Neural language model embeddings for Named Entity Recognition: A study from language perspective

Muskaan Maurya
The English & Foreign Languages University
muskaan.maurya06@gmail.com

Anupam Mandal
Centre for AI & Robotics
amandal.cair@gov.in

Manoj Maurya
Centre for AI & Robotics
manoj.cair@gov.in

Naval Gupta
Centre for AI & Robotics
naval.gupta.cair@gov.in

Somya Nayak
The English & Foreign Languages University
somyanayak@efluniversity.ac.in

Abstract

Named entity recognition (NER) models based on neural language models (LMs) exhibit state-of-the-art performance. However, the performance of such LMs have not been studied in detail with respect to finer language related aspects in the context of NER tasks. Such a study will be helpful in effective application of these models for cross-lingual and multilingual NER tasks. In this study, we examine the effects of script, vocabulary sharing, foreign names and pooling of multilanguage training data for building NER models. It is observed that monolingual BERT embeddings show the highest recognition accuracy among all transformer-based LMs for monolingual NER models. It is also seen that vocabulary sharing and data augmentation with foreign named entities (NEs) are most effective towards improving accuracy of cross-lingual NER models. Multilingual NER models trained by pooling data from similar languages can address training data inadequacy and exhibit performance close to that of monolingual models trained with adequate NER-tagged data of a single language.

1 Introduction

Named Entity Recognition (NER) is one of the active areas of research in natural language processing that involves automatic tagging of names of people, places, organizations, *etc*., in natural language text. NER finds its use in applications like information extraction, summarization, and question-answering. As names are out-of-vocabulary words for any given language, NER models try to learn the relationship of these Named Entity (NE) words with the surrounding words from the context. Both statistical (Saha et al., 2008) and deep neural network based techniques have been used for NER modeling. However, deep neural networks have been found to be effective in learning latent relationships from long context (Li et al., 2023). Therefore it has been the subject of much of NER research in recent past. In this context, NER models based on bi-directional Long-Short-Term Memory networks (bi-LSTM) (Huang et al., 2015) coupled with Conditional Random Fields (CRF) have shown impressive performance. However, such architectures are limited by sequential processing resulting in slower performance. NER models trained using embeddings generated by neural LMs based on transformer architecture (Vaswani et al., 2017) have shown state-of-the-art performance on NER tasks. The improvement in accuracy has been due to the attention mechanism being able to capture the relevant relationships effectively, while inherent parallelism in transformer architecture has been able to improve the processing speed.

Despite transformer architectures based LMs being the mainstay of the current state-of-the-art NER models, the focus of research has been primarily to improve these LMs or extend applications of such pre-trained LMs to different languages. This assumes availability of adequate NER tagged training data for such languages. However, for low-resource scenarios where adequate amount of NER tagged training data is not available, cross-lingual and multilingual NER models are generally resorted to (Conneau et al., 2020).

The previous studies (Fu et al., 2023) in this context have examined the performance of NER models in cross-lingual scenarios for homogenous and heterogenous languages. The homogeneity of these languages is determined based on their membership in different language families. However, the effect of finer language related aspects such as vocabulary, scripts, context, and language structure has not been studied with respect to NER tasks. For example, how does a NER model trained with data of a particular context having context-specific names perform when applied to a different context and unseen names. It is felt that such a study is needed not only to answer such questions but also for the effective application of the NER models in cross-lingual and multilingual scenarios. In this work, we study the performance of transformer based language models in the context of NER tasks considering different language-related aspects namely, script similarity, vocabulary sharing, cross-language, and multilingual knowledge sharing. This study involves NER models trained with embeddings obtained using different transformer based architectures.

The paper is organized as follows. The different language related aspects considered in this study that are likely to affect NER performance are discussed in Section 2. The languages, transformer-based LMs, and datasets used in this study are described in Section 3. The experiments and results are presented in Section 4. Finally, we conclude the paper in Section 5 with key findings of the study and directions for future work.

2 Language-related aspects and NER models

The performance of NER models is influenced directly and indirectly by different language related aspects mentioned in Section 1. In this section, we discuss in detail how these aspects impact both the development and performance of NER models, particularly when they are applied in multilingual and cross-lingual scenarios.

