Judicious Selection of Training Data in Assisting Language for Multilingual Neural NER

Rudra Murthy V[†], **Anoop Kunchukuttan**[‡][*], **Pushpak Bhattacharyya**[†]
[†] Center for Indian Language Technology (CFILT)
Department of Computer Science and Engineering
IIT Bombay, India.
[‡]Microsoft AI & Research, Hyderabad, India.
{rudra,pb}@cse.iitb.ac.in, ankunchu@microsoft.com

Abstract

Multilingual learning for Neural Named Entity Recognition (NNER) involves jointly training a neural network for multiple languages. Typically, the goal is improving the NER performance of one of the languages (the primary language) using the other assisting languages. We show that the divergence in the tag distributions of the common named entities between the primary and assisting languages can reduce the effectiveness of multilingual learning. To alleviate this problem, we propose a metric based on symmetric KL divergence to filter out the highly divergent training instances in the assisting language. We empirically show that our data selection strategy improves NER performance in many languages, including those with very limited training data.

1 Introduction

Neural NER trains a deep neural network for the NER task and has become quite popular as they minimize the need for hand-crafted features and, learn feature representations from the training data itself. Recently, multilingual learning has been shown to benefit Neural NER in a resource-rich language setting (Gillick et al., 2016; Yang et al., 2017). Multilingual learning aims to improve the NER performance on the language under consideration (primary language) by adding training data from one or more assisting languages. The neural network is trained on the combined data of the primary (D_P) and the assisting languages (D_A). The neural network has a combination of language-dependent and language-independent layers, and, the network learns better cross-lingual features via these language-independent layers.

Existing approaches add all training sentences from the assisting language to the primary language and train the neural network on the combined data. However, data from assisting languages can introduce a drift in the tag distribution for named entities, since the common named entities from the two languages may have vastly divergent tag distributions. For example, the entity *China* appears in training split of Spanish (primary) and English (assisting) (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003) with the corresponding tag frequencies, Spanish = $\{$ *Loc* : 20, *Org* : 49, *Misc* : 1 $\}$ and English = $\{$ *Loc* : 91, *Org* : 7 $\}$. By adding English data to Spanish, the tag distribution of *China* is skewed towards *Location* entity in Spanish. This leads to a drop in named entity recognition performance. In this work, we address this problem of drift in tag distribution owing to adding training data from a supporting language.

The problem is similar to the problem of data selection for domain adaptation of various NLP tasks, except that additional complexity is introduced due to the multilingual nature of the learning task. For domain adaptation in various NLP tasks, several approaches have been proposed to address drift in data distribution (Moore and Lewis, 2010; Axelrod et al., 2011; Ruder and Plank, 2017). For instance, in machine translation, sentences from out-of-domain data are selected based on a suitably defined metric (Moore and Lewis, 2010; Axelrod et al., 2011). The metric attempts to capture similarity of the out-of-domain sentences with the in-domain data. Out-of-domain sentences most similar to the in-domain data are added.

Like the domain adaptation techniques summarized above, we propose to judiciously add sentences from the assisting language to the primary language data based on the divergence between the tag distributions of named entities in the train-

[*] This work began when the second author was a research scholar at IIT Bombay

Language	Source	Train (#Tokens)	Test (#Tokens)	Word Embeddings
English	Tjong Kim Sang and De Meulder (2003)	204,567	46,666	Dhillon et al. (2015) (Spectral embeddings)
Spanish	Tjong Kim Sang (2002)	264,715	51,533	
Dutch	Tjong Kim Sang (2002)	202,931	68,994	
Italian	Speranza (2009)	149,651	86,420	
German	Faruqui and Padó (2010)	74,907	20,696	
Hindi	Lalitha Devi et al. (2014)	81,817	23,696	Bojanowski et al. (2017) (fastText embeddings)
Marathi	In-house	71,299	36,581	
Tamil	Lalitha Devi et al. (2014)	66,143	18,646	
Bengali	Lalitha Devi et al. (2014)	34,387	7,614	
Malayalam	Lalitha Devi et al. (2014)	26,295	8,275	

Table 1: Dataset Statistics

ing instances. Adding assisting language sentences with lower divergence reduces the possibility of entity drift enabling the multilingual model to learn better cross-lingual features.

