Parts of Speech (PoS) and Universal Parts of Speech (UPoS) Tagging: A Critical Review with Special Reference to Low Resource Languages

Kuwali Talukdar, Shikhar Kumar Sarma, Manash Pratim Bhuyan
Department of Information Technology, Gauhati University, India
kuwalitalukdar@gmail.com, sks001@gmail.com, mpratim250@gmail.com

Abstract

Universal Parts of Speech (UPoS) tags are parts of speech annotations used in Universal Dependencies. Universal Dependency (UD) helps in developing cross-linguistically consistent treebank annotations for multiple languages with a common framework and standard. For various Natural Language Processing (NLP) tasks and research such as semantic parsing, syntactic parsing as well as linguistic parsing, UD treebanks are becoming increasingly important resources. A lot of interest has been seen in adopting UD and UPoS standards and resources for integrating with various NLP techniques, including Machine Translations, Question Answering, Sentiment Analysis etc. Consequently, a wide variety of Artificial Intelligence (AI) and NLP tools are being created with UD and UPoS standards on board. Part of Speech (PoS) tagging is one of the fundamental NLP tasks, which labels a specific sentence or set of words in a paragraph with lexical and grammatical annotations, based on the context of the sentence. Contemporary Machine Learning (ML) and Deep Learning (DL) techniques require god quality tagged resources for training potential tagger models. Low resource languages face serious challenges here. This paper discusses about the UPoS in UD and presents a concise yet inclusive piece of literature regarding UPoS, PoS, and various taggers for multiple languages with special reference to various low resource languages. Already adopted approaches and models developed for different low resource languages are included in this review, considering representations from a wide variety of languages. Also, the study offers a comprehensive classification based on the well-known ML and DL techniques used in the development of part-of-speech taggers. This will serve as a ready-reference for understanding nuances of PoS and UPoS tagging.

Keywords

Parts of Speech, Universal Parts of Speech, Universal Dependencies, PoS Taggers, Machine Learning, Deep Learning.

1 Introduction

Natural language processing (NLP) is a complex procedure of stages of computational modules, and it belongs to the artificial intelligence field. As human languages are inherently ambiguous, it is incredibly challenging for machines to comprehend. Therefore, researchers have come up with various tools and techniques both for generation, as well as understanding/processing natural languages both for text and speech forms. NLP research includes various tasks to improve the machine understanding/generation processes such as machine translation, information retrieval, information extraction, question-answering, speech synthesis and recognition etc. To carry out these tasks, Parts-of-Speech (PoS) tagging or grammatical/lexical annotation is an important and fundamental requirement, as most of the NLP tasks involve rigorous dependencies on syntactic characteristics.

The term "parts of speech" refers to a word's grammatical characteristic and how they relate to other words in a sentence. There are eight main parts of speech in the English language, including the noun, pronoun, verb, adjective, adverb, preposition, conjunction, and interjection. With the exception of the word classes "noun" and "verb," which are present in practically every language, there are substantial differences in the parts of speech among different languages. For instance, in Assamese, an Indo Aryan language, there are eight parts of speech given in the standard grammar books namely Noun, Pronoun, Adjective, Adverb, Verb, Preposition, Conjunction and Interjection, but some other categories are also there such as Postposition, Auxiliary Verb, Pronominal, Numeral, Subordinating, Punctuation, Symbols, Unknown, Particles etc. which are required to have a proper hold on the Assamese grammar

and on the parts of speech. Universal parts of speech standard is defined as the core parts of speech categories alongwith a set of universal features used to distinguish additional lexical and grammatical properties of words. There are 17 core tags in UPoS, shown in Table 1.

	PoS	Tag
1	adjective	ADJ
2	adposition	ADP
3	adverb	ADV
4	auxiliary	AUX
5	coordinating conjunction	CCONJ
6	determiner	DET
7	interjection	INTJ
8	noun	NOUN
9	numeral	NUM
10	particle	PART
11	pronoun	PRON
12	proper noun	PROPN
13	punctuation	PUNCT
14	subordinating conjunction	SCONJ
15	symbol	SYM
16	verb	V
17	others	X

Table 1: 17 Core UPoS Categories

UPoS standard is becoming increasingly popular across the languages, as it gives a uniform set of annotation conventions, which was originally designed for cross-linguistically consistent treebank annotation for multiple languages. Apart from the 17 core universal categories, the UPoS standard defines universal features, 7 as lexical features, and 17 as inflectional features, allowing addition of a broad range of language specific features under these feature categories.

Parts of speech tagging to a word manually is tiresome and time-consuming work, thus there is growing interest in automating the tagging process (Pisceldo et al. 2009). Automatic POS tagging applies different techniques and over last many decades various models such as rule based, statistical, hybrid, machine learning and deep learning have been evolved through extensive experimentations (Antony et al.,2011). There are many performing taggers developed over time, for example, for English language there are various POS taggers– Brill tagger, Tree tagger, ENGTWOL, CLAWS tagger etc. Various innovative techniques also have been applied in order to increase performances of such taggers, like using character representation of POS tags for high positive impact resulting higher predicted POS accuracy. (Dozat et al, 2017; Smith et al., 2018).

