Proceedings of the WILDRE-6 Workshop @LREC2020, pages 29–34
Marseille, 20 June 2022

L3Cube-MahaNER: A Marathi Named Entity Recognition Dataset and BERT models

Parth Patil [* 1,3], **Aparna Ranade**[* 1,3], **Maithili Sabane**[* 1,3], **Onkar Litake**[* 1,3], **Raviraj Joshi** [2,3]
[1] Pune Institute of Computer Technology, [2] Indian Institute of Technology Madras, [3] L3Cube Pune
[1,3] Pune, Maharashtra India, [2] Chennai, Tamilnadu India,
{parthpatil8399,aparna.ar217,msabane12}@gmail.com
onkarlitake@ieee.org, ravirajoshi@gmail.com

Abstract

Named Entity Recognition (NER) is a basic NLP task and finds major applications in conversational and search systems. It helps us identify key entities in a sentence used for the downstream application. NER or similar slot filling systems for popular languages have been heavily used in commercial applications. In this work, we focus on Marathi, an Indian language, spoken prominently by the people of Maharashtra state. Marathi is a low resource language and still lacks useful NER resources. We present L3Cube-MahaNER, the first major gold standard named entity recognition dataset in Marathi. We also describe the manual annotation guidelines followed during the process. In the end, we benchmark the dataset on different CNN, LSTM, and Transformer based models like mBERT, XLM-RoBERTa, IndicBERT, MahaBERT, etc. The MahaBERT provides the best performance among all the models. The data and models are available at https://github.com/l3cube-pune/MarathiNLP .

Keywords: Named Entity Recognition, NER, Marathi Dataset, Transformers

1. Introduction

A principal technique of information extraction is Named Entity Recognition. It is an integral part of natural language processing systems. The technique involves the identification and categorization of the named entity (Marrero et al., 2013; Lample et al., 2016). These categories include entities like people's names, locations, numerical values, and temporal values. NER has a myriad of applications like customer service, text summarization, etc. Through the years, a large amount of work has been done for Named Entity Recognition in the English language (Yadav and Bethard, 2018). The work is very mature and the functionality comes out of the box with NLP libraries like NLTK (Bird et al., 2009) and spacy (Honnibal and Montani, 2017). In contrast, limited work is done in the Indic languages like Hindi and Marathi (Kale and Govilkar, 2017). (Patil et al., 2016) addresses the problems faced by Indian languages like the presence of abbreviations, ambiguities in named entity categories, different dialects, spelling variations, and the presence of foreign words. (Shah, 2016) elaborates on these issues along with others like the lack of well-annotated data, fewer resources, and tools, etc. Furthermore, the existing resources for NER in Marathi released in (Murthy et al., 2018) titled IIT Bombay Marathi NER Corpus has only 3588 train sentences and 3 target named entities. Also, about 39 percent of sentences in this dataset contain O tags only further reducing the number of useful tokens. Moreover, many datasets aren't available publicly or contain fewer sample sentences. We aim to build a much bigger Marathi NER corpora with a variety of labels currently missing in the literature. The FIRE 2010 dataset is a comparable dataset with 27,177 sentences but is not publicly available. Although, text classification in Hindi and Marathi has recently received some attention (Joshi et al., 2019; Kulkarni et al., 2022; Kulkarni et al., 2021; Velankar et al., 2021), however the same is not true for NER.

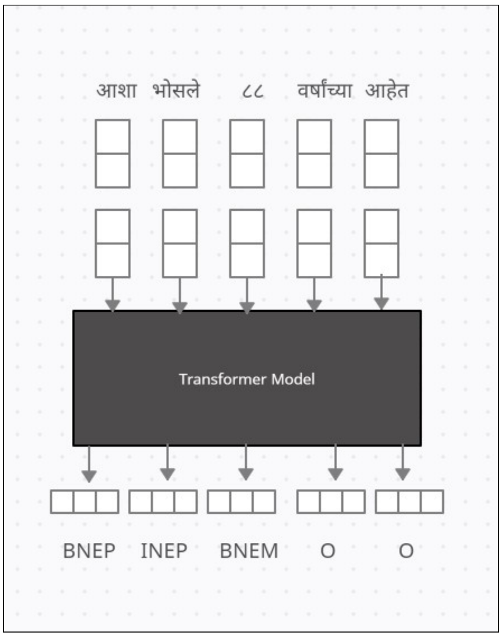


Figure 1: Model Architecture

In this paper, we present our dataset L3Cube-MahaNER. This dataset has been manually annotated and compiled in-house. It is a large dataset annotated

