Development of a Dataset and a Deep Learning Baseline Named Entity Recognizer for Three Low Resource Languages: Bhojpuri, Maithili and Magahi

RAJESH KUMAR MUNDOTIYA, Department of Computer Science and Engineering, IIT(BHU)
SHANTANU KUMAR, Department of Linguistics, Banaras Hindu University
AJEET KUMAR, Cognizant
UMESH CHANDRA CHAUDHARY, Cognizant
SUPRIYA CHAUHAN, Department of Linguistics, Banaras Hindu University
SWASTI MISHRA, Department of Humanistic Studies, IIT(BHU)
PRAVEEN GATLA, Department of Linguistics, Banaras Hindu University
ANIL KUMAR SINGH, Department of Computer Science and Engineering, IIT(BHU)

In Natural Language Processing (NLP) pipelines, Named Entity Recognition (NER) is one of the preliminary problems, which marks proper nouns and other named entities such as Location, Person, Organization, Disease etc. Such entities, without an NER module, adversely affect the performance of a machine translation system. NER helps in overcoming this problem by recognising and handling such entities separately, although it can be useful in Information Extraction systems also. Bhojpuri, Maithili and Magahi are low resource languages, usually known as Purvanchal languages. This paper focuses on the development of an NER benchmark dataset for Machine Translation systems developed to translate from these languages to Hindi by annotating parts of the available corpora with named entities. Bhojpuri, Maithili and Magahi corpora of sizes 228373, 157468 and 56190 tokens, respectively, were annotated using 22 entity labels. The annotation considers coarse-grained annotation labels followed by the tagset used in one of the Hindi NER datasets. We also report a Deep Learning baseline that uses an LSTM-CNNs-CRF model. The lower baseline F1-scores from the NER tool obtained by using Conditional Random Fields models are 70.56% for Bhojpuri, 73.19% for Maithili and 84.18% for Magahi. The Deep Learning-based technique (LSTM-CNNs-CRF) achieved 61.41% for Bhojpuri, 71.38% for Maithili and 86.39% for Magahi. As the results show, LSTM-CNNs-CRF fails to outperform the lower baseline in the case of Bhojpuri and Maithili, which have more data in terms of the number of tokens, but not in terms of the number of named entities. However, the cross-lingual model training of LSTM-CNNs-CRF for Bhojpuri and Maithili performed better than the CRF.

CCS Concepts: · Computing methodologies → Language resources; Machine translation; Information extraction; Lexical semantics.

Authors’ addresses: Rajesh Kumar Mundotiya, rajeshkm.rs.cse16@iitbhu.ac.in, Department of Computer Science and Engineering, IIT(BHU), Varanasi, Uttar Pradesh, 221005; Shantanu Kumar, Department of Linguistics, Banaras Hindu University, Varanasi, Uttar Pradesh, 221005, shantanubhu97@gmail.com; Ajeet kumar, Cognizant, Banglore, Karnataka, ajeet786sanjay@gmail.com; Umesh Chandra Chaudhary, Cognizant, Banglore, Karnataka, umeshchandra8294@gmail.com; Supriya Chauhan, Department of Linguistics, Banaras Hindu University, Varanasi, Uttar Pradesh, 221005, supriyachauhanaisha@gmail.com; Swasti Mishra, Department of Humanistic Studies, IIT(BHU), Varanasi, Uttar Pradesh, 221005, swastimish@iitbhu.ac.in; Praveen Gatla, Department of Linguistics, Banaras Hindu University, Varanasi, Uttar Pradesh, 221005, praveengatla@gmail.com; Anil Kumar Singh, Department of Computer Science and Engineering, IIT(BHU), Varanasi, Uttar Pradesh, 221005, aksingh.cse@iitbhu.ac.in.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2022 Association for Computing Machinery.
2375-4699/2022/7-ART $15.00
https://doi.org/10.1145/3533428

ACM Trans. Asian Low-Resour. Lang. Inf. Process.
1 INTRODUCTION

Named Entity Recognition (NER) is the process of identification of named entities (Person, Organization, Location etc.) in natural language text. The present paper concentrates on three low resource languages (LRLs): Bhojpuri, Maithili and Magahi (BMM), which belong to the Indo-Aryan language family. This work may be seen as the first attempt to develop an NER tool for Bhojpuri and Magahi. There is no previous work on NER for these languages as far as we know. Recently, Priyadarshi and Saha [2021] have developed the NER tool for Maithili. The main aim of the present paper is to start with insights from the NER systems that are developed for Indian Languages with more resources and based on that we try to develop an NER System for BMM.

Named Entity was introduced in the Sixth Message Understanding Conference (MUC-6) [Grishman and Sundheim 1996]. It was often seen as part of an Information Extraction system, which refers to the automatic extraction of structured information such as entities, relationships between entities and attributes describing entities from unstructured sources. The NEs were identified in two conventional ways, before the recent success of machine learning and then Deep Learning based techniques:

1. A raw sentence was compared with gazetteer’s lists to identify the NEs, where the gazetteer’s lists are created manually for person names, location names and organization names etc.
2. One may identify the Named Entities based on language specific linguistic rules. For example, proper nouns always start with capital letters in English, e.g. London, Shakespeare, Darwin etc.

It is a challenging task to implement NER for Indian languages due to the absence of capitalization in their writing systems. On the other hand, these systems are phonetically organized and designed, which makes it easy possible to use phonetic features for NER for Indian languages. Preparing a gazetteer’s list for all nouns is impossible because there can be a vast number of unknown named entities in the world in, since a corpus is ultimately a small sample of all possible sentences of a language. Even though named entities may be finite in number at a point in time, but they are rarely, if ever, all present in the same corpus that the machine learns from.

Not much work has been reported for NER for most Low Resource languages due to insufficient annotated corpus (or no corpus at all). There have been efforts on major Indian languages, i.e., Hindi, Tamil, Telugu, Urdu, Punjabi, but no efforts on Low Resource Indian languages such as BMM.

1.1 Bhojpuri, Maithili, Magahi : An Introduction

Bhojpuri is often considered a major ‘sub-language’ of Hindi. It is not only a language which is spoken in various states of India but in other countries as well, viz. Nepal, Mauritius, Fiji, Surinam etc. The writing system of Bhojpuri was earlier the Kaithi script but now Devanagari script is used more to write Bhojpuri. According to 2011 census[1], there are 5,05,79,447 Bhojpuri speakers.

Maithili, now a language, belongs to the Indo-Aryan language family, while Bhojpuri and Magahi are considered ‘sub-languages’ (or even dialects) of Hindi and are mainly spoken in Eastern Uttar Pradesh, Bihar and Jharkhand states of India. Maithili is included in the 22 ‘scheduled’ languages of the Republic of India (1950, Constitution, Article 343). It was added in the Constitution of India in 2003 by the 92nd Constitutional Amendment Act. It is spoken mainly in India in Bihar, Jharkhand, Uttar Pradesh etc. as well as in Nepal. It is the only language in the Bihari sub-family that is included in the eighth schedule of the Indian constitution. There are 1,35,83,464 Maithili speakers (Census, 2011). It is also one of the 122 recognised languages of Nepal. In 2007, Maithili was included in the interim Constitution of Nepal and in March 2018, it received the second official language status in the Jharkhand state of India. It too was earlier considered a sub-language or a dialect.

[1]https://www.censusindia.gov.in/2011Census/Language-2011/Statement-1.pdf
Magahi or Magadhi, also considered a major sub-language of Hindi, is chiefly spoken in some districts of Bihar, Jharkhand, and also in the Maldah district of West Bengal. Magahi was also written in the Kaithi script in earlier days, but at present it is usually written in the Devanagari script. There are 1,27,06,825 Magahi speakers (Census, 2011).

Earlier work on machine translation (particularly rule-based or transfer-based) had reported that proper handling of named tokens could improve the translation quality and performance [Babych and Hartley 2003; Bhalla et al. 2013; Vu et al. 2020]. These named tokens would have been (mis)translated during source to target translation without an NER module, but with an NER module they can instead be simply transliterated. Even though the MT systems are based on a transfer approach, the NER module (like the POS tagging and Chunking modules) is based on machine learning or Deep Learning, not a rule-based approach. Due to this, we have annotated some corpus and developed an NER system for these three languages and have reported the lower and a higher baseline results. The former is based on CRF and the latter on a combination of Long Short Term Memory (LSTM), Convolutional Neural Network (CNN) and Conditional Randon Fields (CRF), called LSTM-CNNs-CRF.