The advantages of POS tagging has been described by Niladri Sekhar Dash (2013) which has been divided into three levels:

i. Lexical level: It looks into the surface form of word representation by analyzing its morphological structure.
ii. Orthographic level: It shows the distinction in the homographic and also in the semantic role in the same text or in other similar text.
iii. Syntactic level: In this level to assign the POS entities the syntactic-o-grammatical functions of words are identified.

PoS tagging researchers started looking at the use of Machine Learning (ML) and Deep Learning (DL) approaches in recent years, to serve the needs of effective PoS taggers to be integrated with various NLP applications and tasks. Many ML and DL-based approaches were proposed by various researchers to improve the effectiveness of PoS taggers in classifying words according to their parts of speech in their context. But the complexities associated with PoS tagging for unknown and ambiguous words have made it difficult for achieving full accuracy, and therefore various newer techniques and approaches are being explored and experimenting.

2 Types of PoS taggers and approaches

There are two types of POS tagging: supervised and unsupervised. Rule-based, Stochastic, Transformation based, and Memory based fall under supervised learning, and Neural categories including ML and DL fall under unsupervised learning.

i. Rule-based tagger: This tagger uses dictionary or lexicon where words are stored alongwith the syntactic information. If a word has more than one possible syntactic categories, the grammar based hand-crafted rules are used to find the correct tag. It is the oldest technique in tagging.
ii. Statistical tagger: This tagger is also known as stochastic or probabilistic tagger as it selects the most probable sequence of tags

from a text corpus, based on past statistical evidences.

iii. Memory-based tagger: In memory based, a set of cases are stored in memory. Each case contains the word, its context and suitable tag. It is a combination of rule based and stochastic method.

iv. Transformation based tagger: This is also a rule-based tagging method, but rules are applied in multiple cycles. Transformation based tagging is also called Brill tagging.

v. Neural tagger- It consist of series of algorithms that recognizes the relationship in the dataset through a process that mimics the way human brain works.

2.1 Machine Learning Algorithms

2.1.1 Hidden Markov Model: The most frequently used PoS tagging technique in the stochastic approach is the Hidden Markov model (Mathew et al., 2012). It is a statistical model and utilized to determine the most frequent tag sequence for a given word sequence in a phrase. When employing a Hidden Markov Model, one well-known technique for tagging the most probable tag sequence for each word in a phrase is the Viterbi algorithm.

2.1.2 Naive Bayes: Naive Bayesian Networks are a type of probabilistic network model that can be utilized to take advantage of these haphazard connections or correlations between a problem's variables (Tseng et al. 2012).

2.1.3 Conditional Random Field*:* A technique for creating discriminative probabilistic models that segment and label a set of sequential data is called a conditional random field (CRF) (Gashaw and Shashirekha, 2018). An undirected x, y graphical model known as a conditional random field is one in which each vertex denotes a random variable whose distribution is dependent on a particular observation variable.

2.1.4 Support Vector Machine: SVM is a machine learning technique that is employed in applications that need binary classification, including NLP (Surahio and Mahar 2018). In essence, an SVM algorithm learns a linear hyperplane that, with the highest boundary, divides the set of positive collections from the set of negative collections.

2.2 Deep Learning algorithms:

2.2.1 Recurrent neural network (RNN): One type of artificial neural network model where connections between the processing units take the form of cyclic routes is the recurrent neural network (RNN). They accept inputs, update the hidden layers based on previous calculations, and forecast each element of a sequence, making it recurrent (Banga and Mehndiratta, 2017).

2.2.2 Convolutional neural network: A deep learning network structure that is better suited for the data included in the array's data structure is the convolutional neural network (CNN). The CNN has an input layer, a memory stack of pooling and convolutional layers for extracting feature sets, and a fully connected layer with a softmax classifier in the classification layer, similar to other neural network designs (Gupta et al., 2020).

2.2.3 Long Short-Term Memory: An RNN network architecture with the ability to learn long-term dependencies is known as a Long Short-Term Memory (LSTM). Additionally, an LSTM can be trained to bridge time gaps in more than 1000 steps (Deshmukh and Kiwelekar, 2020).

2.2.4 Bidirectional Long Short Term Memory: Bidirectional LSTM includes two distinct hidden layers to handle input in both ways. The first hidden layer processes the forward input sequences, while the second hidden layer processes the backward input sequences. Both are coupled to the same output layer, which gives access to both the future and past context of each point in the sequence (Deshmukh and Kiwelekar, 2020).

2.2.5 Gate recurrent unit: A recurrent neural network extension called a Gated Recurrent Unit (GRU) tries to handle memories of sequences of data by storing the network's past input state and planning to target vectors depending on that input (Deshmukh and Kiwelekar, 2020).

2.2.6 Feed forward neural network: One artificial neural network in which connections between the neuron units do not form a cycle is a feed-forward neural network (FNN). Additionally, information processing in feedforward neural networks is transferred from network input levels to output layers (Anastasyev et al., 2018).

