Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 2953–2961
Marseille, 11–16 May 2020

SiNER: A Large Dataset for Sindhi Named Entity Recognition

Wazir Ali, Junyu Lu, Zenglin Xu
School of Computer Science and Engineering
University of Electronic Science and Technology of China, 611731
aliwazirjam, cs.junyu, zenglin.xu@gmail.com

Abstract

We introduce the SiNER: a named entity recognition (NER) dataset for low-resourced Sindhi language with quality baselines. It contains 1,338 news articles and more than 1.35 million tokens collected from Kawish and Awami Awaz Sindhi newspapers using the begin-inside-outside (BIO) tagging scheme. The proposed dataset is likely to be a significant resource for statistical Sindhi language processing. The ultimate goal of developing SiNER is to present a gold-standard dataset for Sindhi NER along with quality baselines. We implement several baseline approaches of conditional random field (CRF) and recent popular state-of-the-art bi-directional long-short term memory (Bi-LSTM) models. The promising F1-score of 89.16% outputted by the Bi-LSTM-CRF model with character-level representations demonstrates the quality of our proposed SiNER dataset.

Keywords: Language Resources, SiNER, Sindhi Language, Named Entity Recognition

1. Introduction

Named entity recognition is an essential lower-level task (Ma and Hovy, 2016) in natural language processing (NLP), used to extract and categorize naming entities into a predefined set of classes such as person, location, organization (Sang and De Meulder, 2003), numeral and temporal entities (dos Santos et al., 2015). It is essential to have a high-quality NER system for downstream NLP tasks such as information extraction (Grishman and Sundheim, 1996; Neudecker, 2016), question answering (Moldovan, 2002) and machine translation (Babych and Hartley, 2003). The NER task traditionally requires a large amount of knowledge in the form of lexicons and feature engineering to achieve high performance (Chiu and Nichols, 2016).

The remarkable development has been made in the NER task since the message understanding conference (Grishman and Sundheim, 1996). Later, (Sang and Erik, 2002; Sang and De Meulder, 2003) introduced quality datasets for European languages along with exact match evaluation matrices. Numerous methods have been employed for the NER task, which can be broadly categorized into the rule-based, and language-independent statistical approaches. Recently, state-of-the-art language-independent deep learning models proposed by (Chiu and Nichols, 2016; Ma and Hovy, 2016; Lample et al., 2016; Tran et al., 2017; Kuru et al., 2016) have been successfully and extensively opted to address NER related problems with unsupervised word embeddings. But these language-independent neural models can be exploited on the substantial amount of training and evaluation datasets.

The language resources (LRs) play an essential role in the digital survival of natural languages because of the ever-increasing usage of web-based technologies in daily life. Most of the European and East Asian languages are rich in such LRs, but most of the South and Southeast Asian Languages (SSALs) including Sindhi are still under-resourced (Ekbal et al., 2008; Singh, 2008). Sindhi has the status of an official language in the Sindh province of Pakistan and one of the national languages in India (Motlani, 2016) with the total number of 75 million speakers. The Sindhi NER task was initially coined by (Ali et al., 2015) and (Nawaz et al., 2017) by discussing the challenges and future research opportunities. Later, Hakro et al., (2017) and Jumani et al. (2018) proposed Sindhi NER systems using a rule-based approach on a small of the corpus.

In this paper, we introduce novel gold-standard SiNER dataset using a large number of news articles obtained from most circulated Kawish and Awami-Awaz Sindhi newspapers (Ali et al., 2019) with eleven entity classes. The annotation task is performed by three native Sindhi speakers using the Doccano (Nakayama et al., 2018) a web-based text annotation tool. After the annotation, preprocessing and manual evaluation of proposed SiNER, we employ language-independent approaches of CRF (Sutton et al., 2012) for initial baseline, and Bi-LSTM, Bi-LSTM-CRF (Huang et al., 2015; Lample et al., 2016), Bi-LSTM-CRF with character-level representations (Kuru et al., 2016; Tran et al., 2017; Misawa et al., 2017) for quality baselines. To the best of our knowledge, we are the first to develop and evaluate the SiNER dataset for Sindhi language along with quality baselines. The synopsis of our novel contributions is given as follows:

- We reveal a novel SiNER dataset for low-resourced Sindhi language.
- We present quality baselines for SiNER by employing CRF and state-of-the-art language-independent Bi-LSTM and Bi-LSTM-CRF approaches.
- The performance comparison of Bi-LSTM models with GloVe and fastText word embeddings on SiNER dataset.

The remaining sections of the paper are organized in the following sequence: Section 2. presents a brief overview of Sindhi language for linguistic awareness. The related work regarding NER datasets and state-of-the-art neural algorithms is given in Section 3. Whereas, Section 4. consists of the employed methodology for the development and evaluation of SiNER dataset. Moreover, Section 5. comprised of experiments and results, and lastly, Section 6. covers the discussion and conclusion, respectively.

2. Sindhi language

Historically, Sindhi belongs to the Indo-Aryan language family passed through many literary evolutions. It has some unique linguistic characteristics such as rich morphological structure, multiple writings systems, and dialects with the historical linguistic and cultural background (Motlani, 2016; Jamro, 2017). Presently the Sindhi language is an official language in the Sindh province of Pakistan, also being taught as a compulsory subject from primary to higher education. It is also one of the national languages in India with Devanagari (सिन्धी) script. However, Sindhi Persian-Arabic (سنڌي) is the standard writing system. Both scripts differ from each other in terms of writing script, grammar, and vocabulary. Persian and Arabic languages influence Sindhi Persian-Arabic, while the writing system of Hindi influences Sindhi-Devanagari script.