1.2 Contributions
As there is no prior work on the NER problem for Bhojpuri and Magahi. Although a notable work found for Maithili, the contributions in this paper are as follows:

- Annotation of NEs in Bhojpuri, Maithili and Magahi corpora, with the sizes being 228373, 157468 and 56190 tokens, respectively, using 22 entity labels at the fine-grained level.
- Provide benchmarking results (F1-score) on these annotated datasets by using a conventional machine learning technique (CRF).
- Apply a state-of-the-art technique (LSTM-CNNs-CRF) to improve the benchmarking scores and provide an upper baseline for future NER for these languages.
- Analyzed the effect of cross-lingual training of the LSTM-CNNs-CRF model.

2 RELATED WORK
An NER module can be a part of several Natural Language Processing and Understanding systems. Earlier work relied on rule-based techniques, which used orthographic features, lexicons and ontologies. Rau [1991] reported an initial work for extracting company names by using heuristics and a hand-crafted feature-based algorithm. Later, feature engineering techniques evolved with machine learning. Weak supervision was also a promising approach, so a bootstrapping method was used, which found contextual patterns through seed entities and ranked them [Riloff et al. 1999]. However, more accurate contextual information could be gathered from syntactic relations [Cucchiarelli and Velardi 2001]. Pasca et al. [2006] generated synonyms by distributional similarity for generalized contextual patterns. Zhang and Elhadad [2013] used Inverse Document Frequency (IDF) and shallow syntactic knowledge, which filters lower IDF ranked terms before the prediction of the classifier. Apart from this, pointwise mutual information that is commonly used for Information Retrieval, was exploited to classify name entities [Etzioni et al. 2005]. WordNet [Miller 1995] was also employed for labelling NEs by selecting the NE synsets from WordNet, based on the frequent appearance of entities in the corpus [Alfonseca and Manandhar 2002].

The utility of machine learning techniques was demostrantrated in the CoNLL-2003 shared task 2, organised on four different languages: Spanish, Dutch, German and English, each with four entities (Person, Location, Organization and Miscellaneous). On the CoNLL task, various machine learning techniques have been evaluated which cover AdaBoost, Hidden Markov Model, Maximum Entropy, CRF, Memory-based Learning, Transformation-based learning, Support Vector Machine (SVM), Recurrent Neural Networks, Voted Perceptron and combinations

2https://www.clips.uaantwerpen.be/conll2003/ner/
of them with rules or handcrafted features\(^3\). Similarly, SVM, neural networks and Decision Trees were exploited for Hungarian named entity classification at the phrase level [Farkas and Szarvas 2006]. Semantic features and gazetteers have been used with the Bayesian network to recognize NEs of the Spanish language [Padró and Padró 2005]. Ando and Zhang [2005] used a structured machine learning algorithm for performing a multi-task learning based approach, where label prediction was considered as the primary task and masking of the current word was an auxiliary task. The classifier was selected based on the performance on the auxiliary task.

A neural network architecture for NER was developed in 2008 by Collobert and Weston [2008], which relied on feature engineering, a dictionary, lexicon and orthographic features. Later, this architecture was modified with automatic feature extraction (at the level of word embedding) instead of using a feature engineered method [Collobert et al. 2011]. Deep learning models take input with units as words, characters, affixes and combinations of them, or even bytes. Collobert et al. [2011] model comprised of word-based features that are passed to the CRF layer via a convolutional layer. Later, the same model was enhanced by sequential features (an LSTM layer) on the English CoNLL-2003 dataset [Huang et al. 2015]. Kim et al. [2016] exploited character level features by generating word embeddings using bidirectional LSTM with Convolutional Neural Network (CNN) as a highway network. The predictions were made with softmax instead of CRF layer. Ma and Hovy [2016a] used a combination of characters and a word level representation and analyzed the impact on the Out-Of-Vocabulary (OOV) words, since the model does not perform well on OOV. Dernoncourt et al. [2017] followed the same model architecture to train the NER. Similarly, Santos and Guimaraes [2015] obtained the word representation from characters by CNN and concatenated the embeddings of words before feeding to bidirectional LSTM. The Viterbi algorithm has also been used for inferencing for NER. Bharadwaj et al. [2016] has extended the model by the integration of phonemes information as an additional feature. Similarly, Yadav et al. [2018] integrated \(n\)-gram based most frequent affixes with the concatenated word representation for the same model for NER.

2.1 NER So Far on Indian Languages

One of the earliest works on NER for Indian languages was reported by Cucerzan and Yarowsky [1999], mainly for Hindi. The author used a bootstrapping algorithm and an iterative learning algorithm to classify the names (both first and last name) and places on the 18806 tokens and achieved 41.70% and 79.04% \(F_1\)-score and accuracy, respectively. The IJCNLP 2008 workshop on NER for South and South East Asian Languages\(^4\) was the first shared task for NER for Indian (or South Asian) languages and it also reported perhaps the first dataset (in the public domain) for NER for Indian languages: it included five Indian languages: Hindi, Urdu, Bengali, Oriya and Telugu [Singh 2008]. Of these, the Bengali dataset was developed by Jadavpur University and IIIT, Hyderabad, Urdu by CRULP, Lahore and IIIT, Allahabad and the rest was developed by workshop organising institute (IIIT, Hyderabad). These datasets were annotated with 12 NE tags. Ekal and Bandyopadhyay [2008] achieved 91.8% as the best \(F_1\)-score from SVM on the annotated Bengali news corpus of 467858 tokens by 16 entities.

Saha et al. [2008] worked on Hindi NER by using Maximum Entropy Model (MaxEnt), which assigns an outcome for each token based on its history and features. They used about 243K words for training purposes, which was taken from the Dainik Jagran\(^5\) (a popular Hindi newspaper), out of which about 16482 belonged to 4 named entities. Their MaxEnt based NER system (with beam search) was able to achieve a \(F_1\)-score of 81.52%, using a hybrid set of Gazetteer, patterns, and lexical and contextual features.

Morwal et al. [2012] worked on NER using Hidden Markov Model (HMM) for Hindi, Urdu and Punjabi. They used different number of tags for different corpora belonging to different domains. For example, they used Person,
Location, River and Country tags in the tourism corpus; and Person, Time, Month, Dry-fruits and Food items tags in the story corpus.

The Punjabi language is mainly written in two scripts: Gurmukhi (of Brahmi origin, like Devanagari) and Shahmukhi (a variant of the Persio-Arabic script). Most of the earlier works [Kaur et al. 2009; Kaur and Josan 2015; Morwal et al. 2012] on NER for this language were on data in Gurmukhi script and used statistical algorithms. Recent work on NER for Punjabi language using Shahamukhi script was explored by using 318275 tokens, out of which 16300 are entities such as Person, Location and Organization. The authors obtained 85.2% as best F1-score after applying a Recurrent Neural Network (RNN) over other classical and neural network techniques [Ahmad et al. 2020].

Sambalpuri is another low resource language with very little NLP work. It is an Indo-Aryan language spoken in parts of the Indian state of Odisha. Behera et al. (2017) worked on Sambalpuri and Odia NER using SVM [Behera and Muzaffar 2017]. They took 112K words for Sambalpuri and 250K words for Odia. They made 7 labels for annotating named entities. The F1-score measure obtained for Sambalpuri was 96.72% and 98.10% for Odia.

Lalitha Devi et al. (2008) worked on 94K words of Tourism domain for Tamil NER by using CRF. They used a total of 106 tags divided into three categories of ENAMEX, TIMEX and NUMEX. They obtained 80.44% F1-score [R and Lalitha Devi 2008]. C.S et al. [2012] also worked on Tamil NER using CRF model. They observed challenges in NER which occur due to several factors and are also applicable to other Indian languages as follows: agglutination, ambiguity, nested entities, spelling variations, name variations and the lack of capitalization.

The Assamese language has also been evaluated for the NER system [Sharma et al. 2010, 2014; Talukdar et al. 2014]. News domain data, ‘Asomiya Pratidin’ and ‘Emille corpus’ with 0.15 million tokens and 0.1 million tokens have been considered for CRF model training and testing respectively which has yielded 83% accuracy [Sharma et al. 2014]. Later, the output of CRF was post-processed by linguistic rules and obtained 93.22% as F1-score. Similarly, [Priyadarshini and Saha 2021] have developed the NER system for Maithili. A baseline classifier was prepared using CRF by manually annotating 200 thousand words of Maithili corpus. Later, they extended many experiments using various RNN. RNN models were found better than CRF. The authors investigated the effectiveness of gazetteer lists in RNN and found that incorporation of the gazetteer layer caused the improvement. Their final system achieved an F1-score of 91.6% with 94.9% precision and 88.53% recall.