2.2.7 Deep neural network: Normal Recurrent Neural Networks (RNN) only pass input through one layer before processing it at the output layer. However, Deep Neural Networks (DNN) combine the principles of RNNs and deep neural networks (DNN) (Srivastava et al., 2018).

3 Universal Dependencies and UPoS

Prior to the 17 POS tags, there were 12 POS tags in the universal parts of speech: Noun, Verb, Adjective, Adverb, Pronoun, Determiner, Adposition, Numeral, Conjunction, and Particles. A project called Universal Dependencies (UD) uses universal parts of speech tags to create treebank annotation that is consistent across many different languages. The annotation scheme is based on an evolution of (Universal) Stanford dependencies (Marneffe et al., 2014; Slav Petrov et.al, 2012). They have opted unsupervised approach to evaluate their cross-lingual Parts-of-speech projection system for six different languages. They have also proposed 12 universal POS; using those tagset and mapping, they compared POS tag accuracies for 25 different treebank to evaluate POS tagging accuracies on a single tagset. Secondly, they combine the grammar induction system of Snyder et al. (2008) with their cross lingual project PoS tagger. Sampo pyysalo et al. (2015) have presented universal dependencies (UD) for Finnish. They have mapped according to the previously introduced annotation to the UD standard. Parsing experiment comparing the performance of a state of-the-art parser trained on a language specific annotation to perform on the corresponding UD annotation was also done. A multilingual multitask model called Udify, created by Dan Kondratyuk et al., (2017) is able to accurately anticipate Universal POS.

In order to benefit the low resource languages by the cross-linguistic annotation, they evaluated Udify for multilingual languages. Udify could generate UD annotations automatically for languages that allow UD. It also offers the best results in dependency parsing; however, it gives inaccurate results for the Lemmas and Universal features. In order to better understand how multilingual training produces UD predictions even for languages for which Udify and BERT models are not trained, they have also evaluated zero shot learning. They observed a little improvement in the BERT model that was trained on 104 languages before being fine-tuned across all datasets.

In their study, Anderson et al. (2021) tested the effectiveness of anticipated UPoS tags as inputs for dependency parsers using a low resource universal dependency treebank with varied tree bank sizes. Predicted UPoS tags have been proven helpful for treebanks with few resources, but as data volumes expand, their beneficial effects diminish. They examined (Tiedemann, 2015) and discovered that low resource parsers performed poorly even when using gold standard PoS tagged data. Even with cross-lingual approaches and a low resource setting in (Kann et al., 2020), the tagger performance is unsatisfactory. The transition-based parser for multilingual universal dependence (UD) demonstrates a slight improvement by utilizing Universal PoS tags (UPoS).

4 Taggers for various languages

Indian languages are morphologically rich, agglutinative and ambiguous, resulting challenges regularly encountered while assigning correct tags to the words as per the sentences, as the words may behave differently in different context.

An *Awngi* language part of speech tagger utilizing the Hidden Markov Model was proposed by Demilie (2019). Awngi is very low resource Ethiopian language. They gathered 94,000 sentences and used 23 individually made tags. Performance of the *Awngi* HMM POS tagger was assessed using a tenfold cross-validation mechanism. The empirical finding demonstrates that uni-gram and bi-gram taggers achieve tagging accuracy of 93.64% and 94.77%, respectively.

Atmakuri et al. (2018) have discussed different combination of features on different size corpus to develop a CRF based POS tagger.

They have achieved maximum accuracy of 89% on their largest corpus. They have compared different taggers developed by researchers for Kannada language. For Kannada, Antony P. J and Soman K.P. (Antony et al., 2011) created a PoS tagger with 30 tags using support vector machines (SVM). Using the 25 tagset PoS tags produced by Bharati and others (2006) compared two probabilistic models (HMM and CRF) on the EMILLE dataset. CRF model was the best, with 84% accuracy, they found. Without employing a probabilistic model, M. C. Padma and R. J. Pratibha (Padma and Pratibha, 2016) created another POS tagger for Kannada. They employed a rule-based strategy and used the BIS Dravidian tag set. Additionally, they ran four small datasets on a tiny fraction of the EMILLE corpus to evaluate their model, and they obtained an average precision of 88.75%. Using CRF for PoS tagging, K.P. Pallavi and Anitha S. Pillai (Pallavi and Pillai, 2016) created a tagger for Kannada that achieves a precision of 92.4%.

Kushagra et al. (2018) have developed a unique code-mixed PoS-tagged dataset of English and Hindi tweets. They have also claimed that their dataset is larger than the previous annotated dataset and closely resemble real world tweets. Additionally, they have described how their PoS tagged dataset can be used by developing and evaluating an automatic PoS-tagging model. A set of self-made features are presented which is used by the model to predict PoS tags of a token. They have developed a sequence labelling task PoS tagging using conditional Random field and LSTM recurrent neural networks. The outcome of models are- PoS CRF achieves F1 score of 90.20% and PoS LSTM achieves F1 scores of 82.51%. They also observed that above models show low performance for PoS tags when it comes to Hindi adjectives and adverb, as there are different grammatical rules for Hindi language.