Following are the contributions of the paper: (a) We present a simple approach to select assisting language sentences based on symmetric KL-Divergence of overlapping entities (b) We demonstrate the benefits of multilingual Neural NER on low-resource languages. We compare the proposed data selection approach with monolingual Neural NER system, and the multilingual Neural NER system trained using all assisting language sentences. To the best of our knowledge, ours is the first work for judiciously selecting a subset of sentences from an assisting language for multilingual Neural NER.

2 Judicious Selection of Assisting Language Sentences

For every assisting language sentence, we calculate the sentence score based on the average symmetric KL-Divergence score of overlapping entities present in that sentence. By overlapping entities, we mean entities whose surface form appears in both the languages' training data. The symmetric KL-Divergence $SKL(x)$, of a named entity x, is defined as follows,

$$SKL(x) = \big[\, KL(\, P_p(x) \,||\, P_a(x)\,) + KL(\, P_a(x) \,||\, P_p(x)\,)\,\big]/2 \tag{1}$$

where $P_p(x)$ and $P_a(x)$ are the probability distributions for entity x in the primary (p) and the assisting (a) languages respectively. KL refers to the standard KL-Divergence score between the two probability distributions.

KL-Divergence calculates the distance between the two probability distributions. Lower the KL-Divergence score, higher is the tag agreement for an entity in both the languages thereby, reducing the possibility of entity drift in multilingual learning. Assisting language sentences with the sentence score below a threshold value are added to the primary language data for multilingual learning. If an assisting language sentence contains no overlapping entities, the corresponding sentence score is zero resulting in its selection.

Network Architecture

Several deep learning models (Collobert et al., 2011; Ma and Hovy, 2016; Murthy and Bhattacharyya, 2016; Lample et al., 2016; Yang et al., 2017) have been proposed for monolingual NER in the literature. Apart from the model by Collobert et al. (2011), remaining approaches extract sub-word features using either Convolution Neural Networks (CNNs) or Bi-LSTMs. The proposed data selection strategy for multilingual Neural NER can be used with any of the existing models. We choose the model by Murthy and Bhattacharyya (2016)[1] in our experiments.

Multilingual Learning

We consider two parameter sharing configurations for multilingual learning (i) sub-word feature extractors shared across languages (Yang et al., 2017) (*Sub-word*) (ii) the entire network trained in a language independent way (*All*). As Murthy and Bhattacharyya (2016) use CNNs to extract sub-word features, only the character-level CNNs are shared for the *Sub-word* configuration.

[1] The code is available here: https://github.com/murthyrudra/NeuralNER

Primary	**Assisting**	**Layers**	**Data Selection**		**Primary**	**Assisting**	**Layers**	**Data Selection**	
Language	**Language**	**Shared**	**All**	**SKL**	**Language**	**Language**	**Shared**	**All**	**SKL**
German	Monolingual	None	87.64	-	**Italian**	Monolingual	None	75.98	-
	English	All	89.08	**89.46**		English	All	76.22	**76.91**†
		Sub-word	88.76	**89.10**			Sub-word	79.44	79.44
	Spanish	All	89.02	**91.61**†		Spanish	All	74.94	**76.92**†
		Sub-word	88.37	**89.10**†			Sub-word	76.99	**77.45**†
	Dutch	All	89.66	**90.85**†		Dutch	All	75.59	**77.29**†
		Sub-word	89.94	**90.11**			Sub-word	77.38	**77.56**

Table 2: F-Score for German and Italian Test data using Monolingual and Multilingual learning strategies. † indicates that the *SKL* results are statistically significant compared to adding all assisting language data with p-value < 0.05 using two-sided Welch t-test.

3 Experimental Setup

In this section we list the datasets used and the network configurations used in our experiments.

3.1 Datasets

The Table 1 lists the datasets used in our experiments along with pre-trained word embeddings used and other dataset statistics. For German NER, we use *ep-96-04-16.conll* to create train and development splits, and use *ep-96-04-15.conll* as test split. As Italian has a different tag set compared to English, Spanish and Dutch, we do not share output layer for *All* configuration in multilingual experiments involving Italian. Even though the languages considered are resource-rich languages, we consider German and Italian as primary languages due to their relatively lower number of train tokens. The German NER data followed *IO* notation and for all experiments involving German, we converted other language data to *IO* notation. Similarly, the Italian NER data followed *IOBES* notation and for all experiments involving Italian, we converted other language data to *IOBES* notation.