Rao et al. [2015] conducted a shared task as part of the FIRE 2015 conference on Entity Extraction From Social Media using Named Entity Recognizer for Indian languages. They collected their corpus using the Twitter API in the period of May-June 2015 for training data and August-September for testing data of Tamil, Malayalam, Hindi and English languages. They used 22 tags for different kinds of names. Different participating teams used various machine learning methods (CRF, SVM, HMM, Naive Bayes, Decision Trees) with diverse sets of features. The baseline F1-score 47.10 for Hindi, 19.05 for Tamil 31.24 for Malayalam and 40.56 for English was reported.

Ali et al. [2020] have reported very recent work on the NER corpus for Sindhi, in which they annotated over 1.35 million tokens with eleven entity classes using corpus derived from Awami Awaz and Kavish newspapers and reported a best benchmark F1-score of 89.16% by using CNN-LSTM-CRF model on it.

3 DIFFICULTIES WITH BHOJPURI, MAITHILI, MAGAHI

Like many other modern Indo-Aryan languages, Bhojpuri, Maithili and Magahi are also non-tonal languages. They have Subject-Object-Verb (SOV) word order. Word-formation in these languages is somewhere between synthetic and analytical typology. There are a number of challenges while creating the corpus of these less-resourced languages. Despite being labelled as dialects of Hindi, morphological constructions of Bhojpuri, Maithili and
Magahi considerably differ from that of Hindi. Linguistic differences and lexical ambiguity of these languages create challenges for machine learning and are responsible for many problems in NER.

### 3.1 Morphologically Rich Languages

Also like many other Indian languages, Bhojpuri, Maithili and Magahi are morphologically rich, so the identification of the ‘root’ or lemma is challenging for these languages. They are partially synthetic languages. Hence, the use of embedded case markers, emphatic markers, classifiers, determiners etc. is frequent in these languages. These markers are responsible for problems and errors in NER. Some examples are: `mEWilIka` (Maithili’s), `Garasaz` (from home: -saz), `jAwika` (caste of: -ka), `gAmaka` (of village: -ka), `gAmakez` (of village: -kez), `rAmanAmIka` (name called Ram: -ika), `mircAI` (chillies: -Al) are from Maithili, `SakunwaloM` (Sankunla too/also: -oM), `majaXAre` (in dilemma: -e), `baniyavoM` (seller too: -oM), `GarahUz` (home as well: -Uz), `Kewavo` (field too: -vo), `surenxaravo` (surendra too: -vo), `ekke` (one too: -ke; an example of gemination), `sistAme` (in the system: -Ame), `sahebAina` (feminine of Sahib, Memsaib: -Aina) in Bhojpuri and , `rAwe` (in the night: -e), `xuariyA` (door: -iyA), `sonalo` (sonal too: -o), `saberahIM` (in morning: -hIM), `bABano` (Brahmin too: -o), `Gare` (home too: -e) in Magahi. As can be seen from some of these examples, names can also be inflected in the three languages, creating problems for the algorithm. Some other markers in these examples do not apply to names, but by their very frequent occurrence and by their appearance on too many words, they pose challenges for NER.

### 3.2 Ambiguity

Bhojpuri, Maithili and Magahi, like many non-standardised (or less-standardised) languages, also have lexical ambiguity in unusual abundance. This applies to many other Indian languages. These ambiguous words pose problems for NLP problems like POS tagging and NER, especially for machine learning, both while performing annotation and for modeling by machine learning. Similar to many other languages Bhojpuri, Maithili and Magahi also have a surprising amount of ambiguity even among proper names. According to our analysis, some of the examples are mentioned here in two categories: the ambiguity between the common and proper noun, and the ambiguity within the class of proper nouns.

1. **Ambiguity between Common and Proper Nouns**
   - Bhojpuri:
     - `cunarI` is “A type of cloth [Common Noun]” and “Name of a Person [Proper Noun]” as well.
     - `GAGarA` is “A type of cloth [Common Noun]” and “Name of a River [Proper Noun]” as well.
     - `kisalaya` is “Young shoot/Bud [Common Noun]” and “Name of a Person [Proper Noun]” as well.
   - Maithili:
     - `xaNdaka` is “of) Punishment [Common Noun]” and is “Name of a Forest (vana) [Proper Noun]” also.
     - `GUgarA` is “Anklet bells [Common Noun]” as well as "Name of a Person [Proper Noun]”.

2. **Ambiguity of a Proper Name**
   - People vs. Months:
     - `kAwika` and `sAvana` are “Name of a Person” and “Month of Indian Calendar” as well.
   - People vs. Locations:
     - `surEyA`, `bEjanAWa` and `kexArAnAwa` are “Name of a person” and “Name of a place” as well.
   - People vs. Seasons:
     - `basanwa`, `hemaMwa` and `SiSiRa` are “Name of a person” and “Name of a Season” as well.

---

7The example words are written in the WX notation

8Compared to other languages with very high numbers of speakers, as these latter languages tend to reduce lexical ambiguity in written language through means like standardization.
3.3 Spelling Variations
Like many other Indian (less-standardised) languages, Bhojpuri, Maithili and Magahi also have the problem of spelling variation. In Bhojpuri, Maithili and Magahi speech communities, different people spell the same words differently, partly because they are pronounced differently in different dialectal regions and the fact that writing systems are not standardised at all. Because of this, a number of spelling variations of a single word create confusion and problems for the NER task. For example, laikiya, laiki, laikiva, laikaniya, laikaniya are variations of Boy and surasawiyA, sarasawiyA, sarasawiyA are variations of person name. These are not the usual inflections as given in examples in section 3.1, but represent features like familiarity or informality or deprecative usage.

3.4 Other Challenges
Whether it is a matter of traditional resources such as grammar books, dictionaries, textbooks, magazines, newspapers etc. or modern resources such as websites, blogs, emails, chats, digitized information has increased in recent decades. Yet, compared to the requirements for successful NLP, especially data-driven NLP, the amount of linguistic resources (even simple text corpora) are still not available in sufficient quantities or with good enough quality for these languages due to their social status as dialects of Hindi. Even though Maithili has attained the status of language, it is still as resource scarce as Bhojpuri and Magahi. The lack of standardization and formal or official usage is one of the major problems for them, which poses difficulty in NER.

Apart from the scarcity of annotated resources, Bhojpuri, Maithili and Magahi also lack in terms of tools required for preprocessing such as Part-of-Speech tagger and Chunker that help in recognizing NEs are not available. Or tools which are available have relatively poor performance so far.

4 ANNOTATION
4.1 NER Guidelines
We have used Indian Language named entities tagset and annotation guidelines which were prepared under the Development of Cross Lingual Information Access (CLIA) system Phase-II Consortium Project funded by Ministry Communication and Information Technology (MCIT), Department of Information Technology, Government of India Version: 1.1. In this tagset, there are three main categories, viz. ENAMEX, NUMEX and TIMEX. There are a total of 22 tags in ENAMEX, NUMEX and TIMEX combined. In ENAMEX, there are eleven NER tags, viz. Person, Organization, Location, Facilities, Locomotives, Artifacts, Entertainment, Materials, Organisms, Plants and Disease. NUMEX consists of Distance, Money, Quantity and Count. TIMEX includes Time, Year, Month, Day, Date, Period and Special Day.

http://tdil-dc.in/index.php?option=com_download&task=showresourceDetails&toolid=815&lang=en
4.2 About Corpus
For creating annotated corpus for NER, we used the BMM corpus [Mundotiya et al. 2020] that was collected as part of the project on Bhojpuri, Maithili, Magahi to Hindi Machine Translation System under Project Varanasi. The main goal of this project was to develop MT systems from Bhojpuri, Maithili, Magahi to Hindi. We have used 16492, 9815 and 5320 sentences from the BMM corpus to create the annotated data for NER. These sentences have 228373, 157468 and 56190 tokens and 32091, 23338 and 10175 types for respective languages. After annotation, 12351 named entities in Bhojpuri, 19809 in Maithili and 7152 in Magahi were encountered, as mentioned in Table 1.