Tham (2020) has developed a hybrid tagger for Khasi language integrating HMM (Hidden Markov Model) and CRF (Conditional Random Fields). As Khasi is an under resourced language without CRF integration which incorporates the language features, it is not feasible only with HMM tagger, they observed.

While developing the Khasi PoS tagger, HMM approach was used, and along with that, a tenfold cross validation was also carried out to check the performance. As per available sources on Khasi language, two HMM PoS taggers were used so far and trained and tested on two different datasets. The first dataset consisting of 86,087 tokens where HMM tagger was trained; it gave accuracy of 95.68% whereas the second one gave accuracy of 76.7% with 7,500 words and custom made tagset of 54 tags. Here the Corpus consist of 94,651 tokens and while performing the ten-fold cross validation the Baseline tagger gave accuracy of 84.05%, the NLTK tagger gave accuracy of 87.58% and HMM PoS tagger gave 93.39% which is the best among the three.

Banga and Mehndiratta (2017) evaluated different taggers on different data size. The output of all the taggers are compared on the basis of accuracy and total time required for training data. Some taggers give good performance in terms of accuracy and some in terms of computation times.

Straka and Straková (2020) have created a system that uses lemmatization and PoS tagging and is built on UDPipe 2.0. Additionally, they evaluated the BERT and XLM-RoBERTa, and their impact on the contextualized embedding and treebank encoding, contributing to the Eva Latin shared job. Some of the related task similar to EvaLatin shared task has also been discussed, such as SIGMORPHON2019 hard task (McCarthy et al., 2019) worked on 107 different Corpora and 66 languages, work by Zeman et al. (2017) and Multilingual Parsing from raw text to Universal dependencies shared tasks etc. Here, an overview of the employed architecture is covered. The multitask architecture is built on UDPipe 2.0. The embedding input words are processed through three shared bidirectional LSTM layers, and the output is subsequently processed by softmax classifiers, which produce lemmas and POS tags. All three UD 2.5 Latin treebanks are trained for

treebank embedding. To compare BERT with XLM RoBERTa, an ablation experiment was run. BERT embedding demonstrates only a slight improvement in accuracy, although XLM-RoBERTa performs better.

Alharbi et al. (2019) suggested utilizing Bi-LSTM for PoS tagging for Arabic Gulf Dialect. For the goals of sequence modelling, the Support Vector Machine (SVM) classifier and Bi-Directional Long Short Term Memory (Bi-LSTM) machine learning methods are used. With the use of a Bi-LSTM, the POS tagging model was enhanced from a 75% state-of-the-art accuracy to over 91% accuracy for Gulf dialect. Additionally, they build a dataset for PoS tagging and several sets of characteristics to test the models.

Kumar et al. (2018) in their paper have discussed about Malayalam tweets POS tagging. They have considered 17 coarse tags and 9915 tweets were tagged manually and evaluated by using the methods of Deep Learning such as Recurrent Neural Network (RNN), Gated Recurrent Units (GRU), Long Short Term Memory (LSTM) and Bidirectional LSTM (BiLSTM). Theano framework, which is a library of deep learning, was used for the experiment and evaluation. After model training and experiments at word level, GRU based deep learning gave the highest F1-measure of 0.9254 at character level, BLSTM based sequential model gave F1 measure of 0.8739. They have found that with increase in the hidden state there is improvement in the state of the tagger. The authors have mentioned that foregoing task is difficult to understand and to natural language processes such as parsing, PoS tagging, Named Entity Recognition seems to be challenging as the task differs in lexical, syntactic and orthographic patterns. Initially tweeter PoS tagging work in English was done by Gimpel et al. where a coarse PoS tagset was designed to show the main parts of the speech.

Sinha et al. (2018) presented a tagger for Chhattisgarhi language. The tagger is a hybrid tagger with combination of Rule based techniques and Statistical techniques. The tools used for database findings are statistical and conditional random field approach. Their system has two parts– first searching the word in the database and second is if the word is found then tag it. They have achieved the accuracy of 70% and if the words present in the sentences are also present in the database then the accuracy is higher than 90%.

Prabha et al. (2018) have developed a PoS tagger for Nepali language, which belongs to Indo-Aryan family. Their approach was deep learning and the methods like – RNN, LSTM, GRU and Bi-directional variants were used.

The work of Anderson and Gomez-Rodriguez (2020) shows that high tagging accuracy is required to show effective performance of UPoS tags as features for neural parsers. The author here also did analysis on the effect of UPoS accuracy on the parser performance. The impact of PoS tag on parsing performance considering different UD treebanks using gold PoS tags and also predicted PoS tags were evaluated. It has been observed that using gold PoS tags results more accuracy in predicting. Here the author has considered various UD treebanks such as Catalans, AnCora, Japanese GSD etc. and PoS taggers are trained on them using sequence labelling framework NCRF++ (Yang and Zhang, 2018) to find different accuracies. Also, two dependency parsers were trained with and without predicted PoS tags.

