

Dedicated to my Family, Friends and Guruji. . .

Certificate

This is to certify that this thesis entitled “Better Sequence Labeling for Low Resource Languages Using Attention and Transfer Learning” submitted by “Rajesh Kumar Mundotiya” (Roll No.: 16071001) for the award of the degree of doctor of philosophy to the Indian Institute of Technology (Banaras Hindu University), Varanasi, is a record of bona fide research works carried out by him under my direct supervision and guidance and it has not been submitted elsewhere for a degree. It is further certified that the student has fulfilled all the requirements of Comprehensive Examination, Candidacy and SOTA for the award of Ph.D. Degree.



Signature of Supervisor

Anil Kumar Singh

Associate Professor

Department of Computer Science and Engineering

Indian Institute of Technology (BHU)

Varanasi - 221005, India

Declaration

I, Rajesh Kumar Mundotiya, certify that the work embodied in this thesis is my own bonafide work and carried out by me under the supervision of Anil Kumar Singh from July-2016 to February 2022, at the Department of Computer Science and Engineering, Indian Institute of Technology (BHU), Varanasi. The matter embodied in this thesis has not been submitted for the award of any other degree/diploma. I declare that I have faithfully acknowledged and given credits to the research workers wherever their works have been cited in my work in this thesis. I further declare that I have not willfully copied any other's work, paragraphs, text, data, results, etc., reported in journals, books, magazines, reports dissertations, theses, etc., or available at websites and have not included them in this thesis and have not cited as my own work.



Date: 10/02/2022

Place: Varanasi

Signature of Student
(Rajesh Kumar Mundotiya)

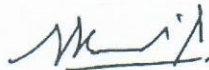
Certificate by the Supervisor

It is certified that the above statement made by the student is correct to the best of my/our knowledge.



Signature of Supervisor

सह आचार्य/Associate Professor
(Anil Kumar Singh) B.Sc. & Engg.
संगणक विज्ञान एवं अभियांत्रिकी विभाग
भारतीय प्रौद्योगिकी संस्थान / Indian Institute of Technology
(बनारस हिन्दू यूनिवर्सिटी) / (Banaras Hindu University)
वाराणसी / Varanasi-221005



Signature of Head of Department

आचार्य व विभागाध्यक्ष
(Prof. Sanjay Kumar Singh)
Professor & Head

संगणक विज्ञान एवं अभियांत्रिकी विभाग
Department of Computer Sc. & Engg.
भारतीय प्रौद्योगिकी संस्थान

Indian Institute of Technology
(बनारस हिन्दू यूनिवर्सिटी)

(Banaras Hindu University)

वाराणसी-221005 / Varanasi-221005

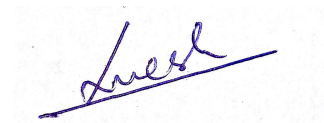
Copyright Transfer Certificate

Title of the Thesis: **Better Sequence Labeling for Low Resource Languages Using Attention and Transfer Learning**

Name of Student: **Rajesh Kumar Mundotiya**

Copyright Transfer

The undersigned hereby assigns to the Indian Institute of Technology (Banaras Hindu University), Varanasi, all rights under copyright that may exist in and for the above thesis submitted for the award of the Doctor of Philosophy.



Date: 10/02/2022

Place: Varanasi

Signature of Student
(Rajesh Kumar Mundotiya)

Note: However, the author may reproduce or authorize others to reproduce material extracted verbatim from the thesis or derivative of the thesis for author's personal use provided that the source and the Institute's copyright notice are indicated.

Acknowledgments

I will always remember my doctoral years as an intensive period of my life, full of both amazing and tough experiences. I could have never finished my thesis without the inspiration and support and encouragement of many excellent people including my well-wishers, my friends, colleagues and various institutions. At the end of my thesis, it is a pleasant task to express my thanks to all those who contributed in many ways to the success of this study and made it an unforgettable experience for me.

First of all, I would like to express my thanks to my supervisor Dr Anil Kumar Singh for his excellent guidance, exceptional encouragement, and his willingness to help. Getting to know him is one of my biggest fortunes and the most precious part of my PhD study. I cannot express my gratitude enough for his support and extraordinary care over the years. I could not have imagined having a better advisor and mentor for my PhD study.

I especially thank the members of my Research Progress Evaluation Committee, Dr Ravindranath Chowdary C, Dr Sukhada and Dr Anil Kumar Singh for their invaluable suggestions regarding the thesis, their insightful comments and encouragement, but also for the hard question which incited me to widen my research from various perspectives.