[*] Equal contribution of the authors.

according to the IOB, non-IOB, and binary entity notation for Marathi NER. It contains 25,000 manually tagged sentences categorized according to the eight entity classes. The original sentences have been taken from a news domain corpus (Joshi, 2022a) and the average length of these sentences is 9 words. These entities annotated in the dataset include names of locations, organizations, people, and numeric quantities like time, measure, and other entities like dates and designations. The paper also describes the dataset statistics and the guidelines that have been followed while tagging these sentences.

We also present the results of deep-learning models like Convolutional Neural Network (CNN), Long-Short Time Memory (LSTM), biLSTM, and Transformer models like mBERT (Devlin et al., 2019a), IndicBERT (Kakwani et al., 2020), XLM-RoBERTa, RoBERTa-Marathi, MahaBERT (Joshi, 2022a), MahaROBERTa, MahaALBERT that have been trained on the L3Cube-MahaNER dataset. We experiment on all major multi-lingual and Marathi BERT models to establish a benchmark for future comparisons. The dataset and resources will be publicly shared on Github.

2. Related Work

Named Entity Recognition is a concept that originated at the Message Understanding Conferences (Grishman and Sundheim, 1996) in 1995. Machine learning techniques and linguistic techniques were the two major techniques used to perform NER. Handmade rules (Abdallah et al., 2012) developed by experienced linguists were used in the linguistic techniques. These systems, which included gazetteers, dictionaries, and lexicalized grammar, demonstrated good accuracy levels in English. However, these strategies had the disadvantage of being difficult to transfer to other languages or professions. Decision Trees (Paliouras et al., 2000), Conditional Random Field, Maximum Entropy Model (Bender et al., 2003), Hidden Markov Model, and Support Vector Machine were included in machine learning techniques. To attain better competence, these supervised learning algorithms make use of massive volumes of NE annotated data.

A comparative study by training the models on the same data using Support Vector Machine (SVM) and Conditional Random Field(CRF) was carried out by (Krishnarao et al., 2009). It was concluded that the CRF model was superior. A more effective hybrid system consisting of the Hidden Markov Model, a combination of handmade rules and MaxEnt was introduced by (Srihari, 2000) for performing NER. Deep learning models were then utilized to complete the NER problem as technology progressed. CNN (Albawi et al., 2017), LSTM (Hochreiter and Schmidhuber, 1997), biLSTM (Yang and Xu, 2020), and Transformers were among the most popular models.

NER for Indian languages is a comparatively difficult task due to a lack of capitalization, spelling variances, and uncertainty in the meaning of words. The structure of the language is likewise difficult to grasp. Furthermore, the lack of a well-ordered labeled dataset makes advanced approaches such as deep learning methods difficult to deploy. (Bhattacharjee et al., 2019) has described various problems faced while implementing NER for Indian languages.

(Murthy et al., 2018) introduced Marathi annotated dataset named IIT Bombay Marathi NER Corpus for Named Entity Recognition consisting of 5591 sentences and 108359 tags. They considered 3 main categories named Location, Person, and Organization for training the character-based model on the dataset. They made use of multilingual learning to jointly train models for multiple languages, which in turn helps in improving the NER performance of one of the languages.

(Pan et al., 2017) in 2017 released a dataset named WikiAnn NER Corpus consisting of 14,978 sentences and 3 tags labeled namely Organization, Person, and Location. It is however a silver-standard dataset for 282 different languages including Marathi. This project aims to create a cross-lingual name tagging and linking framework for Wikipedia's 282 languages.

3. Compilation of dataset

3.1. Data Collection

Our dataset consists of 25,000 sentences in the Marathi language. We have used the base sentences from the L3Cube-MahaCorpus (Joshi, 2022a), which is a monolingual Marathi dataset majorly from the news domain. The sentences in the dataset are in the Marathi language with minimal appearance of English words and numerics as present in the original news. However, while annotating the dataset, these English words have not been considered as a part of the named entity categories. Furthermore, the dataset does not preserve the context of the news, such as the publication profiles, regions, and so on.