Specifically, deep belief networks and the Deep Learning methodology are used to create a Bengali PoS tagger. Using the corpus, they have developed a word dictionary for PoS tagging. The PoS tagging dictionary can reduce the ambiguity of tagging procedures. On the corpus, the deep learning-based Bengali PoS tagger achieves an accuracy rate of 93.33% (Patoary et al., 2020).

A DL-based PoS tagging approach for Bengali (Kabir et al. 2016) utilized the language's suffixes. An experiment is carried out using a labelled corpus of 2927 words. The accuracy of the suggested DL-based POS

tagging model was 93.90%. Additionally, the deep learning model outperformed earlier models like rule-based and global linear models in terms of accuracy. Additionally, the suggested model is integrated into the free and open-source Bengali NLP toolbox written in Python.

A HMM based tagger was also developed by Kartik et al. (2020) and they have used Indian Language standard PoS tagset, and also trained their program using 5000 sentences from the tourism domain. Also, another chunk of 4500 sentences were later on combined with the corpus to prepare the system. The HMM could pick the tag for a word by looking the pre and post words. They have achieved accuracy of 85.40%.

Lohe and Pandey (2020) in their paper described the PoS tagging processes using Rule Based approach. The author has discussed the Rule-based parts of speech tagger for Hindi language. They have also reviewed various papers based on Hindi PoS taggers.

Singh et al. (2013) have designed a PoS tagger for Marathi language using Trigram method of Statistical approach. The model provides automatic tagging of Marathi words and to derive the best sequence it checks the probability values of the previous two tags.

Sayar and Singh (2018) have discussed about PoS tagging for Hindi language. Their approach was HMM and they have used Hindi WordNet dictionary. The performance analysis was done considering the parameters such as Precision, Recall and F1 measure. They have obtained precision – 93.17%, Recall- 96.46%, F-measure- 90.13%.

Jamatia and Das (2014) have focused on developing a PoS tagging system considering Hindi and Bengali tweets. Also, they have discussed about social media text (SMT) and the challenges they have faced as SMT has different writing practices. A PoS tagger for English tweets was developed by Gimpel et al. (2011), and then combining it with Indian language standard PoS tagset (LDC-IL) the author have designed coarse grain POS tagset. Their corpus consists of 3488 tweets. For annotating their corpus, they opted for the popular and fastest crowd source service of Amazon Mechanical Turk but the outcome was poor, then they tried bootstrapping. They have tried various ML experiments on Hindi tweet PoS tagging and they also proposed to work for developing PoS tagger for code-mix SMT.

Mishra and Jain (2018) in their paper presents a PoS tagger for Hindi language using hybrid approach. Initially the authors have tagged around 1 lakh unique class category of words with the help of WordNet dictionary. For some untagged words, rule based approach is used to assign tags to the words. To remove ambiguity, HMM model is used as a statistical approach. Also, the corpus is evaluated using seven standard parts of speech tags and the accuracy rate is 92%. The PoS tagging identifies the lexical categories of Hindi words present. The corpus consists of 1000 sentences and 15000 words.

For Malayalam, a DL based PoS tagger is suggested (Akhil et al., 2020). In order to develop the PoS tagger, the studies make use of the Gated Recurrent Unit (GRU), Long Short-Term Memory (LSTM), Recurrent Neural Networks (RNN), and Bi-directional Long Short Term Memory (BiLSTM). When compared to earlier models, the suggested model performed better. This results in the model's precision, recall, and F-measure being 0.9878, 0.9788, and 0.9832 respectively.

Turkish deep neural network language models were suggested as a solution to the PoS tagging issue. Long Short-Term Memory (LSTM) and Recurrent Neural Network are used in the experiment (RNN). There is a performance comparison with cutting-edge techniques. According to the experiment's findings, LSTM outperforms RNN with an F-measure metric of 88.7% (Bahcevan et al., 2018).

Aziz and Sunitha (2015) have used hybrid approach for Parts of Speech tagging in Malayalam language. The rules set consist of 267 rules and is applied with Morph analyzer. They have achieved accuracy of 90.5%.

Since relatively little work has been done on the digitalization of the Assamese language corpora, Roy and Purkayastha (2016) have created an Assamese PoS tagged corpus for enhanced language comprehension. The Assamese tagset they used consist of 31 tags incorporating BIS tagset and 10000 words of Assamese text have been tagged manually. The customized tagset consists of 31 tags with 11 top-level categories and corresponding subtypes. In their system, an Assamese model was created using a manually tagged dataset as input to the Tnt tagger, and an Assamese parser was created using CFG and NLTK to produce the parse tree. The top-down, recursive-decent parsing method used by the parser is created by a series of recursive procedures.

Another paper by Roy and Purkayastha (2016) have discussed the various approaches for Parts of Speech tagging used by the researchers and also have given a brief overview of the computational works done by them for the upliftment of the Assamese language in terms of language technologies.