Moreover, Sindhi-Roman[1] writing script is also receiving acceptance because of the online usage of Sindhi. Previously, Gujrati(گجراتي), Khudabadi (خدابادي), Gurumukhi (گرومکي), and Landa (لاندا) writing systems were also used for Sindhi writing (Motlani, 2016). The Persian-Arabic is standard and most famous writing scripts recognized in British-Colonial rule in 1852. Sindhi has six local and major dialects spoken in various regions of Pakistan and India, which differ in terms of pronunciation, vocabulary, and grammar. The major dialects [2] include Sindhi-Siraiki (سرائيڪي), Vicholi (وچولي), Laari (لاڙي), Laasi (لاسي), Thari (ٿري), and Kachi (ڪچي) respectively. Except from the above six major dialects, Macharia (مچيارا), Musalmani (مسلماني), and Dukslinu (Hindu-Sindhi) are also spoken in some regions of Pakistan and India. The Vicholi dialect is standard, widely spoken, recognized for administrative, literature and educational purposes.

Sindh province in Pakistan is the largest area of Sindhi native speakers. Also, a good number of Sindhi native speakers reside in Rajasthan, Ulhasnagar, Maharashtra, and Gujrat in India. Moreover, Sindhi is also the first language of native speakers who migrated to America, the United Kingdom, Tanzania, Hong Kong, Canada, Singapore, Philippines, Kenya, Uganda, South, and East Africa. The total number of Sindhi speakers is around 75 million (Motlani, 2016; Jamro, 2017) across the world. At present[3] many news literary, academic, and official blogs and websites in Pakistan and India have become a good source for text generation. Sindhi is a rich morphological cursive language like Arabic and Urdu. Its alphabet consists of 52 letters, 29 letters borrowed from Arabic, four from Persian, and 18 are modified letters. Sindhi words have the capacity to have multiple meanings, such polysemous situations are discussed in Section 5. Moreover, the absence of diacritic signs and many possible ways for word-formation make it a morphologically complex language.

3. Related Work

A large amount of NER resources is available for English and other European, East Asian languages. As a result, extensive research efforts have been made by using hand-crafted, language-independent, and hybrid approaches. In this section, we present the related work initiated on NER corpus development, including SSALs, along with the use of CRF and neural hybrid approaches in the NER task.

CoNLL: The CoNLL shared task (Sang and Erik, 2002; Sang and De Meulder, 2003) presented well-known NER datasets using British newswire for English, German, Dutch, Spanish European languages. Both datasets consist of four entity types of person, location, organization, and miscellaneous with the BIO labelling scheme.

MUC: The message understanding conference MUC-6 (Grishman and Sundheim, 1996) contains 318 annotated English news articles of wall street journal (WSJ) with seven named entity (NE) types of Person, Organization, Location, Date, Time, Money, and Percent.

IJCNLP: The IJCNLP-2008 workshop on SSALs low-resourced languages provided NER datasets for Hindi, Urdu, Bengali Oriya and Telugu languages with 12 predefined tags. Five teams took part in the manual annotation of datasets (Singh, 2008) by assigning one language to each team. A separate team created the corpus for each language using Shakti standard format (Bharati et al., 2007).

QUAERO-2009: The Quaero project reveals news NER corpus (Galibert et al., 2010) with a baseline for the French language. The news corpus is collected using an optical character recognition method with the participation of four groups mainly to develop and evaluate NE dataset.

AnCora: Bilingual multi-purpose annotated corpus (Taule´ et al., 2008) developed for Catalan and Spanish languages from journalist text for NER and other NLP tasks.

A gold-standard multilingual NER (Neudecker, 2016) corpus is proposed by collecting from Europeana newspapers for Dutch, French and Austrian languages using optical character recognition tool on newspaper pages. (Piskorski et al., 2017) created the first multilingual NER corpus for 7 Slavic languages namely Russian, Polish, Czech, Slavik, Ukrainian, Slovene, Croatian, by collecting news and web documents. More recently, Ghukasyan et al. (2018) propose a gold and silver standard NER datasets for the Armenian language with baseline. Initially, Ali et al. (2015); Nawaz et al. (2017) coined the related challenges and future research opportunities in Sindhi NER. Later, Hakro et al. (2017) propose a Sindhi NER system using a rule-based approach. Jumani et al. (2018) also applied the rule-based approach only on a small number of 936 words due to the lack of annotated corpora to address the problem in the Sindhi NER task. However, to the best of our knowledge, the work on the annotation of Sindhi NER does not exist.