Table 1. Language-wise dataset statistics used for annotation of named entities

| Lang    | #Sentences | #Tokens   | #Types  | #Entities | #Others     |
|---------|------------|-----------|---------|-----------|-------------|
| Bhojpuri| 16492      | 228373    | 32091   | 12351     | 216022      |
| Maithili| 9815       | 157468    | 23338   | 19809     | 137659      |
| Magahi  | 5320       | 56190     | 10175   | 7152      | 49038       |

4.3 BMM NER Annotation
On the basis of the above mentioned guidelines, the corpora were tagged with hierarchical tags. For BMM, we mainly follow the guidelines used for Hindi for the purpose of NE annotation. The broad categories and their statistics for each language are outlined in Table 2. And each category with their statistics of hierarchical entities are summarised in Table 3.

Table 2. Frequencies of the three broad NE categories for Bhojpuri, Maithili and Magahi

| Lang    | ENAMEX | NUMEX | TIMEX |
|---------|--------|-------|-------|
| Bhojpuri| 10504  | 1152  | 695   |
| Maithili| 15861  | 2214  | 1734  |
| Magahi  | 5790   | 725   | 637   |

5 ALGORITHMS
For performing the baseline experiments on the prepared NER dataset of Bhojpuri, Maithili and Magahi, we have used two standard techniques which were the earlier state-of-the-art results for NER for other languages. These techniques are: CRF, a statistical algorithm and LSTM-CNNs-CRF [Ma and Hovy 2016b], based on a deep learning algorithm. From our study on related work of NER, a statistical algorithm (CRF) yielded comparative results to the Deep Learning method, perhaps because of the lack of sufficient data.

5.1 Conventional Machine Learning Algorithm
Conditional Random Field (CRF) is a discriminative model that uses conditional probability and is well suited for performing sequential prediction. Coined as CRF by Lafferty et al. [Lafferty et al. 2001; Rozenfeld et al. 2006], this undirected graphical model learns the dependency between each state and the entire input sequence. An input of each feature function has multiple input values in terms of the sentence, the word positions in the sentence, and...
Table 3. The frequencies of annotated hierarchical entities for Bhojpuri, Maithili and Magahi. The ENAMEX, NUMEX and TIMEX categories contained 11, 4 and 7 fine-grained named entities. ‘Other’ denotes regular words or tokens which are not named entities.

| Entities   | Bhojpuri | Maithili | Magahi |
|------------|----------|----------|--------|
| **ENAMEX** |          |          |        |
| Artifact   | 635      | 752      | 638    |
| Disease    | 34       | 9        | 18     |
| Entertainment | 347   | 532      | 31     |
| Facility   | 121      | 784      | 123    |
| Location   | 985      | 4330     | 763    |
| Locomotive | 112      | 157      | 58     |
| Material   | 278      | 481      | 379    |
| Organism   | 481      | 222      | 566    |
| Organization | 109   | 2081     | 20     |
| Person     | 7244     | 6462     | 3145   |
| Plant      | 158      | 51       | 49     |
| **NUMEX**  |          |          |        |
| Count      | 685      | 1797     | 558    |
| Distance   | 14       | 24       | 2      |
| Money      | 166      | 131      | 112    |
| Quantity   | 287      | 262      | 53     |
| **TIMEX**  |          |          |        |
| Date       | 69       | 48       | 5      |
| Day        | 36       | 99       | 169    |
| Month      | 66       | 281      | 52     |
| Period     | 279      | 491      | 28     |
| Special_Day | 8      | 210      | 1      |
| Time       | 175      | 413      | 337    |
| Year       | 62       | 192      | 45     |
| **OTHER**  |          |          |        |
| Other      | 216022   | 137659   | 49038  |

The labels of the current and the previous word that provide the contextual information. The feature function is expected to express some kind of characteristic of the sequence. A set of weights are assigned (initializing to random values) to the feature function to build the conditional field. Maximum Likelihood Estimation estimates the parameters. Finally, gradient descent updates the parameter values iteratively until the values converge.

5.2 Deep Learning Algorithm

LSTM-CNNs-CRF model [Ma and Hovy 2016b] consists of three components which are sequentially arranged. These components are: CNN, LSTM and CRF. They are responsible for capturing the character-level information, word-level information and the dependency information. The architecture is depicted in Figure 1.
The CNN Layer

The CNN model helps to extract character n-gram information of the given input words for our problem. It has three essential operations, which are Convolution, Pooling and Feed-forward. The convolution layer performs convolution operations in which filters are applied over the fixed-length sequential input, where each filter has a certain window size to produce a new feature.

Let an input word $W_1$ has $c_1, \ldots, c_m$ characters, encoded by a one-hot vector of $o_1, \ldots, o_m$, over which $k$ filters have been applied with $h$ window size so that the generated feature becomes:

$$C = [C_1, C_2, \ldots, C_{m-h+1}]$$

where, $C_i = \text{ReLU}[F_{o_{i+h-1}} + b]$

Here, $C$ is feature map produced by $k$ filters. The Pooling operation is performed by the max-pooling layer on the feature maps to extract the most relevant information to get a word vector.

$$\hat{C} = \max(C)$$

These obtained features $\{\hat{C}_1, \hat{C}_2, \hat{C}_3, \ldots, \hat{C}_k\}$ are fed to the fully connected layer for obtaining the desired vector size of a word from characters.

$$\text{Char}_\text{emb}_1 = W^T.\hat{C}_{1:k} + b$$

The input layer at the word-level encodes the word into a fixed-length real-valued vector ($e_1$), which is learnt during the model training. The final word representation is obtained after combining the actual word
representation and character-level word representation. These encoded-word vector and character-level word vector are concatenated as the final word embedding (WE).

\[ WE_{1:n} = \text{Char}_{emb_{1:n}} \oplus e_{1:n} \]

### The LSTM Layer

The final word embeddings are then passed to the LSTM layer for capturing the longer dependencies among the words of the input sentence. BMM languages tend to have long-distance dependencies. Here, the bidirectional LSTM layer has been used for modelling the longer dependencies by applying LSTM on the forward and the backward directions. The obtained hidden states from both directions are concatenated and are fed to the next layer:

\[
\begin{align*}
\overrightarrow{h_{1:n}} &= \text{LSTM}(WE_{1:n}) \\
\overleftarrow{h_{1:n}} &= \text{LSTM}(WE_{1:n}) \\
 h &= \overrightarrow{h_{1:n}} \oplus \overleftarrow{h_{1:n}}
\end{align*}
\]

### The CRF Layer

After capturing dependencies among the words by the LSTM layer, a CRF layer is used to capture dependencies of the labels. The CRF layer is applied over the representations obtained from the bidirectional LSTM output for obtaining the label dependencies of the input sentences. Linear CRF models the linear relationships with the previous labels to generate the probability score of the current labels by the potential function \( \psi(.) \):

\[
\psi(h_i, y_i, y_{i-1}) = \exp(y_i^T W_1^T h_i + y_{i-1}^T W_2 y_i)
\]

\[
\pi(y|h; \theta) = \frac{\prod_{i=1}^{n} \psi(h_i, y_i, y_{i-1})}{\sum_{y'} \prod_{i=1}^{n} \psi(h_i, y'_i, y'_{i-1})}
\]

Here, \( Y \) denotes all possible labels and \( \theta \) is the learning parameter.

### 6 EXPERIMENTS

The dataset was divided into training and testing splits with a ratio of 80-20. While splitting, it was ensured that the testing dataset included all the named entities with frequency of at least one. The dataset statistics after splitting are shown in Table 4. The training strategy and the obtained results have been explained in the following sections.

#### 6.1 Training the CRF Model

There are several implementations of CRF that are publicly available. We have used the CRFsuit\(^{10}\) implementation with the training algorithm of L-BFGS, executed up to a maximum of 100 iterations. To avoid overfitting and underfitting issues of CRF, C1 and C2 regularization parameters with random search cross-validation have been used for training, where the value of cross-validation is 3, and the number iterations are 50. The current word, the neighbouring words with the adjacency of 2, affixes of the current word with a window size of 3, whether the current word and the neighbouring words are digits and whether the current word is first, or the last word of a sentence are considered as hand-crafted features for training the CRF model. The optimal values of C1 and C2

\(^{10}\)https://sklearn-crfsuite.readthedocs.io/en/latest/
Table 4. The dataset sizes for each language. The OOV percentage is calculated by token-type differences between test data and the training data.

| Language | Data-Mode | Sentences | Tokens | Types | OOV (%) |
|----------|-----------|-----------|--------|-------|---------|
| Bhojpuri | Train     | 13193     | 160226 | 19642 | 25.50   |
|          | Test      | 3299      | 68147  | 12449 |         |
| Maithili | Train     | 7849      | 125442 | 15859 | 27.94   |
|          | Test      | 1966      | 32026  | 7479  |         |
| Magahi   | Train     | 4256      | 44833  | 6868  | 22.34   |
|          | Test      | 1065      | 11357  | 3307  |         |

are 0.178 and 0.006 for Magahi, 0.440 and 0.018 for Maithili and 0.481 and 0.003 for Bhojpuri as obtained after training.