Early's Parsing method was used in the work "Parsing of Parts of Speech Tagged Assamese Texts" by Mirzanur et al (2009). They have created a method that uses Assamese sentence structure analysis to create grammatical rules. Bipul et al. discussed their work to develop an effective syntactic analyzer and annotated corpora for the Assamese language's rich morphological features. Barman et al. (2013) created an Assamese PoS tagger using CRF++ and fnTBL in their study. Their accuracy rate when utilizing CRF++ was 67.33%, and when using TBL, it was 87.17%.

Daimary et al. (2018) in their paper have discussed about developing a PoS tagger for Assamese language, using Stochastic approach based on HMM model. The model is trained on 256,690 words, and their system gives accuracy of 89.21%. Assamese language PoS tagging issues, as well as different PoS tagging methodologies and Assamese language characteristics, have been examined by Boro and Sharma (2020). A comprehensive review is also done by Kuwali and Shikhar (2023) for different approaches and techniques applied for PoS tagging of Indo Aryan languages.

5 Summary

The majority of researchers have preferred Deep Learning (DL) methods over the past ten years while constructing PoS tagging models, according to the works we have reviewed. It is observed that next frequent solutions use hybrid approaches by fusing machine learning and deep learning algorithms, and rest of the PoS tagger models are implemented based on standalone machine learning techniques and conventional rules based and statistical approaches.

This review study provides interesting and new researches information with up-to-date knowledge, recent researcher's inclinations, and development of the field by providing a thorough assessment of the part of speech tagging approaches based on deep learning (DL) and machine learning (ML) methods. Though for Indian Languages Rule-based, Transformation based, supervised and unsupervised approaches gives good performances but it has been observed that the recent state of art models viz. HMM model, Deep learning models will be suitable for further research in the UPOS based tagging. The challenging task, which was encountered while developing a tagger, is the language ambiguity and many languages are still low resource language. Our future work is to develop a POS tagger giving better performances for Assamese language.

6 Conclusion:

A comprehensive review on Parts of Speech tagging using both PoS tagset standard as well as Universal PoS tagset has been attempted in this paper. While efforts have been given to include a wide variety of approaches and techniques, special attention has been given to include works on low resource languages. Contemporary as well as relatively earlier works are included so that evolutionary impressions could be drawn

from the contents. This will serve as a ready reference for early researchers planning experimental works on sequential labelling tasks using PoS an UPoS tagsets.

References

Pisceldo F, Adriani M, and R. Manurung R 2009. *Probabilistic Part of Speech Tagging for Bahasa Indonesia*. In: Proc. 3rd Int. MALINDO Work. Coloca. event ACL-IJCNLP.

Antony PJ, Dr Sonam KP. 2011. *Parts of speech tagging for Indian languages: A literature survey* International journal of Computer application (0975-8887), Volume 34-No.8

Timothy Dozat, Peng Qi, and Christopher D Manning. 2017. Stanford's graph-based neural dependency parser at the CoNLL 2017 shared task. *In Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies,* pages 20–30.

Aaron Smith, Bernd Bohnet, Miryam de Lhoneux, Joakim Nivre, Yan Shao, and Sara Stymne. 2018. *82 treebanks, 34 models: Universal dependency parsing with multi-treebank models. In Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 113–123.

Niladri Sekhar Dash, 2013 *Parts-of-Speech (POS) Tagging of Bengali Written Text Corpus* Bhasa Bijnan o Prayukti: An International Journal on linguistics and language Technology. Vol 1,No. 1, Pp 53-96.

Mathew W, Raposo R, Martins B.2012 Predicting future locations with hidden Markov models. *In: Proceedings of the 2012 ACM conference on ubiquitous computing*, p. 911–18.

Tseng C, Patel N, Paranjape H, Lin TY, Teoh S. 2012 *Classifying Twitter Data with Naive Bayes Classifier*. In: 2012 IEEE International Conference on Granular Computing Classifying, pp. 1–6.

Gashaw I, Shashirekha H. 2018 *Machine Learning Approaches for Amharic Parts-of-speech Tagging* in Proc. of ICON-2018, Patiala, India, pp.69–74, December.

Srivastava P, Chauhan K, Aggarwal D, Shukla A, Dhar J, Jain VP 2018. *Deep learning based unsupervised POS tagging for Sanskrit*. In: Proceedings of the 2018 International Conference on Algorithms, Computing and Artificial Intelligence; pp. 1-6.

Hirpssa S, Lehal GS.2020 *POS tagging for Amharic text: a machine learning approach*. INFOCOMP.; 19(1):1–8.

Surahio FA, Mahar JA. 2018 *Prediction system for Sindhi parts of speech tags by using support vector machine.* In: 2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET); pp. 1-6

Gupta V, Juyal S, Singh GP, Killa C, Gupta N. 2020. *Emotion recognition of audio/speech data using deep learning approaches*. J. Inf. Optim Sci.; 41(6):1309–17.

