1, 0.3, 0.6), but in other cases performance is poor (i.e., 0.4, 0.5). The cases with which it does not improve have a high dropout rate, the models tend overfitting and tend to lower their performance.

TABLE II
COMBINATIONS OF EVALUATION SETUPS

Combination tag	Locations Dictionary	Denoting Locations
nothing	0	0
dicc	0	1
denote	1	0
all	1	1

In the case of the experimentation with dropout rates equal to 0.9, f-measure lower than 0.4 were obtained. In consequence, the models with this configuration were discarded. However, for the optimal dropout rate (0.3), it is observed that there is a greater f-measure when applying any combination. Regarding the use of a *locations dictionary* or *denoting locations*, it is evident that the combination of the two processes significantly improves the f-measure. However, if the techniques are applied individually, the best performance is achieved with *denoting locations.*

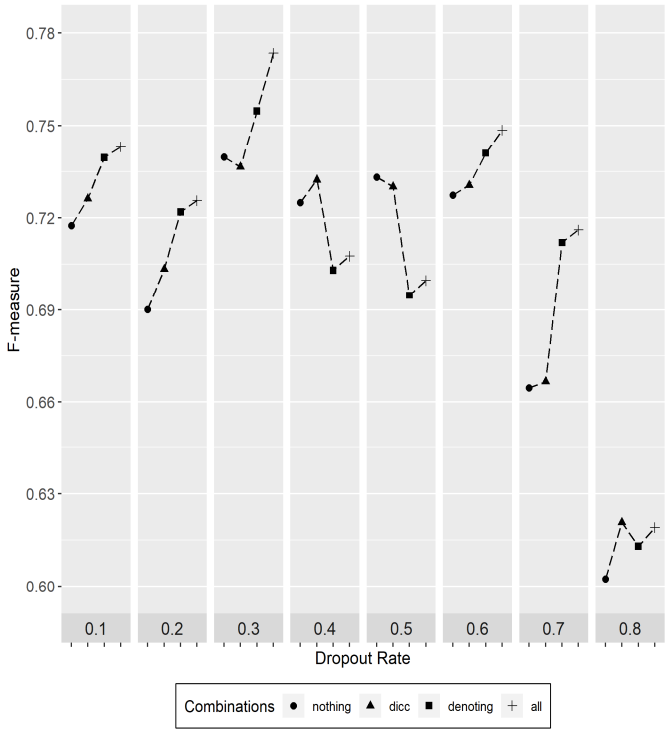

Fig. 3 Dropout rate vs denoting locations a locations dictionary

IV. Conclusions

According to the results obtained in the methodology of study, the following conclusions are described:

- The size of the training dataset definitively influences the performance of NER models applying CNN and LSTM. The greater the number of data, the greater the performance of the models.
- It is important to evaluate relevant parameters such as the dropout rate that can generate overfitting in the models. In this study case better performance is obtained with a dropout rate equal to 0.3.

- The use of a locations dictionary does not ensure better performance in extracting locations; however, the use of dictionary with adequate setting others parameter, reduces the task of processing in the text. This is because the replacing common phrases that denote entities give a better f-measure.
- Finally, combining a dictionary of locations and replacing phrases denoting entities improves the performance of the models.

For future work, it is proposed to expand a dictionary of locations and a precise corpus for the replacement of phrases that denote entities that are linked to the emergency call environment. Also, it is necessary to perform experiments with alternative datasets for evaluating the parameters rate and the application of the same techniques.

ACKNOWLEDGMENT

This research was supported by the vice-rectorate of investigations of the Universidad del Azuay. We thank our colleagues from Laboratorio de Investigación y Desarrollo en Informática (LIDI) de la Universidad del Azuay who provided insight and expertise that greatly assisted this research.

REFERENCES

[1] Ecuador, "Decreto N° 988." Quito, 2011.

[2] Y. Vikas and B. Steven, "A survey on recent advances in named entity recognition from deep learning models," in *Proceedings of the 27th International Conference on Computational Linguistics*, 2018, vol. 59, no. 1, pp. 2145–2158.

[3] X. Liu, Y. Zhou, and Z. Wang, "Recognition and extraction of named entities in online medical diagnosis data based on a deep neural network," *J. Vis. Commun. Image Represent.*, vol. 60, pp. 1–15, 2019.

[4] J. Linet and C. Zea, "Reconocimiento de entidades nombradas para el idioma espa ˜ nol utilizando Conditional Random Fields con caracter ´ ısticas no supervisadas."

[5] M. Gridach, "Character-level neural network for biomedical named entity recognition," *J. Biomed. Inform.*, vol. 70, pp. 85–91, 2017.