6.2 Training the LSTM-CNNs-CRF Model

The LSTM-CNNs-CRF model takes input in the form of characters and words to generate word embeddings to overcome the scarcity of annotated data. As Deep Learning models are very sensitive to the data size as well as the values of the parameters, it is not guaranteed that the same value of the parameter will provide optimal results for another language. The word embeddings, character embeddings and the hidden representations play a vital role in obtaining the best possible results. For our experiments, the sizes of the word embedding, the character embedding and the hidden representations are 100, 20 and 25, respectively for Bhojpuri. Similarly, the values are 100, 30 and 50 for Magahi and 200, 20 and 50 for Maithili. The number of convolutional layers is 4. The model training is performed with the Stochastic Gradient Descent (SGD) optimizer, where the learning rate is 0.015, which decays over the epoch by 0.05 for constraint learning. During training, the L2 regularizer and dropout are also used with the values of 0.5 and 1e−8, respectively to prevent model overfitting.

7 RESULT AND ANALYSIS

By considering the parameters defined in the previous section, the results obtained on CRF and LSTM-CNNs-CRF are mentioned in Tables 5 and 6, respectively. Bhojpuri has a large amount of annotated data than the remaining two languages. The obtained results on the test data for CRF and LSTM-CNNs-CRF are 70.56% and 61.41% (F1-score), respectively. The entity-wise score given in those tables shows that LSTM-CNNs-CRF struggles to learn for low-frequency entities such as Day, Disease, Distance, Organization, Special_Day and Year. However, CRF is suffered for the Special_Day only. In the CRF training, the best transition is obtained from I-Money→I-Money, B-Organization→I-Organization, B-Period→I-Period, B-Facility→I-Facility and B-Year→I-Year. Similarly, the worst transition represents the wrong interpretation as B-Count→I-Person, B-Count→B-Count, I-Person→B-Artifact and B-Artifact→B-Artifact.

For Maithili, the obtained F1-score is 73.19%, 71.38% for CRF and LSTM-CNNs-CRF respectively. Disease and Distance are two NEs that were wrongly predicted by the LSTM-CNNs-CRF model, whereas Disease prediction was difficult for the CRF model due to less frequency. Some of the remaining entities (Day, Date, Month, Year, Distance, Organism and Plant) have rare intermediates. The optimal transitions from this language’s annotated dataset are B-Entertainment→I-Entertainment, B-Facility→I-Facility and B-Organism→I-Organism, and worst transitions are B-Period→B-Count, B-Location→I-Person and B-Location→B-Period for the CRF model.

For Magahi, the obtained F1-score of 84.18% and 86.39% for CRF and LSTM-CNNs-CRF respectively. As, Magahi has smaller annotated data as well as a high number of low-frequency entities such as Date, Distance,
Table 5. NER tag-wise scores obtained by CRF for Bhojpuri, Maithili and Magahi. The metrics, which are Precision, Recall and F1-score.

| NER-Tag   | Bhojpuri P  | Bhojpuri R | Bhojpuri F1 | Maithili P  | Maithili R | Maithili F1 | Magahi P   | Magahi R | Magahi F1 |
|-----------|-------------|------------|-------------|-------------|------------|-------------|------------|----------|----------|
| Artifact  | 96.77       | 34.88      | 51.28       | 77.78       | 31.46      | 44.80       | 100.00     | 54.44    | 70.50    |
| Disease   | 100         | 77.78      | 87.50       | 00.00       | 00.00      | 00.00       | 00.00      | 00.00    | 00.00    |
| Entertainment | 84.00 | 20.79      | 33.33       | 87.80       | 62.07      | 72.73       | 100.00     | 40.00    | 57.14    |
| Facility  | 36.00       | 23.08      | 28.12       | 80.77       | 47.19      | 59.57       | 100.00     | 72.22    | 83.87    |
| Location  | 95.36       | 54.14      | 69.06       | 91.96       | 68.90      | 78.78       | 98.33      | 67.82    | 80.27    |
| Locomotive| 73.91       | 42.50      | 53.97       | 93.75       | 45.45      | 61.22       | 75.00      | 30.00    | 42.86    |
| Material  | 91.67       | 23.16      | 28.12       | 80.77       | 47.19      | 59.57       | 100.00     | 72.22    | 83.87    |
| Organism  | 100.00      | 18.46      | 31.17       | 66.67       | 05.41      | 10.00       | 100.00     | 75.81    | 86.24    |
| Organization | 100.00 | 18.18      | 30.77       | 93.64       | 55.86      | 69.98       | 100.00     | 50.00    | 66.67    |
| Person    | 98.97       | 79.01      | 87.87       | 97.11       | 75.62      | 85.03       | 97.59      | 84.98    | 90.85    |
| Plant     | 84.62       | 56.41      | 67.69       | 100.00      | 36.36      | 53.33       | 100.00     | 62.50    | 76.92    |
| Count     | 71.62       | 30.81      | 43.09       | 87.69       | 69.80      | 77.73       | 96.67      | 75.32    | 84.67    |
| Distance  | 50.00       | 20.00      | 28.57       | 50.00       | 33.33      | 40.00       | 00.00      | 00.00    | 00.00    |
| Money     | 88.46       | 38.33      | 53.49       | 62.50       | 38.46      | 47.62       | 100.00     | 95.00    | 97.44    |
| Quantity  | 48.78       | 28.17      | 35.71       | 83.33       | 07.94      | 14.49       | 100.00     | 54.55    | 70.59    |
| Date      | 100.00      | 23.53      | 33.18       | 100.00      | 60.00      | 75.00       | 100.00     | 66.67    | 80.00    |
| Day       | 100.00      | 16.67      | 28.57       | 85.71       | 30.00      | 44.44       | 94.44      | 89.47    | 91.89    |
| Month     | 100.00      | 27.78      | 43.48       | 97.44       | 88.73      | 92.68       | 100.00     | 72.73    | 84.21    |
| Period    | 86.67       | 17.11      | 28.57       | 90.32       | 75.68      | 82.35       | 100.00     | 33.33    | 50.00    |
| Special_Day | 100.00 | 00.00      | 00.00       | 95.00       | 65.52      | 77.55       | 00.00      | 00.00    | 00.00    |
| Time      | 100.00      | 26.79      | 42.25       | 88.89       | 26.23      | 40.51       | 100.00     | 70.27    | 82.54    |
| Year      | 57.14       | 40.00      | 47.06       | 100.00      | 59.26      | 74.42       | 100.00     | 62.50    | 76.92    |
| Avg. score| 93.64       | 59.90      | 70.56       | 91.53       | 62.86      | 73.19       | 97.67      | 75.00    | 84.18    |
| OTHER     | 97.94       | 99.08      | 98.51       | 95.52       | 98.33      | 96.90       | 95.11      | 95.29    | 95.04    |
| Avg. score (including OTHER) | 96.59 | 96.99 | 96.73       | 93.32 | 93.87 | 93.33       | 95.11 | 95.29 | 95.04 |

Special_Day\footnote{The Special_Day entity has 1 frequency; hence it has not appeared in the validation.}, etc. Moreover, the LSTM-CNNs-CRF technique suffers from learning those entities. Even though, few entities such as Disease, Entertainment, Organisation, Plant, Year and Distance have metric scores around 0 to intermediate tag for the CRF. Most of the entities are not longer than a single token; hence intermediates of these tags are rare. During the evaluation of the model, we found that the optimal transitions between B-Quantity→I-Quantity, B-Plant→I-Plant and I-Quantity→I-Material, where B-Person→B-Facility, B-Day→B-Person and OTHER→I-Quantity are worst transitions which indicate wrong annotations.

LSTM-CNNs-CRF fails to outperform the CRF baseline in the case of Bhojpuri and Maithili, which have more data in terms of the number of tokens, but not in terms of the number of named entities.