I would like to express my sincere thanks to Prof. Sanjay Kumar Singh, Head of Computer Science and Engineering Department, for his kindness and valuable support in carrying out the research. I express my sincere thanks to the faculty members, Prof. K. K. Shukla, Dr S. Pal, Dr H. P. Gupta and staff of the department.

I am greatly indebted to Dr Swasti Mishra, Department of Humanistic Studies, IIT (BHU), Dr Praveen Gatla, Department of Linguistics, BHU and Dr Manish Kumar Singh, Cerence Inc. for their kind support, guidelines, and got benefited from their vast knowledge of NLP.

Many thanks to all my friends and colleagues at Computer Science and Engineering, for being such friendly people with whom it is such a pleasure to work. I cannot forget all the invaluable time we spent discussing research and life, especially during our after-lunch coffee expeditions.

Special thanks go to all the inhabitants of office CSE. I wouldn't have made it without your encouragement and support.

Most importantly, my deepest gratitude is for my family for their constant support, inspiration, guidance, and sacrifices. My parents were a constant source of motivation and

inspiration. Their affection and guidance were instrumental in me choosing Engineering and eventually continuing on to my PhD. I would like to thank my mother, who was also my childhood teacher and is always there to stand by my side.

Last but not the least, I would like to give special thanks to my elder brother Mr Mukesh Kumar for supporting me unconditionally throughout my study. I sincerely thank all of them who contributed in helping me to see the light at the end of every scary tunnel during my PhD.

- Rajesh Kumar Mundotiya

Preface

Since the popular spread of the Internet, natural language content in the form of suggestions, opinions, news, and tweets, apart from formal documents, has proliferated at an increasing rate. This has helped in the development of natural language processing (NLP) tools for diverse languages in a variety of domains. Language analysis and understanding is an application of NLP, where machines interpret linguistic information and many other facets of language as humans do. Text classification, syntactic, semantic analysis and entity recognition or linking are examples of problems where errorless decisions are predicted based on language understanding.

Sequence labeling plays a vital role in solving numerous NLP applications such as Machine Translation and Information Extraction etc. In sequence labeling, a category or label is assigned to each member of an instance, where instance denotes a sequence, for example, part of speech (POS) tagging, chunking and named entity recognition (NER). POS tagging assigns a sequence of grammatical categories to the given sentence and Chunking groups them into ‘chunks’ or what can be called minimal phrases. NER is another sequence labeling problem, which marks proper nouns and other named entities such as Location, Person, Organization, Disease etc. Existing studies have shown that convolutional neural network, recurrent neural network and its variants and conditional random field are the key components for deep learning models. They, along with pre-trained embedding, provide good results for these sequence labeling problems. However, these solutions do not generalize for low resource languages, which are characterized by very little benchmarking seed amount of (un)-annotated data. Transfer learning and multilingual learning do, however, enable gains in performance. Improving the performance of sequence labeling models using existing knowledge is one of the prominent research areas in NLP for low resource languages. This thesis focuses on four aspects of this: proposing deep learning architectures, integrating handcrafted features, performing transfer learning, and providing baseline systems using benchmark seed datasets for some low resource languages.

Bhojpuri, Maithili and Magahi are low resource languages and widely spoken in central north-eastern India, belonging to the Indo-Aryan language family. Creating an annotated corpus for POS tagging and chunking and then building an initial automatic tool for these languages is the first attempt towards the development of language technology tools for these languages. The annotated corpus used to develop POS taggers and chunkers, based on various traditional machine learning algorithms (trigrams ‘n’ tags, conditional random fields (CRF), maximum entropy Markov model and structured support vector machine) and more recent state-of-the-art deep learning model which comprise long short-term memory (LSTM), convolution neural network (CNN) and CRF, named as LSTM-CNN-CRF model, has been used.

A robust deep learning model has been proposed that leverages character level information to deal with out-of-vocabulary and an attention mechanism to capture the interdependence of input words known as Self-Attention-based Hierarchical Bi-LSTM CRF (SAHBiLC) model. Model transfer is one of the ways of transfer learning by which the learned knowledge (parameters of deep learning model) can transfer from one model to another model; hence initially, the model has been trained on Hindi and then transferred to Bhojpuri, Maithili and Magahi: this setting being referred to as the fine-tuned version of SAHBiLC model, named as Fine-SAHBiLC. SAHBiLC and Fine-SAHBiLC outperform previous results for Bhojpuri and Maithili, Magahi, respectively, for both tasks. This shows that finetuning is effective for such languages which have complex morphology and lesser training data.