For low-resource language setup, we consider the following Indian languages: Hindi, Marathi[2], Bengali, Tamil and Malayalam. Except for Hindi all are low-resource languages. We consider only *Person, Location* and *Organization* tags. Though the scripts of these languages are different, they share the same set of phonemes making script mapping across languages easier. We convert Tamil, Bengali and Malayalam data to the Devanagari script using the *Indic NLP library*[3] (Kunchukuttan et al., 2015) thereby, allowing sharing of sub-word features across the Indian languages. For Indian languages, the annotated data followed the *IOB* format.

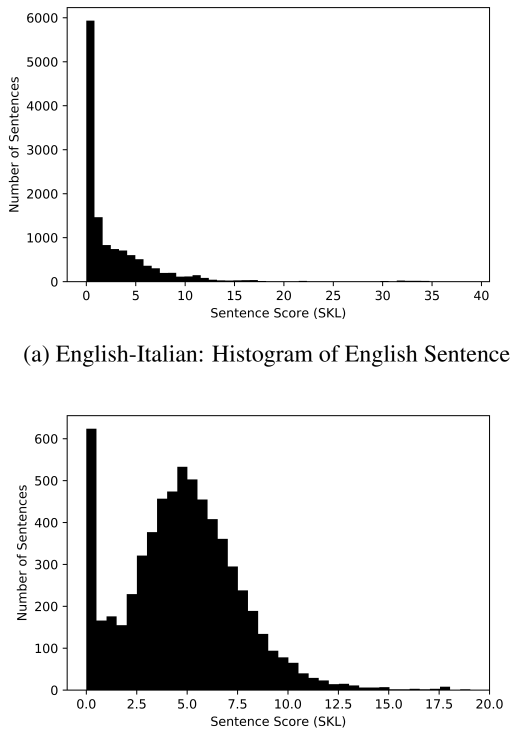
3.2 Network Hyper-parameters

With the exception of English, Spanish and Dutch, remaining language datasets did not have official train and development splits provided. We randomly select 70% of the train split for training the model and remaining as development split. The threshold for sentence score *SKL*, is selected based on cross-validation for every language pair. The dimensions of the Bi-LSTM hidden layer are 200 and 400 for the monolingual and multilingual experiments respectively. We extract 20 features per convolution filter, with width varying from 1 to 9. The initial learning rate is 0.4 and multiplied by 0.7 when validation error increases. The training is stopped when the learning rate drops below 0.002. We assign a weight of 0.1 to assisting language sentences and oversample primary language sentences to match the assisting language sentence count in all multilingual experiments.

For European languages, we have performed hyper-parameter tuning for both the monolingual and multilingual learning (with all assisting language sentences) configurations. The best hyper-parameter values for the language pair involved were observed to be within similar range. Hence, we chose the same set of hyper-parameter values for all languages.

[2] Data is available here: `http://www.cfilt.iitb.ac.in/ner/annotated_corpus/`

[3] `https://github.com/anoopkunchukuttan/indic_nlp_library`



(a) English-Italian: Histogram of English Sentences

(b) Spanish-Italian: Histogram of Spanish Sentences

Figure 1: Histogram of assisting language sentences ranked by their sentence scores

4 Results

We now present the results on both resource-rich and resource-poor languages.

4.1 Resource-Rich Languages

Table 2 presents the results for German and Italian NER. We consistently observe improvements for German and Italian NER using our data selection strategy, irrespective of whether only sub-word features are shared (*Sub-word*) or the entire network (*All*) is shared across languages.

Adding all Spanish/Dutch sentences to Italian data leads to drop in Italian NER performance when all layers are shared. Label drift from overlapping entities is one of the reasons for the poor results. This can be observed by comparing the histograms of English and Spanish sentences ranked by the SKL scores for Italian multilingual learning (Figure 1). Most English sentences have lower SKL scores indicating higher tag agreement for overlapping entities and lower drift in tag distribution. Hence, adding all English sentences improves Italian NER accuracy. In contrast, most Spanish sentences have larger SKL scores and adding these sentences adversely impacts Italian NER performance. By judiciously selecting assisting language sentences, we eliminate sentences which are responsible for drift occurring during multilingual learning.