The CRF framework for segmentation and labelling of sequential data (Lafferty et al., 2001) revealed significant research directions for the utilization of statistical models in classification problems. Later Chen et al. (2006) tackled the problem of Chinese NER using CRF, and Ekbal et al. (2008) also employed CRF model on five low-resourced SSALs namely Hindi, Urdu, Bengali, Telugu, and Oriya in IJCNLP-2008 shared task by showing that CRF approach can deal with diverse overlapping and non-independent features especially in inflective languages. The Bi-LSTM-CRF was first employed by Huang et al. (2015) to address the sequence tagging problem. Moreover, Kuru et al. (2016)

[1] https://sindhyat.com/database/SindhiRomanDictionary/

[2] https://www.indianmirror.com/languages/sindhi-language

[3] http://www.abyznewslinks.com/pakis.htm

proposed a character-level language-independent stack Bi-LSTM model with the Viterbi algorithm for counting probabilities converted to word-level NE tags. Later, the Bi-LSTM-CRF approach (Lample et al., 2016) yields state-of-the-art performance in Dutch, German, and Spanish languages. The hybrid Bi-LSTM-CNN-CRF model (Ma and Hovy, 2016) use the word and character-level input representations on CoNLL and WSJ datasets and achieves state-of-the-art performance without relying on any external task-specific resources. Furthermore, Misawa et al. (2017) also proposed the word- and character-level Bi-LSTM-CRF model by showing that CNN is not suitable for Japanese NER to extract sub-word information efficiently. Recently an LSTM-CRF model (Tran et al., 2017) with the utilization of bias decoding method also yielded state-of-the-art results on the CoNLL-2003 shared task.

4. SiNER: Our New Dataset

We introduce the first large SiNER dataset, which will be a sophisticated addition in the computational resources of Sindhi language. Most of the research on English NER datasets has focused on common entity types of Persons, Organizations, and Locations (Sang and De Meulder, 2003; Finkel et al., 2005; Derczynski et al., 2017) and numeral expressions (Strotgen and Gertz, 2013). Only a few corpora cover other entity types, such as Geopolitical entities and facilities (Doddington et al., 2004; Weischedel et al., 2011). Our proposed SiNER is large dataset and covers eleventh NE classes of Person (PERSON), Title (TITLE), Organization (ORG), country/states (GPE), location (LOC), parties/groups/agencies (NORP), government buildings (FAC), incidents (EVENT), languages (LANGUAGE), artwork (ART) and miscellaneous (OTHERS) depicted in Table 1. This section describes the methodology employed in the SiNER development and validation processes.

4.1. Corpus Acquisition

The recent work on Sindhi corpus development and neural word embeddings (Ali et al., 2019) propose a large amount of Sindhi corpus obtained from multiple web resources. We utilized the corpus (see Table 2) of Kawish and Awami Awaz Sindhi newspapers for the annotation project. The news corpus contains the latest vocabulary, rich in NEs, comprise events and regional, international news also well proofed before the publication of newspapers.

4.2. Preprocessing

The text preprocessing is a task-specific problem, and especially it becomes more challenging while working on low-resourced language like Sindhi because of the different writing styles of authors in news articles and borrowed words from other languages. Therefore, the Sindhi news corpus contains a little amount of unwanted data such as some vocabulary of other languages, mainly English words/acronyms, Urdu, and occasionally verses of Holy Quran and poetry. Therefore, it is essential to normalize the text or filter out such unwanted data to get a more authentic vocabulary. Firstly, we cleaned news articles for annotation, and secondly, the raw corpus (Ali et al., 2019) for training Sindhi word embeddings by designing a preprocessing pipeline described as follows:

Label	Description
PERSON	Names, including fictional
TITLE	Titles of person, designation or rank, etc.
ORG	Companies, Institutions
GPE	Continents, States, Countries Cities
LOC	Towns, villages, Non-GPE location of mountain ranges and bodies of water
NORP	Nationalities, agencies, political parties, religious groups etc.
FAC	Government buildings, airports
EVENT	Incidents, Wars, battles, sports festivals and special days
LANGUAGE	Names of languages
ART	Title of books, songs, movies, and any other art-related work
OTHERS	Date, time, percentage, quantity, money, abbreviations, disease, seasons, games, ordinal, and cardinal numerals, etc.

Table 1: Description of named entity labels in proposed SiNER dataset

Resource	Articles	Sentences	Tokens
Kawish	791	26,170	906,915
Awami awaz	547	15,579	451,809
Total	1,338	41,749	1,358,724

Table 2: News articles used for annotation of SiNER dataset

Input: Concatenated all the corpus files and input in UTF-8 text format.

Replacement symbols: The punctuation marks including, hyphen, apostrophe, comma, quotation, and exclamation marks replaced with white space for authentic tokenization.

Filtration of noisy data: We filtered out unimportant data such as the rest of the punctuation marks, special characters, HTML tags, all types of numeric entities, email, and web addresses.

Normalization: We tokenized the corpus then normalized to lower-case for the filtration of multiple white spaces, unwanted borrowed words from other languages such as English and duplicate words. Sindhi stop words were filtered for learning Sindhi GloVe word representations.