NER models extract named entities from text based on the relationship between named entities and other words in a sentence. Embeddings obtained from language models that are used to train NER models implicitly capture different aspects of the language structure. For example, the positioning of proper nouns (named entities) in a valid sentence construct of a language is captured using positional embeddings. The positional embeddings are language-specific and depend on the grammar and structure of a language. Hence, training data containing NEs in different sentence constructs (e.g., active and passive voice form) can capture linguistic variations of a language and is likely to have a positive impact on the accuracy of the NER models.

Another important aspect is the context of usage of named entities in the text of a language. For instance, the Hindi word *Muskaan* means *smile* while the same word is also used as name of a person of feminine gender in Indian sub-continent. The word in the later use case is a NE while the former is not. This understanding can be derived only from context that plays an important role in determining NEs. LM embeddings are expected to capture such contextual information, that are exploited during training of NER models. The ability of a NER model to differentiate between the above two forms of usage of the same word (one as an NE and the other as non-NE) depends on whether such variations in context of usage has been presented during training. As such context related variations are language-specific, performance needs to be studied for such models in cross-lingual scenarios.

Many languages often share a common script. However, vocabulary sharing between the languages may be low (*e.g.,* Urdu and Arabic) or high (*e.g.,* Urdu and Persian). On the other hand, the script is different for Hindi and Urdu. However, there is a high degree of vocabulary sharing between the two languages. A specific challenge for cross-lingual NER is related to processing of scripts, that is described as follows. The generation of word embeddings requires a tokenizer specific to the script of a language. Therefore, the performance of NER models trained with word embeddings of a language that shares vocabulary but not the script with another language needs to be investigated. The performance of monolingual NER models (those trained with word embeddings of respective languages) in such cross-lingual scenarios carrying similar and dissimilar language pairs needs to be studied. The outcome of the study can be used to identify attributes that can help in deciding whether an NER model can be shared across languages with acceptable performance. Such information can be leveraged for low-resource languages where building of NER models from scratch is difficult due to the absence of adequate training data. In such a scenario, a NER model trained on a high resource language can be alternatively used for a low resource language with reasonable accuracy.

Multi-lingual training (Conneau et al., 2020) involves the pooling of training data across multiple languages when the amount of training data from a single language is inadequate to meet training requirements. However, the choice of languages to be pooled is an important consideration and needs to be studied. The proportion of individual languages in the pooled dataset is also likely to affect the overall NER performance. A study in this aspect can provide insight into the language bias (if any) caused in the model due to the varying distribution of individual language data in the training dataset.

3 Design of experiments

3.1 Languages selected for the study

In order to examine the effect of different language related aspects on NER performance, languages from major non-latin languages of the world were selected for the study. The languages chosen for the study, all from Asia exhibit diversity in terms of vocabulary, writing system (scripts), and language structure are used by a sizeable population. The language families, member languages, and corresponding writing systems are described in Table 1.

Language Families	Members	Writing System (scripts)
Indo-Aryan	Hindi	Abugida
Indo-Aryan	Urdu	Abjad
Iranian	Persian	Abjad
Sino-Tibetan	Chinese	Logographic
Dravidian	Tamil	Abugida
Semitic	Arabic	Abjad

Table 1: Language families chosen for the study

An NER model was trained for each of the above languages using NER tagged data of the corresponding language. However, in order to study the performance of these NER models in cross-lingual scenarios, a study was conducted by grouping these languages into pairs considering similarity of different language-related attributes that are likely to be relevant to NER task. These groups are shown in Table 2.

Language Pairs	Language related attributes		
	Similar language family	Script similarity	Vocabulary sharing
Hindi - Urdu	✓	✗	high
Urdu - Persian	✓	✓	high
Arabic - Urdu	✗	✓	low
Chinese - Tamil	✗	✗	low

Table 2: Language groups based on language attributes

3.2 Transformer-based LM architectures

As discussed in Section 3.1, NER models were built using tagged text data of respective languages. The training of NER models was done with embeddings generated by different LMs. The language models in turn were trained with the following transformer architectures:

1. **BERT** (Devlin et al., 2019)
2. **ALBERT** (Lan et al., 2019)
3. **RoBERTa** (Liu et al., 2019)
4. **XLM-R** (Conneau et al., 2020)

3.3 Datasets

The experiments in this study were carried out using both publicly available and custom created datasets. Language-specific NER models were trained using publicly available datasets. Table 3 shows these datasets for different languages used in this study along with their domain and sentence count.