Deshmukh RD, Kiwelekar A. 2020 *Deep Learning Techniques for Part of Speech Tagging by Natural Language Processing.* In: 2nd Int. Conf. Innov. Mech. Ind. Appl. ICIMIA 2020 - Conf. Proc., no. Icimia, pp. 76–81.

Anastasyev D, Gusev I, Indenbom E.2018 *Improving part-of-speech tagging via multi-task learning and character-level word representations*. Komp'juternaja Lingvistika i Intellektual'nye Tehnol., vol. 2018-May, no. 17, pp. 14–27.

Slav Petrov and Dipanjan Das, Ryan McDonald 2012 *A Universal Part-of-Speech Tagset*{A Universal Part-of-Speech Tagset}, {http://petrovi.de/data/universal.pdf},Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC '12)}

Sampo Pyysalo 1 Jenna Kanerva 1, 2 Anna Missilä 4 Veronika Laippala 3, 4 Filip Ginter 1.2015 *Universal Dependencies for Finnish*

Marneffe, M.-C & Dozat, T. & Silveira, N. & Haverinen, K. & Ginter, F. & Nivre, Joakim & Manning, C.D.. (2014). Universal Stanford Dependencies: A cross-linguistic typology. Proceedings of the 9Th International Conference on Language Resources and Evaluation (LREC). 4585-4592.

Snyder, Benjamin & Naseem, Tahira & Eisenstein, Jacob & Barzilay, Regina. (2008). Unsupervised Multilingual Learning for POS Tagging.. Proceedings of EMNLP. 1041-1050. 10.3115/1613715.1613851.

Dan Kondratyuk 1, 2 and Milan Straka. 2019. *75 Languages, 1 Model: Parsing Universal Dependencies Universally*

Mark Anderson, Mathieu Dehouck, Carlos Gómez-Rodríguez. 2021 *The Efficacy of Predicted UPOS Tags for Low Resource UD Parsing.*

Jorg Tiedemann. 2015 *Cross-lingual dependency parsing with universal dependencies and predicted pos labels*, Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015) pages 340–349.

Katharina Kann, Ophe ́lie Lacroix, and Anders Søgaard. 2020 "*Weakly supervised POS taggers perform poorly on truly low-resource languages*.",

Proceedings of the AAAI Conference on Artificial Intelli- gence,34(5):8066–8073.

Demilie WB .2019 *Parts of Speech Tagger for Awngi Language.* Int J Eng Sci Comput. 9:1

Shriya Atmakuri, Bhavya Shahi, Ashwath Rao B and Muralikrishna SN. 2018 *A comparison of features for POS tagging in Kannada*

A. Bharati, R. Sangal, D. M. Sharma, and L. Bai. 2006 Anncorra: Annotating corpora guidelines for POS and chunk annotation for Indian languages. In Technical Report (TR-LTRC-31), LTRC, IIIT-Hyderabad.

Prathibha R.J and Padma M.C., "Morpheme based parts of speech tagger for Kannada language", published in International Journal of Management and Applied Science (IJMAS), Volume 2, Issue 7, ISSN: 2394 - 7926, pp. 202-206, 2016.

K. P. Pallavi, Anitha S. Pillai, "Kannpos-Kannada Parts of Speech Tagger Using Conditional Random Fields", Emerging Research in Computing, Information, Communication and Applications. Springer, pp 479-491, 2016.

Kushagra Singh, Indira Sen, Ponnurangam Kumaraguru, 2018 *A Twitter Corpus for Hindi-English Code Mixed POS Tagging* {Association for Computational Linguistics}

Medari Janai Tham. 2020 *A Hybrid POS Tagger for Khasi, an Under Resourced Language* International Journal of Advanced Computer Science and Applications(IJACSA), Volume 11 Issue 10.

Ritu Banga, Pulkit Mehndiratta. 2017 *Tagging Efficiency Analysis on Part of Speech Tagger.* International Conference on Information Technology

Milan Straka, Jana Straková "UDPipe at EvaLatin 2020: Contextualized Embedding and Treebank Embedding"

McCarthy, A. D., Vylomova, E., Wu, S., Malaviya, C., Wolf-Sonkin, L., Nicolai, G., Kirov, C., Silfverberg, M., Mielke, S. J., Heinz, J., Cotterell, R., and Hulden, M.2019. *The SIGMORPHON 2019 Shared Task: Morphological Analysis in Context and Cross-Lingual Transfer for Inflection*. [In Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology] pages 229–244

Zeman, D., Popel, M., Straka, M., Hajicˇ, J., Nivre, J., Gin- ter, F., et al. 2017. *CoNLL 2017 Shared Task: Multi- lingual Parsing from Raw Text to Universal Dependencies*.[In Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, pages 1–19]

Alharbi R, Magdy W, Darwish K, AbdelAli A, Mubarak H. *2019 Part-of-speech tagging for Arabic Gulf dialect using Bi-LSTM*. Int Conf Lang Resour Eval. 2018;3925–3932:.