4.3. Annotation Methodology

The corpus annotation is an expensive activity; we use Doccano (Nakayama et al., 2018) a text annotation tool for annotation of SiNER. It is a web-based open source annotation platform for sequence labelling, sentiment analysis, and machine translation. The native graduate students of linguistics performed the annotation task with

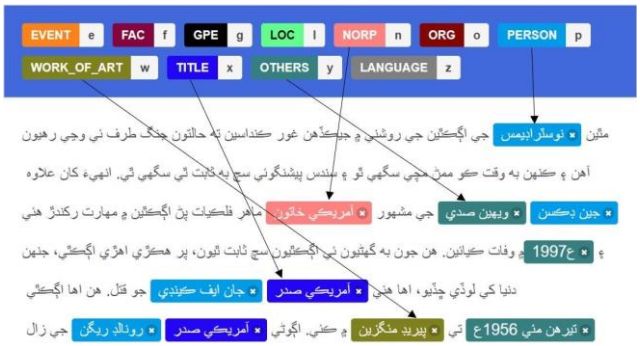

Figure 1: Graphical user interface of the web-based Doccano annotation platform used for the annotation of SiNER dataset.

the supervision of NLP and linguist expert for the authentication in assigning labels. However, the project supervisor worked with language-specific lead annotators to develop and maintain formal annotation task definitions, guidelines, train annotators and monitor annotation quality. The annotation interface of Doccano is shown in Figure 1, a detailed label distribution of NEs in SiNER is depicted in Figure 2, and complete statistics of the proposed dataset used in experimental setup is given in Table 4, respectively.

4.4. Manual Evaluation

We manually validate the SiNER dataset by checking NEs and assigned tags after completing the annotation task. As we mentioned earlier that Sindhi news articles contain some loan words of other languages mainly, English in the form of abbreviations, titles of books/songs, and names of movies etc. Therefore, we replaced such types of loan/borrowed words with Sindhi words. In this step, ambiguous, missing and improper tags were manually corrected in the proposed dataset. The complete manual evaluation process consists of the following steps:

Validation of ambiguous entities: Although the annotators were native Sindhi speakers, some ambiguous NEs require validation to authenticate the labels. In this step, the labels were validated according to their contextual meaning.

Missing Labels: Due to the large annotation task, there are many possibilities of missing entities which may lead to poor quality of dataset as well as the accuracy of the NER system. Therefore, nearly 2% of missing entities were labelled in the final validation process.

Correction of improper tags: Different annotators perform the annotation task. Therefore, we manually corrected and validated the NEs to ensure the authentication of labels.

Replacement of English acronyms: Some writers use English abbreviations in Sindhi news articles. However, it is not common practice to use English abbreviations instead of Sindhi. Therefore, for the consistency of proposed dataset, we manually replaced such abbreviations with Sindhi by following the rules of acronyms used in Sindhi language, such as BBC was replaced with (بي بي سي) CNN with (سي اين اين), UNO with (يو اين او) and UNESCO (يونيسڪو).

4.5. SiNER Format

We follow the standard BIO2 format (Sang and Veenstra, 1999; Sang and Erik, 2002) for the SiNER dataset depicted

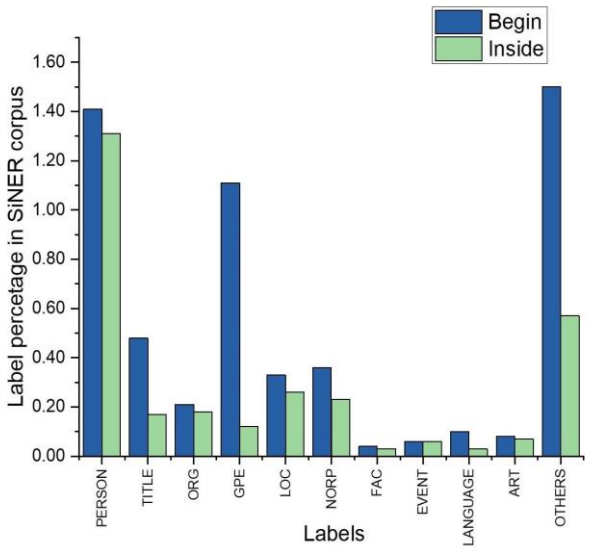

Figure 2: The label distribution in SiNER dataset. The number of single entities is larger in GPE and OTHERS labels. However, the number of nested NEs is higher in PERSON tags.

in Table 3. The given example of a Sindhi sentence " جين ڊڪسن ويهين صدي جي مشهور آمريڪي خاتون ماهرِ فلڪيات اڳڪٿين ۾ مهارت رکندڙ هئي." means "*The famous American lady astronomer named Jane Dickens was adept in the art of predictions in the twentieth century*". The BIO tagging scheme is short for Begin, Inside and Outside, commonly used for the tagging of tokens in chunking and NER datasets. The B-prefix at the beginning of NE labels indicates the beginning of a name in NER system, and an I-prefix before a label indicates the nested name and O-tag indicates that a token does not belong to NEs.