However, in order to study the performance of NER models for specific scenarios *e.g.,* when the context of NEs in the training data differs from that in evaluation data or when foreign names are present, custom evaluation datasets were created. One such study was done for Chinese NER models trained with data having contextual relevance to Chinese society. The NER models were trained using openly available Chinese datasets mentioned in Table 3. However, the evaluation was performed on a custom created dataset, having contextual relevance to both Indian and Chinese society. This dataset referred to as *Chinese-Hindi-Evaluation-Data* (CHED) is created in Chinese script with the help of linguist. The CHED dataset consisting of 790 sentences is divided into two parts:

- The first part referred to as CHED-I consists of 400 Chinese sentences having Chinese names and content typical of Chinese society.
- The second part referred to as CHED-II consists of 390 Chinese sentences containing Indian names and content typical of Indian (*e.g.,* Urdu and Arabic) society.

In order to study the efficacy of cross-lingual knowledge transfer of NER models, evaluation datasets were created with foreign names that are not usually seen in the given language. Such a dataset named as *Chinese-Hindi-foreign* (CHF) was created by substituting Chinese names with Indic names in the Chinese NER dataset. The In-

Language	Name of dataset	No of sentences	Tagged NEs	Domain coverage
Hindi	Naamapadam (Mhaske et al., 2023)	999684	PER, LOC, ORG	-
Chinese	MSRA (Levow, 2006)	48444	PER, LOC, ORG, GPE	-
	People's Daily (PD) (Xu)	27821	PER, LOC, ORG	News
	Wikiann (Pan et al., 2017)	40000	PER, LOC, ORG	Wikipedia
Urdu	MC-PUCIT (Irfan)	362257	PER, LOC, ORG	-
	Jahangir (Irfan)	1662	PER, LOC, ORG, DAT, TIM	-
	IJNLP (Irfan)	2005	PER, LOC, ORG, TIM, NUM, DES	-
	Wikiann (Pan et al., 2017)	22000	PER, LOC, ORG	Wikipedia
Arabic	AQMAR (Mohit et al., 2012)	2298	PER, LOC, ORG, MISC	Wikipedia
	ANER (Benajiba et al., 2007)	4871	PER, LOC, ORG, MISC	Wikipedia
	Wikiann (Pan et al., 2017)	40000	PER, LOC, ORG	Wikipedia
Persian	ARMAN (Poostchi et al., 2018a)	7682	PER, LOC, ORG, FAC, EVE, PRO	-
	PEYMA (Shahshahani et al., 2018)	7145	PER, LOC, ORG, MON, DAT, TIM, PER	-
	ParsNER (Poostchi et al., 2018b)	40324	PER, LOC, ORG, MON, DAT, TIM, PER, FAC, EVE, PRO	-
	PersianNER (Ehsan Asgarian)	100000	PER, LOC, ORG, DAT, TIM, EVE	Persian Wikipedia
	Wikiann (Pan et al., 2017)	40000	PER, LOC, ORG	Wikipedia
	FarsiNER (Taghizadeh et al., 2019)	510299	PER, LOC, ORG	-
Tamil	Naamapadam (Mhaske et al., 2023)	500726	PER, LOC, ORG	-

Table 3: Dataset information: *PER: Person, LOC: Location, ORG: Organization, GPE: Geo-PoliticaL Entity, DAT: Date, TIM: Time, NUM: Number, DES: Designation, MISC: Miscellaneous, FAC: Facility, EVE: Event, PRO: Product*

dic names are represented with Chinese logograms. The process of substitution is explained below.

Assume a NE tagged Chinese sentence in the dataset:

玛	丽	的	中	文	不	太	好
B-PER	I-PER	O	O	O	O	O	O

The corresponding sentence having an Indian name (in Chinese logograms) replacing the Chinese name is given as:

帕	哈	斯	的	中	文	不	太	好
B-PER	I-PER	I-PER	O	O	O	O	O	O

4 Experiments and Results

The first set of three experiments (Experiments 1, 2, and 3) were conducted with monolingual NER models trained with NER-tagged data of a single language.