[6] R. Gutiérrez, A. Castillo, V. Bucheli, and O. Solarte, "Named Entity Recognition for Spanish language and applications in technology forecasting Reconocimiento de entidades nombradas para el idioma Español y su aplicación en la vigilancia tecnológica," *Rev. Antioqueña las Ciencias Comput. y la Ing. Softw.*, vol. 5, pp. 43–47, 2015.

[7] M. Won, P. Murrieta-Flores, and B. Martins, "Ensemble Named Entity Recognition (NER): Evaluating NER Tools in the Identification of Place Names in Historical Corpora," *Front. Digit. Humanit.*, vol. 5, no. March, pp. 1–12, 2018.

[8] M. Khalifa and K. Shaalan, "Character convolutions for Arabic Named Entity Recognition with Long Short-Term Memory Networks," *Comput. Speech Lang.*, vol. 58, pp. 335–346, 2019.

[9] M. H. Bokaei and M. Mahmoudi, "Improved Deep Persian Named Entity Recognition," in *9th International Symposium on Telecommunication: With Emphasis on Information and Communication Technology, IST 2018*, 2019, pp. 381–386.

[10] L. Derczynski *et al.*, "Analysis of named entity recognition and linking for tweets," *Inf. Process. Manag.*, vol. 51, no. 2, pp. 32–49, 2015.

[11] I. Moreno and T. Rom, "A Domain and Language Independent Named Entity Classification Approach Based on Profiles and Local Information A Domain and Language Independent Named Entity Classification Approach Based on Profiles and Local Information," no. September, 2017.

[12] W. G. Aguilar, D. Alulema, A. Limaico, and D. Sandoval, "Development and Verification of a Verbal Corpus Based on Natural Language for Ecuadorian Dialect," in *Proceedings - IEEE 11th International Conference on Semantic Computing, ICSC 2017*, 2017, pp. 515–519.

[13] R. Jain, D. S. Anand, and V. Janakiraman, "Scrubbing Sensitive PHI Data from Medical Records made Easy by SpaCy - A Scalable Model Implementation Comparisons," *CoRR*, vol. abs/1906.0, 2019.

[14] L. G. Moreno-sandoval, S. Carolina, K. Esp, A. Pomares-quimbaya, and J. C. Garcia, "Spanish Twitter Data Used as a Source of Information About Consumer Food Choice," in *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, 2018, vol. 2, pp. 134–146.

[15] B. Kleinberg, M. Mozes, A. Arntz, and B. Verschuere, "Using Named Entities for Computer- Automated Verbal Deception Detection," pp. 1–10, 2017.

[16] N. Limsopatham and N. Collier, "Bidirectional LSTM for named entity recognition in ttwitter messages," in *Proceedings of the 2nd Workshop on Noisy User-generated Text ({WNUT})*, 2016, pp. 145–152.

[17] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," *CoRR*, vol. CoRR, pp. 260–270, 2016.

[18] V. Yadav, R. Sharp, and S. Bethard, "Deep Affix Features Improve Neural Named Entity Recognizers," in *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, 2018, pp. 167–172.

[19] M. H. Bokaei and M. Mahmoudi, "Improved Deep Persian Named Entity Recognition," *2018 9th Int. Symp. Telecommun.*, pp. 381–386, 2018.

[20] H. Chen, Z. Lin, G. Ding, J. Lou, Y. Zhang, and B. Karlsson, "GRN: Gated Relation Network to Enhance Convolutional Neural Network for Named Entity Recognition," in *Proceedings of AAAI*, 2019.

[21] Q. Lu, Y. Xu, R. Yang, N. Li, and C. Wang, "Serial and Parallel Recurrent Convolutional Neural Networks for Biomedical Named Entity Recognition," in *International Conference on Database Systems for Advanced Applications*, 2019, pp. 439--443.

[22] G. Aguilar, A. P. López Monroy, F. González, and T. Solorio, "Modeling Noisiness to Recognize Named Entities using Multitask Neural Networks on Social Media," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2018, pp. 1401–1412.

[23] R. Jiang, A. Star, and R. E. Banchs, "Evaluating and Combining Named Entity Recognition Systems," in *Proceedings of the Sixth Named Entity Workshop*, 2016, pp. 21–27.

[24] Pires André, Devezas José Luís, Nunes Sérgio, A. Pires, J. Devezas, and S. Nunes, "Benchmarking Named Entity Recognition Tools for Portuguese," *Proc. Ninth INForum Simp. Informática*, pp. 111--121, 2017.