Named entity	Roman Transliteration	Tag
جين	Jane	B-PERSON
ڊڪسن	Dickens	I-PERSON
ويهين	veehen	B-OTHERS
صدي	sadee	I-OTHERS
جي	gi	O
مشهور	mashahoor	O
آمريڪي	aamreeki	B-NORP
خاتون	khatoon	I-NORP
ماهرِ	mahirai	O
فلڪيات	falakyat	O
اڳڪٿين	agkathyune	O
۾	mein	O
مهارت	maharat	O
رکندڙ	rakhandar	O
هئي	hui	O
.	.	O

Table 3: The format of SiNER dataset, similar to the CoNLL-2003 shared task. The Roman transliteration is given for the ease of reading

5. Experiments and Results

5.1. Conditional Random Field (CRF)

We initially evaluate the SiNER dataset using a CRF based approach, widely used in sequence classification problems (Chen et al., 2006; Sutton et al., 2012). The CRF proposed by (Lafferty et al., 2001) for modelling sequential data such as word labels in a given input sentence offers several advantages for sequence segmentation and labelling tasks. It is useful to consider the relationship between surrounding labels and jointly decode the most suitable chain of labels for an input sentence (Ma and Hovy, 2016). Another advantage of using CRF is its rich feature sets, e.g., overlapping features using conditional probability (Sutton et al., 2012). Such as, given an input sequence $X = x_1, x_2 \dots \dots x_n$ and sequence of NE tags $Y = y_1, y_2 \dots \dots y_n$ and $P(Y|X)$ conditional probability is defined by CRF as follows:

$$P(Y|X) \propto \exp\left(w^T f(y_n, Y_{n-1}, x)\right) \quad (1)$$

Where w is a weight vector $w = (w_1, w_2 \dots w_m)^T$ that maps entire X input sequence to entire Y into $\mathbb{R}^d$ as a log-linear model with parameter vector $w \in \mathbb{R}^d$. The regularization log-likelihood $L(w)$ function can be defined as:

$$\sum_{i=1}^{n} \log P\left(y^i \middle| x^i; w\right) - \frac{\lambda_2}{2}\|w\|_2^2 - \lambda_1\|w\|_1 \quad (2)$$

The vector parameters are forced to be trivial by terms $\frac{\lambda_2}{2}\|w\|_2^2$ and $\lambda_1\|w\|_1$ in normalization. The vector parameter w^* is estimated as:

$$w^* = \arg\ ma\, x_{w\in\mathbb{R}} d\ L(w) \quad (3)$$

After the estimation of w^* the most likely tag of y^* can be found by y^*

$$s^* = \arg\ ma\, x_s P(y|x; w^*) \quad (4)$$

We choose exact match matrices introduced in the CoNLL shared-task (Sang and De Meulder, 2003) for the evaluation of SiNER using a 5-fold cross-validation scheme. The 80% data is utilized for training and 20% for testing of the CRF model. We evaluate the model with three evaluation measures of Precision, Recall, and F1-Score. The label wise detailed observed results are presented in Table 5.

5.2. Bi-Directional Long Short-Term Memory Network

In this section, we briefly describe the opted neural models to evaluate SiNER for quality baselines. The dataset is divided into train, validation, and test sets. We use PyTorch (Paszke et al., 2017) deep learning framework for the implementation of neural models on GTX 1080-TITAN GPU for all the experiments.

5.3. Bi-LSTM Architecture

The LSTM belongs to recurrent neural network family proposed by (Hochreiter and Schmidhuber, 1997) since then, it is widely used to address sequence-tagging problems in NLP applications. The Bi-LSTM predict sequences by giving an input words sequence ($x_1, x_2 \dots x_n$) of a sentence containing n words, each represented as N-dimensional vector, returns another sequence ($h_1, h_2, \dots h_n$), which represents sequence information at every time step. A forward LSTM compute left context $\overrightarrow{h}$ of the given input sentence and backward LSTM compute right context $\overleftarrow{h}$ of every word then combine both $\overrightarrow{h_t}, \overleftarrow{h_t}$ to generate output y_t. This way, Bi-LSTM can capture more information with two separate hidden states to predict past and future information efficiently. In the Bi-LSTM architecture (Figure 3) we use word embeddings or character-representations as an input to Bi-LSTM encoder. The set of characters consists of all unique characters in the SiNER dataset. As we mentioned earlier, Sindhi Persian-Arabic is written in the right to left direction. Therefore, the forward LSTM runs from the end of a sentence, and backward LSTM runs from the beginning. The output of both forward and backward states is concatenated to use as input for classifier either softmax or CRF. The Bi-LSTM output of the softmax classifier is mapped through a dense layer with a softmax activation function. In such a way, each token in a sentence is given a probability distribution for the possible labels to select a label with maximum probability. However, with CRF, the output of Bi-LSTM encoder is mapped to the number of tags through a dense layer and linear activation function to the number of labels for CRF-classifier. Afterwards, linear CRF chain maximizes the label probability of whole sentence.

5.4. Word Embeddings

Collobert et al., 2011 showed that neural models could gain better performance with word embeddings. Recently, word embedding learned on large unlabeled corpus has become an integral part of neural models in NLP applications with great ability to improve the performance of the neural models. The GloVe (Pennington et al., 2014) is a log-bilinear regression model that combines two methods of context window and global matrix factorization for training word embeddings of a given vocabulary in an unsupervised

SiNER	Training	Development	Test
Sentences	28,259	7,465	6,006
Tokens	791,948	285,203	281,540
Entities	68,195	37,241	18,003
Nil entities	723,753	247,962	263,537

Table 4: The complete statistics of SiNER dataset used in training, development and test experiments

Tags	Precision	Recall	F1
PERSON	93.21	88.31	90.45
TITLE	92.64	92.53	91.54
ORG	88.52	75.30	81.60
GPE	91.47	85.12	88.50
LOC	86.52	59.29	71.62
NORP	94.26	93.38	94.87
FAC	96.59	93.77	94.62
EVENT	74.54	64.92	69.85
LANGUAGE	93.58	96.67	94.65
ART	43.66	84.83	52.59
OTHERS	85.92	86.76	85.63
Average	84.77	83.257	82.54