Despite the success of the proposed model, we compare the results of CRF and LSTM-CNN-CRF for the NER datasets of these languages since the number of named entities are relatively few in the data, that is, the datasets of these languages are highly imbalanced. Here, the deep learning baseline provides a better result for Magahi only due to fewer intermediate entities (tags). We observe that the obtained results are consistent with the number of named entities in the datasets rather than with the total size of the dataset in the number of tokens.

Traditional machine learning algorithms require feature engineering. Linguistic information such as affixes, lemma, adjacent words is commonly used for these sequence labeling problems. Recent deep learning models combine the forward and backward word information captured by the recurrent neural network or its variants for POS tagging. However, it assumes an identical magnitude of words of a sentence to capture sequential information. The information of contextual words to the current word plays a vital role in capturing the

non-continuous relationship. Based on the same assumption, we have proposed a Monotonic chunk-wise attention model with CNN-GRU-Softmax (MCCGS), a deep learning model that captures this essential information. This architecture consists of Input Encoder (IE), which encodes word and character-level information, Contextual Encoder (CE), which assigns weightage to adjacent word and Disambiguator (D), which resolves intra-label dependencies as core components. Later on, this proposed model gets integrated with morphological information as an external feature. Tense, Case, Gender, Number, Person and Lemma considered external features integrated into the core components of MCCGS architecture as MCCGS-IE, MCCGS-CE and MCCGS-D. The MCCGS architecture and its variants are validated on datasets for 21 languages from the universal dependency (UD) treebank. Compared to the state-of-the-art position-aware self-attention-based model, the MCCGS model has improved by 0.29% in mean accuracy. The MCCGS model improved further by 2.98% due to unification of features at the Disambiguator component (MCCGS-D).

If we assume that the distribution of source and target data for a language is inequivalent, then modeling over these settings is widely known as domain adaptation. The contrastive training minimizes the difference of inequivalent distribution. Contrastive training has been tried as a robust approach that captures the essential features during model training, and based on this, contrastive-MCCGS model architecture has been proposed for POS tagging. It learns optimal features in a low resource regime.

We experimented on the datasets of four domains, Article, Conversation, Disease and Tourism, of the Hindi treebank, Tweet domain from TweepBank, Newswire domain from Penn TreeBank (PTB) and Tweet domain from ARK and compared it with several state-of-the-art models. The CMCCGS model has been further extended to domain adaptation by using single and multi-source domain adaptation to allow fine-tuning. Compared to single-source adaptation, significant improvements have been observed after using multi-source domain adaptation. Since it is multi-source domain adaptation, for experiments based on the deep learning models, the effect of layer freezing was observed. Multi-source domain adaptation provides competitive results over the baselines. Very low resource domains such as Tourism, Disease and tweet domain of TweepBank and ARK have shown improvement in accuracy by Article, Article and Tourism (multi-source), PTB, and PTB and TweepBank (multi-source) as source domain, respectively. However, the Conversation domain has a negative impact on domain adaptation.

Devising deep learning networks that perform well on low resource languages is challenging

since very little labelled data or human annotated data is available for these languages. One of the possible solutions to overcome this challenge is to use multilingual training with accurate parameter transfer. We present a multilingual meta learning based algorithm to learn representations for POS tagging of low resource languages. Meta learning algorithms learn a model from the distribution of tasks that can only adapt to a previously unseen task with a few examples. Two meta-learning algorithms, model agnostic meta learning (MAML) and Reptile, are used to train the model. We conduct extensive experiments on POS tagging using nine languages from two different scripts, out of which the model is trained on eight languages via meta learning and fine-tuned on an unseen low resource language. Data for all nine languages is obtained from the Universal Dependencies Treebank. Empirical results have shown that a model trained via multilingual meta learning can learn cross-lingual features, cross-script features and task-specific features that can easily be transferred into previously unseen low resource languages. Our Approach gains significant improvement over a robust baseline on a low resource language. An experiment for few shot POS tagging verifies the learned representations.