7.1 Experiments on Cross-lingual Training

For performing the cross-lingual model training, we have followed Bari et al. [2020] procedure to the LSTM-CNNs-CRF model. They have experimented with variants of the proposed model based on the cross-lingual
Table 6. NER tag-wise scores obtained by LSTM-CNNs-CRF for Bhojpuri, Maithili and Magahi. The metrics, which are Precision, Recall and $F_1$-score.

| NER-Tag    | Bhojpuri |          |          | Maithili |          |          | Magahi |          |          |
|------------|----------|----------|----------|----------|----------|----------|--------|----------|----------|
|            | P        | R        | F$_1$    | P        | R        | F$_1$    | P      | R        | F$_1$    |
| Artifact   | 84.91    | 26.16    | 40.00    | 83.33    | 28.09    | 42.02    | 100.00 | 58.89    | 74.13    |
| Disease    | 00.00    | 00.00    | 00.00    | 00.00    | 00.00    | 00.00    | 00.00  | 00.00    | 00.00    |
| Entertainment | 77.78    | 06.93    | 12.73    | 70.18    | 68.97    | 69.57    | 100.00 | 40.00    | 57.14    |
| Facility   | 25.00    | 10.26    | 14.55    | 81.25    | 43.82    | 56.93    | 100.00 | 72.22    | 83.87    |
| Location   | 97.20    | 39.10    | 55.76    | 93.88    | 66.72    | 78.01    | 96.83  | 70.11    | 81.33    |
| Locomotive | 65.00    | 32.50    | 43.33    | 100.00   | 15.15    | 26.32    | 80.00  | 40.00    | 53.33    |
| Material   | 77.78    | 07.37    | 13.46    | 88.57    | 39.74    | 54.87    | 94.29  | 91.67    | 92.96    |
| Organism   | 100.00   | 07.69    | 14.29    | 100.00   | 05.41    | 10.26    | 97.96  | 77.42    | 86.49    |
| Organization | 00.00    | 00.00    | 00.00    | 92.31    | 57.93    | 71.19    | 100.00 | 25.00    | 40.00    |
| Person     | 98.88    | 72.45    | 83.62    | 97.08    | 70.89    | 81.94    | 100.00 | 88.29    | 93.78    |
| Plant      | 90.91    | 25.64    | 40.00    | 100.00   | 45.45    | 62.50    | 100.00 | 62.50    | 76.92    |
| Count      | 71.43    | 23.26    | 35.09    | 83.57    | 72.65    | 77.73    | 95.65  | 85.71    | 90.41    |
| Distance   | 00.00    | 00.00    | 00.00    | 00.00    | 00.00    | 00.00    | 00.00  | 00.00    | 00.00    |
| Money      | 100.00   | 06.67    | 12.50    | 66.67    | 15.38    | 25.00    | 100.00 | 95.00    | 97.44    |
| Quantity   | 55.56    | 14.08    | 22.47    | 77.78    | 11.11    | 19.44    | 100.00 | 54.55    | 70.59    |
| Date       | 100.00   | 05.88    | 11.11    | 100.00   | 20.00    | 33.33    | 00.00  | 00.00    | 00.00    |
| Day        | 00.00    | 00.00    | 00.00    | 69.23    | 45.00    | 54.55    | 94.44  | 89.47    | 91.89    |
| Month      | 00.00    | 00.00    | 00.00    | 100.00   | 88.37    | 93.83    | 100.00 | 72.73    | 84.21    |
| Period     | 100.00   | 14.47    | 25.29    | 92.16    | 63.51    | 75.20    | 100.00 | 33.33    | 50.00    |
| Special_Day | 00.00    | 00.00    | 00.00    | 100.00   | 62.07    | 76.60    | 00.00  | 00.00    | 00.00    |
| Time       | 100.00   | 28.57    | 44.44    | 83.33    | 24.59    | 37.97    | 100.00 | 72.97    | 84.37    |
| Year       | 00.00    | 00.00    | 00.00    | 100.00   | 66.67    | 80.00    | 100.00 | 75.00    | 85.71    |
| Avg. score | 91.14    | 50.51    | 61.41    | 91.60    | 60.65    | 71.38    | 97.82  | 78.42    | 86.39    |
| OTHER      | 97.39    | 99.46    | 98.41    | 95.27    | 98.61    | 96.91    | 96.92  | 98.28    | 97.60    |
| Avg. score (including OTHER) | 96.15 | 96.84 | 96.25 | 93.34 | 93.85 | 93.33 | 95.44 | 95.62 | 95.44 |

Transfer. Among them, the Cross-Word has been considered for our experiments. In this, source and target word embedding are mapped and projected into a common space with the adversarial training. The described monolingual corpus of these languages in Mundotiya et al. [2020] have been exploited to generate the fast text embedding with its default settings, which is further used for cross-lingual information learning at the word level. We have considered Bhojpuri as the source language for Maithili and Magahi. We considered Maithili as a source language for Bhojpuri only.

The obtained results through cross-lingual training are mentioned in Table 7. The mean $F_1$-score for Bhojpuri and Maithili NEs are 78.95% and 77.10%, showing an improvement of 17.51% and 5.72% compared to the deep learning baseline model. Although, Magahi’s performance has been declined by 0.69%. This indicates that the cross-lingual result depends on the size of the annotated data of the target language, as well as the size of annotated data of the source language. Cross-lingual training has also affected the recall value of the less frequent tags.
Table 7. NER tag-wise scores obtained by cross-lingual LSTM-CNNs-CRF for Bhojpuri, Maithili and Magahi. The metrics, which are Precision, Recall and $F_1$-score.

| NER-Tag   | Bhojpuri | | | | Maithili | | | | | Magahi | | |
|-----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|
|           | P        | R        | $F_1$    | P        | R        | $F_1$    | P        | R        | $F_1$    | P        | R        | $F_1$    |
| Artifact  | 52.54    | 50.82    | 51.67    | 63.79    | 58.73    | 61.16    | 81.25    | 54.17    | 65.00    | 0.00    | 0.00    | 0.00    |
| Disease   | 100.00   | 50.00    | 66.67    | 0.00     | 0.00     | 0.00     | 100.00   | 50.00    | 66.67    | 100.00   | 50.00    | 66.67    |
| Entertainment | 60.53   | 44.23    | 51.11    | 59.26    | 34.78    | 43.84    | 100.00   | 50.00    | 66.67    | 100.00   | 50.00    | 66.67    |
| Facility  | 08.00    | 40.00    | 13.33    | 49.43    | 72.88    | 58.90    | 81.82    | 75.00    | 78.26    | 76.92    | 79.37    | 78.13    |
| Location  | 100.00   | 33.33    | 50.00    | 56.52    | 100.00   | 72.22    | 100.00   | 50.00    | 66.67    | 92.11    | 70.00    | 79.55    |
| Locomotive | 44.44   | 53.33    | 48.48    | 75.00    | 54.55    | 63.16    | 0.00     | 0.00     | 0.00     | 0.00     | 0.00     | 0.00     |
| Material  | 24.56    | 31.82    | 27.72    | 34.78    | 32.00    | 33.33    | 81.97    | 80.65    | 81.30    | 81.97    | 80.65    | 81.30    |
| Organisation | 33.33   | 16.67    | 22.22    | 80.67    | 58.54    | 67.84    | 0.00     | 0.00     | 0.00     | 0.00     | 0.00     | 0.00     |
| Person    | 87.25    | 83.18    | 85.17    | 83.95    | 76.06    | 79.81    | 90.97    | 80.00    | 85.14    | 90.97    | 80.00    | 85.14    |
| Plant     | 61.54    | 72.73    | 66.67    | 83.33    | 83.33    | 83.33    | 100.00   | 40.00    | 57.14    | 100.00   | 40.00    | 57.14    |
| Count     | 29.11    | 45.10    | 35.38    | 73.53    | 71.02    | 72.25    | 92.31    | 85.71    | 88.89    | 92.31    | 85.71    | 88.89    |
| Distance  | 00.00    | 00.00    | 00.00    | 100.00   | 16.67    | 28.57    | 0.00     | 0.00     | 0.00     | 0.00     | 0.00     | 0.00     |
| Money     | 70.00    | 35.00    | 46.67    | 41.18    | 70.00    | 51.85    | 90.91    | 83.33    | 86.96    | 90.91    | 83.33    | 86.96    |
| Quantity  | 40.00    | 29.63    | 34.04    | 11.76    | 06.67    | 08.51    | 0.00     | 0.00     | 0.00     | 0.00     | 0.00     | 0.00     |
| Date      | 00.00    | 00.00    | 00.00    | 50.00    | 25.00    | 33.33    | 0.00     | 0.00     | 0.00     | 0.00     | 0.00     | 0.00     |
| Day       | 00.00    | 00.00    | 00.00    | 33.33    | 21.43    | 26.09    | 92.31    | 75.00    | 82.76    | 92.31    | 75.00    | 82.76    |
| Month     | 00.00    | 00.00    | 00.00    | 87.50    | 84.00    | 85.71    | 100.00   | 100.00   | 100.00   | 100.00   | 100.00   | 100.00   |
| Period    | 50.00    | 29.63    | 37.21    | 79.07    | 55.74    | 65.38    | 100.00   | 66.67    | 80.00    | 100.00   | 66.67    | 80.00    |
| Special_day | 00.00  | 00.00    | 00.00    | 80.00    | 52.17    | 63.16    | 0.00     | 0.00     | 0.00     | 0.00     | 0.00     | 0.00     |
| Time      | 53.33    | 72.73    | 61.54    | 56.41    | 43.14    | 48.89    | 80.65    | 80.65    | 80.65    | 80.65    | 80.65    | 80.65    |
| Year      | 25.00    | 20.00    | 22.22    | 100.00   | 58.33    | 73.68    | 100.00   | 75.00    | 85.71    | 100.00   | 75.00    | 85.71    |
| Avg. score | 94.80    | 69.60    | 78.95    | 90.58    | 68.20    | 77.10    | 99.19    | 75.97    | 85.70    | 99.19    | 75.97    | 85.70    |
| OTHER     | 98.48    | 98.68    | 98.58    | 96.72    | 98.05    | 97.38    | 96.62    | 98.48    | 97.54    | 96.62    | 98.48    | 97.54    |
| Avg. score (including OTHER) | 97.19    | 97.10    | 97.10    | 94.34    | 94.56    | 94.35    | 95.46    | 95.64    | 95.44    | 95.46    | 95.64    | 95.44    |