Table 5: The label wise initial baseline results in the macro average score using CRF approach

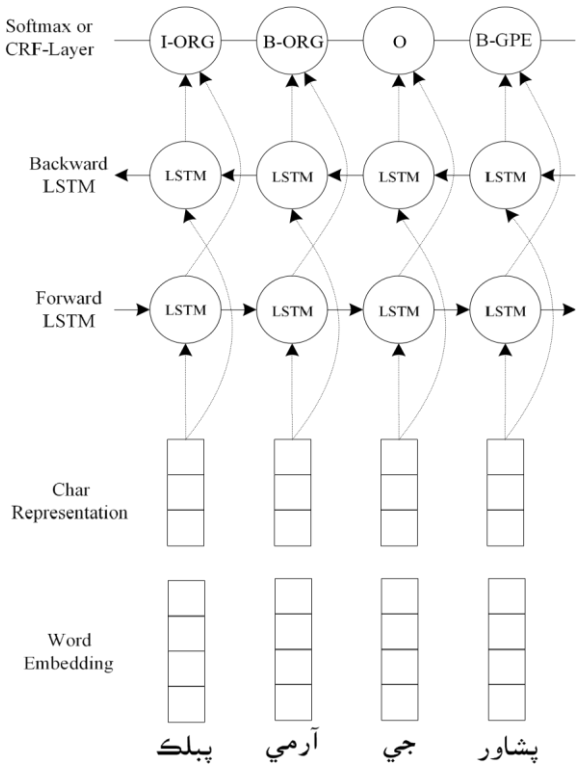

Figure 3: The label distribution in SiNER dataset. The number of single entities is larger in GPE and OTHERS labels.

way. The fastText (Bojanowski et al., 2017) is a recent model for learning neural word representations by representing each word as a bag-of-character n-grams instead of directly learning the single word. Such as a word "sindhi" with n=3, the fastText representation of this word is <si, sin, ind, ndh, dhi, hi>, where the angular brackets show the beginning and end of the word. In this way, fasText captures the meaning of words as bag-of-character n-gram, which allows the word embeddings to understand suffixes and prefixes. More recently, (Grave et al., 2018) proposed Sindhi word embeddings by only extracting Wikipedia and common crawl data. Hence the vocabulary is limited, and the quality is also lower than most recently proposed (Ali et al., 2019) Sindhi word embeddings. We utilize the proposed news corpus (Ali et al., 2019) for the annotation of SiNER dataset and training of GloVe and fastText Sindhi word embeddings in the experimental setup.

5.5. Hyperparameters

We choose similar hyper-parameters in all the experiments to analyze the performance difference by using 100-dimensional GloVe, fastText word embeddings and 25-dimensional character representations. The hidden size of Bi-LSTM is set to 200 and learning rate to 0.025, respectively. We apply the drop-out of 0.5 to avoid the overfitting problem (Srivastava et al., 2014) through all the experiments. Moreover, early stopping is used for regularization. We use Adamax optimizer for all Bi-LSTM models.

5.6. Results

Firstly, we run the experiments using the CRF approach, which consequently performs well on all the NE classes in terms of precision, recall, and F1-score, respectively. The CRF forms a strong baseline with the average F1-score of 82.54% on the test dataset. Moreover, the Bi-LSTM and BiLSTM-CRF neural network models with and without using character-level representations produce excellent results as compared to the CRF approach. It is encouraging that without relying on any language-specific setting, we achieve the F1-score of 84.67% with the Bi-LSTM-CRF-Char model using Glove Sindh word representations and, the same model achieves high F1-score of 89.16% on the fastText Sindhi word representations by showing that the character-level representations are important for sequence labelling tasks. We present the detailed macro average scores yield by all the employed models in Table 6. The CRF is dominant over softmax classifier, and character representations also helped in performance gain in deep learning setup. The fastText word embeddings are enriched with sub-word representations, which helped in performance gain over the word-level approach.

5.7. Challenges in Sindhi Named Entity Recognition

Sindhi language contains a huge number of polysemous words bearing different meanings which change their meaning according to grammatical positions in the sentences. Additionally, the absence of diacritic symbols also creates many ambiguous situations in dealing with polysemous words because modern Sindhi Persian-Arabic is written without assigning diacritic symbols in daily life. In this section, some analyzed ambiguities in SiNER are discussed briefly as follows:

Lack of Capitalization: There is no difference between plain text and NEs in Sindhi, while English has the capitalization rule as an important feature to enhance the accuracy of the NER system.

Multi-type entities: There are a lot of examples for such polysemous NEs, such as (Sindhu-سنڌو) is the name of girl tagged as PERSON. Also, the name of a river (Sindhu-سنڌو) tagged as LOC, and a verb (Sandho-سنڌو) means partition.