Contents

Certificate	ii
Acknowledgments	v
Preface	vii
Contents	xi
List of Figures	xiv
List of Tables	xvi
Abbreviations	xviii
Symbols	xx
1 Introduction	1
1.1 Research Objectives	4
1.2 Scope	5
1.3 Contributions	6
1.4 Chapterization	7
2 Backgrounds and Literature survey	9
2.1 Low Resource Languages	9
2.2 Traditional Learning Techniques in NLP	11
2.2.1 Trigrams ‘n’ Tags	11
2.2.2 Maximum Entropy Markov Model	11
2.2.3 Conditional Random Fields	12
2.2.4 Structured Support Vector Machine	13
2.3 Deep Learning Models in NLP	14
2.3.1 Convolutional Neural Network	14
2.3.2 Recurrent Neural Network	15
2.3.3 Attention Mechanisms	19

2.4	Transfer Learning	22
2.5	Distributional Vector Representation	24
2.5.1	Context Independent Embedding	25
2.5.2	Context Dependent Embedding	28
2.6	Hand Crafted Features in Neural Network	31
2.7	Literature Survey on Sequence Labeling	31
2.7.1	Sequence Labeling Work on Indian Languages	32
2.7.2	Sequence Labeling Models with Feature Engineering	37
3	Baseline Experiments on Low Resource Languages for Sequence Labeling	41
3.1	Introduction	41
3.2	Contributions of Chapter	44
3.3	Methodology	45
3.4	Experiment-I: POS Tagging and Chunking	48
3.4.1	Dataset Description	48
3.4.2	Machine Learning Strategy	49
3.4.3	Deep Learning Strategy	50
3.4.3.1	Feature transfer	50
3.4.4	Results and Analysis	51
3.4.4.1	Error analysis	58
3.5	Experiment-II: NER	63
3.5.1	Dataset Description	63
3.5.2	Experimental Settings	65
3.5.2.1	CRF model training	65
3.5.2.2	Deep learning model training	66
3.5.3	Results and Analysis	66
3.5.3.1	Effect of epoch in LSTM-CNNs-CRF model	70
3.5.3.2	Error analysis	70
3.6	Summary	74
4	Leveraging Contextual Information for Sequential Labeling	75
4.1	Introduction	75
4.2	Contributions of Chapter	76
4.3	Methodology	77
4.3.1	Input Encoder	77
4.3.2	Contextual Encoder	81
4.3.3	Disambiguator	83
4.4	Inclusion of Morphological Features	84
4.5	Contrastive Training Component	86
4.6	Experiment-I: Feature Integration in MCCGS	88
4.6.1	Dataset Description	90
4.6.2	Experimental Settings	91
4.6.3	Results and Analysis	92

4.6.3.1	Feature inclusion	93
4.6.3.2	Effect of chunk size	96
4.6.3.3	Experiments on Indo-Aryan languages	97
4.6.3.4	Analysis	97
4.7	Experiment-II: Contrastive MCCGS	99
4.7.1	Domain Adaptation	101
4.7.2	Dataset Description	102
4.7.3	Experimental Settings	103
4.7.4	Results and Analysis	103
4.7.4.1	Single-source domain adaptation	105
4.7.4.2	Multi-source domain adaptation	107
4.7.4.3	Effect of layers freezing in domain adaptation	107
4.8	Summary	109
5	Multilingual Learning based Sequence Labeling	111
5.1	Introduction	111
5.2	Contribution of Chapter	113
5.3	Methodology	113
5.3.1	Residual BiLSTM-CRF Model	113
5.3.2	Meta Learning Algorithms	116
5.3.2.1	Model agnostic meta learning	117
5.3.2.2	Reptile	118
5.4	Dataset Description	119
5.5	Experimental Details	119
5.6	Results and Discussion	121
5.6.1	Results for Few Shot POS Tagging	121
5.6.2	Results for Low Resource POS Tagging	122
5.6.3	Ablation Study	124
5.6.4	Analysis	124
5.7	Summary	126
6	Conclusions and Future Scopes	127
6.1	Proposed Methods and Findings	127
6.2	Future Scopes	129
A	List of Publications	132
	Bibliography	134