3.2. Dataset Annotation

We have manually tagged the entire dataset into eight named entity classes. These classes include Person (NEP), Location(NEL), Organization(NEO), Measure(NEM), Time(NETI), Date(NED), and Designation(ED). While tagging the sentences, we established an annotation guideline to ensure consistency. The first 200 sentences were tagged together to further establish consistency among four annotators proficient in Marathi reading and writing. Post this the tagging was performed in parallel except for ambiguous sentences which were separately handled. Firstly, the sentences were relieved of any contextual associations. Then, the approach for the contents of the named entity classes was decided as follows. Proper nouns involving persons' names are tagged as NEP and places are tagged as NEL. All kinds of organizations like companies, councils, political parties, and government departments are

Link to the dataset

Dataset	Sentence Count	Tag Count
Train	21500	27300
Test	2000	2472
Validation	1500	1847

Table 1: Count of sentences and tags in the dataset.

Tags	Train	Test	Validation
NEM	7052	620	488
NEP	6910	611	457
NEL	4949	447	329
NEO	4176	385	268
NED	2466	244	182
ED	1003	92	75
NETI	744	73	48

Table 2: Count of individual tags of L3Cube-MahaNER.

tagged as NEO. Numeric quantities of all kinds are tagged as NEM concerning the context. Furthermore, temporal values like time are tagged as NETI, and dates are tagged as NED. Apart from that, individual titles and designations, which precede proper nouns in the sentences are tagged as ED. Despite maintaining these guidelines, some entities had ambiguous meanings and were difficult to tag. In these circumstances, we resolved the intricacies unanimously by taking a vote amongst the annotators. The sentences were tagged according to the predominant vote.

3.3. Dataset Statistics

For more clarity, some example sentences with tagged entities are mentioned in Table 6.

4. Experimental Techniques

4.1. Model Architectures

The deep learning models are trained using large labeled datasets and the neural network architectures learn features from the data effectively, without the need for feature extraction to be done manually.

Tags	Train	Test	Validation
B-NEM	5824	523	404
I-NEM	1228	97	84
B-NEP	4775	428	322
I-NEP	2135	183	135
B-NEL	4461	407	293
I-NEL	488	40	36
B-NEO	2741	256	178
I-NEO	1435	129	90
B-NED	1937	191	141
I-NED	529	53	41
B-ED	838	74	61
I-ED	165	18	14
B-NETI	633	63	43
I-NETI	111	10	5

Table 3: Count of individual tags of L3Cube-MahaNER.

Similarly, the transformer aims to address sequence-to-sequence problems while also resolving long-range relationships in natural language processing. The transformer model contains a ”self-attention” mechanism that examines the relationship between all of the words in a phrase. It provides differential weightings to indicate which phrase components are most significant in determining how a word should be read. Thus the transformer identifies the context that assigns each word in the sentence its meaning. The training time also is lowered as the feature enhances parallelization.

CNN: This model uses a single 1D convolution over the 300-dimensional word embeddings. These embeddings are fed into a Conv1D layer having 512 filters and a filter size of 3. The output at each timestep is subjected to a dense layer of size 8. The dense layer size is equal to the size of the output labels. There are 8 output labels for non-IOB notation and 15 output labels for IOB notation. The activation function used is relu. All the models have the same optimizer and loss functions. The optimizer used is RMSPROP. The embedding layer for all the word-based models is initialized using fast text word embeddings.

LSTM: This model uses a single LSTM layer to process the 300-dimensional word embeddings. The LSTM layer has 512 hidden units followed by a dense layer similar to the CNN model.

biLSTM: It is analogous to the CNN model with the single 1D convolution substituted by a biLSTM layer. An embedding vector of dimension 300 is used in this model and the biLSTM has 512 hidden units. A batch size of 16 is used.