	Model	Precision	Recall	F1-Score
	CRF	84.77	83.25	82.54
Glove	Bi-LSTM	82.33	84.38	83.34
	Bi-LSTM +Char	83.69	85.65	84.64
	Bi-LSTM +CRF	86.84	81.74	84.21
	Bi-LSTM +CRF+Char	84.40	84.93	84.67
fastText	Bi-LSTM	86.87	**87.82**	87.07
	Bi-LSTM +Char	87.24	87.59	87.42
	Bi-LSTM +CRF	**89.72**	86.94	**88.09**
	Bi-LSTM +CRF+Char	**90.83**	87.54	**89.16**

Table 6: Comparison of the CRF and Bi-LSTM models on SiNER test dataset using GloVe and fastText Sindhi word representations. The bold results highlight the best results

Another example is (Bihar-بهار) GPE, a state in India. However, it is also the name of season (Bahar- بهار) means spring, labelled as OTHERS and, thirdly, the same word (bahar-بهار) is used to express happiness.

Name of person: Ambiguity in the person names is also a common phenomenon in Sindhi, such as the name of a person (Wazeer-وزير) is also the title of a person which means minister. The name (Suhni-سهڻي) is also an adjective that means beautiful.

Country names: There is no difference between country name Syria (شام-Sham) tagged as GPE and evening (شام) in Sindhi. There is plenty of such polysemous words which also create ambiguous situations.

Location: An example of a word president (Sadar-صدر) is labelled as TITLE. However, it is also the name of a town that lies in the LOC category.

Cardinal numbers: The numbers also can an ambiguous situation with common words such as eight (atha-اٺ) is a number and (uthu-اٺ) also means camel another example is ten (daha-ڏهہ) and (duho-ڏهہ) means milking. Such ambiguities can only be handled by assigning the diacritic signs.

6. Discussion and Conclusions

The corpus acquisition, preprocessing, annotation and evaluation of a low-resourced language like Sindhi is a challenging task. We utilized news corpus for the annotation of SiNER dataset because Sindhi language does not have any authentic software for spell checking, so the news articles contain least spelling errors. Therefore, news corpus is the most suitable choice for the annotation purpose. Moreover, the text annotation for a specific NLP application is a labour-intensive task that requires careful assessment to maintain the quality of the gold-standard dataset. Therefore, graduate students of linguistics took part in the annotation project. Afterwards, we manually validate the SiNER dataset for the authentication of NE tags, which is also an expensive activity. But such a validation process is essential in the development and evaluation of a novel gold-standard dataset.

In this paper, we present the first large SiNER dataset for low-resourced Sindhi language with quality baselines. Our work mainly consists of three novel contributions. Firstly, a gold-standard SiNER dataset is annotated using web-based Doccano text annotation tool. Secondly, we present quality baselines using the CRF and recent state-of-the-art deep neural sequence classification models of Bi-LSTM, Bi-LSTM-CRF, and Bi-LSTM-CRF-Char using GloVe and fastText word representations. Thirdly, we compare the performance of the CRF with the Bi-LSTM models. The Bi-LSTM-CRF-Char model yields an encouraging F1-score of 84.67% with GloVe, and the best F1-score of 89.16% with fasText word embeddings. Conclusively, we address the problem of NER in Sindhi language by proposing a novel gold-standard dataset and utilize deep learning machinery for the first time for low-resourced Sindhi language. In the future, we will design a deep neural algorithm to analyze the impact of internal and external word embeddings and joint learning model of Sindhi parts-of-speech tagging and NER tasks.

7. Acknowledgements

This work was funded by the National Key R&D Program of China (No. 2018YFB1005100 & No. 2018YFB1005104).

8. Bibliographical References

Ali, W., Kehar, A., and Shaikh, H. (2015). Towards Sindhi named entity recognition: Challenges and opportunities. In *1st National Conference on Trends and Innovations in Information Technology*.

Ali, W., Kumar, J., Lu, J., and Xu, Z. (2019). A new corpus for low-resourced Sindhi language with word embeddings. *arXiv preprint arXiv:1911.12579*.

Babych, B. and Hartley, A. (2003). Improving machine translation quality with automatic named entity recognition. In *Proceedings of the 7th International EAMT workshop on MT and other language technology tools, Improving MT through other language technology tools, Resource and tools for building MT at EACL 2003*.

Bharati, A., Sangal, R., and Sharma, D. M. (2007). Ssf: Shakti standard format guide. *Language Technologies Research Centre, International Institute of Information Technology, Hyderabad, India*, pages 1–25.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Chen, W., Zhang, Y., and Isahara, H. (2006). Chinese named entity recognition with conditional random fields. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, pages 118–121.

Chiu, J. P. and Nichols, E. (2016). Named entity recognition with bidirectional LSTM-CNNS. *Transactions of the Association for Computational Linguistics*, 4:357–370.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of machine learning research*, 12(Aug):2493–2537.

Derczynski, L., Nichols, E., van Erp, M., and Limsopatham, N. (2017). Results of the WNUT-2017 shared task on novel and emerging entity recognition. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 140–147.

Doddington, G. R., Mitchell, A., Przybocki, M. A., Ramshaw, L. A., Strassel, S. M., and Weischedel, R. M. (2004). The automatic content extraction (ace) program tasks, data, and evaluation. In *LREC*, volume 2, page 1. Lisbon.

dos Santos, C., Guimaraes, V., Niteroi, R., and de Janeiro,´ R. (2015). Boosting named entity recognition with neural character embeddings. In *Proceedings of NEWS 2015 The Fifth Named Entities Workshop*, page 25.