To understand how overlapping entities impact the NER performance, we study the statistics of overlapping named entities between Italian-English and Italian-Spanish pairs. 911 and 916 unique entities out of 4061 unique Italian entities appear in the English and Spanish data respectively. We had hypothesized that entities with divergent tag distribution are responsible for hindering the performance in multilingual learning. If we sort the common entities based on their SKL divergence value. We observe that 484 out of 911 common entities in English and 535 out of 916 common entities in Spanish have an SKL score greater than 1.0. 162 out of 484 common entities in English-Italian data having SKL divergence value greater than 1.0 also appear more than 10 times in the English corpus. Similarly, 123 out of 535 common entities in Spanish-Italian data having SKL divergence value greater than 1.0 also appear more than 10 times in the Spanish corpus. However, these common 162 entities have a combined frequency of 12893 in English, meanwhile the 123 common entities have a combined frequency of 34945 in Spanish. To summarize, although the number of overlapping entities is comparable in English and Spanish sentences, entities with larger SKL divergence score appears more frequently in Spanish sentences compared to English sentences. As a consequence, adding all Spanish sentences leads to significant drop in Italian NER performance which is not the case when all English sentences are added.

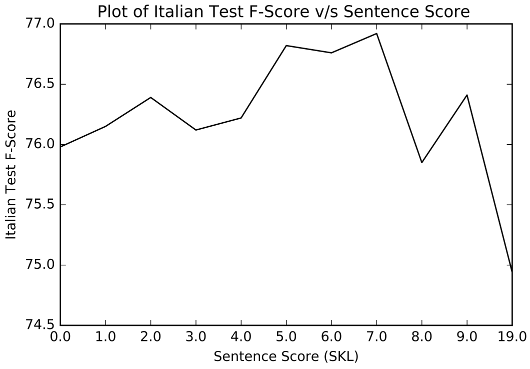


Figure 2: Spanish-Italian Multilingual Learning: Influence of Sentence score (SKL) on Italian NER

Primary Language	Assisting Language									
	Hindi		Marathi		Bengali		Malayalam		Tamil	
	ALL	SKL	ALL	SKL	ALL	SKL	ALL	SKL	ALL	SKL
Hindi	64.93	-	59.30	**66.33**	58.51	59.30	58.21	59.13	56.75	58.75
Marathi	54.46	**63.30**	61.46	-	47.67	61.28	50.13	61.05	59.04	58.62
Bengali	44.34	**51.05**†	41.28	**55.77**†	40.02	-	48.79	**49.84**†	38.38	**44.14**†
Malayalam	59.74	**64.00**†	65.88	**66.42**†	58.01	**63.65**†	57.94	-	58.25	**58.92**
Tamil	60.13	**61.51**†	60.54	**61.67**†	53.27	**60.32**†	61.03	**61.45**	53.13	-

Table 3: Test set F-Score from monolingual and multilingual learning on Indian languages. Result from monolingual training on the primary language is underlined. † indicates *SKL* results statistically significant compared to adding all assisting language data with p-value < 0.05 using two-sided Welch t-test.

4.2 Resource-Poor Languages

As Indian languages exhibit high lexical overlap (Kunchukuttan and Bhattacharyya, 2016) and syntactic relatedness (V Subbãrão, 2012), we share all layers of the network across languages. Table 3 presents the results. Bengali, Malayalam, and Tamil (low-resource languages) benefits from our data selection strategy. Hindi and Marathi NER performance improves when the other is used as assisting language.

Bengali, Malayalam, and Tamil have weaker baselines compared to Hindi and Marathi, and are benefited from our approach irrespective of the assisting language chosen. However, Hindi and Marathi are not benefited from multilingual learning with Bengali, Malayalam and Tamil. Malayalam and Tamil being morphologically rich have low entity overlap (surface level) with Hindi and Marathi. As a result, only 2-3% of Malayalam and Tamil sentences are eliminated from our approach, leading to no gains from multilingual learning. Hindi and Marathi are negatively impacted by noisy Bengali data. Bengali has less training sentences compared to other languages and, choosing a low SKL threshold results in selecting very few Bengali sentences for multilingual learning.