7.2 Error Analysis
The cross-lingual experiment using LSTM-CNNs-CRF provides best results compare to other models. Hence, the top-10 most wrong NE prediction along with their frequency for all three languages have been illustrated in Figure 2. The OTHER is the most frequent tag in the annotated datasets so, the prediction of OTHER from any NE have not considered for analysis except Magahi due to limited number of error patterns found for it. As shown in Figure 2a the Bhojpuri model confused with the NE prediction for numbers such as Count is predicted as Money, Quantity, Year and vice-versa also. Apart from Count, Artifact, Facility and Locomotive are the NE tags where model fails, for example, Ayodeksa was annotated by Material but model predict it as Artifact and trEktrara was tagged as Locomotive but predict as Facility. Similarly, according to Figure 2b, Maithili suffers with recognition of Location, Organization and Facility mostly for example, nAtya vixyAlaya was tagged as Facility, however the model predicts Organization. Apart from this, Count, Quantity and Distance were also ambiguous for recognition. Since the Maithili is complex language then the rest two languages hence it has more errors then remaining two. All the highlighted top-10 frequent errors for Magahi in Figure 2c are having most frequent NE
tags in the annotated dataset (as seen in Table 3). musahari is an example of case where it is tagged as Location but predicted as Artifact.

However, some NEs were annotated wrong in annotated dataset that also corrected by the model prediction such as kirAsana, xahl-cUdzA were tagged as OTHER but model predicts those as Material in Bhojpuri and Maithili, respectively. Similarly, the word xasa correctly predicted as Count but was tagged as OTHER in Magahi. The incorrect prediction due to the above described BMM difficulties (in section 3) are also discussed with the few examples with respect to each language such as:

- In Bhojpuri, xuKa proper noun is ambiguous with common noun. Hence, It is predicted as OTHER. ego is a COUNT but it is predicted as OTHER due to morphological richness. xu hajAra cAra is YEAR but it is predicted as COUNT due to morphological ambiguity. grlna cillI mIdiyA is an ORGANIZATION in a chunk but it is predicted as OTHER by considering it separately but it leads to a syntactic ambiguity. javAharo is a PERSON but predicted it as OTHER due to spelling variation. LAKana is PERSON but it is predicted as FACILITY due morphological richness. kalaCule-kalaCule is OTHER (adverb) but predicted as MATERIAL due to the succeeding verb parosanA.

- In Maithili, yAwrl is proper noun but tagged as OTHER because it is also a common noun. Hence it is considered as ambiguity. vINA is both proper noun and AERTIFACT hence it is considered as ambiguous word. SarmA is PERSON due to being a sir name but also as a verb. Hence ambiguity is seen. ahamaxaka is PERSON but tagged as OTHER due to morphological richness of language. Since it is a synthetic language with suffix attached with the name. kawaya is not a place but tagged as LOCATION due to the succeeding word being a verb which appears as a syntactic ambiguity. takA is quantity in chunk but tagged as OTHER. Since it is considered separately as OTHER which is because of the morphological richness of the language. arihAnA is LOCATION but tagged as OTHER due to ambiguity with the verb.

- In Magahi, homlpawWI is an ARTIFACT. However, it is predicted as OTHER due to spelling variation. Pulesariye is Person but predicted it as OTHER due to spelling variation. xuie is a COUNT but it is predicted as OTHER due to spelling variation. babuA is a PERSON, but it is also considered as a common name which creates ambiguity. sAdZI-sAyA is ARTIFACT but predicted as OTHER which is due to morphological richness. bihAra-sarakAra is an ORGANIZATION when it is used in a chunk and LOCATION and other respectively when considered separately so here tagged as OTHER.

ACM Trans. Asian Low-Resour. Lang. Inf. Process.
8 CONCLUSION

Bhojpuri, Maithili and Magahi are Purvanchal languages which are often considered dialects of Hindi, even though they are widely spoken in parts of India. Bhojpuri is spoken even outside India. Partly due to their dialectal nature, they show more linguistic variations such as nominal case inflection, emphatic expressions. Like other computational resources, there is a lack of any NER system for these languages. We describe a first attempt at this. This attempt includes the creation of a dataset as well as reporting the results for two baseline systems, one that uses CRF and the other that uses an LSTM-CNNs-CRF model. These NER systems are planned to be used in machine translation system for Bhojpuri, Maithili and Magahi to Hindi. The NER dataset, prepared by native speaker linguists, consists of 228373, 157468 and 56190 tokens, out of which 12351, 19809 and 7152 are NE’s. The tagset used is a union of ENAMEX, TIMEX and NUMEX tagsets, having a total of 22 labels. The results obtained (in terms of $F_1$-score) are 70.56% for 61.41% for Bhojpuri with CRF and LSTM-CNNs-CRF, respectively. The results for Maithili are 73.19% and 71.38% and for Magahi, they are 84.18% and 86.39% for the two models. Even though the total data size is more for Bhojpuri, the scores are lower as the number of NEs in the dataset of this language is relatively much less than for the other two languages. In other words, the results are consistent with the number of NEs in the datasets, rather than with the total size of the dataset in terms of the number of tokens. This claim was also supported by cross-lingual experimentation.

ACKNOWLEDGMENTS

We would like to thank our NER annotators Ajeet Kumar and Umesh Kumar Chaudhary for Bhojpuri, Nihal Kumar for Maithili, Rishav Raj Singh, Supriya Chauhan and Manu Mashani for Magahi. All the annotators are native speakers of the respective languages and did their Masters in Linguistics from Banaras Hindu University, Varanasi.

REFERENCES

Muhammad Tayyab Ahmad, Muhammad Kamran Malik, Khurram Shahzad, Faisal Aslam, Asif Iqbal, Zubahar Nawaz, and Faisal Bukhari. 2020. Named Entity Recognition and Classification for Punjabi Shahmukhi. ACM Trans. Asian Low-Resour. Lang. Inf. Process. 19, 4, Article 51 (April 2020), 13 pages. https://doi.org/10.1145/3383306

Enrique Alfonseca and Suresh Manandhar. 2002. An unsupervised method for general named entity recognition and automated concept discovery. In Proceedings of the 1st international conference on general WordNet, Mysore, India. 34–43.

Wazir Ali, Junyu Lu, and Zenglin Xu. 2020. SiNER: A Large Dataset for Sindhi Named Entity Recognition. In Proceedings of The 12th Language Resources and Evaluation Conference. 2953–2961.

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. Journal of Machine Learning Research 6, Nov (2005), 1817–1853.

Bogdan Babych and Anthony Hartley. 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.

M Saiful Bari, Shafiq Joty, and Prathyusha Jwalapuram. 2020. Zero-resource cross-lingual named entity recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34. 7415–7423.

Pitambar Behera and Sharmin Muzaffar. 2017. 1.1 Approaches to named entity recognition. (2017).

Deepthi Bhalla, Nidhieh Joshi, and Iti Mathur. 2013. Improving the quality of MT output using novel name entity translation scheme. In 2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI). IEEE, 1548–1553.