Ekbal, A., Haque, R., Das, A., Poka, V., and Bandyopadhyay, S. (2008). Language independent named entity recognition in Indian languages. In *Proceedings of the IJCNLP-08 Workshop on Named Entity Recognition for South and South East Asian Languages*.

Finkel, J. R., Grenager, T., and Manning, C. (2005). Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 363–370. Association for Computational Linguistics.

Galibert, O., Quintard, L., Rosset, S., Zweigenbaum, P., Nedellec, C., Aubin, S., Gillard, L., Raysz, J.-P., Pois,´ D., Tannier, X., et al. (2010). Named and specific entity detection in varied data: The Quæro named entity baseline evaluation. In *LREC*.

Ghukasyan, T., Davtyan, G., Avetisyan, K., and Andrianov, I. (2018). Pioner: Datasets and baselines for Armenian named entity recognition. In *2018 Ivannikov Ispras Open Conference (ISPRAS)*, pages 56–61. IEEE.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., and Mikolov, T. (2018). Learning word vectors for 157 languages. In *LREC*.

Grishman, R. and Sundheim, B. M. (1996). Message understanding conference-6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

Hakro, M. A., Lashari, I. A., et al. (2017). Sindhi named entity recognition (SNER). *The Government-Annual Research Journal of Political Science.*, 5(5).

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory.Neural computation, 9(8):1735–1780.

Huang, Z., Xu, W., and Yu, K. (2015). Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Jamro, W. A. (2017). Sindhi language processing: A survey. In *2017 International Conference on Innovations in Electrical Engineering and Computational Technologies (ICIEECT)*, pages 1–8. IEEE.

Jumani, A. K., Memon, M. A., Khoso, F. H., Sanjrani, A. A., and Soomro, S. (2018). Named entity recognition system for Sindhi language. In *International conference for emerging technologies in computing*, pages 237–246. Springer.

Kuru, O., Can, O. A., and Yuret, D. (2016). Charner: Character-level named entity recognition. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 911–921.

Lafferty, J., McCallum, A., and Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C. (2016). Neural architectures for named entity recognition. In *Proceedings of NAACLHLT*, pages 260–270.

Ma, X. and Hovy, E. (2016). End-to-end sequence labeling via bi-directional LSTM-CNNS-CRF. *arXiv preprint arXiv:1603.01354*.

Misawa, S., Taniguchi, M., Miura, Y., and Ohkuma, T. (2017). Character-based bidirectional LSTM-CRF with words and characters for Japanese named entity recognition. In *Proceedings of the First Workshop on Subword and Character Level Models in NLP*, pages 97–102.

Moldovan, D. (2002). LCC tools for question answering Dan Moldovan, Sanda Harabagiu, Roxana Girju, Paul Morarescu, Finley Lacatusu, Adrian Novischi, Adriana Badulescu and Orest Bolohan Language Computer Corporation Richardson, tx 75080.

Motlani, R. (2016). Developing language technology tools and resources for a resource-poor language: Sindhi. In *Proceedings of the NAACL Student Research Workshop*, pages 51–58.

Nakayama, H., Kubo, T., Kamura, J., Taniguchi, Y., and Liang, X. (2018). Doccano: Text annotation tool for human. Software available from https://github.com/doccano/doccano.

Nawaz, D., Awan, S., Bhutto, Z., Memon, M., and Hameed, M. (2017). Handling ambiguities in Sindhi named entity recognition (SNER). *Sindh University Research JournalSURJ (Science Series)*, 49(3):513–516.

Neudecker, C. (2016). An open corpus for named entity recognition in historic newspapers. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4348–4352.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. (2017). Automatic differentiation in PyTorch. In *NIPS-W*.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532– 1543.

Piskorski, J., Pivovarova, L., Snajder, J., Steinberger, J., and˘ Yangarber, R. (2017). The first cross-lingual challenge on recognition, normalization, and matching of named entities in Slavic languages. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*, pages 76–85.

Sang, E. T. K. and De Meulder, F. (2003). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLTNAACL 2003*, pages 142–147.

Sang, T. K. and Erik, F. (2002). Memory-based named entity recognition. In *proceedings of the 6th conference on Natural language learning-Volume 20*, pages 1–4. Association for Computational Linguistics.

Sang, E. F. and Veenstra, J. (1999). Representing text chunks. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, pages 173–179. Association for Computational Linguistics.

Schuster, M. and Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681.

Singh, A. K. (2008). Named entity recognition for South and South East Asian languages: taking stock. In *Proceedings of the IJCNLP-08 Workshop on Named*

Entity Recognition for South and South East Asian Languages.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Strotgen, J. and Gertz, M. (2013). Multilingual and cross-¨ domain temporal tagging. *Language Resources and Evaluation*, 47(2):269–298.

Sutton, C., McCallum, A., et al. (2012). An introduction to conditional random fields. *Foundations and Trends* R *in Machine Learning*, 4(4):267–373.

Taule, M., Martí, M. A., and Recasens, M. (2008). Ancora: Multilevel annotated corpora for Catalan and Spanish. In *LREC*.

Tran, Q. H., MacKinlay, A., and Yepes, A. J. (2017). Named entity recognition with stack residual LSTM and trainable bias decoding. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 566–575.

Weischedel, R., Pradhan, S., Ramshaw, L., et al. (2011). Ontonotes release 4.0.