4.3 Influence of SKL Threshold

Here, we study the influence of SKL score threshold on the NER performance. We run experiments for Italian NER by adding Spanish training sentences and sharing all layers except for output layer across languages. We vary the threshold value from 1.0 to 9.0 in steps of 1, and select sentences with score less than the threshold. A threshold of 0.0 indicates monolingual training and threshold greater than 9.0 indicates all assisting language sentences considered. The plot of Italian test F-Score against SKL score is shown in the Figure 2. Italian test F-Score increases initially as we add more and more Spanish sentences and then drops due to influence of drift becoming significant. Finding the right SKL threshold is important, hence we use a validation set to tune the SKL threshold.

5 Conclusion

In this paper, we address the problem of divergence in tag distribution between primary and assisting languages for multilingual Neural NER. We show that filtering out the assisting language sentences exhibiting significant divergence in the tag distribution can improve NER accuracy. We propose to use the symmetric KL-Divergence metric to measure the tag distribution divergence. We observe consistent improvements in multilingual Neural NER performance using our data selection strategy. The strategy shows benefits for extremely low resource primary languages too.

This problem of drift in data distribution may not be unique to multilingual NER, and we plan to study the influence of data selection for multilingual learning on other NLP tasks like sentiment analysis, question answering, neural machine translation, *etc*. We also plan to explore more metrics for multilingual learning, specifically for morphologically rich languages.

Acknowledgements

We thank Gajanan Rane and Geetanjali Rane for annotating the Marathi data, which was created as part of the CLIA project.

References

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, Edinburgh, United Kingdom.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*.

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*.

Paramveer S. Dhillon, Dean P. Foster, and Lyle H. Ungar. 2015. Eigenwords: Spectral word embeddings. *Journal of Machine Learning Research*.

Manaal Faruqui and Sebastian Padó. 2010. Training and evaluating a German Named Entity Recognizer with semantic generalization. In *Proceedings of KONVENS*.

Dan Gillick, Cliff Brunk, Oriol Vinyals, and Amarnag Subramanya. 2016. Multilingual language processing from bytes. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics*, San Diego, US.

Anoop Kunchukuttan and Pushpak Bhattacharyya. 2016. Orthographic syllable as basic unit for SMT between related languages. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Texas, USA.

Anoop Kunchukuttan, Ratish Puduppully, and Pushpak Bhattacharyya. 2015. Brahmi-net: A transliteration and script conversion system for languages of the Indian subcontinent. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, Colorado, USA.

Shobha Lalitha Devi, Pattabhi RK Rao, Malarkodi C.S, and R Vijay Sundar Ram. 2014. Indian language NER annotated FIRE 2014 corpus (FIRE 2014 NER Corpus). In *Named-Entity Recognition Indian Languages FIRE 2014 Evaluation Track*.

Guillaume Lample, Miguel Ballesteros, Kazuya Kawakami, Sandeep Subramanian, and Chris Dyer. 2016. Neural architectures for Named Entity Recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics*, San Diego, US.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via Bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Berlin, Germany.

Robert C. Moore and William Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden.

Rudra V. Murthy and Pushpak Bhattacharyya. 2016. A deep learning solution to Named Entity Recognition. In *CICLing*, Konya, Turkey.

Sebastian Ruder and Barbara Plank. 2017. Learning to select data for transfer learning with Bayesian Optimization. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark.

Manuela Speranza. 2009. The Named Entity Recognition task at EVALITA 2009. In *Proceedings of the Workshop Evalita*.

Erik F. Tjong Kim Sang. 2002. Introduction to the conll-2002 shared task: Language-independent Named Entity Recognition. In *Proceedings of the 6th Conference on Natural Language Learning at COLING-02*, Taipei, Taiwan.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent Named Entity Recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, Edmonton, Canada.

Kārumūri V Subbārāo. 2012. *South Asian Languages: A Syntactic Typology*. Cambridge University Press.

Zhilin Yang, Ruslan Salakhutdinov, and William Cohen. 2017. Multi-task cross-lingual sequence tagging from scratch. In *International Conference on Learning Representations*, Toulon, France.