The training was carried out using a DGX-A-100 workstation with the following hyper-parameters: Number of epochs: 8, Number of GPU cards: 2, Batch per GPU device: 50, Total train batch size: 100. The performance accuracy is evaluated in terms of F1 score that is defined as:

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

4.1 Experiment 1

The first experiment was to study the performance of embeddings obtained from different monolingual and multi-lingual transformer based LMs for a NER task. The study was conducted for Hindi, Chinese, Persian, Arabic, and Urdu languages where the NER models were trained on NE-tagged datasets of respective languages with embeddings derived using the different LMs mentioned in Section 3.2. The details of the languages, corresponding transformer based language models, and NER training and evaluation datasets used in the study are summarized in Table 4. The models were identified based on their availability in the domain and prior application in NER studies. Table 5 shows the performance of monolingual NER models in terms of F1 scores for each language.

It is seen that monolingual NER models trained using monolingual BERT embeddings exhibit the highest recognition accuracy among all transformer models for four out of five languages.

4.2 Experiment 2

The next experiment was to study the efficacy of NER models in a cross-lingual scenario. Towards this, NER models trained in one language were evaluated with test data of another language. In this experiment, language pairs were chosen considering the grouping of languages mentioned in Section 3.1. The evaluation was done using the test data set mentioned in Table 4.

The results of each train-test language pair is given in Table 6. It is observed that cross-lingual knowledge transfer in context of NER task is high when embeddings are generated using multilingual LMs. These embeddings also demonstrate marginal improvement in NER accuracy over em-

Language	Model Type	Model Name	No training sentences	No validation sentences	No test sentences
Hindi	BERT multilingual ALBERT multilingual RoBERTa	hindiBERT (Joshi, 2022) IndicBERT (Kakwani et al., 2020) XLM-R-base (Conneau et al., 2020)	949698	24993	24993
Chinese	BERT ALBERT multilingual RoBERTa	ckiplab-bert (ckiplab) ckiplab-albert (ckiplab) XLM-R-base (Conneau et al., 2020)	85866	12319	18080
Persian	BERT ALBERT multilingual RoBERTa	bert-fa-base-uncased (Farahani et al., 2020) albert-base-fa (Team, 2021) XLM-R-base (Conneau et al., 2020)	705450	18565	18564
Arabic	BERT multilingual RoBERTa	bert-base-ar (Inoue et al., 2021) XLM-R-base (Conneau et al., 2020)	45095	1187	1187
Urdu	monolingual RoBERTa multilingual RoBERTa	roberta-base-ur (UrduHack) XLM-R-base (Conneau et al., 2020)	349131	19396	19397
Tamil	BERT multilingual RoBERTa	tamilBERT (Joshi, 2022) XLMR-base (Conneau et al., 2020)	475654	12536	12536

Table 4: Datasets and models used in this study

	Hindi	Chinese	Persian	Arabic	Urdu
	monolingual language models				
BERT	**0.804**	**0.8880**	**0.9717**	**0.9634**	-
ALBERT	-	0.8154	0.9279	-	-
RoBERTa	-	-	-	-	**0.9892**
	multilingual language models				
mALBERT	0.8037	-	-	-	-
XLM-R	0.8032	0.8657	0.8024	0.9536	0.9889

Table 5: F1 scores of different monolingual and multilingual models based on different LM architectures for different languages

Language	Model	Test data Language	No of test data sentences	F1
similar script - high vocabulary sharing				
Persian	BERT	Urdu	19397	0.1259
	XLM-R			0.5261
Urdu	RoBERTa	Persian	18564	0.1069
	XLM-R			0.3952
similar script - low vocabulary sharing				
Arabic	BERT	Urdu	19397	0.1421
	XLM-R			0.1158
Urdu	RoBERTa	Arabic	1187	0.0558
	XLM-R			0.1318
different script - high vocabulary sharing				
Hindi	BERT	Urdu	19397	0.1199
	XLM-R			0.4425
Urdu	RoBERTa	Hindi	24993	0.0004
	XLM-R			0.5561
different script - low vocabulary sharing				
Chinese	BERT	Tamil	12536	0.3158
	XLM-R			0.4471
Tamil	BERT	Chinese	18080	0.0031
	XLM-R			0.1701

Table 6: F1 scores of NER models in cross-lingual scenario for different language pairs

beddings generated using monolingual LMs for language pairs that have low vocabulary sharing. Further, knowledge transfer is most effective for language pairs that have a high degree of vocabulary sharing.