Akash Bharadwaj, David R Mortensen, Chris Dyer, and Jaime G Carbonell. 2016. Phonologically aware neural model for named entity recognition in low resource transfer settings. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. 1462–1472.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In Proceedings of the 25th international conference on Machine learning. 160–167.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. Journal of machine learning research 12, Aug (2011), 2493–2537.
Malarikodi C.S., Pattabhi RK Rao, and Sobha Lalitha Devi. 2012. Tamil NER - Coping with Real Time Challenges. In Proceedings of the Workshop on Machine Translation and Parsing in Indian Languages. The COLING 2012 Organizing Committee, Mumbai, India, 23–38. https://www.aclweb.org/anthology/W12-5603

Alessandro Cucchiarelli and Paola Velardi. 2001. Unsupervised named entity recognition using syntactic and semantic contextual evidence. Computational Linguistics 27, 1 (2001), 123–131.

Silvio Cucerzan and David Yarowsky. 1999. Language independent named entity recognition combining morphological and contextual evidence. In 1999 joint SIGDAT conference on empirical methods in natural language processing and very large corpora.

Franck Dernoncourt, Ji Young Lee, and Peter Szolovits. 2017. NeurOnER: an easy-to-use program for named-entity recognition based on neural networks. arXiv preprint arXiv:1705.05487 (2017).

Asif Ekbal and Sivaji Bandyopadhyay. 2008. Development of bengali named entity tagged corpus and its use in ner systems. In Proceedings of the 6th Workshop on Asian Language Resources.

Oren Etzioni, Michael Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S Weld, and Alexander Yates. 2005. Unsupervised named-entity extraction from the web: An experimental study. Artificial intelligence 165, 1 (2005), 91–134.

Richard Farkas and György Szarvas. 2006. Statistical Named Entity Recognition for Hungarian--analysis of the impact of feature space characteristics. (2006).

Ralph Grishman and Beth M Sundheim. 1996. Message understanding conference-6: A brief history. In COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv:1508.01991 (2015).

Ananddeep Kaur, G Josan, and Jagroop Kaur. 2009. Named entity recognition for punjabi: A conditional random field approach. In Proceedings of 7th international conference on Natural Language ProcessingICON-09. Macmillan Publishers, India.

Ananddeep Kaur and Gurpreet Singh Josan. 2015. Evaluation of Named Entity Features for Punjabi Language. Procedia Computer Science 46 (2015), 159 – 166. https://doi.org/10.1016/j.procs.2015.02.007 Proceedings of the International Conference on Information and Communication Technologies, ICTT 2014, 3-5 December 2014 at Bolgatty Palace & Island Resort, Kochi, India.

Yoon Kim, Yavincen Jerne, David Sontag, and Alexander M Rush. 2016. Character-aware neural language models. In Thirtieth AAAI Conference on Artificial Intelligence.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. (2001).

Xuezhe Ma and Eduard Hovy. 2016a. End-to-end sequence labeling via bi-directional lstm-cnns-crf. arXiv preprint arXiv:1603.01354 (2016).

Xuezhe Ma and Eduard Hovy. 2016b. End-to-End Sequence Labeling via Bi-Directional LSTM-CNNS-CRF. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Berlin, Germany, 1064–1074. https://doi.org/10.18653/v1/P16-1101

George A. Miller. 1995. WordNet: A Lexical Database for English. Commun. ACM 38, 11 (Nov. 1995), 39–41. https://doi.org/10.1145/219717.219748

Sudha Morwal, Nusrat Jahan, and Deeptri Chopra. 2012. Named entity recognition using hidden Markov model (HMM). International Journal on Natural Language Computing (IJNLC) 1, 4 (2012), 15–23.

Rajesh Kumar Mundotiya, Manish Kumar Singh, Rahul Kapur, Swasti Mishra, and Anil Kumar Singh. 2020. Basic Linguistic Resources and Baselines for Bhojpuri, Magahi and Maithili for Natural Language Processing. arXiv preprint arXiv:2004.13945 (2020).

Muntas Padró and Lluis Padró. 2005. A named entity recognition system based on a finite automata language acquisition algorithm. Procesamiento del Lenguaje Natural 35 (2005), 319–328.

Marius Pasca, Dekang Lin, Jeffrey Bigham, Andrei Lifchits, and Alpa Jain. 2006. Organizing and searching the world wide web of facts-step one: the one-million fact extraction challenge. In AAAI, Vol. 6. 1400–1405.

Ankur Priyadarshi and Sujuan Kumar Saha. 2021. The first named entity recognizer in Maithili: Resource creation and system development. Journal of Intelligent & Fuzzy Systems Preprint (2021), 1–13.

Vijayakrishna R and Sobha Lalitha Devi. 2008. Domain Focused Named Entity Recognizer for Tamil Using Conditional Random Fields. In Proceedings of the IJCNLP-08 Workshop on Named Entity Recognition for South and South East Asian Languages. https://www.aclweb.org/anthology/I08-5009

Pattabhi RK Rao, CS Malarkodi, R Vijay Sundar Ram, and Sobha Lalitha Devi. 2015. ESM-IL: Entity Extraction from Social Media Text for Indian Languages@ FIRE 2015-An Overview.. In FIRE workshops. 74–80.

Lisa F Rau. 1991. Extracting company names from text. In [1991] Proceedings. The Seventh IEEE Conference on Artificial Intelligence Application, Vol. 1. IEEE, 29–32.

Ellen Riloff, Rosie Jones, et al. 1999. Learning dictionaries for information extraction by multi-level bootstrapping. In AAAI/IJAI. 474–479.

Binyamin Rosenfeld, Ronen Feldman, and Moshe Fresko. 2006. A systematic cross-comparison of sequence classifiers. In Proceedings of the 2006 SIAM International Conference on Data Mining, SIAM, 564–568.

Sujan Kumar Saha, Sudeshna Sarkar, and Pabitra Mitra. 2008. A Hybrid Feature Set based Maximum Entropy Hindi Named Entity Recognition. In Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I. https://www.aclweb.org/anthology/I08-
Fig. 3. Confusion matrix for Bhojpuri for the LSTM-CNNs-CRF (above) and CRF (below) models; The \( \times \) refers to correctly prediction

A APPENDIX

The following tables give the confusion matrices for the reported prediction experiments. While plotting a confusion matrix, we have ignored NEs that: (i) have both actual and predicted counts as zeros, (ii) or are all predicted correctly. For example, Year, Special_Day, Distance and Disease in Bhojpuri, and Distance and Special_Day in Magahi belong to the case (i), whereas Time and Period in Bhojpuri, and Month, Time, Year, Person, Facility, Organization and Entertainment in Magahi, and Month in Maithili are all predicted correctly by the LSTM-CNNs-CRF model.

1045

Cicero Nogueira dos Santos and Victor Guimaraes. 2015. Boosting named entity recognition with neural character embeddings. arXiv preprint arXiv:1505.05008 (2015).

Padmaja Sharma, Utpal Sharma, and Jugal Kalita. 2010. The first steps towards Assamese named entity recognition. In Brisbane Convention Center, Vol. 1. 1–11.

Padmaja Sharma, Utpal Sharma, and Jugal Kalita. 2014. Named entity recognition in Assamese using CRFS and rules. In 2014 International Conference on Asian Language Processing (IALP). IEEE, 15–18.

Anil Kumar Singh. 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.

Gitimoni Talukdar, Pranjal Protim Borah, and Arup Barua. 2014. A survey of named entity recognition in assamese and other indian languages. arXiv preprint arXiv:1407.2918 (2014).

Van-Hai Vu, Quang-Phuoc Nguyen, Kiem-Hieu Nguyen, Joon-Choul Shin, and Cheol-Young Ock. 2020. Korean-Vietnamese Neural Machine Translation with Named Entity Recognition and Part-of-Speech Tags. IEEE Transactions on Information and Systems 103, 4 (2020), 866–873.

Vikas Yadav, Rebecca Sharp, and Steven Bethard. 2018. Deep affix features improve neural named entity recognizers. In Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics. 167–172.

Shaodian Zhang and Noémie Elhadad. 2013. Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts. Journal of Biomedical Informatics 46, 6 (2013), 1088–1098.
Fig. 4. Confusion matrix for Maithili for the LSTM-CNNs-CRF (above) and CRF (below) models; The × refers to correctly prediction.

Fig. 5. Confusion matrix for Magahi for the LSTM-CNNs-CRF (above) and CRF (below) models; The × refers to correctly prediction.