List of Figures

2.1	CNN over the text [171]	15
2.2	Folded RNN with feedback loop to a input sequence	16
2.3	Unfolded RNN to a input sequence (x) along with its time steps (t)	16
2.4	LSTM block at t^{th} time step [98]	17
2.5	GRU block at t^{th} time step	18
2.6	Seq2seq model with encoder and decoder to translate the input sequence ABC to $WXYZ$	19
2.7	Bahdanau’s attention model to a encoder-decoder based model	21
2.8	Vaswani et al. [242] transformer architecture which comprise the encoder-decoder	22
2.9	Transfer learning taxonomy for NLP	23
2.10	The continuous skip-gram model	27
2.11	The continuous bag-of-words model	27
2.12	ELMo model architecture for pre-training representation	29
2.13	BERT model architecture for pre-training representation	30
3.1	SAHBiLC model architecture for POS Tagging	46
3.2	F-score result of POS tagging for (a) Bhojpuri, (b) Maithili and (c) Magahi	53
3.3	F-score results of Chunk tagging for (a) Bhojpuri and (b) Maithili	54
3.4	F-scores of POS tagging after applying SAHBiLC model and Fine-SAHBiLC model on (a) Bhojpuri and (b) Maithili	56
3.5	Most affected POS tags after applying SAHBiLC model and Fine-SAHBiLC model on (a) Bhojpuri, (b) Maithili and (c) Magahi	56
3.6	The F-scores of chunk tagging after applying SAHBiLC model and Fine-SAHBiLC model on (a) Bhojpuri and (b) Maithili	57
3.7	Effect of epochs on LSTM-CNNs-CRF model’s accuracy on Bhojpuri, Maithili and Magahi.	70
3.8	Confusion matrix for Bhojpuri for the LSTM-CNNs-CRF (above) and CRF (below) models; The  refers to correctly prediction	71
3.9	Confusion matrix for Maithili for the LSTM-CNNs-CRF (above) and CRF (below) models; The  refers to correctly prediction	72
3.10	Confusion matrix for Magahi for the LSTM-CNNs-CRF (above) and CRF (below) models; The  refers to correctly prediction	73
4.1	Overview of the MCCGS architecture with its components	78
4.2	Character level CNN model architecture to generate word vector	80

4.3	MCCGS model architecture with symbolic feature inclusion	86
4.4	Impact of morphological features and longer contextual dependencies	88
4.5	Comparison of mean accuracy scores	96
4.6	Example of models prediction to POS tagging	99
4.7	Domain adaptation	101
4.8	Accuracy comparison of SOTA and CMCCGS models	104
4.9	Accuracy comparison of the CMCCGS model with (part(i) and part(ii)) and without (none) freezing layers'	110
5.1	Residual BiLSTM-CRF model for POS tagging	114
5.2	Meta learning based training for POS tagging	116
5.3	Sanskrit POS tagger's performance as a function of the number of Hindi training examples used in the meta pretraining stage	125
5.4	Marathi POS tagger's performance as a function of the number of Hindi training examples used in the meta pretraining stage	125

List of Tables

3.1	Annotation statistics of POS tagging datasets	48
3.2	Annotation statistics of Chunking datasets	48
3.3	Feature set used for POS Tagging and Chunking in machine learning techniques and contextual boundary value is upto 3, represented by j	49
3.4	Parameters and Hyper-parameters employed during training the SAHBiLC model	51
3.5	Results of traditional machine learning techniques (%) in terms of Accuracy, Precision, Recall and F-score for POS Tagging	52
3.6	Results for traditional machine learning techniques in terms of Accuracy, Precision, Recall and F-score for Chunking	53
3.7	Results for Deep Learning techniques in terms of Accuracy, Precision, Recall and F-score for POS Tagging	55
3.8	Results for Deep Learning techniques in terms of Accuracy, Precision, Recall and F-score for Chunking	55
3.9	Language-wise dataset statistics used for annotation of named entities	63
3.10	The statistics of annotated for the three broad categories for Bhojpurī, Maithilī and Magahī	64
3.11	The statistics of annotated hierarchical entities for Bhojpurī , Maithilī and Magahī . The ENAMEX, NUMEX and TIMEX categories contained 11, 4 and 7 hierarchical named entities. ‘Other’ denotes regular words or tokens which are not named entities.	64
3.12	The dataset sizes for each language. The OOV percentage is calculated by token-type differences between test data and the training data.	65
3.13	The value of (hyper-)parameters used for training of the LSTM-CNNs-CRF model	66
3.14	NER tag-wise scores obtained by CRF and LSTM-CNNs-CRF for Bhojpurī. The metrics, which are P recision, R ecall and F ₁ -score	67
3.15	NER tag-wise scores obtained for CRF and LSTM-CNNs-CRF for Maithilī	68
3.16	NER tag-wise scores obtained for CRF and LSTM-CNNs-CRF for Magahī	69
4.1	The languages with their statistics, obtained from the UD treebank. Some languages have more than one treebank; hence the related treebank information is mentioned after - in the language name.	91
4.2	Obtained results using the MCCGS model	93