4.3 Experiment 3

This experiment was conducted to study the effect of content and context with respect to dissimilar languages that have significant differences in vocabulary and writing systems. Chinese and Hindi were identified as dissimilar languages based on the above criteria. In this study, three different NER models were built corresponding to the following cases:

1. Text data having Chinese names and context was used to train *baseline* NER Chinese models.
2. Chinese names in the training dataset were substituted with Indic names and a new *substitution* NER model was built. In this case, the context of the NEs was preserved in the substituted dataset, however with a change in content (foreign names).
3. In this case, a NER model was trained by combining both the original and the substituted training datasets as described above. This is referred to as *augmentation* model.

The training datasets for the above three cases are shown in Table 7.

Dataset category	Named entity context	No of train data sentences	No of validation data sentences
Baseline	Chinese	110119	5796
Substitution (CHF)	Indian	76151	9008
Augmentation	Chinese + Indian	176466	19608

Table 7: Datasets used in Experiment 3

The evaluation of the three NER models described above was carried out on CHED dataset (refer to Section 3.3). The results are given in Table 8.

Training Data	Transformer LMs	F1	
		Chinese-NEs Chinese-context (CHED - I)	Indic-NEs Chinese-context (CHED - II)
Baseline	BERT	0.7896	0.6824
	ALBERT	0.7344	0.6002
	XLM-R	0.7577	0.6788
Substitution	BERT	0.029	0.7370
	ALBERT	0.0527	0.6822
	XLM-R	0.0352	0.7053
Augmentation	BERT	0.7881	0.7860
	ALBERT	0.7330	0.7178
	XLM-R	0.7604	0.7616

Table 8: F1 scores of NER models in recognition of foreign named entities

The results indicate that NER models trained on datasets containing native names in a given language perform poorly when used for the recognition of foreign (unseen) names. This can be addressed by augmenting the training dataset with foreign names which leads to an improvement in the NER performance.

The second set of experiments was to study how language-related factors affect the performance of multilingual NER models. In this study, the NER models were trained with NER-tagged data of more than one language. A cross-lingual LM, namely XLM-R was used for the generation of embedding as the training dataset consisted of content pooled from more than one language. Here, the performance of NER models was studied with respect to the following:

1. Type of languages pooled together to create the training dataset.
2. Proportion (in terms of volume/ number of sentences) of individual languages in the pooled dataset.

4.4 Experiment 4

In this experiment, bilingual and multilingual NER models were trained by pooling NER-tagged data from two and four languages respectively. These NER models are trained with embeddings generated by XLM-R model from the pooled dataset consisting of NER-tagged data of corresponding languages. The language pairs and the corresponding bilingual NER models along with their NER performance figures evaluated on individual languages, are shown in Table 9.

Next, the number of languages in the NER training dataset was increased to four consisting of dissimilar languages. The list of languages used for training the multilingual NER model and its performance is reported in Table 10.

Lang 1	Lang 2	NER model	F1 on Lang1	F1 on Lang2
similar language pairs				
Persian	Urdu	Fa-Ur	0.7446	0.9891
Urdu	Hindi	Ur-Hi	0.9885	0.8156
Arabic	Urdu	Ar-Ur	0.9539	0.9894
dissimilar language pairs				
Arabic	Chinese	Ar-Zh	0.9537	0.8646
Chinese	Hindi	Zh-Hi	0.8677	0.8187
Tamil	Hindi	Ta-Hi	0.7594	0.8000
Arabic	Tamil	Ar-Ta	0.9525	0.7584
Chinese	Tamil	Zh-Ta	0.8677	0.7591
Arabic	Hindi	Ar-Hi	0.9550	0.8158

Table 9: F1 scores of bi-lingual NER models with XLM-R embedding

Languages	F1
Arabic	0.9541
Hindi	0.8000
Tamil	0.7573
Chinese	0.8691

Table 10: F1 scores of multilingual NER model with XLM-R embedding

It can be observed that the accuracy of the monolingual NER models (refer to Table 5) trained on datasets of respective languages matches closely with that of multilingual NER models trained on a pooled NER dataset of similar language pairs (refer to Table 9). This shows that in the event of inadequate availability of NER training data in a given language, such datasets of similar languages can be pooled together to train models that can exhibit NER accuracy close to that of monolingual NER models.

