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

Association for Computational Linguistics (ACL) 2018

Rudra Murthy V
CFILT Lab,
Indian Institute Of
Technology Bombay
rudra@cse.iitb.ac.in

Anoop Kunchukuttan
Microsoft AI & Research,
Hyderabad, India
ankunchu@microsoft.com

Pushpak Bhattacharyya
CFILT Lab,
Indian Institute Of
Technology Bombay
pb@cse.iitb.ac.in

Outline

Table of Contents

Problem Statement

Judiciously select labeled data from assisting language to improve the NER performance in the primary language for multilingual learning

Table of Contents

Why need to judiciously select data from assisting language?

- Many language have less named entity annotated data
- Several approaches have explored use of data from one or more languages (assisting languages) [Gillick et al. [2016], Yang et al. [2017]]
- However, annotated data from assisting languages might negatively influence the performance on the primary language

What can go wrong in multilingual learning for NER?

- Vocabulary
    - False Friends
    - Dataset Characteristics
- Sub-word features
    - Capitalization feature
        - Religions, Languages, Nationalities, *etc.* uppercase in English but not in Spanish
- Contextual features
    - Different Word Order

What can go wrong in multilingual learning for NER?

- Vocabulary
  - False Friends
  - Dataset Characteristics
- Sub-word features
  - Capitalization feature
    - Religions, Languages, Nationalities, *etc.* uppercase in English but not in Spanish
- Contextual features
  - Different Word Order

Why need to judiciously select data from assisting language?

- Vocabulary
  - False Friends
  - Dataset Characteristics

English

Word	Per	Loc	Org	Misc
China	-	91	7	-
France	-	123	4	1
Reuters	-	40	18	-

⋮

Spanish

Word	Per	Loc	Org	Misc
China	-	20	49	1
France	-	-	10	-
Reuters	-	3	1	-

⋮

Table of Contents

Related Work

Axelrod et al. [2011] Moore and Lewis [2010]	• Select **sentences** from **general domain** data **most similar** to **in-domain data** • Used **language model** to **measure similarity** of **general domain data** with the **in-domain training data**
Ruder and Plank [2017]	• **Learn to weigh** various **data selection measures** using **Bayesian Optimization**
Zhao et al. [2018]	• Select **assisting data** for **multi-task domain adaptation** • **Assisting language sentences** with **highest log likelihood value** were selected
Ponti et al. [2018]	• Measure **cross-lingual syntactic variation** considering both **morphological** and **structural properties** • **Selecting** a **assisting language** with a **lower degree** of **anisomorphism** is crucial for knowledge transfer

Table 1: Literature most relevant to our work

Table of Contents

Proposed Approach

Select sentences based on the agreement in tag distribution of common entities

Goal: Improve Spanish NER performance by adding English NER annotated data

Proposed Approach

Select sentences based on the agreement in tag distribution of common entities

Goal: Improve Spanish NER performance by adding English NER annotated data

English

Word	Per	Loc	Org	Misc
China	-	91	7	-
France	-	123	4	1
Reuters	-	40	18	-

⋮

Spanish

Word	Per	Loc	Org	Misc
China	-	20	49	1
France	-	-	10	-
Reuters	-	3	1	-

⋮

Proposed Approach

Select sentences based on the agreement in tag distribution of common entities

Goal: Improve Spanish NER performance by adding English NER annotated data

English

Word	Per	Loc	Org	Misc
China	-	91	7	-
France	-	123	4	1
Reuters	-	40	18	-
⋮				

Spanish

Word	Per	Loc	Org	Misc
China	-	20	49	1
France	-	-	10	-
Reuters	-	3	1	-
⋮				

Proposed Approach

Select sentences based on the agreement in tag distribution of common entities

Goal: Improve Spanish NER performance by adding English NER annotated data

English

Word	Per	Loc	Org	Misc
China	-	91	7	-
France	-	123	4	1

Spanish

Word	Per	Loc	Org	Misc
China	-	20	49	1
[illegible]	[illegible]	-	10	-

Select English sentences containing entities with similar tag distribution

Select sentences based on the agreement in tag distribution of common entities

Goal: Improve Spanish NER performance by adding English NER annotated data

English

Word	Per	Loc	Org	Misc
China	-	91	7	-
France	-	123	4	1

Spanish

Word	Per	Loc	Org	Misc
China	-	20	49	1
France	-	-	10	-

Use Symmetric Kl-Divergence to calculate the tag disagreement for common entities between English and Spanish

Proposed Approach

Select sentences based on the agreement in tag distribution of common entities

Goal: Improve Spanish NER performance by adding English NER annotated data

	English				Spanish						
Word	Per	Loc	Org	Misc	Per	Loc	Org	Misc	KL(Eng‖Esp)	KL(Esp‖Eng)	SKL
China	-	91	7	-	-	20	49	1	0.9314	1.3972	2.3287
France	-	123	- 4	1	-	-	10	-	10.4332	2.6388	13.0721
Reuters	-	**40**	**18**	-	-	**3**	**1**	-	**0.1088**	**0.1531**	**0.2620**

Proposed Approach

for every sentence X, in assisting language **do**
  Score(X) $\leftarrow$ 0.0
  for every word x_i, in sentence X **do**
    if word x_i appears in primary language **then**
      $\text{SKL}(x_i) \leftarrow \left[KL(P_p(x_i)||P_a(x_i)) + KL(P_a(x)||P_p(x))\right]/2$ {$P_p(x_i)$ and $P_a(x_i)$ are tag distributions of x_i in primary and assisting languages}
      Score(X) $\leftarrow$ Score(X) + SKL(x_i)
    end if
  end for
end for