4.3	Obtained scores after inclusion of morphological features in the MCCGS model. Here, IE, CE and D stand for MCCGS-IE, MCCGS-CE and MCCGS-D models respectively.	94
4.4	Effect of the chunk size on the model performance in terms of mean accuracy	96
4.5	Obtained scores on the Indo-Aryan languages of UD treebank and HUTB dataset	98
4.6	Dataset statistics for each domain	103
4.7	Results obtained on each domain of HT dataset	105
4.8	Results obtained on PTB, TweepBank and ARK datasets	105
4.9	Domain adaptation on the HT dataset	106
4.10	Domain adaptation on ARK and TweepBank datasets	106
4.11	Results of Multi-source domain adaptation by CMCCGS model	108
4.12	Results of the domain adaptation based on freezing layers' by CMCCGS model	109
5.1	Data statistics	119
5.2	Results for few shot POS tagging	122
5.3	Results for low resource POS tagging	122
6.1	Mapping of chapters with their associated research questions	128

Abbreviations

AI	A rtificial I ntelligence
BERT	B idirectional E ncoder R epresentations from T ransformers
biLMs	bi -directional L anguage M odels
Bi-GRU	B idirectional- G ated R ecurrent U nit Network
Bi-LSTM	B idirectional- L ong S hort- T erm M emory N etwork
Bi-RNN	B idirectional- R ecurrent N eural N etwork
BMM	B hojpuri M aithili M agahi
CBOW	C ontinuous B ag- O f- W ords model
CE	C ontextual E ncoder
CLIA	C ross L ingual I nformation A ccess
CNN	C onvolutional N eural N etwork
CRF	C onditional R andom F ields
D	D isambiguation
DARPA	D efence A dvanced R esearch P rojects Agency
DLDP	D igital L anguage D iversity P roject
DNNs	D eep N eural N etworks
DsDs	D istant S upervision from D isparate S ources
ELLORA	E nabling L anguages with L Ow R esources
ELMo	E mbeddings F rom L anguage M odel
F	F -score
GloVe	G lobal V ectors
GRUs	G ated R ecurrent U nit Networks
HUTB	H indi- U rdu multi-representational T ree B ank

IE	Information E ncoder
LORELEI	L ow R esource L anguages for E mergent I ncidents
LRE	Language R ecognition
LS	Lexical S imilarity
LSTMs	Long S hort- T erm Memory N etworks
MAML	Model A gnostic M eta L earning
MCC	Matthews C orrelation C oefficient
MCCGS	Monotonic C hunk-wise attention with CNN-GRU-Softmax
MCIT	Ministry C ommunication and I nformation T echnology
MEMM	Maximum E ntropy M arkov M odel
MOCHA	M onotonic C hunk-wise A ttention
MT	Machine T ranslation
NER	Named E ntity R ecognition
NEs	Named E ntities
NIST	National I nstitute of S tandards and T echnologies
NLP	Natural L anguage P rocessing
OOV	O ut- O f- V ocabulary
P	P recision
POS	P art- O f- S peech
PSA	P osition-aware S elf A ttention
R	R ecall
REFLEX	R esearch on E nglish and F oreign L anguage E Xploitation
RNNs	Recurrent N eural N etworks
SAHBiLC	Self Attention based Hierarchical Bi-LSTM CRF
SOV	S ubject- O bject- V erb
SSVMs	Structured S upport V ector M achines
TC	T ype C onstraints
UD	Universal D ependency
UPOS	Universal P OS

Symbols

P	Probability
S	State sequence
O	Observation sequence
W	Word sequence
T	Total labels
t	Timestep
λ	Learning weight to each feature
X	Feature space
D_s	Source domain
D_t	Target domain
T_s	Source task
T_t	Target task
\mathcal{L}	Loss
C	Co-occurrence matrix
θ	Learning parameters
h	Hidden state
x	Word representation
c	Character representation
e	Energy score
ξ	Gaussian noise
η	Perturbation
γ	Weighted factor
\mathcal{F}	Feature set

τ	Task
\overrightarrow{LSTM}	Forward long short-term memory
\overleftarrow{LSTM}	Backward long short-term memory
\overrightarrow{GRU}	Forward gated recurrent unit
\overleftarrow{GRU}	Backward gated recurrent unit
CONV	Convolution operation
BiLSTM	Bidirectional LSTM