4.5 Experiment 5

In this experiment, the effect of the volume of language data on the performance of multilingual NER models was studied. This is particularly useful when a small amount of NER training data available for a target language can be mixed with a bigger training dataset of another language in order to build a NER model in the target language. The proportion of the smaller and bigger language datasets (in terms of number of sentences) needs to be studied at which the NER model exhibits an acceptable recognition performance. For this, NER models were built by progressively increasing the proportion of one of the languages in the pooled training dataset. This was done by increasing the number of training sentences in the target language (smaller dataset) while keeping the number of sentences in the other language (bigger dataset) constant. As the nature (similarity) of the languages pooled to create the training set may have a bearing on NER performance, the study was carried out separately for NER models trained with similar and dissimilar language pairs.

From language pairs mentioned in Table 2, Persian-Urdu (similar) and Tamil-Chinese (dissimilar) were selected. The criteria of similarity were based on similarity in script, vocabulary, and membership of same language family. NER models were also built by reversing the roles of languages in a given language pair. Table 11 shows NER performance for Persian-Urdu and Chinese-Tamil models when evaluated on test datasets mentioned in Table 4.

Language 1	Language 2	No of training sentences of language 2	F1 on language 2
similar script - high vocabulary sharing			
Persian (acc: 0.75)	Urdu	0	0.5261
		1000	0.6922
		5000	0.7504
		25000	0.8198
		125000	0.9546
		349131	0.9891
Urdu (acc: 0.98)	Persian	0	0.3952
		1000	0.4551
		5000	0.5000
		25000	0.5789
		125000	0.6575
		705450	0.7446
dissimilar script - low vocabulary sharing			
Tamil (acc: 0.76)	Chinese	0	0.1701
		1000	0.4225
		5000	0.5035
		25000	0.6575
		85866	0.8677
Chinese (acc: 0.87)	Tamil	0	0.3158
		1000	0.6184
		5000	0.6678
		25000	0.7006
		125000	0.7328
		476363	0.7591

Table 11: F1 scores showing cross-lingual inferencing strength of XLM-R for different language pairs

It is observed that the accuracy of the NER model for a language increases progressively with an increase in the amount of training data of the target language in the pooled dataset. The quantum of increase in accuracy depends on the nature of languages mixed to create the pooled dataset. Furthermore, the ratio vs the accuracy changes when the languages in language pairs are reversed.

5 Discussion and Conclusion

In this work, we have studied the effect of language related aspects on the performance accuracy of monolingual and multilingual NER models based on different transformer based neural LMs. We have specifically studied the effect of finer language-related aspects namely script similarity, vocabulary sharing, and content and context dependency on NER performance.

It was observed that, for each of the chosen languages, monolingual BERT embeddings gave the highest accuracy among all the embedding models. It was also found that the performance of cross-lingual NER models was highly dependent on vocabulary sharing and the presence of foreign NEs in training data. Further, it was seen that data inadequacy can be handled by pooling data of similar languages.

In future, work can be undertaken with respect to transfer learning techniques for cross-lingual NER models considering different language-related aspects mentioned in this study. The scope of this study can be extended to cover other NLP tasks such as machine translational as well.

Acknowledgements

The authors would like to thank Director CAIR for his support in carrying out the work.

References

Yassine Benajiba, Paolo Rosso, and José BenedíRuiz. 2007. Anersys: An arabic named entity recognition system based on maximum entropy. pages 143–153.

ckiplab. Chinese Transformer-based Language models. https://github.com/ckiplab/ckip-transformers.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the ACL*, pages 8440–8451, Online. ACL.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc of the 2019 Conf. of the NACACL: Human Language Technologies*, pages 4171–4186, Minneapolis, Minnesota. ACL.

Soheil Alizadeh Ehsan Asgarian, Hamed. Persian NER Corpus. https://github.com/Text-Mining/Persian-NER.