BERT: BERT (Devlin et al., 2019b) is a Google-developed transformer-based approach for NLP pre-training that was inspired by pre-training contextual representations. It's a deep bidirectional model, which means it's trained on both sides of a token's context. BERT's most notable feature is that it can be fine-tuned by adding a few output layers.

mBERT: mBERT (Pires et al., 2019), which stands for multilingual BERT is the next step in constructing models that understand the meaning of words in context. A deep learning model was built on 104 languages by concurrently encoding all of their information on mBERT.

ALBERT: ALBERT (Lan et al., 2020) is a transformer design based on BERT that requires many fewer parameters than the current state-of-the-art model BERT. These models can train around 1.7 times quicker than BERT models and have greater data

Model	F1	Precision	Recall	Accuracy
mBERT	82.82	82.63	83.01	96.75
Indic BERT	84.66	84.10	85.22	97.09
XLM-RoBERTa	84.19	83.42	84.97	97.12
RoBERTa-Marathi	81.93	81.58	82.29	96.67
MahaBERT	84.81	84.55	85.07	97.10
MahaRoBERTa	**85.30**	**84.27**	**86.36**	**97.18**
MahaAlBERT	84.50	84.54	84.45	96.98
CNN	72.2	81.0	66.6	97.16
LSTM	70.0	77.1	64.8	94.46
biLSTM	73.7	77.2	77.6	94.99

Table 4: F1 score(macro), precision and recall of various transformer and normal models for IOB notation using the Marathi dataset.

Model	F1	Precision	Recall	Accuracy
mBERT	85.3	82.83	97.94	96.92
Indic BERT	86.56	85.86	87.27	97.15
XLM-RoBERTa	85.69	84.21	87.22	97.07
RoBERTa-Marathi	83.86	82.22	85.57	96.92
MahaBERT	**86.80**	**84.62**	**89.09**	**97.15**
MahaRoBERTa	86.60	84.30	89.04	97.24
MahaAlBERT	85.96	84.32	87.66	97.32
CNN	79.5	82.1	77.4	97.28
LSTM	74.9	84.1	68.5	94.89
biLSTM	80.4	83.3	77.6	94.99

Table 5: F1 score(macro), precision and recall of various transformer and normal models for non-IOB notation using the Marathi dataset.

throughput than BERT models. IndicBERT is a multilingual ALBERT model that includes 12 main Indian languages and was trained on large-scale datasets. Many public models, such as mBERT and XLM-R, have more parameters than IndicBERT, although the latter performs exceptionally well on a wide range of tasks.

RoBERTa: RoBERTa (Liu et al., 2019) is an unsupervised transformers model that has been trained on a huge corpus of English data. This means it was trained exclusively on raw texts, with no human labeling, and then utilized an automated approach to generate labels and inputs from those texts. The multilingual model XLM-RoBERTa has been trained in 100 languages. Unlike certain XLM multilingual models, it does not require lang tensors to detect which language is being used. It can also deduce the correct language from the supplied ids.

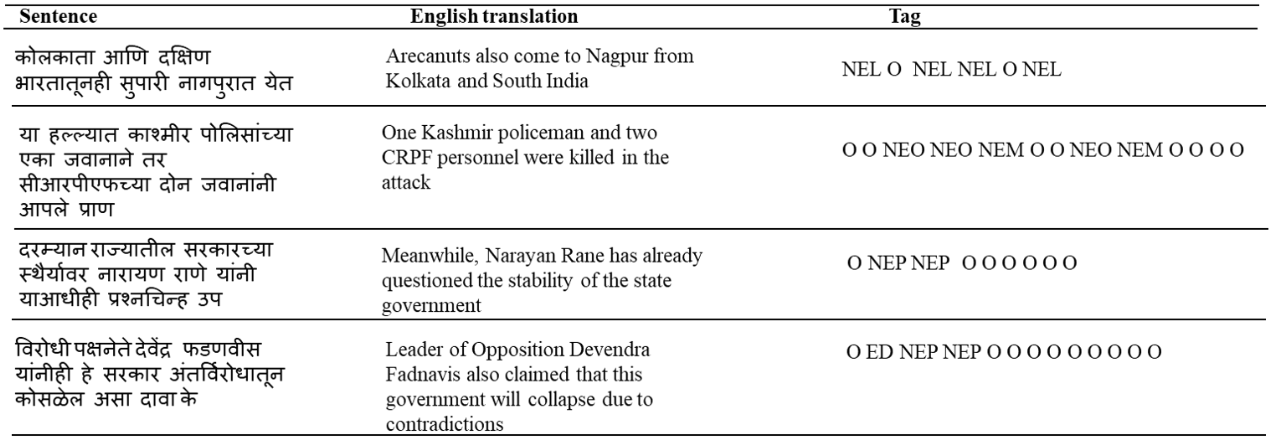
MahaBERT: MahaBERT (Joshi, 2022b) is a 752 million token multilingual BERT model fine-tuned using L3Cube-MahaCorpus as well as other Marathi monolingual datasets that are available publicly.