[25] J. P. C. Chiu and E. Nichols, "Named Entity Recognition with Bidirectional LSTM-CNNs," *Trans. Assoc. Comput. Linguist.*, vol. 4, no. 2003, pp. 357–370, 2016.

[26] A. Goyal, V. Gupta, and M. Kumar, "Recent Named Entity Recognition and Classification techniques: A systematic review," *Comput. Sci. Rev.*, vol. 29, pp. 21–39, 2018.

[27] L. Azzopardi, B. Stein, N. Fuhr, P. Mayr, C. Hauff, and D. Hiemstra, "Impact of Training Dataset Size on Neural Answer Selection Models," in *European Conference on Information Retrieval*, 2019, vol. 11437, pp. 828–835.

[28] T. M. Ma, Yukun Kim, Jung-jae Bigot, Benjamin, Khan, "Featured-enriched word embeddings for naemd entity recognition in open-domain conversations," *Icassp 2016*, pp. 6055–6059, 2016.

[29] N. Srivastava *et al.*, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, pp. 1929–1958, 2014.

[30] D. Bonadiman, A. Severyn, and A. Moschitti, "Deep Neural Networks for Named Entity Recognition in Italian," *Proc. Second Ital. Conf. Comput. Linguist. CLiC-it 2015*, pp. 51–55, 2016.

[31] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth, "Occam's razor," in *In- formation processing letters*, 1987, vol. 24, no. April, pp. 377–380.

[32] F. Ruiz and J. Verdugo, "Guía de Uso de SPEM 2 con EPF Composer," *Univ. Castilla-La Mancha Esc. Super. Informática Dep. Tecnol. y Sist. Inf. Grup. Alarcos*, vol. 3, p. 93, 2008.

[33] R. Gutiérrez, A. Castillo, V. Bucheli, and O. Solarte, "Named Entity Recognition for Spanish language and applications in technology forecasting," *Rev. Antioqueña las Ciencias Comput. y la Ing. Softw.*, vol. 5, pp. 43–47, 2015.

[34] C. A. C. Molina, R. E. Gutierrez, and O. Solarte, "Prototipo para el reconocimiento de entidades nombradas en el idioma Español," in

2015 10th Colombian Computing Conference, 10CCC 2015, 2015, pp. 364–371.
[35] G. Crichton, S. Pyysalo, B. Chiu, and A. Korhonen, "A neural network multi-task learning approach to biomedical named entity recognition," *BMC Bioinformatics*, vol. 18, no. 1, pp. 1–14, 2017.
[36] M. Gridach, "Character-level neural network for biomedical named entity recognition," *J. Biomed. Inform.*, vol. 70, no. May, pp. 85–91, 2017.
[37] J. Zhang *et al*., "Enable Automated Emergency Responses Through an Agent-Based Computer-Aided Dispatch System," in *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, 2018, pp. 1844–1846.
[38] I. Moreno, M. T. Romá-Ferri, and P. Paloma, "A domain and language independent named entity classification approach based on profiles and local information," in *RANLP*, 2017, no. September, pp. 510–518.
[39] Docaano, "Text annotation for Human," 2019. [Online]. Available: https://doccano.herokuapp.com/. [Accessed: 09-Sep-2019].
[40] Al Explosion, "Prodigy," *Named Entity Recognition*, 2019. [Online]. Available: https://prodi.gy/features/. [Accessed: 09-Sep-2019].
[41] Neuroner, "NeuroNER," 2019. [Online]. Available: http://neuroner.com/. [Accessed: 09-Sep-2019].
[42] R. Jiang, R. E. Banchs, and H. Li, "Evaluating and Combining Name Entity Recognition Systems," pp. 21–27, 2016.
[43] T. Tran and R. Kavuluru, "An end-to-end deep learning architecture for extracting protein-protein interactions affected by genetic mutations," *Database*, vol. 2018, no. 2018, pp. 1–13, 2018.
[44] X. Liu and M. Zhou, "Two-stage NER for tweets with clustering," *Inf. Process. Manag.*, vol. 49, no. 1, pp. 264–273, 2013.
[45] M. Tkachenko and A. Simanovsky, "Named entity recognition: Exploring features," in *Proceedings of KONVENS*, 2012, vol. 2012, pp. 118–127.
[46] Y. Zhu, G. Wang, and B. F. Karlsson, "CAN-NER: Convolutional Attention Network for Chinese Named Entity Recognition," in *NAACL-HLT*, 2019.
[47] Y. Zeng, H. Yang, and Y. F. B, "A convolution BiLSTM neural network model for Chinese event extraction," in *Natural Language Understanding and Intelligent Applications*, vol. 1, 2016, pp. 275–287.