Mehrdad Farahani, Mohammad Gharachorloo, Marzieh Farahani, and Mohammad Manthouri. 2020. Parsbert: Transformer-based model for Persian Language Understanding. *ArXiv*, abs/2005.12515.

Yingwen Fu, Nankai Lin, Boyu Chen, Ziyu Yang, and Shengyi Jiang. 2023. Cross-lingual named entity recognition for heterogenous languages. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:371–382.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *CoRR*, abs/1508.01991.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in Arabic pre-trained language models. In *Proc of the Sixth Arabic NLP Workshop*, Kyiv, Ukraine (Online). ACL.

Muhammad Irfan. Urdu NER Corpus. https://github.com/mirfan899/Urdu.

Raviraj Joshi. 2022. L3Cube-HindBERT and DevBERT: Pre-Trained BERT Transformer models for Devanagari based Hindi and Marathi Languages. *arXiv preprint arXiv:2211.11418*.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In *Findings of the ACL: EMNLP 2020*, pages 4948–4961, Online. ACL.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A lite BERT for self-supervised learning of language representations. *CoRR*, abs/1909.11942.

Gina-Anne Levow. 2006. The third international Chinese language processing bakeoff: Word segmentation and named entity recognition. In *Proc of the Fifth SIGHAN Workshop on Chinese Language Processing*, pages 108–117, Sydney, Australia. ACL.

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2023. A survey on deep learning for named entity recognition: Extended abstract. In *2023 IEEE 39th Int. Conf. on Data Engineering (ICDE)*, pages 3817–3818.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Arnav Mhaske, Harshit Kedia, Sumanth Doddapaneni, Mitesh M. Khapra, Pratyush Kumar, Rudra Murthy, and Anoop Kunchukuttan. 2023. Naamapadam: A large-scale named entity annotated data for Indic languages. In *Proc of the 61st Annual Meeting of the ACL*, pages 10441–10456, Toronto, Canada. ACL.

Behrang Mohit, Nathan Schneider, Rishav Bhowmick, Kemal Oflazer, and Noah A. Smith. 2012. Recall-oriented learning of named entities in arabic wikipedia. EACL '12, page 162–173, USA. ACL.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proc of the 55th Annual Meeting of the ACL*, pages 1946–1958, Vancouver, Canada. ACL.

Hanieh Poostchi, Ehsan Zare Borzeshi, and Massimo Piccardi. 2018a. BiLSTM-CRF for Persian Named-Entity Recognition ArmanPersoNERCorpus: the First Entity-Annotated Persian Dataset. In *LREC*.

Hanieh Poostchi, Ehsan Zare Borzeshi, and Massimo Piccardi. 2018b. BiLSTM-CRF for Persian named-entity recognition ArmanPersoNERCorpus: the first entity-annotated Persian dataset. In *Proc of the Eleventh Int. Conf. on LREC 2018*, Miyazaki, Japan. ELRA.

Sujan Kumar Saha, Sanjay Chatterji, Sandipan Dandapat, Sudeshna Sarkar, and Pabitra Mitra. 2008. A hybrid named entity recognition system for south and south East Asian languages. In *Proc of the IJCNLP-08 Workshop on NER for South and South East Asian Languages*.

Mahsa Sadat Shahshahani, Mahdi Mohseni, Azadeh Shakery, and Heshaam Faili. 2018. Peyma: A tagged corpus for Persian Named Entities. *ArXiv*, abs/1801.09936.

Nasrin Taghizadeh, Zeinab Borhani-fard, Melika Golestani Pour, Mojgan Farhoodi, Maryam Mahmoudi, Masoumeh Azimzadeh, and Heshaam Faili. 2019. NSURL-2019 task 7: Named Entity Recognition (NER) in Farsi. In *Proc of the first Int. Workshop on NSURL*, NSURL '19, Trento, Italy.

Hooshvare Team. 2021. Albert-persian: A lite bert for self-supervised learning of language representations for the persian language. https://github.com/m3hrdadfi/albert-persian.

UrduHack. Urdu Transformer-based Language model. https://github.com/urduhack/urduhack.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Kevin Canwen Xu. People's Daily Chinese NER Corpus. https://huggingface.co/datasets/peoples_daily_ner.