S. Kumar, M. Anand Kumar and K.P. Soman 2018 "*Deep Learning Based Part-of-Speech Tagging for Malayalam Twitter Data* .

Manish Kumar Sinha, Shubham kumar sahu, Sachin ther, 2018 *Parts Of Speech Tagging for Chhattisgarhi Language* © 2018 IJCRT | Volume 6, Issue 1 February | ISSN: 2320-2882

Greeshma Prabha, Jyotsna P V, Shahina KK, Premjith B, Soman K P. 2018 *A deep learning approach for Parts of speech tagging in Nepali Language*.

Mark Anderson, Carlos Gomez-Rodriguez, 2020 *On the Frailty of Universal POS tags for neural UD Parsers*.

Yang, Jie and Yue Zhang. "NCRF++: An Open-source Neural Sequence Labeling Toolkit." ArXiv abs/1806.05626 (2018): n. pag.

Patoary AH, Kibria MJ, Kaium A. 2020;. *Implementation of Automated Bengali Parts of Speech Tagger: An Approach Using Deep Learning Algorithm*. In: 2020 IEEE Region 10 Symposium (TENSYMP) pp. 308-311

Kabir MF, Abdullah-Al-Mamun K, Huda MN. 2016 *Deep learning based parts of speech tagger for Bengali.* In: 2016 5th International Conference on Informatics, Electronics and Vision (ICIEV); pp. 26-29.

Yadav, Kartik & Mishra, Saumya & Srivastava, Awadhesh. (2020). *A Review Paper on Part of Speech Tagger for Hindi*. International Journal of Engineering Applied Sciences and Technology. 5. 525-530. 10.33564/IJEAST.2020.v05i02.089.

Priyanka Lohe, Vikas Pandey. 2020 *Survey on Part of Speech Tagger for Hindi Language using Rule based Approach* International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 07 Issue: 11 | Nov p-ISSN: 2395-0072

Jyoti Singh, Nisheeth Joshi, Iti Mathur.2013, *Parts Of Speech Tagging Of Marathi Text using Trigram method*, international Journal of Advanced Information technology(Ijait) Vol. 3,No. 2, April

Rajesh kumar Sayar and Singh Shekhawat 2018, *Parts of speech tagging for Hindi language using HMM*.

Anupam Jamatia, Amitava Das.2014. "*Parts of speech tagging system for Indian social media text on Twitter*"

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilmann, Dani Yogatama, Jeffrey Flanagan and Noah A. Smith. *2011 Part-Of-Speech tagging for Twitter: Annotation, Features and Experiments.* [Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: short papers] paged 42-47

Nidhi Mishra, Simpal Jain.2018. *POS tagging of Hindi language using hybrid approach,* 2018. Smart and Innovative Trends in Next Generation Computing Technologies, Springer Singapore

Akhil KK, Rajimol R, Anoop VS. 2020 *Parts-of-Speech tagging for Malayalam using deep learning techniques*. Int J Inf Tech;12(3):741–8.

Bahcevan CA, Kutlu E, Yildiz T.2018 *Deep Neural Network Architecture for Part-of-Speech Tagging for Turkish Language*. UBMK 2018 - 3rd Int. Conf. Comput. Sci. Eng., pp. 235–238,

Anisha Aziz T, Sunitha C. 2015 *A Hybrid Parts of Speech Tagger for Malayalam Language*. International Conference on Advances in Computing, Communications and Informatics (ICACCI).

Bipul Roy and Bipul Syam Purkayastha 2016 *Parsing and Part-of-Speech tagging for Assamese Texts* Advances in Computer Science and Information Technology (ACSIT) 3(6); 517-521

Bipul Roy, Bipul Syam Purkayastha. 2016 *A Study on Different Part of Speech (POS) Tagging Approaches in Assamese Language* IJARCCE 5(3)934-938.

Mirzanur Rahman, Sufal Das and Utpal Sharma. 2009 *Parsing of Parts-of -Speech tagged Assamese Text*} {International Journal Of computer Science Issues,Vol 6}

Surjya Kanta Daimary, Vishal Goyal, Madhumita Barbora,.2018 *Development of Part of Speech Tagger for Assamese Using HMM* International Journal of Synthetic Emotions (IJSE), IGI Global, 9(1), 23-32, January.

Karabi Kherkatary Boro, Uzzal Sharma, 2020.*An In-depth Study on POS Tagging for Assamese Language*.AJET5, 9 (2)

Anup Kumar Barman, Jumi Sarmah, Prof. Shikhar Kr Sarma . 2013 *POS Tagging of Assamese Language and Performance Analysis of CRF++ and fnTBL approaches*.

Khan W, et al. 2019 *Part of speech tagging in urdu: comparison of machine and deep learning approaches*. IEEE Access.;7:38918–36.

Kuwali Talukdar and Shikhar Kumar Sarma, "Parts of Speech Taggers for Indo Aryan Languages: A critical Review of Approaches and Performances," 2023 4th International Conference on Computing and Communication Systems (I3CS), Shillong, India, 2023, pp. 1-6, doi: 10.1109/I3CS58314.2023.10127336.