MahaROBERTA: MahaROBERTA (Joshi, 2022b)is a MarathiRoBERTa model that is based on a multilingual RoBERTa (xlm-roberta-base) framework that has been fine-tuned using L3Cube-MahaCorpus and other publicly released Marathi monolingual corpora.

MahaALBERT: MahaALBERT (Joshi, 2022b) is an AlBERT-based Marathi monolingual model trained using L3Cube-MahaCorpus as well as other Marathi monolingual datasets available publicly.

5. Results

In this study, we have experimented with various model architectures like CNN, LSTM, biLSTM, and transformers like BERT, and RoBERTa to perform named entity recognition on our dataset. This section presents the F1 scores attained by training these models on our dataset for IOB and non-IOB notations. The results have been reported in Table 4 and Table 5 respectively. Among the CNN and LSTM-based models, the biLSTM model with the trainable word embeddings gives the best results on the L3Cube-MahaNER dataset for IOB as well as non-IOB notations. Moreover, for the transformers-based models, it is observed that the Ma-



Sentence	English translation	Tag
कोलकाता आणि दक्षिण भारतातूनही सुपारी नागपुरात येत	Arecanuts also come to Nagpur from Kolkata and South India	NEL O NEL NEL O NEL
या हल्ल्यात काश्मीर पोलिसांच्या एका जवानाने तर सीआरपीएफच्या दोन जवानांनी आपले प्राण	One Kashmir policeman and two CRPF personnel were killed in the attack	O O NEO NEO NEM O O NEO NEM O O O O
दरम्यान राज्यातील सरकारच्या स्थैर्यावर नारायण राणे यांनी याआधीही प्रश्नचिन्ह उप	Meanwhile, Narayan Rane has already questioned the stability of the state government	O NEP NEP O O O O O O
विरोधी पक्षनेते देवेंद्र फडणवीस यांनीही हे सरकार अंतर्विरोधातून कोसळेल असा दावा के	Leader of Opposition Devendra Fadnavis also claimed that this government will collapse due to contradictions	O ED NEP NEP O O O O O O O O O

Table 6: Sample Tagged Sentences

haRoBERTa model yields the best results for IOB and MahaBERT provides the best results for non-IOB notations. The LSTM and the RoBERTa-Marathi models report the lowest scores among all models for both.

6. Conclusion

In this paper, we hold forth on the problem of scarcity of annotated corpora and hence present L3Cube-MahaNER which is a large dataset for Marathi Named Entity, containing 25000 distinct sentences. We achieved the results using IOB and non-IOB notations on deep learning models such as CNN, LSTM, biLSTM, and transformers in BERT as listed above, to set the basis for future work. We observed the highest accuracy on MahaRoBERTa for IOB notations and model MahaBERT for non-IOB notations. We believe that our corpus will play a pivotal role in expanding conversational AI for the Marathi Language.

7. Bibliographical References

Abdallah, S., Shaalan, K., and Shoaib, M. (2012). Integrating rule-based system with classification for arabic named entity recognition. volume 7181, pages 311–322, 03.

Albawi, S., Mohammed, T. A., and Al-Zawi, S. (2017). Understanding of a convolutional neural network. In *2017 International Conference on Engineering and Technology (ICET)*, pages 1–6.

Bender, O., Och, F. J., and Ney, H. (2003). Maximum entropy models for named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 148–151.