Add assisting language sentences with sentence score Score(X) less than a threshold θ to the primary language data

Table of Contents

Dataset Statistics

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 2: Dataset Statistics

Network Details

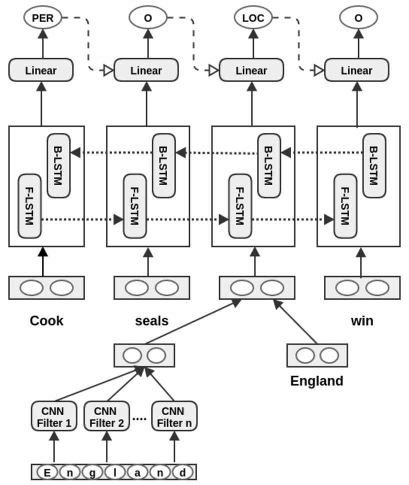

Figure 1: Architecture of the Neural Network (Murthy and Bhattacharyya [2016])

Parameter sharing configurations considered

- Sub-word feature extractors shared across languages (Yang et al. [2017])
- Neural network trained in language independent way

Results

Primary Language	Assisting Language	Layers Shared	Data Selection All	Data Selection SKL	Primary Language	Assisting Language	Layers Shared	Data Selection All	Data Selection 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 3: 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 $p\text{-value} < 0.05$ using two-sided Welch t-test.

Histogram of assisting language sentences ranked by their sentence scores

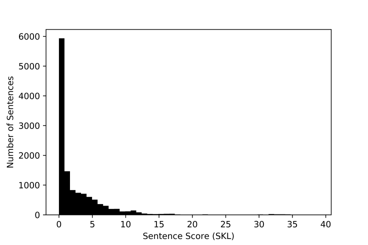

Figure 2: English-Italian: Histogram of English Sentences

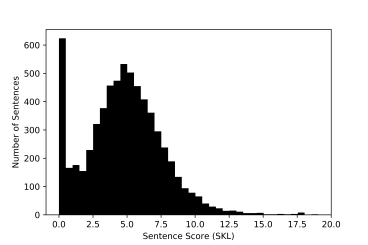

Figure 3: Spanish-Italian: Histogram of Spanish Sentences

Analysis: European Languages

- Adding **all Spanish/Dutch** sentences to **Italian** data, leads to **drop in Italian NER** performance
- **Label drift** from **overlapping entities** is one of the reasons for the poor results
- We compare the **histograms** of **English** and **Spanish sentences ranked** by the **SKL scores** for **Italian** multilingual learning
- Similar pattern is observed in the case of Dutch sentences

Results

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

Table 4: 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.

Analysis: Indian Languages

- **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**
- **Hindi** and **Marathi** are **not benefited** from multilingual learning with **Bengali, Malayalam** and **Tamil**

Influence of SKL Threshold

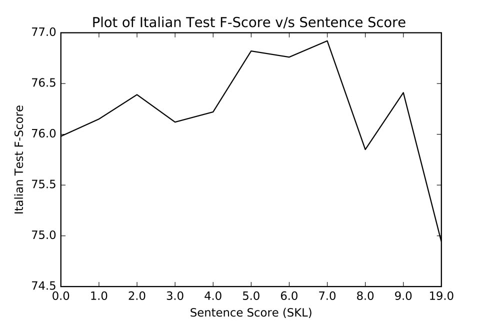

Figure 4: Spanish-Italian Multilingual Learning: Influence of Sentence score (SKL) on Italian NER

Analysis: Influence of SKL Threshold

- Train for Italian NER by adding Spanish training sentences and sharing all layers except for output layer across languages
- We vary the threshold value from 0.0 to 9.0 in steps of 1
- **Italian** test **F-Score increases initially** as we add more and more Spanish sentences and **then drops** due to **influence** of **drift becoming significant**

Conclusion And Future Work

- 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
- A more principled approach for data selection would be exploring the work of Ponti et al. [2018]
- We plan to study the influence of data selection for multilingual learning on other NLP tasks like sentiment analysis, question answering, neural machine translation
- We also plan to explore more metrics for multilingual learning, specifically for morphologically rich languages

Thank You

References I

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

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

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

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

References II

Dan Gillick, Cliff Brunk, Oriol Vinyals, and Amarnag Subramanya. 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, 2016.

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

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

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

Edoardo Maria Ponti, Roi Reichart, Anna Korhonen, and Ivan Vulic. Isomorphic transfer of syntactic structures in cross-lingual nlp. In *ACL 2018*, 2018.

Sebastian Ruder and Barbara Plank. 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, 2017.

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

References IV

Erik F. Tjong Kim Sang. 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, 2002.

Erik F. Tjong Kim Sang and Fien De Meulder. 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, 2003.

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

Huasha Zhao, Yi Yang, Qiong Zhang, and Luo Si. Improve neural entity recognition via multi-task data selection and constrained decoding. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. Association for Computational Linguistics, 2018.

Why need to judiciously select data from assisting language?

- Vocabulary
    - False Friends
    - Dataset Characteristics
- Sub-word features
    - Capitalization feature
        - Religions, Languages, Nationalities, *etc.* uppercase in English but not in Spanish
- Contextual features
    - Different Word Order
        - I am going to Washington

में	वाशिंगटन	जा	रहा	हूँ
mein	washington	jaa	raha	hun
me	washington	going to		