Bhattacharjee, K., S, S. K., Mehta, S., Kumar, A., Mehta, R., Pandya, D., Chaudhari, P., and Verma, D. (2019). Named entity recognition: A survey for indian languages. In *2019 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT)*, volume 1, pages 217–220.

Bird, S., Klein, E., and Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.".

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019a). Bert: Pre-training of deep bidirectional transformers for language understanding.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019b). Bert: Pre-training of deep bidirectional transformers for language understanding.

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

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

Honnibal, M. and Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Joshi, R., Goel, P., and Joshi, R. (2019). Deep learning for hindi text classification: A comparison. In *International Conference on Intelligent Human Computer Interaction*, pages 94–101. Springer.

Joshi, R. (2022a). L3cube-mahacorpus and mahabert: Marathi monolingual corpus, marathi bert

Multicase BERT
Indic BERT
Xlm-roberta
Roberta-Marathi
Roberta-Hindi
indic-transformers-hi-roberta
MahaBERT
MahaRoBERTa
MahaAlBERT

language models, and resources. *arXiv preprint arXiv:2202.01159*.

Joshi, R. (2022b). L3cube-mahacorpus and mahabert: Marathi monolingual corpus, marathi bert language models, and resources.

Kakwani, D., Kunchukuttan, A., Golla, S., N.C., G., Bhattacharyya, A., Khapra, M. M., and Kumar, P. (2020). IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online, November. Association for Computational Linguistics.

Kale, S. and Govilkar, S. (2017). Survey of named entity recognition techniques for various indian regional languages. *International Journal of Computer Applications*, 164(4):37–43.

Krishnarao, A. A., Gahlot, H., Srinet, A., and Kushwaha, D. S. (2009). A comparative study of named entity recognition for hindi using sequential learning algorithms. In *2009 IEEE International Advance Computing Conference*, pages 1164–1169.

Kulkarni, A., Mandhane, M., Likhitkar, M., Kshirsagar, G., and Joshi, R. (2021). L3cubemahasent: A marathi tweet-based sentiment analysis dataset. *arXiv preprint arXiv:2103.11408*.

Kulkarni, A., Mandhane, M., Likhitkar, M., Kshirsagar, G., Jagdale, J., and Joshi, R. (2022). Experimental evaluation of deep learning models for marathi text classification. In *Proceedings of the 2nd International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications*, pages 605–613. Springer.

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

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2020). Albert: A lite bert for self-supervised learning of language representations.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach.

Marrero, M., Urbano, J., Sánchez-Cuadrado, S., Morato, J., and Gómez-Berbís, J. M. (2013). Named entity recognition: fallacies, challenges and opportunities. *Computer Standards & Interfaces*, 35(5):482–489.

Murthy, R., Kunchukuttan, A., and Bhattacharyya, P. (2018). Judicious selection of training data in assisting language for multilingual neural NER. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 401–406.

Paliouras, G., Karkaletsis, V., Petasis, G., and Spyropoulos, C. D. (2000). Learning decision trees for named-entity recognition and classification. In *In ECAI Workshop on Machine Learning for Information Extraction*.

Pan, X., Zhang, B., May, J., Nothman, J., Knight, K., and Ji, H. (2017). Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada, July. Association for Computational Linguistics.

Patil, N., Patil, A., and B.V, P. (2016). Issues and challenges in marathi named entity recognition. *International Journal on Natural Language Computing*, 5.

Pires, T., Schlinger, E., and Garrette, D. (2019). How multilingual is multilingual bert?

Shah, H. (2016). Study of named entity recognition for indian languages. *International Journal of Information Sciences and Techniques*, 6.

Srihari, R. (2000). A hybrid approach for named entity and sub-type tagging. In *Sixth Applied Natural Language Processing Conference*, pages 247–254, Seattle, Washington, USA, April. Association for Computational Linguistics.

Velankar, A., Patil, H., Gore, A., Salunke, S., and Joshi, R. (2021). Hate and offensive speech detection in hindi and marathi. *arXiv preprint arXiv:2110.12200*.

Yadav, V. and Bethard, S. (2018). A survey on recent advances in named entity recognition from deep learning models. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2145–2158.

Yang, G. and Xu, H. (2020). A residual bilstm model for named entity recognition. *IEEE Access*, 8:227710–227718.